Emerald Books

Advances in Econometrics
Volume 25

Nonparametric Econometric Methods

Qi Li
Jeffrey S. Racine
Editors

Emerald

NONPARAMETRIC ECONOMETRIC METHODS

ADVANCES IN ECONOMETRICS

Series Editors: Thomas B. Fomby and R. Carter Hill

Recent Volumes:

Volume 18:	Spatial and Spatiotemporal Econometrics, Edited by J. P. LeSage and R. Kelley Pace
Volume 19:	Applications of Artificial Intelligence in Finance and Economics, Edited by J. M. Binner, G. Kendall and S. H. Chen
Volume 20A:	Econometric Analysis of Financial and Economic Time Series, Edited by Dek Terrell and Thomas B. Fomby
Volume 20B:	Econometric Analysis of Financial and Economic Time Series, Edited by Thomas B. Fomby and Dek Terrell
Volume 21:	Modelling and Evaluating Treatment Effects in Econometrics, Edited by Daniel L. Millimet, Jeffrey A. Smith and Edward J. Vytlacil
Volume 22:	Econometrics and Risk Management, Edited by Thomas B. Fomby, Knut Solna and Jean-Pierre Fouque
Volume 23:	Bayesian Econometrics, Edited by Siddhartha Chib, William Griffiths, Gary Koop and Dek Terrell
Volume 24:	Measurement Error: Consequences, Applications and Solutions, Edited by Jane M. Binner, David L. Edgerton and Thomas Elger

ADVANCES IN ECONOMETRICS VOLUME 25

NONPARAMETRIC ECONOMETRIC METHODS

EDITED BY

QI LI
Department of Economics,
Texas A&M University

JEFFREY S. RACINE
Department of Economics,
McMaster University, Canada

United Kingdom – North America – Japan
India – Malaysia – China

Emerald Group Publishing Limited
Howard House, Wagon Lane, Bingley BD16 1WA, UK

First edition 2009

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-84950-623-6
ISSN: 0731-9053 (Series)

Awarded in recognition of Emerald's production department's adherence to quality systems and processes when preparing scholarly journals for print

INVESTOR IN PEOPLE

CONTENTS

LIST OF CONTRIBUTORS

Peter Bearse	Department of Economics, University of North Carolina at Greensboro, Greensboro, NC, USA
Zongwu Cai	Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, USA; The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian, China
Raymond J. Carroll	Department of Statistics, Texas A&M University, TX, USA
Ye Chen	Department of Economics, Princeton University, Princeton, NJ, USA
Chinman Chui	Institute for Financial and Accounting Studies, Xiamen University, China
Yanqin Fan	Department of Economics, Vanderbilt University, Nashville, TN, USA
Jingping Gu	Department of Economics, University of Arkansas, Fayetteville, AR, USA
Daniel J. Henderson	Department of Economics, State University of New York at Binghamton, Binghamton, NY, USA
Yongmiao Hong	Department of Economics, Cornell University, Ithaca, NY, USA; The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian, China

Kim P. Huynh	Department of Economics, Indiana University, Bloomington, IN, USA
David T. Jacho-Chávez	Department of Economics, Indiana University, Bloomington, IN, USA
Gaosheng Ju	Department of Economics, Texas A&M University, TX, USA
Subal C. Kumbhakar	Department of Economics, State University of New York at Binghamton, Binghamton, NY, USA
Dingding Li	Department of Economics, University of Windsor, Canada
Qi Li	Department of Economics, Texas A&M University, TX, USA
Rui Li	School of Economics and Management, Beijing University of Aeronautics and Astronautics, China
Zhongwen Liang	Department of Economics, Texas A&M University, TX, USA
Patrick McCann	Department of Statistics, Virginia Tech University, VA, USA
Sang Soo Park	Department of Economics, University of North Carolina, Chapel Hill, NC, USA
Christopher F. Parmeter	Department of Agricultural and Applied Economics, Virginia Tech University, VA, USA
Krishna P. Paudel	Department of Agricultural Economics and Agribusiness, Louisiana State University AgCenter, Baton Rouge, LA, USA
Jeffrey S. Racine	Department of Economics, McMaster University, Canada
Paul Rilstone	Department of Economics, York University, Canada

Thanasis Stengos	Department of Economics, University of Guelph, Canada
Liangjun Su	Singapore Management University, Singapore
Yiguo Sun	Department of Economics, University of Guelph, Canada
Brennan S. Thompson	Department of Economics, Ryerson University, Canada
Efhtymios G. Tsionas	Athens University of Economics and Business, Greece
Aman Ullah	Department of Economics, University of California, Riverside, CA, USA
Ximing Wu	Department of Agricultural Economics, Texas A&M University, TX, USA
Hector O. Zapata	Department of Agricultural Economics and Agribusiness, Louisiana State University AgCenter, Baton Rouge, LA, USA
Zhiyuan Zheng	Department of Economics, Virginia Tech University, VA, USA

CALL FOR PAPERS

The editors of Advances in Econometrics, a research annual published by Emerald Group Publishing Limited, are currently soliciting abstracts and papers covering applied or theoretical topics relevant to the application of maximum simulated likelihood estimation and inference. Papers chosen will appear in the volume Advances in Econometrics: Maximum Simulated Likelihood Methods and Applications (Volume 26, 2010). The volume will be edited by Professor William Greene, Department of Economics, New York University.

A special conference for contributors is planned for November 6–8, 2009 at the Lod and Carole Cook Conference Center http://cookconferencecenter.com/ on the Louisiana State University campus in Baton Rouge, Louisiana. Financial support to attend the conference will be provided to the authors chosen to present their papers at the conference. This will be the eighth such conference held by *Advances in Econometrics* on the topics of the volume. See http://www.bus.lsu.edu/hill/aie/aie_main.htm for information on the previous conferences.

The research annual's editorial policy is to publish papers that are in sufficient detail so that econometricians who are not experts in the topics of the volume will find them useful in their research. To that end, authors should provide, upon request, computer programs utilized in their papers. For more information on the *Advances in Econometrics* series and the titles and contents of its previous volumes go to http://faculty.smu.edu/tfomby/aie.htm. Please e-mail your abstracts or papers no later than August 24, 2009 to Professor Thomas B. Fomby (tfomby@smu.edu), Department of Economics, Southern Methodist University, Dallas, TX 75275 (phone: 214-768-2559, fax: 214-768-1821) or Professor R. Carter Hill (eohill@lsu.edu), Department of Economics, Louisiana State University, Baton Rouge, LA 70803 (phone: 225-578-1490; fax: 225.578.3807).

INTRODUCTION

The field of nonparametric econometrics continues to grow at an exponential rate. The field has matured significantly in the past decade, and many nonparametric techniques are now commonplace in applied research. However, many challenges remain, and the papers in this Volume address some of them.[1]

Below we present a brief overview of the papers accepted in this Volume, and we shall group the papers into six categories, namely, (1) Model identification and testing of econometric models, (2) Estimation of semiparametric models, (3) Empirical applications of nonparametric methods, (4) Copula and density estimation, (5) Computation, and (6) Surveys.

1. MODEL IDENTIFICATION AND TESTING OF ECONOMETRIC MODELS

Identification and inference are central to applied analysis, and two papers examine these issues, the first being theoretical in nature and the second being simulation based.

The evaluation of treatment effects has permeated the social sciences and is no longer confined to the medical sciences. The first paper, "Partial identification of the distribution of treatment effects and its confidence sets" by Yanqin Fan and Sang Soo Park, investigates partial identification of the distribution of treatment effects of a binary treatment under various assumptions. The authors propose nonparametric estimators of the sharp bounds and construct asymptotically uniform confidence sets for the distribution of treatment effects. They also propose bias-corrected estimators of the sharp bounds. This paper provides a complete study on partial identification of and inference for the distribution of treatment effects for randomized experiments.

The link between the magnitude of a bandwidth and the relevance of the corresponding covariate in a regression has received much deserved attention as of late. The second paper, "Cross-validated bandwidths and

significance testing'' by Christopher Parmeter, Zhiyuan Zheng, and Patrick McCann employs simulation to examine two methods for nonparametric selection of significant variables, one being a standard bootstrap-based nonparametric significance test, and the other being based on least squares cross-validation (LSCV) smoothing parameter selection. The simulation results show that the two methods perform similarly when testing for a single variable's significance, while for a joint test, the formal testing procedure appears to perform better than that based on the LSCV procedure. Their findings underscore the importance of testing for joint significance when choosing variables in a nonparametric framework.

2. ESTIMATION OF SEMIPARAMETRIC MODELS

Semiparametric models are popular in applied settings as they are relatively easy to interpret and deal directly with the curse-of-dimensionality issue. Two papers address semiparametric methods.

Panel data settings present a range of interesting problems. Linear parametric panel methods often rely on a range of devices including linear differencing for removing fixed effects and so forth. Linear models may be overly restrictive, however, while fully nonparametric methods may be unreliable due to the so-called curse-of-dimensionality. The first paper, ''Semiparametric estimation of fixed effects panel data varying coefficient models'' by Yiguo Sun, Raymond Carroll, and Dingding Li, proposes a kernel method for estimating a semiparametric varying coefficient model with fixed effects. Their method can identify an additive intercept term, while the conventional method based on first differences fails to do so. The authors establish the asymptotic normality result of the proposed estimator and also propose a procedure for testing the null hypothesis of fixed effects against the alternative of random effects varying coefficient models. They also point out that future research is warranted for reducing size distortions present in the proposed test.

The functional coefficient model constitutes a flexible approach toward semiparametric estimation, and this model nests a range of models including the linear parametric model and partially linear models, by way of example. The second paper, ''Functional coefficient estimation with both categorical and continuous data'' by Liangjun Su, Ye Chen, and Aman Ullah, considers the problem of estimating a semiparametric varying coefficient model that admits a mix of discrete and continuous covariates for stationary time series data. They establish the asymptotic normality result for the proposed local

linear estimator, and apply their procedure to analyze a wage determination equation. They detect complex interaction patterns among the regressors in the wage equation including increasing returns to education when experience is very low, high returns for workers with several years of experience, and diminishing returns when experience is high.

3. EMPIRICAL APPLICATIONS OF NONPARAMETRIC METHODS

The application of nonparametric methods to substantive problems is considered in three papers.

Though human development is an extremely broad concept, two fundamental components that receive widespread attention are health and living standards. However, much current research is based upon unconditional estimates of joint distributions. The first paper, "The evolution of the conditional joint distribution of life expectancy and per capita income growth" by Thanasis Stengos, Brennan Thompson, and Ximing Wu, examines the joint conditional distribution of health (life expectancy) and income growth and its evolution over time. Using nonparametric estimation methods the authors detect second-order stochastic dominance of the non-OECD countries over the OECD countries. They also find strong evidence of first-order stochastic dominance of the earlier years over the later ones.

Conventional wisdom dictates that there is a positive relationship between governance and economic growth. The second paper, "A nonparametric quantile analysis of growth and governance" by Kim Huynh and David Jacho-Chávez, reexamines the empirical relationship between governance and economic growth using nonparametric quantile methods. The authors detect a significant nonlinear relationship between economic growth and governance (e.g., political stability, voice, and accountability) and conclude that the empirical relationship between voice and accountability, political stability, and growth are highly nonlinear at different quantiles. They also detect heterogeneity in these effects across indicators, regions, time, and quantiles, which ought to be of interest to practitioners using parametric quantile methods.

Risk in production theory is typically analyzed under either output price uncertainty or production uncertainty (commonly known as "production risk"). Input allocation decisions in the presence of price uncertainty and production risk are key aspects of production theory. The third paper, "Nonparametric estimation of production risk and risk preference

functions'' by Subal Kumbhakar and Efthymios Tsionas, uses nonparametric kernel methods to estimate production functions, risk preference functions, and risk premium. They applied their proposed method to Norwegian salmon farming data and found that labor is risk decreasing while capital and feed are risk increasing. They conclude by identifying fruitful areas for future research, in particular, the estimation of nonparametric system models that involve cross-equation restrictions.

4. COPULA AND DENSITY ESTIMATION

The nonparametric estimation of density functions is perhaps the most popular of all nonparametric procedures. There are three papers that deal with this fundamental topic.

Copula methods are receiving much attention as of late from applied analysts. A copula is a means of expressing a multivariate distribution such that a range of dependence structures can be represented. The first paper, ''Exponential series estimation of empirical copulas with application to financial returns'' by Chinman Chui and Ximing Wu, proposes using a multivariate exponential series estimator (ESE) to estimate copula densities nonparametrically. Conventional nonparametric methods can suffer from the so-called boundary bias problem, and the authors demonstrate that the ESE method overcomes this problem. Furthermore, simulation results show that the ESE method outperforms kernel and log-spline estimators, while it also provides superior estimates of tail dependence compared to the empirical tail index coefficient that is popular in applied settings.

The nonparametric estimation of multivariate cumulative distribution functions (CDFs) has also received substantial attention as of late. The second paper, ''Nonparametric estimation and multivariate CDF with categorical and continuous data'' by Gaosheng Ju, Rui Li, and Zhongwen Liang, considers the problem of estimating a multivariate CDF with mixed continuous and discrete variables. They use the cross-validation method to select the smoothing parameters and provide the asymptotic theory for the resulting estimator. They also apply the proposed estimator to empirical data to estimate the joint CDF of the unemployment rate and city size.

The presence of boundary bias in nonparametric settings is undesirable, and a range of methods have been proposed to mitigate such bias. In a density estimation context, perhaps the most popular methods involve the use of ''boundary kernels'' and ''data reflection.'' The third paper, ''Higher order bias reduction of kernel density and density derivative estimators at

boundary points'' by Peter Bearse and Paul Rilstone, proposes a new method that can reduce the boundary bias in kernel density estimation. The asymptotic properties of the proposed method are derived and simulations are used to compare the finite-sample performance of the proposed method against several existing alternative methods.

5. COMPUTATION

Computational issues involving semiparametric and nonparametric methods can be daunting for some practitioners. In the paper ''Nonparametric and semiparametric methods in R'' by Jeffrey S. Racine, the use of the R environment for estimating nonparametric and semiparametric models is outlined. Many of the facilities in R are summarized, and a range of packages that handle semiparametric nonparametric methods are outlined. The ease with which a range of methods can be deployed by practitioners is highlighted.

6. SURVEYS

Four papers that survey recent developments in nonparametric methods are considered.

Financial data often necessitates some of the most sophisticated approaches toward estimation and inference. The first paper, ''Some recent developments in nonparametric finance'' by Zongwu Cai and Yongmiao Hong, surveys many of the important recent developments in nonparametric estimation and inference applied to financial data, and provide an overview of both continuous and discrete time processes. They focus on nonparametric estimation and testing of diffusion processes including nonparametric testing of parametric diffusion models, nonparametric pricing of derivative, and nonparametric predictability of asset returns. The authors conclude that much theoretical and empirical research remains to be done in this area, and they identify a set of topics that are deserving of attention.

The ability to impose constraints in nonparametric settings has received much attention as of late. The second paper, ''Imposing economic constraints in nonparametric regression: survey, implementation, and extension'' by Daniel Henderson and Christopher Parmeter, surveys recent developments on the nonparametric estimation of regression models under constraints such as convexity, homogeneity, and monotonicity. Their survey includes isotonic regression, constrained splines, Matzkin's approach, data

rearrangement, data sharpening, and constraint weighted bootstrapping. They focus on the computational implementation under linear constraints, and then discuss extensions that allow for nonlinear constraints.

Simon Kuznets proposed a theory stating that, over time, economic inequality increases while a country is developing and then decreases when a critical level of average income is attained. Researchers allege that the "Kuznets curve" (inverted U shape) also appears in the environment. The environmental Kuznets curve estimation literature is vast, and conflicting evidence exists on its empirical validity. The third paper, "Functional form of the environmental Kuznets curve" by Hector Zapata and Krishna Paudel, provides an overview of recent developments on testing functional forms with semiparametric and nonparametric methods, and then discusses applications employing semiparametric and nonparametric methods to examine the relationship between environmental pollution and economic growth.

A number of recent advances in nonparametric estimation and inference have extended the reach of these methods, particularly for practitioners. The fourth paper, "Some recent developments on nonparametric econometrics" by Zongwu Cai, Jingping Gu, and Qi Li, provides a selected review of nonparametric estimation and testing of econometric models. They summarize the recent developments on (i) nonparametric regression models with mixed discrete and continuous data, (ii) nonparametric models with nonstationary data, (iii) nonparametric models with instrumental variables, and (iv) nonparametric estimation of conditional quantile functions. They also identify a number of open research problems that are deserving of attention.

NOTE

1. The papers in this Volume of Advances in Econometrics were presented initially at the 7th Annual Advances in Econometrics Conference held on the LSU campus in Baton Rouge Louisiana during November 14–16 2008. The theme of the conference was "Nonparametric Econometric Methods" and the editors would like to acknowledge generous financial support provided by the LSU Department of Economics, the Division of Economic Development and Forecasting, and the LSU Department of Agricultural Economics and Agribusiness.

Qi Li
Jeffrey S. Racine

PART I
MODEL IDENTIFICATION AND TESTING OF ECONOMETRIC MODELS

PARTIAL IDENTIFICATION OF THE DISTRIBUTION OF TREATMENT EFFECTS AND ITS CONFIDENCE SETS

Yanqin Fan and Sang Soo Park

ABSTRACT

In this paper, we study partial identification of the distribution of treatment effects of a binary treatment for ideal randomized experiments, ideal randomized experiments with a known value of a dependence measure, and for data satisfying the selection-on-observables assumption, respectively. For ideal randomized experiments, (i) we propose nonparametric estimators of the sharp bounds on the distribution of treatment effects and construct asymptotically valid confidence sets for the distribution of treatment effects; (ii) we propose bias-corrected estimators of the sharp bounds on the distribution of treatment effects; and (iii) we investigate finite sample performances of the proposed confidence sets and the bias-corrected estimators via simulation.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 3–70

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025004

1. INTRODUCTION

Evaluating the effect of a treatment or a social program is important in diverse disciplines including the social and medical sciences. The central problem in the evaluation of a treatment is that any potential outcome that program participants would have received without the treatment is not observed. Because of this missing data problem, most work in the treatment effect literature has focused on the evaluation of various average treatment effects such as the mean of treatment effects. See Lee (2005), Abbring and Heckman (2007), Heckman and Vytlacil (2007a, 2007b) for discussions and references. However, empirical evidence strongly suggests that treatment effect heterogeneity prevails in many experiments and various interesting effects of the treatment are missed by the average treatment effects alone. See Djebbari and Smith (2008) who studied heterogeneous program impacts in social experiments such as PROGRESA; Black, Smith, Berger, and Noel (2003) who evaluated the Worker Profiling and Reemployment Services system; and Bitler, Gelbach, and Hoynes (2006) who studied the welfare effect of the change from Aid to Families with Dependent Children (AFDC) to Temporary Assistance for Needy Families (TANF) programs. Other work focusing on treatment effect heterogeneity includes Heckman and Robb (1985), Manski (1990), Imbens and Rubin (1997), Lalonde (1995), Dehejia (1997), Heckman and Smith (1993), Heckman, Smith, and Clements (1997), Lechner (1999), and Abadie, Angrist, and Imbens (2002).

When responses to treatment differ among otherwise observationally equivalent subjects, the entire distribution of the treatment effects or other features of the treatment effects than its mean may be of interest. Two general approaches have been proposed in the literature to study the distribution of treatment effects. In the first approach, the distribution of treatment effects is partially identified, see Manski (1997a,1997b), Fan and Park (2010), Fan and Wu (2007), Fan (2008), and Firpo and Ridder (2008). Assuming monotone treatment response, Manski (1997a) developed sharp bounds on the distribution of treatment effects, while (i) assuming the availability of ideal randomized data,[1] Fan and Park (2010) developed estimation and inference tools for the sharp bounds on the distribution of treatment effects and (ii) assuming that data satisfy the selection-on-observables or the strong ignorability assumption, Fan and Park (2010) and Firpo and Ridder (2008) established sharp bounds on the distribution of treatment effects and Fan (2008) proposed nonparametric estimators of the sharp bounds and constructed asymptotically valid confidence sets (CSs) for the distribution of treatment effects. In the context of switching regimes

models, Fan and Wu (2007) studied partial identification and inference for conditional distributions of treatment effects. In the second approach, restrictions are imposed on the dependence structure between the potential outcomes such that distributions of the treatment effects are point identified, see, for example, Heckman et al. (1997), Biddle, Boden, and Reville (2003), Carneiro, Hansen, and Heckman (2003), Aakvik, Heckman, and Vytlacil (2005), and Abbring and Heckman (2007), among others. In addition to the distribution of treatment effects, Fan and Park (2007b) studied partial identification of and inference for the quantile of treatment effects for randomized experiments; Fan and Zhu (2009) investigated partial identification of and inference for a general class of functionals of the joint distribution of potential outcomes including the correlation coefficient between the potential outcomes and many commonly used inequality measures of the distribution of treatment effects under the selection-on-observables assumption. Firpo and Ridder (2008) also presented some partial identification results for functionals of the distribution of treatment effects under the selection-on-observables assumption.

The objective of this paper is threefold. First, this paper provides a review of existing results on partial identification of the distribution of treatment effects in Fan and Park (2010) and establishes similar results for randomized experiments when the value of a dependence measure between the potential outcomes such as Kendall's τ is known. Second, this paper relaxes two strong assumptions used in Fan and Park (2010) to derive the asymptotic distributions of nonparametric estimators of sharp bounds on the distribution of treatment effects and constructs asymptotically valid CSs for the distribution of treatment effects. Third, as evidenced in the simulation results presented in Fan and Park (2010), the simple plug-in nonparametric estimators of the sharp bounds on the distribution of treatment effects tend to have upward/downward bias in finite samples. In this paper, we confirm this analytically and construct bias-corrected estimators of these bounds. We present an extensive simulation study of finite sample performances of the proposed CSs and of the bias-corrected estimators. The issue of constructing CSs for the distribution of treatment effects belongs to the recently fast growing area of inference for partially identified parameters, see for example, Imbens and Manski (2004), Bugni (2007), Canay (2007), Chernozhukov, Hong, and Tamer (2007), Galichon and Henry (2009), Horowitz and Manski (2000), Romano and Shaikh (2008), Stoye (2009), Rosen (2008), Soares (2006), Beresteanu and Molinari (2008), Andrews (2000), Andrews and Guggenberger (2007), Andrews and Soares (2007), Fan and Park (2007a), and Moon and Schorfheide (2007). Like Fan and Park

(2007b), we follow the general approach developed in Andrews and Guggenberger (2005a, 2005b, 2005c, 2007) for nonregular models.

The rest of this paper is organized as follows. In Section 2, we review sharp bounds on the distribution of treatment effects and related results for randomized experiments in Fan and Park (2010). In Section 3, we present improved bounds when additional information is available. In Section 4, we first revisit the nonparametric estimators of the distribution bounds proposed in Fan and Park (2010) and their asymptotic properties. Motivated by the restrictive nature of the unique, interior assumption of the sup and inf in Fan and Park (2010), we then provide asymptotic properties of the estimators with a weaker assumption. Section 5 constructs asymptotically valid CSs for the bounds and the true distribution of treatment effects under much weaker assumptions than those in Fan and Park (2010). Section 6 provides bias-corrected estimators of the sharp bounds in Fan and Park (2010). Results from an extensive simulation study are provided in Section 7. Section 8 concludes. Some technical proofs are collected in Appendix A. Appendix B presents expressions for the sharp bounds on the distribution of treatment effects in Fan and Park (2010) for certain known marginal distributions.

Throughout the paper, we use $\Rightarrow$ to denote weak convergence. All the limits are taken as the sample size goes to ∞.

2. SHARP BOUNDS ON THE DISTRIBUTION OF TREATMENT EFFECTS AND BOUNDS ON ITS *D*-PARAMETERS FOR RANDOMIZED EXPERIMENTS

In this section, we review the partial identification results in Fan and Park (2010). Consider a randomized experiment with a binary treatment and continuous outcomes. Let Y_1 denote the potential outcome from receiving the treatment and Y_0 the potential outcome without receiving the treatment. Let $F(y_1, y_0)$ denote the joint distribution of Y_1, Y_0 with marginals $F_1(\cdot)$ and $F_0(\cdot)$, respectively. It is well known that with randomized data, the marginal distribution functions $F_1(\cdot)$ and $F_0(\cdot)$ are identified, but the joint distribution function $F(y_1, y_0)$ is not identified. The characterization theorem of Sklar (1959) implies that there exists a copula[2] $C(u, v)$: $(u, v) \in [0,1]^2$ such that $F(y_1, y_0) = C(F_1(y_1), F_0(y_0))$ for all y_1, y_0. Conversely, for any marginal distributions $F_1(\cdot)$, $F_0(\cdot)$ and any copula function C, the function $C(F_1(y_1), F_0(y_0))$ is a bivariate distribution function with given

marginal distributions F_1, F_0. This theorem provides the theoretical foundation for the widespread use of the copula approach in generating multivariate distributions from univariate distributions. For reviews, see Joe (1997) and Nelsen (1999). Since copulas connect multivariate distributions to marginal distributions, the copula approach provides a natural way to study the joint distribution of potential outcomes and the distribution of treatment effects when the marginal distributions are identified.

For $(u, v) \in [0, 1]^2$, let $C^{\mathrm{L}}(u, v) = \max(u + v - 1, 0)$ and $C^{\mathrm{U}}(u, v) = \min(u, v)$ denote the Fréchet–Hoeffding lower and upper bounds for a copula, that is, $C^{\mathrm{L}}(u, v) \leq C(u, v) \leq C^{\mathrm{U}}(u, v)$. Then for any (y_1, y_0), the following inequality holds:

$$C^{\mathrm{L}}(F_1(y_1), F_0(y_0)) \leq F(y_1, y_0) \leq C^{\mathrm{U}}(F_1(y_1), F_0(y_0)) \tag{1}$$

The bivariate distribution functions $C^{\mathrm{L}}(F_1(y_1), F_0(y_0))$ and $C^{\mathrm{U}}(F_1(y_1), F_0(y_0))$ are referred to as the Fréchet–Hoeffding lower and upper bounds for bivariate distribution functions with fixed marginal distributions F_1 and F_0. They are distributions of perfectly negatively dependent and perfectly positively dependent random variables, respectively, see Nelsen (1999) for more discussions.

For randomized experiments, the marginals F_1 and F_0 are identified and Eq. (1) partially identifies $F(y_1, y_0)$. See Heckman and Smith (1993), Heckman et al. (1997), Manski (1997b), and Fan and Wu (2007) for applications of Eq. (1) in the context of program evaluation. Lee (2002) used Eq. (1) to bound correlation coefficients in sample selection models.

2.1. Sharp Bounds on the Distribution of Treatment Effects

Let $\Delta = Y_1 - Y_0$ denote the individual treatment effect and $F_\Delta(\cdot)$ its distribution function. For randomized experiments, the marginals F_1 and F_0 are identified. Given F_1 and F_0, sharp bounds on the distribution of Δ can be found in Williamson and Downs (1990).

Lemma 1. Let

$$F^{\mathrm{L}}(\delta) = \max\left(\sup_y\{F_1(y) - F_0(y - \delta)\}, 0\right) \text{ and}$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\left(\inf_y\{F_1(y) - F_0(y - \delta)\}, 0\right)$$

Then $F^{\mathrm{L}}(\delta) \leq F_\Delta(\delta) \leq F^{\mathrm{U}}(\delta)$.

At any given value of δ, the bounds $(F^{\mathrm{L}}(\delta), F^{\mathrm{U}}(\delta))$ are informative on the value of $F_\Delta(\delta)$ as long as $[F^{\mathrm{L}}(\delta), F^{\mathrm{U}}(\delta)] \subset [0,1]$ in which case, we say $F_\Delta(\delta)$ is partially identified. Viewed as an inequality among all possible distribution functions, the sharp bounds $F^{\mathrm{L}}(\delta)$ and $F^{\mathrm{U}}(\delta)$ cannot be improved, because it is easy to show that if either F_1 or F_0 is the degenerate distribution at a finite value, then for all δ, we have $F^{\mathrm{L}}(\delta) = F_\Delta(\delta) = F^{\mathrm{U}}(\delta)$. In fact, given any pair of distribution functions F_1 and F_0, the inequality: $F^{\mathrm{L}}(\delta) \leq F_\Delta(\delta) \leq F^{\mathrm{U}}(\delta)$ cannot be improved, that is, the bounds $F^{\mathrm{L}}(\delta)$ and $F^{\mathrm{U}}(\delta)$ for $F_\Delta(\delta)$ are point-wise best-possible, see Frank, Nelsen, and Schweizer (1987) for a proof of this for a sum of random variables and Williamson and Downs (1990) for a general operation on two random variables.

Let $\gtrsim_{\mathrm{FSD}}$ and $\gtrsim_{\mathrm{SSD}}$ denote the first-order and second-order stochastic dominance relations, that is, for two distribution functions G and H,

$$G \gtrsim_{\mathrm{FSD}} H \text{ iff } G(x) \leq H(x) \text{ for all } x$$
$$G \gtrsim_{\mathrm{SSD}} H \text{ iff } \int_{-\infty}^{x} G(v)dv \leq \int_{-\infty}^{x} H(d)dv \text{ for all } x$$

Lemma 1 implies: $F^{\mathrm{L}} \gtrsim_{\mathrm{FSD}} F_\Delta \gtrsim_{\mathrm{FSD}} F^{\mathrm{U}}$. We note that unlike sharp bounds on the joint distribution of Y_1, Y_0, sharp bounds on the distribution of Δ are not reached at the Fréchet–Hoeffding lower and upper bounds for the distribution of Y_1, Y_0. Let Y_1', Y_0' be perfectly positively dependent and have the same marginal distributions as Y_1, Y_0, respectively. Let $\Delta' = Y_1' - Y_0'$. Then the distribution of Δ' is given by:

$$F_{\Delta'}(\delta) = E1\{Y_1' - Y_0' \leq \delta\} = \int_0^1 1\{F_1^{-1}(u) - F_0^{-1}(u) \leq \delta\}du$$

where $1\{\cdot\}$ is the indicator function the value of which is 1 if the argument is true, 0 otherwise. Similarly, let Y_1'', Y_0'' be perfectly negatively dependent and have the same marginal distributions as Y_1, Y_0, respectively. Let $\Delta'' = Y_1'' - Y_0''$. Then the distribution of Δ'' is given by:

$$F_{\Delta''}(\delta) = E1\{Y_1'' - Y_0'' \leq \delta\} = \int_0^1 1\{F_1^{-1}(u) - F_0^{-1}(1-u) \leq \delta\}du$$

Interestingly, we show in the next lemma that there exists a second-order stochastic dominance relation among the three distributions $F_\Delta, F_{\Delta'}, F_{\Delta''}$.

Lemma 2. Let $F_\Delta, F_{\Delta'}, F_{\Delta''}$ be defined as above. Then $F_{\Delta'} \gtrsim_{\mathrm{SSD}} F_\Delta \gtrsim_{\mathrm{SSD}} F_{\Delta''}$.

Theorem 1 in Stoye (2008), see also Tesfatsion (1976), shows that $F_{\Delta'} \gtrsim_{\text{SSD}} F_{\Delta}$ is equivalent to $E[U(\Delta')] \leq E[U(\Delta)]$ or $E[U(Y'_1 - Y'_0)] \leq E[U(Y_1 - Y_0)]$ for every convex real-valued function U. Corollary 2.3 in Tchen (1980) implies the conclusion of Lemma 2, see also Cambanis, Simons, and Stout (1976).

2.2. *Bounds on D*-Parameters

The sharp bounds on the treatment effect distribution implies bounds on the class of "*D*-parameters" introduced in Manski (1997a), see also Manski (2003). One example of "*D*-parameters" is any quantile of the distribution. Stoye (2008) introduced another class of parameters, which measure the dispersion of a distribution, including the variance of the distribution. In this section, we show that sharp bounds can be placed on any dispersion or spread parameter of the treatment effect distribution in this class. For convenience, we restate the definitions of both classes of parameters from Stoye (2008). He refers to the class of "*D*-parameters" as the class of "D_1-parameters."

Definition 1. A population statistic θ is a D_1-parameter, if it increases weakly with first-order stochastic dominance, that is, $F \gtrsim_{\text{FSD}} G$ implies $\theta(F) \geq \theta(G)$.

Obviously if θ is a D_1-parameter, then Lemma 1 implies: $\theta(F^{\text{L}}) \geq \theta(F_{\Delta}) \geq \theta(F^{\text{U}})$. In general, the bounds $\theta(F^{\text{L}}), \theta(F^{\text{U}})$ on a D_1-parameter may not be sharp, as the bounds in Lemma 1 are point-wise sharp, but not uniformly sharp, see Firpo and Ridder (2008) for a detailed discussion on this issue. In the special case where θ is a quantile of the treatment effect distribution, the bounds $\theta(F^{\text{L}}), \theta(F^{\text{U}})$ are known to be sharp and can be expressed in terms of the quantile functions of the marginal distributions of the potential outcomes. Specially, let $G^{-1}(u)$ denote the generalized inverse of a nondecreasing function G, that is, $G^{-1}(u) = \inf\{x|G(x) \geq u\}$. Then Lemma 1 implies: for $0 \leq q \leq 1, (F^{\text{U}})^{-1}(q) \leq F_{\Delta}^{-1}(q) \leq (F^{\text{L}})^{-1}(q)$ and the bounds are known to be sharp. For the quantile function of a distribution of a sum of two random variables, expressions for its sharp bounds in terms of quantile functions of the marginal distributions are first established in Makarov (1981). They can also be established via the duality theorem, see Schweizer and Sklar (1983). Using the same tool, one can establish the following expressions for sharp bounds on the quantile function of the distribution of treatment effects, see Williamson and Downs (1990).

Lemma 3. For $0 \le q \le 1, (F^{\mathrm{U}})^{-1}(q) \le F_{\Delta}^{-1}(q) \le (F^{\mathrm{L}})^{-1}(q)$, where

$$(F^{\mathrm{L}})^{-1}(q) = \begin{cases} \inf_{u\in[q,1]}[F_1^{-1}(u) - F_0^{-1}(u-q)] & \text{if } q \neq 0 \\ F_1^{-1}(0) - F_0^{-1}(1) & \text{if } q = 0 \end{cases}$$

$$(F^{U})^{-1}(q) = \begin{cases} \sup_{u\in[0,q]}[F_1^{-1}(u) - F_0^{-1}(1+u-q)] & \text{if } q \neq 1 \\ F_1^{-1}(1) - F_0^{-1}(0) & \text{if } q = 1 \end{cases}$$

Like sharp bounds on the distribution of treatment effects, sharp bounds on the quantile function of Δ are not reached at the Fréchet–Hoeffding bounds for the distribution of (Y_1, Y_0). The following lemma provides simple expressions for the quantile functions of treatment effects when the potential outcomes are either perfectly positively dependent or perfectly negatively dependent.

Lemma 4. For $q \in [0,1]$, we have (i) $F_{\Delta'}^{-1}(q) = [F_1^{-1}(q) - F_0^{-1}(q)]$ if $[F_1^{-1}(q) - F_0^{-1}(q)]$ is an increasing function of q; (ii) $F_{\Delta''}^{-1}(q) = [F_1^{-1}(q) - F_0^{-1}(1-q)]$.

The proof of Lemma 4 follows that of the proof of Proposition 3.1 in Embrechts, Hoeing, and Juri (2003). In particular, they showed that for a real-valued random variable Z and a function φ increasing and left continuous on the range of Z, it holds that the quantile of $\varphi(Z)$ at quantile level q is given by $\varphi(F_Z^{-1}(q))$, where F_Z is the distribution function of Z. For (i), we note that $F_{\Delta'}^{-1}(q)$ equals the quantile of $[F_1^{-1}(U) - F_0^{-1}(U)]$, where U is a uniform random variable on [0,1]. Let $\varphi(U) = F_1^{-1}(U) - F_0^{-1}(U)$. Then $F_{\Delta'}^{-1}(q) = \varphi(q) = F_1^{-1}(q) - F_0^{-1}(q)$ provided that $\varphi(U)$ is an increasing function of U. For (ii), let $\varphi(U) = F_1^{-1}(U) - F_0^{-1}(1-U)$. Then $F_{\Delta''}^{-1}(q)$ equals the quantile of $\varphi(U)$. Since $\varphi(U)$ is always increasing in this case, we get $F_{\Delta''}^{-1}(q) = \varphi(q)$.

Note that the condition in (i) is a necessary condition; without this condition, $[F_1^{-1}(q) - F_0^{-1}(q)]$ can fail to be a quantile function. Doksum (1974) and Lehmann (1974) used $[F_1^{-1}(F_0(y_0)) - y_0]$ to measure treatment effects. Recently, $[F_1^{-1}(q) - F_0^{-1}(q)]$ has been used to study treatment effects heterogeneity and is referred to as the quantile treatment effects (QTE), see for example, Heckman et al. (1997), Abadie et al. (2002), Chernozhukov and Hansen (2005), Firpo (2007), Firpo and Ridder (2008), and Imbens and Newey (2009), among others, for more discussion and references on the estimation of QTE. Manski (1997a) referred to QTE as ΔD-parameters and the quantile of the treatment effect distribution as $D\Delta$-parameters.

Assuming monotone treatment response, Manski (1997a) provided sharp bounds on the quantile of the treatment effect distribution.

It is interesting to note that Lemma 4 (i) shows that QTE equals the quantile function of the treatment effects only when the two potential outcomes are perfectly positively dependent AND QTE is increasing in q. Example 1 below illustrates a case where QTE is decreasing in q and hence is not the same as the quantile function of the treatment effects even when the potential outcomes are perfectly positively dependent. In contrast to QTE, the quantile of the treatment effect distribution is not identified, but can be bounded, see Lemma 3. At any given quantile level, the lower quantile bound $(F^{\mathrm{U}})^{-1}(q)$ is the smallest outcome gain (worst case) regardless of the dependence structure between the potential outcomes and should be useful to policy makers. For example, $(F^{\mathrm{U}})^{-1}(0.5)$ is the minimum gain of at least half of the population.

Definition 2. A population statistic θ is a D_2-parameter, if it increases weakly with second-order stochastic dominance, that is, $F \gtrsim_{\mathrm{SSD}} G$ implies $\theta(F) \geq \theta(G)$.

If θ is a D_2-parameter, then Lemma 2 implies $\theta(F_{\Delta'}) \leq \theta(F_{\Delta}) \leq \theta(F_{\Delta''})$. Stoye (2008) defined the class of D_2-parameters in terms of mean-preserving spread. Since the mean of Δ is identified in our context, the two definitions lead to the same class of D_2-parameters. In contrast to D_1-parameters of the treatment effect distribution, the above bounds on D_2-parameters of the treatment effect distribution are reached when the potential outcomes are perfectly dependent on each other and they are known to be sharp. For a general functional of F_{Δ}, Firpo and Ridder (2008) investigated the possibility of obtaining its bounds that are tighter than the bounds implied by F^{L}, F^{U}. Here we point out that for the class of D_2-parameters of F_{Δ}, their sharp bounds are available. One example of D_2-parameters is the variance of the treatment effect Δ. Using results in Cambanis et al. (1976), Heckman et al. (1997) provided sharp bounds on the variance of Δ for randomized experiments and proposed a test for the common effect model by testing the value of the lower bound of the variance of Δ. Stoye (2008) presents many other examples of D_2-parameters, including many well-known inequality and risk measures.

2.3. An Illustrative Example: Example 1

In this subsection, we provide explicit expressions for sharp bounds on the distribution of treatment effects and its quantiles when $Y_1 \sim N(\mu_1, \sigma_1^2)$ and

$Y_0 \sim N(\mu_0, \sigma_0^2)$. In addition, we provide explicit expressions for the distribution of treatment effects and its quantiles when the potential outcomes are perfectly positively dependent, perfectly negatively dependent, and independent.

2.3.1. Distribution Bounds

Explicit expressions for sharp bounds on the distribution of a sum of two random variables are available for the case where both random variables have the same distribution which includes the uniform, the normal, the Cauchy, and the exponential families, see Alsina (1981), Frank et al. (1987), and Denuit, Genest, and Marceau (1999). Using Lemma 1, we now derive sharp bounds on the distribution of $\Delta = Y_1 - Y_0$.

First consider the case $\sigma_1 = \sigma_0 = \sigma$. Let $\Phi(\cdot)$ denote the distribution function of the standard normal distribution. Simple algebra shows

$$\sup_y \{F_1(y) - F_0(y-\delta)\} = 2\Phi\left(\frac{\delta - (\mu_1 - \mu_0)}{2\sigma}\right) - 1 \text{ for } \delta > \mu_1 - \mu_0,$$

$$\inf_y \{F_1(y) - F_0(y-\delta)\} = 2\Phi\left(\frac{\delta - (\mu_1 - \mu_0)}{2\sigma}\right) - 1 \text{ for } \delta < \mu_1 - \mu_0$$

Hence,

$$F^{\mathrm{L}}(\delta) = \begin{cases} 0, & \text{if } \delta < \mu_1 - \mu_0 \\ 2\Phi\left(\dfrac{\delta - (\mu_1 - \mu_0)}{2\sigma}\right) - 1, & \text{if } \delta \geq \mu_1 - \mu_0 \end{cases} \tag{2}$$

$$F^{\mathrm{U}}(\delta) = \begin{cases} 2\Phi\left(\dfrac{\delta - (\mu_1 - \mu_0)}{2\sigma}\right) & \text{if } \delta < \mu_1 - \mu_0 \\ 1, & \text{if } \delta \geq \mu_1 - \mu_0 \end{cases} \tag{3}$$

When[3] $\sigma_1 \neq \sigma_0$, we get

$$\sup_y \{F_1(y) - F_0(y-\delta)\} = \Phi\left(\frac{\sigma_1 s - \sigma_0 t}{\sigma_1^2 - \sigma_0^2}\right) + \Phi\left(\frac{\sigma_1 t - \sigma_0 s}{\sigma_1^2 - \sigma_0^2}\right) - 1$$

$$\inf_y \{F_1(y) - F_0(y-\delta)\} = \Phi\left(\frac{\sigma_1 s + \sigma_0 t}{\sigma_1^2 - \sigma_0^2}\right) - \Phi\left(\frac{\sigma_1 t + \sigma_0 s}{\sigma_1^2 - \sigma_0^2}\right) + 1$$

where $s = \delta - (\mu_1 - \mu_0)$ and $t = \sqrt{s^2 + (\sigma_1^2 - \sigma_0^2)\ \ln(\sigma_1^2/\sigma_0^2)}$. For any δ, one can show that $\sup_y\{F_1(y) - F_0(y-\delta)\} > 0$ and $\inf_y\{F_1(y) - F_0(y-\delta)\} < 0$. As a result,

$$F^{\mathrm{L}}(\delta) = \Phi\left(\frac{\sigma_1 s - \sigma_0 t}{\sigma_1^2 - \sigma_0^2}\right) + \Phi\left(\frac{\sigma_1 t - \sigma_0 s}{\sigma_1^2 - \sigma_0^2}\right) - 1$$

$$F^{\mathrm{U}}(\delta) = \Phi\left(\frac{\sigma_1 s + \sigma_0 t}{\sigma_1^2 - \sigma_0^2}\right) + \Phi\left(\frac{\sigma_1 t + \sigma_0 s}{\sigma_1^2 - \sigma_0^2}\right) + 1$$

For comparison purposes, we provide expressions for the distribution F_Δ in three special cases.

Case I. Perfect positive dependence. In this case, Y_0 and Y_1 satisfy $Y_0 = \mu_0 + (\sigma_0/\sigma_1)Y_1 - (\sigma_0/\sigma_1)\mu_1$. Therefore,

$$\Delta = \begin{cases} \left(\dfrac{\sigma_1 - \sigma_0}{\sigma_1}\right) Y_1 + \left(\dfrac{\sigma_0}{\sigma_1}\mu_1 - \mu_0\right), & \text{if } \sigma_1 \neq \sigma_0 \\ \mu_1 - \mu_0, & \text{if } \sigma_1 = \sigma_0 \end{cases}$$

If $\sigma_1 = \sigma_0$, then

$$F_\Delta(\delta) = \begin{cases} 0 \text{ and } \delta < \mu_1 - \mu_0 \\ 1 \text{ and } \mu_1 - \mu_0 \leq \delta \end{cases} \tag{4}$$

If $\sigma_1 \neq \sigma_0$, then

$$F_\Delta(\delta) = \Phi\left(\frac{\delta - (\mu_1 - \mu_0)}{|\sigma_1 - \sigma_0|}\right)$$

Case II. Perfect negative dependence. In this case, we have $Y_0 = \mu_0 - (\sigma_0/\sigma_1)Y_1 + (\sigma_0/\sigma_1)\mu_1$. Hence,

$$\Delta = \frac{\sigma_1 + \sigma_0}{\sigma_1} Y_1 - \left(\frac{\sigma_0}{\sigma_1}\mu_1 + \mu_0\right)$$

$$F_\Delta(\delta) = \Phi\left(\frac{\delta - (\mu_1 - \mu_0)}{\sigma_1 + \sigma_0}\right)$$

Case III. Independence. This yields

$$F_{\Delta}(\delta) = \Phi\left(\frac{\delta - (\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \tag{5}$$

Fig. 1 below plots the bounds on the distribution F_{Δ} (denoted by F_L and F_U) and the distribution F_{Δ} corresponding to perfect positive dependence, perfect negative dependence, and independence (denoted by F_PPD, F_PND, and F_IND, respectively) of potential outcomes for the case $Y_1 \sim N(2,2)$ and $Y_0 \sim N(1,1)$. For notational compactness, we use (F_1, F_0) to signify $Y_1 \sim F_1$ and $Y_0 \sim F_0$ throughout the rest of this paper.

First, we observe from Fig. 1 that the bounds in this case are informative at all values of δ and are more informative in the tails of the distribution F_{Δ} than in the middle. In addition, Fig. 1 indicates that the distribution of the treatment effects for perfectly positively dependent potential outcomes is most concentrated around its mean 1 implied by the second-order stochastic

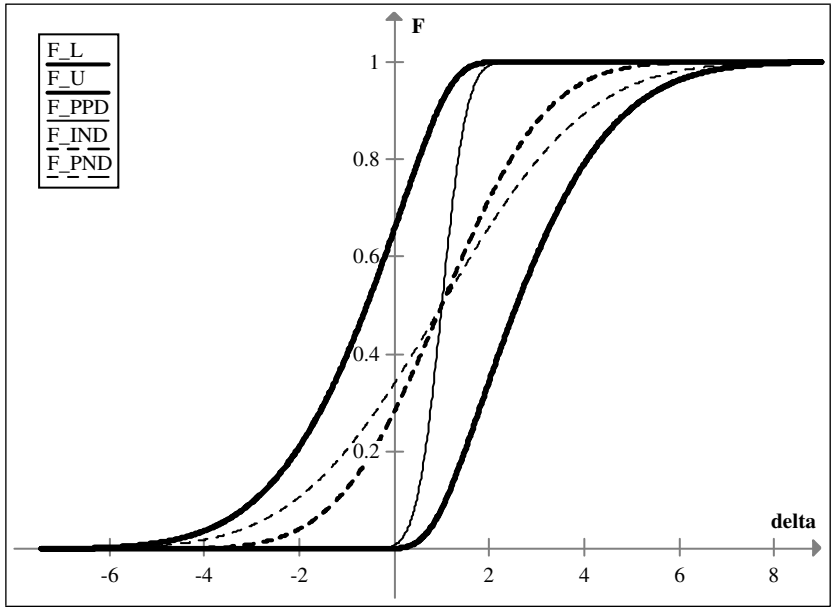


Fig. 1. Bounds on the Distribution of the Treatment Effect: ($N(2,2)$, $N(1,1)$).

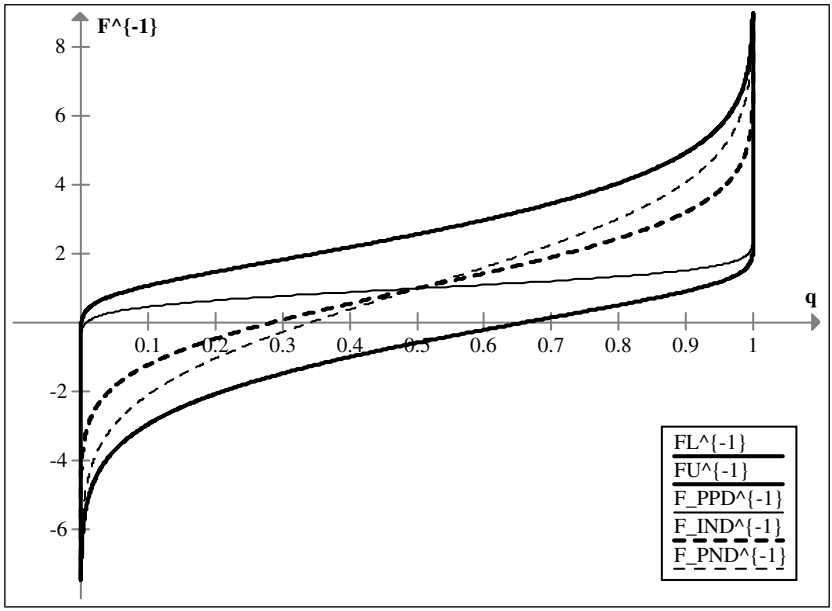


Fig. 2. Bounds on the Quantile Function of the Treatment Effect: ($N(2,2)$, $N(1,1)$).

relation $\text{F\_PPD} \gtrsim_{\text{SSD}} \text{F\_IND} \gtrsim_{\text{SSD}} \text{F\_PPD}$. In terms of the corresponding quantile functions, this implies that the quantile function corresponding to the perfectly positively dependent potential outcomes is flatter than the quantile functions corresponding to perfectly negatively dependent and independent potential outcomes, see Fig. 2 above.

2.3.2. Quantile Bounds

By inverting Eqs. (2) and (3), we obtain the quantile bounds for the case $\sigma_1 = \sigma_0 = \sigma$:

$$(F^{\text{L}})^{-1}(q) = \begin{cases} \text{any value in } (-\infty, \mu_1 - \mu_0] & \text{for } q = 0 \\ (\mu_1 - \mu_0) + 2\sigma\, \Phi^{-1}\left(\dfrac{1+q}{2}\right) & \text{otherwise} \end{cases}$$

$$(F^{\text{U}})^{-1}(q) = \begin{cases} (\mu_1 - \mu_0) + 2\sigma\, \Phi^{-1}\left(\dfrac{q}{2}\right) & \text{for } q \in [0, 1) \\ \text{any value in } [\mu_1 - \mu_0, \infty) & \text{for } q = 1 \end{cases}$$

When $\sigma_1 \neq \sigma_0$, there is no closed-form expression for the quantile bounds. But they can be computed numerically by either inverting the distribution bounds or using Lemma 3. We now derive the quantile function for the three special cases.

Case I. Perfect positive dependence. If $\sigma_1 = \sigma_0$, we get

$$F_{\Delta}^{-1}(q) = \begin{cases} \text{any value in } (-\infty, \mu_1 - \mu_0) & \text{for } q = 0, \\ \text{any value in } [\mu_1 - \mu_0, \infty) & \text{for } q = 1, \\ \text{undefined for } q \in (0,1). \end{cases}$$

When $\sigma_1 \neq \sigma_0$, we get

$$F_{\Delta}^{-1}(q) = (\mu_1 - \mu_0) + |\sigma_1 - \sigma_0|\Phi^{-1}(q) \text{ for } q \in [0,1]$$

Note that by definition, QTE is given by:

$$F_1^{-1}(q) - F_0^{-1}(q) = (\mu_1 - \mu_0) + (\sigma_1 - \sigma_0)\Phi^{-1}(q)$$

which equals $F_{\Delta}^{-1}(q)$ only if $\sigma_1 > \sigma_0$, that is, only if the condition of Lemma 4 (i) holds. If $\sigma_1 < \sigma_0$, $[F_1^{-1}(q) - F_0^{-1}(q)]$ is a decreasing function of q and hence cannot be a quantile function.

Case II. Perfect negative dependence.

$$F_{\Delta}^{-1}(q) = (\mu_1 - \mu_0) + (\sigma_1 + \sigma_0)\Phi^{-1}(q) \text{ for } q \in [0,1]$$

Case III. Independence.

$$F_{\Delta}^{-1}(q) = (\mu_1 - \mu_0) + \sqrt{\sigma_1^2 + \sigma_0^2}\Phi^{-1}(q) \text{ for } q \in [0,1]$$

In Fig. 2, we plot the quantile bounds for Δ (FL^{−1} and FU^{−1}) when $Y_1 \sim N(2, 2)$ and $Y_0 \sim N(1, 1)$ and the quantile functions of Δ when Y_1 and Y_0 are perfectly positively dependent, perfectly negatively dependent, and independent (F_PPD^{−1}, F_PND^{−1}, and F_IND^{−1}, respectively).

Again, Fig. 2 reveals the fact that the quantile function of Δ corresponding to the case that Y_1 and Y_0 are perfectly positively dependent is flatter than that corresponding to all the other cases. Keeping in mind that in this case, $\sigma_1 > \sigma_0$, we conclude that the quantile function of Δ in the perfect positive dependence case is the same as QTE. Fig. 2 leads to the conclusion that QTE is a conservative measure of the degree of heterogeneity of the treatment effect distribution.

3. MORE ON SHARP BOUNDS ON THE JOINT DISTRIBUTION OF POTENTIAL OUTCOMES AND THE DISTRIBUTION OF TREATMENT EFFECTS

For randomized experiments, Eq. (1) and Lemma 1, respectively, provide sharp bounds on the joint distribution of potential outcomes and the distribution of treatment effects. When additional information is available, these bounds are no longer sharp. In this section, we consider two types of additional information. One is the availability of a known value of a dependence measure between the potential outcomes and the other is the availability of covariates ensuring the validity of the selection-on-observables assumption.

3.1. Randomized Experiments with a Known Value of Kendall's τ

In this subsection, we first review sharp bounds on the joint distribution of the potential outcomes Y_1, Y_0 when the value of a dependence measure such as Kendall's τ between the potential outcomes is known. Then we point out how this information can be used to tighten the bounds on the distribution of Δ presented in Lemma 1. We provide details for Kendall's τ and point out relevant references for other measures including Spearman's ρ.

To begin, we introduce the notation used in Nelsen, Quesada-Molina, Rodriguez-Lallena, and Ubeda-Flores (2001). Let (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3) be three independent and identically distributed random vectors of dimension 2 whose joint distribution is H. Kendall's τ and Spearman's ρ are defined as:

$$\tau = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

$$\rho = 3\{\Pr[(X_1 - X_2)(Y_1 - Y_3) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_3) < 0]\}$$

For any $t \in [-1,1]$, let $\mathcal{T}_t$ denote the set of copulas with a common value t of Kendall's τ, that is,

$$\mathcal{T}_t = \{C | C \text{ is a copula such that } \tau(C) = t\}$$

Let $\underline{T}_t$ and $\bar{T}_t$ denote, respectively, the point-wise infimum and supremum of $\mathcal{T}_t$. The following result presents sharp bounds on the joint distribution of the potential outcomes Y_1, Y_0. It can be found in Nelsen et al. (2001).

Lemma 5. Suppose that the value of Kendall's τ between Y_1 and Y_0 is t. Then

$$\underline{T}_t(F_1(y_1), F_0(y_0)) \leq F(y_1, y_0) \leq \bar{T}_t(F_1(y_1), F_0(y_0))$$

where, for any $(u, v) \in [0,1]^2$;

$$\underline{T}_t(u, v) = \max\left(0, u + v - 1, \frac{1}{2}\left[(u+v) - \sqrt{(u-v)^2 + 1 - t}\right]\right)$$

$$\overline{T}_t(u, v) = \min\left(u, v, \frac{1}{2}\left[(u+v-1) + \sqrt{(u+v-1)^2 + 1 + t}\right]\right)$$

As shown in Nelsen et al. (2001),

$$\begin{aligned} \underline{T}_t(u, v) &= C^{\mathrm{L}}(u, v) \quad \text{if } t \in [-1, 0] \\ \underline{T}_t(u, v) &\geq C^{\mathrm{L}}(u, v) \quad \text{if } t \in [0, 1] \end{aligned} \tag{6}$$

and

$$\bar{T}_t(u, v) = C^{\mathrm{U}}(u, v) \quad \text{if } t \in [0, 1]$$

$$\bar{T}_t(u, v) \leq C^{\mathrm{U}}(u, v) \quad \text{if } t \in [-1, 0]$$

Hence, for any fixed (y_1, y_0), the bounds $[\underline{T}_t(F_1(y_1), F_0(y_0)), \bar{T}_t(F_1(y_1), F_0(y_0))]$ are in general tighter than the bounds in Eq. (1) unless $t = 0$. The lower bound on $F(y_1, y_0)$ can be used to tighten bounds on the distribution of treatment effects via the following result in Williamson and Downs (1990).

Lemma 6. Let $\underline{C}_{XY}$ denote a lower bound on the copula C_{XY} and F_{X+Y} denote the distribution function of $X + U$. Then

$$\sup_{x+y=z} \underline{C}_{XY}(F(x), G(y)) \leq F_{X+Y}(z) \leq \inf_{x+y=z} \underline{C}^d_{XY}(F(x), G(y))$$

where $\underline{C}^d_{XY}(u, v) = u + v - \underline{C}_{XY}(u, v)$.

Let $Y_1 = X$ and $Y_0 = -Y$ in Lemma 6. By using Lemma 5 and the duality theorem, we can prove the following proposition.

Proposition 1. Suppose the value of Kendall's τ between Y_1 and Y_0 is t. Then

(i) $\sup_x \underline{T}_{-t}(F_1(x), 1 - F_0(x - \delta)) \leq F_\Delta(\delta) \leq \inf_x \underline{T}^d_{-t}(F_1(x), 1 - F_0(x - \delta))$, where

$$\underline{T}_{-t}(u, v) = \max\left\{0, u + v + 1, \frac{1}{2}\left[(u + v) - \sqrt{(u - v)^2 + 1 + t}\right]\right\}$$

$$\underline{T}^d_{-t}(u, v) = \max\left\{u + v, 1, \frac{1}{2}\left[(u + v) + \sqrt{(u - v)^2 + 1 + t}\right]\right\}$$

(ii) $\sup_{\underline{T}^d_{-t}(u,v)=q}[F_1^{-1}(u) - F_0^{-1}(1 - v)] \leq F_\Delta^{-1}(q) \leq \inf_{\underline{T}_{-t}(u,1-v)=q}[F_1^{-1}(u) - F_0^{-1}(1 - v)]$.

Proposition 1 and Eq. (6) imply that the bounds in Proposition 1 (i) are sharper than those in Lemma 1 if $t \in [-1, 0]$ and are the same as those in Lemma 1 if $t \in [0, 1]$. This implies that if the potential outcomes Y_1 and Y_0 are positively dependent in the sense of having a nonnegative Kendall's τ, then the information on the value of Kendall's τ does not improve the bounds on the distribution of treatment effects. On contrary, if they are negatively dependent on each other, then knowing the value of Kendall's τ will in general improve the bounds.

Remark 1. If instead of Kendall's τ, the value of Spearman's ρ between the potential outcomes is known, one can also establish tighter bounds on $F_\Delta(z)$ by using Theorem 4 in Nelsen et al. (2001) and Lemma 6.

Remark 2. Other dependence information that may be used to tighten bounds on the joint distribution of potential outcomes and thus the distribution of treatment effects include known values of the copula function of the potential outcomes at certain points, see Nelsen and Ubeda-Flores (2004) and Nelsen, Quesada-Molina, Rodriguez-Lallena, and Ubeda-Flores (2004).

3.2. Selection-on-Observables

In many applications, observations on a vector of covariates for individuals in the treatment and control groups are available. In this subsection, we extend sharp bounds for randomized experiments in Lemma 1 to take into account these covariates. For notational compactness, we let $n = n_1 + n_0$ so that there are n individuals altogether. For $i = 1, \ldots, n$, let X_i denote the

observed vector of covariates and D_i the binary variable indicating participation; $D_i = 1$ if individual i belongs to the treatment group and $D_i = 0$ if individual i belongs to the control group. Let $Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i)$ denote the observed outcome for individual i. We have a random sample $\{Y_i, X_i, D_i\}_{i=1}^n$. In the literature on program evaluation with selection-on-observables, the following two assumptions are often used to evaluate the effect of a treatment or a program, see for example, Rosenbaum and Rubin (1983), Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998), Dehejia and Wahba (1999), and Hirano, Imbens, and Ridder (2003), to name only a few.

C1. Let (Y_1, Y_0, D, X) have a joint distribution. For all $x \in \mathcal{X}$ (the support of X), (Y_1, Y_0) is jointly independent of D conditional on $X = x$.

C2. For all $x \in \mathcal{X}$, $0 < p(x) < 1$, where $p(x) = P(D = 1|x)$.

In the following, we present sharp bounds on the joint distribution of potential outcomes and the distribution of Δ under (C1) and (C2). For any fixed $x \in \mathcal{X}$, Eq. (1) provides sharp bounds on the conditional joint distribution of Y_1, Y_0 given $X = x$:

$$C^{\mathrm{L}}(F_1(y_1|x), F_0(y_0|x)) \le F(y_1, y_0|x) \le C^{\mathrm{U}}(F_1(y_1|x), F_0(y_0|x))$$

and Lemma 1 provides sharp bounds on the conditional distribution of Δ given $X = x$:

$$F^{\mathrm{L}}(\delta|x) \le F_\Delta(\delta|x) \le F^{\mathrm{U}}(\delta|x)$$

where

$$F^{\mathrm{L}}(\delta|x) = \sup_y \max(F_1(y|x) - F_0(y - \delta|x), 0)$$

$$F^{\mathrm{U}}(\delta|x) = 1 + \inf_y \min(F_1(y|x) - F_0(y - \delta|x), 0)$$

Here, we use $F_\Delta(\cdot|x)$ to denote the conditional distribution function of Δ given $X = x$. The other conditional distributions are defined similarly. Conditions (C1) and (C2) allow the identification of the conditional distributions $F_1(y|x)$ and $F_0(y|x)$ appearing in the sharp bounds on $F(y_1, y_0|x)$ and $F_\Delta(\delta|x)$. To see this, note that

$$\begin{aligned} F_1(y|x) &= P(Y_1 \le y|X = x) = P(Y_1 \le y|X = x, D = 1) \\ &= P(Y \le y|X = x, D = 1) \end{aligned} \tag{7}$$

where (C1) is used to establish the second equality. Similarly, we get

$$F_0(y|x) = P(Y \le y|X = x, D = 0) \tag{8}$$

Sharp bounds on the unconditional joint distribution of Y_1, Y_0 and the unconditional distribution of Δ follow from those of the conditional distributions:

$$E[C^{\mathrm{L}}(F_1(y_1|X), F_0(y_0|X))] \le F(y_1, y_0) \le C^{\mathrm{U}}(F_1(y_1|X), F_0(y_0|X))$$

$$E(F^{\mathrm{L}}(\delta|X)) \le F_\Delta(\delta) = E(F_\Delta(\delta|X)) \le E(F^{\mathrm{U}}(\delta|X))$$

We note that if X is independent of (Y_1, Y_0), then the above bounds on $F(y_1, y_0)$ and $F_\Delta(\delta)$ reduce, respectively, to those in Eq. (1) and Lemma 1. In general, X is not independent of (Y_1, Y_0) and the above bounds are tighter than those in Eq. (1) and Lemma 1, see Fan (2008) for a more detailed discussion on the sharp bounds with covariates. Under the selection on observables assumption, Fan and Zhu (2009) established sharp bounds on a general class of functionals of the joint distribution $F(y_1, y_0)$ including the correlation coefficient between the potential outcomes and the class of D_2-parameters of the distribution of treatment effects.

4. NONPARAMETRIC ESTIMATORS OF THE SHARP BOUNDS AND THEIR ASYMPTOTIC PROPERTIES FOR RANDOMIZED EXPERIMENTS

Suppose random samples $\{Y_{1i}\}_{i=1}^{n_1} \sim F_1$ and $\{Y_{0i}\}_{i=1}^{n_0} \sim F_0$ are available. Let $\mathcal{Y}_1$ and $\mathcal{Y}_0$ denote, respectively, the supports[4] of F_1 and F_0. Note that the bounds in Lemma 1 can be written as:

$$F^{\mathrm{L}}(\delta) = \sup_{y \in \mathcal{R}} \{F_1(y) - F_0(y - \delta)\}, F^{\mathrm{U}}(\delta) = 1 + \inf_{y \in \mathcal{R}} \{F_1(y) - F_0(y - \delta)\} \tag{9}$$

since for any two distributions F_1 and F_0, it is always true that $\sup_{y \in \mathcal{R}} \{F_1(y) - F_0(y - \delta)\} \ge 0$ and $\inf_{y \in \mathcal{R}} \{F_1(y) - F_0(y - \delta)\} \le 0$.

When $\mathcal{Y}_1 = \mathcal{Y}_0 = \mathcal{R}$, Eq. (9) suggests the following plug-in estimators of $F^{\mathrm{L}}(\delta)$ and $F^{\mathrm{U}}(\delta)$:

$$F_n^{\mathrm{L}}(\delta) = \sup_{y \in \mathcal{R}} \{F_{1n}(y) - F_{0n}(y - \delta)\}, F_n^{\mathrm{U}}(\delta) = 1 + \inf_{y \in \mathcal{R}} \{F_{1n}(y) - F_{0n}(y - \delta)\} \tag{10}$$

where $F_{1n}(\cdot)$ and $F_{0n}(\cdot)$ are the empirical distributions defined as:

$$F_{kn}(y) = \frac{1}{n_k}\sum_{i=1}^{n_k} 1\{Y_{ki} \le y\}, \quad k = 1, 0$$

When either $\mathcal{Y}_1$ or $\mathcal{Y}_0$ is not the whole real line, we derive alternative expressions for $F^{\mathrm{L}}(\delta)$ and $F^{\mathrm{U}}(\delta)$ which turn out to be convenient for both computational purposes and for asymptotic analysis. For illustration, we look at the case: $\mathcal{Y}_1 = \mathcal{Y}_0 = [0, 1]$ in detail and provide the results for the general case afterwards.

Suppose $\mathcal{Y}_1 = \mathcal{Y}_0 = [0, 1]$. If $1 \ge \delta \ge 0$, then Eq. (9) implies:

$$\begin{aligned}
F^{\mathrm{L}}(\delta) &= \max\left\{ \sup_{y\in[\delta,1]} \{F_1(y) - F_0(y-\delta)\}, \sup_{y\in(-\infty,\delta)} \{F_1(y) - F_0(y-\delta)\}, \right. \\
&\qquad \left. \sup_{y\in(1,\infty)} \{F_1(y) - F_0(y-\delta)\} \right\} \\
&= \max\left\{ \sup_{y\in[\delta,1]} \{F_1(y) - F_0(y-\delta)\}, \sup_{y\in(-\infty,\delta)} F_1(y), \sup_{y\in(1,\infty)} \{1 - F_0(y-\delta)\} \right\} \\
&= \max\left\{ \sup_{y\in[\delta,1]} \{F_1(y) - F_0(y-\delta)\}, F_1(\delta), 1 - F_0(1-\delta) \right\} \\
&= \sup_{y\in[\delta,1]} \{F_1(y) - F_0(y-\delta)\} \qquad (11)
\end{aligned}$$

and

$$\begin{aligned}
F^{\mathrm{U}}(\delta) &= 1 + \min\left\{ \inf_{y\in[\delta,1]} \{F_1(y) - F_0(y-\delta)\}, \inf_{y\in(-\infty,\delta)} \{F_1(y) - F_0(y-\delta)\}, \right. \\
&\qquad \left. \inf_{y\in(1,\infty)} \{F_1(y) - F_0(y-\delta)\} \right\} \\
&= 1 + \min\left\{ \inf_{y\in[\delta,1]} \{F_1(y) - F_0(y-\delta)\}, \inf_{y\in(-\infty,\delta)} F_1(y), \inf_{y\in(1,\infty)} \{1 - F_0(y-\delta)\} \right\} \\
&= 1 + \min\left\{ \inf_{y\in[\delta,1]} \{F_1(y) - F_0(y-\delta)\}, 0 \right\}
\end{aligned}$$

If $-1 \le \delta < 0$, then

$$\begin{aligned} F^{\mathrm{L}}(\delta) &= \max\left\{ \sup_{y\in[0,1+\delta]} \{F_1(y) - F_0(y-\delta)\}, \sup_{y\in(-\infty,0)} \{F_1(y) - F_0(y-\delta)\}, \right. \\ &\qquad \left. \sup_{y\in(1+\delta,\infty)} \{F_1(y) - F_0(y-\delta)\} \right\} \\ &= \max\left\{ \sup_{y\in[0,1+\delta]} \{F_1(y) - F_0(y-\delta)\}, \sup_{y\in(-\infty,0)} \{-F_0(y-\delta)\}, \right. \\ &\qquad \left. \sup_{y\in(1+\delta,\infty)} \{F_1(y) - 1)\} \right\} \\ &= \max\left\{ \sup_{y\in[0,1+\delta]} \{F_1(y) - F_0(y-\delta)\}, 0 \right\} \end{aligned} \tag{12}$$

and

$$\begin{aligned} F^{\mathrm{U}}(\delta) &= 1 + \min\left\{ \inf_{y\in[0,1+\delta]} \{F_1(y) - F_0(y-\delta)\}, \inf_{y\in(-\infty,0)} \{F_1(y) - F_0(y-\delta)\}, \right. \\ &\qquad \left. \inf_{y\in(1+\delta,\infty)} \{F_1(y) - F_0(y-\delta)\} \right\} \\ &= 1 + \min\left\{ \inf_{y\in[0,1+\delta]} \{F_1(y) - F_0(y-\delta)\}, \inf_{y\in(-\infty,0)} \{-F_0(y-\delta)\}, \right. \\ &\qquad \left. \inf_{y\in(1+\delta,\infty)} \{F_1(y) - 1\} \right\} \\ &= 1 + \inf_{y\in[0,1+\delta]} \{F_1(y) - F_0(y-\delta)\} \end{aligned}$$

Based on Eqs. (11) and (12), we propose the following estimator of $F^{\mathrm{L}}(\delta)$:

$$F_n^{\mathrm{L}}(\delta) = \begin{cases} \sup_{y\in[\delta,1]}\{F_{1n}(y) - F_{0n}(y-\delta)\} & \text{if } 1 \ge \delta \ge 0 \\ \max\{\sup_{y\in[0,1+\delta]}\{F_{1n}(y) - F_{0n}(y-\delta), 0\} & \text{if } -1 \le \delta < 0 \end{cases}$$

Similarly, we propose the following estimator for $E^{\mathrm{U}}(\delta)$:

$$F_n^{\mathrm{U}}(\delta) = \begin{cases} 1 + \min\ \{\inf_{y\in[\delta,1]}\{F_{1n}(y) - F_{0n}(y-\delta)\}, 0\} & \text{if } 1 \ge \delta \ge 0 \\ 1 + \inf_{y\in[0,1+\delta]}\{F_{1n}(y) - F_{0n}(y-\delta)\} & \text{if } -1 \le \delta < 0 \end{cases}$$

We now summarize the results for general supports $\mathcal{Y}_1$ and $\mathcal{Y}_0$. Suppose $\mathcal{Y}_1 = [a, b]$ and $\mathcal{Y}_0 = [c, d]$ for $a, b, c, d \in \bar{\mathcal{R}} \equiv \mathcal{R} \cup \{-\infty, +\infty\}, a < b, c < d$ with $F_1(a) = F_0(c) = 0$ and $F_1(b) = F_0(d) = 1$. It is easy to see that

$$F^{\mathrm{L}}(\delta) = F^{\mathrm{U}}(\delta) = 0, \quad \text{if } \delta \le a - d \quad \text{and} \quad F^{\mathrm{L}}(\delta) = F^{\mathrm{U}}(\delta) = 1, \quad \text{if } \delta \ge b - c$$

For any $\delta \in [a - d, b - c] \cap \mathcal{R}$ let $\mathcal{Y}_\delta = [a, b] \cap [c + \delta, d + \delta]$. A similar derivation to the case $\mathcal{Y}_1 = \mathcal{Y}_0 = [0, 1]$ leads to

$$F^{\mathrm{L}}(\delta) = \max\left\{ \sup_{y \in \mathcal{Y}_\delta} \{F_1(y) - F_0(y - \delta)\}, 0 \right\}$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\left\{ \inf_{y \in \mathcal{Y}_\delta} \{F_1(y) - F_0(y - \delta)\}, 0 \right\}$$

which suggest the following plug-in estimators of $F^{\mathrm{L}}(\delta)$ and $F^{\mathrm{U}}(\delta)$:

$$F_n^{\mathrm{L}}(\delta) = \max\left\{ \sup_{y \in \mathcal{Y}_\delta} \{F_{1n}(y) - F_{0n}(y - \delta)\}, 0 \right\} \tag{13}$$

$$F_n^{\mathrm{U}}(\delta) = 1 + \min\left\{ \inf_{y \in \mathcal{Y}_\delta} \{F_{1n}(y) - F_{0n}(y - \delta)\}, 0 \right\} \tag{14}$$

By using $F_n^{\mathrm{L}}(\delta)$ and $F_n^{\mathrm{U}}(\delta)$, we can estimate bounds on effects of interest other than the average treatment effects including the proportion of people receiving the treatment who benefit from it, see Heckman et al. (1997) for discussion on some of these effects. In the rest of this section, we review the asymptotic distributions of $\sqrt{n_1}(F_n^{\mathrm{L}}(\delta) - F^{\mathrm{L}}(\delta))$ and $\sqrt{n_1}(F_n^{\mathrm{U}}(\delta) - F^{\mathrm{U}}(\delta))$ established in Fan and Park (2010), provide two numerical examples to demonstrate the restrictiveness of two assumptions used in Fan and Park (2010), and then establish asymptotic distributions of $\sqrt{n_1}(F_n^{\mathrm{L}}(\delta) - F^{\mathrm{L}}(\delta))$ and $\sqrt{n_1}(F_n^{\mathrm{U}}(\delta) - F^{\mathrm{U}}(\delta))$ with much weaker assumptions.

4.1. Asymptotic Distributions of $F_n^{\mathrm{L}}(\delta), F_n^{\mathrm{U}}(\delta)$

Define

$$\mathcal{Y}_{\sup,\delta} = \arg\sup_{y \in \mathcal{Y}_\delta} \{F_1(y) - F_0(y - \delta)\}, \quad \mathcal{Y}_{\inf,\delta} = \arg\inf_{y \in \mathcal{Y}_\delta} \{F_1(y) - F_0(y - \delta)\}$$

$$M(\delta) = \sup_{y \in \mathcal{Y}_\delta} \{F_1(y) - F_0(y - \delta)\}, \quad m(\delta) = \inf_{y \in \mathcal{Y}_\delta} \{F_1(y) - F_0(y - \delta)\}$$

$$M_n(\delta) = \sup_{y \in \mathcal{Y}_\delta} \{F_{1n}(y) - F_{0n}(y-\delta)\}, \quad m_n(\delta) = \inf_{y \in \mathcal{Y}_\delta} \{F_{1n}(y) - F_{0n}(y-\delta)\}$$

Then

$$F_n^{\mathrm{L}}(\delta) = \max\{M_n(\delta), 0\}, \quad F_n^{\mathrm{U}}(\delta) = 1 + \min\{m_n(\delta), 0\}$$

Fan and Park (2010) assume that $\mathcal{Y}_{\sup,\delta}$ and $\mathcal{Y}_{\inf,\delta}$ are both singletons. Let $y_{\sup,\delta}$ and $y_{\inf,\delta}$ denote, respectively, the elements of $\mathcal{Y}_{\sup,\delta}$ and $\mathcal{Y}_{\inf,\delta}$. The following assumptions are used in Fan and Park (2010).

A1. (i) The two samples $\{Y_{1i}\}_{i=1}^{n_1}$ and $\{Y_{0i}\}_{i=1}^{n_0}$ are each i.i.d. and are independent of each other; (ii) $n_1/n_0 \to \lambda$ as $n_1 \to \infty$ with $0 < \lambda < \infty$.

A2. The distribution functions F_1 and F_0 are twice differentiable with bounded density functions f_1 and f_0 on their supports.

A3. (i) For every $\varepsilon > 0$, $\sup_{y \in \mathcal{Y}_\delta : |y - y_{\sup,\delta}| \geq \varepsilon} \{F_1(y) - F_0(y-\delta)\} < \{F_1(y_{\sup,\delta}) - F_0(y_{\sup,\delta} - \delta)\}$; (ii) $f_1(y_{\sup,\delta}) - f_0(y_{\sup,\delta} - \delta) = 0$ and $f_1'(y_{\sup,\delta}) - f_0'(y_{\sup,\delta} - \delta) < 0$.

A4. (i) For every $\varepsilon > 0$, $\inf_{y \in \mathcal{Y}_\delta : |y - y_{\inf,\delta}| \geq \varepsilon} \{F_1(y) - F_0(y-\delta)\} < \{F_1(y_{\inf,\delta}) - F_0(y_{\inf,\delta} - \delta)\}$; (ii) $f_1(y_{\inf,\delta}) - f_0(y_{\inf,\delta} - \delta) = 0$ and $f_1'(y_{\inf,\delta}) - f_0'(y_{\inf,\delta} - \delta) > 0$.

The independence assumption of the two samples in (A1) is satisfied by data from ideal randomized experiments. (A2) imposes smoothness assumptions on the marginal distribution functions. (A3) and (A4) are identifiability assumptions. For a fixed $\delta \in [a-d, b-c] \cap \mathcal{R}$, (A3) requires the function $y \to \{F_1(y) - F_0(y-\delta)\}$ to have a well-separated interior maximum at $y_{\sup,\delta}$ on $\mathcal{Y}_\delta$, while (A4) requires the function $y \to \{F_1(y) - F_0(y-\delta)\}$ to have a well-separated interior minimum at $y_{\inf,\delta}$ on $\mathcal{Y}_\delta$. If $\mathcal{Y}_\delta$ is compact, then (A3) and (A4) are implied by (A2) and the assumption that the function $y \to \{F_1(y) - F_0(y-\delta)\}$ have a unique maximum at $y_{\sup,\delta}$ and a unique minimum at $y_{\inf,\delta}$ in the interior of $\mathcal{Y}_\delta$.

The following result is provided in Fan and Park (2010).

Theorem 1. Define

$$\sigma_{\mathrm{L}}^2 = F_1(y_{\sup,\delta})[1 - F_1(y_{\sup,\delta})] + \lambda F_0(y_{\sup,\delta} - \delta)[1 - F_0(y_{\sup,\delta} - \delta)] \quad \text{and}$$

$$\sigma_{\mathrm{U}}^2 = F_1(y_{\inf,\delta})[1 - F_1(y_{\inf,\delta})] + \lambda F_0(y_{\inf,\delta} - \delta)[1 - F_0(y_{\inf,\delta} - \delta)]$$

(i) Suppose (A1)–(A3) hold. For any $\delta \in [a-d, b-c] \cap \mathcal{R}$

$$\sqrt{n_1}[F_n^{\mathrm{L}}(\delta) - F^{\mathrm{L}}(\delta)] \Rightarrow \begin{cases} N(0, \sigma_{\mathrm{L}}^2), & \text{if } M(\delta) > 0 \\ \max\{N(0, \sigma_{\mathrm{L}}^2), 0\} & \text{if } M(\delta) = 0 \end{cases}$$

$$\text{and } \Pr(F_n^{\mathrm{L}}(\delta) = 0) \to 1 \text{ if } M(\delta) < 0$$

(ii) Suppose (A1), (A2), and (A4) hold. For any $\delta \in [a-d, b-c] \cap \mathcal{R}$,

$$\sqrt{n_1}[F_n^{\mathrm{U}}(\delta) - F^{\mathrm{U}}(\delta)] \Rightarrow \begin{cases} N(0, \sigma_{\mathrm{U}}^2) & \text{if } m(\delta) > 0 \\ \min\{N(0, \sigma_{\mathrm{U}}^2), 0\} & \text{if } m(\delta) = 0 \end{cases}$$

$$\text{and } \Pr(F_n^{\mathrm{U}}(\delta) = 1) \to 1 \text{ if } m(\delta) > 0$$

Theorem 1 shows that the asymptotic distribution of $F_n^{\mathrm{L}}(\delta)(F_n^{\mathrm{U}}(\delta))$ depends on the value of $M(\delta)$ ($m(\delta)$). For example, if δ is such that $M(\delta) > 0$ ($m(\delta) < 0$), then $F_n^{\mathrm{L}}(\delta)$ ($F_n^{\mathrm{U}}(\delta)$) is asymptotically normally distributed, but if δ is such that $M(\delta) = 0$ ($m(\delta) = 0$), then the asymptotic distribution of $F_n^{\mathrm{L}}(\delta)(F_n^{\mathrm{U}}(\delta))$ is truncated normal.

Remark 3. Fan and Park (2010) proposed the following procedure for computing the estimates $F_n^{\mathrm{L}}(\delta), F_n^{\mathrm{U}}(\delta)$ and estimates of σ_{L}^2 and σ_{U}^2 in Theorem 1. Suppose we know $\mathcal{Y}_\delta$. If $\mathcal{Y}_\delta$ is unknown, we can estimate it by:

$$\mathcal{Y}_{\delta n} = [Y_{1(1)}, Y_{1(n_1)}] \cap [Y_{0(1)} + \delta, Y_{0(n_0)} + \delta]$$

where $\{Y_{1(i)}\}_{i=1}^{n_1}$ and $\{Y_{0(i)}\}_{i=1}^{n_0}$ are the order statistics of $\{Y_{1(i)}\}_{i=1}^{n_1}$ and $\{Y_{0(i)}\}_{i=1}^{n_0}$, respectively (in ascending order). In the discussion below, $\mathcal{Y}_\delta$ can be replaced by $\mathcal{Y}_{\delta n}$ if $\mathcal{Y}_\delta$ is unknown.

We define a subset of the order statistics $\{Y_{1(i)}\}_{i=1}^{n_1}$ denoted as $\{Y_{1(i)}\}_{i=r_1}^{s_1}$ as follows:

$$r_1 = \arg\min_i [\{Y_{1(i)}\}_{i=1}^{n_1} \cap \mathcal{Y}_\delta] \quad \text{and} \quad s_1 = \arg\max_i [\{Y_{1(i)}\}_{i=1}^{n_1} \cap \mathcal{Y}_\delta]$$

In words, $Y_{1(r_1)}$ is the smallest value of $\{Y_{1(i)}\}_{i=1}^{n_1} \cap \mathcal{Y}_\delta$ and $Y_{1(s_1)}$ is the largest. Then,

$$M_n(\delta) = \max_i \left\{ \frac{i}{n_1} - F_{0n}(Y_{1(i)} - \delta) \right\} \quad \text{for } i \in \{r_1, r_1 + 1, \ldots, s_1\} \tag{15}$$

$$m_n(\delta) = \min_i \left\{ \frac{i}{n_1} - F_{0n}(Y_{1(i)} - \delta) \right\} \quad \text{for } i \in \{r_1, r_1 + 1, \ldots, s_1\} \tag{16}$$

The estimates $F_n^{\mathrm{L}}(\delta)$, $F_n^{\mathrm{U}}(\delta)$ are given by: $F_n^{\mathrm{L}}(\delta) = \max\{M_n(\delta), 0\}$, $F_n^{\mathrm{U}}(\delta) = 1 + \min\{m_n(\delta), 0\}$.

Define two sets I_M and I_m such that

$$I_M = \left\{ i : i = \arg\max_i \left\{ \frac{i}{n_1} - F_{0n}(Y_{1(i)} - \delta) \right\} \right\} \quad \text{and}$$

$$I_m = \left\{ i : i = \arg\min_i \left\{ \frac{i}{n_1} - F_{0n}(Y_{1(i)} - \delta) \right\} \right\}$$

Then the estimators σ^2_{Ln} and σ^2_{Un} can be defined as:

$$\sigma^2_{\mathrm{Ln}} = \frac{i}{n_1}\left(1 - \frac{i}{n_1}\right) + \lambda F_{0n}(Y_{1(i)} - \delta)(1 - F_{0n}(Y_{1(i)} - \delta)) \quad \text{and}$$

$$\sigma^2_{\mathrm{Un}} = \frac{j}{n_1}\left(1 - \frac{j}{n_1}\right) + \lambda F_{0n}(Y_{1(j)} - \delta)(1 - F_{0n}(Y_{1(j)} - \delta))$$

for $i \in I_M$ and $j \in I_m$. Since I_M or I_m may not be singleton, we may have multiple estimates of σ^2_{Ln} or σ^2_{Un}. In such a case, we may use $i = \min_k\{k \in I_M\}$ and $j = \min_k\{k \in I_m\}$.

Remark 4. Alternatively we can compute $F_n^{\mathrm{L}}(\delta), F_n^{\mathrm{U}}(\delta)$ as follows. Note that for $0 < q < 1$, Lemma 3 (the duality theorem) implies that the quantile bounds $(F_n^{\mathrm{U}})^{-1}(q)$ and $(F_n^{\mathrm{L}})^{-1}(q)$ can be computed by:

$$\begin{aligned}(F_n^{\mathrm{L}})^{-1}(q) &= \inf_{u \in [q,1]} [F_{1n}^{-1}(u) - F_{0n}^{-1}(u - q)], (F_n^{\mathrm{U}})^{-1}(q) \\ &= \sup_{u \in [0,q]} [F_{1n}^{-1}(u) - F_{0n}^{-1}(1 + u - q)]\end{aligned}$$

where $F_{1n}^{-1}(\cdot)$ and $F_{0n}^{-1}(\cdot)$ represent the quantile functions of $F_{1n}(\cdot)$ and $F_{0n}(\cdot)$, respectively. To estimate the distribution bounds, we compute the values of $(F_n^{\mathrm{L}})^{-1}(q)$ and $(F_n^{\mathrm{U}})^{-1}(q)$ a evenly spaced values of q in (0, 1). One choice that leads to easily computed formulas for $(F_n^{\mathrm{L}})^{-1}(q)$ and

$(F_n^{\mathrm{U}})^{-1}(q)$ is $q = r/n_1$ for $r = 1, \ldots, n_1$, as one can show that

$$(F_n^{\mathrm{L}})^{-1}\left(\frac{r}{n_1}\right) = \min_{l=r,\ldots,(n_1-1)} \min_{s=j,\ldots,k} [Y_{1(l+1)} - Y_{0(s)}] \tag{17}$$

where $j = [n_0((l-r)/n_1)] + 1$ and $k = [n_0((l-r+1)/n_1)]$, and

$$(F_n^{\mathrm{U}})^{-1}\left(\frac{r}{n_1}\right) = \max_{l=0,\ldots,(r-1)} \max_{s=j',\ldots,k'} [Y_{1(l+1)} - Y_{0(s)}] \tag{18}$$

where $j' = [n_0((n_1+l-r)/n_1)] + 1$ and $k' = [n_0((n_1+l-r+1)/n_1)]$. In the case where $\mathrm{n}_1 = \mathrm{n}_0 = \mathrm{n}$, Eqs. (17) and (18) simplify:

$$(F_n^{\mathrm{L}})^{-1}\left(\frac{r}{n}\right) = \min_{l=r,\ldots,(n-1)} [Y_{1(l+1)} - Y_{0(l-r+1)}]$$

$$(F_n^{\mathrm{U}})^{-1}\left(\frac{r}{n}\right) = \max_{l=0,\ldots,(r-1)} [Y_{1(l+1)} - Y_{0(n+l-r+1)}]$$

The empirical distribution of $(F_n^{\mathrm{L}})^{-1}(r/n_1), r = 1, \ldots, n_1$, provides an estimate of the lower bound distribution and the empirical distribution of $(F_n^{\mathrm{U}})^{-1}(r/n_1), r = 1, \ldots, n_1$, provides an estimate of the upper bound distribution. This is the approach we used in our simulations to compute $F_n^{\mathrm{L}}(\delta), F_n^{\mathrm{U}}(\delta)$.

4.2. Two Numerical Examples

We present two examples to illustrate the various possibilities in Theorem 1. For the first example, the asymptotic distribution of $F_n^{\mathrm{L}}(\delta)(F_n^{\mathrm{U}}(\delta))$ is normal for all δ. For the second example, the asymptotic distribution of $F_n^{\mathrm{L}}(\delta)(F_n^{\mathrm{U}}(\delta))$ is normal for some δ and nonnormal for some other δ. More examples can be found in Appendix B.

Example 1 (Continued). Let $Y_j \sim N(\mu_j, \sigma_j^2)$ for $j = 0, 1$ with $\sigma_1^2 \neq \sigma_0^2$. As shown in Section 2.3, $M(\delta) > 0$ and $m(\delta) < 0$ for all $\delta \in \mathcal{R}$. Moreover,

$$y_{\sup,\delta} = \frac{\sigma_1^2 s + \sigma_1\sigma_0 t}{\sigma_1^2 - \sigma_0^2} + \mu_1 \quad \text{and} \quad y_{\inf,\delta} = \frac{\sigma_1^2 s + \sigma_1\sigma_0 t}{\sigma_1^2 - \sigma_0^2} + \mu_1$$

are unique interior solutions, where $s = \delta - (\mu_1 - \mu_0)$ and $\sqrt{s^2 + 2(\sigma_1^2 - \sigma_0^2)\ln(\sigma_1/\sigma_0)}$. Theorem 1 implies that the asymptotic

distribution of $F_n^{\mathrm{L}}(\delta)(F_n^{\mathrm{U}}(\delta))$ is normal for all $\delta \in \mathcal{R}$. Inferences can be made using asymptotic distributions or standard bootstrap with the same sample size.

Example 2. Consider the following family of distributions indexed by $a \in (0, 1)$. For brevity, we denote a member of this family by $C(a)$. If $X \sim C(a)$, then

$$F(x) = \begin{cases} \dfrac{1}{a}x^2 & \text{if } x \in [0, a] \\ 1 - \dfrac{(x-1)^2}{(1-a)} & \text{if } x \in [a, 1] \end{cases} \quad \text{and} \quad f(x) = \begin{cases} \dfrac{2}{a}x & \text{if } x \in [0, a] \\ \dfrac{2(1-x)}{(1-a)} & \text{if } x \in [a, 1] \end{cases}$$

Suppose $Y_1 \sim C(1/4)$ and $Y_0 \sim C(3/4)$. The functional form of $F_1(y) - F_0(y-\delta)$ differs according to δ. For $y \in \mathcal{Y}_\delta$, using the expressions for $F_1(y) - F_0(y-\delta)$ provided in Appendix B, one can find $y_{\sup,\delta}$ and $M(\delta)$. They are:

$$y_{\sup,\delta} = \begin{cases} \dfrac{1+\delta}{2} & \text{if } -1 + \dfrac{1}{2}\sqrt{2} < \delta \leq 1 \\ \left\{0, \dfrac{1+\delta}{2}, 1+\delta\right\} & \text{if } \delta = -1 + \dfrac{1}{2}\sqrt{2} \\ \{0, 1+\delta\} & \text{if } -1 \leq \delta < -1 + \dfrac{1}{2}\sqrt{2} \end{cases}$$

$$M(\delta) = \begin{cases} 4(\delta+1)^2 - 1 & \text{if } -1 \leq \delta \leq -\dfrac{3}{4} \\ -\dfrac{4}{3}\delta^2 & \text{if } -\dfrac{3}{4} \leq \delta \leq -1 + \dfrac{1}{2}\sqrt{2} \\ -\dfrac{3}{2}(\delta-1)^2 + 1 & \text{if } -1 + \dfrac{1}{2}\sqrt{2} \leq \delta \leq 1 \end{cases}$$

Fig. 3 plots $y_{\sup,\delta}$ and $M(\delta)$ against δ.

Fig. 4 plots $F_1(y) - F_0(y-\delta)$ against $y \in [0, 1]$ for a few selected values of δ. When $\delta = -(5/8)$ (Fig. 4(a)), the supremum occurs at the boundaries of $\mathcal{Y}_\delta$. When $\delta = -1 + (\sqrt{2}/2)$ (Fig. 4(b)), $\{y_{\sup,\delta}\} = \{0, ((1+\delta)/2), 1+\delta\}$, that is, there are three values of $y_{\sup,\delta}$; one interior and two boundary solutions. When $\delta > -1 + (\sqrt{2}/2)$, $y_{\sup,\delta}$ becomes a unique interior solution. Fig. 4(c) plots the case where the interior solution leads to a value 0 for $M(\delta)$ and

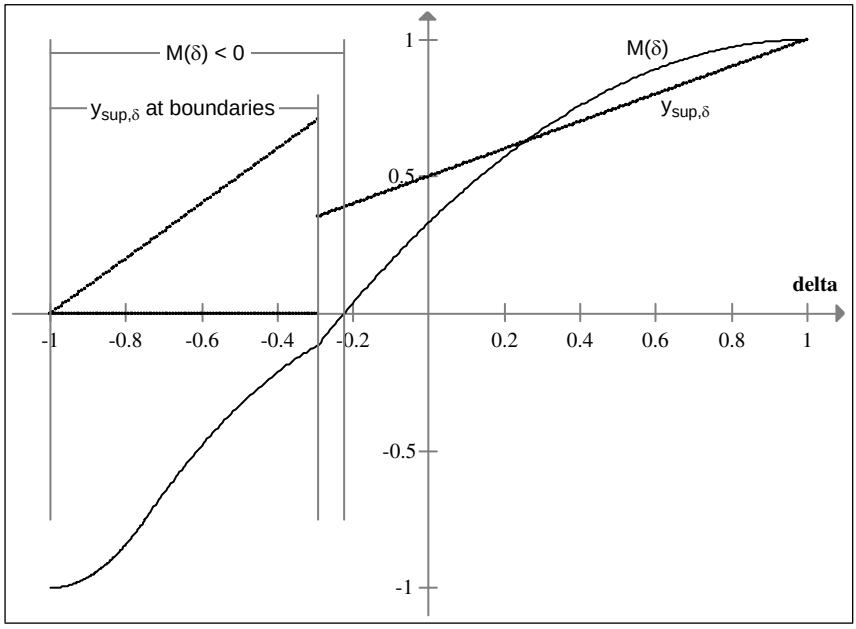


Fig. 3. Graphs of $M(\delta)$ and $y_{\sup,\delta}$: $(C(1/4), C(3/4))$.

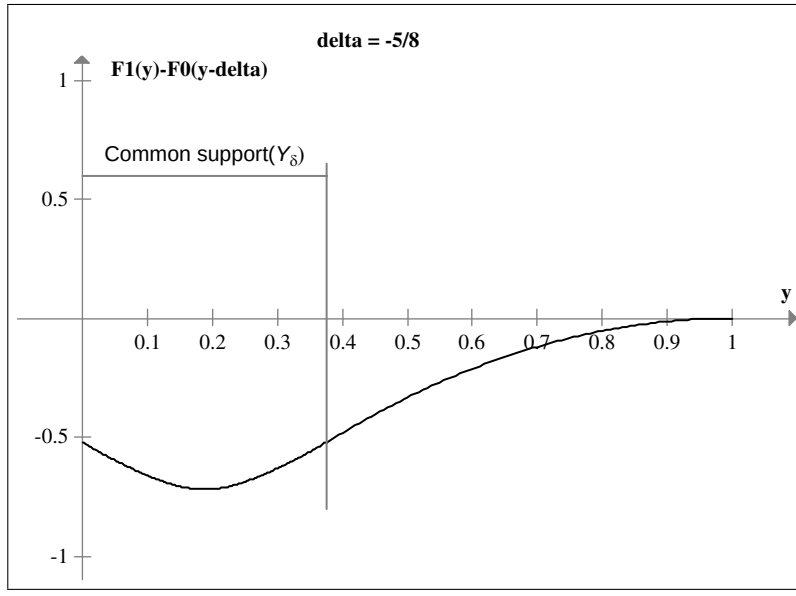
Fig. 4(d) a case where the interior solution corresponds to a positive value for M (δ).

Depending on the value of δ, $M(\delta)$ can have different signs leading to different asymptotic distributions for $F_n^{\mathrm{L}}(\delta)$. For example, when $\delta = 1-(\sqrt{6}/2)$ (Fig. 4(c)), $M(\delta) = 0$ and for $\delta > 1-(\sqrt{6}/2), M(\delta) > 0$. Since $M(\delta) = 0$ when $\delta = 1-(\sqrt{6}/2), y_{\sup,\delta} = 1-(\sqrt{6}/4)$ is in the interior, and $f_1'(y_{\sup,\delta}) - f_0'(y_{\sup,\delta} - \delta) = -(16/3) < 0$, Theorem 1 implies that at $\delta = 1-(\sqrt{6}/2)$,

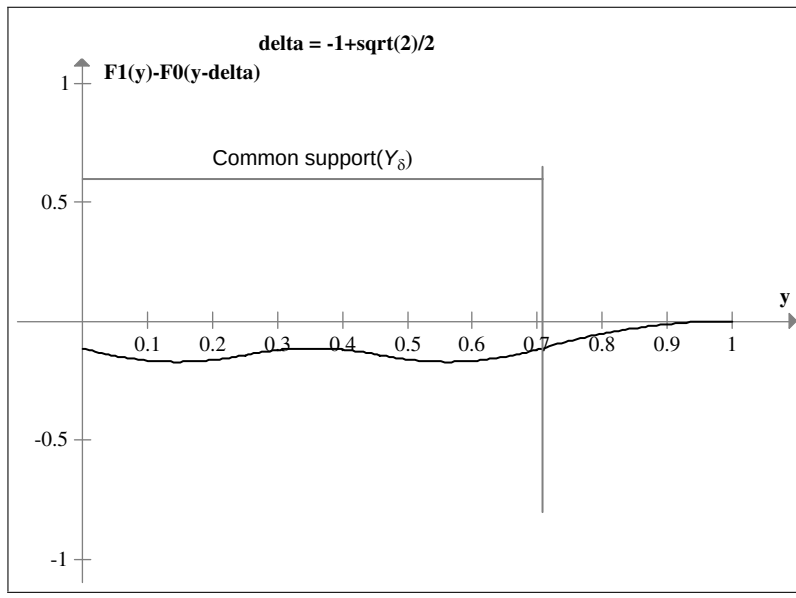
$$\sqrt{n_1}[F_n^{\mathrm{L}}(\delta) - F^{\mathrm{L}}(\delta)] \Rightarrow \max(N(0, \sigma_{\mathrm{L}}^2), 0) \quad \text{where} \quad \sigma_{\mathrm{L}}^2 = \frac{(1+\lambda)}{4}$$

When $\delta = 1/8$ (Fig. 4(d)),

$$y_{\sup,\delta} = \frac{9}{16}, M(\delta) = \frac{47}{96} > 0, \; f_1'(y_{\sup,\delta}) - f_0'(y_{\sup,\delta} - \delta) = -\frac{16}{3} < 0$$

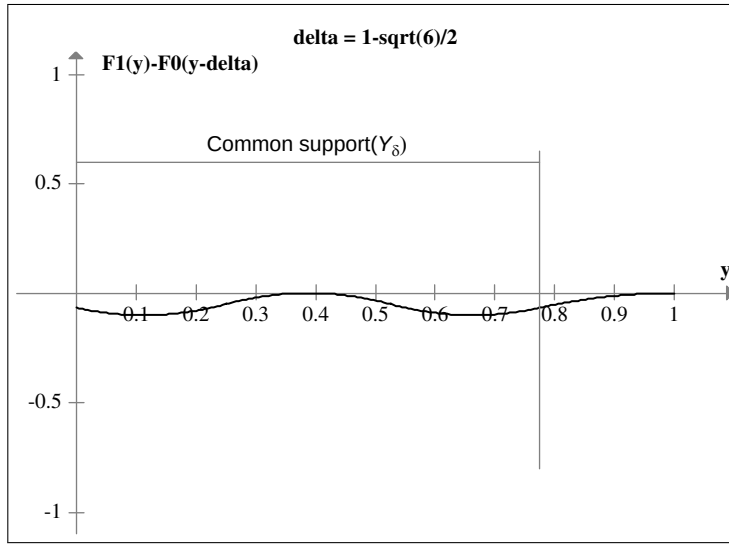


(a)

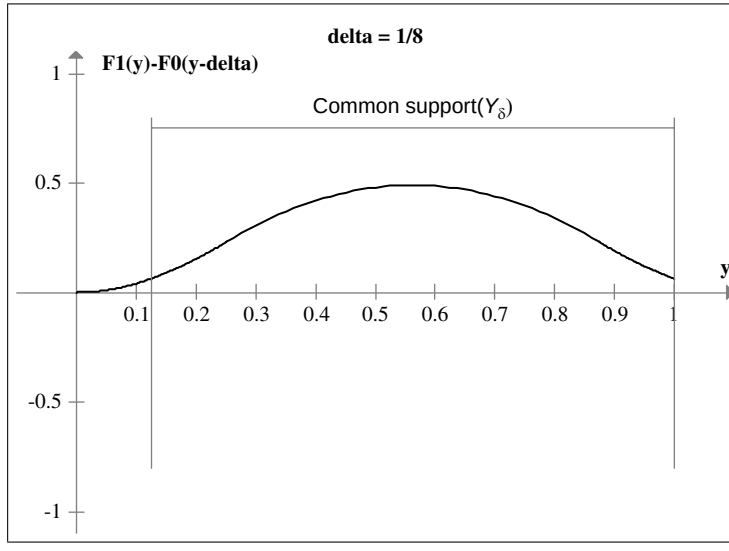


(b)

Fig. 4. Graphs of $[F_1(y) - F_0(y - \delta)]$ and Common Supports for Various δ; (a) $\delta = -(5/8)$; (b) $\delta = -1 + (\sqrt{2}/2)$; (c) $\delta = 1 - (\sqrt{6}/2)$; and (d) $\delta = 1/8$.



(c)



(d)

Fig. 4. (*Continued*)

Theorem 1 implies that when $\delta = 1/8$,

$$\sqrt{n_1}[F_n^{\mathrm{L}}(\delta) - F^{\mathrm{L}}(\delta)] \Rightarrow N(0, \sigma_{\mathrm{L}}^2) \quad \text{where} \quad \sigma_{\mathrm{L}}^2 = (1+\lambda)\frac{7,007}{36,864}$$

We now illustrate both possibilities for the upper bound $F^{\mathrm{U}}(\delta)$. Suppose $Y_1 \sim C(3/4)$ and $Y_0 \sim C(1/4)$. Then using the expressions for $F_1(y) - F_0(y-\delta)$ provided in Appendix B, we obtain

$$y_{\inf,\delta} = \begin{cases} \dfrac{1+\delta}{2} & \text{if } -1 \le \delta \le 1 - \dfrac{\sqrt{2}}{2} \\ \left\{\delta, \dfrac{1+\delta}{2}, 1\right\} & \text{if } \delta = 1 - \dfrac{\sqrt{2}}{2} \\ \{\delta, 1\} & \text{if } 1 - \dfrac{1}{2}\sqrt{2} \le z \le 1 \end{cases}$$

$$m(\delta) = \begin{cases} \dfrac{2}{3}(\delta+1)^2 - 1 & \text{if } -1 \le \delta \le 1 - \dfrac{\sqrt{2}}{2} \\ \dfrac{4\delta^2}{3} & \text{if } 1 - \dfrac{\sqrt{2}}{2} \le \delta \le \dfrac{3}{4} \\ -4(1-\delta)^2 + 1 & \text{if } \dfrac{3}{4} \le \delta \le 1 \end{cases}$$

Fig. 5 shows $y_{\inf,\delta}$ and $m(\delta)$.

Graphs of $F_1(y) - F_0(y-\delta)$ against y for selective δ's are presented in Fig. 6. Fig. 6(a) and (b) illustrate two cases each having a unique interior minimum, but in Fig. 6(a), $m(\delta)$ is negative and in Fig. 6(b), $m(\delta)$ is 0. Fig. 6(c) illustrates the case with multiple solutions: one interior minimizer and two boundary ones, while Fig. 6(d) illustrates the case with two boundary minima.

4.3. Asymptotic Distributions of $F_n^{\mathrm{L}}(\delta), F_n^{\mathrm{U}}(\delta)$ Without (A3) and (A4)

As Example 2 illustrates, assumptions (A3) and (A4) may be violated. Figs. 4 or 6 provide us with cases where multiple interior maximizers or minimizers exist. In Fig. 6(b) and (c), there are two interior maximizers when $\delta = (\sqrt{6}/2) - 1$ or $\delta = 1 - (\sqrt{2}/2)$ with $a_1 = 3/4$ and $a_0 = 1/4$. When $\delta = (\sqrt{6}/2) - 1$, $M(\delta) = (\sqrt{6-2})^2/2$ and $\mathcal{Y}_{\sup,\delta} = \{((6-\sqrt{6})/4), ((3\sqrt{6-6})/4)\}$. When $\delta = 1 - (\sqrt{2}/2), M(\delta) = ((2-\sqrt{2})^2)/2$ and $\mathcal{Y}_{\sup,\delta} = \{((\sqrt{2+2})/4), ((6-3\sqrt{2})/4)\}$. Shown in Fig. 4(b) and (c) are

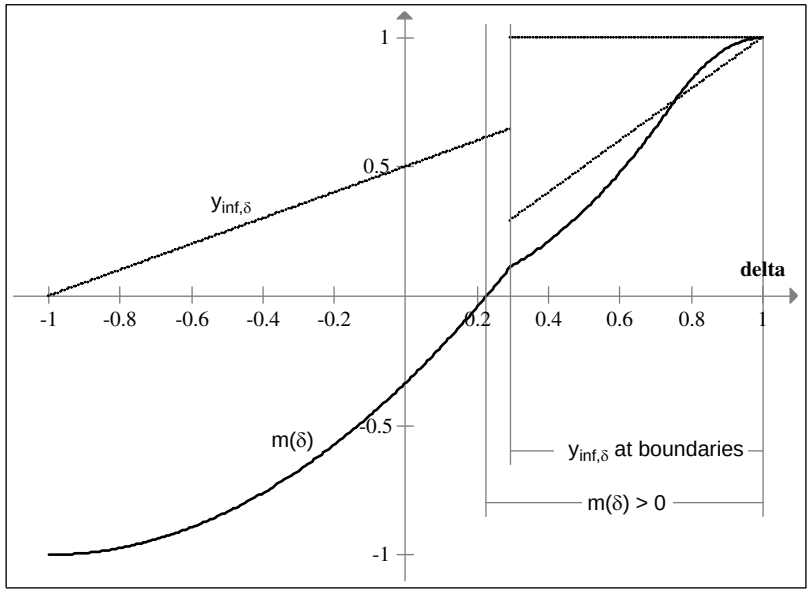


Fig. 5. Graphs of $m(\delta)$ and $y_{\inf,\delta} : (C(3/4), C(1/4))$.

cases with multiple interior minimizers for $a_1 = 1/4$ and $a_0 = 3/4$. When $\delta = (\sqrt{2}/2) - 1, m(\delta) = -((2-\sqrt{2})^2/2)$ and $\mathcal{Y}_{\inf,\delta} = \{((2-\sqrt{2})/4), ((3\sqrt{2}-2)/4)\}$. When $\delta = 1 - (\sqrt{6}/2), m(\delta) = -(\sqrt{6}-2)^2/2$ and $\mathcal{Y}_{\inf,\delta} = \{((\sqrt{6}-2)/4), ((10-3\sqrt{6})/4)\}$.

We now dispense with assumptions (A3) and (A4). Recall that

$$\mathcal{Y}_{\sup,\delta} = \{y \in \mathcal{Y}_\delta : F_1(y) - F_0(y-\delta) = M(\delta)\}$$

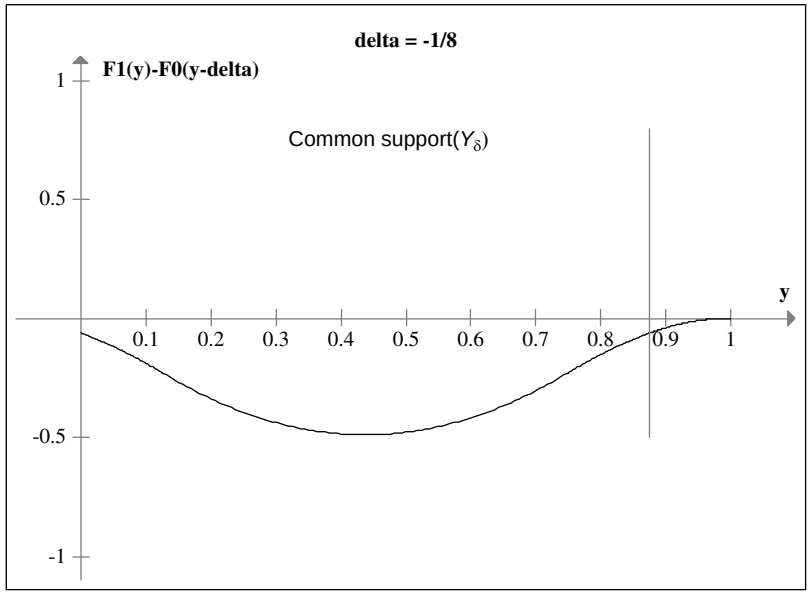
$$\mathcal{Y}_{\inf,\delta} = \{y \in \mathcal{Y}_\delta : F_1(y) - F_0(y-\delta) = m(\delta)\}$$

For a given $b>0$, define

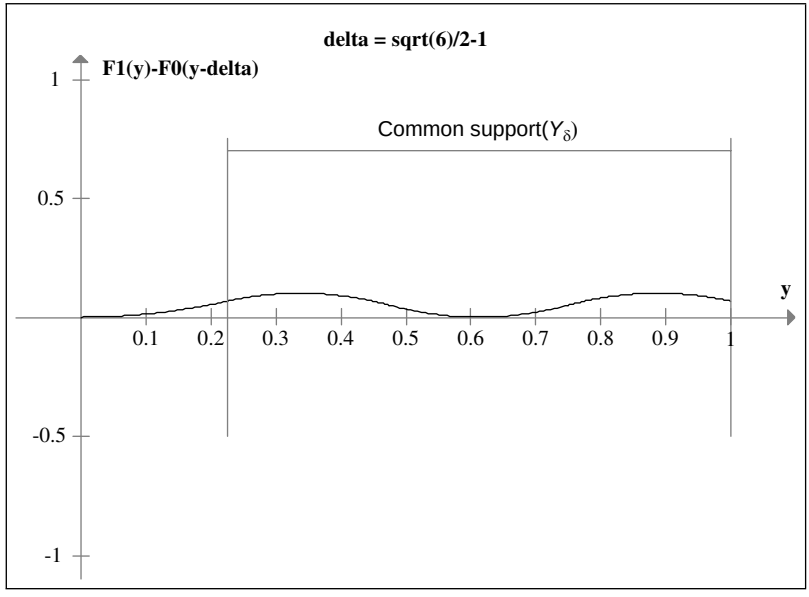
$$\mathcal{Y}^b_{\sup,\delta} = \{y \in \mathcal{Y}_\delta : F_1(y) - F_0(y-\delta) \geq M(\delta) - b\}$$

$$\mathcal{Y}^b_{\inf,\delta} = \{y \in \mathcal{Y}_\delta : F_1(y) - F_0(y-\delta) \leq m(\delta) + b\}$$

A3′. There exists $K>0$ and $0<\eta<1$ such that for all $y \in \mathcal{Y}^b_{\sup,\delta}$, for $b>0$ sufficiently small, there exists a $y_{\sup,\delta} \in \mathcal{Y}_{\sup,\delta}$ such that $y_{\sup,\delta} \leq y$ and $(y - y_{\sup,\delta}) \leq Kb^\eta$.

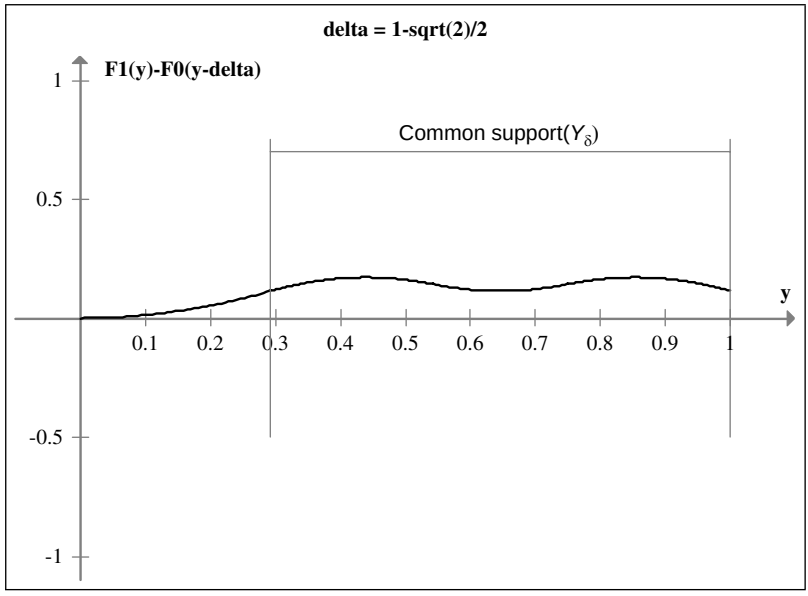


(a)

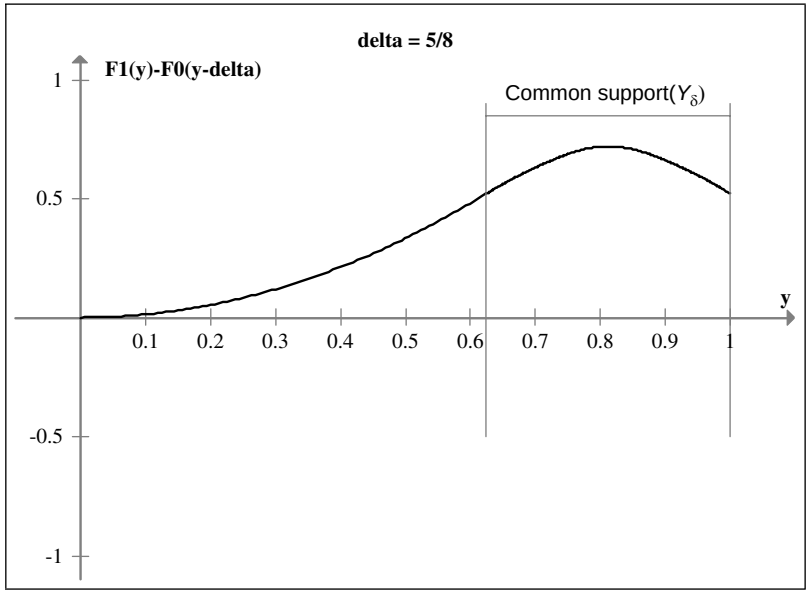


(b)

Fig. 6. Graphs of $[F_1(y) - F_0(y - \delta)]$ and Common Supports for Various δ; (a) $\delta = -(1/8)$; (b) $\delta = (\sqrt{6}/2) - 1$; (c) $\delta = 1 - (\sqrt{2}/2)$; and (d) $\delta = 5/8$.



(c)



(d)

Fig. 6. (*Continued*)

A4′. There exists $K>0$ and $0<\eta<1$ such that for all $y \in \mathcal{Y}^b_{\text{inf},\delta}$ for $b>0$ sufficiently small, there exists a $y_{\text{inf},\delta} \in \mathcal{Y}_{\text{inf},\delta}$ such that $y_{\text{inf},\delta} \leq y$ and $(y - y_{\text{inf},\delta}) \leq Kb^\eta$.

Assumptions (A3)′ and (A4)′ adapt Assumption (1) in Galichon and Henry (2009). As discussed in Galichon and Henry (2009), they are very mild assumptions. By following the proof of Theorem 1 in Galichon and Henry (2009), we can show that under conditions stated in the theorem below,

$$\sqrt{n_1}[M_n(\delta) - M(\delta)] \Rightarrow \sup_{y \in \mathcal{Y}_{\text{sup},\delta}} G(y,\delta), \sqrt{n_1}[m_n(\delta) - m(\delta)] \Rightarrow \inf_{y \in \mathcal{Y}_{\text{inf},\delta}} G(y,\delta)$$

where $\{G(y,\delta) : y \in \mathcal{Y}_\delta\}$ is a tight Gaussian process with zero mean. Thus the theorem below holds.

Theorem 2.

(i) Suppose (A1) and (A3)′ hold. For any $\delta \in [a-d, b-c] \cap \mathcal{R}$, we have

$$\sqrt{n_1}[F^{\text{L}}_n(\delta) - F^{\text{L}}(\delta)] \Rightarrow \begin{cases} \sup_{y \in \mathcal{Y}_{\text{sup},\delta}} G(y,\delta), & \text{if } M(\delta)>0 \\ \max\{\sup_{y \in \mathcal{Y}_{\text{sup},\delta}} G(y,\delta), 0\} & \text{if } M(\delta)=0 \end{cases}$$

$$\text{and} \quad \Pr(F^{\text{L}}_n(\delta)=0) \to 1 \text{ if } M(\delta)<0$$

where $\{G(y,\delta) : y \in \mathcal{Y}_\delta\}$ is a tight Gaussian process with zero mean.

(ii) Suppose (A1) and (A4)′ hold. For any $\delta \in [a-d, b-c] \cap \mathcal{R}$, we get

$$\sqrt{n_1}[F^{\text{U}}_n(\delta) - F^{\text{U}}(\delta)] \Rightarrow \begin{cases} \inf_{y \in \mathcal{Y}_{\text{inf},\delta}} G(y,\delta), & \text{if } m(\delta)<0 \\ \min\{\inf_{y \in \mathcal{Y}_{\text{inf},\delta}} G(y,\delta), 0\} & \text{if } m(\delta)=0 \end{cases}$$

$$\text{and} \quad \Pr(F^{\text{U}}_n(\delta)=1) \to 1 \text{ if } m(\delta)>0$$

When (A3) and (A4) hold, $\mathcal{Y}_{\text{sup},\delta}$ and $\mathcal{Y}_{\text{inf},\delta}$ are singletons and Theorem 2 reduces to Theorem 1.

5. CONFIDENCE SETS FOR THE DISTRIBUTION OF TREATMENT EFFECTS FOR RANDOMIZED EXPERIMENTS

5.1. Confidence Sets for the Sharp Bounds

First, we consider the lower bound. Let

$$G_n(y,\delta)=\sqrt{n_1}[F_{1n}(y)-F_1(y)]-\sqrt{n_1}[F_{0n}(y-\delta)-F_0(y-\delta)]$$

Then

$$\begin{aligned}&\sqrt{n_1}[F_n^{\mathrm{L}}(\delta)-F^{\mathrm{L}}(\delta)]\\&=\max\left\{\sup_{y\in\mathcal{Y}_\delta}\{G_n(y,\delta)+\sqrt{n_1}[F_1(y)-F_0(y-\delta)]\},0\right\}-\max\{\sqrt{n_1}M(\delta),0\}\\&\Rightarrow\max\left\{\sup_{y\in\mathcal{Y}_\delta}[G(y,\delta)+h_{\mathrm{L}}(y,\delta)]+\min\{h_{\mathrm{L}}(\delta),0\},-\max\{h_{\mathrm{L}}(\delta),0\}\right\}(\equiv W^1_{\mathrm{L},\delta})\end{aligned}\tag{19}$$

$$=\max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}}G(y,\delta)+\min\{h_{\mathrm{L}}(\delta),0\},-\max\{h_{\mathrm{L}}(\delta),0\}\right\}(\equiv W^2_{\mathrm{L},\delta})\tag{20}$$

where $h_{\mathrm{L}}(y,\delta)=\lim\sqrt{n_1}[F_1(y)-F_0(y-\delta)-M(\delta)]\le 0$ and $h_{\mathrm{L}}(\delta)=\lim[\sqrt{n_1}M(\delta)]$.

Define $h^*_{\mathrm{L}}(\delta)=\sqrt{n_1}M_n(\delta)I\{|M_n(\delta)|>b_n\}$ and

$$\begin{aligned}h^*_{\mathrm{L}}(y,\delta)=\sqrt{n_1}[F_{1n}(y)-F_{0n}(y-\delta)-M_n(\delta)]I\{[F_{1n}(y)\\-F_{0n}(y-\delta)-M_n(\delta)]<-b'_n\}\end{aligned}$$

where b_n is a prespecified deterministic sequence satisfying $b_n\to 0$ and $\sqrt{n_1}b_n\to\infty$ and b'_n is a prespecified deterministic sequence satisfying $b'_n\ln\ln\ n_1+(\sqrt{n_1}b')^{-1}\sqrt{\ln\ln\ n_1}\to 0$. In the simulations, we considered $b_n=cn_1^{-a},0<a<(1/2),c>0$ and $b'_n=c'n_1^{-(1-a')/2},0<a'<1,c'>0$. For such b'_n, we have

$$b'_n\ln\ln\ n_1+(\sqrt{n_1}b'_n)^{-1}\sqrt{\ln\ln\ n_1}=c'\frac{\ln\ln\ n_1}{\sqrt{n_1^{1-a'}}}+\frac{1}{c'}\frac{\sqrt{\ln\ln\ n_1}}{\sqrt{n_1^{a'}}}\to 0$$

Based on Eqs. (19) and (20), we propose two bootstrap procedures to approximate the distribution of $\sqrt{n_1}[F_n^{\mathrm{L}}(\delta)-F^{\mathrm{L}}(\delta)]$. In the first procedure,

we approximate the distribution of $W^1_{\mathrm{L},\delta}$ and in the second procedure, we approximate the distribution of $W^2_{\mathrm{L},\delta}$. Draw bootstrap samples with replacement from $\{Y_{1i}\}_{i=1}^{n_1}$ and $\{Y_{0i}\}_{i=1}^{n_0}$, respectively. Let $F^*_{1n}(y)$, $F^*_{0n}(y)$ denote the empirical distribution functions based on the bootstrap samples, respectively. Define

$$G^*_n(y,\delta) = \sqrt{n_1}[F^*_{1n}(y) - F_{1n}(y)] - \sqrt{n_1}[F^*_{0n}(y-\delta) - F_{0n}(y-\delta)]$$

In the first bootstrap approach, we use the distribution of the following random variable conditional on the original sample to approximate the quantiles of the limiting distribution of $\sqrt{n_1}[F^{\mathrm{L}}_n(\delta) - F^{\mathrm{L}}(\delta)]$:

$$W^{1*}_{\mathrm{L},\delta} = \max\left\{\sup_{y\in\mathcal{Y}_\delta}\{G^*_n(y,\delta) + h^*_{\mathrm{L}}(y,\delta)\} + \min\{h^*_{\mathrm{L}}(\delta),0\}, -\max\{h^*_{\mathrm{L}}(\delta),0\}\right\}$$

In the second bootstrap approach, we estimate $\mathcal{Y}_{\sup,\delta}$ directly and approximate the distributions of $W_{\mathrm{L},\delta}$. Define

$$\mathcal{Y}_{n\sup,\delta} = \{y_i \in \{Y_{1i}\}_{i=1}^{n_1} \cup \{Y_{0i}\}_{i=1}^{n_0} : M_n(\delta) - (F_{n1}(y_i) - F_{n0}(y_i-\delta)) \le b'_n\}$$

Then the distribution of the following random variable conditional on the original sample can be used to approximate the quantiles of the limiting distribution of $\sqrt{n_1}[F^{\mathrm{L}}_n(\delta) - F^{\mathrm{L}}(\delta)]$:

$$W^{2*}_{\mathrm{L},\delta} = \max\left\{\sup_{y\in\mathcal{Y}_{n\ \sup,\delta}} G^*_n(y,\delta), -h^*_{\mathrm{L}}(\delta)\right\} + \min\{h^*_{\mathrm{L}}(\delta),0\}$$

The upper bound can be dealt with similarly. Note that

$$\begin{aligned}
&\sqrt{n_1}[F^{\mathrm{U}}_n(\delta) - F^{\mathrm{U}}(\delta)] \\
&\quad\Rightarrow \min\left\{\inf_{y\in\mathcal{Y}_\delta}\{G_n(y,\delta) + h_{\mathrm{U}}(y,\delta)\} + \max\{h_{\mathrm{U}}(\delta),0\}, -\min\{h_{\mathrm{U}}(\delta),0\}\right\} \\
&\quad\Rightarrow \min\left\{\inf_{y\in\mathcal{Y}_\delta}[G(y,\delta) + h_{\mathrm{U}}(y,\delta)] + \max\{h_{\mathrm{U}}(\delta),0\}, -\min\{h_{\mathrm{U}}(\delta),0\}\right\} (\equiv W^1_{\mathrm{U},\delta}) \\
&\quad= \min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta) + \max\{h_{\mathrm{U}}(\delta),0\}, -\min\{h_{\mathrm{U}}(\delta),0\}\right\} (\equiv W^2_{\mathrm{U},\delta})
\end{aligned}$$

where $h_{\mathrm{U}}(y,\delta) = \lim\sqrt{n_1}[F_1(y) - F_0(y-\delta) - m(\delta)] \ge 0$ and $h_{\mathrm{U}}(\delta) = \lim[\sqrt{n_1}m(\delta)]$.

Define $h^*_{\mathrm{U}}(\delta) = \sqrt{n_1} m_n(\delta) I\{|m_n(\delta)| > b_n\}$ and

$$h^*_{\mathrm{U}}(y,\delta) = \sqrt{n_1}[F_{1n}(y) - F_{0n}(y-\delta) - m_n(\delta)] I\{[F_{1n}(y) - F_{0n}(y-\delta) - m_n(\delta)] > b'_n\}$$

We propose to use the distribution of $W^{1*}_{\mathrm{U},\delta}$ or $W^{2*}_{n\mathrm{U},\delta}$ conditional on the original sample to approximate the quantiles of the distribution of $\sqrt{n_1}[F^{\mathrm{U}}_n(\delta) - F^{\mathrm{U}}(\delta)]$, where

$$W^{1*}_{\mathrm{U},\delta} = \min\left\{ \inf_{y \in \mathcal{Y}_\delta} \{G^*_n(y,\delta) + h^*_{\mathrm{U}}(y,\delta)\} + \max\{h^*_{\mathrm{U}}(\delta), 0\}, -\min\{h^*_{\mathrm{U}}(\delta), 0\} \right\}$$

$$W^{2*}_{\mathrm{U},\delta} = \min\left\{ \inf_{y \in \mathcal{Y}_{n\inf\,\delta}} G^*_n(y,\delta), -h^*_{\mathrm{U}}(\delta) \right\} + \max\{h^*_{\mathrm{U}}(\delta), 0\}$$

in which

$$\mathcal{Y}_{n\inf,\,\delta} = \{y_i \in \{Y_{1i}\}_{i=1}^{n_1} \cup \{Y_{0i}\}_{i=1}^{n_0} : m_n(\delta) - (F_{n1}(y_i) - F_{n0}(y_i - \delta)) \geq -b'_n\}$$

Throughout the simulations presented in Section 7, we used $W^{2*}_{\mathrm{L},\delta}$ and $W^{2*}_{\mathrm{U},\delta}$.

5.2. *Confidence Sets for the Distribution of Treatment Effects*

For notational simplicity, we let $\theta_0 = F_\Delta(\delta)$, $\theta_{\mathrm{L}} = F^{\mathrm{L}}(\delta)$, and $\theta_{\mathrm{U}} = F^{\mathrm{U}}(\delta)$. Also let $\Theta = [0, 1]$. This subsection follows similar ideas to Fan and Park (2007b). Noting that

$$\theta_0 = \arg\min_{\theta \in \Theta} \{(\theta_{\mathrm{L}} - \theta)^2_+ + (\theta_{\mathrm{U}} - \theta)^2_-\}$$

where $(x)_- = \min\{x, 0\}$ and $(x)_+ = \max\{x, 0\}$, we define the test statistic

$$T_n(\theta_0) = n_1(\hat{\theta}_L - \theta_0)^2_+ + n_1(\hat{\theta}_{\mathrm{U}} - \theta_0)^2_- \tag{21}$$

where $\hat{\theta}_{\mathrm{L}} = F^{\mathrm{L}}_n(\delta)$ and $\hat{\theta}_{\mathrm{U}} = F^{\mathrm{U}}_n(\delta)$. Then a $(1-\alpha)$ level CS for θ_0 can be constructed as,

$$\mathrm{CS}_n = \{\theta \in \Theta : T_n(\theta) \leq c_{1-\alpha}(\theta)\} \tag{22}$$

for an appropriately chosen critical value $c_{1-\alpha}(\theta)$.

To determine the critical value $c_{1-\alpha}(\theta)$, the limiting distribution of $T_n(\theta)$ under an appropriate local sequence is essential. We introduce some necessary notation. Let

$$h^{\mathrm{L}}(\theta_0) = -\lim_{n\to\infty} \sqrt{n}[\theta_{\mathrm{L}} - \theta_0] \quad \text{and} \quad h^{\mathrm{U}}(\theta_0) = \lim_{n\to\infty} \sqrt{n}[\theta_{\mathrm{U}} - \theta_0]$$

Then $h^{\mathrm{L}}(\theta_0) \geq 0, h^{\mathrm{U}}(\theta_0) \geq 0$, and $h^{\mathrm{L}}(\theta_0) + h^{\mathrm{U}}(\theta_0) = \lim_{n\to\infty}(\sqrt{n}\nabla)$, where $\nabla \equiv \theta_{\mathrm{U}} - \theta_{\mathrm{L}}$ is the length of the identified interval. As proposed in Fan and Park (2007b), we use the following shrinkage "estimators" of $h^{\mathrm{L}}(\theta_0)$ and $h^{\mathrm{U}}(\theta_0)$.

$$h^{\mathrm{L}*}(\theta_0) = -\sqrt{n}[\widehat{\theta}_{\mathrm{L}} - \theta_0]I\{[\theta_0 - \widehat{\theta}_{\mathrm{L}}] > b_n\}$$

$$h^{\mathrm{U}*}(\theta_0) = \sqrt{n}[\widehat{\theta}_{\mathrm{U}} - \theta_0]I\{[\widehat{\theta}_{\mathrm{U}} - \theta_0] > b_n\}$$

It remains to establish the asymptotic distribution of $T_n(\theta_0)$:

$$\begin{aligned} T_n(\theta_0) &= (\sqrt{n_1}[\widehat{\theta}_{\mathrm{L}} - \theta_{\mathrm{L}}] - \sqrt{n_1}[\theta_0 - \theta_L])_+^2 + (\sqrt{n_1}[\widehat{\theta}_{\mathrm{U}} - \theta_{\mathrm{U}}] + \sqrt{n_1}[\theta_{\mathrm{U}} - \theta_0])_-^2 \\ &\Rightarrow (W_{\mathrm{L},\delta} - h^{\mathrm{L}}(\theta_0))_+^2 + (W_{\mathrm{U},\delta} - h^{\mathrm{U}}(\theta_0))_-^2 \end{aligned}$$

Let

$$T_n^*(\theta_0) = (W_{\mathrm{L},\delta}^* - h^{\mathrm{L}}(\theta_0))_+^2 + (W_{\mathrm{U},\delta}^* - h^{\mathrm{U}}(\theta_0))_-^2$$

and $cv_{1-\alpha}^*(h^{\mathrm{L}}(\theta_0), h^{\mathrm{U}}(\theta_0))$ denote the $1-\alpha$ quantile of the bootstrap distribution of $T_n^*(\theta_0)$, where $W_{\mathrm{L},\delta}^*$ and $W_{\mathrm{U},\delta}^*$ are either $W_{\mathrm{L},\delta}^{1*}$ and $W_{\mathrm{U},\delta}^{1*}$ or $W_{\mathrm{L},\delta}^{2*}$ and $W_{\mathrm{U},\delta}^{2*}$ defined in the previous subsection. The following theorem holds for a $\underline{p} \in [0, 1]$.

Theorem 3. Suppose (A1), (A3)′, and (A4)′ hold. Then, for $\alpha \in [0, \underline{p}]$,

$$\lim_{n_1\to\infty} \inf_{\theta_0 \in [\theta_{\mathrm{L}}, \theta_{\mathrm{U}}]} \Pr(\theta_0 \in \{\theta : T_n(\theta) \leq cv_{1-\alpha}^*(h^{\mathrm{L}*}(\theta), h^{\mathrm{U}*}(\theta))\}) \geq 1 - \alpha$$

The coverage rates presented in Section 7 are results of the confidence sets of Theorem 3. The presence of $\underline{p}$ in Theorem 3 is due to the fact that $T_n(\theta_0)$ is nonnegative and so is $cv_{1-\alpha}^*(h^{\mathrm{L}*}(\theta), h^{\mathrm{U}*}(\theta))$. In Appendix A, we show that one can take $\underline{p}$ as,

$$\underline{p} = 1 - \Pr\left[\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta) \leq 0, \inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta) \geq 0\right] \tag{23}$$

In actual implementation, $\underline{p}$ has to be estimated. We propose the following estimator $\hat{\underline{p}}$:

$$\hat{\underline{p}} = 1 - \frac{1}{B}\sum_{b=1}^{B} 1\left\{\sup_{y\in\mathcal{Y}_{n\sup,\delta}} G_n^{(b)}(y,\delta) \leq 0, \inf_{y\in\mathcal{Y}_{n\inf,\delta}} G_n^{(b)}(y,\delta) \geq 0\right\}$$

where $G_n^{(b)}(y,\delta)$ is $G_n^*(y,\delta)$ from bth bootstrap samples.

6. BIAS-CORRECTED ESTIMATORS OF SHARP BOUNDS ON THE DISTRIBUTION OF TREATMENT EFFECTS FOR RANDOMIZED EXPERIMENTS

In this section, we demonstrate that the plug-in estimators $F_n^{\mathrm{L}}(\delta)$, $F_n^{\mathrm{U}}(\delta)$ tend to have nonnegligible bias in finite samples. In particular, $F_n^{\mathrm{L}}(\delta)$ tends to be biased upward and $F_n^{\mathrm{U}}(\delta)$ tends to be biased downward. We show this analytically when (A3) and (A4) hold. In particular, when (A3) and (A4) hold, we provide closed-form expressions for the first-order asymptotic biases of $F_n^{\mathrm{L}}(\delta), F_n^{\mathrm{U}}(\delta)$ and use these expressions to construct bias-corrected estimators for $F^{\mathrm{L}}(\delta)$ and $F^{\mathrm{U}}(\delta)$. When (A3) and (A4) fail, we propose bootstrap bias-corrected estimators of the sharp bounds $F^{\mathrm{L}}(\delta)$ and $F^{\mathrm{U}}(\delta)$.

Recall

$$F_n^{\mathrm{L}}(\delta) = \max\{M_n(\delta), 0\} \quad \text{and} \quad F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\}$$

$$F_n^{\mathrm{U}}(\delta) = 1 + \min\{m_n(\delta), 0\} \quad \text{and} \quad F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\}$$

where under (A3) and (A4), we have

$$\sqrt{n_1}(M_n(\delta) - M(\delta)) \Rightarrow N(0, \sigma_{\mathrm{L}}^2) \quad \text{and} \quad \sqrt{n_1}(m_n(\delta) - m(\delta)) \Rightarrow N(0, \sigma_{\mathrm{U}}^2)$$

First, we consider the lower bound. Ignoring the second-order terms, we get:

$$\begin{aligned}
E[F_n^{\mathrm{L}}(\delta)] &= E[M_n(\delta) I_{\{M_n(\delta)\geq 0\}}] \\
&= E\left[\left\{M(\delta) + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}} Z\right\} I_{\{M(\delta)+(\sigma_{\mathrm{L}}/\sqrt{n_1})Z\geq 0\}}\right] \quad \text{where} \quad Z \sim N(0,1) \\
&= M(\delta) E[I_{\{M(\delta)+(\sigma_{\mathrm{L}}/\sqrt{n_1})Z\geq 0\}}] + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}} E[Z I_{\{M(\delta)+(\sigma_{\mathrm{L}}/\sqrt{n_1})Z\geq 0\}}] \\
&= M(\delta) E[I_{\{z\geq -(\sqrt{n_1}/\sigma_{\mathrm{L}})M(\delta)\}}] + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}} E[Z I_{\{Z\geq -(\sqrt{n_1}/\sqrt{\sigma_{\mathrm{L}}})M(\delta)\}}] \\
&= M(\delta) \int_{-(\sqrt{n_1}/\sigma_{\mathrm{L}})M(\delta)}^{\infty} \phi(z)dz + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}} \int_{-(\sqrt{n_1}/\sigma_{\mathrm{L}})M(\delta)}^{\infty} z\phi(z)dz \\
&= M(\delta)\left\{1 - \Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}} M(\delta)\right)\right\} \\
&\quad - \frac{1}{\sqrt{2\pi}} \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}} \int_{-(\sqrt{n_1}/\sigma_{\mathrm{L}})M(\delta)}^{\infty} \exp\left(-\frac{z^2}{2}\right) d\left(-\frac{z^2}{2}\right) \\
&= M(\delta)\Phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}} M(\delta)\right) + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}} \phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}} M(\delta)\right)
\end{aligned}$$

Case I. Suppose $M(\delta) \geq 0$. Then ignoring second-order terms, we obtain

$$\begin{aligned} E[F_n^{\mathrm{L}}(\delta)] - F^{\mathrm{L}}(\delta) &= M(\delta)\Phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}}\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) - M(\delta) \\ &= M(\delta)\left\{\Phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) - 1\right\} + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}}\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) \\ &= -M(\delta)\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}}\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) \\ &= \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}}\left\{\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right)\right\} \\ &> 0 \text{ (positive bias)} \end{aligned}$$

because

$$\lim_{x\to 0}\{\phi(-x) - x\Phi(-x)\} = \phi(0) = \frac{1}{\sqrt{2\pi}}$$

$$\begin{aligned} \lim_{x\to+\infty}\{\phi(-x) - x\Phi(-x)\} &= \lim_{x\to-\infty}\{\phi(x) + x\Phi(x)\} = \lim_{x\to-\infty}\frac{d}{dx}\left(\frac{\Phi(x)}{x^{-1}}\right) \\ &= -\lim_{x\to-\infty}\left(\frac{\Phi(x)}{x^{-2}}\right) = 0 \end{aligned}$$

$$\frac{d}{dx}\{\phi(-x) - x\Phi(-x)\} = -\Phi(-x) < 0 \text{ for all } x \in R_+ \cap \{0\}$$

Case II. Suppose $M(\delta) < 0$. Then ignoring second-order terms, we obtain

$$\begin{aligned} E[F_n^{\mathrm{L}}(\delta)] - F^{\mathrm{L}}(\delta) &= M(\delta)\Phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) + \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}}\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) \\ &= \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}}\left\{\phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right) + \frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\Phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}M(\delta)\right)\right\} \\ &= \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}}\left\{\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}|M(\delta)|\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}|M(\delta)|\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}|M(\delta)|\right)\right\} \\ &> 0 \text{ (positive bias)} \end{aligned}$$

Summarizing Case I and Case II, we obtain the first-order asymptotic bias of $F_n^{\mathrm{L}}(\delta)$:

$$E[F_n^{\mathrm{L}}(\delta)] - F^{\mathrm{L}}(\delta) = \frac{\sigma_{\mathrm{L}}}{\sqrt{n_1}}\left\{\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}|M(\delta)|\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}|M(\delta)|\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}}}|M(\delta)|\right)\right\}$$

regardless of the sign of $M(\delta)$, an estimator of which is

$$\widehat{\mathrm{Bias}}_{\mathrm{L}} = \frac{\sigma_{\mathrm{L}n}}{\sqrt{n_1}}\left\{\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}n}}|M_n^*(\delta)|\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{L}n}}|M_n^*(\delta)|\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}n}}|M_n^*(\delta)|\right)\right\}$$

where $M_n^*(\delta) = M_n(\delta)I\{|M_n(\delta)| > b_n\}$ in which $b_n \to 0$ and $\sqrt{n_1}b_n \to \infty$. We define the bias-corrected estimator of $F^{\mathrm{L}}(\delta)$ as,

$$\begin{aligned} F_{n\mathrm{BC}}^{\mathrm{L}}(\delta) &= \max\{F_n^{\mathrm{L}}(\delta) - \widehat{\mathrm{Bias}}_{\mathrm{L}}, 0\} \\ &= \max\left\{F_n^{\mathrm{L}}(\delta) - \frac{\sigma_{\mathrm{L}n}}{\sqrt{n_1}}\left\{\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}n}}|M_n^*(\delta)|\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{L}n}}|M_n^*(\delta)|\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{L}n}}|M_n^*(\delta)|\right)\right\}, 0\right\} \\ &\le F_n^{\mathrm{L}}(\delta) \end{aligned}$$

Now consider the upper bound. The following holds:

$$\begin{aligned} E[F_n^{\mathrm{U}}(\delta)] &= 1 + E[m_n(\delta)I_{\{m_n(\delta)\le 0\}}] \\ &= 1 + E\left[\left\{m(\delta) + \frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}Z\right\}I_{\{m(\delta)+(\sigma_{\mathrm{U}}/\sqrt{n_1})Z\le 0\}}\right] \\ &= 1 + m(\delta)E[I_{\{m(\delta)+(\sigma_{\mathrm{U}}/\sqrt{n_1})Z\le 0\}}] + \frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}E[ZI_{\{m(\delta)+(\sigma_{\mathrm{U}}/\sqrt{n_1})Z\le 0\}}] \\ &= 1 + m(\delta)\int_{-\infty}^{-(\sqrt{n_1}/\sigma_{\mathrm{U}})m(\delta)}\phi(z)dz + \frac{1}{\sqrt{2\pi}}\frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\int_{-\infty}^{-(\sqrt{n_1}/\sigma_{\mathrm{U}})m(\delta)} z\exp\left(-\frac{z^2}{2}\right)dz \\ &= 1 + m(\delta)\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\sigma)\right) - \frac{1}{\sqrt{2\pi}}\frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\int_{-\infty}^{-(\sqrt{n_1}/\sigma_{\mathrm{U}})m(\delta)}\exp\left(-\frac{z^2}{2}\right)d\left(-\frac{z^2}{2}\right) \\ &= 1 + m(\delta)\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) - \frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) \end{aligned}$$

Case I. Suppose $m(\delta) \leq 0$. Then ignoring second-order terms, we obtain

$$\begin{aligned} E[F_n^{\mathrm{U}}(\delta)] - F^{\mathrm{U}}(\delta) &= m(\delta)\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) - \frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) - m(\delta) \\ &= -m(\delta)\Phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) - \frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) \\ &= -m(\delta)\Phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) - \frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\phi\left(\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) \\ &= -\frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\left(\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}|m(\delta)|\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}|m(\delta)|\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}|m(\delta)|\right)\right) \\ &< 0 \text{ (negative bias)} \end{aligned}$$

Case II. Suppose $m(\delta) > 0$. Then ignoring second-order terms, we obtain

$$\begin{aligned} E[F_n^{\mathrm{U}}(\delta)] - F^{\mathrm{U}}(\delta) &= m(\delta)\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) - \frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) \\ &= -\frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\left(\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}m(\delta)\right)\right) \\ &< 0 \,(\text{negative bias}) \end{aligned}$$

Therefore, the first-order asymptotic bias of $F_n^{\mathrm{U}}(\delta)$ is given by:

$$E[F_n^{\mathrm{U}}(\delta)] - F^{\mathrm{U}}(\delta) = -\frac{\sigma_{\mathrm{U}}}{\sqrt{n_1}}\left(\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}|m(\delta)|\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}|m(\delta)|\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}}}|m(\delta)|\right)\right)$$

regardless of the sign of $m(\delta)$, an estimator of which is

$$\widehat{\mathrm{Bias}}_{\mathrm{U}} = -\frac{\sigma_{\mathrm{U}n}}{\sqrt{n_1}}\left(\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}n}}|m_n^*(\delta)|\right) - \frac{\sqrt{n_1}}{\sigma_{\mathrm{U}n}}|m_n^*(\delta)|\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}n}}|m_n^*(\delta)|\right)\right)$$

where $m_n^*(\delta) = m_n(\delta)I\{|m_n(\delta)| > b_n\}$. A bias corrected estimator of $F^{\mathrm{U}}(\delta)$ is defined as,

$$\begin{aligned} F_{n\mathrm{BC}}^{\mathrm{U}}(\delta) = \min\{F_n^{\mathrm{U}}(\delta) - \widehat{\mathrm{Bias}}, 1\} = \min\Bigg\{ & F_n^{\mathrm{U}}(\delta) + \frac{\sigma_{\mathrm{U}n}}{\sqrt{n_1}}\left(\phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}n}}|m_n^*(\delta)|\right)\right. \\ & \left. - \frac{\sqrt{n_1}}{\sigma_{\mathrm{U}n}}|m_n^*(\delta)|\Phi\left(-\frac{\sqrt{n_1}}{\sigma_{\mathrm{U}n}}|m_n^*(\delta)|\right)\right), 1\Bigg\} \geq F_n^{\mathrm{U}}(\delta) \end{aligned}$$

The bias-corrected estimators we just proposed depend on the validity of (A3) and (A4). Without these assumptions, the analytical expressions

derived for the bias may not be correct. Instead, we propose the following bootstrap bias-corrected estimators. Define

$$\widehat{\text{Bias}}(F_n^{\text{L}}(\delta)) = \frac{1}{B}\sum_{b=1}^{B}\frac{W_{\text{L},\delta}^{(b)}}{\sqrt{n_1}} \quad \text{and} \quad \widehat{\text{Bias}}(F_n^{\text{U}}(\delta)) = \frac{1}{B}\sum_{b=1}^{B}\frac{W_{\text{U},\delta}^{(b)}}{\sqrt{n_1}}$$

where $W_{\text{L},\delta}^{(b)}(W_{\text{U},\delta}^{(b)})$ are $W_{\text{L},\delta}^{F*}(W_{\text{U},\delta}^{F*})$ or $W_{\text{L},\delta}^{*}(W_{\text{U},\delta}^{*})$ from bth bootstrap samples, where $W_{\text{L},\delta}^{F*}$, $W_{\text{U},\delta}^{F*}$, $W_{\text{L},\delta}^{*}$, and $W_{\text{U},\delta}^{*}$ are defined in the previous subsections. The bootstrap bias-corrected estimators of $F^{\text{L}}(\delta)$ and $F^{\text{U}}(\delta)$ are, respectively,

$$\widehat{F}_{n\text{BC}}^{\text{L}}(\delta) = \max\{F_n^{\text{L}}(\delta) - \widehat{\text{Bias}}(F_n^{\text{L}}(\delta)), 0\} \quad \text{and}$$
$$\widehat{F}_{n\text{BC}}^{\text{U}}(\delta) = \min\{F_n^{\text{U}}(\delta) - \widehat{\text{Bias}}(F_n^{\text{U}}(\delta)), 1\}$$

7. SIMULATION

In this section, we examine the finite sample accuracy of the nonparametric estimators of the treatment effect distribution bounds, investigate the coverage rates of the proposed CSs for the distribution of treatment effects at different values of δ, and the finite sample performance of the bootstrap bias-corrected estimators of the sharp bounds on the distribution of treatment effects. We focus on randomized experiments.

The data generating processes (DGP) used in this simulation study are, respectively, Example 1 and Example 2 introduced in Sections 2.3 and 4.2. The detailed simulation design will be described in Section 7.1 together with estimates F_n^{L} and F_n^{U}. Section 7.2 presents results on the coverage rates of the CSs for the distribution of treatment effects and Section 7.3 presents results on the bootstrap bias-corrected estimators.

7.1. The Simulation Design and Estimates F_n^{L} and F_n^{U}

The DGPs used in the simulations are: (i) (Case C1) $(F_1, F_0, \delta) = (C(1/4), C(3/4), (1/8))$; (ii) (Case C2) $(F_1, F_0, \delta) = (C(1/4), C(3/4), 1 - (\sqrt{6}/2))$; (iii) (Case C3) $(F_1, F_0, \delta) = (C(3/4), C(1/4), (\sqrt{6}/2) - 1)$; and (iv) (Case C4) $(F_1, F_0, \delta) = (C(3/4), C(1/4), -(1/8))$.

(Case C1) is aiming at the case where $M(\delta)>0$ with a singleton $\mathcal{Y}_{\sup,\delta}$ so that we have a normal asymptotic distribution for $\sqrt{n_1}(F_n^{\mathrm{L}}(\delta)-F^{\mathrm{L}}(\delta))$. The $m(\delta)$ for this case is greater than zero so $F^{\mathrm{U}}(\delta)=1$ and $\Pr(F_n^{\mathrm{U}}(\delta)=1)\to 1$. In this case, $\mathcal{Y}_{\inf,\delta}$ consists of two boundary points of $\mathcal{Y}_{\delta}$.

In (Case C2), $M(\delta)=0$ and $\mathcal{Y}_{\sup,\delta}$ is a singleton so we have a truncated normal asymptotic distribution for $\sqrt{n_1}(F_n^{\mathrm{L}}(\delta)-F^{\mathrm{L}}(\delta))$. The $m(\delta)$, however, is less than zero and has two interior maximizers. So the asymptotic distribution of $\sqrt{n_1}(F_n^{\mathrm{U}}(\delta)-F^{\mathrm{U}}(\delta))$ is $\sup_{y\in\mathcal{Y}_{\inf,\delta}}G(y,\delta)$.

(Case C3) is opposite to (Case C2). In (Case C3), $\sqrt{n_1}(F_n^{\mathrm{L}}(\delta)-F^{\mathrm{L}}(\delta))$ has an asymptotic distribution of $\sup_{y\in\mathcal{Y}_{\sup,\delta}}G(y,\delta)$ because $M(\delta)>0$ and $\mathcal{Y}_{\sup,\delta}$ has two interior points whereas $\sqrt{n_1}(F_n^{\mathrm{U}}(\delta)-F^{\mathrm{U}}(\delta))$ has a truncated normal asymptotic distribution since $m(\delta)=0$ and $\mathcal{Y}_{\inf,\delta}$ is a singleton.

Finally, (Case C4) is the opposite of (Case C1). In (Case C4), $M(\delta)<0$ so $\Pr(F_n^{\mathrm{L}}(\delta)=0)\to 1$ and $m(\delta)<0$ with $\mathcal{Y}_{\inf,\delta}$ being a singleton so $\sqrt{n_1}(F_n^{\mathrm{U}}(\delta)-F^{\mathrm{U}}(\delta))$ has a normal asymptotic distribution. Table 1 summarizes these DGPs.

We also generated DGPs for two normal marginal distributions. Table 2 summarizes the cases considered in the simulation. In all of these cases, $\sqrt{n_1}(F_n^{\mathrm{L}}(\delta)-F^{\mathrm{L}}(\delta))$ and $\sqrt{n_1}(F_n^{\mathrm{U}}(\delta)-F^{\mathrm{U}}(\delta))$ have asymptotic normal distributions but we include these DGPs in order to see the finite sample

Table 1. DGPs (Case C1)–(Case C4).

	(Case C1)	(Case C2)
(F_1, F_0, δ)	$\left(C(1/4), C(3/4), \frac{1}{8}\right)$	$\left(C(1/4), C(3/4), 1-\frac{\sqrt{6}}{2}\right)$
F^{L}	$\mathrm{M}(\delta)=F^{\mathrm{L}}(\delta)\approx 0.49$	$M(\delta)=F^{\mathrm{L}}(\delta)=0$
$\mathcal{Y}_{\sup,\delta}$	Singleton, interior point	Singleton, interior point
$\mathrm{W}_{L,}\delta$	$N(0,\sigma_{\mathrm{L}}^2)$	$\max\{N(0,\sigma_{\mathrm{L}}^2),0\}$
F^{U}	$m(\delta)\approx 0.06$, $F^{\mathrm{U}}(\delta)=1$	$1-m(\delta)=F^{\mathrm{U}}(\delta)\approx 0.9$
$\mathcal{Y}_{\inf,\delta}$	Two boundary points	Two interior points
$W_{U,\delta}$	$\Pr(F_n^{\mathrm{U}}(\delta)=1)\to 1$	$\inf_{y\in\mathcal{Y}_{\inf,\delta}}G(y,\delta)$
	(Case C3)	(Case C4)
(F_1, F_0, δ)	$\left(C(3/4), C(1/4), \frac{\sqrt{6}}{2}-1\right)$	$\left(C(3/4), C(1/4), -\frac{1}{8}\right)$
F^{L}	$M(\delta)=F^{\mathrm{L}}(\delta)\approx 0.1$	$M(\delta)\approx -0.06$, $F^{\mathrm{L}}(\delta)=0$
$\mathcal{Y}_{\sup,\delta}$	Two interior points	Two boundary points
$W_{L,\delta}$	$\sup_{y\in\mathcal{Y}_{\sup,\delta}}G(y,\delta)$	$\Pr(F_n^{\mathrm{L}}(\delta)=0)\to 1$
F^{U}	$1-m\ (\delta)=F^{\mathrm{U}}(\delta)=1$	$1-m(\delta)=F^{\mathrm{U}}(\delta)\approx 0.51$
$\mathcal{Y}_{\inf,\delta}$	Singleton, interior point	Singleton, interior point
$W_{U,\delta}$	$\min\{N(0,\sigma_{\mathrm{U}}^2),0\}$	$N(0,\sigma_{\mathrm{U}}^2)$

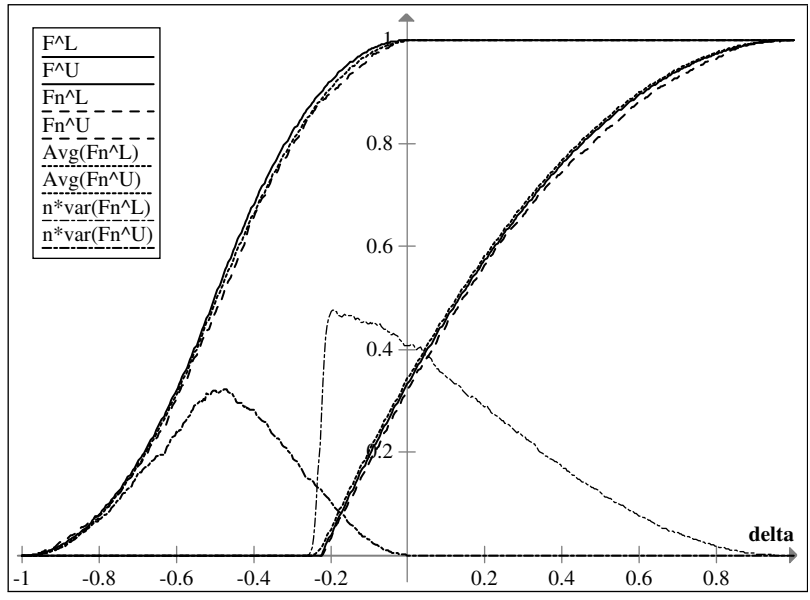
Table 2. DGPs (Case N1)–(Case N6).

	(Case N1)	(Case N2)	(Case N3)
(F_1, F_0, δ)	$(N(2,2), N(1,1), 1.3)$	$(N(2,2), N(1,1), 2.6)$	$(N(2,2), N(1,1), 4.5)$
F^{L}	$M(\delta) = F^{\mathrm{L}}(\delta) \approx 0.15$	$M(\delta) = F^{\mathrm{L}}(\delta) \approx 0.51$	$M(\delta) = F^{\mathrm{L}}(\delta) \approx 0.86$
$\mathcal{Y}_{\sup,\delta}$	Singleton	Singleton	Singleton
$W_{L,\delta}$	$N(0, \sigma_{\mathrm{L}}^2)$	$N(0, \sigma_{\mathrm{L}}^2)$	$N(0, \sigma_{\mathrm{L}}^2)$
F^{U}	$1-m(\delta) = F^{\mathrm{U}}(\delta) \approx 0.97$	$1-m(\delta) = F^{\mathrm{U}}(\delta) \approx 1$	$1-m(\delta) = F^{\mathrm{U}}(\delta) \approx 1$
$\mathcal{Y}_{\inf,\delta}$	Singleton	Singleton	Singleton
$W_{U,\delta}$	$N(0, \sigma_{\mathrm{U}}^2)$	$N(0, \sigma_{\mathrm{U}}^2)$	$N(0, \sigma_{\mathrm{U}}^2)$
	(Case N4)	(Case N5)	(Case N6)
(F_1, F_0, δ)	$(N(2,2), N(1,1), -2.4)$	$(N(2,2), N(1,1), -0.6)$	$(N(2,2), N(1,1), 0.7)$
F^{L}	$M(\delta) = F^{\mathrm{L}}(\delta) \approx 0$	$M(\delta) = F^{\mathrm{L}}(\delta) \approx 0$	$M(\delta) = F^{\mathrm{L}}(\delta) \approx 0.04$
$\mathcal{Y}_{\sup,\delta}$	Singleton	Singleton	Singleton
$W_{L,\delta}$	$N(0, \sigma_{\mathrm{L}}^2)$	$N(0, \sigma_{\mathrm{L}}^2)$	$N(0, \sigma_{\mathrm{L}}^2)$
F^{U}	$1-m(\delta) = F^{\mathrm{U}}(\delta) \approx 0.16$	$1-m(\delta) = F^{\mathrm{U}}(\delta) \approx 0.49$	$1-m(\delta) = F^{\mathrm{U}}(\delta) \approx 0.85$
$\mathcal{Y}_{\inf,\delta}$	Singleton	Singleton	Singleton
$W_{U,\delta}$	$N(0, \sigma_{\mathrm{U}}^2)$	$N(0, \sigma_{\mathrm{U}}^2)$	$N(0, \sigma_{\mathrm{U}}^2)$

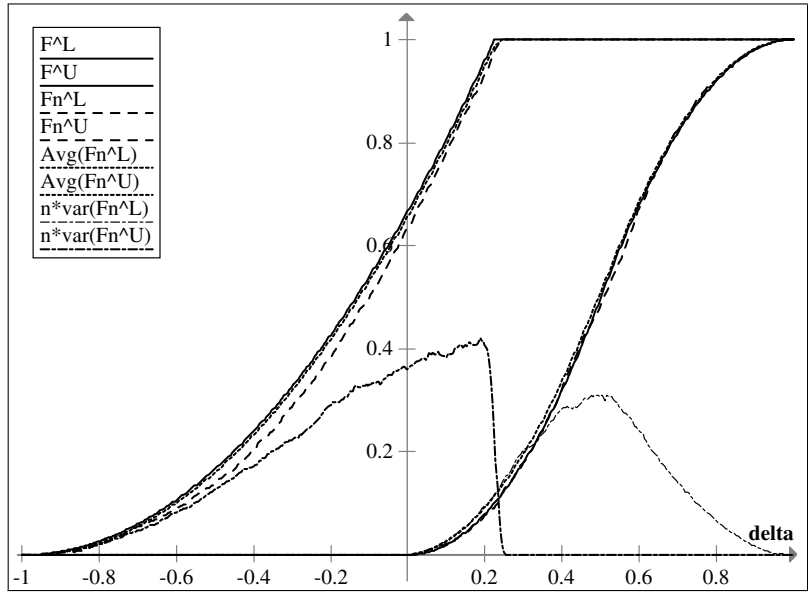
performance of our bootstrap procedures for different values of $F^{\mathrm{L}}(\delta)$ and $F^{\mathrm{U}}(\delta)$. From (Case N1) to (Case N6), $F^{\mathrm{L}}(\delta)$ ranges from being very close to zero to about 0.86 and $F^{\mathrm{U}}(\delta)$ from 0.16 to almost 1.

We now present F_n^{L} and F_n^{U} for the normal marginals (DGPs (Case N1)–(Case N6)) and C (α) class of marginals (DGPs (Case C1)–(Case C4)). For each set of marginal distributions, random samples of sizes $n_1 = n_0 = n = 1{,}000$ are drawn and F_n^{L} and F_n^{U} are computed. This is repeated for 500 times. Below we present four graphs. In each graph, we plotted F_n^{L} and F_n^{U} randomly chosen from the 500 estimates, the averages of 500 F_n^{L}s and F_n^{U}s, and the simulation variances of F_n^{L} and F_n^{U} multiplied by n. Each graph consists of eight curves. The true distribution bounds F^{L} and F^{U} are denoted as F^L and F^U, respectively. Their estimates (F_n^{L} and F_n^{U}) are Fn^L and Fn^U. The lines denoted by avg(Fn^L) and avg(Fn^U) show the averages of 500 F_n^{L}s and F_n^{U}s. The simulation variances of F_n^{L} and F_n^{U} multiplied by n are denoted as n*var(Fn^L) and n*var(Fn^U).

Fig. 7(a) and (b) correspond to (Case C1)–(Case C4), while Fig. 7(c) corresponds to (Case N1)–(Case N6). In all cases, we observe that Fn^L and avg(Fn^L) are very close to F^L at all points of its support (the same holds true for F^U). In fact, these curves are barely distinguishable from each other. The largest variance in all cases for all values of δ is less than 0.0005.



(a)



(b)

Fig. 7. (a) Estimates of the Distribution Bounds: ($C(1/4)$, $C(3/4)$); (b) Estimates of the Distribution Bounds: ($C(3/4)$, $C(1/4)$); and (c) Estimates of the Distribution Bounds: ($N(2,2)$, $N(1,1)$).

(c)

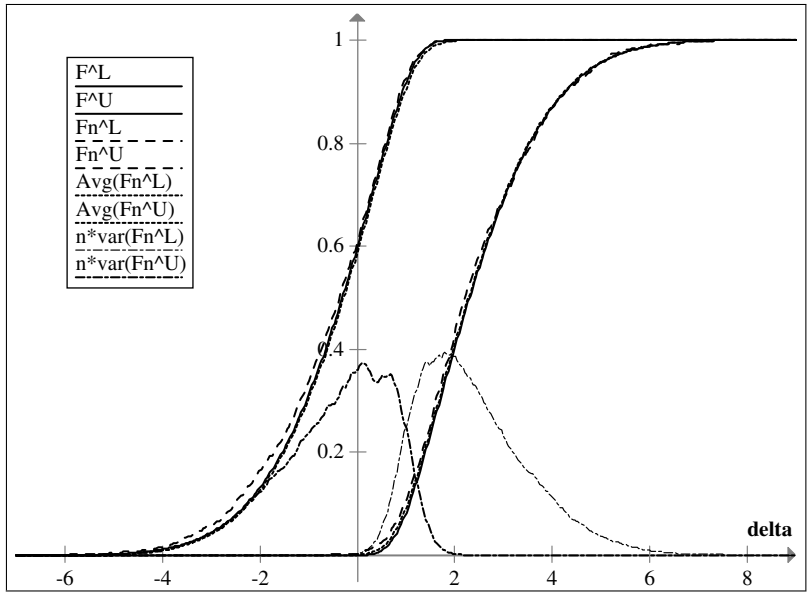


Fig. 7. (*Continued*)

7.2. Simulation Results for Coverage Rates

In this and the next subsections, we present simulation results for the bootstrap CSs and the bootstrap bias-corrected estimators. For each DGP, we generated random samples of sizes $n_1 = n_0 = 300$ and 1,000, respectively. The number of replications we used is 2,500 and the number of bootstrap repetitions is $B = 1{,}999$ as suggested in Davidson and Mackinnon (2004, pp. 163–165). The shrinkage parameters are: $b_n = n_1^{-(1/3)}$ and $b'_n = 0.3n_1^{-(0.95/2)}$, that is, $c = 1.0$, $a = 1/3$, $c' = 0.3$, and $a' = 0.05$ in the expressions in Section 5.1. We used the second procedure based on $W^*_{\mathrm{L},\delta}$ and $W^*_{\mathrm{U},\delta}$. We set $\alpha = 0.05$ throughout the simulations.

Table 3 presents the minimum values of coverage rates of the CSs defined in Theorem 3 ($F_\Delta(\delta)$ columns) and the average values of $\underline{\hat{p}}$ with DGPs (Case C1)–(Case C4).

The CSs for DGPs (Case C2) and (Case C4) perform very well. As n grows, the coverage rates for DGPs (Case C2) and (Case C3) become closer to the nominal level $1-\alpha = 0.95$. Considering that (Case C2) and (Case C3) are cases where the estimator of one of the two bounds follows a normal

Table 3. Coverage Rates and avg($\hat{\underline{p}}$) for (Case C1)–(Case C4).

	(Case C1)		(Case C2)		(Case C3)		(Case C4)	
	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)
$n = 300$	0.9320	0.9220	0.9360	0.9762	0.9356	0.9766	0.9312	0.9203
$n = 1{,}000$	0.9376	0.9228	0.9488	0.9780	0.9540	0.9786	0.9384	0.9213

Table 4. Coverage Rates and avg($\hat{\underline{p}}$) for (Case N1)–(Case N6).

	(Case N1)		(Case N2)		(Case N3)	
	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)
$n = 300$	0.9304	0.9628	0.9252	0.929	0.9332	0.9007
$n = 1{,}000$	0.9536	0.9626	0.9508	0.9479	0.9492	0.9050
	(Case N4)		(Case N5)		(Case N6)	
	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)	$F_\Delta(\delta)$	avg($\hat{\underline{p}}$)
$n = 300$	0.950	0.9182	0.9176	0.9717	0.9444	0.9629
$n = 1{,}000$	0.9492	0.9293	0.950	0.9869	0.9492	0.9643

distribution asymptotically but the estimator of the other bound violates (A3) and (A4), our bootstrap procedure seems to perform very well. The minimum coverage rates for (Case C1) and (Case C4) in which the estimator of one of the two bounds degenerates asymptotically are about 0.93–0.94. They improve slowly as the sample size becomes larger. When $n = 1{,}000$, the coverage rates are still less than 0.94 but a little better than the coverage rates with $n = 300$. The average $\hat{\underline{p}}$ differs from DGP to DGP. (Case C1) and (Case C4), where $F_n^{\mathrm{L}}(\delta)$ or $F_n^{\mathrm{U}}(\delta)$ has a degenerate asymptotic distribution, have $\hat{\underline{p}}$ as low as about 0.92. (Case C2) and (Case C3) have $\hat{\underline{p}}$ about 0.98. In both cases, $\hat{\underline{p}}$ is far greater than $\alpha = 0.05$.

The coverage rates for DGPs (Case N1)–(Case N6) are in Table 4. Recall that in all of these cases, $\sqrt{n_1}(F_n^{\mathrm{L}}(\delta) - F^{\mathrm{L}}(\delta))$ and $\sqrt{n_1}(F_n^{\mathrm{U}}(\delta) - F^{\mathrm{U}}(\delta))$ have asymptotic normal distributions.

The coverage rates for $F_\Delta(\delta)$ increased from about 0.92–0.93 when $n = 300$ to almost 0.95 when $n = 1{,}000$. For (Case N4) and (Case N6), the coverage rates for $n = 300$ are already very good. As in DGPs (Case C1)–(Case C4), the average $\hat{\underline{p}}$ differs from DGP to DGP. Nonetheless, $\hat{\underline{p}}$ is greater than 0.05 for all cases.

7.3. Simulation Results for Bias-Corrected Estimators

In each replication, we computed the bootstrap biases and mean squared errors of F_n^{L} and F_n^{U} as well as $\widehat{F}_{n\mathrm{BC}}^{\mathrm{L}}$ and $\widehat{F}_{n\mathrm{BC}}^{\mathrm{U}}$, where we used the bootstrap bias-correction with the second bootstrap procedure discussed in Section 5.1. "Bias" and "$\sqrt{\mathrm{MSE}}$" in Table 5 represent the average bias and the square roots of the mean squared errors (MSE).

The direction of the bias without correction is as expected. The bias estimates are positive for F_n^{L} and negative for F_n^{U} for all DGPs except for the cases that $\sqrt{n_1}(F_n^{\mathrm{L}}(\delta) - F^{\mathrm{L}}(\delta))$ and $\sqrt{n_1}(F_n^{\mathrm{U}}(\delta) - F^{\mathrm{U}}(\delta))$ degenerate asymptotically (Case C1 for F_n^{L} and Case C4 for F_n^{U}). The bias-correction took

Table 5. Bias and MSE Reduction for (Case C1)–(Case C4).

		(Case C1)		(Case C2)	
		$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$	$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$
$n = 300$	Bias	0.0190	0.0003	0.0305	0.0142
	$\sqrt{\mathrm{MSE}}$	0.0382	0.0352	0.0429	0.0263
$n = 1{,}000$	Bias	0.0095	−0.0009	0.0152	0.0066
	$\sqrt{\mathrm{MSE}}$	0.0211	0.0197	0.0220	0.0130
		$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$	$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$
$n = 300$	Bias	0	0	−0.0292	−0.0064
	$\sqrt{\mathrm{MSE}}$	0	0	0.0361	0.0253
$n = 1{,}000$	Bias	0	0	−0.0150	−0.0031
	$\sqrt{\mathrm{MSE}}$	0	0	0.0187	0.0134
		(Case C3)		(Case C4)	
		$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$	$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$
$n = 300$	Bias	0.0292	0.0064	0	0
	$\sqrt{\mathrm{MSE}}$	0.0348	0.0247	0	0
$n = 1{,}000$	Bias	0.0144	0.0024	0	0
	$\sqrt{\mathrm{MSE}}$	0.0182	0.0131	0	0
		$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$	$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$
$n = 300$	Bias	−0.0306	−0.0141	−0.0192	−0.0004
	$\sqrt{\mathrm{MSE}}$	0.0430	0.0265	0.0382	0.0349
$n = 1{,}000$	Bias	−0.0159	−0.0070	−0.0099	0.0004
	$\sqrt{\mathrm{MSE}}$	0.0228	0.0136	0.0211	0.0194

effect with $n = 300$ quite dramatically already. In (Case C1) for F_n^{L} and (Case C4) for F_n^{U}, where the asymptotic distributions of those estimators are normal, the magnitude of the bias reduces to roughly about 1/50–1/60 of the bias of F_n^{L} or F_n^{U}. For other DGPs, the magnitude of the bias-reduction is not as great but still the biases reduced by roughly about 1/1.5–1/4.5 of the bias of F_n^{L} or F_n^{U}. The relative magnitude of bias-reduction is similar in $n = 1{,}000$ for (Case C2) or (Case C3). It is roughly about $1/2 \sim 1/5$ of the bias of F_n^{L} or F_n^{U}. The bias estimates of $\widehat{F}_{n\mathrm{BC}}^{\mathrm{L}}$ for (Case C1) and $\widehat{F}_{n\mathrm{BC}}^{\mathrm{U}}$ (Case C4) changed sign when $n = 1{,}000$. The bootstrap bias-corrected estimators work quite well and we can see huge reduction in bias and changes of signs in (Case C1) for F_n^{L} and (Case C4) for F_n^{U} (where the normal asymptotics holds). We will see the sign change with the DGPs (Case N1)–(Case N6) as well. The bootstrap bias-corrected estimators also have smaller MSEs than F_n^{L} and F_n^{U} as shown in the table. The $\sqrt{\mathrm{MSE}}$ of $\widehat{F}_{n\mathrm{BC}}^{\mathrm{L}}$ and $\widehat{F}_{n\mathrm{BC}}^{\mathrm{U}}$ are roughly 2/3 of the $\sqrt{\mathrm{MSE}}$ of F_n^{L} and F_n^{U} for (Case C2) and (Case C3) but the reduction in $\sqrt{\mathrm{MSE}}$ is not as great in (Case C1) for F_n^{L} and (Case C4) for F_n^{U} as in other DGPs.

Table 6 show that results for (Case N1)–(Case N6) are similar. The sign change happened in all DGPs except for those in which $F^{\mathrm{L}}(\delta) \approx 0$ or $F^{\mathrm{U}}(\delta) \approx 1$. The relative magnitude of the bias in $\widehat{F}_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$ or $\widehat{F}_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$ to the bias in $F_n^{\mathrm{L}}(\delta)$ or $F_n^{\mathrm{U}}(\delta)$ ranges from 1/2 to 1/13. The reduction in $\sqrt{\mathrm{MSE}}$ is not sizable.

8. CONCLUSION

In this paper, we have provided a complete study on partial identification of and inference for the distribution of treatment effects for randomized experiments. For randomized experiments with a known value of a dependence measure between the potential outcomes such as Kendall's τ, we established tighter bounds on the distribution of treatment effects. Estimation of these bounds and inference for the distribution of treatment effects in this case can be done by following Sections 4 and 5 in this paper. When observable covariates are available such that the selection-on-observables assumption holds, Fan (2008) developed estimation and inference procedures for the distribution of treatment effects and Fan and Zhu (2009) established estimation and inference procedures for a general class of functionals of the joint distribution of potential outcomes

Table 6. Bias and MSE Reduction for (Case N1)–(Case N6).

		(Case N1)		(Case N2)		(Case N3)	
		$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$	$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$	$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$
$n = 300$	Bias	0.0233	0.0023	0.0187	0.0011	0.0108	−0.0023
	$\sqrt{\mathrm{MSE}}$	0.0397	0.0354	0.0376	0.0343	0.0226	0.0214
$n = 1{,}000$	Bias	0.0106	−0.0008	0.0088	−0.0011	0.0049	−0.0024
	$\sqrt{\mathrm{MSE}}$	0.0207	0.0187	0.0205	0.0193	0.0121	0.0118
		$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$	$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$	$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$
$n = 300$	Bias	−0.0182	0.0017	−0.0011	−0.0001	0	0
	$\sqrt{\mathrm{MSE}}$	0.0276	0.0207	0.0024	0.0005	0.0001	0
$n = 1{,}000$	Bias	−0.0087	0.0024	−0.0005	0.0	0.0	0.0
	$\sqrt{\mathrm{MSE}}$	0.0144	0.0120	0.0010	0.0001	0.0	0.0
		(Case N4)		(Case N5)		(Case N6)	
		$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$	$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$	$F_n^{\mathrm{L}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{L}}(\delta)$
$n = 300$	Bias	0.0	0.0	0.0013	0.0001	0.0192	−0.0009
	$\sqrt{\mathrm{MSE}}$	0.0002	0.0	0.0026	0.0005	0.0286	0.0210
$n = 1{,}000$	Bias	0.0	0.0	0.0005	0.0	0.0089	−0.0021
	$\sqrt{\mathrm{MSE}}$	0.0001	0.0	0.0005	0.0	0.0145	0.0118
		$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$	$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$	$F_n^{\mathrm{U}}(\delta)$	$F_{n\mathrm{BC}}^{\mathrm{U}}(\delta)$
$n = 300$	Bias	−0.0111	0.0024	−0.0195	−0.0017	−0.0229	−0.0019
	$\sqrt{\mathrm{MSE}}$	0.0228	0.0213	0.0381	0.0344	0.0385	0.0344
$n = 1{,}000$	Bias	−0.0055	0.0019	−0.0085	0.0014	−0.0104	0.0009
	$\sqrt{\mathrm{MSE}}$	0.0127	0.012	0.02	0.0187	0.0209	0.0189

including many commonly used inequality measures of the distribution of treatment effects.

This paper has focused on binary treatments. The results can be easily extended to multivalued treatments. For example, consider a randomized experiment on a treatment taking values in $\{0, 1, \ldots, T\}$. Define the treatment effect between t and t' as $\Delta_{t',t} = Y_{t'} - Y_t$ for any $t, t' \in \{0, 1, \ldots, T\}$ and $t \neq t'$. Then by substituting Y_1 with $T_{t'}$ and Y_0 with Y_t, the results in this paper apply to $F_{\Delta t',t}$. The results in this paper can also be extended to continuous treatments, provided that the marginal distribution of the potential outcome corresponding to a given level of treatment intensity is identified.

NOTES

1. In the rest of this paper, we refer to ideal randomized experiments (data) as randomized experiments (data).
2. A copula is a bivariate distribution with uniform marginal distributions on [0,1].
3. Frank et al. (1987) provided expressions for the sharp bounds on the distribution of a sum of two normal random variables. We believe there are typos in their expressions, as a direct application of their expressions to our case would lead to different expressions from ours. They are:

$$F^{\mathrm{L}}(\delta) = \Phi\left(\frac{-\sigma_1 s - \sigma_0 t}{\sigma_0^2 - \sigma_1^2}\right) + \Phi\left(\frac{\sigma_0 s - \sigma_1 t}{\sigma_0^2 - \sigma_1^2}\right) - 1$$

$$F^{\mathrm{U}}(\delta) = \Phi\left(\frac{-\sigma_1 s + \sigma_0 t}{\sigma_0^2 - \sigma_1^2}\right) + \Phi\left(\frac{\sigma_0 s + \sigma_1 t}{\sigma_0^2 - \sigma_1^2}\right)$$

4. In practice, the supports of F_1 and F_0 may be unknown, but can be estimated by using the corresponding univariate order statistics in the usual way. This would not affect the results to follow. For notational compactness, we assume that they are known.

ACKNOWLEDGMENTS

We thank the editors of the *Advances in Econometrics*, Vol. 24, T. Fomby, R. Carter Hill, Q. Li, and J. S. Racine, participants of the 7th annual Advances in Econometrics Conference, and two referees for helpful comments that improved both the exposition and content of this paper.

REFERENCES

Aakvik, A., Heckman, J., & Vytlacil, E. (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary among observationally identical persons: An application to Norwegian vocational rehabilitation programs. *Journal of Econometrics*, *125*, 15–51.

Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental variables estimation of quantile treatment effects. *Econometrica*, *70*, 91–117.

Abbring, J. H., & Heckman, J. (2007). Econometric evaluation of social programs, Part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. *The Handbook of Econometrics*, *6B*, 5145–5301.

Alsina, C. (1981). Some functional equations in the space of uniform distribution functions. *Equationes Mathematicae*, *22*, 153–164.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, *68*, 399–405.

Andrews, D. W. K., & Guggenberger, P. (2005a). *The limit of finite-sample size and a problem with subsampling*. Unpublished Manuscript, Cowles Foundation, Yale University, New Haven, CT.

Andrews, D. W. K., & Guggenberger, P. (2005b). *Hybrid and size-corrected subsampling methods*. Unpublished Manuscript, Cowles Foundation, Yale University, New Haven, CT.

Andrews, D. W. K., & Guggenberger, P. (2005c). *Applications of subsampling, hybrid, and size-correction methods*. Cowles Foundation Discussion Paper No. 1608, Yale University, New Haven, CT.

Andrews, D. W. K., & Guggenberger, P. (2007). *Validity of subsampling and 'plug-in asymptotic' inference for parameters defined by moment inequalities*. Unpublished Manuscript, Cowles Foundation, Yale University, New Haven, CT.

Andrews, D. W. K., & Soares, G. (2007). *Inference for parameters defined by moment inequalities using generalized moment selection*. Cowles Foundation Working Paper no. 1631, Yale University, New Haven, CT.

Beresteanu, A., & Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, *76*, 763–814.

Biddle, J., Boden, L., & Reville, R. (2003). *A method for estimating the full distribution of a treatment effect, with application to the impact of workfare injury on subsequent earnings*. Mimeo, Michigan State University.

Bitler, M., Gelbach, J., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, *96*, 988–1012.

Black, D. A., Smith, J. A., Berger, M. C., & Noel, B. J. (2003). Is the threat of reemployment services more effective than the services themselves? Experimental evidence from the UI system. *American Economic Review*, *93*(3), 1313–1327.

Bugni, F.A. (2007). Bootstrap inference in partially identified models. Mimeo, Northwestern University.

Cambanis, S., Simons, G., & Stout, W. (1976). Inequalities for $\varepsilon k(X, Y)$ when the marginals are fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *36*, 285–294.

Canay, I. A. (2007). *EL inference for partially identified models: Large deviations optimality and bootstrap validity*. Manuscript, University of Wisconsin.

Carneiro, P., Hansen, K. T., & Heckman, J. (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review*, *44*(2), 361–422.

Chernozhukov, V., & Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, *73*, 245–261.

Chernozhukov, V., Hong, H., & Tamer, E. (2007). Parameter set inference in a class of econometric models. *Econometrica*, *75*, 1243–1284.

Davidson, R., & Mackinnon, J. G. (2004). *Econometric theory and method*. New York, NY: Oxford University Press.

Dehejia, R. (1997). *A decision-theoretic approach to program evaluation*. Ph.D. Dissertation, Department of Economics, Harvard University.

Dehejia, R., & Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Denuit, M., Genest, C., & Marceau, E. (1999). Stochastic bounds on sums of dependent risks. *Insurance: Mathematics and Economics*, *25*, 85–104.

Djebbari, H., & Smith, J. A. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, *145*, 64–80.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Annals of Statistics*, *2*, 267–277.

Embrechts, P., Hoeing, A., & Juri, A. (2003). Using copulae to bound the value-at-risk for functions of dependent risks. *Finance & Stochastics*, *7*(2), 145–167.

Fan, Y. (2008). *Confidence sets for distributions of treatment effects with covariates*. Vanderbilt University, Nashville, TN (work in progress).

Fan, Y., & Park, S. (2007a). *Confidence sets for the quantile of treatment effects*. Manuscript, Vanderbilt University.

Fan, Y., & Park, S. (2007b). *Confidence sets for some partially identified parameters*. Manuscript, Vanderbilt University.

Fan, Y., & Park, S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, *26* (forthcoming).

Fan, Y., & Wu, J. (2007). *Sharp bounds on the distribution of the treatment effect in switching regimes models*. Manuscript, Vanderbilt University.

Fan, Y., & Zhu, D. (2009). *Partial identification and confidence sets for parameters of the joint distribution of the potential outcomes*. Working Paper, Vanderbilt University, Nashville, TN.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, *75*, 259–276.

Firpo, S., & Ridder, G. (2008). *Bounds on functionals of the distribution of treatment effects*. Institute of Economic Policy Research (IEPR) Working Paper no. 08-09. University of Southern California, CA.

Frank, M. J., Nelsen, R. B., & Schweizer, B. (1987). Best-possible bounds on the distribution of a sum – a problem of Kolmogorov. *Probability Theory and Related Fields*, *74*, 199–211.

Galichon, A., & Henry, M. (2009). A test of non-identifying restrictions and confidence regions for partially identified parameters. *Journal of Econometrics*, *152*, 186–196.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*, 315–331.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*, 1017–1098.

Heckman, J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In: J. Heckman & B. Singer (Eds), *Longitudinal analysis of labor market data*. New York: Cambridge University Press.

Heckman, J., & Smith, J. (1993). Assessing the case for randomized evaluation of social programs. In: K. Jensen & P. K. Madsen (Eds), *Measuring labour market measures: Evaluating the effects of active labour market policies* (pp. 35–96). Copenhagen, Denmark: Danish Ministry of Labor.

Heckman, J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for Heterogeneity in Programme Impacts. *Review of Economic Studies*, *64*, 487–535.

Heckman, J., & Vytlacil, E. (2007a). Econometric evaluation of social programs. Part I: Causal models, structural models and econometric policy evaluation. *The Handbook of Econometrics*, *6B*, 4779–4874.

Heckman, J., & Vytlacil, E. (2007b). Econometric evaluation of social programs. Part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *The Handbook of Econometrics*, *6B*, 4875–5143.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161–1189.

Horowitz, J. L., & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, *95*, 77–84.

Imbens, G. W., & Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, *72*, 1845–1857.

Imbens, G. W., & Newey, W. (2009). *Identification and estimation of triangular simultaneous equations models without additivity*. *Econometrica* (forthcoming).

Imbens, G. W., & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, *64*, 555–574.

Joe, H. (1997). *Multivariate models and dependence concepts*. London, UK: Chapman & Hall/CRC.

Lalonde, R. (1995). The promise of public sector-sponsored training programs. *Journal of Economic Perspectives*, *9*, 149–168.

Lechner, M. (1999). Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business and Economic Statistics*, *17*, 74–90.

Lee, L. F. (2002). *Correlation bounds for sample selection models with mixed continuous, discrete and count data variables*. Manuscript, The Ohio State University, Athens, OH.

Lee, M. J. (2005). *Micro-econometrics for policy, program, and treatment effects*. New York, NY: Oxford University Press.

Lehmann, E. L. (1974). *Nonparametrics: Statistical methods based on ranks*. San Francisco, CA: Holden-Day Inc.

Makarov, G. D. (1981). Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability and its Applications*, *26*, 803–806.

Manski, C. F. (1990). Non-parametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, *80*, 319–323.

Manski, C. F. (1997a). Monotone treatment effect. *Econometrica*, *65*, 1311–1334.

Manski, C. F. (1997b). The mixing problem in programme evaluation. *Review of Economic Studies*, *64*, 537–553.

Manski, C. F. (2003). *Partial identification of probability distributions*. New York: Springer-Verlag.

Moon, R., & Schorfheide, F. (2007). *A Bayesian look at partially-identified models*. Manuscript, University of Pennsylvania, Philadelphia, PA.

Nelsen, R. B. (1999). *An introduction to copulas*. New York: Springer.

Nelsen, R. B., Quesada-Molina, J. J., Rodriguez-Lallena, J. A., & Ubeda-Flores, M. (2001). Bounds on bivariate distribution functions with given margins and measures of association. *Communications in Statistics: Theory and Methods*, *30*, 1155–1162.

Nelsen, R. B., Quesada-Molina, J. J., Rodriguez-Lallena, J. A., & Ubeda-Flores, M. (2004). Best-possible bounds on sets of bivariate distribution functions. *Journal of Multivariate Analysis*, *90*, 348–358.

Nelsen, R. B., & Ubeda-Flores, M. (2004). A comparison of bounds on sets of joint distribution functions derived from various measures of association. *Communications in Statistics: Theory and Methods*, *33*, 2299–2305.

Romano, J., & Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, *138*, 2786–2807.

Rosen, A. (2008). Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics*, *146*, 107–117.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*, *45*, 212–218.

Schweizer, B., & Sklar, A. (1983). *Probabilistic metric spaces*. New York: North-Holland.

Sklar, A. (1959). Functions de réartition à n dimensions et leures marges. *Publications de l'Institut de Statistique de L'Université de Paris*, *8*, 229–231.

Soares, G. (2006). *Inference for partially identified models with inequality moment constraints*. Working Paper, Yale University, New Haven, CT.

Stoye, J. (2008). *Partial identification of spread parameters*. Working Paper, New York University, New York, NY.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica* (forthcoming).

Tchen, A. H. (1980). Inequalities for distributions with given marginals. *Annals of Probability*, *8*, 814–827.

Tesfatsion, L. (1976). Stochastic dominance and the maximization of expected utility. *Review of Economic Studies*, *43*, 301–315.

Williamson, R. C., & Downs, T. (1990). Probabilistic arithmetic I: Numerical Methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, *4*, 89–158.

APPENDIX A. PROOF OF EQ. (23)

Obviously, one can take $1 - \underline{p} = \lim_{n_1 \to \infty} \inf_{\theta_0 \in [\theta_L, \theta_U]} \Pr(\theta_0 \in \{\theta : T_n(\theta) \leq 0)\}$. Now,

$$\lim_{n_1 \to \infty} \inf_{\theta_0 \in [\theta_L, \theta_U]} \Pr(\theta_0 \in \{\theta : T_n(\theta) \leq 0) = \inf \Pr[(W_{L,\delta} - h^L(\theta_0))_+^2 + (W_{U,\delta} + h^U(\theta_0))_-^2 = 0]$$

We need to show that

$$\inf \Pr[(W_{L,\delta} - h^L(\theta_0))_+^2 + (W_{U,\delta} + h^U(\theta_0))_-^2 = 0] = \Pr\left[\sup_{y \in \mathcal{Y}_{\sup,\delta}} G(y,\delta) \leq 0, \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta) \geq 0\right]$$

First, we consider the case with $W_{\mathrm{L},\delta} - h^{\mathrm{L}}(\theta_0) \le 0$. We have:

$$
\begin{aligned}
& W_{\mathrm{L},\delta} - h^{\mathrm{L}}(\theta_0) \le 0 \\
& \quad \Leftrightarrow \max\left\{ \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta), -h_{\mathrm{L}}(\delta) \right\} \le -\min\{h_{\mathrm{L}}(\delta), 0\} + h^{\mathrm{L}}(\theta_0) \\
& \quad \Leftrightarrow \max\left\{ \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta), -h_{\mathrm{L}}(\delta) \right\} \le -h_{\mathrm{L}}(\delta) + \lim_{n_1 \to \infty} \sqrt{n_1} F_{\Delta}(\delta) \\
& \quad \Leftrightarrow \max\left\{ \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta), -\lim_{n_1 \to \infty} \sqrt{n_1} M(\delta) \right\} \le \lim_{n_1 \to \infty} \sqrt{n_1} [F_{\Delta}(\delta) - M(\delta)]
\end{aligned}
$$

since

$$
\begin{aligned}
h^{\mathrm{L}}(\theta_0) &= -\lim_{n_1 \to \infty} [\sqrt{n_1} F^{\mathrm{L}}(\delta) - \sqrt{n_1} F_{\Delta}(\delta)] \\
&= -\lim_{n_1 \to \infty} [\max\{\sqrt{n_1} M(\delta), 0\} - \sqrt{n_1} F_{\Delta}(\delta)] \\
&= -\max\left\{ \lim_{n_1 \to \infty} \sqrt{n_1} M(\delta), 0 \right\} + \lim_{n_1 \to \infty} \sqrt{n_1} F_{\Delta}(\delta)
\end{aligned}
$$

(i) If $F_{\Delta}(\delta) = F^{\mathrm{L}}(\delta) = 0 > M(\delta)$, then

$$
\begin{aligned}
& \max\left\{ \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta), -\lim_{n_1 \to \infty} \sqrt{n_1} M(\delta) \right\} \le \lim_{n_1 \to \infty} \sqrt{n_1} [F_{\Delta}(\delta) - M(\delta)] \\
& \quad \Leftrightarrow \max\left\{ \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta), \infty \right\} \le \infty \\
& \quad \Leftrightarrow \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta) < \infty
\end{aligned}
$$

which holds trivially.

(ii) If $F_{\Delta}(\delta) = F^{\mathrm{L}}(\delta) = 0 = M(\delta)$, then

$$
\begin{aligned}
& \max\left\{ \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta), -\lim_{n_1 \to \infty} \sqrt{n_1} M(\delta) \right\} \le \lim_{n_1 \to \infty} \sqrt{n_1} [F_{\Delta}(\delta) - M(\delta)] \\
& \quad \Leftrightarrow \max\left\{ \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta), 0 \right\} \le 0 \\
& \quad \Leftrightarrow \sup_{y \in \mathcal{Y}_{\mathrm{sup},\delta}} G(y,\delta) \le 0
\end{aligned}
$$

(iii) If $F_\Delta(\delta) = F^{\mathrm{L}}(\delta) = M(\delta) > 0$, then

$$\max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta), -\lim_{n_1\to\infty}\sqrt{n_1}M(\delta)\right\} \leq \lim_{n_1\to\infty}\sqrt{n_1}[F_\Delta(\delta) - M(\delta)]$$

$$\Leftrightarrow \max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta), -\infty\right\} \leq 0$$

$$\Leftrightarrow \sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta) \leq 0$$

(iv) If $F_\Delta(\delta) = F^{\mathrm{L}}(\delta) = 0 > M(\delta)$, then

$$\max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta), -\lim_{n_1\to\infty}\sqrt{n_1}M(\delta)\right\} \leq \lim_{n_1\to\infty}\sqrt{n_1}[F_\Delta(\delta) - M(\delta)]$$

$$\Leftrightarrow \max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta), \infty\right\} \leq \infty$$

$$\Leftrightarrow \sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta) < \infty$$

which holds trivially.

(v) If $F_\Delta(\delta) > F^{\mathrm{L}}(\delta) = 0 = M(\delta)$, then

$$\max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta), -\lim_{n_1\to\infty}\sqrt{n_1}M(\delta)\right\} \leq \lim_{n_1\to\infty}\sqrt{n_1}[F_\Delta(\delta) - M(\delta)]$$

$$\Leftrightarrow \max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta), 0\right\} \leq \infty$$

$$\Leftrightarrow \sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta) < \infty$$

which holds trivially.

(vi) If $F_\Delta(\delta) > F^{\mathrm{L}}(\delta) = M(\delta) > 0$, then

$$\max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta), -\lim_{n_1\to\infty}\sqrt{n_1}M(\delta)\right\} \leq \lim_{n_1\to\infty}\sqrt{n_1}[F_\Delta(\delta) - M(\delta)]$$

$$\Leftrightarrow \max\left\{\sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta), \infty\right\} \leq \infty$$

$$\Leftrightarrow \sup_{y\in\mathcal{Y}_{\sup,\delta}} G(y,\delta) < \infty$$

which holds trivially.

Summarizing (i)–(vi), we have

$$W_{\mathrm{L},\delta} - h^{\mathrm{L}}(\theta_0) \le 0 \Leftrightarrow \sup_{y \in \mathcal{Y}_{\sup,\delta}} G(y,\delta) \le 0$$

if $F_\Delta(\delta) = F^{\mathrm{L}}(\delta) = M(\delta) \ge 0$; otherwise it holds trivially.

Similarly to the $W_{\mathrm{L},\delta} - h^{\mathrm{L}}(\theta_0) \ge 0$ case, we get

$$\begin{aligned}
&W_{\mathrm{U},\delta} + h^{\mathrm{U}}(\theta_0) \ge 0 \\
&\Leftrightarrow \min\left\{ \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta), -h_{\mathrm{U}}(\delta) \right\} + \max\{h_{\mathrm{U}}(\delta), 0\} + h^{\mathrm{U}}(\theta_0) \ge 0 \\
&\Leftrightarrow \min\left\{ \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta), -h_{\mathrm{U}}(\delta) \right\} \ge -\max\{h_{\mathrm{U}}(\delta), 0\} - \lim_{n \to \infty} \sqrt{n}[F^{\mathrm{U}}(\delta) - F_\Delta(\delta)] \\
&\Leftrightarrow \min\left\{ \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta), -\lim_{n_1 \to \infty} \sqrt{n_1} m(\delta) \right\} \ge -\lim_{n_1 \to \infty} [1 + m(\delta) - F_\Delta(\delta)]
\end{aligned}$$

since

$$\begin{aligned}
h^{\mathrm{U}}(\theta_0) &= \lim_{n_1 \to \infty} [\sqrt{n_1} F^{\mathrm{U}}(\delta) - \sqrt{n_1} F_\Delta(\delta)] \\
&= \lim_{n_1 \to \infty} \sqrt{n_1} \min\{m(\delta), 0\} + \lim_{n_1 \to \infty} \sqrt{n_1}(1 - F_\Delta(\delta)) \\
&= \min\{h_{\mathrm{U}}(\delta), 0\} + \lim_{n_1 \to \infty} \sqrt{n_1}(1 - F_\Delta(\delta))
\end{aligned}$$

(i) If $1 + m(\delta) > 1 = F^{\mathrm{U}}(\delta) = F_\Delta(\delta)$, then

$$\begin{aligned}
&\min\left\{ \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta), -\lim_{n_1 \to \infty} \sqrt{n_1} m(\delta) \right\} \ge -\lim_{n_1 \to \infty} [1 + m(\delta) - F_\Delta(\delta)] \\
&\quad \Leftrightarrow \min\left\{ \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta), -\infty \right\} \ge -\infty \\
&\quad \Leftrightarrow \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta) \ge -\infty
\end{aligned}$$

which holds trivially.

(ii) If $1 + m(\delta) = 1 = F^{\mathrm{U}}(\delta) = F_\Delta(\delta)$, then

$$\begin{aligned}
&\min\left\{ \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta), -\lim_{n_1 \to \infty} \sqrt{n_1} m(\delta) \right\} \ge -\lim_{n_1 \to \infty} [1 + m(\delta) - F_\Delta(\delta)] \\
&\quad \Leftrightarrow \min\left\{ \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta), 0 \right\} \ge 0 \\
&\quad \Leftrightarrow \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta) \ge 0
\end{aligned}$$

(iii) If $1 > 1 + m(\delta) = F^{\mathrm{U}}(\delta) = F_{\Delta}(\delta)$, then

$$\begin{aligned}
&\min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta), -\lim_{n_1\to\infty}\sqrt{n_1}m(\delta)\right\} \geq -\lim_{n_1\to\infty}[1+m(\delta)-F_{\Delta}(\delta)] \\
&\quad\Leftrightarrow \min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta), \infty\right\} \geq 0 \\
&\quad\Leftrightarrow \inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta) \geq 0
\end{aligned}$$

(iv) If $1 + m(\delta) > 1 = F^{\mathrm{U}}(\delta) > F_{\Delta}(\delta)$, then

$$\begin{aligned}
&\min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta), -\lim_{n_1\to\infty}\sqrt{n_1}m(\delta)\right\} \geq -\lim_{n_1\to\infty}[1+m(\delta)-F_{\Delta}(\delta)] \\
&\quad\Leftrightarrow \min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta), -\infty\right\} \geq -\infty \\
&\quad\Leftrightarrow \inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta) \geq -\infty
\end{aligned}$$

which holds trivially.

(v) If $1 + m(\delta) = 1 = F^{\mathrm{U}}(\delta) > F_{\Delta}(\delta)$, then

$$\begin{aligned}
&\min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta), -\lim_{n_1\to\infty}\sqrt{n_1}m(\delta)\right\} \geq -\lim_{n_1\to\infty}[1+m(\delta)-F_{\Delta}(\delta)] \\
&\quad\Leftrightarrow \min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta), 0\right\} \geq -\infty \\
&\quad\Leftrightarrow \inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta) \geq -\infty
\end{aligned}$$

which holds trivially.

(vi) If $1 > 1 + m(\delta) = F^{\mathrm{U}}(\delta) > F_{\Delta}(\delta)$, then

$$\begin{aligned}
&\min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta), -\lim_{n_1\to\infty}\sqrt{n_1}m(\delta)\right\} \geq -\lim_{n_1\to\infty}[1+m(\delta)-F_{\Delta}(\delta)] \\
&\quad\Leftrightarrow \min\left\{\inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta), \infty\right\} \geq -\infty \\
&\quad\Leftrightarrow \inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta) \geq -\infty
\end{aligned}$$

which holds trivially. Summarizing (i)–(vi), we get

$$W_{\mathrm{U},\delta} + h^{\mathrm{U}}(\theta_0) \geq 0 \Leftrightarrow \inf_{y\in\mathcal{Y}_{\inf,\delta}} G(y,\delta) \geq 0$$

if $1 \geq 1 + m(\delta) = F^{\mathrm{U}}(\delta) = F_{\Delta}(\delta)$; otherwise it holds trivially.

Finally, we obtain:

$$\begin{aligned}
&\inf\ \Pr[(W_{\mathrm{L},\delta} - h^{\mathrm{L}})(\theta_0)_+^2 + (W_{\mathrm{U},\delta} + h^{\mathrm{U}}(\theta_0))_-^2 = 0] \\
&\quad = \inf\ \Pr[W_{\mathrm{L},\delta} - h^{\mathrm{L}}(\theta_0) \le 0, W_{\mathrm{U},\delta} + h^{\mathrm{U}}(\theta_0) \ge 0] \\
&\quad = \Pr\left[\sup_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta) \le 0, \inf_{y \in \mathcal{Y}_{\inf,\delta}} G(y,\delta) \ge 0\right]
\end{aligned}$$

APPENDIX B. EXPRESSIONS FOR $y_{\sup,\delta}$, $y_{\inf,\delta}$, $m(\delta)$ AND $m(\delta)$ FOR SOME KNOWN MARGINAL DISTRIBUTIONS

Denuit et al. (1999) provided the distribution bounds for a sum of two random variables when they both follow shifted exponential distributions or both follow shifted Pareto distributions. Below, we augment their results with explicit expressions for $y_{\sup,\delta}$, $y_{\inf,\delta}$, $M(\delta)$, and $m(\delta)$ which may help us understand the asymptotic behavior of the nonparametric estimators of the distribution bounds when the true marginals are either shifted exponential or shifted Pareto.

First, we present some expressions used in Example 2.

Example 2 (continued). In Example 2, we considered the family of distributions denoted by $C(a)$ with $a \in (0,1)$. If $X \sim C(a)$, then

$$F(x) = \begin{cases} \dfrac{1}{a}x^2 & \text{if } x \in [0,a] \\ 1 - \dfrac{(x-1)^2}{(1-a)} & \text{if } x \in [a,1] \end{cases} \quad \text{and} \quad f(x) = \begin{cases} \dfrac{2}{a}x & \text{if } x \in [0,a] \\ \dfrac{2(1-x)}{(1-a)} & \text{if } x \in [a,1] \end{cases}$$

Suppose $Y_1 \sim C(\alpha_1)$ and $Y_0 \sim C(\alpha_0)$. We now provide the functional form of $F_1(y) - F_0(y-\delta)$.

1. Suppose $\delta < 0$. Then $\mathcal{Y}_\delta = [0, 1+d]$.
   (a) If $a_0 + \delta \le 0 < a_1 \le 1 + \delta$, then

$$F_1(y) - F_0(y-\delta) = \begin{cases} \dfrac{y^2}{a_1} - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } 0 \le y \le a_1 \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_1 \le y \le 1+\delta \end{cases}$$

(b) If $0 \le a_0 + \delta \le a_1 \le 1 + \delta$, then

$$F_1(y) - F_0(y-\delta) = \begin{cases} \dfrac{y^2}{a_1} - \dfrac{(y-\delta)^2}{a_0} & \text{if } 0 \le y \le a_0 + \delta \\ \dfrac{y^2}{a_1} - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_0 + \delta \le y \le a_1 \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_1 \le y \le 1 + \delta \end{cases}$$

(c) If $a_0 + \delta \le 0 \le 1 + \delta \le a_1$, then

$$F_1(y) - F_0(y-\delta) = \frac{y^2}{a_1} - \left(1 - \frac{(y-\delta-1)^2}{(1-a_0)}\right) \quad \text{if } 0 \le y \le 1 + \delta$$

(d) If $0 \le a_0 + \delta < 1 + \delta \le a_1$, then

$$F_1(y) - F_0(y-\delta) = \begin{cases} \dfrac{y^2}{a_1} - \dfrac{(y-\delta)^2}{a_0} & \text{if } 0 \le y \le a_0 + \delta \\ \dfrac{y^2}{a_1} - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_0 + \delta \le y \le 1 + \delta \end{cases}$$

(e) If $0 < a_1 \le a_0 + \delta \le 1 + \delta$, then

$$F_1(y) - F_0(y-\delta) = \begin{cases} \dfrac{y^2}{a_1} - \dfrac{(y-\delta)^2}{a_0} & \text{if } 0 \le y \le a_1 \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \dfrac{(y-\delta)^2}{a_0} & \text{if } a_1 \le y \le a_0 \le \delta \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_0 + \delta \le y \le 1 + \delta \end{cases}$$

2. Suppose $\delta \ge 0$. Then $\mathcal{Y}_\delta = [\delta, 1]$.
   (a) If $\delta < a_0 + \delta \le a_1 < 1$, then
      (i) if $a_1 \ne a_0$ and $\delta \ne 0$, then

$$F_1(y) - F_0(y-\delta) = \begin{cases} \dfrac{y^2}{a_1} - \dfrac{(y-\delta)^2}{a_0} & \text{if } \delta \le y \le a_0 + \delta \\ \dfrac{y^2}{a_1} - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_0 + \delta \le y \le a_1 \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_1 \le y \le 1 \end{cases}$$

(ii) $a_1 = a_0 = a$ and $\delta = 0$, then

$$F_1(y) - F_0(y-\delta) = 0 \quad \text{for} \quad \text{all} \quad y \in [0,1]$$

(b) If $\delta \le a_1 \le a_0 + \delta \le 1$, then

$$F_1(y) - F_0(y-\delta) = \begin{cases} \dfrac{y^2}{a_1} - \dfrac{(y-\delta)^2}{a_0} & \text{if } \delta \le y \le a_1 \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \dfrac{(y-\delta)^2}{a_0} & \text{if } a_1 + \le y \le a_0 \le \delta \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_0 + \delta \le y \le 1 \end{cases}$$

(c) If $\delta \le a_1 < 1 \le a_0 + \delta$, then

$$F_1(y) - F_0(y-\delta) = \begin{cases} \dfrac{y^2}{a_1} - \dfrac{(y-\delta)^2}{a_0} & \text{if } \delta \le y \le a_1 \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \dfrac{(y-\delta)^2}{a_0} & \text{if } a_1 \le y \le 1 \end{cases}$$

(d) If $a_1 < \delta < a_0 + \delta \le 1$, then

$$F_1(y) - F_0(y-\delta) = \begin{cases} \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \dfrac{(y-\delta)^2}{a_0} & \text{if } \delta \le y \le a_0 + \delta \\ \left(1 - \dfrac{(y-1)^2}{(1-a_1)}\right) - \left(1 - \dfrac{(y-\delta-1)^2}{(1-a_0)}\right) & \text{if } a_0 + \delta \le y \le 1 \end{cases}$$

(e) If $a_1 < \delta < 1 \le a_0 + \delta$, then

$$F_1(y) - F_0(y-\delta) = \left(1 - \frac{(y-1)^2}{(1-a_1)}\right) - \frac{(y-\delta)^2}{a_0} \quad \text{if } \delta \le y \le 1$$

(Shifted) Exponential marginals. The marginal distributions are:

$$F_1(y) = 1 - \exp\left(-\frac{y-\theta_1}{\alpha_1}\right) \text{ for } y \in [\theta_1, \infty) \quad \text{and}$$

$$F_0(y) = 1 - \exp\left(-\frac{y-\theta_0}{\alpha_0}\right) \quad \text{for} \quad y \in [\theta_0, \infty), \quad \text{where } \alpha_1, \theta_1, \alpha_0, \theta_0 > 0$$

Let $\delta_c = (\theta_1 - \theta_0) - \min\{\alpha_1, \alpha_0\}(\ln\ \alpha_1 - \ln\ \alpha_0)$.

1. Suppose $\alpha_1 < \alpha_0$.
   (a) If $\delta \le \delta_c$,

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = 0$$

$$\text{where } M(\delta) = \left(\left(\frac{\alpha_0}{\alpha_1}\right)^{\alpha_1/(\alpha_1-\alpha_0)} - \left(\frac{\alpha_0}{\alpha_1}\right)^{\alpha_0/(\alpha_1-\alpha_0)} \right) \exp\left(-\frac{\delta - (\theta_1 - \theta_0)}{\alpha_1 - \alpha_0} \right) < 0$$

$$\text{and } y_{\inf,\delta} = \frac{\alpha_0\alpha_1(\ln\ \alpha_1 - \ln\ \alpha_0) + \alpha_1\theta_0 - \alpha_0\theta_1 + \alpha_1\delta}{\alpha_1 - \alpha_0} \text{(an interior solution)}$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1 + m(\delta)$$

$$\text{where } m(\delta) = \min\left\{ \exp\left(-\frac{\max\{\theta_1 - (\delta + \theta_0), 0\}}{\alpha_0} \right) - \exp\left(-\frac{\max\{\theta_0 + \delta - \theta_1, 0\}}{\alpha_1} \right), 0 \right\}$$

$$\text{and } y_{\sup,\delta} = \max\{\theta_1, \theta_0 + \delta\} \quad \text{or} \quad \infty \text{ (boundary solution)}$$

   (b) If $\delta > \delta_c$,

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = M(\delta) > 0$$

$$\text{where } M(\delta) = 1 - \exp\left(-\frac{\delta + \theta_0 - \theta_1}{\alpha_1} \right) \text{and } y_{\inf,\delta} = \theta_0 + \delta$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1$$

$$\text{since } m(\delta) = 0 \quad \text{and} \quad y_{\sup,\delta} = \infty$$

2. Suppose $\alpha_1 = \alpha_0 = \alpha$. Then

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = M(\delta)$$

$$\text{where } M(\delta) = \begin{cases} 0 & \text{if } \delta \le \theta_1 - \theta_0 \\ 1 - \exp\left(-\dfrac{\delta - (\theta_1 - \theta_0)}{\alpha} \right) > 0 & \text{if } \delta > \theta_1 - \theta_0 \end{cases}$$

$$\text{and } y_{\inf,\delta} = \begin{cases} \infty & \text{if } \delta < \theta_1 - \theta_0 \\ \text{any point in } \mathcal{R} & \text{if } \delta = \theta_1 - \theta_0 \\ \theta_0 + \delta & \text{if } \delta > \theta_1 - \theta_0 \end{cases}$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1 + m(\delta)$$

$$\text{where } m(\delta) = \begin{cases} \exp\left(-\dfrac{\theta_1 - (\delta + \theta_0)}{\alpha} \right) - 1 < 0 & \text{if } \delta < \theta_1 - \theta_0 \\ 0 & \text{if } \delta \ge \theta_1 - \theta_0 \end{cases}$$

$$\text{and } y_{\sup,\delta} = \begin{cases} \theta_1 & \text{if } \delta < \theta_1 - \theta_0 \\ \text{any point in } \mathcal{R} & \text{if } \delta = \theta_1 - \delta_0 \\ \infty & \text{if } \delta > \theta_1 - \theta_0 \end{cases}$$

3. Suppose $\alpha_1 > \alpha_0$.
   (a) If $\delta < \delta_c$,

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = 0, \quad \text{since } M(\delta) = 0 \text{ and } y_{\inf,\delta} = \infty$$

$$F_{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1 + m(\delta)$$

$$\text{where } m(\delta) = \exp\left(-\frac{\theta_1 - (\delta + \theta_0)}{\alpha_0}\right) - 1 < 0, \quad y_{\sup,\delta} = \theta_1$$

   (b) If $\delta \geq \delta_c$,

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = M(\delta)$$

$$\text{where } M(\delta) = \max\left\{\exp\left(-\frac{\max\{\theta_1 - (\delta + \theta_0), 0\}}{\alpha_0}\right) - \exp\left(-\frac{\max\{\theta_0 + \delta - \theta_1, 0\}}{\alpha_1}\right), 0\right\}$$

$$\text{and } y_{\inf,\delta} = \max\{\theta_1, \theta_0 + \delta\} \quad \text{or} \quad \infty \text{ (boundary solution)}$$

$$F^{\mathrm{U}} = 1 + \min\{m(\delta), 0\} = 1 + m(\delta)$$

$$\text{where } m(\delta) = \left(\left(\frac{\alpha_0}{\alpha_1}\right)^{\alpha_1/(\alpha_1 - \alpha_0)} - \left(\frac{\alpha_0}{\alpha_1}\right)^{\alpha_0/(\alpha_1 - \alpha_0)}\right) \exp\left(-\frac{\delta - (\theta_1 - \theta_0)}{\alpha_1 - \alpha_0}\right) < 0$$

$$\text{and } y_{\sup,\delta} = \frac{\alpha_0\alpha_1(\ln\ \alpha_1 - \ln\ \alpha_0) + \alpha_1\theta_0 - \alpha_0\theta_1 + \alpha_1\delta}{\alpha_1 - \alpha_0} \text{(an interior solution)}$$

(Shifted) Pareto marginals. The marginal distributions are:

$$F_1(y) = 1 - \left(\frac{\lambda_1}{\lambda_1 + y - \theta_1}\right)^{\alpha} \quad \text{for} \quad y \in [\theta_1, \infty) \quad \text{and}$$

$$F_0(y) = 1 - \left(\frac{\lambda_0}{\lambda_0 + y - \theta_0}\right)^{\alpha} \quad \text{for} \quad y \in [\theta_0, \infty), \quad \text{where } \alpha, \lambda_1, \theta_1, \lambda_0, \theta_0 > 0$$

Define

$$\delta_c = (\theta_1 - \theta_0) - (\max\{\lambda_1, \lambda_0\})^{\alpha/(\alpha+1)}(\lambda_1^{1/(\alpha+1)} - \lambda_0^{1/(\alpha+1)})$$

1. Suppose $\lambda_1 < \lambda_0$.
   (a) If $\delta \le \delta_c$, then

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = M(\delta)$$

$$\text{where } M(\delta) = (\lambda_0^{\alpha/(\alpha+1)} - \lambda_1^{\alpha/(\alpha+1)})\left(\frac{\lambda_1^{\alpha/(\alpha+1)} - \lambda_0^{\alpha/(\alpha+1)}}{\delta - \lambda_0 + \lambda_1 - \theta_1 + \theta_0}\right)^{\alpha} > 0$$

$$\text{and } y_{\inf,\delta} = \frac{(\delta + \theta_0 - \lambda_0)\lambda_1^{\alpha/(\alpha+1)} + (\lambda_1 - \theta_1)\lambda_0^{\alpha/(\alpha+1)}}{\lambda_1^{\alpha/(\alpha+1)} - \lambda_0^{\alpha/(\alpha+1)}} \text{(an interior solution)}$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1 + m(\delta)$$

$$\text{where } m(\delta) = \min\left\{\left(\frac{\lambda_0}{\lambda_0 + \max\{\theta_1 - \delta - \theta_0, 0\}}\right)^{\alpha} - \left(\frac{\lambda_1}{\lambda_1 + \max\{\theta_0 + \delta - \theta_1, 0\}}\right)^{\alpha}, 0\right\}$$

$$\text{and } y_{\sup,\delta} = \max\{\theta_1, \theta_0 + \delta\} \quad \text{or} \quad \infty \text{ (boundary solution)}$$

   (b) If $\delta > \delta_c$, then

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = M(\delta)$$

$$\text{where } M(\delta) = 1 - \left(\frac{\lambda_1}{\lambda_1 + \theta_0 + \delta - \theta_1}\right)^{\alpha} \ge 0 \quad \text{and} \quad y_{\inf,\delta} = \theta_0 + \delta$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1$$

$$\text{since } m(\delta) = 0 \quad \text{and} \quad y_{\sup,\delta} = \infty$$

2. Suppose $\lambda_1 = \lambda_0 = \lambda$. Then

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = M(\delta)$$

$$\text{where } M(\delta) = \begin{cases} 0 & \text{if } \delta \le \theta_1 - \theta_0 \\ 1 - \left(\dfrac{\lambda}{\lambda + \delta - (\theta_1 - \theta_0)}\right)^{\alpha} \ge 0 & \text{if } \delta > \theta_1 - \theta_0 \end{cases}$$

$$\text{and } y_{\inf,\delta} = \begin{cases} \infty & \text{if } \delta < \theta_1 - \theta_0 \\ \text{any point in } \mathcal{Y} & \text{if } \delta = \theta_1 - \theta_0 \\ \theta_0 + \delta & \text{if } \delta > \theta_1 - \theta_0 \end{cases}$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1 + m(\delta)$$

$$\text{where } m(\delta) = \begin{cases} \left(\dfrac{\lambda}{\lambda - \delta + (\theta_1 - \theta_0)}\right)^{\alpha} - 1 & \text{if } \delta < \theta_1 - \theta_0 \\ 0 & \text{if } \delta \geq \theta_1 - \theta_0 \end{cases}$$

$$\text{and } y_{\sup,\delta} = \begin{cases} \theta_1 & \text{if } \delta < \theta_1 - \theta_0 \\ \text{any point in } \mathcal{Y} & \text{if } \delta = \theta_1 - \theta_0 \\ \infty & \text{if } > \theta_1 - \theta_0 \end{cases}$$

3. Suppose $\lambda_1 > \lambda_0$.

(a) If $\delta < \delta_c$, then

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = 0 \text{ since } M(\delta) = 0, \quad \text{and} \quad y_{\inf,\delta} = \infty$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1 + m(\delta)$$

$$\text{where } m(\delta) = \left(\frac{\lambda_0}{\lambda_0 + \theta_1 - \delta - \theta_0}\right)^{\alpha} - 1 \leq 0 \text{ and } y_{\sup,\delta} = \theta_1$$

(b) If $\delta \geq \delta_c$, then

$$F^{\mathrm{L}}(\delta) = \max\{M(\delta), 0\} = M(\delta)$$

$$\text{where } M(\delta) = \max\left\{\left(\frac{\lambda_0}{\lambda_0 + \max\{\theta_1 - \delta - \theta_0, 0\}}\right)^{\alpha} - \left(\frac{\lambda_1}{\lambda_1 + \max\{\theta_0 + \delta - \theta_1, 0\}}\right)^{\alpha}, 0\right\}$$

$$\text{and } y_{\inf,\delta} = \max\{\theta_1, \theta_0 + \delta\} \text{ or } \infty \text{ (boundary solution)}$$

$$F^{\mathrm{U}}(\delta) = 1 + \min\{m(\delta), 0\} = 1 + m(\delta)$$

$$\text{where } m(\delta) = (\lambda_0^{\alpha/(\alpha+1)} - \lambda_1^{\alpha/(\alpha+1)})\left(\frac{\lambda_1^{\alpha/(\alpha+1)} - \lambda_0^{\alpha/(\alpha+1)}}{\delta - \lambda_0 + \lambda_1 - \theta_1 + \theta_0}\right)^{\alpha} < 0$$

$$\text{and } y_{\sup,\delta} = \frac{(\delta + \theta_0 - \lambda_0)\lambda_1^{\alpha/(\alpha+1)} + (\lambda_1 - \theta_1)\lambda_0^{\alpha/(\alpha+1)}}{\lambda_1^{\alpha/(\alpha+1)} - \lambda_0^{\alpha/(\alpha+1)}} \text{ (an interior solution)}$$

CROSS-VALIDATED BANDWIDTHS AND SIGNIFICANCE TESTING

Christopher F. Parmeter, Zhiyuan Zheng and Patrick McCann

ABSTRACT

The link between the magnitude of a bandwidth and the relevance of the corresponding covariate in a regression has recently garnered theoretical attention. Theory suggests that variables included erroneously in a regression will be automatically removed when bandwidths are selected via cross-validation procedure. However, the connections between the bandwidths of the variables that are smoothed away *and the insights from these same variables when properly* tested *for statistical significance have not been previously studied. This paper proposes a variety of simulation exercises to examine the relative performance of both cross-validated bandwidths and individual and joint tests of significance. We focus on settings where the hypothesis of interest may focus on a single data type (e.g., continuous only) or a mix of discrete and continuous variables. Moreover, we propose an extension of a well-known kernel smoothing significance test to handle mixed data types. Our results suggest that individual tests of significance and variable-specific bandwidths are very close in performance, but joint tests and joint bandwidth recognition*

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 71–98

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025005

produce substantially different results. This underscores the importance of testing for joint significance when one is trying to arrive at the final nonparametric model of interest.

1. INTRODUCTION

Recent research by Hall, Li, and Racine (2007) has documented that least squares cross validation (LSCV) has the asymptotic capability to automatically remove irrelevant variables erroneously included in a local constant regression. Rather than the bandwidths going to zero as the sample size increases, as one would expect under the classical analysis of a data-driven bandwidth selection procedure, the bandwidths associated with the *irrelevant* variables progress toward their theoretical upper bounds (bandwidths for continuous variables have upper bound ∞, whereas discrete variables have an upper bound of 1) as the sample grows. In a local constant setting, this removes continuous variables from the regression, while in a local linear setting, this forces the continuous variable to enter the model linearly.[1] In any setting (local constant, local linear, or local polynomial), a discrete variable whose bandwidth hits its upper bound is deemed irrelevant.

Even with this appealing feature of bandwidths selected via data-driven methods, cross-validated bandwidths are not a panacea for erroneous inclusion of irrelevant variables; the method can assign a large bandwidth to a relevant variable or place a small bandwidth on an irrelevant variable. Thus, the process of testing for variable significance is paramount in applied work. Here, the use of standard nonparametric significance tests (e.g., Racine, 1997; Lavergne & Vuong, 2000; Racine, Hart, & Li, 2006; Gu, Li, & Liu, 2007) allow the researcher to formally test for significance of a regressor, or set of regressors, rather than relying on the relative magnitude of the bandwidth(s). While the performance of these tests is well known, less is understood about the relationship of these tests with the recent results related to the ''smoothing away'' irrelevant variables. This paper considers how standard nonparametric tests of significance compare with respect to raw interpretation of cross-validated bandwidths, both in individual and joint settings.

While the past literature on bandwidth selection is well understood and the literature on significance testing has burgeoned, there does not yet exist a synthesis of the methods when used in conjunction with one another.

For example, simulation results in Gu et al. (2007) suggest that their bootstrap test of significance displays robust size properties for the two data-generating processes considered with respect to their bandwidth choice[2]; however, their supplied bandwidths were selected to satisfy theoretical concerns for the proposed test statistic as opposed to being data driven. As we will argue below, while rule-of-thumb thresholds for cross-validated bandwidths can be used to determine which variables are irrelevant, it is also important to test the significance of any variables not smoothed out of the model. Cai, Gu, and Li (2009) suggest first using local constant estimation to determine the variables that are irrelevant, then testing those variables to ensure statistically that they do not belong in the model and then performing local linear estimation on the *potentially* reduced subset of covariates. Our work here attempts to discern how well the first stage of this approach works in the presence of numerous irrelevant variables.[3]

Given our discussion so far, this paper attempts to present simulation evidence regarding bandwidth estimation in the presence of irrelevant variables and how it contrasts with a standard nonparametric omitted variable test. We focus solely on LSCV given the theoretical results of Hall et al. (2007) and show that the bootstrap test of Gu et al. (2007) can be applied in the presence of mixed data, a ubiquitous feature of economic datasets.[4] Our simulations will be conducted using local constant kernel methods considering both individual and joint tests of significance for continuous, discrete, or mixed continuous/discrete settings under a variety of realistic regression models that include both a high number of irrelevant and relevant variables to mimic settings likely to dominate applied work. Additionally, we wish to determine the ability of using LSCV bandwidths to determine variable relevance in a joint setting. Simulation results in Hall et al. (2007) suggest that the bandwidths, considered individually, display a remarkable ability to detect irrelevant variables. Overall, our simulations will allow us to make broad comments on a number of ad hoc suggestions as to the approach researchers should take to engage in nonparametric model reduction.

The remainder of our paper is structured as follows. Section 2 provides discussion on nonparametric estimation in the presence of mixed discrete–continuous data, LSCV bandwidth selection, and the bootstrap omitted variable test used for our simulations to investigate individual and joint significance. Section 3 provides the details of our simulation study and summarizes our findings. Section 4 discusses future issues that need to be considered when considering nonparametric model selection issues.

2. NONPARAMETRIC ESTIMATION AND SIGNIFICANCE TESTING

2.1. General Nonparametric Kernel Regression

We begin with a generic regression setup:

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where y_i is our response variable, $x_i \in \mathbb{R}^q$ is a vector of covariates, and ε_i represents a random disturbance. Our interest lies in testing significance (individual or joint) for a (set of) covariate(s) in x_i. We use Li-Racine generalized kernels (see Li & Racine, 2004; Racine & Li, 2004). These kernels admit a mix of discrete and continuous covariates which are ubiquitous in applied econometric settings.

Ignoring for the moment the fact that irrelevant regressors may have been included in Eq. (1), we model the unknown relationship through the conditional mean, that is, $m(x_i) = E[y_i|x_i]$ using a method known as local constant regression (see Nadaraya, 1964; Watson, 1964). This allows us to write the regression equation at a given point as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i K_h(x, x_i)}{\sum_{i=1}^{n} K_h(x, x_i)} = \sum_{i=1}^{n} A_i(x) y_i \tag{2}$$

where

$$K_h(x, x_i) = \prod_{s=1}^{q_c} h_s^{-1} l^c\left(\frac{x_s^c - x_{si}^c}{h_s}\right) \prod_{s=1}^{q_u} l^u(x_s^u, x_{si}^u, \lambda_s^u) \prod_{s=1}^{q_o} l^o(x_s^o, x_{si}^o, \lambda_s^o) \tag{3}$$

$K_h(x, x_i)$ is the commonly used product kernel (see Pagan & Ullah, 1999). We have used the notation x_s^c, x_s^o and x_s^u to denote variables that are continuous, ordered, and unordered. Additionally, we have q_c continuous variables, q_u unordered variables, and q_o ordered variables in our regression framework ($q_c + q_u + q_o = q$). We elect to employ smoothing kernels for the discrete data because Racine and Li (2004) have shown that sample splitting (commonly known as the frequency approach) as opposed to smoothing categorical variables can lead to large losses in efficiency. They advocate the use of special kernels designed explicitly for the type of variable being smoothed. In this setting, l^c can be taken to be the standard normal kernel function[5] used for continuous variables with window width $h_s^c = h_s(n)$ associated with the sth component of x^c. l^u is a variation of Aitchison and Aitken's (1976) kernel function for use with

unordered data types:

$$l^{\mathrm{u}}(x_s^{\mathrm{u}}, x_{si}^{\mathrm{u}}, \lambda_s^{\mathrm{u}}) = \begin{cases} 1 - \lambda_s^{\mathrm{u}} & \text{if } x_{si}^{\mathrm{u}} = x_s^{\mathrm{u}} \\ \dfrac{\lambda_s^{\mathrm{u}}}{c_s - 1} & \text{if } x_{si}^{\mathrm{u}} \neq x_s^{\mathrm{u}} \end{cases} \tag{4}$$

where c_s comes from the fact that $x_{\mathrm{u}}^s \in \{0, 1, \ldots c_s - 1\}$. The range of λ_s^{u} is $[0, (c_s - 1)/c_s]$. For an indicator variable, $c_s = 2$ and the largest value that λ_s^{u} can take is $1/2$. l^{o} is the Wang and Ryzin (1981) kernel function designed for smoothing ordered discrete variables, defined as

$$l^{\mathrm{o}}(x_s^{\mathrm{o}}, x_{si}^{\mathrm{o}}, \lambda_s^{\mathrm{u}}) = (\lambda_s^{\mathrm{u}})^{|x_s^{\mathrm{o}} - x_{si}^{\mathrm{o}}|} \tag{5}$$

where the range of λ_s^{o} is $[0, 1]$. This kernel function is slightly different from the original kernel proposed by Wang and Ryzin (1981). Li and Racine (2006, p. 145) show that Wang and Ryzin's (1981) kernel function does not possess the ability to smooth away irrelevant ordered discrete variables when that variable has at least three categories.

Eq. (2) can be written in matrix notation to display it in a more compact form. Let $\mathbf{i}$ denote an $n \times 1$ vector of ones and let $\mathcal{K}(x)$ denote the diagonal matrix with jth element $K_h(x, x_j)$. Also, denote by y the $n \times 1$ vector of responses. Then, we can express our LCLS estimator as

$$\hat{m}(x) = (\mathbf{i}'\mathcal{K}(x)\mathbf{i})^{-1}\mathbf{i}'\mathcal{K}(x)y \tag{6}$$

The name local constant comes from the fact that our estimator is a weighted regression of a *constant* on our response vector. The weights are determined locally by the associated covariates and the bandwidths. This is similar to generalized least squares, except our weights change for each point on our regression curve as opposed to being globally determined as they are in standard least squares approaches.

2.2. Cross-Validated Bandwidth Selection

Estimation of the bandwidths (h, λ^{u}, λ^{o}) is typically the most salient factor when performing nonparametric estimation. For example, choosing a very small h means that there may not be enough points in a neighborhood of the point being smoothed and thus we may get an undersmoothed estimate (low bias, high variance). On the other hand, choosing a very large h, we may smooth over too many points and thus get an oversmoothed estimate (high bias, low variance). This trade-off is a well-known dilemma in applied

nonparametric econometrics and thus we usually resort to automatic selection procedures to obtain the bandwidths. Although there exist many selection methods, Hall et al. (2007) (HLR hereafter) have shown that LSCV has the ability to smooth away irrelevant variables that may have been erroneously included into the unknown regression function. Specifically, the bandwidths are chosen to minimize

$$\mathrm{CV}(h,\lambda) = \underset{\{h,\lambda\}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2 \tag{7}$$

where $\hat{m}_{-i}(x_i)$ is the common leave-one-out estimator. An alternative data-driven approach with impressive finite sample performance is known as improved $\mathrm{AIC_c}$ and was proposed by Hurvich, Simonoff, and Tsai (1998). Li and Racine (2004) show that in small samples improved $\mathrm{AIC_c}$ performs admirably compared to LSCV when one employs a local linear least squares approach. Even though the performance of smoothing parameters estimated via the $\mathrm{AIC_c}$ criterion have desirable features, we elect to use the standard LSCV criterion to estimate our bandwidths given the theoretical work of HLR.

For the discrete variables, the bandwidths indicate which variables are relevant, as well as the extent of smoothing in the estimation. From the definitions for the ordered and unordered kernels, it follows that if the bandwidth for a particular unordered or ordered discrete variable equals zero, then the kernel reduces to an indicator function and no weight is given to observations for which $x_i^{\mathrm{o}} \neq x_j^{\mathrm{o}}$ or $x_i^{\mathrm{u}} \neq x_j^{\mathrm{u}}$; in this case it is as if the research had engaged in sample splitting. On the other hand, if the bandwidth for a particular unordered or ordered discrete variable reaches its upper bound, then equal weight is given to observations with $x_i^{\mathrm{o}} = x_j^{\mathrm{o}}$ and $x_i^{\mathrm{o}} \neq x_j^{\mathrm{o}}$. In this case, the variable is completely smoothed out (and thus does not impact the estimation results). For unordered discrete variables, the upper bound is given by $(c_r - 1)/c_r$ where c_r represents the number of unique values taken on by the variable. For example, a categorical variable for geographic location which takes on 5 values would have an upper bound for its bandwidth of $4/5 = 0.8$. For ordered discrete variables, the upper bound is always unity. See HLR for further details.

HLR have shown that the inclusion of *irrelevant* regressors does not add to the "curse of dimensionality." Their paper shows that when one uses cross-validation procedures to select the appropriate amount of smoothness of the unknown function, the covariates that are irrelevant are eliminated from the conditional mean relationship. In essence, instead of the

bandwidth decreasing to zero at an appropriate rate when the sample is increased, the bandwidths move toward their theoretical upper bounds. A large bandwidth effectively suggests that the associated variable is being smoothed out as the product kernel in Eq. (3) can be rewritten as two distinct product kernels, one for the relevant variables and another for the irrelevant variables. The large bandwidths force the product kernel pertaining to the irrelevant variables to be constant across all observations. Thus, given that our conditional mean is a ratio, the irrelevant variables cancel out of the formula and it is as if the researcher had failed to include them in the first place. This property allows nonparametric estimators to not only allow for functional form misspecification, but relevant covariate selection at the same time.

However, there is no free lunch for this method as it hinges on several facets that need to be considered on a case-by-case basis. First, the key assumption used by HLR asks that the irrelevant regressors are independent of the relevant regressors, something unlikely to hold in practice.[6] Second, it is not entirely clear how well this method works as the set of relevant regressors is increased. HLR's finite sample simulations investigated at most two relevant regressors while their empirical application considered six variables for 561 observations in which only two regressors were deemed relevant according to their procedure. Clearly more work needs to be done to assess the performance of the bandwidths for very small sample sizes and for large sets of potential regressors, a task we take up in our simulations.[7]

What is noteworthy of the HLR finding is that the cross-validated bandwidths provide a cheap and easy way of assessing individual significance. However, three core issues remain. First, as our simulations show, the method does not perform well when a large number of irrelevant variables are included, a not uncommon feature of applied work. Second, ignoring the number of irrelevant variables included, a large bandwidth does not provide a *p*-value to assess the level of significance. The HLR theory only provides a rule of thumb for saying yes or no to a variable's relevance. Lastly, while the theory predicts that all irrelevant variables are smoothed away simultaneously, there has been no simulation study to determine if the impressive finite sample performance of LSCV bandwidths holds when one looks for joint significance. Moreover, there is no appropriate rule of thumb in this case, as a "test" for three variables being insignificant is confusing if two of the variables are smoothed away but one is not, how does one draw conclusions from this type of setup?

2.3. Testing for Variable Significance

While the properties of LSCV discovered by HLR suggest that irrelevant variables are removed, statistically there is no way to determine joint (in)significance by simply appealing to the bandwidths returned. A formal test for joint significance of variables is thus warranted to make statistically precise statements about the relevance of variables entering into the model.

To determine whether or not a set of variables are jointly significant, we utilize the tests of Lavergne and Vuong (2000) and Gu et al. (2007). Consider a nonparametric regression model of the form

$$y_i = m(w_i,\ z_i) + u_i \tag{8}$$

Here, we discuss in turn the case where the variables in z are all continuous (Gu et al., 2007), are all discrete (Racine et al., 2006), or a mixture of discrete and continuous insignificant variables, but w may contain mixed data. In what follows, let w have dimension r and z have dimension $q-r$. The null hypothesis is that the conditional mean of y does not depend on z.

$$H_0 : E(y|w,z) = E(y|w) \tag{9}$$

2.3.1. All Continuous Case

Define $u = y - E(y|w)$. Then $E(u|x) = 0, \quad x = (w,z)$, under the null we can construct a test statistic based on

$$E\{u f_w(w) E[u f_w(w)|x] f(x)\} \tag{10}$$

where $f_w(w)$ and $f(x)$ are the pdfs of w and x, respectively. A feasible test statistic is given by

$$\hat{I}_n^c = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} (y_i - \hat{y}_i)\hat{f}_w(w_i)(y_j - \hat{y}_j)\hat{f}_w(w_j) W(x_i, x_j, h, \lambda^{\mathrm{o}}, \lambda^{\mathrm{u}}) \tag{11}$$

where $W(x_i,\ x,\ h,\ \lambda^{\mathrm{o}},\ \lambda^{\mathrm{u}})$ is the Li-Racine generalized product kernel discussed in Eq. (3) and

$$\hat{f}_w(w_i) = \frac{1}{n-1} \sum_{j=1,\, j\neq i}^{n} W(w_i, w_j, h_w, \lambda_w^{\mathrm{o}}, \lambda_w^{\mathrm{u}})$$

is the leave-one-out estimator of $f_w(w_i)$. The leave-one-out estimator of $E(y_i|w_i)$ is

$$\hat{y}_i = \frac{1}{(n-1)\hat{f}_w(w_i)} \sum_{j=1,\, j\neq i}^{n} y_j W(w_i, w_j, h_w, \lambda_w^{\mathrm{o}}, \lambda_w^{\mathrm{u}})$$

One shortcoming of this test is that it requires the researcher to estimate (or determine) two sets of bandwidths, one for the model under the null and another for the model under the alternative. For large samples this may be computationally expensive. Under the null hypothesis, a studentized version of the statistic presented in Eq. (11) is

$$T_n^{\mathrm{c}} = (nh_1h_2\ldots h_q)^{1/2}\hat{I}_n^{\mathrm{c}}/\hat{\sigma}_n^{\mathrm{c}} \to N(0,1) \tag{12}$$

where

$$(\hat{\sigma}_n^{\mathrm{c}})^2 = \frac{2h_1h_2\cdots h_q}{n^2}\sum_{i=1}^{n}\sum_{j=1,\,j\neq i}^{n}(y_i-\hat{y}_i)^2\hat{f}_w(w_i) \times (y_j-\hat{y}_j)^2\hat{f}_w(w_j)W(x_i,x_j,h,\lambda^{\mathrm{o}},\lambda^{\mathrm{u}}) \tag{13}$$

In a small-scale simulation study, Gu et al. (2007) show that use of the asymptotic distribution for this test statistic has inaccurate size and poor power. A bootstrap procedure is suggested instead. The bootstrap test statistic is obtained via the following steps:

(i) For $i = 1, 2, \ldots, n$, generate the two-point wild bootstrap error $u_i^* = [(1-\sqrt{5})/2]\hat{u}_i$, where $\hat{u}_i = y_i - \hat{y}_i$ with probability $r = (1-\sqrt{5})/2\sqrt{5}$ and $u_i^* = [(1+\sqrt{5})/2]\hat{u}_i$ with probability $1-r$.

(ii) Use the wild bootstrap error u_i^* to construct $y_i^* = \hat{y}_i + u_i^*$, then obtain the kernel estimator of $E^*(y_i^*|w_i)f_w(w_i)$ via

$$\hat{y}_i^*\hat{f}_w(w_i) = \frac{1}{n-1}\sum_{j=1,\,j\neq i}^{n} y_j^*W(w_i,w_j,h_w,\lambda_w^{\mathrm{o}},\lambda_w^{\mathrm{u}})$$

The estimated density-weighted bootstrap residual is

$$\hat{u}_i^*\hat{f}_w(w_i) = (y_i^*-\hat{y}_i^*)\hat{f}_w(w_i) = y_i^*\hat{f}_w(w_i) - \hat{y}_i^*\hat{f}_w(w_i)$$

(iii) Compute the standardized bootstrap test statistic $T_n^{\mathrm{c}*}$, where y^* and $\hat{y}^*$ replace y and $\hat{y}$ wherever they occur.

(iv) Repeat steps (i)–(iii) B times and obtain the empirical distribution of the B bootstrap test statistics. Let $T_{n(\alpha B)}^{\mathrm{b}*}$ denote the α-percentile of the bootstrap distribution. We will reject the null hypothesis at significance level α if $T_n^{\mathrm{c}} > T_{n(\alpha B)}^{\mathrm{c}*}$.

In practice, researchers may use any set of bandwidths for estimation of the test statistic. However, for the test to be theoretically consistent, the bandwidths used for the model under the alternative need to have a slower rate than those used for the model under the null hypothesis if

$\dim(w) = r \leq q/2$ (see Gu et al., 2007, Assumption A2). This guarantees that the mean-square error of the null model is smaller than that coming from the alternative model. In essence, the residuals used in Eq. (11) or Eq. (12) need to converge at a faster rate than the rate on the bandwidths used for the estimation of $E(u|x) = 0$ to ensure that the test statistic is properly capturing this relationship.

An empirical approach would be to use LSCV to estimate the scale factors of the bandwidths in each stage. However, this procedure has two shortcomings. First, the theory in HLR suggests that the bandwidths associated with irrelevant variables do not converge to zero at any rate, inconsistent with Assumption A2 of Gu et al. (2007). Second, ignoring theoretical rates the bandwidths are supposed to possess, the test statistics in Eqs. (11) and (12) do not incorporate the presence of the variables smoothed away with LSCV bandwidths. In the simulations reported in Gu et al. (2007), they smoothed both relevant and irrelevant variables with similar bandwidths.

2.3.2. All Discrete Case

While the nonparametric significance test of Gu et al. (2007) was initially designed and studied theoretically for the case of continuous regressors, computationally the test can easily be generalized to handle mixed discrete–continuous data, both for testing and estimation by simple appeal to the generalized product kernels provided in Racine and Li (2004). In our simulations, we report size and power by simply using the bootstrap test of Lavergne and Vuong (2000) and Gu et al. (2007). While their theory pertains only to continuous variables, the null hypothesis of interest does not depend on the data type, and it is easy to replace the continuous product kernels with generalized Li-Racine kernels.

In Racine and Li (2004), it was shown that the optimal rate for continuous variable bandwidths for consistent estimation of a regression function in the local constant setting was not affected by the presence of discrete variables. Moreover, they also showed that the optimal rate for the bandwidths associated with discrete variables were only dependent upon the number of continuous variables. To be explicit, the bandwidths associated with continuous variables have optimal rate $n^{-1/(4+q_c)}$ where q_c is the number of continuous variables. Moreover, the bandwidths pertaining to discrete covariates have optimal rate $n^{-2/(4+q_c)}$. Thus, a strategy for implementing the aforementioned omitted variable test in the presence of discrete variables in the null hypothesis would be to use the rates consistent with Racine and Li (2004) and Assumption A2 of Gu et al. (2007)

(guaranteeing that the mean-square error of the restricted model goes to zero faster than that of the unrestricted model), which is what we take up in our simulations.

2.3.3. Mixed Discrete–Continuous Case

To the authors knowledge no formal test that admits both discrete and continuous variables to be tested jointly exists in the literature. We determine the appropriateness of the Gu et al. (2007) test when both discrete and continuous variables enter into the null hypothesis. While their theory for the bootstrap test statistic focuses solely on continuous variables, our conjecture is that in finite samples, there is no reason why one cannot include discrete variables into the discussion. The key difference with the test statistic's construction is that generalized kernels will need to be used as opposed to the standard continuous product kernels used in Gu et al. (2007).

While no formal theory exists for the distribution of the test statistic under the null in the presence of mixed data, it is hypothesized that the asymptotic properties of the test can be uncovered using stochastic equicontinuity arguments similar to those in Hsiao, Li, and Racine (2007, Theorem 2.1). The reason for this is that the test of correct functional form in the presence of mixed data proposed by Hsiao et al. (2007) has exactly the same form as the test proposed by Lavergne and Vuong (2000) except that the residuals that enter into the test statistic come from a nonparametric model as opposed to a parametric model (for the functional form test). Moreover, this same rational suggests that the asymptotic distribution of the bootstrap version of Hsiao et al.'s (2007, Theorem 2.2) model specification test will hold as well. While our arguments for the use of the Lavergne and Vuong's (2000) significance test are heuristic, as we will see, our size and power appear to confirm that the use of this test can perform admirably in the face of mixed data. Additionally, as Lavergne and Vuong (2000) show in the model with only continuous covariates, a standardized test statistic has limiting standard normal distribution. In our simulations, we too standardize our test statistic in exactly the same fashion, except that no formal theory exists to show that this standardization is correct.

3. MONTE CARLO ILLUSTRATION

As discussed earlier, a majority of the proposed tests of significance in the literature, while capable of handling multiple variables, provide

simulation studies that focus solely on a single regressor (either continuous or discrete). Table 1 lists many of the recent simulation studies for varying nonparametric significance tests and highlights the sample sizes used and the number of variables in the model. The w in the table refers to variables that are always significant, while z represents the potentially irrelevant variable used for assessing size and power properties of the test.

Outside of Racine et al. (2006) and HLR, all of the papers listed use only continuous variables and consider only a single relevant regressor coupled with a single irrelevant regressor. Also, most of the simulation studies use sample sizes of 50 and 100 to assess the properties of the test under study. Additionally, there is no consensus in this literature as to the appropriate data generating process (DGP). Several authors have used high-frequency DGPs while others have employed simple linear terms. Also, a majority of the papers have used ad hoc bandwidths selected to meet the theoretical underpinnings of their test as opposed to investigating the properties of the test in likely encountered applied settings. The simulation studies of Racine (1997) and Racine et al. (2006) have used data-driven methods with notable success as the test statistics in these settings appear to be independent of the bandwidth choice.

Our simulations are designed to include both low- and high-frequency settings and are similar to the DGPs used by the studies listed in Table 1. They will allow us to gauge how the tests will work when multiple continuous and discrete regressors are present and one is interested in joint significance testing, a common occurrence in applied econometric work. We also perform individual tests as well to compare them directly to the bandwidths obtained via cross validation. Additionally, we allow for nonlinearities both through interactions across variables as well as directly via nonlinear terms of the covariate(s). The beauty of nonparametric methods (and the bandwidths) is that regardless of the type of nonlinearity, the method is capable of detecting it. Thus, suppose one posited that wages were nonlinearly related to education and the impact of education varied across race. Here, we have that wages are directly nonlinear in education and indirectly nonlinear across race. In either (both) setting(s), bandwidths obtained via data-driven methods will detect if these variables (race and education) are relevant, but they do not suggest *which* type of nonlinearity is present. To uncover the interaction effect between race and education, one could use the nonparametric Chow test of either Lavergne (2001) or Racine et al. (2006).

Table 1. Characterization of Previous Simulation Studies Regarding Tests of Significance.

Racine (1997, Table 1)	
DGP	$y = \sin(2\pi w) + \varepsilon$
R.V.	w and z continuous
Sample sizes	$n = 50$
Bandwidth	LSCV
Lavergne and Vuong (2000, Tables 1 and 2)	
DGP	$y = w + w^3 + d(z) + \varepsilon$ $d(z) = \alpha z$ or $d(z) = \sin(\alpha \pi z)$
R.V.	w and z continuous
Sample sizes	$n = 50, 200$
Bandwidth	Rule of thumb
Delgado and González-Manteiga (2001, Table 1)	
DGP	$y = m(w) + d(z) + \varepsilon$ $m(w) = 1 + w$ or $m(w) = 1 + \sin(10w)$ $d(z) = \alpha \sin(z)$
R.V.	w and z continuous
Sample sizes	$n = 50, 100$
Bandwidth	Rule of thumb
Racine et al. (2006, Tables 1 and 2)	
DGP	$y = 1 + z_2 + w + d(z) + \varepsilon$ $d(z) = \alpha z_1(1 + w^2)$
R.V.	z_1, z_2 discrete, w continuous
Sample sizes	$n = 50, 100$
Bandwidth	LSCV
Gu et al. (2007, Tables 3–8)	
DGP	$y = w + w^3 + d(z) + \varepsilon$ $d(z) = \alpha z$ or $d(z) = \alpha \sin(2\pi z)$
R.V.	w and z continuous
Sample sizes	$n = 50, 100$
Bandwidth	Rule of thumb
Hall et al. (2007, Table 2)	
DGP	$y = w_1 + w_2 + \varepsilon$
R.V.	w_1, z_1 discrete and w_2, z_2 continuous
Sample sizes	$n = 100, 250$
Bandwidth	LSCV

We conduct Monte Carlo simulations according to the following data-generating processes:

DGP_1: $y = x_1 + \delta x_2 + \delta x_3 + \varepsilon$.
DGP_2: $y = x_1 + \delta x_1 x_2 + \delta x_1 x_3^2 + \varepsilon$.
DGP_3: $y = x_1 + x_2 + x_3 + \delta x_1(1 + x_2^2)\sin(0.5\pi x_3) + \delta x_3 \sin(x_2^3) + \varepsilon$.
DGP_4: $y = x_1 + x_2 + x_1 x_2 + \delta x_1 x_3^2 + x_1^2 x_4 + \delta x_2 x_3 x_5 + \delta x_6^3 + \varepsilon$.

Our DGPs are given in increasing order of complexity, with DGP_3 indicative of a high-frequency model. DGP_1 and DGP_2 are similar to the main DGP used in Lavergne and Vuong (2000). The key difference is that we have added an additional variable, and we allowed for interactions between them, potentially making it harder to determine significance. DGP_3 is consistent with many of the simulation studies listed in Table 1. To appropriately determine the size properties of Gu et al.'s (2007) bootstrap test, we set $\delta = 0$. To determine power properties, we set $\delta = 0.1$, 0.5, or 1. We consider both continuous-only and discrete-only settings for DGP_1–DGP_3 and use DGP_4 for our mixed discrete–continuous setting. We determine both size and power for samples sizes of $n = 100$ and 200. We use 399 bootstrap replications to determine the bootstrap p-value of all test statistics and use 399 Monte Carlo simulations for each scenario considered.

In our continuous-only setting, we generate all variables as independent $N(0,1)$, including ε. In our discrete-only setting, we change x_2 from a continuous variable to an unordered variable with $\Pr[x_{i2} = 1] = 0.35$ and x_3 from a continuous variable to an ordered categorical variable with $P(x_{i3} = 0) = 0.25$, $P(x_{i3} = 1) = 0.4$, and $P(x_{i3} = 2) = 0.35$.[8]

Since the testing properties of the continuous-only and discrete-only case have been canvassed in the literature, we use an expanded DGP that includes mixed data to determine the ability of the Gu et al. (2007) test. DGP_4 is *only* studied in our simulations involving mixed discrete–continuous null hypotheses. The addition of an additional continuous regressor suggests that the size properties of the test will likely be effected given our use of small sample sizes. To generate data from this DGP, we draw x_1, x_2, x_3, and ε independent of each other from a standard normal. x_4 is generated as an unordered categorical variable with $\Pr[x_{i4} = 1] = 0.35$, while x_5 and x_6 are ordered categorical variables with $\Pr[x_{i5} = 0] = 0.25$, $\Pr[x_{i5} = 1] = 0.4$ and $\Pr[x_{i5} = 2] = 0.35$ and $\Pr[x_{i6} = 0] = \Pr[x_{i6} = 1] = 0.25$ and $\Pr[x_{i6} = 2] = 0.5$, respectively.

We consider two rule-of-thumb metrics regarding the LSCV bandwidths for the continuous covariates to determine if a variable (or set thereof) is irrelevant, either two standard deviations (2 SD) or the interquartile range

(IQR) for each variable. For discrete predictors, we use 80% of the LSCV bandwidths' theoretical upper bounds. For example, a dummy variable has a bandwidth with upper bound 0.5, so our rule for assessing this variable's irrelevance would be a bandwidth larger than 0.4. When assessing joint insignificance, we use a box-type method where all variables under consideration must be smoothed out individually to be deemed jointly irrelevant.

3.1. Continuous-Only Case

Tables 2–4 display our results in the continuous variable setting. These tables contain quite a lot of information and as such we describe in detail what we are reporting. First, we report the raw results from the Gu et al. (2007) test statistic using their ad hoc bandwidth selection procedure. Their selection of the bandwidths, when only continuous variables are present, is to construct individual bandwidths as $c \cdot \mathrm{SD}_j n^{-1/(4+d)}$ where c is a scaling factor common to all variables, SD_j the in-sample standard deviation of the jth variable being smoothed and d is a variable used to control the rate of decay of the bandwidth to ensure consistency with Assumption A2 of Gu et al. (2007). We note that the theory underlying Gu et al. (2007) suggests that the bandwidths used for the unrestricted model be smaller than what is theoretically consistent. To do this, one can keep the scaling portion of the bandwidth fixed ($c \cdot \mathrm{SD}_j$) but change the rate on the bandwidth (d). Our reported results come from undersmoothing the unrestricted model while using optimal smoothing for the restricted model as is consistent with Gu et al. (2007, Theorems 2.1 and 2.2). We use the same set of scaling constants as in Gu et al. (2007) ($c = 0.25, 0.5, 1, 2$). We report size ($\delta = 0$) and power ($\delta = 0.1$, 0.5, or 1) in the first block at the 1%, 5%, and 10% levels. The second block of our table looks at the performance of the LSCV bandwidths using our ad hoc rules for assessing irrelevance (individual or joint) as gauged by either 2 sd (columns labeled 2 SD) of each variable or the interquartile range of the variable (column labeled IQR).

We see from these simulation results several interesting features. First, the size of the Gu et al. (2007) is very close to nominal levels using their bandwidth selection measure which is encouraging given that we are including an additional continuous covariate beyond what their simulations investigated. As noted earlier, the power of the test appears to depend somewhat on the choice of smoothing coefficient chosen, although the power increases as the sample size goes up across all three of our DGPs.

Table 2. DGP_1.

(a) Gu et al. (2007) Bandwidths

	$c = 0.25$			$c = 0.5$			$c = 1$			$c = 2$		
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.050	0.073	0.013	0.050	0.108	0.018	0.038	0.103	0.005	0.053	0.105
$\delta = 0.1$	0.003	0.045	0.110	0.013	0.065	0.120	0.013	0.050	0.123	0.038	0.103	0.163
$\delta = 0.5$	0.015	0.080	0.155	0.080	0.228	0.346	0.378	0.612	0.742	0.832	0.965	0.987
$\delta = 1$	0.020	0.168	0.318	0.366	0.659	0.784	0.967	1.000	1.000	1.000	1.000	1.000
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.053	0.0080	0.013	0.053	0.05	0.015	0.063	0.125	0.015	0.058	0.103
$\delta = 0.1$	0.010	0.028	0.090	0.015	0.063	0.113	0.035	0.103	0.155	0.040	0.138	0.223
$\delta = 0.5$	0.023	0.103	0.188	0.158	0.373	0.489	0.722	0.892	0.945	0.987	1.000	1.000
$\delta = 1$	0.090	0.358	0.524	0.799	0.957	0.980	1.000	1.000	1.000	1.000	1.000	1.000

(b) LSCV Bandwidth Results

	2 SD				IQR			
	x_1	x_2	x_3	Joint	x_1	x_2	x_3	Joint
$n = 100$								
$\delta = 0$	0.000	0.687	0.604	0.426	0.000	0.757	0.712	0.561
$\delta = 0.1$	0.000	0.551	0.564	0.318	0.000	0.669	0.659	0.446
$\delta = 0.5$	0.000	0.018	0.013	0.000	0.000	0.048	0.043	0.000
$\delta = 1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$n = 200$								
$\delta = 0$	0.000	0.669	0.639	0.441	0.000	0.900	0.865	0.769
$\delta = 0.1$	0.000	0.486	0.489	0.263	0.000	0.822	0.837	0.687
$\delta = 0.5$	0.000	0.000	0.000	0.000	0.000	0.168	0.153	0.008
$\delta = 1$	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000

We do not present testing results for the bandwidths obtained via LSCV as they were inappropriately sized,[9] and per the earlier discussion, do not satisfy the necessary theoretical underpinnings of the asymptotic validity of the test.

Our bandwidth results suggest that data-driven methods successfully remove irrelevant variables, although the percentage of times both variables are removed jointly is, as expected, lower than how often each variable is smoothed away. Additionally, we note that using the IQR of a variable seems to consistently determine the appropriate irrelevant variables

Table 3. DGP_2.

(a) Gu et al. (2007) Bandwidths

	$c = 0.25$			$c = 0.5$			$c = 1$			$c = 2$		
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.050	0.073	0.013	0.050	0.108	0.018	0.038	0.103	0.005	0.053	0.105
$\delta = 0.1$	0.008	0.065	0.093	0.010	0.065	0.120	0.005	0.048	0.103	0.010	0.065	0.108
$\delta = 0.5$	0.003	0.088	0.078	0.020	0.088	0.155	0.058	0.148	0.218	0.073	0.228	0.323
$\delta = 1$	0.008	0.175	0.143	0.043	0.175	0.301	0.263	0.536	0.657	0.499	0.769	0.857
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.038	0.080	0.013	0.053	0.105	0.015	0.063	0.125	0.015	0.058	0.103
$\delta = 0.1$	0.003	0.038	0.080	0.015	0.045	0.108	0.023	0.088	0.128	0.015	0.058	0.100
$\delta = 0.5$	0.005	0.055	0.095	0.020	0.088	0.158	0.068	0.213	0.346	0.213	0.469	0.612
$\delta = 1$	0.033	0.143	0.256	0.198	0.466	0.602	0.732	0.902	0.952	0.952	0.992	0.995

(b) LSCV Bandwidth Results

	2 SD				IQR			
	x_1	x_2	x_3	Joint	x_1	x_2	x_3	Joint
$n = 100$								
$\delta = 0$	0.000	0.687	0.604	0.426	0.000	0.757	0.712	0.561
$\delta = 0.1$	0.000	0.554	0.586	0.341	0.000	0.662	0.672	0.451
$\delta = 0.5$	0.000	0.100	0.185	0.018	0.000	0.221	0.203	0.030
$\delta = 1$	0.000	0.020	0.038	0.005	0.000	0.053	0.043	0.005
$n = 200$								
$\delta = 0$	0.000	0.669	0.639	0.441	0.000	0.900	0.865	0.769
$\delta = 0.1$	0.000	0.491	0.617	0.343	0.000	0.872	0.825	0.724
$\delta = 0.5$	0.000	0.018	0.035	0.003	0.000	0.514	0.070	0.018
$\delta = 1$	0.000	0.000	0.000	0.000	0.000	0.133	0.003	0.000

(both individually and jointly) beyond that of using 2 SD of the variable. However, this comes at a cost as the IQR also erroneously smooths away relevant variables at a higher frequency that does using 2 SD. This is due to the fact that in general, our IQR was narrower than 2 SD and as such this resulted in better performance for appropriately smoothing away irrelevant variables but poorer performance when considering relevant variables.

What is interesting from these simulations is that while on an individual basis using the bandwidths to determine which variables to formally test,

Table 4. DGP_3.

(a) Gu et al. (2007) Bandwidths

	$c = 0.25$			$c = 0.5$			$c = 1$			$c = 2$		
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.050	0.073	0.013	0.050	0.108	0.018	0.038	0.103	0.005	0.053	0.105
$\delta = 0.1$	0.005	0.060	0.095	0.008	0.060	0.123	0.005	0.060	0.125	0.025	0.090	0.138
$\delta = 0.5$	0.008	0.080	0.165	0.083	0.206	0.308	0.271	0.481	0.607	0.579	0.837	0.917
$\delta = 1$	0.028	0.160	0.333	0.396	0.664	0.769	0.957	0.985	0.992	1.000	1.000	1.000
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.038	0.080	0.013	0.053	0.105	0.015	0.063	0.125	0.015	0.058	0.103
$\delta = 0.1$	0.008	0.028	0.085	0.018	0.055	0.103	0.025	0.100	0.143	0.038	0.118	0.198
$\delta = 0.5$	0.023	0.103	0.170	0.138	0.318	0.429	0.586	0.772	0.880	0.942	0.980	0.985
$\delta = 1$	0.085	0.346	0.509	0.797	0.927	0.960	1.000	1.000	1.000	1.000	1.000	1.000

(b) LSCV Bandwidth Results

	2 SD				IQR			
	x_1	x_2	x_3	Joint	x_1	x_2	x_3	Joint
$n = 100$								
$\delta = 0$	0.000	0.687	0.604	0.426	0.000	0.757	0.712	0.561
$\delta = 0.1$	0.000	0.554	0.586	0.341	0.000	0.662	0.672	0.451
$\delta = 0.5$	0.000	0.100	0.185	0.018	0.000	0.221	0.203	0.030
$\delta = 1$	0.000	0.020	0.038	0.005	0.000	0.053	0.043	0.005
$n = 200$								
$\delta = 0$	0.000	0.669	0.639	0.441	0.000	0.900	0.865	0.769
$\delta = 0.1$	0.000	0.491	0.617	0.343	0.000	0.872	0.825	0.724
$\delta = 0.5$	0.000	0.018	0.035	0.003	0.000	0.514	0.070	0.018
$\delta = 1$	0.000	0.000	0.000	0.000	0.000	0.133	0.003	0.000

if they are indeed irrelevant, this does not appear to be the case jointly. When it comes to a joint decision, using the bandwidths to determine irrelevance results in a lower total percentage of the number of times the bandwidths jointly arrive at the appropriate set of irrelevant variables, using our joint rule-of-thumb method. For example, in Table 4 using $2 \cdot \text{SD}$ and $n = 200$, we see that in 66.9% of all the simulations x_2 is correctly smoothed out of the regression while 63.9% of all the simulations x_3 is appropriately removed, but jointly they are correctly removed in only 44% of the

simulations. Alternatively, using the IQR rule of thumb, x_2 is removed 90% of the time and x_3 is removed 86.5% of the time, resulting in them being jointly removed 76.9% of the time. As noted earlier though, the IQR seems to penalize too much when indeed the variables are relevant. Also, when n increases from 100 to 200, we see that for $\delta = 0.1$ and 0.5 the percentage of times a variable that is relevant is deemed irrelevant using the IQR has increased. This appears to be the case for $\delta = 0.1$ using 2 SD as a rule of thumb as well.

Overall, these simulations suggest that a sound empirical strategy would be to use local constant regression coupled with LSCV bandwidth selection to determine the variables that are initially smoothed away (based on the results here using 2 SD as a gauge) and then to use the test of Gu et al. (2007) to determine which of the remaining variables whose relevance is under consideration is actually significant. This strategy will potentially circumvent the use of "extreme" bandwidths in the construction of the test statistic that resulted in the poor size properties that we found in our simulations.

3.2. Discrete-Only Case

Testing significance of discrete variables provides an opportunity to gauge how a finite upper bound on a bandwidth impacts the test results as opposed to an infinite upper bound. We saw that in the continuous-only case that our rule-of-thumb methods were able to detect individual irrelevance but refocusing our attention toward joint relevance resulted in diminished performance relative to the testing results. Tables 5–7 provide size and power results for our test statistic using only discrete variables in the null hypothesis and a threshold of relevance set at 80% of the upper bound using bandwidths determined via LSCV. Since this test has not been used in practice before, we examine individual tests of significance as well as joint tests of significance.

The first thing we note is that across the three DGPs, the test has impressive size and power using the ad hoc bandwidths in both the individual and joint testing setups. Again, we follow closely the theory laid out in Gu et al. (2007) and undersmooth our bandwidths in the unrestricted model estimation while using the standard level of smoothing in the restricted model. When we consider the determination of relevance as gauged via 80% of the theoretical upper bounds, we see that individually the bandwidths determine a high percentage of the simulations that the appropriate variables are smoothed out and this percentage is increasing as n increase.

Table 5. DGP_1 Where x_2 and x_3 are Discrete Variables.

(a) Gu et al. (2007) Bandwidths

	$c = 0.25$			$c = 0.5$			$c = 1$			$c = 2$		
x_2 and x_3 joint significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.015	0.053	0.105	0.023	0.063	0.130	0.015	0.080	0.135	0.020	0.075	0.125
$\delta = 0.1$	0.020	0.070	0.110	0.023	0.078	0.135	0.023	0.088	0.160	0.035	0.088	0.175
$\delta = 0.5$	0.135	0.328	0.461	0.258	0.506	0.619	0.424	0.662	0.767	0.544	0.777	0.872
$\delta = 1$	0.727	0.925	0.972	0.965	0.992	0.995	0.997	0.997	1.000	0.997	1.000	1.000
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.003	0.050	0.083	0.015	0.045	0.073	0.008	0.055	0.088	0.003	0.050	0.103
$\delta = 0.1$	0.005	0.050	0.110	0.025	0.053	0.118	0.015	0.070	0.118	0.023	0.110	0.165
$\delta = 0.5$	0.386	0.664	0.764	0.642	0.832	0.907	0.792	0.947	0.972	0.925	0.980	0.990
$\delta = 1$	0.997	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
x_2 individual significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.055	0.090	0.005	0.038	0.108	0.005	0.048	0.108	0.015	0.050	0.108
$\delta = 0.1$	0.010	0.050	0.095	0.005	0.053	0.125	0.013	0.063	0.140	0.018	0.083	0.135
$\delta = 0.5$	0.038	0.160	0.261	0.098	0.241	0.356	0.135	0.338	0.471	0.198	0.444	0.571
$\delta = 1$	0.013	0.058	0.128	0.080	0.293	0.451	0.238	0.576	0.729	0.404	0.752	0.870
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.058	0.095	0.015	0.048	0.100	0.020	0.050	0.083	0.010	0.055	0.095
$\delta = 0.1$	0.013	0.053	0.095	0.018	0.055	0.118	0.025	0.055	0.130	0.015	0.078	0.150
$\delta = 0.5$	0.093	0.286	0.401	0.198	0.439	0.574	0.358	0.617	0.742	0.471	0.712	0.842
$\delta = 1$	0.504	0.764	0.842	0.779	0.907	0.950	0.922	0.967	0.987	0.957	0.987	0.992
x_3 individual significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.050	0.073	0.013	0.050	0.108	0.018	0.038	0.103	0.005	0.053	0.105
$\delta = 0.1$	0.018	0.063	0.115	0.015	0.068	0.135	0.023	0.073	0.148	0.023	0.078	0.150
$\delta = 0.5$	0.123	0.331	0.444	0.236	0.474	0.579	0.353	0.609	0.727	0.474	0.724	0.830
$\delta = 1$	0.759	0.915	0.947	0.927	0.990	0.995	0.985	1.000	1.000	1.000	1.000	1.000
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.040	0.115	0.005	0.050	0.110	0.003	0.038	0.095	0.010	0.045	0.100
$\delta = 0.1$	0.008	0.063	0.130	0.018	0.075	0.130	0.015	0.085	0.153	0.020	0.110	0.175
$\delta = 0.5$	0.363	0.627	0.742	0.576	0.810	0.895	0.762	0.942	0.970	0.885	0.970	0.985
$\delta = 1$	0.995	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 5. (*Continued*).

(b) LSCV Bandwidth Results using 80% of the Upper Bound

	$n = 100$			$n = 200$		
	x_2	x_3	Joint	x_2	x_3	Joint
$\delta = 0$	0.697	0.551	0.411	0.772	0.602	0.501
$\delta = 0.1$	0.684	0.506	0.378	0.707	0.521	0.398
$\delta = 0.5$	0.363	0.030	0.010	0.211	0.000	0.000
$\delta = 1$	0.028	0.000	0.000	0.000	0.000	0.000

For example, in Table 6 we see that 69.7% of the time x_2 is appropriately smoothed away when $n = 100$ but this number increases to 77.2% of the time when we use samples of 200. As expected for models further away from the null, $\delta = 0.5$ and 1, as n increases the probability that a variable, or set of variables, is smoothed away is decreasing. We note that for all of our DGPs that when $\delta = 0.1$ this model is extremely close to the null and is hard to detect why the bandwidths suggest that a large portion of the time the variable is smoothed away erroneously. Interestingly, our test results seem to do a remarkable job of detecting even small departures from the null hypothesis when the bandwidths do not, providing even more evidence that one should formally test for insignificance.

Overall, we see that using the bootstrap test of Gu et al. (2007) using only discrete variables in the null hypothesis results in remarkable size and power properties, whereas raw interpretation of the bandwidths suggests that when the null is false our joint bandwidth measure does a good job of not smoothing out all variables simultaneously. However, when we examine our measure when the null is true we see that indeed, as the sample size increases the performance of this baseline measure is improving, it does not mimic the desirable behavior of the formal test. Again, the results in Racine and Li (2004) suggest that inclusion of discrete variables does not add to the curse of dimensionality so it is natural that the test results are better than in the continuous setting where all variables contributed to the dimensionality of the model.

3.3. Mixed Discrete–Continuous Case

In this setting, we try to mimic traditional applied milieus where there are a variety of covariates which are of mixed type. More importantly, we are

Table 6. DGP_2, Where x_2 and x_3 are Discrete Variables.

(a) Gu et al. (2007) Bandwidths

	$c = 0.25$			$c = 0.5$			$c = 1$			$c = 2$		
x_2 and x_3 joint significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.015	0.053	0.105	0.023	0.063	0.130	0.015	0.080	0.135	0.020	0.075	0.125
$\delta = 0.1$	0.015	0.063	0.110	0.020	0.073	0.150	0.015	0.100	0.163	0.023	0.090	0.145
$\delta = 0.5$	0.188	0.409	0.514	0.308	0.566	0.694	0.436	0.699	0.789	0.414	0.722	0.812
$\delta = 1$	0.759	0.910	0.932	0.887	0.967	0.990	0.962	0.997	1.000	0.962	0.997	0.997
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.003	0.050	0.083	0.015	0.045	0.073	0.008	0.055	0.088	0.003	0.050	0.103
$\delta = 0.1$	0.010	0.045	0.100	0.018	0.063	0.118	0.018	0.075	0.138	0.010	0.090	0.153
$\delta = 0.5$	0.509	0.707	0.799	0.692	0.845	0.915	0.822	0.952	0.980	0.865	0.980	0.992
$\delta = 1$	0.997	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
x_2 individual significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.040	0.078	0.020	0.050	0.090	0.005	0.045	0.123	0.005	0.033	0.110
$\delta = 0.1$	0.013	0.040	0.083	0.018	0.053	0.088	0.005	0.043	0.120	0.008	0.040	0.113
$\delta = 0.5$	0.018	0.075	0.118	0.023	0.100	0.158	0.030	0.098	0.185	0.023	0.108	0.195
$\delta = 1$	0.073	0.168	0.298	0.123	0.281	0.431	0.145	0.371	0.531	0.133	0.401	0.559
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.005	0.053	0.095	0.005	0.050	0.115	0.000	0.043	0.113	0.005	0.030	0.073
$\delta = 0.1$	0.010	0.060	0.085	0.013	0.050	0.090	0.015	0.048	0.088	0.010	0.045	0.105
$\delta = 0.5$	0.020	0.073	0.135	0.028	0.105	0.178	0.025	0.128	0.218	0.018	0.095	0.203
$\delta = 1$	0.075	0.223	0.393	0.160	0.381	0.519	0.233	0.494	0.637	0.216	0.534	0.687
x_3 individual significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.058	0.108	0.010	0.055	0.113	0.005	0.055	0.108	0.013	0.068	0.110
$\delta = 0.1$	0.018	0.063	0.133	0.015	0.075	0.150	0.015	0.078	0.160	0.018	0.088	0.165
$\delta = 0.5$	0.281	0.506	0.622	0.454	0.689	0.772	0.596	0.812	0.890	0.544	0.835	0.917
$\delta = 1$	0.865	0.960	0.980	0.962	0.992	0.995	0.990	0.997	1.000	0.977	0.997	1.000
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.040	0.115	0.005	0.050	0.110	0.003	0.038	0.095	0.010	0.045	0.100
$\delta = 0.1$	0.020	0.070	0.148	0.013	0.080	0.165	0.018	0.083	0.185	0.020	0.088	0.193
$\delta = 0.5$	0.629	0.810	0.890	0.817	0.955	0.980	0.937	0.987	0.992	0.957	0.992	0.997
$\delta = 1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 6. (*Continued*).

(b) LSCV Bandwidth Results using 80% of the Upper Bound

	$n = 100$			$n = 200$		
	x_2	x_3	Joint	x_2	x_3	Joint
$\delta = 0$	0.697	0.551	0.411	0.772	0.602	0.501
$\delta = 0.1$	0.699	0.409	0.311	0.762	0.298	0.233
$\delta = 0.5$	0.599	0.000	0.000	0.546	0.000	0.000
$\delta = 1$	0.378	0.000	0.000	0.103	0.000	0.000

interested in a mixed hypothesis which the current menu of available tests does not formally allow for. Again, as mentioned earlier, theoretical backing aside, there is no reason the test of Lavergne and Vuong (2000) and Gu et al. (2007) cannot include discrete variables. We present several testing scenarios, including bandwidth rules, for DGP_4, in Table 8.

Our joint significance test under the appropriate null, $H_0 : x_3, x_5, x_6$ are insignificant, reveals that the test appears to be oversized across all levels of the bandwidth. The results for $c = 0.25$, however, seem to display uniformly better size at our conventional testing levels than our other scaling setups. Here, we posit that the size of the test suffers due to the inclusion of an additional, relevant covariate. This adds to the curse of dimensionality and having a sample size of $n = 100$ is not enough to overcome the additional covariate. However, we see that doubling of our sample size to $n = 200$ dramatically improves the performance of the test and that the size of the test is almost exact in this finite sample setting. This suggests that the nonparametric test of omitted variables can be used to test significance of mixed joint hypothesis in practice.

Switching to the performance of the LSCV bandwidths, we note that, as before, using IQR results in a higher proportion of the simulations with the appropriate continuous variables smoothed out, but with this specific DGP we do not notice the erroneous smoothing out that occurred in our previous simulations. We note that our DGP in the mixed setting results in x_5 having a hard time being determined to be relevant even when it is true. This is because our model is close to the null even when $\delta = 0.1$, 0.5, or 1. What is striking is that our joint measure of determination is worse than in our other setups because our null hypothesis involves three covariates as opposed to two. This highlights the difficulty of assessing irrelevance in a joint fashion based on the LSCV bandwidths. Note that in only 34% of our simulations

Table 7. DGP_3, Where x_2 and x_3 are Discrete Variables.

(a) Gu et al. (2007) Bandwidths

	$c = 0.25$			$c = 0.5$			$c = 1$			$c = 2$		
x_2 and x_3 joint significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.015	0.053	0.105	0.023	0.063	0.130	0.015	0.080	0.135	0.020	0.075	0.125
$\delta = 0.1$	0.018	0.060	0.120	0.028	0.080	0.145	0.018	0.095	0.158	0.038	0.108	0.185
$\delta = 0.5$	0.150	0.333	0.449	0.258	0.499	0.609	0.409	0.659	0.797	0.544	0.805	0.872
$\delta = 1$	0.832	0.935	0.965	0.955	0.992	0.997	0.987	1.000	1.000	0.995	1.000	1.000
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.003	0.050	0.083	0.015	0.045	0.073	0.008	0.055	0.088	0.003	0.050	0.103
$\delta = 0.1$	0.008	0.060	0.093	0.018	0.053	0.110	0.018	0.070	0.128	0.020	0.088	0.158
$\delta = 0.5$	0.366	0.617	0.742	0.586	0.837	0.917	0.815	0.947	0.977	0.920	0.982	0.992
$\delta = 1$	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
x_2 individual significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.055	0.090	0.005	0.038	0.108	0.005	0.048	0.108	0.015	0.050	0.108
$\delta = 0.1$	0.010	0.048	0.088	0.005	0.040	0.120	0.010	0.055	0.123	0.020	0.080	0.130
$\delta = 0.5$	0.030	0.133	0.236	0.078	0.211	0.311	0.113	0.328	0.434	0.175	0.409	0.536
$\delta = 1$	0.190	0.381	0.484	0.318	0.544	0.697	0.496	0.724	0.815	0.589	0.807	0.887
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.058	0.095	0.015	0.048	0.100	0.020	0.050	0.083	0.010	0.055	0.095
$\delta = 0.1$	0.010	0.055	0.098	0.015	0.060	0.103	0.020	0.063	0.123	0.015	0.075	0.145
$\delta = 0.5$	0.100	0.246	0.378	0.165	0.391	0.546	0.318	0.574	0.707	0.409	0.687	0.789
$\delta = 1$	0.516	0.742	0.837	0.752	0.902	0.940	0.880	0.952	0.972	0.917	0.970	0.982
x_3 individual significance test												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.058	0.108	0.010	0.055	0.113	0.005	0.055	0.108	0.013	0.068	0.110
$\delta = 0.1$	0.015	0.068	0.120	0.013	0.073	0.123	0.015	0.083	0.155	0.023	0.088	0.158
$\delta = 0.5$	0.113	0.293	0.421	0.213	0.439	0.594	0.311	0.602	0.702	0.464	0.692	0.787
$\delta = 1$	0.609	0.817	0.887	0.787	0.927	0.965	0.920	0.982	0.997	0.972	0.997	0.997
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.008	0.040	0.115	0.005	0.050	0.110	0.003	0.038	0.095	0.010	0.045	0.100
$\delta = 0.1$	0.015	0.060	0.120	0.008	0.075	0.125	0.020	0.083	0.143	0.015	0.100	0.165
$\delta = 0.5$	0.291	0.506	0.639	0.456	0.702	0.789	0.639	0.832	0.920	0.764	0.932	0.975
$\delta = 1$	0.955	0.992	1.000	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 7. (*Continued*).

(b) LSCV Bandwidth Results using 80% of the Upper Bound

	$n = 100$			$n = 200$		
	x_2	x_3	Joint	x_2	x_3	Joint
$\delta = 0$	0.697	0.551	0.411	0.772	0.602	0.501
$\delta = 0.1$	0.682	0.476	0.356	0.719	0.454	0.353
$\delta = 0.5$	0.393	0.023	0.003	0.251	0.000	0.000
$\delta = 1$	0.050	0.000	0.000	0.000	0.000	0.000

were x_3, x_5, and x_6 smoothed away simultaneously according to our standard deviation determination rule. For $\delta = 0.5$, 1, the data-driven bandwidths never jointly remove the three variables under investigation. We also note that x_1 and x_4 are never smoothed out in any of these simulations.

These results, while limited in scope, provide two key insights for applied econometricians. First, the standard, continuous-only nonparametric omitted variable test can be modified to handle a joint hypothesis involving mixed data. Second, data-driven bandwidths can be used as an effective screen for removing irrelevant variables in a local constant setting, but they do not preclude the use of a formal statistical test.

4. CONCLUSION

This research has focused on two broad aspects of assessing variable irrelevance in multivariate nonparametric kernel regression in the presence of mixed data types. First, we discussed the lack of a theoretically consistent test that allows joint hypothesis testing involving both continuous and categorical data. We then discussed a currently existing test of significance, which can include both types of data simultaneously, and its performance when either discrete or mixed data enter into the null hypothesis. Second, we investigated the performance of several suggested ad hoc means of using LSCV bandwidths to determine variable irrelevance prior to testing.

Our results revealed that implementing the test of Gu et al. (2007) using mixed data types did not harm its performance with respect to size or power. Additionally, we provided evidence that while using cross-validated bandwidths on an individual basis resulted in good detection of variable

Table 8. DGP_4, Where x_4, x_5, and x_6 are Discrete Variables.

(a) Gu et al. (2007) Bandwidths

	$c = 0.25$			$c = 0.5$			$c = 1$			$c = 2$		
Joint significance test H_0: x_3, x_5, and x_6 are insignificant												
$n = 100$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.065	0.133	0.028	0.095	0.158	0.025	0.115	0.188	0.025	0.108	0.201
$\delta = 0.1$	0.018	0.090	0.165	0.048	0.173	0.263	0.173	0.348	0.469	0.356	0.619	0.729
$\delta = 0.5$	0.173	0.494	0.704	0.885	0.980	0.987	1.000	1.000	1.000	1.000	1.000	1.000
$\delta = 1$	0.223	0.609	0.832	0.970	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$												
α	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\delta = 0$	0.013	0.055	0.102	0.010	0.055	0.090	0.010	0.049	0.096	0.010	0.047	0.101
$\delta = 0.1$	0.035	0.153	0.236	0.190	0.386	0.509	0.637	0.825	0.907	0.900	0.982	0.995
$\delta = 0.5$	0.464	0.820	0.930	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\delta = 1$	0.647	0.937	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

(b) LSCV Bandwidth Results

Variable	Continuous (2 SD)			Discrete (0.8)			Joint (2 SD, 0.8)	Continuous (IQR)			Joint (IQR, 0.8)
	x_1	x_2	x_3	x_4	x_5	x_6	Joint	x_1	x_2	x_3	Joint
$n = 100$											
$\delta = 0$	0.000	0.000	0.647	0.000	0.734	0.694	0.341	0.000	0.000	0.774	0.401
$\delta = 0.1$	0.000	0.000	0.632	0.000	0.674	0.311	0.093	0.000	0.000	0.767	0.108
$\delta = 0.5$	0.000	0.000	0.183	0.000	0.569	0.000	0.000	0.000	0.003	0.308	0.000
$\delta = 1$	0.000	0.000	0.028	0.000	0.471	0.000	0.000	0.000	0.000	0.053	0.000
$n = 200$											
$\delta = 0$	0.000	0.000	0.685	0.004	0.765	0.779	0.420	0.000	0.000	0.882	0.461
$\delta = 0.1$	0.000	0.000	0.621	0.000	0.690	0.309	0.102	0.000	0.000	0.860	0.138
$\delta = 0.5$	0.000	0.000	0.111	0.000	0.230	0.000	0.000	0.000	0.000	0.250	0.000
$\delta = 1$	0.000	0.000	0.002	0.000	0.176	0.000	0.000	0.000	0.000	0.038	0.000

irrelevance, the same measures applied jointly are not as successful at uncovering irrelevance. This suggests that in the presence of multiple irrelevant regressors formal testing should always be used as a backdrop for determining if a set of variables should be included in one's final nonparametric model. One should use economic theory to guide them toward the appropriate set of covariates to test for joint significance.

Further research should focus on the construction of and distribution theory for a test to formally handle mixed data types in null hypotheses, preferably a test that only involves estimation of the unrestricted model. Additionally, simulation results comparing test performance across local constant and local linear methodologies would be insightful as the cross-validated bandwidths obtained when one uses local linear (or any other order polynomial) are not directly related to variable relevance for continuous regressors. Also, the use of bandwidths obtained through other cross-validation methods, such as improved AIC_c, would prove useful since LSCV is known to produce bandwidths that lead to undersmoothing in finite samples.

NOTES

1. It is hypothesized that for local polynomial estimation with polynomial degree p, as the bandwidth diverges, the associated variable enters the model in a polynomial of order p fashion.
2. Their power is influenced directly via the bandwidth used to perform the test (Gu et al., 2007, Table 6).
3. See also Li and Racine (2006, p. 373) for a related discussion.
4. Their bootstrap theory only pertains to continuous variables, however.
5. One could also use the Epanechnikov or biweight kernel as well.
6. This is not entirely damning as it was shown in finite samples that the LSCV bandwidths continued to smooth away irrelevant variables when dependence was allowed between relevant and irrelevant regressors. The assumption was made for ease of proof of the corresponding theorems in the paper.
7. See Henderson, Papageorgiou, and Parmeter (2008) for additional simulation results with a large number of irrelevant variables.
8. We still have x_1 continuous in these settings.
9. This was due to the fact that LSCV was providing scale factors on the order of 100 or 1000 as opposed to 0.25 or 2 for the irrelevant variables.

ACKNOWLEDGMENTS

We would like to acknowledge the insightful comments we received during attendance at the 7th Annual Advances in Econometrics Conference (November 2008 at Louisiana State University) by conference participants. The comments of two anonymous referees and thoughtful advice from Daniel Henderson and Jeffrey Racine are warmly appreciated. We alone are responsible for errors and omissions.

REFERENCES

Aitchison, J., & Aitken, C. B. B. (1976). Multivariate binary discrimination by kernel method. *Biometrika*, *63*, 413–420.

Cai, Z., Gu, J., & Li, Q. (2009). Some recent developments on nonparametric econometrics. In: Q. Li & J. S. Racine (Eds), *Advances in econometrics: Nonparametric methods* (Vol. 25, this volume). Bingley, UK: Emerald.

Delgado, M. A., & González-Manteiga, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *The Annals of Statistics*, *29*(5), 1469–1507.

Gu, J., Li, D., & Liu, D. (2007). A bootstrap nonparametric significance test. *Journal of Nonparametric Statistics*, *19*(6–8), 215–230.

Hall, P., Li, Q., & Racine, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics*, *89*, 784–789.

Henderson, D. J., Papageorgiou, C., & Parmeter, C. F. (2008). Are any growth theories linear? Why we should care about what the evidence tells us. Munich Personal RePEc Archive Paper no. 8767.

Hsiao, C., Li, Q., & Racine, J. S. (2007). A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics*, *140*, 802–826.

Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B*, *60*, 271–293.

Lavergne, P. (2001). An equality test across nonparametric regressions. *Journal of Econometrics*, *103*, 307–344.

Lavergne, P., & Vuong, Q. (2000). Nonparametric significance testing. *Econometric Theory*, *16*(4), 576–601.

Li, Q., & Racine, J. S. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, *14*, 485–512.

Li, Q., & Racine, J. S. (2006). *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.

Nadaraya, E. A. (1964). On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability*, *10*, 186–190.

Pagan, A., & Ullah, A. (1999). *Nonparametric Econometrics*. New York: Cambridge University Press.

Racine, J. S. (1997). Consistent significance testing for nonparametric regression. *Journal of Business and Economic Statistics*, *15*(3), 369–379.

Racine, J. S., Hart, J., & Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, *25*(4), 523–544.

Racine, J. S., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130.

Wang, M. C., & Ryzin, J. V. (1981). A class of smooth estimators for discrete estimation. *Biometrika*, *68*, 301–309.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, *26*(15), 359–372.

PART II
ESTIMATION OF SEMIPARAMETRIC MODELS

SEMIPARAMETRIC ESTIMATION OF FIXED-EFFECTS PANEL DATA VARYING COEFFICIENT MODELS

Yiguo Sun, Raymond J. Carroll and Dingding Li

ABSTRACT

We consider the problem of estimating a varying coefficient panel data model with fixed-effects (FE) using a local linear regression approach. Unlike first-differenced estimator, our proposed estimator removes FE using kernel-based weights. This results a one-step estimator without using the backfitting technique. The computed estimator is shown to be asymptotically normally distributed. A modified least-squared cross-validatory method is used to select the optimal bandwidth automatically. Moreover, we propose a test statistic for testing the null hypothesis of a random-effects varying coefficient panel data model against an FE one. Monte Carlo simulations show that our proposed estimator and test statistic have satisfactory finite sample performance.

1. INTRODUCTION

Panel data traces information on each individual unit across time. Such a two-dimensional information set enables researchers to estimate complex

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 101–129

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025006

models and extract information and inferences, which may not be possible using pure time-series data or cross-section data. With the increased availability of panel data, both theoretical and applied work in panel data analysis have become more popular in the recent years.

Arellano (2003), Baltagi (2005) and Hsiao (2003) provide excellent overview of parametric panel data model analysis. However, it is well known that a misspecified parametric panel data model may give misleading inferences. To avoid imposing the strong restrictions assumed in the parametric panel data models, econometricians and statisticians have worked on theories of nonparametric and semiparametric panel data regression models. For example, Henderson, Carroll, and Li (2008) considered the fixed-effects (FE) nonparametric panel data model. Henderson and Ullah (2005), Lin and Carroll (2000, 2001, 2006), Lin, Wang, Welsh, and Carroll (2004), Lin and Ying (2001), Ruckstuhl, Welsh, and Carroll (2000), Wang (2003), and Wu and Zhang (2002) considered the random-effects (RE) nonparametric panel data models. Li and Stengos (1996) considered a partially linear panel data model with some regressors being endogenous via instrumental variable (IV) approach, and Su and Ullah (2006) investigated an FE partially linear panel data model with exogenous regressors.

A purely nonparametric model suffers from the 'curse of dimensionality' problem, while a partially linear semiparametric model may be too restrictive as it only allows for some additive nonlinearities. The varying coefficient model considered in this paper includes both pure nonparametric model and partially linear regression model as special cases. Moreover, we assume an FE panel data model. By FE we mean that the individual effects are correlated with the regressors in an unknown way. Consistent with the well-known results in parametric panel data model estimation, we show that RE estimators are inconsistent if the true model is one with FE, and that FE estimators are consistent under both RE- and FE panel data model, although the RE estimator is more efficient than the FE estimator when the RE model holds true. Therefore, estimation of RE models is appropriate only when individual effects are uncorrelated with regressors. As, in practice, economists often view the assumptions required for the RE model as being unsupported by the data, this paper emphasizes more on estimating an FE panel data varying coefficient model, and we propose to use the local linear method to estimate unknown smooth coefficient functions. We also propose a test statistic for testing an RE varying coefficient panel data model against an FE one. Simulation results show that our proposed estimator and test statistic have satisfactory finite sample performances.

Recently, Cai, and Li (2008) studied a dynamic nonparametric panel data model with unknown varying coefficients. As Cai and Li (2008) allow the regressors not appearing in the varying coefficient curves to be endogenous, the generalized method of moments-based IV estimation method plus local linear regression approach is used to deliver consistent estimator of the unknown smooth coefficient curves. In this paper, all the regressors are assumed to be exogenous. Therefore, the least-squared method combining with local linear regression approach can be used to produce consistent estimator of the unknown smoothing coefficient curves. In addition, the asymptotic results are given when the time length is finite.

The rest of the paper is organized as follows. In Section 2 we set up the model and discuss transformation methods that are used to remove FE. Section 3 proposes a nonparametric FE estimator and studies its asymptotic properties. In Section 4 we suggest a statistic for testing the null hypothesis of an RE varying coefficient model against an FE one. Section 5 reports simulation results that examine the finite sample performance of our semiparametric estimator and the test statistic. Finally we conclude the paper in Section 6. The proofs of the main results are collected in the appendix.

2. FIXED-EFFECTS VARYING COEFFICIENT PANEL DATA MODELS

We consider the following FE varying coefficient panel data regression model

$$Y_{it} = X_{it}^{\top}\theta(Z_{it}) + \mu_i + v_{it} \qquad i = 1, \ldots, n;\ t = 1, \ldots, m \tag{1}$$

where the covariate $Z_{it} = (Z_{it,1}, \ldots, Z_{it,q})^{\top}$ is of dimension q, $X_{it} = (X_{it,1}, \ldots, X_{it,p})^{\top}$ is of dimension p, $\theta(\cdot) = \{\theta_1(\cdot), \ldots, \theta_p(\cdot)\}^{\top}$ contains p unknown functions; and all other variables are scalars. None of the variables in X_{it} can be obtained from Z_{it} and vice versa. The random errors v_{it} are assumed to be independently and identically distributed (i.i.d.) with a zero mean, finite variance $\sigma_v^2 > 0$ and independent of μ_j, Z_{js} and X_{js} for all i, j, s and t. The unobserved individual effects μ_i are assumed to be i.i.d. with a zero mean and a finite variance $\sigma_\mu^2 > 0$. We allow for μ_i to be correlated with Z_{it} and/or X_{it} with an unknown correlation structure. Hence, model (1) is an FE model. Alternatively, when μ_i is uncorrelated with Z_{it} and X_{it}, model (1) becomes an RE model.

A somewhat simplistic explanation for consideration of FE models and the need for estimation of the function $\theta(\cdot)$ arises from considerations such as the following. Suppose that Y_{it} is the (logarithm) income of individual i at time period t; X_{it} is education of individual i at time period t, for example, number of years of schooling; and Z_{it} is the age of individual i at time t. The FE term μ_i in Eq. (1) includes the individual's unobservable characteristics such as ability (e.g. IQ level) and characteristics which are not observable for the data at hand. In this problem, economists are interested in the marginal effects of education on income, after controlling for the unobservable individual ability factors. Hence, they are interested in the marginal effects in the income change for an additional year of education regardless of whether the person has high or low ability. In this simple example, it is reasonable to believe that ability and education are positively correlated. If one does not control for the unobserved individual effects, then one would overestimate the true marginal effects of education on income (i.e. with an upward bias).

When $X_{it} \equiv 1$ for all i and t and $p = 1$, model (1) reduces to Henderson et al. (2008) nonparametric panel data model with FE as a special case. One may also interpret $X_{it}^{\top}\theta(Z_{it})$ as an interactive term between X_{it} and Z_{it}, where we allow $\theta(Z_{it})$ to have a flexible format since the popularly used parametric set-up such as Z_{it} and/or Z_{it}^2 may be misspecified.

For a given FE model, there are many ways of removing the unknown fixed effects from the model.

The usual first-differenced (FD) estimation method deducts one equation from another to remove the time-invariant FE. For example, deducting equation for time t from that for time $t-1$, we have for $t = 2, \ldots, m$

$$\tilde{Y}_{it} = Y_{it} - Y_{i,t-1} = X_{it}^{\top}\theta(Z_{it}) - X_{it-1}^{\top}\theta(Z_{it-1}) + \tilde{v}_{it} \quad \text{with } \tilde{v}_{it} = v_{it} - v_{i,t-1} \tag{2}$$

or deducting equation for time t from that for time 1, we obtain for $t = 2, \ldots, m$

$$\tilde{Y}_{it} = Y_{it} - Y_{i1} = X_{it}^{\top}\theta(Z_{it}) - X_{i1}^{\top}\theta(Z_{i1}) + \tilde{v}_{it} \quad \text{with } \tilde{v}_{it} = v_{it} - v_{i1} \tag{3}$$

The conventional FE estimation method, on the other hand, removes the FE by deducting each equation from the cross-time average of the system,

and it gives for $t = 2, \ldots, m$

$$\tilde{Y}_{it} = Y_{it} - \frac{1}{m}\sum_{s=1}^{m} Y_{is} = X_{it}^{\top}\theta(Z_{it}) - \frac{1}{m}\sum_{s=1}^{m} X_{is}^{\top}\theta(Z_{is}) + \tilde{v}_{it}$$
$$= \sum_{s=1}^{m} q_{ts} X_{is}^{\top}\theta(Z_{is}) + \tilde{v}_{it} \quad \text{with } \tilde{v}_{it} = v_{it} - \frac{1}{m}\sum_{s=1}^{m} v_{is} \tag{4}$$

where $q_{ts} = -1/m$ if $s \neq t$ and $1 - 1/m$ otherwise, and $\sum_{s=1}^{m} q_{ts} = 0$ for all t.

Many nonparametric local smoothing methods can be used to estimate the unknown function $\theta(\cdot)$. However, for each i, the right-hand sides of Eqs. (2)–(4) contain linear combination of $X_{it}^{\top}\theta(Z_{it})$ for different time t. If X_{it} contains a time-invariant term, say the first component of X_{it}, and let $\theta_1(Z_{it})$ denote the first component of $\theta(Z_{it})$, then a first difference of $X_{it,1}\theta_1(Z_{it}) \equiv X_{i,1}\theta_1(Z_{it})$ gives $X_{i,1}(\theta_1(Z_{it}) - \theta_1(Z_{i,t-1}))$, which is an additive function with the same function form for the two functions but evaluated at different observation points. Kernel-based estimator usually requires some back-fitting algorithms to recover the unknown function, which will suffer the common problems as indicated in estimating nonparametric additive model. Moreover, if $\theta_1(Z_{it})$ contains an additive constant term, say $\theta(Z_{it}) = c + g_1(Z_{it})$, where c is a constant, then the first difference will wipe out the additive constant c. As a consequence, one cannot consistently estimate $\theta_1(\cdot)$ if one were to estimate an FD model in general (if $X_{i,1} \equiv 1$, one can recover c by averaging $Y_{it} - X_{it}^{\top}\hat{\theta}(Z_{it})$ for all cross-sections and across time).

Therefore, in this paper we consider an alternative way of removing the unknown FE, motivated by a least-squares dummy variable (LSDV) model in parametric panel data analysis. We will describe how the proposed method removes FE by deducting a smoothed version of cross-time average from each individual unit. As we will show later, this transformation method will not wipe the additive constant c in $\theta_1(Z_{it}) = c + g_1(Z_{it})$. Therefore, we can consistently estimate $\theta_1(\cdot)$ as well as other components of $\theta(\cdot)$ when at most one of the variables in X_{it} is time invariant.

We will use I_n to denote an identity matrix of dimension n, and e_m to denote an $m \times 1$ vector with all elements being 1s. Rewriting model (1) in a matrix format yields

$$Y = B\{X, \theta(Z)\} + D_0\mu_0 + V \tag{5}$$

where $Y = (Y_1^{\top}, \ldots, Y_n^{\top})^{\top}$ and $V = (v_1^{\top}, \ldots, v_n^{\top})^{\top}$ are $(nm) \times 1$ vectors; $Y_i^{\top} = (Y_{i1}, \ldots, Y_{in})$ and $v^{\top} = (v_{i1}, \ldots, v_{in})$. $B\{X, \theta(Z)\}$ stacks all $X_{it}^{\top}\theta(Z_{it})$ into an $(nm) \times 1$ vector with the (i, t) subscript matching that of the $(nm) \times 1$

vector of Y; $\mu_0 = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is an $n \times 1$ vector; and $D_0 = I_n \otimes e_m$ is an $(nm) \times n$ matrix with main diagonal blocks being e_m, where $\otimes$ refers to Kronecker product operation. However, we cannot estimate model (5) directly due to the existence of the FE term. Therefore, we need some identification conditions. Su and Ullah (2006) assume $\sum_{i=1}^{n} \mu_i = 0$. We show that assuming an i.i.d sequence of unknown FE μ_i with zero mean and a finite variance is enough to identify the unknown coefficient curves asymptotically. We therefore impose this weaker version of identification condition in this paper.

To introduce our estimator, we first assume that model (1) holds with the restriction $\sum_{i=1}^{n} \mu_i = 0$ (note that we do not impose this restriction for our estimator, and this restriction is added here for motivating our estimator). Define $\mu = (\mu_2, \ldots, \mu_n)^{\top}$. We then rewrite Eq. (5) as

$$Y = B\{X, \theta(Z)\} + D\mu + V \tag{6}$$

where $D = [-e_{n-1}\ I_{n-1}]^{\top} \otimes e_m$ is an $(nm) \times (n-1)$ matrix. Note that $D\mu = \mu_0 \otimes e_m$ with $\mu_0 = (-\sum_{i=2}^{n} \mu_i, \mu_2, \ldots, \mu_n)^{\top}$ so that the restriction $\sum_{i=1}^{n} \mu_i = 0$ is imposed in Eq. (6).

Define an $m \times m$ diagonal matrix $K_H(Z_i, z) = \mathrm{diag}\{K_H(Z_{i1}, z), \ldots, K_H(Z_{im}, z)\}$ for each i, and a $(nm) \times (nm)$ diagonal matrix $W_H(z) = \mathrm{diag}\{K_H(Z_1, z), \ldots, K_H(Z_n, z)\}$, where $K_H(Z_{it}, z) = K\{H^{-1}(Z_{it} - z)\}$ for all i and t, and $H = \mathrm{diag}(h_1, \ldots, h_q)$ is a $q \times q$ diagonal bandwidth matrix. We then solve the following optimization problem:

$$\min_{\theta(Z), \mu} [Y - B\{X, \theta(Z)\} - D\mu]^T W_H(z) [Y - B\{X, \theta(z)\} - D\mu] \tag{7}$$

where we use the local weight matrix $W_H(z)$ to ensure locality of our nonparametric fitting, and place no weight matrix for data variation since the $\{v_{it}\}$ are i.i.d. across equations. Taking first-order condition with respect to μ gives

$$D^{\top} W_H(z) [Y - B\{X, \theta(Z)\} - D\hat{\mu}(z)] = 0 \tag{8}$$

which yields

$$\hat{\mu}(z) = \{D^{\top} - W_H(z) D\}^{-1} D^{\top} W_H(z) [Y - B\{X, \theta(Z)\}] \tag{9}$$

Define $S_H(z) = M_H(z)^{\top} W_H(z) M_H(z)$ and $M_H(z) = I_{n \times m} - D\{D^{\top} W_H(z) D\}^{-1} D^{\top} W_H(z)$, where $I_{n \times m}$ denotes an identity matrix of dimension $(nm) \times (nm)$. Replacing μ in Eq. (7) by $\hat{\mu}(z)$, we obtain the concentrated weighted least squares

$$\min_{\theta(Z)} [Y - B\{X, \theta(Z)\}]^{\top} S_H(Z) [Y - B\{X, \theta(Z)\}] \tag{10}$$

Note that $M_H(z)D\mu \equiv 0_{(nm)\times 1}$ for all z. Hence, the FE term μ is removed in model (10).

To see how $M_H(z)$ transforms the data, simple calculations give

$$M_H(z) = I_{n\times m} - D\left\{\frac{A^{-1} - A^{-1}e_{n-1}e_{n-1}^{\top}A^{-1}}{\sum_{i=1}^{n} c_H(Z_i, z)}\right\}D^{\top}W_H(z)$$

where $c_H(Z_i, z)^{-1} = \sum_{l=1}^{m} K_H(Z_{it}, z)$ for $i = 1, \ldots, n$ and $A = \text{diag}\{c_H(Z_2, z)^{-1}, \ldots, c_H(Z_n, z)^{-1}\}$. We use the formula $(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}$ to derive the inverse matrix, see Appendix B in Poirier (1995).

3. NONPARAMETRIC ESTIMATOR AND ASYMPTOTIC THEORY

A local linear regression approach is commonly used to estimate non-/semiparametric models. The basic idea of this method is to apply Taylor expansion up to the second-order derivative. Throughout the paper we will use the notation $A_n \approx B_n$ to denote that B_n is the leading term of A_n, that is $A_n = B_n + (s.o.)$, where $(s.o.)$ denotes terms having probability order smaller than that of B_n. For each $l = 1, \ldots, p$, we have the following Taylor expansion around z:

$$\theta_l(z_{it}) \approx \theta_l(z) + \left\{H\theta_l'(z)\right\}^{\top}[H^{-1}(z_{it} - z)] + \frac{1}{2}r_{H,l}(z_{it}, z) \tag{11}$$

where $\theta_l'(z) = \partial\theta_l(z)/\partial z$ is the $q \times 1$ vector of the first-order derived function, and $r_{H,l}(z_{it}, z) = \{H^{-1}(z_{it} - z)\}^{\top}\{H((\partial^2\theta_l(z))/(\partial z\partial z^{\top}))H\}\{H^{-1}(z_{it} - z)\}$. Of course, $\theta_l(z)$ approximates $\theta_l(z_{it})$ and $\theta_l'(z)$ approximates $\theta_l'(z_{it})$ when z_{it} is close to z. Define $\beta_l(z) = \{\theta_l(z), [H\theta_l'(z)]^{\top}\}^{\top}$, a $(q+1)\times 1$ column vector for $l = 1, 2, \ldots, p$, and $\beta(z) = \{\beta_1(z), \ldots, \beta_p(z)\}^{\top}$, a $p \times (q+1)$parameter matrix. The first column of $\beta(z)$ is $\theta(z)$. Therefore, we will replace $\theta(Z_{it})$ in Eq. (1) by $\beta(z)G_{it}(z, H)$ for each i and t, where $G_{it}(z, H) = [1, \{H^{-1}(Z_{it} - z)\}^{\top}]^{\top}$ is a $(q+1)\times 1$ vector.

To make matrix operations simpler, we stack the matrix $\beta(z)$ into a $p(q+1)\times 1$ column vector and denote it by $\text{vec}\{\beta(z)\}$. Since $\text{vec}(ABC) = (C^{\top}\otimes A)\text{vec}(B)$ and $(A\otimes B)^{\top} = A^{\top}\otimes B^{\top}$, where $\otimes$ refers to Kronecker product, we have $X_{it}^{\top}\beta(z)G_{it}(z, H) = \{G_{it}(z, H)\otimes X_{it}\}^{\top}\text{vec}\{\beta(z)\}$ for all i and t. Thus, we consider the following minimization problem:

$$\min_{\beta(Z)}[Y - R(z, H)\text{vec}\{\beta(Z)\}]^{\top}S_H(z)[Y - R(z, H)\text{vec}\{\beta(z)\}] \tag{12}$$

where

$$R_i(z,H)=\begin{bmatrix}(G_{i,1}(z,H)\otimes X_{i1})^\top\\ \vdots\\ (G_{i,m}(z,H)\otimes X_{im})^\top\end{bmatrix} \text{ is an } m\times[p(q+1)] \text{ matrix, and}$$

$$R(z,H)=[R_1(z,H)^\top,\ldots,R_n(z,H)^\top]^\top \text{ is an } (nm)\times[p(q+1)] \text{ matrix}$$

Simple calculations give

$$\begin{aligned}\operatorname{vec}\{\hat{\beta}(z)\}&=\{R(z,H)^\top S_H(z)R(z,H)\}^{-1}R(z,H)^\top S_H(z)Y\\&=\operatorname{vec}\{\beta(z)\}+\{R(z,H)^\top S_H(z)R(z,H)\}^{-1}(A_n/2+B_n+C_n)\end{aligned}\quad(13)$$

where $A_n=R(z,H)^\top S_H(z)\Pi(z,H)$, $B_n=R(z,H)^\top S_H(z)D_0\mu_0$, and $C_n=R(z,H)^\top S_H(z)V$. The $\{t+(i-1)m\}$th element of the column vector $\Pi(z,H)$ is $X_{it}^\top r_H(\tilde{Z}_{it},z)$, where $r_H(\cdot,\cdot)=\{r_{H,1}(\cdot,\cdot),\ldots,r_{H,p}(\cdot,\cdot)\}^\top$ and $r_{H,l}(\tilde{Z}_{it},z)=\{H^{-1}(Z_{it}-z)\}^\top\{H((\partial^2\theta_l(\tilde{Z}_{it}))/(\partial z\partial z^\top))H\}\{H^{-1}(Z_{it}-z)\}$ with $\tilde{Z}_{it}$ lying between Z_{it} and z for each i and t. Both A_n and B_n contribute to the bias term of the estimator. Also, if $\sum_{i=1}^n\mu_i=0$ holds true, $B_n=0$; if we only assume μ_i being i.i.d. with zero mean and finite variance, the bias due to the existence of unknown FE can be asymptotically ignored.

To derive the asymptotic distribution of $\operatorname{vec}\{\hat{\beta}(z)\}$, we first give some regularity conditions. Throughout this paper, we use $M>0$ to denote a finite constant, which may take a different value at different places.

Assumption 1. The random variables X_{it} and Z_{it} are i.i.d. across the i index, and

(a) $E||X_{it}||^{2(1+\delta)}\le M<\infty$ and $E||Z_{it}||^{2(1+\delta)}\le M<\infty$ hold for some $\delta>0$ and for all i and t.
(b) The Z_{it} are continuous random variables with a probability density function (pdf) $f_t(z)$. Also, for each $z\in R^q$, $f(z)=\sum_{t=1}^m f_t(z)>0$.
(c) Denote $\lambda_{it}=K_H(Z_{it},z)$ and $\varpi_{it}=\lambda_{it}/\sum_{t=1}^m\lambda_{it}\in(0,1)$ for all i and t. $\Psi(z)=|H|^{-1}\sum_{t=1}^m E[(1-\varpi_{it})\lambda_{it}X_{it}X_{it}^T]$ is a nonsingular matrix.
(d) Let $f_t(z|X_{it})$ be the conditional pdf of Z_{it} at $Z_{it}=z$ conditional on X_{it} and $f_{t,s}(z_1,z_2|X_{it},X_{js})$ be the joint conditional pdf of (Z_{it},Z_{js}) at $(Z_{it},Z_{js})=(z_1,z_2)$ conditional on (X_{it},X_{js}) for $t\neq s$ and any i and j. Also, $\theta(z),f_t(z),f_t(\cdot|X_{it}),f_{t,s}(\cdot,\cdot|X_{it},X_{js})$ are uniformly bounded in the domain of Z, and are all twice continuously differentiable at $z\in R^q$ for all $t\neq s$, i and j.

Assumption 2. Both X and Z have full column rank; $\{X_{it,1}, \ldots, X_{it,p}, \{X_{it,l}Z_{it,j} : l = 1, \ldots, p, j = 1, \ldots, q\}\}$ are linearly independent. If $X_{it,l} \equiv X_{i,l}$ for at most one $l \in \{1, \ldots, p\}$, that is $X_{i,l}$ does not depend on t, we assume $E(X_{i,l}) \neq 0$. The unobserved FE μ_i are i.i.d. with zero mean and finite variance $\sigma_\mu^2 > 0$. The random errors v_{it} are assumed to be i.i.d. with a zero mean, finite variance σ_v^2 and independent of Z_{it} and X_{it} for all i and t. Y_{it} is generated by Eq. (1).

If X_{it} contains a time invariant regressor, say the lth component of X_{it} is $X_{it,l} = W_i$. Then the corresponding coefficient $\theta_l(\cdot)$ is estimable if $M_H(z)(W \otimes e_m) \neq 0$ for a given z, where $W = (W_1, \ldots, W_n)^\top$. Simple calculations give $M_H(z)(W \otimes e_m) = (n^{-1}\sum_{i=1}^n W_i)M_H(z) \times (e_n \otimes e_m)$. The proof of Lemma A.2 in 'Proof of Theorem 1' in the appendix can be used to show that $M_H(z)(e_n \otimes e_m) \neq 0$ for any given z with probability 1. Therefore, $\theta_l(\cdot)$ is asymptotically identifiable if $n^{-1}\sum_{i=1}^n X_{it,l} \equiv n^{-1}\sum_{i=1}^n W_i \nrightarrow 0$ while $\bar{\mu} \xrightarrow{a.s.} 0$. For example, if X_{it} contains a constant, say, $X_{it,1} = W_i \equiv 1$, then $\theta_1(\cdot)$ is estimable because $n^{-1}\sum_{i=1}^n W_i = 1 \neq 0$.

Assumption 3. $K(u) = \prod_{s=1}^q k(u_s)$ is a product kernel, and the univariate kernel function $k(\cdot)$ is a uniformly bounded, symmetric (around zero) pdf with a compact support $[-1, 1]$. In addition, define $|H| = h_1 \cdots h_q$ and $||H|| = \sqrt{\sum_{j=1}^q h_j^2}$. As $n \to \infty, ||H|| \to 0, n|H| \to \infty$.

The assumptions listed above are regularity assumptions commonly seen in nonparametric estimation literature. Assumption 1 apparently excludes the case of either X_{it} or Z_{it} being $I(1)$; other than the moment restrictions, we do not impose $I(0)$ structure on X_{it} across time, since this paper considers the case that m is a small finite number. Also, instead of imposing the smoothness assumption on $f_t(\cdot|X_{it})$ and $f_{t,s}(\cdot, \cdot|X_{it}, X_{is})$ as in Assumption 1(d), we can assume that $f_t(z)E(X_{it}X_{it}^T|z)$ and $f_{t,s}(z_1, z_2)E(X_{it}X_{js}^T|z_1, z_2)$ are uniformly bounded in the domain of Z and are all twice continuously differentiable at $z \in R^q$ for all $t \neq s$ and i and j. Our version of the smoothness assumption simplifies our notation in the proofs.

Assumption 2 indicates that X_{it} can contain a constant term of 1s. The kernel function having a compact support in Assumption 3 is imposed for the sake of brevity of proof and can be removed at the cost of lengthy proofs. Specifically, the Gaussian kernel is allowed.

We use $\hat{\theta}(z)$ to denote the first column of $\hat{\beta}(z)$. Then $\hat{\theta}(z)$ estimates $\theta(z)$.

Theorem 1. Under Assumptions 1–3, we obtain the following bias and variance for $\hat{\theta}(z)$, given a finite integer $m>0$:

$$\text{bias}(\hat{\theta}(z)) = \frac{\Psi(z)^{-1}\Lambda(z)}{2} + O(n^{-1/2}|H|\ln(\ln\ n) + o(||H||^2)$$
$$\text{var}(\hat{\theta}(z)) = n^{-1}|H|^{-1}\sigma_v^2\psi(z)^{-1}\Gamma(z)\psi(z)^{-1} + o(n^{-1}|H|^{-1})$$

where $\psi(z) = |H|^{-1}\sum_{t=1}^{m} E[(1-\varpi_{it})\lambda_{it}X_{it}X_{it}^T]$, $\Lambda(z) = |H|^{-1}\sum_{t=1}^{m} E[(1-\varpi_{it})\lambda_{it}X_{it}X_{it}^T r_H(\tilde{Z}_{it}, z)] = O(||H||^2)$, and $\Gamma(z) = |H|^{-1}\sum_{t=1}^{m} E[(1-\varpi_{it})^2\lambda_{it}^2 X_{it}X_{it}^T]$.

The first term of bias$(\hat{\theta}(z))$ results from the local approximation of $\theta(z)$ by a linear function of z, which is of order $O(||H||^2)$ as usual. The second term of bias$(\hat{\theta}(z))$ results from the unknown FE μ_i: (a) if we assumed $\sum_{i=1}^{n}\mu_i = 0$, this term is zero exactly and (b) the result indicates that the bias term is dominated by the first term and will vanish as $n\to\infty$.

In the appendix, we show that

$$|H|^{-1}\sum_{t=1}^{m} E(\lambda_{it}X_{it}X_{it}^T) = \Phi(z) + o(||H||^2)$$
$$|H|^{-1}\sum_{t=1}^{m} E[\lambda_{it}X_{it}X_{it}^T r_H(\tilde{Z}_{it}, z)] = k_2\Phi(z)\Theta_H(z) + o(||H||^2)$$
$$|H|^{-1}\sum_{t=1}^{m} E(\lambda_{it}^2 X_{it}X_{it}^T) = \left(\int K^2(u)du\right)\Phi(z) + o(||H||^2)$$

where $\kappa_2 = \int k(u)\,u^2 du$, $\Phi(z) = \sum_{t=1}^{m} f_t(z)E(X_{1t}X_{1t}^T|z)$ and $\Theta_H(z) = [tr(H((\partial^2\theta_1(z))/(\partial z\partial z^T))H), \ldots, tr(H((\partial^2\theta_p(z))/(\partial z\partial z^T))H)]^T$. Since $\varpi_{it}\in[0,1)$ for all i and t, the results above imply the existence of $\Psi(z)$, $\Lambda(z)$ and $\Gamma(z)$. However, given a finite integer $m>0$, we cannot obtain explicitly the asymptotic bias and variance due to the random denominator appearing in ϖ_{it}.

Further, the following theorem gives the asymptotic normality results for $\hat{\theta}(z)$.

Theorem 2. Under Assumptions 1–3, and assuming in addition that $E|v_{it}|^{2+\delta}<\infty$ for some $\delta>0$, and that $\sqrt{n|H|}\,||H||^2 = O(1)$ *as* $\to\infty$, we have

$$\sqrt{n|H|}\left\{\hat{\theta}(z) - \theta(z) - \Psi(z)^{-1}\frac{\Lambda(z)}{2}\right\} \xrightarrow{d} N\left(0, \sum_{\theta(z)}\right)$$

where $\Sigma_{\theta(z)} = \sigma_v^2 \lim_{n\to\infty} \Psi(z)^{-1} \Gamma(z) \Psi(z)^{-1}$. Moreover, a consistent estimator for $\Sigma_{\theta(z)}$ is given as follows:

$$\begin{aligned}
\hat{\Sigma}_{\theta(z)} &= S_p \hat{\Omega}(z,H)^{-1} \hat{J}(z,H) \hat{\Omega}(z,H)^{-1} S_p^{\top} \xrightarrow{p} \Sigma_{\theta(z)} \\
\hat{\Omega}(z,H) &= n^{-1}|H|^{-1} R(z,H)^{\top} S_H(z) R(z,H) \\
\hat{J}(z,H) &= n^{-1}|H|^{-1} R(z,H)^{\top} S_H(z) \hat{V}\hat{V}^{\top} S_H(z) R(z,H)
\end{aligned}$$

where $\hat{V}$ is the vector of estimated residuals and S_p includes the first p rows of the identity matrix of dimension $p(q+1)$. Finally, a consistent estimator for the leading bias can be easily obtained based on a nonparametric local quadratic regression result.

4. TESTING RANDOM EFFECTS VERSUS FIXED EFFECTS

In this section we discuss how to test for the presence of RE versus FE in a semiparametric varying coefficient panel data model. The model remains as (1). The RE specification assumes that μ_i is uncorrelated with the regressors X_{it} and Z_{it}, while for the FE case, μ_i is allowed to be correlated with X_{it} and/or Z_{it} in an unknown way.

We are interested in testing the null hypothesis (H_0) that μ_i is a random effect versus the alternative hypothesis (H_1) that μ_i is a fixed effect. The null and alternative hypotheses can be written as

$$H_0 \ : \ \Pr\{E(\mu_i | Z_{i1}, \ldots, Z_{im}, X_{i1}, \ldots, X_{im}) = 0\} = 1 \quad \text{for all } i \tag{14}$$

$$H_1 \ : \ \Pr\{E(\mu_i | Z_{i1}, \ldots, Z_{im}, X_{i1}, \ldots, X_{im}) \neq 0\} > 0 \quad \text{for some } i \tag{15}$$

while we keep the same set-up given in model (1) under both H_0 and H_1.

Our test statistic is based on the squared difference between the FE and RE estimators, which is asymptotically zero under H_0 and positive under H_1. To simplify the proofs and save computing time, we use local constant estimator instead of local linear estimator for constructing our test.

Then following the argument in Section 2 and 'Technical Sketch: Random Effects Estimator' in the appendix, we have

$$\begin{aligned}
\hat{\theta}_{\mathrm{FE}}(z) &= \{X^{\top} S_H(z) X\}^{-1} X^{\top} S_H(z) Y \\
\hat{\theta}_{\mathrm{RE}}(z) &= \{X^{\top} W_H(z) X\}^{-1} X^{\top} W_H(z) Y
\end{aligned}$$

where X is an $(nm)\times p$ with $X=(X_1^\top,\ldots,X_n^\top)$, and for each $i, X_i=(X_{i1},\ldots,X_{im})^\top$ is an $m\times p$ matrix with $X_{it}=[X_{it,1},\ldots,X_{it,p}]^\top$. Motivated by Li, Huang, Li, and Fu (2002), we remove the random denominator of $\hat\theta_{\mathrm{FE}}(z)$ by multiplying $X^\top S_H(z)X$, and our test statistic will be based on

$$\begin{aligned} T_n &= \int \{\hat\theta_{\mathrm{FE}}(z)-\hat\theta_{\mathrm{RE}}(z)\}^\top\{X^\top S_H(z)X\}^\top\{X^\top S_H(z)X\}\{\hat\theta_{\mathrm{FE}}(z)-\hat\theta_{\mathrm{RE}}(z)\}dz \\ &= \int \tilde U(z)^\top S_H(z)XX^\top S_H(z)\,\tilde U(z)dz \end{aligned}$$

since $\{X^\top S_H(z)X\}\{\hat\theta_{\mathrm{FE}}(z)-\hat\theta_{\mathrm{RE}}(z)\}=X^\top S_H(z)\{Y-X\hat\theta_{\mathrm{RE}}(z)\}\equiv X^\top S_H(z)\tilde U(z)$. To simplify the statistic, we make several changes in T_n. First, we simplify the integration calculation by replacing $\tilde U(z)$ by $\tilde U$, where $\hat U\equiv\hat U(Z)=Y-B\{X,\hat\theta_{\mathrm{RE}}(Z)\}$ and $B\{X,\hat\theta_{\mathrm{RE}}(Z)\}$ stacks up $X_{it}^T\hat\theta_{\mathrm{RE}}(Z_{it})$ in the increasing order of i first, then of t. Second, to overcome the complexity caused by the random denominator in $M_H(z)$, we replace $M_H(z)$ by $M_D=I_{n\times m}-m^{-1}I_n\otimes(e_me_m^\top)$ such that the FE can be removed due to the fact that $M_DD_0=0$. With the above modification and also removing the $i=j$ terms in T_n (since T_n contains two summations $\sum_i\sum_j$), our further modified test statistic is given by

$$\tilde T_n \stackrel{\mathrm{def}}{=} \sum_{i=1}^n\sum_{j\neq i}\hat U_i^\top Q_m\int K_H(Z_i,z)X_i^\top X_jK_H(Z_j,z)dzQ_m\hat U_j$$

where $Q_m=I_m-m^{-1}e_me_m^\top$. If $|H|\to0$ as $n\to\infty$, we obtain

$$\begin{aligned} &|H|^{-1}\int K_H(Z_i,z)X_i^\top X_jK_H(Z_j,z)dz \\ &\quad=\begin{bmatrix} \bar K_H(Z_{i,1},Z_{j,1})X_{i,1}^\top X_{j,1} & \cdots & \bar K_H(Z_{i,1},Z_{j,m})X_{i,1}^\top X_{j,m} \\ \vdots & \ddots & \vdots \\ \bar K_H(Z_{i,m},Z_{j,1})X_{i,m}^\top X_{j,1} & \ldots & \bar K_H(Z_{i,m},Z_{j,m})X_{i,m}^\top X_{j,m}\end{bmatrix} \end{aligned} \tag{16}$$

where $\bar K_H(Z_{it},Z_{js})=\int K\{H^{-1}(Z_{it}-Z_{js})+\omega\}K(\omega)d\omega$. We then replace $\bar K_H(Z_{it},Z_{js})$ by $K_H(Z_{it},Z_{js})$; this replacement will not affect the essence of the test statistic since the local weight is untouched. Now, our proposed test statistic is given by

$$\hat T_n=\frac{1}{n^2|H|}\sum_{i=1}^n\sum_{j\neq i}^n\hat U_i^\top Q_mA_{i,j}Q_m\hat U_j \tag{17}$$

where $A_{i,j}$ equals the right-hand side of Eq. (16) after replacing $\bar{K}_H(Z_{it}, Z_{js})$ by $K_H(Z_{it}, Z_{js})$. Finally, to remove the asymptotic bias term of the proposed test statistic, we calculate the leave-one-unit-out RE estimator of $\theta(Z_{it})$; that is for a given pair of (i, j) in the double summation of Eq. (17) with $i \neq j$, $\hat{\theta}_{\mathrm{RE}}(Z_{it})$ is calculated without using the observations on the jth unit, $\{(X_{jt}, Z_{jt}, Y_{jt})\}_{t=1}^{m}$ and $\hat{\theta}_{\mathrm{RE}}(Z_{jt})$ is calculated without using the observations on the ith unit.

We present the asymptotic properties of this test below and delay the proofs to the appendix in 'Proof of Theorem 3'.

Theorem 3. Under Assumptions 1–3, and $f_t(z)$ has a compact support S for all t, and $n\sqrt{|H|}\,||H||^4 \to 0$ as $n \to \infty$, then we have under H_0 that

$$J_n = n\sqrt{|H|}\frac{\hat{T}_n}{\hat{\sigma}_0} \xrightarrow{d} N(0,1) \tag{18}$$

where $\hat{\sigma}_0^2 = \frac{2}{n^2|H|}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(\hat{V}_i^{\top} Q_m A_{i,j} Q_m \hat{V}_j)^2$ is a consistent estimator of

$$\sigma_0^2 = 4\left(1 - \frac{1}{m}\right)^2 \sigma_v^4 \int K^2(u)du \sum_{t=2}^{m}\sum_{s=1}^{t-1} E[f_t(Z_{1s})(X_{1s}^{\top} X_{2t})^2]$$

where $\hat{V}_{it} = Y_{it} - X_{it}^T\hat{\theta}_{\mathrm{FE}}(Z_{it}) - \hat{\mu}_i$ and for each pair of (i, j), $i \neq j$, $\hat{\theta}_{\mathrm{FE}}(Z_{it})$ is a leave-two-unit-out FE estimator without using the observations from the ith and jth units and $\hat{\mu}_i = \bar{Y}_i - m^{-1}\sum_{t=1}^{m} X_{it}^{\top}\hat{\theta}_{\mathrm{FE}}(Z_{it})$. Under H_1, $\Pr[J_n > B_n] \to 1$ as $n \to \infty$, where B_n is any nonstochastic sequence with $B_n = o(n\sqrt{|H|})$.

Assuming that $f_t(z)$ has a compact support S for all t is to simplify the proof of $\sup_{z\in S}||\hat{\theta}_{RE}(z) - \theta(z)|| = o_p(1)$ as $n \to \infty$; otherwise, some trimming procedure has to be placed to show the uniform convergence result and the consistency of $\hat{\sigma}_0^2$ as an estimator of σ_0^2. Theorem 3 states that the test statistic $J_n = n\sqrt{|H|}\hat{T}_n/\hat{\sigma}_0$ is a consistent test for testing H_0 against H_1. It is a one-sided test. If J_n is greater than the critical values from the standard normal distribution, we reject the null hypothesis at the corresponding significance levels.

5. MONTE CARLO SIMULATIONS

In this section we report some Monte Carlo simulation results to examine the finite sample performance of the proposed estimator. The following data

generating process is used:

$$Y_{it} = \theta_1(Z_{it}) + \theta_2(Z_{it}) + X_{it} + \mu_i + v_{it} \tag{19}$$

where $\theta_1(z) = 1 + z + z^2, \theta_2(z) = \sin(z\pi), Z_{it} = \omega_{it} + \omega_{i,t-1,}\omega_{it}$ is i.i.d. uniformly distributed in $[0, \pi/2], X_{it} = 0.5X_{i,t-1} + \xi_{it}, \xi_{it}$ is i.i.d. $N(0, 1)$. In addition, $\mu_i = c_0\bar{Z}_{i.} + \mu_i$ for $i = 2, \ldots, n$ with $c_0 = 0$, 0.5, and 1.0, u_i is i.i.d. $N(0, 1)$. When $c_0 \neq 0$, μ_i and Z_{it} are correlated; we use c_0 to control the correlation between μ_i and $\bar{Z}_i = m^{-1}\sum_{t=1}^{m} Z_{it}$. Moreover, v_{it} is i.i.d. $N(0, 1)$, and ω_{it}, ξ_{it}, u_i and v_{it} are independent of each other.

We report estimation results for both the proposed FE and RE estimators; see 'Technical Sketch: Random Effects Estimator' in the appendix for the asymptotic results of the RE estimator. To learn how the two estimators perform when we have FE model and when we have RE model, we use the integrated squared error as a standard measure of estimation accuracy:

$$\mathrm{ISE}(\hat{\theta}_l) = \int \{\hat{\theta}_l(z) - \theta_l(z)\}^2 f(z)\, dz \tag{20}$$

which can be approximated by the average mean squared error (AMSE)

$$\mathrm{AMSE}(\hat{\theta}_l) = (nm)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{m} [\hat{\theta}_l(Z_{it}) - \theta_l(Z_{it})]^2$$

for $l = 1, 2$. In Table 1 we present the average value of $\mathrm{AMSE}(\hat{\theta}_l)$ from 1,000 Monte Carlo experiments. We choose $m = 3$ and $n = 50$, 100 and 200.

Table 1. Average Mean Squared Errors (AMSE) of the Fixed- and Random-Effects Estimators When the Data Generation Process is a Random Effects Model and When it is a Fixed Effects Model.

Data Process	Random Effects Estimator			Fixed Effects Estimator		
	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
Estimating $\theta_1(\cdot)$:						
$c_0 = 0$	0.0951	0.0533	0.0277			
$c_0 = 0.5$	0.6552	0.5830	0.5544	0.1381	0.1163	0.1021
$c_0 = 1.0$	2.2010	2.1239	2.2310			
Estimating $\theta_2(\cdot)$:						
$c_0 = 0$	0.1562	0.0753	0.0409			
$c_0 = 0.5$	0.8629	0.7511	0.7200	0.1984	0.1379	0.0967
$c_0 = 1.0$	2.8707	2.4302	2.5538			

Since the bias and variance of the proposed FE estimator do not depend on the values of the FE, our estimates are the same for different values of c_0; however, it is not true under the RE model. Therefore, the results derived from the FE estimator are only reported once in Table 1 since it is invariant to different values of c_0.

It is well known that the performance of non/semiparametric models depends on the choice of bandwidth. Therefore, we propose a leave-one-unit-out cross-validation method to automatically select the optimal bandwidth for estimating both the FE and RE models. Specifically, when estimating $\theta(\cdot)$ at a point Z_{it}, we remove $\{(X_{it}, Y_{it}, Z_{it})\}_{t=1}^{m}$ from the data and only use the rest of $(n-1)m$ observations to calculate $\hat{\theta}_{(-i)}(Z_{it})$. In computing the RE estimate, the leave-one-unit-out cross-validation method is just a trivial extension of the conventional leave-one-out cross-validation method. The conventional leave-one-out method fails to provide satisfying results due to the existence of unknown FE. Therefore, when calculating the FE estimator, we use the following modified leave-one-unit-out cross-validation method:

$$\hat{H}_{opt} = \arg\min_{H} [Y - B\{X, \hat{\theta}_{(-1)}(Z)\}]^{\top} M_D^{\top} M_D [Y - B\{X, \hat{\theta}_{(-1)}(Z)\}] \tag{21}$$

where $M_D = I_{n\times m} - m^{-1} I_n \otimes (e_m e_m^{\top})$ satisfies $M_D D_0 = 0$; this is used to remove the unknown FE. In addition, $B\{X, \hat{\theta}_{(-1)}(Z)\}$ stacks up $X_{it}^{\top}\hat{\theta}_{(-i)}(Z_{it})$ in the increasing order of i first, then of t. Simple calculations give

$$\begin{aligned}
&[Y - B\{X, \hat{\theta}_{(-1)}(Z)\}]^{\top} M_D^{\top} M_D [Y - B\{X, \hat{\theta}_{(-1)}(Z)\}] \\
&\quad = [B\{X, \theta(Z)\} - B\{X, \hat{\theta}_{(-1)}(Z)\}]^{\top} M_D^{\top} M_D [B\{X, \theta(Z)\} - B\{X, \hat{\theta}_{(-1)}(Z)\}] \\
&\qquad + 2[B\{X, \theta(Z)\} - B\{X, \hat{\theta}_{(-1)}(Z)\}]^{\top} M_D^{\top} M_D V + V^{\top} M_D M_D V
\end{aligned} \tag{22}$$

where the last term does not depend on the bandwidth. If v_{it} is independent of the $\{X_{js}, Z_{js}\}$ for all i, j, s and t, or (X_{it}, Z_{it}) is strictly exogenous variable, then the second term has zero expectation because the linear transformation matrix M_D removes a cross-time *not* cross-sectional average from each variable, for example $\tilde{Y}_{it} = Y_{it} - m^{-1}\sum_{s=1}^{m} Y_{is}$ for all i and t. Therefore, the first term is the dominant term in large samples and Eq. (21) is used to find an optimal smoothing matrix minimizing a weighted mean squared error of $\{\hat{\theta}(Z_{it})\}$. Of course, we could use other weight matrices in Eq. (21) instead of M_D as long as the weight matrices can remove the FE and do not trigger a non-zero expectation of the second term in Eq. (22).

Table 1 shows that the RE estimator performs better than the FE estimator when the true model is an RE model. However, the FE estimator

performs much better than the RE estimator when the true model is an FE model. This is expected since the RE estimator is inconsistent when the true model is the FE model. Therefore, our simulation results indicate that a test for RE against FE will be always in demand when we analyze panel data models. In Tables 2–4 we report simulation results of the proposed nonparametric test of RE against FE.

For the selection of the bandwidth h, for univariate case, Theorem 3 indicates that $h \to 0$, $nh \to \infty$, and $nh^{9/2} \to 0$ as $n \to \infty$; if we take $h \sim n^{-\alpha}$, Theorem 3 requires $\alpha \in ((2/9), 1)$. To fulfil both conditions $nh \to \infty$ and $nh^{9/2} \to 0$ as $n \to \infty$, we use $\alpha = 2/7$. Therefore, we use $h = c(nm)^{-2/7}\hat{\sigma}_z$ to calculate the RE estimator with c taking a value from .8, 1.0 and 1.2. Since the computation is very time consuming, we only report results for $n = 50$

Table 2. Percentage Rejection Rate When $c_0 = 0$.

C	$n = 50$			$n = 100$		
	1%	5%	10%	1%	5%	10%
0.8	0.007	0.015	0.24	0.21	0.35	0.46
1.0	0.011	0.023	0.041	0.025	0.040	0.062
1.2	0.019	0.043	0.075	0.025	0.054	0.097

Table 3. Percentage Rejection Rate When $c_0 = 0.5$.

C	$n = 50$			$n = 100$		
	1%	5%	10%	1%	5%	10%
0.8	0.626	0.719	0.764	0.913	0.929	0.933
1.0	0.682	0.780	0.819	0.935	0.943	0.951
1.2	0.719	0.811	0.854	0.943	0.962	0.969

Table 4. Percentage Rejection Rate When $c_0 = 1.0$.

C	$n = 50$			$n = 100$		
	1%	5%	10%	1%	5%	10%
0.8	0.873	0.883	0.888	0.943	0.944	0.946
1.0	0.908	0.913	0.921	0.962	0.966	0.967
1.2	0.931	0.938	0.944	0.980	0.981	0.982

and 100. With $m = 3$, the effective sample size is 150 and 300, which is a small but moderate sample size. Although the bandwidth chosen this way may not be optimal, the results in Tables 2–4 show that the proposed test statistic is not very sensitive to the choice of h when c changes, and that a moderate sized distortion and decent power are consistent with the findings in the nonparametric tests literature. We conjecture that some bootstrap procedures can be used to reduce the size distortion in finite samples. We will leave this as a future research topic.

6. CONCLUSION

In this paper we proposed a local linear least-squared method to estimate an FE varying coefficient panel data model when the number of observations across time is finite; a data-driven method was introduced to automatically find the optimal bandwidth for the proposed FE estimator. In addition, we introduced a new test statistic to test for an RE model against an FE model. Monte Carlo simulations indicate that the proposed estimator and test statistic have good finite sample performance.

ACKNOWLEDGMENTS

Sun's research was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC). Carroll's research was supported by a grant from the National Cancer Institute (CA-57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We thank two anonymous referees and Prof. Qi Li for their comments.

REFERENCES

Arellano, M. (2003). *Panel data econometrics*. New York: Oxford University Press.

Baltagi, B. (2005). *Econometrics analysis of panel data* (2nd ed.). New York: Wiley.

Cai, Z., & Li, Q. (2008). Nonparametric estimation of varying coefficient dynamic panel data models. *Econometric Theory*, *24*, 1321–1342.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Annals of Statistics*, *14*, 1–16.

Henderson, D. J., Carroll, R. J., & Li, Q. (2008). Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, *144*, 257–275.

Henderson, D. J., & Ullah, A. (2005). A nonparametric random effects estimator. *Economics Letters*, *88*, 403–407.

Hsiao, C. (2003). *Analysis of panel data* (2nd ed.). New York: Cambridge University Press.
Li, Q., Huang, C. J., Li, D., & Fu, T. (2002). Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics*, *20*, 412–422.
Li, Q., & Stengos, T. (1996). Semiparametric estimation of partially linear panel data models. *Journal of Econometrics*, *71*, 389–397.
Lin, D. Y., & Ying, Z. (2001). Semiparametric and nonparametric regression analysis oí longitudinal data (with discussion). *Journal of the American Statistical Association*, *96*, 103–126.
Lin, X., & Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, *95*, 520–534.
Lin, X., & Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimation equations. *Journal of the American Statistical Association*, *96*, 1045–1056.
Lin, X., & Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B*, *68*, 68–88.
Lin, X., Wang, N., Welsh, A. H., & Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for longitudinal/clustered data. *Biometrika*, *91*, 177–194.
Poirier, D. J. (1995). *Intermediate statistics and econometrics: A comparative approach*. Cambridge, MA: The MIT Press.
Ruckstuhl, A. F., Welsh, A. H., & Carroll, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica*, *10*, 51–71.
Su, L., & Ullah, A. (2006). Profile likelihood estimation of partially linear panel data models with fixed effects. *Economics Letters*, *92*, 75–81.
Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, *90*, 43–52.
Wu, H., & Zhang, J. Y. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, *97*, 883–897.

APPENDIX

Proof of Theorem 1

To make our mathematical formula short, we introduce some simplified notations first: for each i and t, $\lambda_{it} = K_H(Z_{it,}z)$ and $c_H(Z_i, z)^{-1} = \sum_{t=1}^{m} \lambda_{it}$, and for any positive integers i, j, t, s

$$[\cdot]_{it,js} = G_{it}(z,H)G_{js}^T(z,H) = \begin{bmatrix} 1 & G_{js1} & \cdots & G_{jsq} \\ G_{it1} & G_{it1}G_{js1} & \cdots & G_{it1}G_{jsq} \\ \vdots & \vdots & \ddots & \vdots \\ G_{itq} & G_{itq}G_{js1} & \cdots & G_{itq}G_{jsq} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & (H^{-1}(Z_{js} - z))^T \\ H^{-1}(Z_{it} - z) & H^{-1}(Z_{it} - z)(H^{-1}(Z_{js} - z))^T \end{bmatrix} \tag{A.1}$$

where the $(l+1)$th element of $G_{js}(z,H)$ is $G_{jsl}=(Z_{jsl}-z_l)/h_l, l=1,\ldots,q$. Simple calculations show that

$$[\cdot]_{i_1t_1,i_2t_2}[\cdot]_{j_1s_1,j_2s_2}=\left(1+\sum_{j=1}^{q}G_{j_1s_1j}G_{i_2t_2j}\right)[\cdot]_{i_1t_1,j_2s_2}$$

$$R_i(z,H)^TK_H(Z_i,z)e_me_m^TK_H(Z_j,z)R_j(z,H)=\sum_{t=1}^{m}\sum_{s=1}^{m}\lambda_{it}\lambda_{js}[\cdot]_{it,js}\otimes(X_{it}X_{js}^T)$$

In addition, we obtain for a finite positive integer j

$$|H|^{-1}\sum_{t=1}^{m}E[\lambda_{it}^j[\cdot]_{it,it}|X_{it}]=\sum_{t=1}^{m}E[(S_{t,j,1}|X_{it})+O_p(||H||^2)] \tag{A.2}$$

$$|H|^{-1}\sum_{t=1}^{m}E\left[\lambda_{it}^{2j}\sum_{j'=1}^{q}G_{itj'}^2[\cdot]_{it,it}|X_{it}\right]=\sum_{t=1}^{m}E(S_{t,j,2}|X_{it})+O_p(||H||^2) \tag{A.3}$$

where

$$S_{t,j,1}=\begin{bmatrix} f_t(z|X_{it})\int K^j(u)du & \dfrac{\partial f_t(z|X_{it})}{\partial z^T}HR_{K,j} \\ R_{K,j}H\dfrac{\partial f_t(z|X_{it})}{\partial z} & f_t(z|X_{it})R_{K,j}\end{bmatrix} \tag{A.4}$$

$$S_{t,j,2}=\begin{bmatrix} f_t(z|X_{it})\int K^{2j}(u)u^Tudu & \dfrac{\partial f_t(z|X_{1t})}{\partial z^T}H\Gamma_{K,2j} \\ \Gamma_{K,2j}H\dfrac{\partial f_t(z|X_{it})}{\partial z} & f_t(z|X_{it})\Gamma_{K,2j}\end{bmatrix} \tag{A.5}$$

where $R_{K,j}=\int K^j(u)\,uu^Tdu$ and $\Gamma_{K,2j}=\int K^{2j}(u)(u^Tu)(uu^T)du$.

Moreover, for any finite positive integer j_1 and j_2, we have

$$\begin{aligned}&|H|^{-2}\sum_{t=1}^{m}\sum_{s\neq t}^{m}E[\lambda_{it}^{j_1}\lambda_{is}^{j_2}[\cdot]_{it,is}|X_{it},X_{is}] \\ &\quad=\sum_{t=1}^{m}\sum_{s\neq t}^{m}E(T_{j_1,j_2,1}^{(t,s)}|X_{it},X_{is})+O_p(||H||^2)\end{aligned} \tag{A.6}$$

$$\begin{aligned}&|H|^{-2}\sum_{t=1}^{m}\sum_{s\neq t}^{m}E\left[\lambda_{it}^{j_1}\lambda_{is}^{j_2}\left(\sum_{j'=1}^{q}G_{itj'}G_{isj'}\right)[\cdot]_{it,is}|X_{it},X_{is}\right] \\ &\quad=\sum_{t=1}^{m}\sum_{s\neq t}^{m}E(T_{j_1,j_2,2}^{(t,s)}|X_{it},X_{is})+O_p(||H||^2)\end{aligned} \tag{A.7}$$

where we define $b_{j1,j_2,i_1,i_2} = \int K^{j1}(u)u_1^{2i1}du \int K^{j2}(u)u_1^{2i_2}du$

$$T^{(t,s)}_{j_1,j_2,1} = \begin{bmatrix} f_{t,s}(z,z|X_{it},X_{is})b_{j_1,j_2,0,0} & \nabla_s^T f_{t,s}(z,z|X_{it},X_{is})Hb_{j_1,j_2,0,1} \\ H\nabla_t f_{t,s}(z,z|X_{it},X_{is})b_{j_1,j_2,1,0} & H\nabla^2_{t,s} f_{t,s}(z,z|X_{it},X_{is})Hb_{j_1,j_2,1,1} \end{bmatrix}$$

and

$$T^{(t,s)}_{j_1,j_2,2} = \begin{bmatrix} tr(H\nabla^2_{t,s} f_{t,s}(z,z|X_{it},X_{is})H) & \nabla_t^T f_{t,s}(z,z|X_{it},X_{is})H \\ H\nabla_s f_{t,s}(z,z|X_{it},X_{is}) & f_{t,s}(z,z|X_{it},X_{is})I_{q\times q} \end{bmatrix} b_{j_1,j_2,1,1}$$

with $\nabla_s f_{t,s}(z,z|X_{it},X_{is}) = \partial f_{t,s}(z,z|X_{it},X_{is})/\partial z_s$ and $\nabla^2_{t,s} f_{t,s}(z,z|X_{it},X_{is}) = \partial^2 f_{t,s}(z,z|X_{it},X_{is})/(\partial z_t \partial z_s^T)$.

The conditional bias and variance of $\mathrm{vec}(\hat{\beta}(z))$ are given as follows:

$$\begin{aligned}\mathrm{Bias}[\mathrm{vec}(\hat{\beta}(z))|\{X_{it},Z_{it}\}] &= [R(z,H)^T S_H(z)R(z,H)]^{-1}R(z,H)^T S_H(z) \\ &\quad \times \left[\prod(z,H)/2 + D_0\mu_0\right]\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[\mathrm{vec}(\hat{\beta}(z))|\{X_{it},Z_{it}\}] &= \sigma_v^2[R(z,H)^T S_H(z)R(z,H)]^{-1}[R(z,H)^T S_H^2(z)R(z,H)] \\ &\quad \times [R(z,H)^T S_H(z)R(z,H)]^{-1}\end{aligned}$$

Lemma A.1. If Assumption A3 holds, we have

$$\left[\sum_{i=1}^n c_H(Z_i,z)\right]^{-1} = O_p\left(n^{-1}|H|\ln(\ln\ n)\right) \tag{A.8}$$

Proof. Simple calculations give $E\left(\sum_{t=1}^m K_H(Z_{it},z)\right) = |H|f(z) + O(|H|\,||H||^2)$ and $E[K_H(Z_{it},z)] = |H|f_t(z) + O(|H|\,||H||^2)$, where

$f(z) = \sum_{t=1}^{m} f_t(z)$. Next, we obtain for any small $\varepsilon > 0$

$$\begin{aligned}
&\Pr\left\{ \max_{1\le i\le n} \sum_{t=1}^{m} \lambda_{it} > \varepsilon^{-1} f(z)|H| \ln(\ln\ n) \right\} \\
&\quad = 1 - \Pr\left\{ \max_{1\le i\le n} \sum_{t=1}^{m} \lambda_{it} \le \varepsilon^{-1} f(z)|H| \ln(\ln\ n) \right\} \\
&\quad = 1 - \left\{ 1 - \Pr\left\{ \sum_{t=1}^{m} \lambda_{it} > \varepsilon^{-1} f(z)|H| \ln(\ln\ n) \right\} \right\}^n \\
&\qquad \le 1 - \left\{ 1 - \frac{\varepsilon E(\sum_{t=1}^{m} \lambda_{it})}{f(z)|H| \ln(\ln\ n)} \right\}^n \\
&\qquad \le 1 - \{1 - \varepsilon(1 + M||H||^2)/\ln(\ln\ n)\}^n \to 0 \text{ as } n \to \infty
\end{aligned}$$

where the first inequality uses the generalized Chebyshev inequality, and the limit is derived using the l'Hôpital's rule. This will complete the proof of this lemma.

Lemma A.2. Under Assumptions 1–3, we have

$$n^{-1}|H|^{-1} R(z,H)^T S_H(z) R(z,H) \approx |H|^{-1} \sum_{t=1}^{m} E(\varpi_{it}\lambda_{it}[\cdot]_{it,it} \otimes (X_{it}K_{it}^T))$$

where $\varpi_{it} = \lambda_{it}/\sum_{t=1}^{m} \lambda_{it} \in (0,1)$ for all i and t.

Proof. First, simple calculation gives

$$\begin{aligned}
A_n &= R(z,H)^T S_H(z) R(z,H) = R(z,H)^T W_H(z) M_H(z) R(z,H) \\
&= \sum_{i=1}^{n} R_i(z,H)^T K_H(Z_i,z) R_i(z,H) \\
&\quad - \sum_{j=1}^{n}\sum_{i=1}^{n} q_{ij} R_i(z,H)^T K_H(Z_i,z) e_m e_m^T K_H(Z_j,z) R_j(z,H) \\
&= \sum_{i=1}^{n}\sum_{t=1}^{m} \lambda_{it}[\cdot]_{it,it} \otimes (X_{it}X_{it}^T) - \sum_{i=1}^{n} q_{ii} \sum_{t=1}^{m}\sum_{s=1}^{m} \lambda_{it}\lambda_{is}[\cdot]_{it,is} \otimes (X_{it}X_{is}^T) \\
&\quad - \sum_{j=1}^{n}\sum_{i\ne j}^{n} q_{ij} \sum_{t=1}^{m}\sum_{s=1}^{m} \lambda_{it}\lambda_{js}[\cdot]_{it,js} \otimes (X_{it}X_{js}^T) = A_{n1} - A_{n2} - A_{n3}
\end{aligned}$$

where $M_H(z) = I_{n\times m} - [Q \otimes (e_m e_m^T)] W_H(z)$, and the typical elements of Q are $q_{ii} = c_H(Z_i,z) - c_H(Z_i,z)^2/\sum_{i=1}^{n} c_H(Z_i,z)$ and $q_{ij} = -c_H(Z_i,z)c_H(Z_j,z)/\sum_{i=1}^{n} c_H(Z_i,z)$ for $i \ne j$. Here, $c_H(Z_i,z) = \left(\sum_{t=1}^{m} \lambda_{it}\right)^{-1}$ for all i.

Applying (A.2), (A.3), (A.6) and (A.7) to A_{n1}, we have $n^{-1}|H|^{-1}A_{n1} \approx \sum_{t=1}^{m} E[S_{t,1,1} \otimes (X_{it}X_{it}^T)] + O_p(||H||^2) + O_p(n^{-(1/2)}|H|^{-(1/2)})$ if $||H|| \to 0$ and $n|H| \to \infty$ as $n \to \infty$.

Apparently, $\sum_{t=1}^{m} \varpi_{it} = 1$ for all i. In addition, since the kernel function $K(\cdot)$ is zero outside the unit circle by Assumption 3, the summations in A_{n2} are taken over units such that $||H^{-1}(Z_{it} - z)|| \leq 1$. By Lemma A.1 and by the LLN given Assumption 1 (a), we obtain

$$\left\| \frac{1}{n|H| \sum_{i=1}^{n} c_H(Z_i, z)} \sum_{i=1}^{n} \sum_{t=1}^{m} \sum_{s=1}^{m} \varpi_{it}\varpi_{is}[\cdot]_{it,is} \otimes (X_{it}X_{is}^T) \right\| = O_p(n^{-1} \ln(\ln\ n))$$

and

$$\begin{aligned} &\left\| \frac{1}{n|H|} \sum_{i=1}^{n} \sum_{t=1}^{m} \sum_{s \neq t}^{m} \frac{\lambda_{it}\lambda_{is}}{\sum_{t=1}^{m} \lambda_{it}} [\cdot]_{it,is} \otimes (X_{it}X_{is}^T) \right\| \\ &\quad \leq \frac{1}{2n|H|} \sum_{i=1}^{n} \sum_{t=1}^{m} \sum_{s \neq t}^{m} \sqrt{\lambda_{it}\lambda_{is}} \left\| [\cdot]_{it,is} \otimes (X_{it}X_{is}^T) \right\| = O_p(|H|) \end{aligned}$$

where we use $\sum_{t=1}^{m} \lambda_{it} \geq \lambda_{it} + \lambda_{is} \geq 2\sqrt{\lambda_{it}\lambda_{is}}$ for any $t \neq s$.

Hence, we have $n^{-1}|H|^{-1}A_{n2} = n^{-1}|H|^{-1} \sum_{i=1}^{n} \sum_{t=1}^{m} \varpi_{it}\lambda_{it}[\cdot]_{it,it} \otimes (X_{it}X_{it}^T) + O_p(|H|)$. Denote $d_{it} = \varpi_{it}\lambda_{it}[\cdot]_{it,it} \otimes (X_{it}X_{it}^T)$ and $\Delta_n = n^{-1}|H|^{-1} \sum_{i=1}^{n} \sum_{t=1}^{m} (d_{it} - Ed_{it})$. It is easy to show that $n^{-1}|H|^{-1}\Delta_n = O_p(n^{-1/2}|H|^{-1/2})$. Since $E(||d_{it}||) \leq E[\lambda_{it}||[\cdot]_{it,it} \otimes (X_{it}X_{it}^T)||] \leq M|H|$ holds for all i and t, $n^{-1}|H|^{-1}A_{n2} = |H|^{-1} \sum_{t=1}^{m} E[\varpi_{it}\lambda_{it}[\cdot]_{it,it} \otimes (X_{it}X_{it}^T)] + o_p(1)$ exists, but we cannot calculate the exact expectation due to the random denominator.

Consider A_{n3}. We have $n^{-1}|H|^{-1}||A_{n3}|| = O_p(|H|^2 \ln(\ln\ n))$ by Lemma A.1, Assumption 1, and the fact that $n^{-1}|H|^{-1} \sum_{i=1}^{n} \sum_{t=1}^{m} I(||H^{-1}(Z_{it} - z)|| \leq 1) = 2f(z) + O_p(||H||^2) + O_p(n^{-1/2}|H|^{-1/2})$.

Hence, we obtain

$$\begin{aligned} n^{-1}|H|^{-1}A_n &\approx n^{-1}|H|^{-1}A_{n1} - n^{-1}|H|^{-1} \sum_{i=1}^{n} \sum_{t=1}^{m} \varpi_{it}\lambda_{it}[\cdot]_{it,it} \otimes (X_{it}X_{it}^T) \\ &= n^{-1}|H|^{-1} \sum_{i=1}^{n} \sum_{t=1}^{m} (1 - \varpi_{it})\lambda_{it}[\cdot]_{it,it} \otimes (X_{it}X_{it}^T) \\ &= |H|^{-1} \sum_{t=1}^{m} E[(1 - \varpi_{it})\lambda_{it}[\cdot]_{it,it} \otimes (X_{it}X_{it}^T)] + o_p(1) \end{aligned}$$

This will complete the proof of this Lemma.

Lemma A.3. Under Assumptions 1–3, we have

$$n^{-1}|H|^{-1}R(z,H)^T S_H(z)\prod(z,H) \approx |H|^{-1}\sum_{t=1}^{m} E[(1-\varpi_{it})\lambda_{it}(G_{it}\otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it},z)]$$

Proof. Simple calculations give

$$\begin{aligned}
B_n &= R(z,H)^T S_H(z)\prod(z,H)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{m}\lambda_{it}(G_{it}\otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it},z)\\
&\quad -\sum_{j=1}^{n}\sum_{i=1}^{n}q_{ij}\sum_{s=1}^{m}\sum_{t=1}^{m}\lambda_{js}\lambda_{it}(G_{it}\otimes X_{it})X_{js}^T r_H(\tilde{Z}_{js},z)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{m}\lambda_{it}(G_{it}\otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it},z)\\
&\quad -\sum_{i=1}^{n}q_{ii}\sum_{t=1}^{m}\lambda_{it}^2(G_{it}\otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it},z)\\
&\quad -\sum_{i=1}^{n}q_{ii}\sum_{t=1}^{m}\sum_{s\neq t}^{m}\lambda_{is}\lambda_{it}(G_{it}\otimes X_{it})X_{is}^T r_H(\tilde{Z}_{is},z)\\
&\quad -\sum_{j=1}^{n}\sum_{i\neq j}^{n}q_{ij}\sum_{t=1}^{m}\sum_{s=1}^{m}\lambda_{js}\lambda_{it}(G_{it}\otimes X_{it})X_{js}^T r_H(\tilde{Z}_{js},z)\\
&= B_{n1}-B_{n2}-B_{n3}-B_{n4},
\end{aligned}$$

where $\Pi(z, N)$ is defined in Section 3. Using the same method in the proof of Lemma A.2, we show $n^{-1}|H|^{-1}B_n \approx n^{-1}|H|^{-1}\sum_{i=1}^{n}\sum_{t=1}^{m}(1-\varpi_{it})\lambda_{it}(G_{it}\otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it},z)$.

For $l=1,\ldots,k$ we have

$$|H|^{-1}E[\lambda_{it}r_{H,l}(Z_{it},z)|X_{it}] = \kappa_2 f_t\left(\frac{z}{X_{it}}\right)\Theta_H(z)+O_p(||H||^4)$$

$$|H|^{-1}E[\lambda_{it}r_{H,l}(Z_{it},z)H^{-1}(Z_{it}-z)|X_{it}] = O_p(||H||^3)$$

and $E(n^{-1}|H|^{-1}B_{n1}) \approx \{\kappa_2[\Phi(z)\Theta_H(z)]^T, O(||H||^3)\}^T$, where

$$\Theta_H(z) = \left[tr\left(H\frac{\partial^2\theta_1(z)}{\partial z\partial z^T}H\right),\ldots,tr\left(H\frac{\partial^2\theta_k(z)}{\partial z\partial z^T}H\right)\right]^T$$

Similarly, we can show that Var $(n^{-1}|H|^{-1}B_{n1}) = O(n^{-1}|H|^{-1}||H||^4)$ if $E(||X_{it}X_{is}^T X_{it}X_{is}^T||) < M < \infty$ for all t and s.

In addition, it is easy to show that $n^{-1}|H|^{-1}\sum_{i=1}^n\sum_{t=1}^m \varpi_{it}\lambda_{it}$ $(G_{it} \otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it}, z) = n^{-1}|H|^{-1}\sum_{i=1}^n\sum_{t=1}^m E[\varpi_{it}\lambda_{it}(G_{it} \otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it}, z)] + O_p$ $(n^{-1/2}|H|^{-1/2}||H||^2)$, where $|H|^{-1}\sum_{t=1}^m E[\varpi_{it}\lambda_{it}(G_{it} \otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it}, z)] \le |H|^{-1}\sum_{t=1}^m E[\lambda_{it}||(G_{it} \otimes X_{it})X_{it}^T r_H(\tilde{Z}_{it}, z)||] \le M||H||^2 < \infty$ for all i and t.

This will complete the proof of this lemma.

Lemma A.4. Under Assumptions 1–3, we have

$$n^{-1}|H|^{-1}R(z, H)^T S_H(z)D_0\mu_0 = O_p(n^{-1/2}|H|\ln(\ln\ n)).$$

Proof. Simple calculations give $M_H(z)D_0\mu_0 = \bar{\mu}M_H(z)(e_n \otimes e_m)$, where $\bar{\mu} = n^{-1}\sum_{i=1}^n \mu_i$. It follows that

$$\begin{aligned}
C_n &= R(z, H)^T S_H(z)D_0\mu_0 = \bar{\mu}R(z, H)^T S_H(z)(e_n \otimes e_m) \\
&= \bar{\mu}\sum_{i=1}^n\sum_{t=1}^m R_i^T K_i e_m - \bar{\mu}\sum_{j=1}^n\left(\sum_{t=1}^m \lambda_{jt}\right)\sum_{i=1}^n q_{ij}R_i^T K_i e_m \\
&= \bar{\mu}\sum_{i=1}^n\sum_{t=1}^m \lambda_{it}(G_{it} \otimes X_{it}) - \bar{\mu}\sum_{j=1}^n\left(\sum_{t=1}^m \lambda_{jt}\right)\sum_{i=1}^n q_{ij}\sum_{t=1}^m \lambda_{it}(G_{it} \otimes X_{it}) \\
&= n\bar{\mu}\left[\sum_{i=1}^n\left(\sum_{t=1}^m \lambda_{it}\right)^{-1}\right]^{-1}\sum_{i=1}^n\sum_{t=1}^m \varpi_{it}(G_{it} \otimes X_{it})
\end{aligned}$$

and we obtain $n^{-1}|H|^{-1}C_n = \bar{\mu}O_p(|H|\ln(\ln\ n))$ by (a) Lemma A.1, (b) for all $l = 1, \ldots, q, k((Z_{it,l} - z_l)/h) = 0$ if $|Z_{it,l} - z_l| > h$ by Assumption 3, (c) $\varpi_{it} \le 1$ and (d) $E||X_{it}||^{1+\delta} < M < \infty$ for some $\delta > 0$ by Assumption 1. Since $\mu_i \sim$ i.i.d.$(0, \sigma_\mu^2)$, we have $\bar{\mu} = O_p(n^{-1/2})$. It follows that $n^{-1}|H|^{-1}C_n = O_p(n^{-1/2}|H|\ln(\ln\ n))$.

Lemma A.5. Under Assumptions 1–3, we have

$$\begin{aligned}
&n^{-1}|H|^{-1}R(z, H)^T S_H^2(z)R(z, H) \\
&\quad \approx |H|^{-1}\sum_{t=1}^m E[(1 - \varpi_{it})^2\lambda_{it}^2[\cdot]_{it} \otimes (X_{it}X_{it}^T)]
\end{aligned}$$

Proof. Simple calculations give

$$\begin{aligned} D_n &= R(z,H)^T S_H^2(z) R(z,H) = R(z,H)^T W_H(z) M_H(z) W_H(z)^T R(z,H) \\ &= \sum_{i=1}^{n} R_i(z,H)^T K_H^2(Z_i,z) R_i(z,H) \\ &\quad - 2\sum_{j=1}^{n}\sum_{i=1}^{n} q_{ji} R_j(z,H)^T K_H^2(Z_j,z) e_m e_m^T K_H(Z_i,z) R_i(z,H) \\ &\quad + \sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} q_{ij} R_{ji'} R_i(z,H)^T K_H(Z_j,z) \\ &\quad \times e_m e_m^T K_H^2(Z_i,z) e_m e_m^T K_H(Z_{i'},z) R_{i'}(z,H) \\ &= D_{n1} - 2D_{n2} + D_{n3} \end{aligned}$$

Using the same method in the proof of Lemma A.2, we show $D_n \approx \sum_{i=1}^{n}\sum_{t=1}^{m}(1-\varpi_{it})^2 \lambda_{it}^2 [\cdot]_{it,it} \otimes (X_{it} X_{it}^T)$. It is easy to show that $n^{-1}|H|^{-1} D_{n1} = n^{-1}|H|^{-1}\sum_{i=1}^{n}\sum_{t=1}^{m}\lambda_{it}^2[\cdot]_{it,it} \otimes (X_{it}X_{it}^T) = \sum_{t=1}^{m} E[S_{t,2,1} \otimes (X_{it}\, X_{it}^T)] + O_p(||H||^2) + O_p(n^{-1/2}|H|^{-1/2})$.

Also, we obtain $n^{-1}|H|^{-1}\sum_{i=1}^{n}\sum_{t=1}^{m}(1-\varpi_{it})^2\lambda_{it}^2[\cdot]_{it,it} \otimes (X_{it}X_{it}^T) = \kappa(z) + O_p(n^{-1/2}|H|^{-1/2})$, where $\kappa(z) = |H|^{-1}\sum_{t=1}^{m} E\big[(1-\varpi_{it})^2\lambda_{it}^2[\cdot]_{it,it} \otimes (X_{it}Z_{it}^T)\big] \le |H|^{-1}\sum_{t=1}^{m} E\big[\lambda_{it}^2 ||[\cdot]_{it,it} \otimes (X_{it}\, X_{it}^T)||\big] \le M < \infty$ for all i and t.

The four lemmas above are enough to give the result of Theorem 1. Moreover, applying Liaponuov's CLT will give the result of Theorem 2. Since the proof is a rather standard procedure, we drop the details for compactness of the paper.

Technical Sketch: Random Effects Estimator

The RE estimator $\hat{\theta}_{\mathrm{RE}}(\cdot)$ is the solution to the following optimization problem:

$$\min_{\beta(z)} [Y - R(z,H)\mathrm{vec}(\beta(z))]^T W_H(z) [Y - R(z,H)\mathrm{vec}(\beta(z))]$$

that is, we have

$$\begin{aligned} &\mathrm{vec}(\hat{\beta}_{\mathrm{RE}}(z)) \\ &\quad = [R(z,H)^T W_H(z) R(z,H)]^{-1} R(z,H)^T W_H(z) Y \\ &\quad = \mathrm{vec}(\beta(z)) + [R(z,H)^T W_H(z) R(z,H)]^{-1} (\tilde{A}_n/2 + \tilde{B}_n + \tilde{C}_n) \end{aligned}$$

where $\tilde{A}_n = R(z,H)^T W_H(z) \prod(z,H)$, $\tilde{B}_n = R(z,H)^T W_H(z) D_0 \mu_0$, and $\tilde{C}_n = R(z,H)^T W_H(z) V$. Its asymptotic properties are as follows.

Lemma A.6. Under Assumptions 1–3, and $E(X_{it}X_{it}^T|z)$ and $E(\mu_i X_{it}|z)$ have continuous second-order derivative at $z \in R^q$. Also, $\sqrt{n|H|}||H||^2 = O(1)$ as $n \to \infty$, and $E(|v_{it}|^{2+\delta}) < \infty$ and $E(|\mu_i|^{2+\delta}) < M < \infty$ for all i and t and for some $\delta > 0$, we have under H_0

$$\sqrt{n|H|}\left(\hat{\theta}_{\mathrm{RE}}(z) - \theta(z) - \kappa_2 \, \Theta_H \frac{(z)}{2}\right) \xrightarrow{d} N\left(0, \sum_{\theta(Z),\mathrm{RE}}\right) \tag{A.9}$$

where $\kappa_2 = \int k(v) v^2 dv$, $\sum_{\theta(z),\mathrm{RE}} = (\sigma_\mu^2 + \sigma_v^2) \, \Phi \, (z)^{-1} \int K^2(u) du$ and $\Phi(z) = \sum_{t=1}^m f_t(z) E(X_{1t} X_{1t}^T | z)$. Under H_1, we have

$$\begin{aligned} \mathrm{Bias}(\hat{\theta}_{\mathrm{RE}}(z)) &= \Phi(z)^{-1} \sum_{t=1}^m f_t(z) E(\mu_1 X_{1t}|z) + o(1) \\ \mathrm{Var}(\hat{\theta}_{\mathrm{RE}}(z)) &= n^{-1}|H|^{-1}\sigma_v^2 \, \Phi \, (z)^{-1} \int K^2(u) du \end{aligned} \tag{A.10}$$

where $\Theta_H(\mathrm{z})$ is given in the proof of Lemma A.3.

Proof of Lemma A.6. First, we have the following decomposition:

$$\begin{aligned} \sqrt{n|H|}\left[\hat{\theta}_{\mathrm{RE}}(z) - \theta(z)\right] &= \sqrt{n|H|}[\hat{\theta}_{\mathrm{RE}}(z) - E(\hat{\theta}_{\mathrm{RE}}(z))] \\ &\quad + \sqrt{n|H|}[E(\hat{\theta}_{\mathrm{RE}}(z)) - \theta(z)] \end{aligned}$$

where we can show that the first term converges to a normal distribution with mean zero by Liaponuov's CLT (the details are dropped since it is a rather standard proof), and the second term contributes to the asymptotic bias. Since it will cause no notational confusion, we drop the subscription 'RE'. Below, we use $\mathrm{Bias}_i\{\hat{\theta}(z)\}$ and $\mathrm{Var}_i\{\hat{\theta}(z)\}$ to denote the respective bias and variance of $\hat{\theta}_{\mathrm{RE}}(z)$ under H_0 if $i = 0$ and under H_1 if $i = 1$.

First, under H_0, the bias and variance of $\hat{\theta}(z)$ are as follows: $\mathrm{Bias}_0\{\hat{\theta}(z)|\{(X_{it}, Z_{it})\}\} = S_p[R(z,H)^T W_H(z) R(z,H)]^{-1} R(z,H)^T W_H(z) \prod (z,H)/2$ and

$$\begin{aligned} &\mathrm{Var}_0\{\hat{\theta}(z)|\{(X_{it}, Z_{it})\}\} \\ &\quad = S_p[R(z,H)^T W_H(z) R(z,H)]^{-1}[R(z,H)^T W_H(z) \mathrm{Var}(UU^T) W_H(z) R(z,H)] \\ &\qquad \times [R(z,H)^T W_H(z) R(z,H)]^{-1} S_p^T \end{aligned}$$

It is simple to show that $\mathrm{Var}(UU^T) = \sigma_\mu^2 I_n \otimes (e_m e_m^T) + \sigma_v^2 I_{n \times m}$.

Next, under H_1, we notice that $\text{Bias}_1\{\hat{\theta}(z)|\{(X_{it}, Z_{it})\}\}$ is the sum of $\text{Bias}_0\{\hat{\theta}(z)|\{(X_{it}, Z_{it})\}\}$ plus an additional term $S_p[R(z,H)^T W_H(z) R(z,H)]^{-1} R(z,H)^T W_H(z) D_0 \mu_0$, and that

$$\begin{aligned}
&\text{Var}_1\{\hat{\theta}(z)|\{(X_{it}, Z_{it})\}\} \\
&\quad = \sigma_v^2 S_p[R(z,H)^T W_H(z) R(z,H)]^{-1}[R(z,H)^T W_H(z)^2 R(z,H)] \\
&\qquad \times [R(z,H)^T W_H(z) R(z,H)]^{-1} S_p^T
\end{aligned}$$

Noting that $R(z,H)^T W_H(z) R(z,H)$ is A_{n1} in Lemma A.2 and that $R(z,H)^T W_H(z) \prod(z,H)$ is B_{n1} in Lemma A.3, we have

$$\text{Bias}_0\{\hat{\theta}(z)\} = \kappa_2 \Theta_H \frac{(z)}{2} + o(||H||^2) \tag{A.11}$$

In addition, under Assumptions 1–3, and $E(|\mu_i|^{2+\delta}) < M < \infty$ and $E(||X_{it}||^{2+\delta}) < M < \infty$ for all i and t and for some $\delta > 0$, we show that

$$\begin{aligned}
&n^{-1}|H|^{-1} S_p R(z,H)^T W_H(z) D_0 \mu_0 \\
&\quad = n^{-1}|H|^{-1} S_p \sum_{i=1}^{n} \mu_i \sum_{t=1}^{m} \lambda_{it}(G_{it} \otimes X_{it}) \\
&\quad = \sum_{t=1}^{m} f_t(z) E(\mu_1 X_{1t}|z) + O_p(||H||^2) + O_p((n|H|)^{1/2})
\end{aligned} \tag{A.12}$$

which is a non-zero constant plus a term of $o_p(1)$ under H_1. Combining Eqs. (A.11) and (A.12), we obtain Eq. (A.10). Hence, under H_1, the bias of the RE estimator will not vanish as $n \to \infty$, and this leads to the inconsistency of the RE estimator under H_1.

As for the asymptotic variance, we can easily show that under H_0

$$\text{Var}_0\{\hat{\theta}(Z)\} = n^{-1}|H|^{-1}(\sigma_\mu^2 + \sigma_v^2)\Phi(z)^{-1} \int K^2(u) du \tag{A.13}$$

and under $H_1, \text{Var}_1\{\hat{\theta}(z)\} = n^{-1}|H|^{-1}\sigma_v^2 \Phi(z)^{-1} \int K^2(u)\, du$, where we have recognized that $R(z,H)^T W_H(z)^2 R(z,H)$ is D_{n1} in Lemma A.5, and $(\sigma_\mu^2 + \sigma_v^2)$ $R(z,H)^T W_H(z)^2 R(z,H)$ is the leading term of $R(z,H)^T W_H(z)$ $\text{Var}(UU^T)$ $W_H(z) R(z,H)$.

Proof of Theorem 3

Define $\Delta_i = (\Delta_{i1}, \ldots, \Delta_{im})^T$ with $\Delta_{it} = X_{it}^T(\theta(Z_{it}) - \hat{\theta}_{\mathrm{RE}}(Z_{it}))$. Since $M_D D_0 = 0$, we can decompose the proposed statistic into three terms

$$\begin{aligned}
\hat{T}_n &= \frac{1}{n^2|H|}\sum_{i=1}^{n}\sum_{j\neq i}\hat{U}_i^T Q_m A_{i,j} Q_m \hat{U}_j \\
&= \frac{1}{n^2|H|}\sum_{i=1}^{n}\sum_{j\neq i}\Delta_i^T Q_m A_{i,j} Q_m \Delta_j + \frac{2}{n^2|H|}\sum_{i=1}^{n}\sum_{j\neq i}\Delta_i^T Q_m A_{i,j} Q_m V_j \\
&\quad + \frac{1}{n^2|H|}\sum_{i=1}^{n}\sum_{j\neq i} V_i^T Q_m A_{i,j} Q_m V_j \\
&= T_{n1} + 2T_{n2} + T_{n3}
\end{aligned}$$

where $V_i = (v_{i1}, \ldots, v_{im})^T$ is the $m \times 1$ error vector. Since $\hat{\theta}_{\mathrm{RE}}(Z_{it})$ does not depend on the jth unit observation and $\hat{\theta}_{\mathrm{RE}}(Z_{jt})$ does not depend on the ith unit observation for a pair of (i, j), it is easy to see that $E(T_{n2}) = 0$. The proofs fall into the standard procedures seen in the literature of nonparametric tests. We therefore give a very brief proof below.

First, applying Hall's (1984) CLT, we can show that under both H_0 and H_1

$$n\sqrt{|H|}T_{n3} \xrightarrow{d} N(0, \sigma_0^2) \tag{A.14}$$

by defining $H_n(\chi_i, \chi_j) = V_i^T Q_m A_{i,j} Q_m V_j$ with $\chi_i = (X_i, Z_i, V_i)$, which is a symmetric, centred and degenerate variable. We are able to show that

$$\frac{E[G_n^2(\chi_1, \chi_2)] + n^{-1}E[H_n^4(\chi_1, \chi_2)]}{\{E[H_n^2(\chi_1, \chi_2)]\}^2} = \frac{O(|H|^3) + O(n^{-1}|H|)}{O(|H|^2)} \to 0$$

if $|H| \to 0$ and $n|H| \to \infty$ as $n \to \infty$, where $G_n(\chi_1, \chi_2) = E_{\chi i}[H_n(\chi_1, \chi_i) H_n(\chi_2, \chi_i)]$. In addition

$$\begin{aligned}
\mathrm{var}(n\sqrt{|H|}T_{n3}) &= 2|H|^{-1}E(H_n^2(\chi_1, \chi_2)) \\
&\approx 2(1 - m^{-1})^2\sigma_v^4 \sum_{t=1}^{m}\sum_{s=1}^{m}|H|^{-1}E[K_H^2(Z_{1s}, Z_{2t})(X_{1s}^T X_{2t})^2] \\
&= \sigma_0^2 + o(1)
\end{aligned}$$

Second, we can show that $n\sqrt{|H|}T_{n2} = O_p(||H||^2) + O_p(n^{-1/2}|H|^{-1/2})$ under H_0 and $n\sqrt{|H|}T_{n2} = O_p(1)$ under H_1. Moreover, we have, under H_0, $n\sqrt{|H|}T_{n1} = O_p(n\sqrt{|H|}\,||H||^4)$; under $H_1, n\sqrt{|H|}T_{n1} = O_p(n\sqrt{|H|})$.

Finally, to estimate σ_0^2 consistently under both H_0 and H_1, we replace the unknown V_i and V_j in T_{n3} by the estimated residual vectors from the FE estimator. Simple calculations show that the typical element of $\hat{V}_i Q_m$ is $\tilde{\hat{v}}_{it} = y_{it} - X_{it}^T \hat{\theta}_{\mathrm{FE}}(Z_{it}) - v_{it} \quad -(\bar{y}_i - m^{-1}\sum_{t=1}^m X_{it}^T \hat{\theta}_{\mathrm{FE}}(Z_{it}) - \bar{v}_i) = \tilde{\Delta}_{it} - (v_{it} - \bar{v}_i)$, where $\tilde{\Delta}_{it} = X_{it}^T(\theta(Z_{it}) - \hat{\theta}_{\mathrm{FE}}(Z_{it})) - m^{-1}\sum_{t=1}^m X_{it}^T(\theta(Z_{it}) - \hat{\theta}_{\mathrm{FE}}(Z_{it})) = \sum_{l=1}^m q_{lt} X_{il}^T(\theta(Z_{il}) - \hat{\theta}_{\mathrm{FE}}(Z_{il}))$ with $q_{tt} = 1 - 1/m$ and $q_{lt} = -1/m$ for $l \neq t$. The leave-two-unit-out FE estimator does not use the observations from the ith and jth units for a pair (i, j), and this leads to $E(\hat{V}_i^T Q_m A_{i,j} Q_m \hat{V}_j)^2 \approx \sum_{t=1}^m \sum_{s=1}^m E\,[K_H^2(Z_{it,} Z_{js})(X_{it}^T X_{js})^2(\tilde{\Delta}_{it}^2 \tilde{\Delta}_{it}^2 + \tilde{\Delta}_{it}^2 \tilde{v}_{js}^2 + \tilde{\Delta}_{js}^2 \tilde{v}_{it}^2 + \tilde{v}_{it}^2 \tilde{v}_{js}^2)] \approx \sum_{t=1}^m \sum_{s=1}^m E[K_H^2(Z_{it}, Z_{js})(X_{it}^T X_{js})^2 \tilde{v}_{it}^2 \tilde{v}_{js}^2$ where $\tilde{v}_{it} = v_{it} - \bar{v}_i$ and $\bar{v}_i = m^{-1}\sum_{t=1}^m v_{it}$.

FUNCTIONAL COEFFICIENT ESTIMATION WITH BOTH CATEGORICAL AND CONTINUOUS DATA

Liangjun Su, Ye Chen and Aman Ullah

ABSTRACT

We propose a local linear functional coefficient estimator that admits a mix of discrete and continuous data for stationary time series. Under weak conditions our estimator is asymptotically normally distributed. A small set of simulation studies is carried out to illustrate the finite sample performance of our estimator. As an application, we estimate a wage determination function that explicitly allows the return to education to depend on other variables. We find evidence of the complex interacting patterns among the regressors in the wage equation, such as increasing returns to education when experience is very low, high return to education for workers with several years of experience, and diminishing returns to education when experience is high. Compared with the commonly used parametric and semiparametric methods, our estimator performs better in both goodness-of-fit and in yielding economically interesting interpretation.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 131–167

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025007

1. INTRODUCTION

In this paper, we extend the work of Racine and Li (2004) to estimating functional coefficient models with both continuous and categorical data:

$$Y = \sum_{j=1}^{d} a_j(U)X_j + \varepsilon \tag{1}$$

where ε is the disturbance term, X_j a scalar random variable, U a $(p+q) \times 1$ random vector, and $a_j(\cdot)$, $j = 1, \ldots, d$ are unknown smooth functions. As Cai, Fan, and Yao (2000) remark, the idea for this kind of model is not new, but the potential of this modeling techniques had not been fully explored until the seminal work of Cleveland, Grosse, and Shyu (1992), Chen and Tsay (1993), and Hastie and Tibshirani (1993), in which nonparametric techniques were proposed to estimate the unknown functions $a_j(\cdot)$. An important feature of these early works is to assume that the random variable U is continuous, which limits the model's potential applications.

Drawing upon the work of Aitchison and Aitken (1976) and Racine and Li (2004) propose a novel approach to estimate nonparametric regression mean functions with both categorical and continuous data in the i.i.d. setup. They apply their new estimation method to some publicly available data and demonstrate the superb performance of their estimators in comparison with some traditional ones.

In this paper, we consider extending the work of Racine and Li (2004) to the estimation of the functional coefficient model (1) when U contains both continuous and categorical variables. This is important since categorical variables may be present in the functional coefficients. For example, in the study of the output functions for individual firms, firms that belong to different industries may exhibit different output elasticities with respect to labor and capital. So we should allow the categorical variable "industry" to enter U. We will demonstrate that this modeling strategy outperforms the traditional dummy-variable approach widely used in the literature. For the same reason, Li and Racine (2008b) consider a local constant estimation of model (1) by assuming the data are identically and independently distributed (i.i.d.).

Another distinguishing feature of our approach is that we allow for weak data dependence. One of the key applications of nonparametric function estimation is the construction of prediction intervals for stationary time series. The i.i.d. setup of Racine and Li (2004) and Li and Racine (2008b) cannot meet this purpose.

To demonstrate the usefulness of our proposed estimator in empirical applications, we estimate a wage determination equation based on recent CPS data. While in the literature of labor economics, the return to education has already been extensively investigated from various aspects, in this paper, we explicitly allow the return to education to be dependent on other variables, both continuous and discrete, including experience, gender, age, industry, and so forth. Our findings are clearly against the parametric functional form assumption of the most widely used linear separable Mincerian equation, and the return to education does vary substantially with the other regressors. Therefore, our model can help to uncover economically interesting interacting effects among the regressors, and so should have high potential for applications.

The paper is structured as follows. In Section 2, we introduce our functional coefficient estimators and their asymptotic properties. We conduct a small set of Monte Carlo studies to check the relative performance of the proposed estimator in Section 3. Section 4 provides empirical data analysis. Final remarks are contained in Section 5. All technical details are relegated to the appendix.

2. FUNCTIONAL COEFFICIENT ESTIMATION WITH MIXED DATA

2.1. Local Linear Estimator

In this paper, we study estimation of model (1) when U is comprised of a mix of discrete and continuous variables. Let $\{(Y_i, X_i, U_i), i = 1, 2, \ldots\}$ be jointly strictly stationary processes, where (Y_i, X_i, U_i) has the same distribution as (Y, X, U). Let $U_i = (U_i^{c\prime}, U_i^{d\prime})'$, where U_i^c and U_i^d denote a $p \times 1$ vector of continuous regressors and a $q \times 1$ vector of discrete regressors, respectively, $p \geq 1$, and $q \geq 1$. Like Racine and Li (2004), we will use U_{it}^d to denote the tth component of U_i^d, and assume that U_{it}^d can take $c_t \geq 2$ different values, that is, $U_{it}^d \in \{0, 1, \ldots, c_t - 1\}$ for $t = 1, \ldots, q$. Denote $u = (u^c, u^d) \in \mathbb{R}^p \times \mathcal{D}$. We use $f_u(u) = f(u^c, u^d)$ to denote the joint density function of (U_i^c, U_i^d) and $\mathcal{D} = \prod_{t=1}^{q} \{0, 1, \ldots, c_t - 1\}$ to denote the range assumed by U_i^d. With a little abuse of notation, we also use $\{(Y_i, X_i, U_i), i = 1, \ldots, n\}$ to denote the data.

To define the kernel weight function, we focus on the case for which there is no natural ordering in U_i^d. Define

$$l(U_{it}^d, u_t^d, \lambda_t) = \begin{cases} 1 & \text{if } U_{it}^d = u_t^d, \\ \lambda_t & \text{if } U_{it}^d \neq u_t^d, \end{cases} \tag{2}$$

where λ_t is a bandwidth that lies on the interval [0, 1]. Clearly, when $\lambda_t = 0$, $l(U_{it}^{\mathrm{d}}, u_t^{\mathrm{d}}, 0)$ becomes an indicator function, and $\lambda_t = 1$, $l(U_{it}^{\mathrm{d}}, u_t^{\mathrm{d}}, 1)$ becomes an uniform weight function. We define the product kernel for the discrete random variables by:

$$L(U_i^{\mathrm{d}}, u^{\mathrm{d}}, \lambda) = \prod_{t=1}^{q} l(U_{it}^{\mathrm{d}}, u_t^{\mathrm{d}}, \lambda_t) \tag{3}$$

For the continuous random variables, we use $w(\cdot)$ to denote a univariate kernel function and define the product kernel function by $W_{h,iu} = \prod_{t=1}^{p} w((U_{it}^{\mathrm{c}} - u_t^{\mathrm{c}})/h_t)$, where $h = (h_1, \ldots, h_p)$ denotes the smoothing parameters and $U_{it}^{\mathrm{c}}(u_t^{\mathrm{c}})$ is the tth component of $U_i^{\mathrm{c}}(u_t^{\mathrm{c}})$. We then define the kernel weight function K_{iu} by:

$$K_{iu} = L_{\lambda,iu} W_{h,iu} \tag{4}$$

where $L_{\lambda,iu} = L(U_i^{\mathrm{d}}, u^{\mathrm{d}}, \lambda)$.

We now estimate the unknown functional coefficient functions in model (1) by using a local linear regression technique. Suppose that $a_j(\cdot)$ assumes a second-order derivative. Denote by $\dot{a}_j(u) = \partial a_j(u)/\partial u^{\mathrm{c}}$ the $p \times 1$ first-order derivative of $a_j(u)$ with respect to its continuous-valued argument u^{c}. Denote by $\ddot{a}_j(u) = \partial^2 a_j(u)/(\partial u^{\mathrm{c}} \partial u^{\mathrm{c}\prime})$ second-order derivative matrix of $a_j(u)$ with respect to u^{c}. We use $a_{j,ss}(u)$ to denote the sth diagonal element of $\ddot{a}_j(u)$.

For any given u and $\tilde{u}$ in a neighborhood of u, it follows from a first-order Taylor expansion that

$$a_j(\tilde{u}) \approx a_j(u) + \dot{a}_j(u)'(\tilde{u}^{\mathrm{c}} - u^{\mathrm{c}}) \tag{5}$$

for u^{c} in a neighborhood of $\tilde{u}^{\mathrm{c}}$ and $\tilde{u}^{\mathrm{d}} = u^{\mathrm{d}}$. To estimate $\{a_j(u)\}$ (and $\{\dot{a}_j(u)\}$), we choose $\{a_j\}$ and $\{b_j\}$ to minimize

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{\mathrm{d}} \{a_j + b_j'(U_i - u)\} X_{ij} \right]^2 K_{iu} \tag{6}$$

Let $\{(\hat{a}_j, \hat{b}_j)\}$ be the local linear estimator. Then the local linear regression estimator for the functional coefficient is given by

$$\hat{a}_j(u) = \hat{a}_j, \quad j = 1, \ldots, d \tag{7}$$

The local linear regression estimator for the functional coefficient can be easily obtained. To do so, let $\boldsymbol{e}_{j,d(p+1)}$ be the $d(1+p) \times 1$ unit vector of with 1 at the jth position and 0 elsewhere. Let $\tilde{\mathbf{X}}$ denote an $n \times d(1+p)$ matrix with

$$\tilde{X}_i = (X_i', X_i' \otimes (U_i - u)')$$

as its ith row. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. Set $\boldsymbol{W} = \text{diag}\{K_{1u}, \ldots, K_{nu}\}$. Then Eq. (6) can be written as

$$(\boldsymbol{Y} - \tilde{\boldsymbol{X}}\theta)'\boldsymbol{W}(\boldsymbol{Y} - \tilde{\boldsymbol{X}}\theta)$$

where $\theta = (a_1, \ldots, a_d, b_1', \ldots, b_d')'$. So the local linear estimator is simply

$$\hat{\theta} = \hat{\theta}(u) = (\tilde{\boldsymbol{X}}'\boldsymbol{W}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{W}\boldsymbol{Y} \tag{8}$$

which entails that

$$\hat{a}_j = \hat{a}_j(u) = \mathbf{e}_{j,d(1+p)}'\hat{\theta}, \quad j = 1, \ldots, d \tag{9}$$

Let $\theta(u) = (a_1(u), \ldots, a_d(u), \dot{a}_1(u)', \ldots, \dot{a}_d(u)')'$. We will study the asymptotic properties of $\hat{\theta}(u)$.

2.2. Assumptions

To facilitate the presentation, let $\Omega(u) = E(X_i X_i' | U_i = u)$, $\sigma^2(u, x) = E[\varepsilon_i^2 | U_i = u, X_i = x]$, $\Omega^*(u) = E[X_i X_i' \sigma^2(U_i, X_i) | U_i = u]$. Let $f(u, x)$ denote the joint density of (U_i, X_i) and $f_u(u)$ be the marginal density of U_i. Also, let $f_{u|x}(u|x)$ be the conditional density of U_i given $X_i = x$. Let $f_i(u, \tilde{u}|x, \tilde{x})$ be the conditional density of (U_1, U_i) given $(X_1, X_i) = (x, \tilde{x})$.

We now list the assumptions that will be used to establish the asymptotic distribution of our estimator.

Assumption A1.

(i) The process $\{(Y_i, U_i, X_i), i \geq 1\}$ is a strictly stationary α-mixing process with coefficients $\alpha(n)$ satisfying $\sum_{j \geq 1} j^c [\alpha(j)]^{\gamma/(2+\gamma)} < \infty$ for some $\gamma > 0$ and $c > \gamma/(2+\gamma)$.

(ii) $f_{u|x}(u|x) \leq M < \infty$ and $f_i(u, \tilde{u}|x, \tilde{x}) \leq M < \infty$ for all $i \geq 2$ and $u, \tilde{u}, x, \tilde{x}$.

(iii) $\Omega^*(u)$ and $\Omega(u)$ are positive definite.

(iv) The functions $f_u(\cdot, u^d)$, $\sigma^2(\cdot, u^d, x)$, $\Omega(\cdot, u^d)$, and $\Omega^*(\cdot, u^d)$ are continuous for all $u^d \in \mathcal{D}$, and $f_u(u) > 0$.

(v) $a_j(\cdot, u^d)$ has continuous second derivatives for all $u^d \in \mathcal{D}$.

(vi) $E||X||^{2(2+\gamma)} < \infty$, where $||\cdot||$ is the Euclidean norm and γ is given in (i).

(vii) $E[Y_1^2 + Y_i^2 | (U_1, X_1) = (u, x); (U_i, X_i) = (\tilde{u}, \tilde{x})] \leq M < \infty$.

(viii) There exists $\delta > (2+\gamma)$ such that $E[|Y_1|^\delta | (U_1, X_1) = (\tilde{u}, x)] \leq M < \infty$ for all $x \in \mathbb{R}^d$ and all $\tilde{u}$ in the neighborhood of u. $\alpha(j) = O(j^{-\kappa})$, where $\kappa \geq (2+\gamma)\delta/\{2(\delta - 2 - \gamma)\}$.

(ix) There exists a sequence of positive integers s_n such that $s_n \to \infty$, $s_n = o((nh_1 \ldots h_p)^{1/2})$, and $n^{1/2}(h_1 \ldots h_2)^{-1/2}\alpha(s_n) \to 0$.

Assumption A2. The kernel function $w(\cdot)$ is a density function that is symmetric, bounded, and compactly supported.

Assumption A3. As $n \to 0$, the bandwidth sequences $h_s \to 0$ for $s = 1, \ldots, p$, $\lambda_s \to 0$ for $s = 1, \ldots, q$, and (i) $nh_1 \ldots hp \to \infty$, (ii) $(nh_1 \ldots h_p)^{1/2}$ $(||h||^2 + ||\lambda||) = O(1)$.

Assumptions A1–A2 are similar to Conditions A and B in Cai et al. (2000) except that we consider mixed regressors. Assumptions A1(i) is standard in the nonparametric regression for time series. See, for example, Cai et al. (2000) and Cai and Ould-Saïd (2003). It is satisfied by many well-known processes such as linear stationary ARMA processes and a large class of processes implied by numerous nonlinear models, including bilinear, nonlinear autoregressive (NLAR), and ARCH-type models (see Fan & Li, 1999). As Hall, Wolf, and Yao (1999) and Cai and Ould-Saïd (2003) remark, the requirement in Assumption A2 that $w(\cdot)$ is compactly supported can be removed at the cost of lengthier arguments used in the proofs, and in particular, Gaussian kernel is allowed.

Assumption A3 is standard for nonparametric regression with mixed data (see Li & Racine, 2008a).

2.3. Asymptotic Theory for the Local Linear Estimator

To introduce our main results, let $\mu_{s,t} = \int_{\mathbb{R}} v^s w(v)^t dv$, s, $t = 0$, 1, 2. Define two $d(1+p) \times d(1+p)$ diagonal matrices $S = S(u)$ and $\Gamma = \Gamma(u)$ by:

$$S = f_u(u)\begin{pmatrix} \Omega(u) & 0'_{dp\times d} \\ 0_{dp\times d} & \mu_{2,1}\Omega(u) \otimes I_p \end{pmatrix}, \quad \Gamma = f_u(u)\begin{pmatrix} \mu_{0,2}^p \Omega^*(u) & 0'_{dp\times d} \\ 0_{dp\times d} & \mu_{2,2}\Omega^*(u) \otimes I_p \end{pmatrix}$$

where $0_{l\times k}$ is an $l \times k$ matrix of zeros, I_p the $p \times p$ identity matrix, and $\otimes$ the Kronecker product. For any $p \times 1$ vectors $c = (c_1, \ldots, c_p)'$ and $d = (d_1, \ldots, d_p)'$, let $c \odot d \equiv (c_1 d_1, \ldots, c_p d_p)'$.

To describe the leading bias term associated with the discrete random variables, we define

$$I_s(u^{\mathrm{d}}, \tilde{u}^{\mathrm{d}}) = \mathbf{1}(u_s^{\mathrm{d}} \neq \tilde{u}_s^{\mathrm{d}}) \prod_{t\neq s}^{q} \mathbf{1}(u_t^{\mathrm{d}} = \tilde{u}_t^{\mathrm{d}})$$

where $\mathbf{1}(\cdot)$ is the usual indicator function. That is, $I_s(u^{\mathrm{d}}, \tilde{u}^{\mathrm{d}})$ is one if and only u^{d} and $\tilde{u}^{\mathrm{d}}$ differ only in the sth component and is

zero otherwise. Let

$$b(h,\lambda) = H\left\{ \begin{pmatrix} \frac{1}{2}\mu_{2,1} f_u(u)\Omega(u)A \\ 0_{dp\times 1} \end{pmatrix} + \sum_{\tilde{u}^d\in\mathcal{D}} \sum_{s=1}^{q} \lambda_s I_s(u^{\mathrm{d}},\tilde{u}^{\mathrm{d}}) f_u(u^{\mathrm{c}},\tilde{u}^{\mathrm{d}}) \begin{pmatrix} \Omega(u^{\mathrm{c}},\tilde{u}^{\mathrm{d}})(\boldsymbol{a}(u^{\mathrm{c}},\tilde{u}^{\mathrm{d}}) - \boldsymbol{a}(u)) \\ -\mu_{2,1}(\Omega(u^{\mathrm{c}},\tilde{u}^{\mathrm{d}}) \otimes I_p)\boldsymbol{b}(u) \end{pmatrix} \right\} \tag{10}$$

where $H = \sqrt{nh_1 \ldots h_p}$, $A = \left(\sum_{s=1}^{p} h_s^2 a_{1,ss}(u), \ldots, \sum_{s=1}^{p} h_s^2 a_{d,ss}(u)\right)'$, $\boldsymbol{a}(u) = (a_1(u), \ldots, a_d(u))'$, and $\boldsymbol{b}(u) = (\dot{a}_1(u)', \ldots, \dot{a}_d(u)')'$. Define

$$B_{j,1s}(u) = \tfrac{1}{2}\mu_{2,1} a_{j,ss}(u), \quad \text{and}$$
$$B_{j,2s}(u) = f_u(u)^{-1} \boldsymbol{e}_{j,d}' \Omega^{-1}(u) \sum_{\tilde{u}^d\in\mathcal{D}} I_s(u^{\mathrm{d}},\tilde{u}^{\mathrm{d}}) f(u^{\mathrm{c}},\tilde{u}^{\mathrm{d}}) \Omega(u^{\mathrm{c}},\tilde{u}^{\mathrm{d}})[\boldsymbol{a}(u^{\mathrm{c}},\tilde{u}^{\mathrm{d}}) - \boldsymbol{a}(u)]$$

Now we state our main theorem.

Theorem 1. Assume that Assumptions A1–A3 hold. Then for each u that is an interior point

$$HH_1(\widehat{\theta}(u) - \theta(u)) - S^{-1}b(h,\lambda) \xrightarrow{d} N(0, S^{-1}\Gamma S^{-1})$$

where $H_1 = \mathrm{diag}(1, \ldots, 1, h', \ldots, h')$ is a $d(p+1) \times 1$ diagonal matrix with d diagonal elements of 1 and d diagonal elements of h. In particular, for $j = 1, \ldots, d$,

$$\sqrt{nh_1 \ldots h_p}\left(\widehat{a}_j(u) - a_j(u) - \sum_{s=1}^{p} h_s^2 B_{j,1s}(u) - \sum_{s=1}^{q} \lambda_s B_{j,2s}(u)\right) \xrightarrow{d} N\left(0, \frac{\mu_{0,2}^p \boldsymbol{e}_{j,d}' \Omega^{-1}(u)\Omega^*(u)\Omega^{-1}(u)\boldsymbol{e}_{j,d}}{f_u(u)}\right)$$

Remark 1. Noting that S and Γ are both block diagonal matrices, we have asymptotic independence between the estimator of $\boldsymbol{a}(u)$ and that of $\boldsymbol{b}(u)$. Under Assumption A3, the asymptotic bias (Abias) of $\hat{a}_j$ is comprised of two components, $\sum_{s=1}^{p} h_s^2 B_{j,1s}(u)$ and $\sum_{s=1}^{q} \lambda_s B_{j,2s}(u)$, which are associated with the continuous and discrete variables in U_i, respectively. For statistical inference, one needs to estimate $f_u(u)$, $\Omega(u)$, and $\Omega^*(u)$. The procedure is standard and thus is omitted.

Remark 2. It is well known that the two main advantages of a local linear estimate over a local constant estimate are the simpler structure of Abias and the automatic boundary bias correction mechanism for the local linear estimate (see Fan & Gijbels, 1996). Our local linear estimator has the same asymptotic variance as the local constant estimator of Li and Racine (2008b). But the two estimators are different in bias. In our notation, the Abias of Li and Racine's local constant estimator $\hat{a}_j^{(lc)}(u)$ of $a_j(u)$ is given by:

$$\text{Abias}(\hat{a}_j^{(lc)}(u)) = \sum_{s=1}^{p} h_s^2 B_{j,1s}^{(lc)}(u) - \sum_{s=1}^{q} \lambda_s B_{j,2s}^{(lc)}(u)$$

where

$$B_{j,1s}^{(lc)}(u) = \mu_{2,1}\left\{\boldsymbol{e}_{j,d} f_u(u)^{-1}\Omega^{-1}(u)[f_u(u)\Omega_s(u) + \Omega(u) f_{u,s}(u)]\boldsymbol{a}_s(u) + \tfrac{1}{2}a_{j,ss}(u)\right\}$$
$$B_{j,2s}^{(lc)}(u) = B_{j,2s}(u)$$

$\Omega_s(u)$ denotes the first-order partial derivative of $\Omega(u^c, u^d)$ with respect to the sth element in u^c, and $f_{u,s}(u)$ and $\boldsymbol{a}_s(u)$ are similarly defined. Clearly, the continuous element in $u = (u^c, u^d)$ causes the difference in the asymptotic biases of the two types of estimators.

To compare boundary behavior of the two estimators, we focus on the simplest case where there is only one continuous variable in $U_i = (U_i^{c\prime}, U_i^{d\prime})'$, that is, U_i^c is a scalar random variable and $p = 1$. Without loss of generality, we assume that the support of U_i^c is [0, 1]. In this case, we denote the bandwidth simply as $h \equiv h(n)$ and consider the left boundary point $u^c = vh$, where v is a finite positive constant. Following the literature, we assume that $f_u(0, u^d) \equiv \lim_{u^c \downarrow 0} f_u(u^c, u^d)$ exists and is strictly positive for all $u^d \in \mathcal{D}$. Define

$$S_v = \begin{pmatrix} \iota_{v0} & \iota_{v1} \\ \iota_{v1} & \iota_{v2} \end{pmatrix}, \quad \text{and} \quad \Gamma_v = \begin{pmatrix} \kappa_{v0} & \kappa_{v1} \\ \kappa_{v1} & \kappa_{v2} \end{pmatrix} \tag{11}$$

where $\iota_{vj} = \int_{-v}^{\infty} z^j w(z) dz$, and $\kappa_{vj} = \int_{-v}^{\infty} z^j w(z)^2 dz$ for $j = 0, 1$, and 2. Define

$$S(0, u^d; v) = S_v \otimes \Omega(0, u^d) f_u(0, u^d), \quad \text{and}$$
$$\Gamma(0, u^d; v) = \Gamma_v \otimes \Omega^*(0, u^d) f_u(0, u^d)$$

Define

$$b(h,\lambda;v) = H\left\{\frac{1}{2}\begin{pmatrix}\Omega(0,u^{\mathrm{d}})\bar{A}(0,u^{\mathrm{d}})\iota_{v2}\\ \Omega(0,u^{\mathrm{d}})\bar{A}\,(0,u^{\mathrm{d}})\iota_{v3}\end{pmatrix} f_{u}(0,u^{\mathrm{d}}) + \sum_{\tilde{u}^{\mathrm{d}}\in\mathcal{D}}\sum_{s=1}^{q}\lambda_s I_s(u^{\mathrm{d}},\tilde{u}^{\mathrm{d}}) f_{u}(0,\tilde{u}^{\mathrm{d}}) \times \begin{pmatrix}\Omega(0,\tilde{u}^{\mathrm{d}})\{\iota_{v0}[\boldsymbol{a}(0,\tilde{u}^{\mathrm{d}}) - \boldsymbol{a}(0,u^{\mathrm{d}})] - \iota_{v1}\boldsymbol{b}(0,u^{\mathrm{d}})\}\\ \Omega(0,\tilde{u}^{\mathrm{d}})\{\iota_{v1}[\boldsymbol{a}(0,\tilde{u}^{\mathrm{d}}) - \boldsymbol{a}(0,u^{\mathrm{d}})] - \iota_{v2}\boldsymbol{b}(0,u^{\mathrm{d}})\}\end{pmatrix}\right\}$$

where $\iota_{v3} = \int_{-v}^{\infty} z^3 w(z)dz$,

$$\bar{A}(0,u^{\mathrm{d}}) = (h^2 a_1''(0,u^{\mathrm{d}}),\ \ldots,\ h^2 a_d''(0,u^{\mathrm{d}}))' \tag{12}$$

and $a_s''(0,\ u^{\mathrm{d}})$ is the second-order derivative of $a_s(u^{\mathrm{c}},\ u^{\mathrm{d}})$ with respect to u^{c} evaluated at 0. The following corollary summarizes the asymptotic properties of $\widehat{\theta}(u) = \widehat{\theta}(u^{\mathrm{c}},u^{\mathrm{d}})$ for the case where $u^{\mathrm{c}} = vh$.

Corollary 1. Assume that Assumptions A1–A3 hold. If $p = 1$ and the support of U_i^{c} is [0, 1], then for any $u = (u^{\mathrm{c}},\ u^{\mathrm{d}})$ with $u^{\mathrm{c}} = vh$, we have

$$HH_1(\hat{\theta}(u) - \theta(u)) - S(0,u^{\mathrm{d}};v)^{-1}b(h,\lambda;v) \xrightarrow{d} N(0, S(0,u^{\mathrm{d}};v)^{-1}\Gamma(0,u^{\mathrm{d}};v)S(0,u^{\mathrm{d}};v)^{-1})$$

Remark 3. Clearly, for our local linear estimators the biases for the boundary points have the same order as those for the interior points. But the estimators of $\boldsymbol{a}(u)$ and $\boldsymbol{b}(u)$ are generally not asymptotically independent any more because neither $S(0,\ u^{\mathrm{d}};\ v)$, nor $\Gamma(0,\ u^{\mathrm{d}};\ v)$ is block diagonal. As a result, the Abias and variance formulae of $\widehat{a}_j(u)$ are not as simple as those in Theorem 1. Li and Racine (2008b) did not study the boundary behavior of the local constant estimator. Nevertheless, following the arguments used in the proof of the above corollary, we can readily show that their estimator has the same asymptotic variance as ours for boundary points but totally different bias formula. In our notation, the Abias of Li and Racine's local constant estimator $\widehat{\boldsymbol{a}}^{(lc)}(u^{\mathrm{c}},u^{\mathrm{d}})$ of $\boldsymbol{a}(u^{\mathrm{c}},\ u^{\mathrm{d}})$ with $u^{\mathrm{c}} = vh$ (after being scaled by H) is given by

$$\mathrm{Abias}(\widehat{\boldsymbol{a}}^{(lc)}(u^{\mathrm{c}},u^{\mathrm{d}})) = S^{(lc)}(0,u^{\mathrm{d}};v)^{-1}b^{(lc)}(h,\lambda;v)$$

where $S^{(lc)}(0, u^d; v) = \iota_{v0}\, \Omega(0, u^d)\, f_u(0, u^d)$,

$$b^{(lc)}(h, \lambda; v) = H\Bigg\{ f_u(0, u^d)\Omega(0, u^d)\bar{A}^{(lc)}(0, u^d)\iota_{v1} + \sum_{\tilde{u}^d \in \mathcal{D}} \sum_{s=1}^{q} \lambda_s I_s(u^d, \tilde{u}^d) f_u(0, \tilde{u}^d)\Omega(0, \tilde{u}^d)[\boldsymbol{a}(0, \tilde{u}^d) - \boldsymbol{a}(0, u^d)] \Bigg\}$$

and

$$\bar{A}^{(lc)}(0, u^d) = (h\dot{a}_1(0, \ u^d), \ldots, h\dot{a}_d(0, \ u^d))' \tag{13}$$

That is, the contribution of the continuous variable U_i^c to the Abias of the boundary estimator is of order $O(h)$, which is different from the order $O(h^2)$ for interior points. This is a reflection of the main disadvantage of local constant estimators over the local linear estimators.

2.4. Selection of Smoothing Parameters

In this subsection, we focus on how to choose the smoothing parameters to obtain the estimate $\widehat{a}_j$. It is well known that the choice of smoothing parameters is crucial in nonparametric kernel estimation.

Theorem 1. Implies that the leading term for the mean squared error (MSE) of $\widehat{a}_j$ is

$$\mathrm{MSE}(\widehat{a}_j) = \left[\sum_{s=1}^{p} h_s^2 B_{j,1s}(u) + \sum_{s=1}^{q} \lambda_s B_{j,2s}(u) \right]^2 + \frac{1}{nh_1 \ldots h_p} \frac{\mu_{0,2}^p \boldsymbol{e}_{j,d}' \Omega^{-1}(u)\Omega^*(u)\Omega^{-1}(u)\boldsymbol{e}_{j,d}}{f_u(u)}$$

By symmetry, all h_j should have the same order and all λ_s should also have the same order but with $\lambda_s - h_j^2$. By an argument similar to Li and Racine (2008a), it is easy to obtain the optimal rate of bandwidth in terms of minimizing a weighted integrated version of $\mathrm{MSE}(\widehat{a}_j)$. To be concrete, we should choose

$$h_j \sim n^{-1/(4+p)} \quad \text{and} \quad \lambda_j \sim n^{-2/(4+p)}$$

Nevertheless, the exact formula for the optimal smoothing parameters is difficult to obtain except for the simplest cases (e.g., p = 1 and q = 1). This also suggests that it is infeasible to use the plug-in bandwidth in applied setting since the plug-in method would first require the formula for each smoothing parameter and then pilot estimates for some unknown functions in the formula.

In practice, the key in estimating the functional coefficient model is the selection of bandwidth. We propose to use least squares cross-validation (LSCV) to choose the smoothing parameters. We choose (h, λ) to minimize the following LSCV criterion function

$$\mathrm{CV}(h,\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{\mathrm{d}} \widehat{a}_j^{(-i)}(U_i)X_{ij}\right)^2 M(U_i) \tag{14}$$

where $\widehat{a}_j^{(-i)}(U_i)$ is the leave-one-out functional coefficient estimator of $a_j(U_i)$ and $M(U_i)$ is a weight function that serves to avoid division by zero and perform trimming in areas of sparse support. In the following numerical study, we will set $M(U_i) = \Pi_{j=1}^{p} 1(|U_{ij}^{\mathrm{c}} - \bar{U}_j^{\mathrm{c}}| \le 2s_{U_j^{\mathrm{c}}})$, where $\boldsymbol{I}(\cdot)$ is the usual indicator function, and $\bar{U}_j^{\mathrm{c}}$ and $s_{U_j^{\mathrm{c}}}$ denote the sample mean and standard deviation of $\{U_{ij}^{\mathrm{c}}, 1 \le i \le n\}$, respectively. In practice, we can use grid search for (h, λ) when the dimensions of U^{c} and U^{d} are both small. Alternatively, one can apply the minimization function built in various software; but multiple starting values are recommended to reduce the chance of local solutions. In the following simulation study with $p = 1$ and $q = 2$, we try to save time in computation and use the latter method with only one starting value set according to the rule of thumb: $h_0 = S_{U^{\mathrm{c}}} n^{-1/5}$, $\lambda_j = 0.5 S_{U^{\mathrm{c}}} n^{-2/5}$ for $j = 1, 2$, where $S_{U^{\mathrm{c}}}$ is the standard deviation of the scalar random variable U_i^{c}. The performance of our nonparametric estimator is already reasonably well with this simple method.

Nevertheless, if the number of observations in application is large, it is extremely costly to apply the above LSCV method directly on all the observations. So we now propose an alternative way to do the LSCV. But the theoretical justification of this novel approach is beyond the scope of this paper. Let n denote the number of observations in the dataset, which could be as large as 17,446 in our empirical applications. When there is only one continuous variable in U (i.e., U^{c} is a scalar and $p = 1$), we propose the following approach to obtain the data-driven bandwidth:

Step 1. For $b = 1, 2, \ldots, B$, sample $m(\ll n)$ observations randomly from the dataset.

Step 2. Set $h = cS_{U^c}m^{-1/5}$ and $\lambda_j = c_j S_{U^c}m^{-2/5}$ for each and $j = 1, \ldots, q$, where c and c_j take values on [0.2, 4] with increments 0.2 and with the constraint $\lambda_j \leq 1$ satisfied, and S_{U^c} is the standard deviation of U_i^c based on the m observations in Step 1. Find the values of c and c_j that minimize the LSCV criterion function. Denote them as $c^{(b)}$ and $c_j^{(b)}$ for the bth resample.

Step 3. Calculate $\bar{c} = B^{-1}\sum_{b=1}^{B} c^{(b)}$ and $\bar{c}_j = B^{-1}\sum_{b=1}^{B} c_j^{(b)}, j = 1, \ldots, q$. Set $\widehat{h} = \bar{c}S_{U^c}n^{-1/5}$ and $\widehat{\lambda}_j = \bar{c}_j S_{U^c}n^{-2/5}$, where S_{U^c} is the standard deviation of U_i^c based on all n observations.

We will use $\widehat{h}$ and $\widehat{\lambda}_j, j = 1, \ldots, q$, in our empirical applications, where the single continuous variable U^c is *Experience* and U^d is composed of six categorical variables. We choose $m = 400$ and $B = 200$ below. When there are more than one continuous regressor in U, one can modify the above procedure correspondingly.

3. MONTE CARLO SIMULATIONS

We now conduct Monte Carlo experiment to illustrate the finite sample performance of our nonparametric functional coefficient estimators with mixed data. In addition to the proposed estimator, we also include several other parametric and nonparametric estimators.

The first data generating process (DGP) we consider is given by

$$\begin{aligned} Y_i = {} & 0.1(U_{i1}^2 + U_{i2} + U_{i3}) + 0.1(U_{i1}U_{i2} + U_{i3})X_{i1} \\ & + 0.15(U_{i1}U_{i2} + U_{i3})X_{i2} + \varepsilon_i \end{aligned}$$

where $X_{ij} \sim \text{Uniform}(0, 4)$ $(j = 1, 2)$, $U_{i1} \sim \text{Uniform}(0, 4)$, $U_{ij} \in \{0, 1, \ldots, 5\}$ with $P(U_{ij} = l) = 1/6$ for $l = 0, 1, \ldots, 5$ and $j = 2, 3$, and $\varepsilon_i \sim N(0, 1)$. Furthermore, X_{ij}, U_{ij}, and ε_i are i.i.d. and mutually independent.

We consider two nonparametric estimators and three parametric estimators for the conditional mean function $m(x, u) = E(Y_i|X_i = x, U_i = u)$. We first obtain our nonparametric functional coefficient estimator (NP) with mixed data where the smoothing parameters (h, λ) are chosen by the LSCV. Then we obtain the nonparametric frequency estimator (NP-FREQ) with mixed data by using the cross-validated h and setting $\lambda = 0$ (see Li & Racine, 2007, Chapter 3). It is expected that the smaller the ratio of the sample size to the number of "cells," the worse the nonparametric frequency approach relative to our proposed kernel approach.

For the parametric estimation, we consider in practice what an applied econometrician would do when he or she confronts the data $\{(Y_i, X_i, U_i), 1 \le i \le n\}$ and have a strong belief that all the variables in X_i and U_i can affect the dependent variable Y_i. In the first parametric model, we ignore the potential interaction between regressors and estimate a linear model without any interaction (LIN) by regressing Y_i on X_i, U_{i1}, and the two categorical variables U_{i2} and U_{i3}. In the second parametric model, we take into account potential interaction between X_i and U_{1i}, and estimate a linear model with interaction (LIN–INT1) by adding the interaction terms between X_i and U_{1i} into the LIN model. In the third parametric model, we also consider the interaction between X_i and (U_{2i}, U_{3i}), so we estimate a linear model with interaction (LIN–INT2) by adding the interaction terms between X_i and (U_{1i}, U_{2i}, U_{3i}) into the LIN–INT2 model. We expect LIN–INT2 outperforms LIN–INT1, which in turn outperforms LIN in terms of MSEs.

For performance measure, we will generate $2n$ observations $\{(Y_i, X_i, U_i), 1 \le i \le 2n\}$ for $n = 100$, 200, and 400, and use the first n observations for in-sample estimation and evaluation, and the last n observations for out-of-sample evaluation. We consider root-mean-square error (RMSE) for both in-sample and out-of-sample evaluation:

$$\mathrm{RMSE}_{\mathrm{in}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{m(X_i, U_i) - \widehat{m}(X_i, U_i)\}^2}$$

$$\mathrm{RMSE}_{\mathrm{out}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{m(X_{n+i}, U_{n+i}) - \widehat{m}(X_{n+i}, U_{n+i})\}^2 M(X_{n+i}, U^{c}_{n+i})}$$

where, for each method introduced earlier, $\widehat{m}(x, u)$ is an estimate of $m(x, u)$ using the first n observations $\{(Y_i, X_i, U_i), 1 \le i \le n\}$, and $M(\cdot, \cdot)$ is a weight function for the out-of-sample evaluation. We use the weight function here because the out-of-sample observations can lie outside the data range of the in-sample observations, and when this occurs, the nonparametric methods significantly deteriorate. In this simulation study, we set $M(X_{n+i}, U_{c,n+i}) = \Pi_{j=1}^{d+p}\mathbf{1}(|V_{ij} - \bar{V}_j| \le 1.5 s_{V_j})$, where $V_i = (X'_{n+i}, U^{c\prime}_{n+i})'$ and $\bar{V}_j$ and s_{V_j} denote the sample mean and standard deviation of $\{V_{ij}, 1 \le i \le n\}$, respectively. We report the mean, median, standard error, and interquartile range of RMSE over 1,000 Monte Carlo replications.

Table 1 reports the results for all five regression models. We summarize some interesting findings in Table 1. First, our proposed nonparametric functional coefficient estimator dominates all the other parametric or

Table 1. Comparison of Finite Sample Performance of Various Estimators (DGP1).

n	Model	In-Sample RMSE				Out-of-Sample RMSE			
		Mean	Median	SD	IQR	Mean	Median	SD	IQR
100	NP	0.729	0.703	0.160	0.194	1.140	1.096	0.449	0.271
	NP-FREQ	0.993	0.994	0.141	0.193	5.494	3.222	25.067	1.374
	LIN	2.375	2.336	0.521	0.713	2.752	2.684	0.693	0.754
	LIN-INT1	1.686	1.649	0.353	0.495	2.088	2.027	0.464	0.591
	LIN-INT2	1.091	1.072	0.209	0.273	1.531	1.494	0.351	0.427
200	NP	0.523	0.512	0.080	0.097	0.798	0.789	0.114	0.131
	NP-FREQ	0.880	0.876	0.101	0.134	14.325	4.638	67.586	5.410
	LIN	2.436	2.431	0.370	0.503	2.637	2.586	0.390	0.515
	LIN-INT1	1.726	1.726	0.244	0.346	1.941	1.912	0.282	0.363
	LIN-INT2	1.116	1.115	0.154	0.214	1.320	1.304	0.196	0.252
400	NP	0.385	0.374	0.076	0.057	0.591	0.582	0.070	0.078
	NP-FREQ	0.573	0.564	0.074	0.087	7.681	2.144	44.756	2.221
	LIN	2.472	2.460	0.2806	0.361	2.563	2.550	0.280	0.369
	LIN-INT1	1.760	1.757	0.180	0.240	1.860	1.857	0.194	0.259
	LIN-INT2	1.138	1.132	0.109	0.149	1.243	1.235	0.128	0.171

nonparametric estimators in terms of both in-sample RMSE and out-of-sample RMSE. Second, in comparison with the parametric estimators the NP-FREQ behaves reasonably well in terms of in-sample RMSE but not out-of-sample RMSE. The out-of-sample performance of the NP-FREQ is not acceptable even when the sample size is 400, in which case the average number of observations per cell is about 11. Third, as the sample size increases, the in-sample RMSEs of both our nonparametric estimator and the NP-FREQ decrease, but at rate slower than the parametric $n^{-1/2}$-rate as expected. The same is true for the out-of-sample RMSE of our nonparametric estimator. Fourth, the performance of the parametric estimators based on misspecified models may not improve as the sample size increases.

We now consider a second DGP that allows for weak data dependence between observations. The data are generated from the following DGP

$$Y_i = U_{i1}(U_{i1} + U_{i2} + U_{i3}) + U_{i1}(U_{i1} + U_{i2} + U_{i3})X_i + \varepsilon_i$$

where

$$X_i = 0.5X_{i-1} + e_{i1}$$
$$U_{i1} = 0.5 + 0.5U_{i-1,1} + e_{i2}$$

$\varepsilon_i \sim N(0,1)$, $e_{i1} \sim N(0, 1)$, and $e_{i2} \sim$ Uniform$(-0.5, 0.5)$, $U_{ij} \in \{-1, 0, 1\}$ with $P(U_{ij} = l) = 1/3$ for $l = -1, 0, 1$ and $j = 2, 3$. Furthermore, e_{ij} $(j = 1, 2)$, U_{i2}, U_{i3}, and ε_i are i.i.d. and mutually independent.

Like the case for DGP 1, we also consider two nonparametric estimators and three parametric estimators for the conditional mean function $m(x, u) = E(Y_i | X_i = x, U_i = u)$. We denote the corresponding regression models as NP, NP–FREQ, LIN, LIN–INT1, and LIN–INT2, respectively. We again consider the performance measure in terms of RMSE for both in-sample and out-of-sample evaluation and for $n = 100$, 200, and 400. The only difference is that when we generate $\{X_i, U_{i1}\}$, we throw away the first 200 observations to avoid the starting-up effects. We report the mean, median, standard error, and interquartile range of RMSE over 1,000 Monte Carlo replications in Table 2. The findings in Table 2 are similar to those in Table 1. One noticeable difference is that the out-of-sample performance of the NP-FREQ is not bad when $n = 400$ for this DGP. We conjecture this is due to the fact the average number of observations per cell $(400/9 \approx 44)$ is not small in this case.

Table 2. Comparison of Finite Sample Performance of Various Estimators (DGP2).

n	Model	In-Sample RMSE				Out-of-Sample RMSE			
		Mean	Median	SD	IQR	Mean	Median	SD	IQR
100	NP	0.388	0.369	0.122	0.155	0.606	0.575	0.184	0.222
	NP-FREQ	0.491	0.453	0.169	0.224	5.448	0.944	57.263	1.306
	LIN	2.565	2.437	0.856	1.013	3.050	2.871	1.008	1.216
	LIN-INT1	1.999	1.898	0.633	0.789	2.476	2.373	0.7788	1.009
	LIN-INT2	0.400	0.384	0.140	0.171	0.551	0.505	0.249	0.268
200	NP	0.231	0.218	0.070	0.090	0.392	0.372	0.097	0.114
	NP-FREQ	0.266	0.243	0.096	0.122	1.623	0.408	17.987	0.246
	LIN	2.646	2.576	0.639	0.834	2.901	2.847	0.669	0.854
	LIN-INT1	2.059	2.009	0.469	0.550	2.332	2.273	0.534	0.653
	LIN-INT2	0.384	0.369	0.105	0.142	0.468	0.437	0.142	0.168
400	NP	0.129	0.122	0.039	0.046	0.256	0.247	0.049	0.055
	NP-FREQ	0.138	0.126	0.046	0.051	0.274	0.255	0.140	0.062
	LIN	2.690	2.624	0.464	0.630	2.826	2.778	0.439	0.566
	LIN-INT1	2.096	2.066	0.328	0.442	2.240	2.217	0.342	0.459
	LIN-INT2	0.380	0.371	0.076	0.100	0.418	0.411	0.085	0.108

4. AN EMPIRICAL APPLICATION: ESTIMATING THE WAGE EQUATION

In this section, we apply our functional coefficient model to estimate a wage equation embedded in the framework of Mincer's (1974) human capital earning function. The basic Mincer wage function takes the form:

$$\log Y = \beta_0 + \beta_1 S + \beta_2 A + \beta_3 A^2 + \varepsilon \tag{15}$$

where Y is some measure of individual earnings, S is years of schooling, and A is age or work experience. In spite of its simplicity, Mincer equation captures the reality remarkably well (Card, 1999), and has been firmly established as a benchmark in labor economics. Concerning its specification, several extensions have been made to allow more general parametric functional forms (see Murphy & Welch, 1990). Further, a nonparametric analysis has been done in Ullah (1985) and Zheng (2000). And in practice, other control variables, such as indicators of gender, race, occupation, or martial status are routinely included in the wage equation when they are available. Nevertheless, the additive separability assumption of the standard Mincer equation may be too stringent. For instance, it ignores the possibility that higher education results in more return to seniority.[1] Also, it is often of keen economic and policy interest to investigate the differentials among different gender and race groups, where the return to education or experience may differ substantially. Therefore, we intend to estimate the functional coefficient model of the following form:

$$\log Y = a_1(U) + a_2(U)S + \varepsilon \tag{16}$$

where Y and S are as defined above, and U is a vector of mixed variables including one continuous variable – age or work experience, and six categorical variables for gender, race, martial status, veteran status, industry, and geographic location. The specification of Eq. (16) enables us to both study the direct effects of variables in U flexibly and investigate whether and how they influence the return to education. Some past literature has already suggested nonlinear relationship between seniority and wage beyond a quadratic form (Murphy & Welch, 1990; Ullah, 1985; Zheng, 2000), as well as the fact that rising return to education from the 1980s is more drastic in the younger cohorts than in the older ones (Card & Lemieux, 2001).

Our model is also suitable for analyzing the gender and racial wage differentials. In the study of discrimination, it is common practice to

estimate a "gender/racial wage gap" or estimate wage equation in separate samples. (For a survey of race and gender in the labor market, see Altonji & Blank, 1999.) Here the limitation of application of the traditional nonparametric method is the fact that indicators for gender and race are discrete, a problem overcome in our model. Also, compared with estimating wage separately among gender-racial groups or the frequency approach, our approach utilizes the entire dataset, thus achieving efficiency gain. We can also explicitly address other supposedly complicated interaction effects between the variables of interest. Further, unlike a complete nonparametric specification, model (16) has the further advantage that it can be readily extended to instrument variable estimation (Cai, Das, Xiong, & Wu, 2006), provided we have some reasonable instruments to correct the endogeneity in education. To keep our discussion focused, however, this aspect is not further explored in this paper.

The data utilized are drawn from March CPS data of the year 1990, 1995, 2000, and 2005. The earning variable is the weekly earning calculated from annual salary income divided by weeks of work, and deflated by the CPI (1982–1984 = 100). As usual, we exclude observations that are part-time workers, self-employed, over 65, under 18, or earn less than 50 dollars per week. All observations fall into 3 racial categories – White, Hispanic and otherwise, 4 geographic location categories – Northeast, Midwest, South and West, and 10 industrial categories. There are also three dichotomous variables "Female," "Veteran," and "Single." Years of schooling are estimated by records of the highest educational degree attained and experience is approximated by *Age-Schooling*-6. Fig. 1 plots wage against experience and years of schooling for the 4 years under our investigation. The left panel in Fig. 1 suggests the linear relationship (if any) between experience and wage is weak whereas the right panel in Fig. 1 suggests there is a positive relationship between years of schooling and wage.

As a comparison, we also estimate a simple linear wage function, a linear wage function with interacting covariates, and a partially linear model. The results are reported in Tables 3–5 (see also Fig. 2), respectively.

Results in Table 3 are in conformity with some stylized effects in labor economics, including stable return to schooling in the 1990s (Card & DiNardo, 2002; Beaudry & Green, 2004), concavity in return to experience, falling gender–wage gaps (Altonji & Blank, 1999), etc. The returns to schooling appear to range from 9.8 to 10.7% for the data under our investigation. Nevertheless, the inadequacy of a simple linear separable model is made clear in Table 4, since most of the interaction items of the covariates are significantly different from zero. And many of them are of

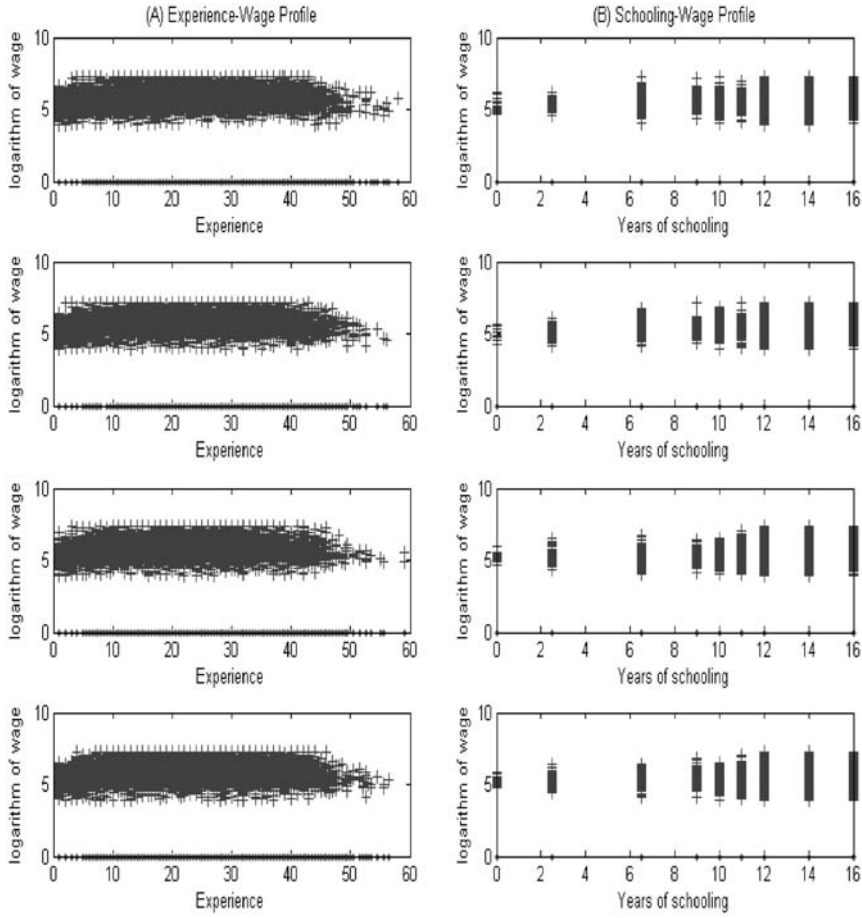


Fig. 1. Experience–Wage and Education–Wage Profiles. *Note:* The four rows correspond to years 1990, 1995, 2000, and 2005 from the top to the bottom. The sparsity of the experience variable is also plotted along the experience axis.

important economic implications, such as the higher return to education for female and higher return to experience for the White. And the goodness-of-fit of the model after accounting for the interaction effects has also increased modestly. Table 4 indicates the omission of these interaction terms may cause significant bias in the estimate of returns to schooling, and the bias can be as large as about 41% for year 2005 if we believe the linear model with interaction terms is correctly specified.

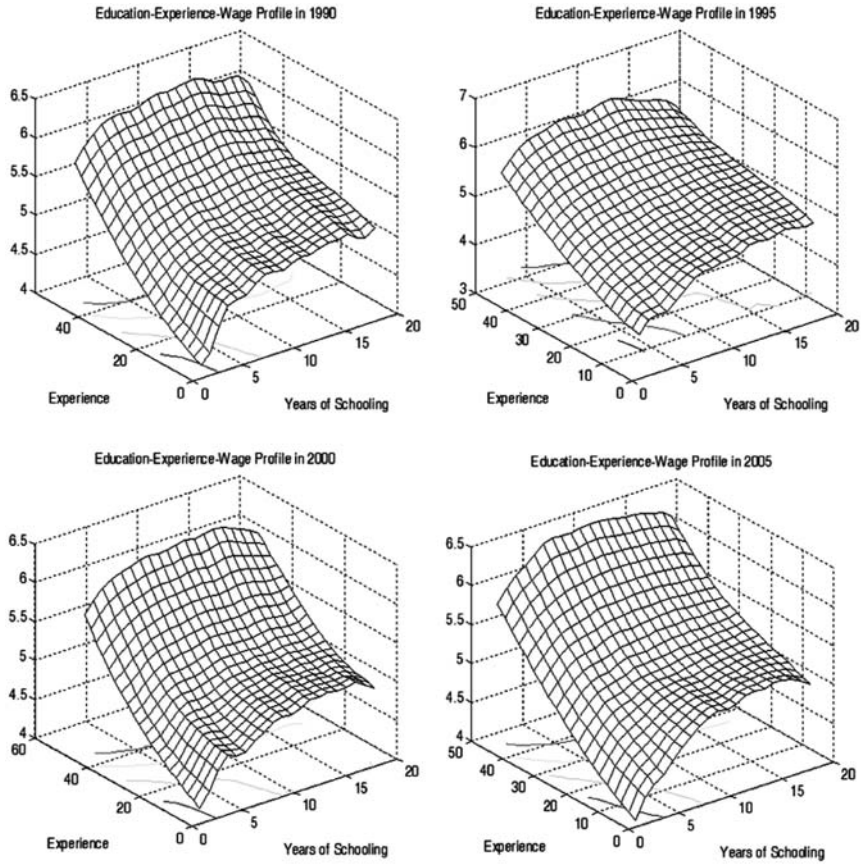


Fig. 2. Education–Experience–Wage Profile Resulting from the Partially Linear Models.

Another extension of Eq. (15) is to consider the partially linear model: $\log Y = m(Schooling,\ Experience) + Z'\beta + \varepsilon$, where Z is a set of dummy variables, and education and experience enter the model nonparametrically. We use the local linear method to estimate this model which is in the spirit of Robinson (1988). A second-order Epanechnikov kernel $w(v) = 0.75(1 - 0.2v^2)1(|v| \leq \sqrt{5})$ is used; and the bandwidth is chosen by a LSCV method. Given the large number of observations in our dataset, it is extremely costly to apply the LSCV method directly on all the observations. So we apply a methodology similar to that proposed at the end of Section 2

Table 3. Linear Wage Equation.

Year	1990		1995		2000		2005	
	(1)		(2)		(3)		(4)	
Education	0.098[a]	(0.002)	0.107[a]	(0.002)	0.105[a]	(0.003)	0.107[a]	(0.002)
Experience	0.029[a]	(0.001)	0.036[a]	(0.002)	0.029[a]	(0.002)	0.031[a]	(0.001)
Experience^2	−0.000[a]	(0.000)	−0.001[a]	(0.000)	−0.001[a]	(0.000)	−0.001[a]	(0.000)
Female	−0.309[a]	(0.010)	−0.290[a]	(0.010)	−0.279[a]	(0.011)	−0.277[a]	(0.008)
White	0.100[a]	(0.013)	0.130[a]	(0.013)	0.097[a]	(0.013)	−0.098[a]	(0.010)
Hispanic	0.034[c]	(0.017)	0.040[c]	(0.022)	0.033[c]	(0.019)	0.034[b]	(0.014)
Single	−0.087[a]	(0.009)	−0.071[a]	(0.010)	−0.097[a]	(0.010)	−0.102[a]	(0.008)
Veteran	−0.013	(0.013)	−0.049[a]	(0.015)	−0.008	(0.016)	−0.031[b]	(0.014)
Observations	12,328		10,834		10,433		17,466	
R2	0.37		0.36		0.33		0.34	

Note: (1) Heteroskedasticity-robust standard errors in parentheses. (2) a, b, and c stand for significance at 1%, 5%, and 10% levels, respectively. (3) Three region indicators, nine industry indicators, and a constant in all specifications.

to choose the bandwidth. As reported in Table 5, the partially linear model performs a little bit better in goodness-of-fit, as expected. However, it is noteworthy that comparing with the simple linear model, accounting for the possibly complex function form of education and experience has also significantly changed the estimates of the coefficients for the other covariates. For instance, the effects of race have drastically dropped in magnitude as well as significance. The difference may be the result of biases induced by the misspecification in a parametric model, and thus indicates the needs for the more general functional form assumption.

In all the above specifications, we use dummy variables to allow different intercepts for different regions and industries, and the majority of them have a significant estimated coefficient. The large number of categories makes it difficult to study their interaction effects with other regressors. In contrast, in the nonparametric framework of mixed regressors, only one categorical variable is necessary to describe such characteristic as industry or location. And this advantage has made our proposed model further suitable for the application.

For a comprehensive presentation of the regression results of model (16), we plot the wage–experience profiles of different cells defined by a discrete characteristic averaged over other categorical covariates. We use the second-order Epanechnikov kernel in our nonparametric estimation, and choose the bandwidth by the LSCV method introduced at the end of Section 2.4.

Table 4. Linear Wage Equation with Interacted Regressors.

Year	1990		1995		2000		2005	
	(1)		(2)		(3)		(4)	
Education	0.133[a]	(0.007)	0.146[a]	(0.007)	0.134[a]	(0.008)	0.151[a]	(0.006)
Experience	0.059[a]	(0.003)	0.071[a]	(0.003)	0.049[a]	(0.004)	0.053[a]	(0.003)
Experience^2	−0.001[a]	(0.000)	−0.001[a]	(0.000)	−0.001[a]	(0.000)	−0.001[a]	(0.000)
Female	−0.349[a]	(0.061)	−0.379[a]	(0.069)	−0.526[a]	(0.074)	−0.353[a]	(0.059)
White	0.039	(0.089)	0.091	(0.091)	−0.077	(0.106)	0.025	(0.082)
Hispanic	0.496[a]	(0.098)	0.551[a]	(0.114)	0.455[a]	(0.111)	0.607[a]	(0.086)
Single	−0.132[a]	(0.013)	−0.128[a]	(0.014)	−0.137[a]	(0.014)	−0.155[a]	(0.012)
Veteran	−0.024[c]	(0.014)	−0.056[a]	(0.015)	−0.010[a]	(0.017)	−0.027[c]	(0.015)
Education × Experience	−0.002[a]	(0.000)	−0.002[a]	(0.000)	−0.001[a]	(0.000)	−0.001[a]	(0.000)
Education × Female	0.014[a]	(0.004)	0.016[a]	(0.004)	0.022[a]	(0.005)	0.009[b]	(0.004)
Education × White	0.009	(0.006)	0.007	(0.006)	0.010	(0.007)	0.006	(0.006)
Education × Hispanic	−0.034[a]	(0.007)	−0.035[a]	(0.008)	−0.039[a]	(0.008)	−0.046[a]	(0.006)
White × Female	−0.135[a]	(0.025)	−0.123[a]	(0.026)	−0.087[a]	(0.026)	−0.098[a]	(0.020)
Hispanic × Female	−0.017	(0.034)	−0.069	(0.043)	−0.035	(0.038)	0.012	(0.028)
Single × Female	0.114[a]	(0.018)	0.141[a]	(0.018)	0.105[a]	(0.020)	0.135[a]	(0.010)
Experience × Female	−0.005[a]	(0.001)	−0.004[a]	(0.001)	−0.002[a]	(0.001)	−0.001[a]	(0.001)
Experience × White	0.000	(0.001)	0.000	(0.001)	0.004[a]	(0.001)	0.002[a]	(0.001)
Experience × Hispanic	−0.003[c]	(0.002)	−0.004[b]	(0.002)	0.001	(0.002)	−0.000	(0.001)
Observations	12,328		10,834		10,433		17,466	
R^2	0.39		0.38		0.34		0.36	

Note: (1) Heteroskedasticity-robust standard errors in parentheses. (2) a, b, and c stand for significance at 1%, 5%, and 10% levels, respectively. (3) Three region indicators, nine industry indicators, and a constant in all specifications.

The $R^{2\prime}s$ of the model have been increased up to 0.66, 0.65, 0.62, 0.68, respectively for the 4 years.

Fig. 3 reports the estimated a_1(*Experience*, *Region*, :) and a_2(*Experience*, *Region*, :) of model (16) for different regions averaged across all other categorical variables. a_1(*Experience*, *Region*, :) can be viewed as the direct effects of experience on wage for the particular region (averaged across all other categorical variables), and a_2(*Experience*, *Region*, :) represents the return to schooling as a function of experience for the particular region. We summarize some interesting findings from Fig. 3. First, while there are considerable variations between regions, we find the direct effects of experience on wage are usually *positive* (upward sloping) but *not necessarily concave*, which is in sharp contrast with the results of the parametric model. Notably, the experience–wage profile estimated here are from cross-sections and cannot be taken as individuals life-cycle earning trend. Second, if the

Table 5. Partially Linear Wage Equation.

Year	1990		1995		2000		2005	
	(1)		(2)		(3)		(4)	
Female	−0.280[a]	(0.010)	−0.265[a]	(0.011)	−0.259[a]	(0.011)	−0.259[a]	(0.008)
White	0.103[a]	(0.012)	0.135[a]	(0.013)	0.096[a]	(0.013)	0.102[a]	(0.010)
Hispanic	0.001	(0.017)	−0.001[a]	(0.022)	−0.017	(0.019)	−0.007	(0.014)
Single	−0.077[a]	(0.009)	−0.058[a]	(0.010)	−0.082[a]	(0.010)	−0.077[a]	(0.008)
Veteran	0.024[a]	(0.013)	−0.009	(0.015)	0.021	(0.016)	−0.001	(0.014)
Observations	12,328		10,834		10,433		17,446	
R^2	0.40		0.39		0.36		0.38	

Note: (1) Heteroskedasticity-robust standard errors in parentheses. (2) a stands for significance at 1% level. (3) Three region indicators, nine industry indicators and a constant in all specifications. (4) The estimate of m (*Schooling*, *Experience*) is plotted in Fig. 2.

standard Mincer equation holds, we expect the estimated a_2(*Experience, Region*, :) to be a horizontal line. But clearly, this is far from reality. The effects of experience on return to schooling are *mainly negative*, which agrees with our previous results from the parametric setting, presented in Table 4. The findings here have interesting econometric interpretation. On the one hand, we may wonder if higher education causes higher return to seniority, or similarly, longer experience leads to higher return to education. On the other hand, it is possible that the young cohorts (implied by shorter experience) have higher return to education, due to cohort supply effects, technological changes or some other reasons. And we need to resort to empirical results to evaluate the overall influence. In the sample studied here, the later force has been found to dominate the former in their direction of impacts. Admittedly, the interacting patterns of the regressors in the wage equation uncovered by this functional coefficient model require further careful investigation.

Fig. 4 reports the estimated a_1(*Experience, Race*, :) and a_2(*Experience*, *Race*, :) of model (16) for different races averaged across all other categorical variables. a_1(*Experience*, *Race*, :) can be viewed as the direct effects of experience on wage for the race, and a_2(*Experience*, *Race*, :) represents the return to schooling as a function of experience for the particular race. The findings are similar to those in Fig. 3. We only mention that the return to schooling seems much higher for White and others (above 0.1 across 2/3 of the range of experience) than Hispanic (below 0.1 in almost all the range of experience).

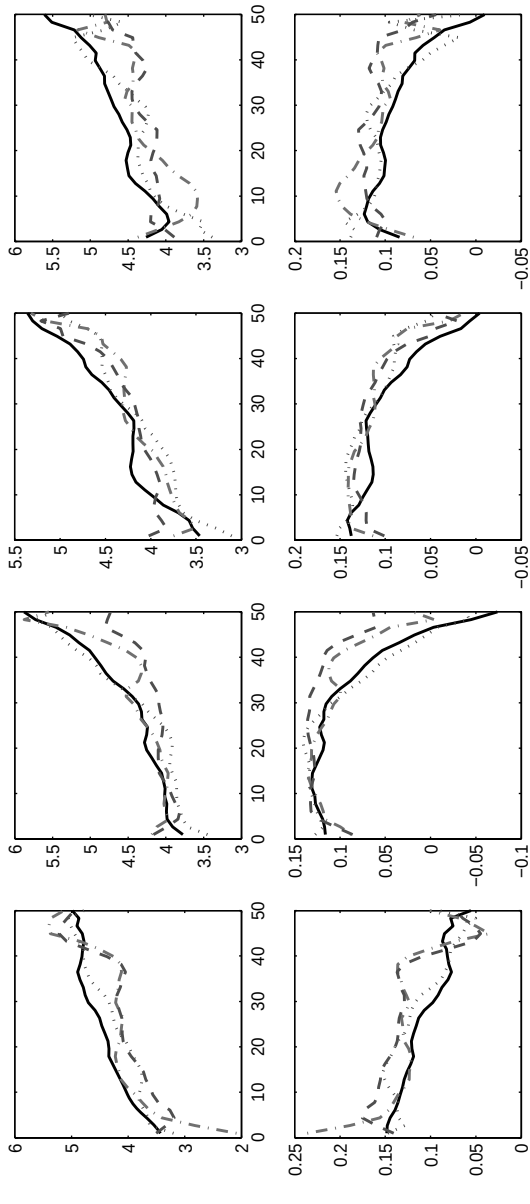


Fig. 3. Plots of $a_1(Experience, Region, \cdot)$ and $a_2(Experience, Region, \cdot)$ averaging over other categorical variables (as represented by '$\cdot$' in the definitions of a_1 and a_2). *Note:* Horizontal axis – *Experience*. Vertical axis – a_1 or a_2. The two rows correspond to a_1 and a_2, respectively, from the top to the bottom. The four columns correspond to *Region* = Northeast, Midwest, South and West from the left to the right column. 1990, solid line; 1995, dotted line; 2000, dashdot line; and 2005, dashed line.

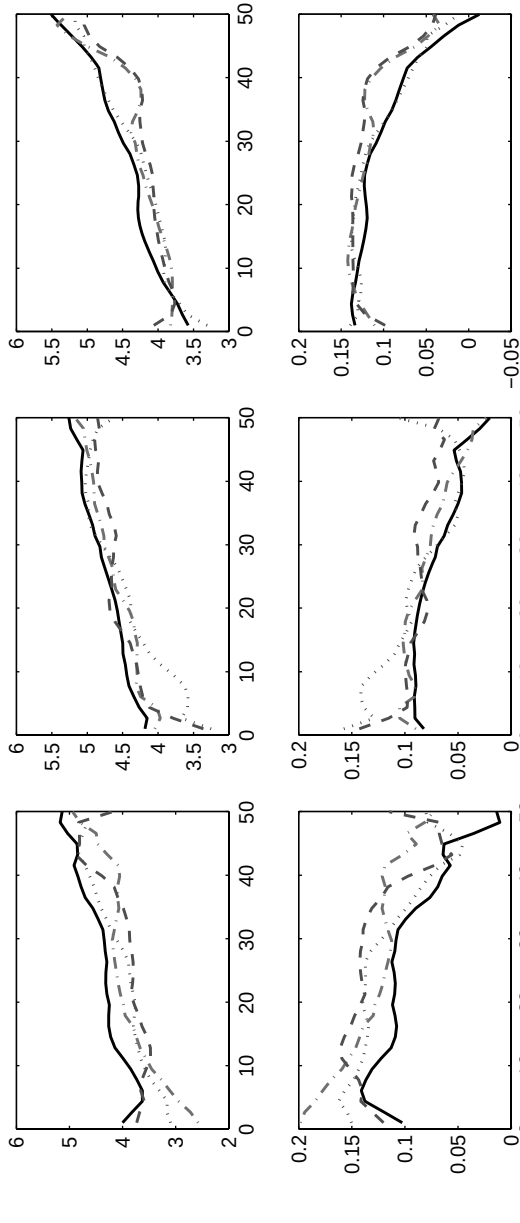


Fig. 4. Plots of a_1 (*Experience*, *Race*, :) and a_2 (*Experience*, *Race*, :) averaging over other categorical variables. *Note:* Horizontal axis – *Experience*. Vertical axis – a_1 or a_2. The two rows correspond to a_1 and a_2 from the top to the bottom. The three columns correspond to *Race* = Otherwise, Hispanic, and White from the left to the right column. 1990, solid line; 1995, dotted line; 2000, dashdot line; and 2005, dashed line.

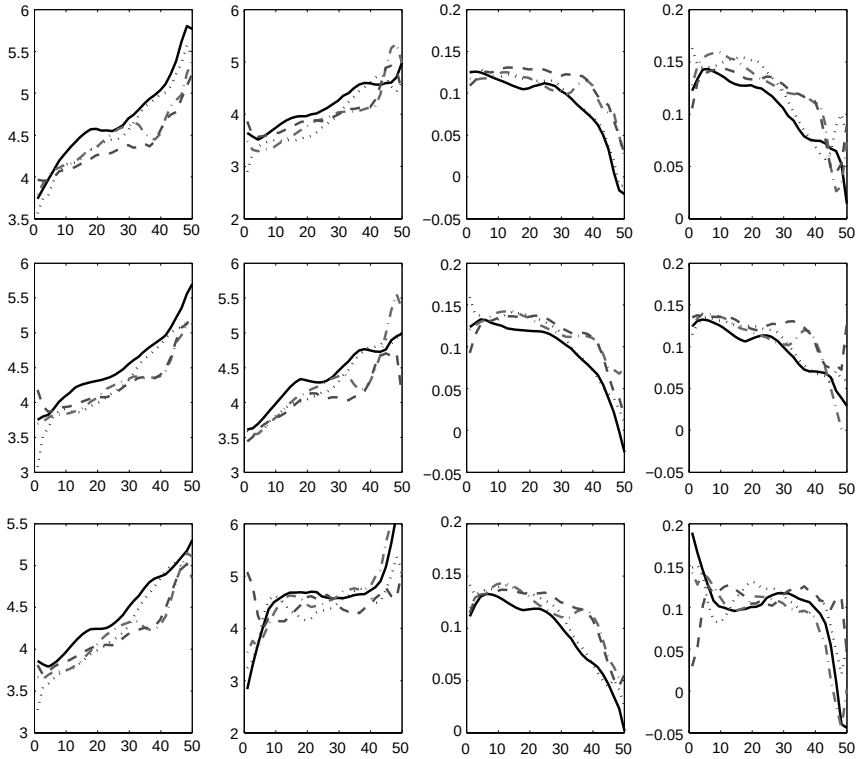


Fig. 5. Plots of a_1(*Experience*, *Gender*, :) and a_2(*Experience*, *Gender*, :) (first row), a_1(*Experience*, *Single*, :) and a_2(*Experience*, *Single*, :) (second row), a_1(*Experience*, *Veteran*, :), and a_2(*Experience*, *Veteran*, :) (third row), averaging over other categorical variables. *Note:* Horizontal axis – *Experience*. Vertical axis – a_1 or a_2. First row: The four columns from the left to the right correspond to a_1 for male, a_1 for female, a_2 for male, and a_2 for female, respectively. Second row: The four columns from the left to the right correspond to a_1 for nonsingle, a_1 for single, a_2 for nonsingle, and a_2 for single, respectively. Third row: The four columns from the left to the right correspond to a_1 for nonveteran, a_1 for veteran, a_2 for nonveteran, and a_2 for veteran, respectively. 1990, solid line; 1995, dotted line; 2000, dashdot line; and 2005, dashed line.

Fig. 5 reports the estimated a_1(*Experience*, :) and a_2(*Experience*, :) depending on whether a person is male or female, single or nonsingle, and veteran or nonveteran. Fig. 6 reports the estimated a_1(*Experience, Industry*, :) and a_2(*Experience*, *Industry*, :) of model (16) for different industries averaged across all other categorical variables. Both figures can be

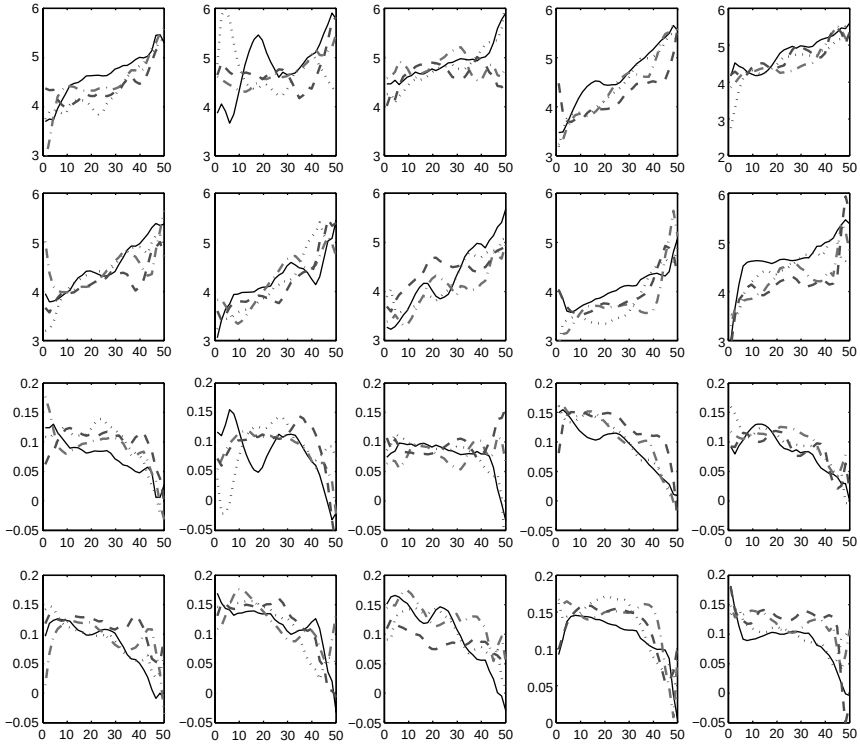


Fig. 6. Plots of a_1(*Experience*, *Industry*, :) and a_2(*Experience*, *Industry*, :) averaging over other categorical variables. *Note:* Horizontal axis – *Experience*. Vertical axis – a_1 or a_2. The first two rows correspond to a_1, and the last two rows correspond to a_2. For rows 1 and 3, the five columns from the left to the right correspond respectively to *Industry* = Agriculture, Mining, Construction, Manufacturing, and Transportation. For rows 2 and 4, the five columns from the left to the right correspond respectively to *Industry* = Wholesale and return, Finance, Personal services, Professional services, and Public administration. 1990, solid line; 1995, dotted line, 2000, dashdot line; and 2005, dashed line.

interpreted similarly to the case of Fig. 3. The most eminent implication by these figures is that return to education does depend heavily upon other variables. In particular, the top panel in Fig. 5 indicates that higher return to education for female across all the range of age or work experience. In addition, we can see substantial variation among the cells which suggests the highly complex functional form of the wage equation.

Fig. 7 reports the estimated a_1(*Experience*, :) and a_2(*Experience*, :) averaged over all categorical variables. Similarly to the cases of Figs. 3–6, we observe that the direct impact of experience on wage is *positive* but the return to schooling as a function of experience tends to be decreasing except when experience is low ($\leq$4 years in 1990, $\leq$12 in 2005). When experience is larger than 37 years, the return to schooling is diminishing very fast as a function of experience. Prior to 37 years, the returns to schooling may vary from 0.105 to 0.145.

Therefore, our empirical application has demonstrated the usefulness of our proposed model in uncovering complicated patterns of interacting effects of the covariates on the dependent variable. And the results are of interesting economic interpretation.

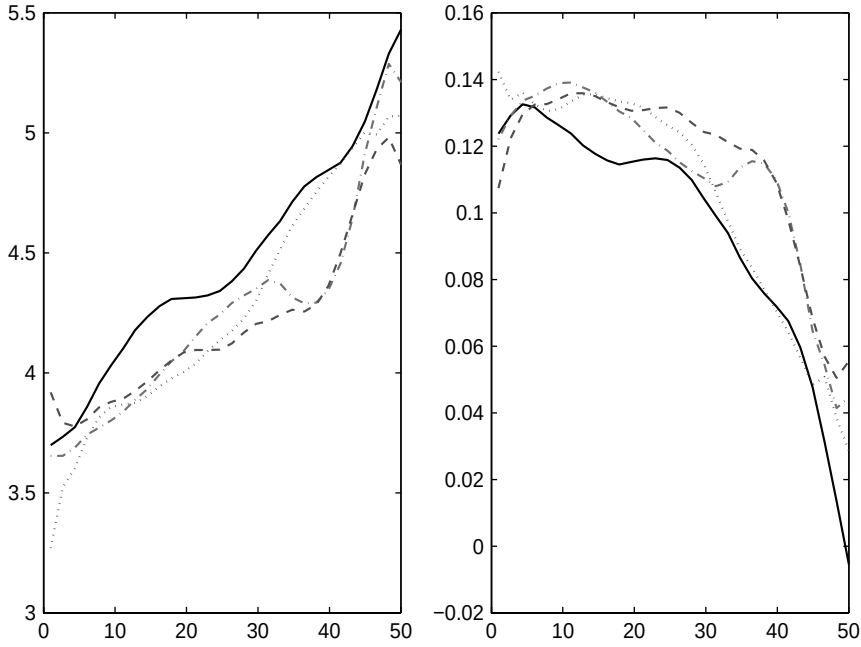


Fig. 7. Plots of a_1(*Experience*, :) and a_2(*Experience*, :) averaging over all categorical variables. *Note:* Horizontal axis – *Experience*. Vertical axis – a_1 or a_2. The two columns from the left to the right correspond to a_1 and a_2, respectively. 1990, solid line; 1995, dotted line; 2000, dashdot line; and 2005, dashed line.

5. CONCLUSIONS

This paper proposes a local linear functional coefficient estimator that admits a mix of discrete and continuous data for stationary time series. Under weak conditions our estimator is asymptotically normally distributed. We also include simulations and empirical applications. We find from the simulations that our nonparametric estimators behave reasonably well for a variety of DGPs.

As an empirical application, we estimate a human capital earning function from the recent CPS data. Unlike the widely used linear separable model, or the frequency approach that conducts estimation in splitted samples, the proposed model enables us to utilize the entire dataset and allows the return to education to vary with the other categorical and continuous variables. The empirical findings show considerable interacting effects among the regressors in the wage equation. For instance, the younger cohorts are found to have higher return to education. While these patterns need further explanation from labor economic theory, the application demonstrates the usefulness of our proposed functional coefficient model due to its flexibility and clear economic interpretation. And thus the model has good potential for applied research. Our future research will address some related problems such as the optimal selection of smoothing parameters. Another extension is to study the estimation of functional coefficient model with both endogeneity and mixed regressors.

NOTE

1. Throughout our paper the use of word return or marginal return from education refers to the functional (varying) coefficient of education that may not be the marginal return if the education is endogenous, an issue not explored in our paper.

ACKNOWLEDGMENTS

The authors gratefully thank the editors and two anonymous referees for their constructive comments and suggestions. They also thank Zongwu Cai for his helpful comment on an early version of this paper. The first author gratefully acknowledges financial support from the NSFC (Project 70501001 and 70601001). The third author gratefully acknowledges the financial support from the Academic Senate, UCR.

REFERENCES

Aitchison, J., & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, *63*, 413–420.

Altonji, J. G., & Blank, R. M. (1999). Race and gender in the labor market", In: O. C. Ashenfelter & D. Card (Eds), *Handbook of Labor Economics* (Vol. 3C, Ch. 48, pp. 3143–3259). North Holland: Elsevier.

Beaudry, P., & Green, D. A. (2004). Changes in US wages, 1976–2000: ongoing skill bias or major technological change? *Journal of Labor Economics*, *23*, 491–526.

Bosq, D. (1996). *Nonparametric statistics for stochastic processes: Estimation and prediction*. New York: Springer.

Cai, Z., Das, M., Xiong, H., & Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics*, *133*, 207–241.

Cai, Z., Fan, J., & Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of American Statistical Association*, *95*, 941–956.

Cai, Z., & Ould-Saïd, E. (2003). Local M-estimator for nonparametric time series. *Statistics and Probability Letters*, *65*, 433–449.

Card, D. (1999). Casual effect of education on earnings. In: O. C. Ashenfelter & D. Card (Eds), *Handbook of Labor Economics* (Vol. 3A, Ch. 48, pp. 1802–1864). North Holland: Elsevier.

Card, D., & DiNardo, J. (2002). Skill biased technological change and rising wage inequality: some problems and puzzles. *Journal of Labor Economics*, *20*, 733–783.

Card, D., & Lemieux, T. (2001). Can falling supply explain the rising return to college for younger men? A cohort-based analysis. *The Quarterly Journal of Economics*, *116*, 705–746.

Chen, R., & Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of American Statistical Association*, *88*, 298–308.

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. In: J. M. Chambers & T. J. Hastie (Eds), *Statistical models in S* (pp. 309–376). Pacific Grove, CA: Wadsworth & Brooks/Cole.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications, vol. 66 of monographs on statistics and applied probability*. London: Chapman and Hall.

Fan, Y., & Li, Q. (1999). Root-*n*-consistent estimation of partially linear time series models. *Journal of Nonparametric Statistics*, *11*, 251–269.

Hall, P., & Heyde, C. C. (1980). *Martingale limit theory and its applications*. New York: Academic Press.

Hall, P., Wolf, R. C. L., & Yao, Q. (1999). Methods of estimating a conditional distribution function. *Journal of the American Statistical Association*, *94*, 154–163.

Hastie, T. J., & Tibshirani, R. J. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B.*, *55*, 757–796.

Li, Q., & Racine, J. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, CA: Princeton University Press.

Li, Q., & Racine, J. (2008a). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, *26*, 423–434.

Li, Q., Racine, J. (2008b). Smoothing varying-coefficient estimation and inference for qualitative and quantitative data. Department of Economics, Texas A&M University, Mimeo.

Mincer, J. (1974). *Schooling, experience and earnings*. New York: National Bureau of Economic Research.

Murphy, K., & Welch, F. (1990). Empirical age-earnings profiles. *Journal of Labor Economics*, *8*, 202–229.

Racine, J., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*, 99–130.

Robinson, P. M. (1988). Root-*n*-consistent semiparametric regression. *Econometrica*, *56*, 931–954.

Ullah, A. (1985). Specification analysis of econometric models. *Journal of Quantitative Economics*, *1*, 187–209.

Zheng, J. (2000). Specification testing and nonparametric estimation of the human capital model. Applying kernel and nonparametric estimation to economic topics. In: T. B. Fomby & R. C. Hill (Eds), *Advances in econometrics* (Vol. 14, pp. 129–154). Stamford, CT: JAI Press Inc.

APPENDIX

We use $||\cdot||$ to denote the Euclidean norm of $\cdot$, C to signify a generic constant whose exact value may vary from case to case, and a' to denote the transpose of a. Let $d_{u_i u} = \sum_{t=1}^{q} \mathbf{1}(U_{it}^d \neq u_t^d)$ where $\mathbf{1}(U_{it}^{\mathrm{d}} \neq u_t^{\mathrm{d}})$ is an indicator function that takes value 1 if $(U_{it}^{\mathrm{d}} \neq u_t^{\mathrm{d}})$ and 0 otherwise. So $d_{u_i u}$ indicates the number of disagreeing components between U_{it}^d and u_t^d.

Proof of Theorem 1. We first define some notation. For any $p \times 1$ vectors $c = (c_1, \ldots, c_p)'$ and $d = (d_1, \ldots, d_p)'$, let $c \odot d = (c_1 d_1, \ldots, c_p d_p)'$ and $c/d = (c_1 d_1, \ldots, c_p d_p)'$ whenever applicable. Let

$$S_n = S_n(u) = \begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1}' & S_{n,2} \end{pmatrix}, \quad T_n = T_n(u) = T_{n,1} + T_{n,2} \tag{A.1}$$

with

$$S_{n,0} = S_{n,0}(u) = n^{-1} \sum_{i=1}^{n} X_i X_i' K_{\mathrm{iu}}$$

$$S_{n,1} = S_{n,1}(u) = n^{-1} \sum_{i=1}^{n} (X_i X_i') \otimes \left(\frac{(U_i^{\mathrm{c}} - u^{\mathrm{c}})}{h} \right)' K_{\mathrm{iu}}$$

$$S_{n,2} = S_{n,2}(u) = n^{-1} \sum_{i=1}^{n} (X_i X_i') \otimes \left(\left(\frac{(U_i^{\mathrm{c}} - u^{\mathrm{c}})}{h} \right) \left(\frac{(U_i^{\mathrm{c}} - u^{\mathrm{c}})}{h} \right)' \right) K_{\mathrm{iu}}$$

$$T_{n,1} = T_{n,1}(u) = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} X_i \varepsilon_i \\ (X_i \varepsilon_i) \otimes ((U_i^{\mathrm{c}} - u^{\mathrm{c}})/h) \end{pmatrix} K_{\mathrm{iu}}, \text{ and}$$

$$T_{n,2} = T_{n,2}(u) = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} (X_i X_i' \boldsymbol{a}(U_i)) \\ (X_i X_i' \boldsymbol{a}(U_i)) \otimes ((U_i^{c} - u^{c})/h) \end{pmatrix} K_{\mathrm{iu}}$$

where recall $\boldsymbol{a}(U_i) = (a_1(U_i), \ldots, a_d(U_i))'$. Then

$$\widehat{\theta} = H_1^{-1} S_n^{-1} T_n$$

where $H_1 = \mathrm{diag}(1, \ldots, 1, h', \ldots, h')$ is a $d(p+1) \times d(p+1)$ diagonal matrix with d diagonal elements of 1 and d diagonal elements of h. Let $H = \sqrt{n h_1 \ \ldots \ h_p}$. Then

$$\begin{aligned} HH_1(\widehat{\theta} - \theta) &= HS_n^{-1}(T_n - S_n\theta) \\ &= HS_n^{-1} T_{n,1} + HS_n^{-1}(T_{n,2} - S_n\theta) \end{aligned}$$

We first prove several lemmas.

Lemma A.1.

(a) $S_{n,0} = \Omega(u) f_u(u) + o_p(1)$,
(b) $S_{n,1} = O_p(||h||^2 + ||h||\,||\lambda||) = o_p(1)$,
(c) $S_{n,2} = \mu_{2,1}(\Omega(u) f_u(u)) \otimes I_p + o_p(1)$.

Proof. We only prove (a) since the proofs of (b) and (c) are similar. First by the stationarity of $\{X_i, U_i\}$

$$\begin{aligned} E(S_{n,0}) &= E(X_i X_i' K_{\mathrm{iu}}) \\ &= E(X_i X_i'\, W_{h,\mathrm{iu}} | d_{u_i u} = 0) p(u^{\mathrm{d}}) \\ &\quad + \sum_{s=1}^{q} E(X_i X_i' W_{h,\mathrm{iu}} L_{\lambda,\mathrm{iu}} | d_{u_i u} = s) P(d_{u_i u} = s) \\ &= E(\Omega(U_i) W_{h,\mathrm{iu}} | d_{u_i u} = 0) p(u^{\mathrm{d}}) + O(||\lambda||) \\ &= \int \Omega(u^{\mathrm{c}} + h \odot v, u^{\mathrm{d}}) f_u(u^{\mathrm{c}} + h \odot v, u^{\mathrm{d}}) W(v) dv + O(||\lambda||) \\ &= \Omega(u) f_u(u) + O(||h||^2 + ||\lambda||) \end{aligned} \tag{A.2}$$

where $p(u^{\mathrm{d}}) = P(U_i^{\mathrm{d}} = u^{\mathrm{d}})$.

Since a typical element of $S_{n,0}$ is

$$s_{n,\mathrm{st}} = n^{-1}\sum_{i=1}^{n} X_{\mathrm{is}} X_{\mathrm{it}} K_{\mathrm{iu}}, s, t = 1, \ldots, d$$

by the Chebyshev's inequality, it suffices to show that

$$\mathrm{var}\,(s_{n,st}) = o(1)\,. \tag{A.3}$$

Let $\xi_i = X_{is}X_{it}K_{\mathrm{iu}}$. By the stationarity of $\{X_i,U_i\}$, we have

$$\mathrm{var}(s_{n,\mathrm{st}}) = \frac{1}{n}\mathrm{var}(\xi_1) + \frac{2}{n}\sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)\mathrm{cov}(\xi_1, \xi_j) \tag{A.4}$$

Clearly,

$$\mathrm{var}(\xi_1) \le E(X_{1s}^2 X_{1t}^2 K_{1u}^2) = O((h_1 \ldots h_n)^{-1}) \tag{A.5}$$

To obtain an upper bound for the second term on the right-hand side of Eq. (A.4), we split it into two terms as follows

$$\sum_{j=1}^{n-1}|\mathrm{cov}(\xi_1, \xi_j)| = \sum_{j=1}^{d_n}|\mathrm{cov}\,(\xi_1, \xi_j)| + \sum_{j=d_n+1}^{n-1}|\mathrm{cov}(\xi_1, \xi_j)| \equiv J_1 + J_2$$

where d_n is a sequence of positive integers such that $d_n h_1 \ldots h_p \to 0$ as $n \to \infty$. Since for any $j > 1$,

$$|E(\xi_1\xi_j)| = |E(X_{1s}X_{1t}K_{1,u}X_{js}X_{jt}K_{j,u})| = O(1)$$

$J_1 = O(d_n)$. For J_2, by the Davydov's inequality (e.g., Hall & Heyde, 1980, p. 278; or Bosq, 1996, p. 19), we have

$$\begin{aligned}\mathrm{cov}(\xi_1, \xi_j) &\le C[\alpha(j-1)]^{\gamma/(2+\gamma)}(E|\xi_1|^{2+\gamma})^{2/(2+\gamma)} \\ &= C[\alpha(j-1)]^{\gamma/(2+\gamma)}\left\{E\left|(X_{1s}X_{1t})^{(2+\gamma)}K_{1,u}^{2+\gamma}\right|\right\}^{2/(2+\gamma)} \\ &= O\left((h_1 \ldots h_p)^{-(2+2\gamma)/(2+\gamma)}\right)[\alpha(j-1)]^{\gamma/(2+\gamma)}\end{aligned} \tag{A.6}$$

So

$$J_2 \leq C(h_1 \ldots h_p)^{-(2+2\gamma)/(2+\gamma)} \sum_{j=d_n}^{n-1} [\alpha(j)]^{\gamma/(2+\gamma)}$$

$$\leq C(h_1 \ldots h_p)^{-(2+2\gamma)/(2+\gamma)} d_n^{-\alpha} \sum_{j=d_n}^{\infty} j^{\alpha} [\alpha(j)]^{\gamma/(2+\gamma)} = o((h_1 \ldots h_p)^{-1}) \qquad \text{(A.7)}$$

by choosing d_n such that $d_n^{-\alpha}(h_1 \ldots h_p)^{-\gamma/(2+\gamma)} = o(1)$. This, in conjunction with Eqs. (A.4) and (A.5), implies, $\text{var}(s_{n,st}) = O((nh_1 \ldots h_p)^{-1}) = o(1)$.

Lemma A.2.

$$HT_{n,1} = n^{-1/2}(h_1 \ldots h_p)^{1/2} \sum_{i=1}^{n} \begin{pmatrix} X_i \varepsilon_i \\ (X_i \varepsilon_i) \otimes ((U_i^{\text{c}} - u^{\text{c}})/h) \end{pmatrix} K_{\text{iu}} \xrightarrow{d} N(0, \Gamma)$$

where $H = \sqrt{nh_1 \ldots h_p}$, $\sigma^2(u, x) = E[\varepsilon_i^2 | U_i = u, X_i = x]$, $\Omega^*(u) = E[X_i X_i' \sigma^2(U_i, X_i) | U_i = u]$, and

$$\Gamma = \Gamma(u) = f_u(u) \begin{pmatrix} \mu_{0,2}^p \Omega^*(u) & 0' \\ 0 & \mu_{2,2} \Omega^*(u) \otimes I_p \end{pmatrix}$$

Proof. Let w be a unit vector on $\mathbb{R}^{d(p+1)}$. Let

$$\zeta_i = (h_1 \ldots h_p)^{1/2} w' \begin{pmatrix} X_i \varepsilon_i \\ (X_i \varepsilon_i) \otimes ((U_i^{\text{c}} - u^{\text{c}})/h) \end{pmatrix} K_{\text{iu}}$$

By the Cramér–Wold device, it suffices to prove

$$I_n = n^{-1/2} \sum_{i=1}^{n} \zeta_i \xrightarrow{d} N(0, w' \Gamma w). \qquad \text{(A.8)}$$

Clearly, by the law of iterated expectation, $E(\zeta_i) = 0$. Now

$$\text{var}(I_n) = \text{var}(\zeta_1) + 2 \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) \text{cov}(\zeta_1, \zeta_j)$$

By arguments similar to those used in the proof of Lemma A.1,

$$\begin{aligned}&\mathrm{var}(\zeta_1)\\&= h_1 \ldots h_p w' E\left\{ \begin{pmatrix} \Omega^*(U_i) & \Omega^*(U_i)\otimes((U_i^{\mathrm{c}}-u^{\mathrm{c}})/h)' \\ \Omega^*(U_i)\otimes((U_i^{\mathrm{c}}-u^{\mathrm{c}})/h) & \Omega^*(U_i)\otimes(((U_i^{\mathrm{c}}-u^{\mathrm{c}})/h)((U_i^{\mathrm{c}}-u^{\mathrm{c}})/h)') \end{pmatrix} K_{\mathrm{iu}}^2 \right\} w\\&= w'\Gamma w + o(1)\end{aligned}$$

and

$$\sum_{j=1}^{n-1} |\mathrm{cov}(\zeta_1, \zeta_j)| = o(1)$$

which implies that

$$\mathrm{var}(I_n) \to w'\Gamma w \quad \text{as} \quad n \to \infty$$

Using the standard Doob's small-block and large-block technique, we can finish the rest of the proof by following the arguments of Cai et al. (2000, pp. 954–955) or Cai and Ould-Saïd (2003, pp. 446–448). ■

Lemma A.3. Let $B_n = H(T_{n,2} - S_n\theta)$. Then $B_n = b(h, \lambda) + o_p(1)$, where $b(h, \lambda)$ is defined in Eq. (10).

Proof. Let

$$\begin{aligned}\varsigma_i = {} & H \begin{pmatrix} (X_i X_i' \boldsymbol{a}(U_i)) \\ (X_i X_i' \boldsymbol{a}(U_i)) \otimes ((U_i^{\mathrm{c}} - u^{\mathrm{c}})/h) \end{pmatrix} K_{\mathrm{iu}} \\ & - H \begin{pmatrix} X_i X_i' & (X_i X_i') \otimes ((U_i^{\mathrm{c}} - u^{\mathrm{c}})/h)' \\ (X_i X_i') \otimes ((U_i^{\mathrm{c}} - u^{\mathrm{c}})/h) & (X_i X_i') \otimes (((U_i^{\mathrm{c}} - u^{\mathrm{c}})/h)((U_i^{\mathrm{c}} - u^{\mathrm{c}})/h)') \end{pmatrix} \theta K_{\mathrm{iu}}\end{aligned}$$

Then we have

$$B_n = \frac{1}{n} \sum_{i=1}^{n} \varsigma_i \tag{A.9}$$

Let $\bar{\varsigma}_i = E(\varsigma_i | U_i)$. Then

$$\begin{aligned}E(B_n) &= E(\bar{\varsigma}_i)\\&= E\{\bar{\varsigma}_i | d_{u_i u} = 0\} p(u^{\mathrm{d}}) + E\{\bar{\varsigma}_i | d_{u_i u} = 1\} P(d_{u_i u} = 1) + O(H||\gamma||^2)\\&\equiv b_{n,1} + b_{n,2} + o(1)\end{aligned}$$

On the set $\{U_i^{\mathrm{d}} = u^{\mathrm{d}}, \ W_{h,iu} > 0\}$,

$$a_j(U_i) = a_j(u) + \dot{a}_j(u)'(U_i^{\mathrm{c}} - u^{\mathrm{c}}) + \tfrac{1}{2}(U_i^{\mathrm{c}} - u^{\mathrm{c}})' \ddot{a}_j(u)(U_i^{\mathrm{c}} - u^{\mathrm{c}}) + o(||h||^2)$$

Let $A(U_i,u)=((U_i^c-u^c)'\ddot{a}_1(u)(U_i^c-u^c),\ldots,(U_i^c-u^c)'\ddot{a}_d(u)(U_i^c-u^c))'$. Recall $A=(\sum_{s=1}^p h_s^2 a_{1,ss}(u),\ldots,\sum_{s=1}^p h_s^2 a_{d,ss}(u))'$, and $\boldsymbol{b}(u)=(\dot{a}_1(u)',\ldots,\dot{a}_d(u)')'$. Then we have

$$b_{n,1}=\frac{1}{2}H\,E\left\{\begin{pmatrix}\Omega(U_i)A(U_i,u)\\(\Omega(U_i)A(U_i,u))\otimes((U_i^c-u^c)/h)\end{pmatrix}W_{h,\mathrm{iu}}\big|d_{u_iu}=0\right\}\times p(u^d)+o(1)$$
$$=\frac{H\mu_{2,1}}{2}\begin{pmatrix}f_u(u)\Omega(u)A\\0\end{pmatrix}+o(1)$$

and

$$\begin{aligned}&b_{n,2}\\ =\;& H\,E\{\xi_i|d_{u_iu}=1\}P(d_{u_iu}=1)\\ =\;& H\,E\left\{\begin{pmatrix}\Omega(U_i)(\mathbf{a}(U_i)-\mathbf{a}(u))-(\Omega(U_i)\otimes((U_i^c-u^c)/h'))\mathbf{b}(u)\\(\Omega(U_i)\,(\mathbf{a}(U_i)-\mathbf{a}(u)))\otimes((U_i^c-u^c)/h)\\-(\Omega(U_i)\otimes(((U_i^c-u^c)/h)((U_i^c-u^c)/h)'))\boldsymbol{b}(u)\end{pmatrix}\right.\\ &\left.\times K_{\mathrm{iu}}|d_{u_iu}=1\right\}P(d_{u_iu}=1)+o(1)\\ =\;& H\sum_{\tilde{u}^d\in\mathcal{D}}\sum_{s=1}^{q}\lambda_s I_s(u^{\mathrm{d}},\tilde{u}^{\mathrm{d}})f_u(u^c,\tilde{u}^{\mathrm{d}})\begin{pmatrix}\Omega(u^c,\tilde{u}^{\mathrm{d}})(\boldsymbol{a}(u^c,\tilde{u}^{\mathrm{d}})-\boldsymbol{a}(u))\\-\mu_{2,1}(\Omega(u^c,\tilde{u}^{\mathrm{d}})\otimes I_p)\boldsymbol{b}(u)\end{pmatrix}+o(1)\end{aligned}$$

Consequently, $E(B_n)=b(h,\lambda)+o(1)$, where $b(h,\lambda)$ is defined in Eq. (10).

To show $\mathrm{var}(B_n)=o(1)$ elementwise, we focus on the first d elements $\varsigma_i^{(1)}$ of ς_i since the other cases are similar, where

$$\varsigma_i^{(1)}=H\left[X_iX_i'(\boldsymbol{a}(U_i)-\boldsymbol{a}(u))-\left(X_iX_i'\otimes\left(\frac{(U_i^c-u^c)}{h}\right)'\right)\boldsymbol{b}(u)\right]K_{\mathrm{iu}}$$

A typical element of $\varsigma_i^{(1)}$ is

$$\varsigma_{i,t}^{(1)}=H\left[X_{it}\sum_{s=1}^{d}X_{is}(a_s(U_i)-a_s(u))-X_{it}\sum_{s=1}^{d}X_{is}\left(\frac{(U_i^c-u^c)}{h}\right)'b_j(u)\right]K_{\mathrm{iu}}$$

$t=1,\ldots,d$.

$$\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n}\varsigma_{i,t}^{(1)}\right)=\frac{1}{n}\mathrm{var}\left(\varsigma_{1,t}^{(1)}\right)+\frac{2}{n}\sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)\mathrm{cov}(\varsigma_{1,t}^{(1)},\varsigma_{j,t}^{(1)})$$

By arguments similar to those used in the proof of Lemma A.1,

$$\frac{1}{n}\mathrm{var}\left(\varsigma_{1,t}^{(1)}\right) = O\left(||h||^4 + ||\lambda||^2\right) = o(1)$$

and

$$\sum_{j=1}^{n-1} |\mathrm{cov}(\varsigma_{1,t}^{(1)}, \varsigma_{j,t}^{(1)})| = o(1)$$

which implies that $\mathrm{var}((1/n)\sum_{i=1}^{n}\varsigma_{i,t}^{(1)}) = o(1)$. Similarly, one can show that the variance of the other elements in B_n is $o(1)$. The conclusion then follows by the Chebyshev's inequality. ■

By Lemmas A.1–A.3,

$$HH_1(\widehat{\theta} - \theta) - B^{-1}b(h, \lambda) \xrightarrow{d} N(0, B^{-1}\Gamma B^{-1})$$

This completes the proof of Theorem 1.

Proof of Corollary 1. Since the proof parallels that of Theorem 1, we only sketch the difference. Recall $S_n(u)$ is defined in (A.1). When $u^{\mathrm{c}} = vh$, we have

$$\begin{aligned} E[S_{n,0}(u)] &= E[X_i X_i' K_{iu}] \\ &= \int \Omega(u^{\mathrm{c}} + hz, u^{\mathrm{d}}) f_u(u^{\mathrm{c}} + hz, u^{\mathrm{d}}) W(z) dz + O(||\lambda||) \\ &= \Omega(0, u^{\mathrm{d}}) f_u(0, u^{\mathrm{d}}) \iota_{v0} + o(1) \end{aligned}$$

where ι_{v0} is defined after Eq. (11). Similarly, $E[S_{n,1}(u)] = \Omega(0, u^{\mathrm{d}}) f_u(0, u^{d}) \iota_{v1} + o(1)$, and $E[S_{n,2}(u)] = \Omega(0, u^{\mathrm{d}}) f_u(0, u^{\mathrm{d}}) \iota_{v2} + o(1)$. It follows that

$$S_n(u^{\mathrm{c}}, u^{\mathrm{d}}) \xrightarrow{d} S_v \otimes \Omega(0, u^{\mathrm{d}}) f_u(0, u^{\mathrm{d}}) \tag{A.10}$$

where S_v is defined in (11). Following the proof of Lemma A.2, with $u^{\mathrm{c}} = vh$ we can show that

$$\mathrm{var}(HT_{n,1}) = \Gamma_v \otimes \Omega^*(0, u^{\mathrm{d}}) f_u(0, u^{\mathrm{d}}) + o(1) \tag{A.11}$$

where Γ_v is defined in Eq. (11). Following the proof of Lemma A.3, when $u^c = vh$, we have

$$b_{n,1} = \frac{1}{2}HE\left\{\begin{pmatrix}\Omega(U_i)A(U_i,u)\\ \Omega(U_i)A(U_i,u)((U_i^c-u^c)/h)\end{pmatrix}W_{h,iu}\big|d_{u_iu}=0\right\}$$
$$\times p(u^d)+o(1)$$
$$= \frac{H}{2}\begin{pmatrix}\Omega(0,u^d)\bar{A}(0,u^d)\iota_{v2}\\ \Omega(0,u^d)\bar{A}(0,u^d)\iota_{v3}\end{pmatrix}f_u(0,u^d)+o(1) \tag{A.12}$$

and

$$b_{n,2} = HE\left\{\begin{pmatrix}\Omega(U_i)(\boldsymbol{a}(U_i)-\boldsymbol{a}(u))-((U_i^c-u^c)/h)\Omega(U_i)\boldsymbol{b}(u)\\ \Omega(U_i)(\boldsymbol{a}(U_i)-\boldsymbol{a}(u))((U_i^c-u^c)/h)-((U_i^c-u^c)/h)^2\Omega(U_i)\boldsymbol{b}(u)\end{pmatrix}\right.$$
$$\left.\times K_{iu}|d_{u_iu}=1\right\}\times p(d_{u_iu}=1)+o(1)$$
$$= H\sum_{\tilde{u}^d\varepsilon D}\sum_{s=1}^{q}\lambda_s I_s(u^d,\tilde{u}^d)f_u(0,\tilde{u}^d)$$
$$\times\begin{pmatrix}\Omega(0,\tilde{u}^d)\{\iota_{v0}[\boldsymbol{a}(0,\tilde{u}^d)-\boldsymbol{a}(0,u^d)]-\iota_{v1}\boldsymbol{b}(0,u^d)\}\\ \Omega(0,\tilde{u}^d)\{\iota_{v1}[\boldsymbol{a}(0,\tilde{u}^d)-\boldsymbol{a}(0,u^d)]-\iota_{v2}\boldsymbol{b}(0,u^d)\}\end{pmatrix}+o(1) \tag{A.13}$$

where $\bar{A}(0,u^d)$ is defined in Eq. (12). Combining Eqs. (A.10)–(A.13) yields the desired result.

PART III
EMPIRICAL APPLICATIONS OF NONPARAMETRIC METHODS

THE EVOLUTION OF THE CONDITIONAL JOINT DISTRIBUTION OF LIFE EXPECTANCY AND PER CAPITA INCOME GROWTH

Thanasis Stengos, Brennan S. Thompson and Ximing Wu

ABSTRACT

In this paper we investigate the joint conditional distribution of health (life expectancy) and income growth, and its evolution over time. The conditional distributions of these two variables are obtained by applying non-parametric methods to a bivariate non-parametric regression system of equations. Analyzing the distributions of the non-parametric fitted values from these models we find strong evidence of movement over time and strong evidence of first-order stochastic dominance of the earlier years over the later ones. We also find strong evidence of second-order stochastic dominance by non-OECD countries over OECD countries in each period. Our results complement the findings of Wu, Savvides and Stengos (2008) who explored the unconditional behaviour of these joint distributions over time.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 171–191

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025008

1. INTRODUCTION

Even though the concept of human development is a very broad concept, it certainly would include health and standard of living as two of its fundamental components. The Human Development Report, first published in 1990, includes the United Nations Development Programme report of a composite index for each member country's average achievements. This index, the Human Development Index (HDI), covers three basic dimensions of human development: health, education and standard of living.

An important question for policy makers is how to improve health, especially in developing countries. Many researchers (see Caldwell, 1986; Musgrove, 1996) argue that development should focus on income growth, since higher incomes indirectly lead to health improvements. Others, Anand and Ravallion (1993) and Bidani and Ravallion (1997) take the stand that income growth alone is not enough as people's ability to function and perform in their economic tasks is affected by their health status and not the other way around. We intend to contribute to this debate by looking at the evolution of per capita income and health as measured by life expectancy over time for a number of countries over a 30-year period.

According to recent World Bank data, over the last 40 years, the world's real GDP has increased by more than 100 percent although there exist important differences among individual country experiences. For the richest country quartile this increase is more than 150 percent, whereas for the poorest quartile this number was only 50 percent. Extreme poverty (the share of population living on less than $1 per day) in developing countries has fallen by about 20 percent over the last 10 years alone, especially in East and South Asia where the accelerating growth of China and India has propelled these regions to be well within the target of the Millennium Development Goals to reduce in half the fraction of people below the cutoff of $1 per day by 2015. Between 1960 and 2000 average life expectancy has increased by 15 years and infant mortality has fallen by more than 50 percent around the world, giving hope that the Millennium Development Goal of reducing infant and child mortality rates to one-third of their 1990 levels would be met.

The rapid health improvements over the last 40 years raise the question of the driving forces behind this trend. Most of the empirical studies (see, e.g. Musgrove, 1996; Filmer & Pritchett, 1999) assume that health improvements are the by-product of higher income as countries with higher income devote more resources for their health services, something that would translate into improved health status for their population.

One of the earlier benchmark studies of the income–health relationship is Preston (1975) who compared different countries' life expectancy and per capita income for different benchmark years (1900, 1930 and 1960) and proposed the 'Preston curve', a non-linear and concave empirical relationship between the two. The concave Preston curve has provided the rationale for much of the empirical work that has followed. However, simple health–per capita income relationships may suffer from endogeneity, especially when it comes to countries on the flat portion of the Preston curve, where health has reached such an advanced stage where additional improvements coming from income growth cannot be attained. In that case it would be the reverse impact from health to income that would be important. Papers such as Pritchett and Summers (1996) address this issue by relying on an instrumental variable (IV) methodology. However, the difficulty here is the choice of instruments as many of those chosen as instruments may not be appropriate or may be weak, for example, the investment ratio (ratio of investment to GDP) will itself be endogenous in a health-type production function.

In a recent paper, Maasoumi, Racine, and Stengos (2007) (MRS hereafter) examined the entire distribution of income growth rates, as well as the distributions of parametrically and non-parametrically fitted and residual growth rates relative to a space of popular conditioning variables in this literature. In that respect they were able to compare convergence in distribution and 'conditional convergence' as they introduced some entropy measures of distance between distributions to statistically examine the question of convergence or divergence. This approach can be viewed as alternative quantifications within a framework of distributional dynamics discussed in Quah (1993, 1997). Quah focused on the distribution of per capita incomes (and relative incomes) by introducing a measure of 'transition probabilities', the stochastic kernel, to analyze their evolution. The MRS paper's focus on significant features of the probability laws that generate growth rates goes beyond both the standard 'β-convergence' and 'σ-convergence' in the literature (see Barro & Sala-i-Martin, 2004). The former concept refers to the possible equality of a single coefficient of a variable in the conditional mean of a distribution of growth rates. The latter, while being derivative of a commonplace notion of 'goodness of fit', also is in reference to the mere fit of a conditional mean regression, and is plagued with additional problems when facing non-linear, non-Gaussian or multi-modal distributions commonly observed for growth and income distributions. As has been pointed out by Durlauf and Quah (1999), the dominant focus in these studies is on certain aspects of estimated conditional means,

such as the sign or significance of the coefficient of initial incomes, how it might change if other conditioning variables are included, or with other functional forms for the production function or regressions. All of the above studies rely on 'correlation' criteria to assess goodness of fit and to evaluate 'convergence'.

In the first study to use a bivariate framework, Wu, Savvides, and Stengos (2008) (WSS hereafter) investigate the unconditional evolution of income per capita and life expectancy using a maximum entropy density estimator. They consider income and life expectancy jointly and estimate their unconditional bivariate distribution for 137 countries for the years 1970, 1980, 1990 and 2000. Their main conclusion is that the world joint distribution has evolved from a bimodal into a unimodal one, that the evolution of the health distribution has preceded that of income and that global inequality and poverty has decreased over time. They also find that global inequality and poverty would be substantially underestimated if the dependence between the income and health distributions is ignored.

In this paper we extend the work of WSS by estimating the joint conditional distribution of health (life expectancy) and income growth, and we examine its evolution over time. The conditional distributions of these two variables are obtained by applying non-parametric regression methods. This generalizes the MRS approach to a multidimensional context. Using a similar data set as WSS, we extend their analysis to go beyond unconditional distributions. As in the MRS univariate framework we will be examining conditional distributions by looking into a bivariate system of per capita income growth and life expectancy growth equations. We will then analyze the distributions of parametrically and non-parametrically fitted values and residuals from these models using a bivariate growth framework relative to the standard conditioning variables that are employed in the literature. The resulting analysis produces 'fitted values' of growth rates and life expectancy as well as 'residual growth rates and life expectancy', which will be used to look at the question of 'conditional' convergence in a bivariate context. Note that in contrast with the WSS study, which was conducted for the unconditional joint distribution of per capita income and life expectancy in levels, our approach will be based on analyzing the conditional joint distribution of growth rates, which provides new insight into the driving forces of their joint evolution over time.

The paper is organized as follows. In Section 2 we discuss the data used. We then proceed to discuss in Section 3 the empirical methodology and results of both the parametric and non-parametric approaches that we pursue. Finally, we conclude in Section 4.

2. DATA

To estimate the global joint distribution of income and life expectancy, we collected data on 124 countries to construct 10-year averages for the 1970s, the 1980s and the 1990s for a total of 372 observations. These countries account for approximately 80 percent of global population. Below we describe in more detail the data that we use and their source. Similar data have been used by WSS.

Data on income per capita are in PPP dollars from the *Penn World Tables 6.2*, and they are used to construct the real per capita GDP growth. This data base provides estimates in 2000 international prices for most countries beginning in 1950 until 2004.

For each country in our sample, the income information is reported in the form of interval summary statistics. In particular, the frequency and average income of each interval are reported. The number of income intervals differs between the first three years (1970, 1980 and 1990) and the final year (2000). Since we construct an average over a 10-year period we do not need to have the same number of intervals to be the same in each year. For 1970, 1980 and 1990, we used income interval data from Bourguignon and Morrisson (2002). We construct an average income observation for each country for each 10-year period. An alternative source of income data for these years would have been the World Development Indicators (WDI). There are two reasons for using the Bourguignon/Morrisson data set: first, it provides a greater number of intervals and thus more detailed information on income distribution; and, second, our results on income distribution can be compared to earlier studies.[1] For 2000, Bourguignon/Morrisson do not provide data and we used income interval data from the WDI.[2] These data are based on household surveys of income (in some cases consumption) from government statistical agencies and World Bank country departments.

Data on life expectancy *at birth* are also in the form of interval statistics. The most detailed division of each country's population by age is in 5-year intervals from the *World Population Prospects* compiled by the Population Division of the United Nations Department of Economic and Social Affairs (2005). This is the most comprehensive collection of demographic statistics. For each of the 124 countries, it provides data on the number of persons in each age group for each of the four years (1970, 1980, 1990 and 2000). The U.N. Population Division begun compiling estimates of life expectancy at 5-year intervals in 1950. For each country we constructed average life expectancy over the relevant 10-year period. For more details about the data construction, see the WSS study.

3. EMPIRICAL RESULTS

In this paper, we use both parametric and non-parametric techniques to estimate a bivariate system of equations that describe real per capita growth and life expectancy growth. The framework of analysis is an extension of the MRS framework to account for the simultaneous evolution of per capita income and life expectancy. We proceed by first estimating a bivariate system of equations parametrically and then continue with the non-parametric analysis.

3.1. Parametric Results

We first consider a bivariate parametric system of seemingly unrelated regressions (SUR) to model the growth path of per capita income and life expectancy. The dependent variables are $Y = (Y_1, Y_2)$, where Y_1 is real GDP per capita growth and Y_2 is life expectancy growth. For each country-year, the list of independent variables is given by $X = (X_1, X_2, \ldots, X_7)$, where X_1 is a dummy variable indicating OECD status, X_2 is a dummy for the 1980s, X_3 is a dummy for the 1990s, X_4 is the log of population growth plus 0.05 to account for a constant rate of technical change of 0.02 and a depreciation rate of 0.03, X_5 is the log of investment share of GDP, X_6 is the log of real GDP at the start of the period and X_7 is the log of life expectancy at the start of the period. The last two variables capture initial conditions and their effect on the transition to a steady state. The specification of the equation describing the evolution of per capita income is a standard growth regression of an extended Solow-type model; the evolution of life expectancy is modelled in a symmetric way.

We begin by estimating a simple benchmark bivariate parametric regression model that is standard for the bivariate extension of the standard workhorse model of the empirical literature,

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon_1 \\ Y_2 &= \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \beta_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 X_6 + \gamma_7 X_7 + \varepsilon_2 \end{aligned} \tag{1}$$

We estimate the above system of equations as an SUR. However, since the right-hand-side variables are identical in the two equations, GLS is identical to estimating each equation separately by OLS. Note that, in each equation, both GDP per capita and life expectancy enter in lagged (i.e., initial) values to guard against endogeneity.

The parameter estimates for specification (1) are given in Tables 1 and 2, and are in line with results from the extensive univariate growth literature. For the per capita income growth regression, we find investment having a positive effect on growth, while population growth seems to have a negative effect. Initial GDP has a negative effect on growth (although not statistically significant) suggesting the presence of (statistically weak) conditional or β-convergence. The initial life expectancy variable also turns out to be statistically insignificant. In the context of an income growth regression, life expectancy stands for a proxy for human capital and as such the latter often does not appear significant in parametric specifications, especially with panel data (see Savvides & Stengos, 2008).

In the life expectancy growth equation, investment is also positive and significant, while population growth is positive but not highly significant.

Table 1. Parameter Estimates for GDP Per Capita Growth Linear Regression.

	Estimate	Std. Error	t-Value	Pr($>\|t\|$)
(Intercept)	3.5256	4.2270	0.83	0.4048
oecd1	−0.7596	0.3882	−1.96	0.0511
d1980	−1.3791	0.2852	−4.84	0.0000
d1990	−1.0124	0.3088	−3.28	0.0011
pop	−2.8323	0.9554	−2.96	0.0032
inv	1.5212	0.2318	6.56	0.0000
initY	−0.0423	0.0668	−0.63	0.5265
initL	0.2787	0.9392	0.30	0.7668

Table 2. Parameter Estimates for Life Expectancy Growth Linear Regression.

	Estimate	Std. Error	t-Value	Pr($>\|t\|$)
(Intercept)	2.0159	0.5371	3.75	0.0002
oecd1	−0.1911	0.0493	−3.87	0.0001
d1980	−0.1053	0.0362	−2.91	0.0039
d1990	−0.2767	0.0392	−7.05	0.0000
Pop	0.2118	0.1214	1.74	0.0819
Inv	0.1556	0.0295	5.28	0.0000
initY	0.0092	0.0085	1.08	0.2805
initL	−0.5441	0.1193	−4.56	0.0000

The initial life expectancy variable has a strongly negative effect which would seem to imply β-convergence in health outcomes. Initial GDP has a significant effect.

Despite its use in the literature, there is evidence that the above parametric linear specification (1) is inadequate and misspecified, especially when it comes to describing the effect of initial conditions on the growth process. Following the per capita income growth literature we allow the initial condition variables X_6 and X_7 to enter as third-degree polynomials (see Liu & Stengos, 1999), that is,

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \\ &\quad + \beta_6 X_6 + \beta_7 X_6^2 + \beta_8 X_6^3 + \beta_9 X_7 + \beta_{10} X_7^2 + \beta_{11} X_7^3 + \varepsilon_1 \\ Y_2 &= \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \beta_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 \\ &\quad + \gamma_6 X_6 + \gamma_7 X_6^2 + \gamma_8 X_6^3 + \gamma_9 X_7 + \gamma_{10} X_7^2 + \gamma_{11} X_7^3 + \varepsilon_2 \end{aligned} \tag{2}$$

The results from the above parametric SUR system are given in Tables 3 and 4. These results are in line with results from the simple parametric specification (1) discussed above. Investment is found to positively affect both per capita GDP and life expectancy growth. Population growth has a negative effect on GDP per capita growth, but an insignificant effect on life expectancy growth. Interestingly, in both of the equations, none of the polynomial terms for either initial GDP per

Table 3. Parameter Estimates for GDP Per Capita Growth Polynomial Regression.

	Estimate	Std. Error	t-Value	Pr($>\|t\|$)
(Intercept)	−973.9962	869.5863	−1.12	0.2634
oecd1	−0.6090	0.4550	−1.34	0.1816
d1980	−1.4345	0.2860	−5.02	0.0000
d1990	−1.0460	0.3183	−3.29	0.0011
Pop	−3.2726	1.0003	−3.27	0.0012
Inv	1.3891	0.2417	5.75	0.0000
initY	−2.4379	5.8822	−0.41	0.6788
initY^2	0.1012	0.3479	0.29	0.7712
initY^3	−0.0012	0.0068	−0.18	0.8590
initL	744.6284	676.8577	1.10	0.2720
initL^2	−184.8683	174.8208	−1.06	0.2910
initL^3	15.2578	15.0334	1.01	0.3108

Table 4. Parameter Estimates for Life Expectancy Growth Polynomial Regression.

	Estimate	Std. Error	t-Value	Pr(>|t|)
(Intercept)	−148.8768	111.0005	−1.34	0.1807
oecd1	−0.1997	0.0581	−3.44	0.0007
d1980	−0.1133	0.0365	−3.10	0.0021
d1990	−0.2873	0.0406	−7.07	0.0000
Pop	0.1723	0.1277	1.35	0.1780
Inv	0.1409	0.0309	4.57	0.0000
initY	−0.7304	0.7508	−0.97	0.3313
initY^2	0.0402	0.0444	0.91	0.3653
initY^3	−0.0007	0.0009	−0.83	0.4094
initL	118.3105	86.3992	1.37	0.1717
initL^2	−30.2379	22.3154	−1.36	0.1763
initL^3	2.5600	1.9190	1.33	0.1830

capita or initial life expectancy appear to be significant, which may suggest overparameterization.

We next test these parametric specifications against some unknown non-parametric alternative. If we denote the parametric model given by the above system of equations as $m_g(x_i, \beta)$, $g = 1, 2$ and the true but unknown regression functions by $E_g(y_{gi}|x_i)$, $g = 1, 2$, then a test for correct specification is a test of the hypothesis H_0: $E_g(y_{gi}|x_i) = m_g(x_i, \beta)$, $g = 1, 2$ almost everywhere versus the alternative H_1: $E_g(y_{gi}|x_i) \neq m_g(x_i, \beta)$, $g = 1, 2$ on a set of positive measure. That is equivalent to testing that $E_g(\varepsilon_{gi}|x_i) = 0$ almost everywhere, where $\varepsilon_{gi} = y_{gi} - m_g(x_i, \beta)$. This implies that for an incorrect specification, $E_g(\varepsilon_{gi}|x_i) \neq 0$ on a set of positive measure. It is important to note that this test is not a joint test, that is, the test is applied to each equation separately.

To avoid problems arising from the presence of a random denominator in the non-parametric estimator of the regression functions $E_g(y_{gi}|x_i)$, the test employs a density weighted estimator of the regression function. To test whether $E_g(\varepsilon_{gi}|x_i) = 0$ holds over the entire support of the regression function, we use the statistic $J = E_g\{[E_g(\varepsilon_{gi}|x_i)]^2 f(x_i)\}$ where $f(x_i)$ denotes the density weighting function. Note that $J = 0$ if and only H_0 is true. The sample analogue of J, J_n is obtained by replacing ε_{gi} with the residuals from the parametric model and both $E_g(\varepsilon_{gi}|x_i)$ and $f(x_i)$ by their respective kernel estimates, and standardizing. The null distribution of the statistic is obtained via bootstrapping (see Hsiao, Li, & Racine, 2008 for details).

For specification (1), we are able to reject the null of correct specification at the 5% and 1% levels, for the income and life expectancy growth equations, respectively (the test statistics J_n are 0.6919 and 4.411, with bootstrap p-values of 0.0276 and 0.0025, respectively). Similarly, for (2), we are able to reject at the 5% and 0.1% levels, for the income and life expectancy growth equations, respectively (the test statistics J_n are 0.3658 and 2.1892, with bootstrap p-values of 0.0401 and 2.22e-16, respectively). We use 399 bootstrap replications throughout the paper.

3.2. Non-Parametric Results

Next, we use local linear estimation to (separately) estimate the non-parametric regression models

$$Y_1 = g_1(X) + \varepsilon_1$$
$$Y_2 = g_2(X) + \varepsilon_2$$

We use least squares cross-validation techniques to obtain the appropriate bandwidths for the discrete and continuous regressors (see Racine & Li, 2004). This approach allows for interactions among all variables and also allows for non-linearities in and among variables. The method has the additional feature that if there is a linear relationship in a variable, then the cross-validated smoothing parameter will automatically detect this. A second-order Gaussian kernel is used for the continuous variables, while the Aitchison and Aitken kernel is used for the unordered categorical variable (OECD status) and the Wang and Van Ryzin kernel is used for the ordered categorical variable (decade). For details, see Racine and Li (2004).

In Figs. 1–4, we summarize the non-parametric results using partial regression plots. These plots simply present the estimated multivariate regression function through a series of bivariate plots in which the regressors not appearing on the horizontal axis of a given plot have been held constant at their respective (within group and decade) medians. For example, in the upper-left plot in Fig. 1, we plot the estimated level of GDP per capita growth conditioned on population growth for just OECD members in the 1970s holding all the other conditioning variables at their respective median levels for OECD members in the 1970s (the estimates are obtained using the pooled sample of OECD and non-OECD members, but the fitted values are plotted for each group separately). In this way we are able to visualize the multivariate regression surface via a series of two-dimensional plots.

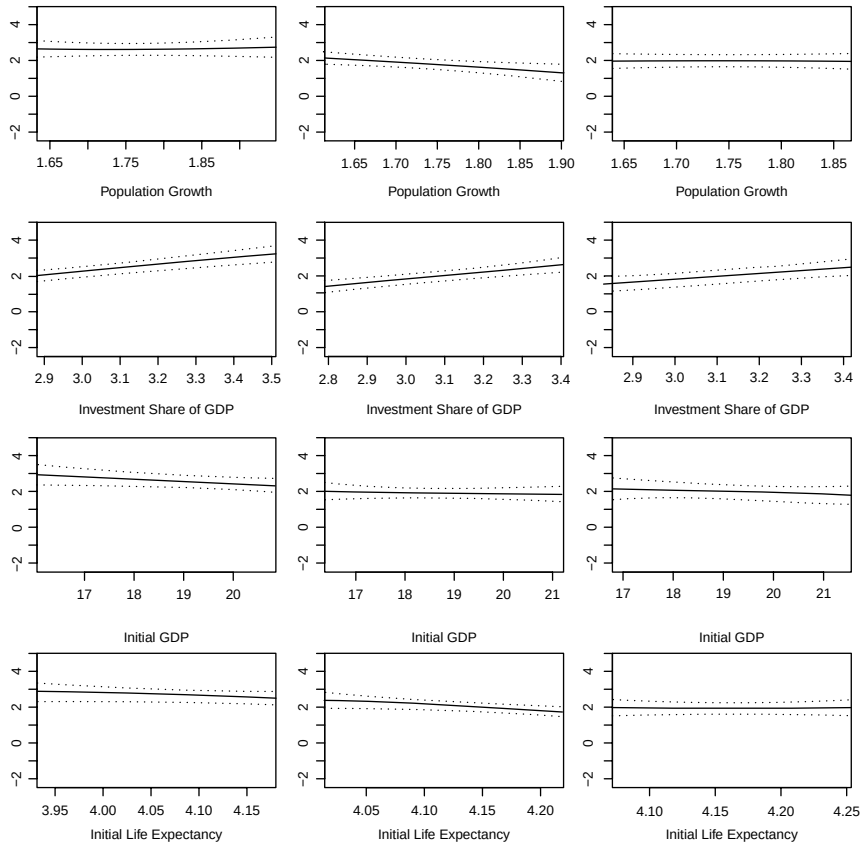


Fig. 1. GDP Per Capita Growth Non-Parametric Partial Regression Plots for OECD Countries. The First, Second and Third Columns are for the 1970s, 1980s and 1990s, Respectively.

The level of investment appears to have a (linearly) positive and stable effect across decade and country group for both equations. Population growth appears to be unrelated to the dependent variables except in the 1980s, where it is slightly negative for the GDP per capita growth equation and slightly positive for the life expectancy growth equation (for both OECD and non-OECD members). For the GDP per capita growth equation, initial GDP appears to have a slightly negative effect in the 1970s, but little effect in either the 1980s or the 1990s (for both OECD and non-OECD members). For OECD members, initial life expectancy seems to

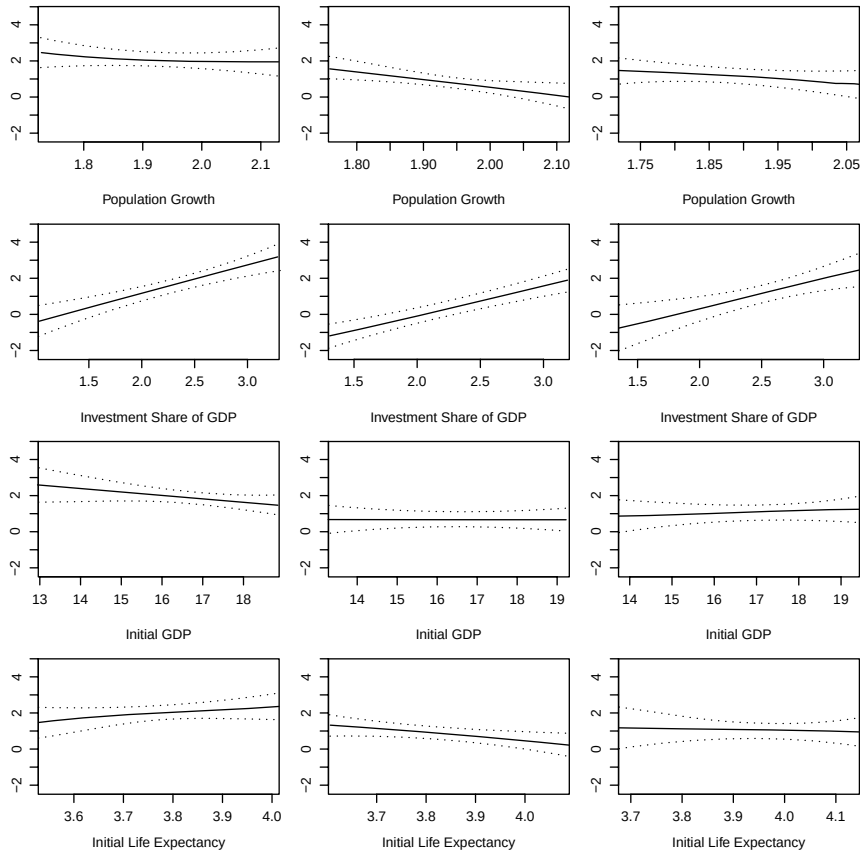


Fig. 2. GDP Per Capita Growth Non-Parametric Partial Regression Plots for Non-OECD Countries. The First, Second and Third Columns are for the 1970s, 1980s and 1990s, Respectively.

have a negative effect on GDP per capita growth in the 1980s, but little effect in the other decades. However, for non-OECD members, the effect of initial life expectancy on GDP per capita growth is mixed: The effect seems to be positive in the 1970s, negative in the 1980s and non-existing in the 1990s. For the life expectancy growth equation, initial GDP appears to have a slight negative effect in all decades and groups. However, initial life expectancy appears to have a generally negative, but non-linear effect in all decades and groups.

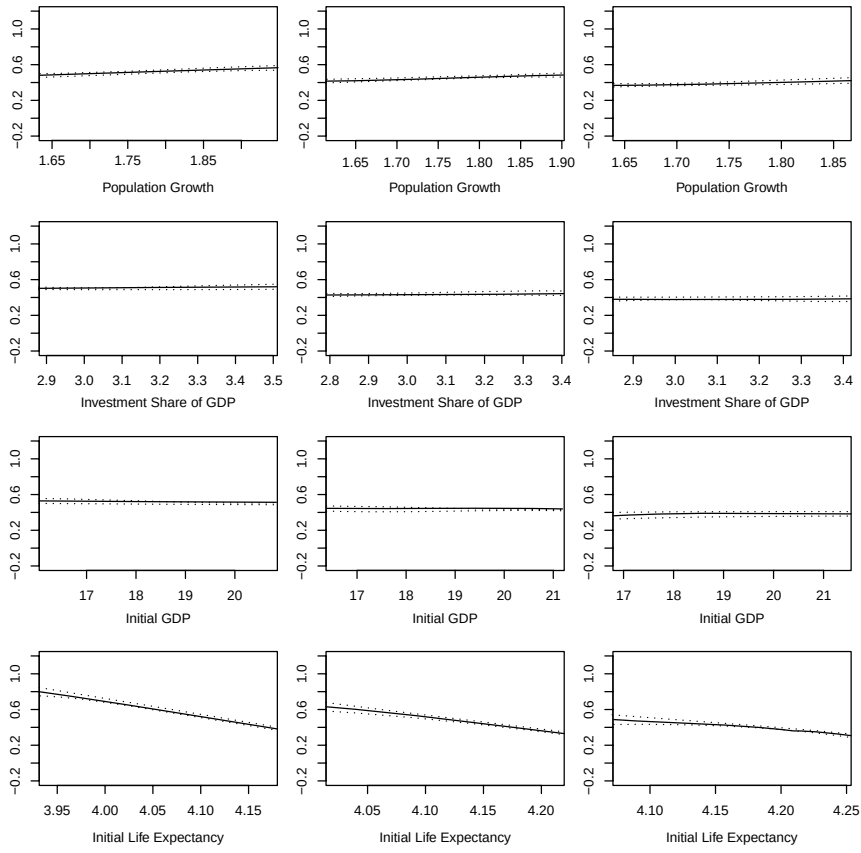


Fig. 3. Life Expectancy Growth Non-Parametric Partial Regression Plots for OECD Countries. The First, Second and Third Columns are for the 1970s, 1980s and 1990s, Respectively.

To further examine how the joint distribution of per capita GDP and life expectancy growth rates differ between groups and over time, we use the notion of stochastic dominance, which is defined as follows. We say distribution G stochastically dominates distribution F at first order if

$$F(x_1, x_2) \geq G(x_1, x_2)$$

for all (x_1, x_2).

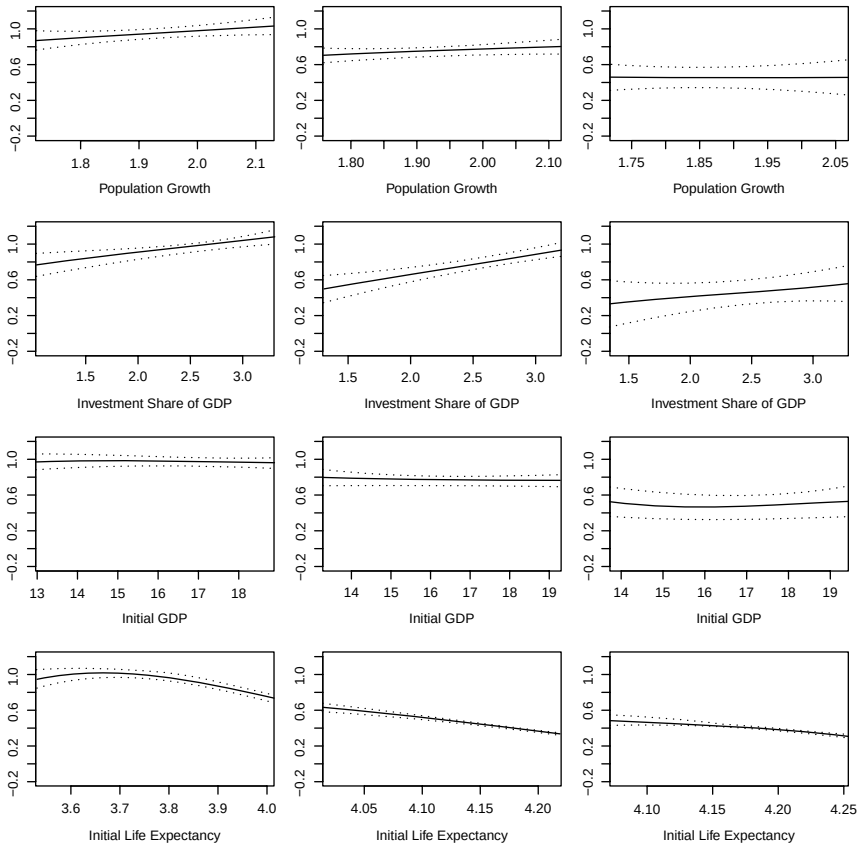


Fig. 4. Life Expectancy Growth Non-Parametric Partial Regression Plots for Non-OECD Countries. The First, Second and Third Columns are for the 1970s, 1980s and 1990s, Respectively.

More generally, we can say that distribution F dominates distribution G stochastically at order s (an integer) if

$$D_F^s(x_1, x_2) \leq D_G^s(x_1, x_2)$$

for all (x_1, x_2), where $D_F^1(x_1, x_2) = F(x_1, x_2)$, and $D_F^s(x_1, x_2)$ is defined recursively as

$$D_F^s(x_1, x_2) = \int_0^{x_1} \int_0^{x_2} D_F^{s-1}(u_1, u_2) du_1 du_2, \quad s \geq 2$$

D_G^1 and D_G^s are defined analogously. In what follows, we will denote this relation by $F \succcurlyeq_s G$.

To empirically test such a relationship, we use the approach of McCaig and Yatchew (2007). To test the null hypothesis that $F \succcurlyeq_s G$, these authors introduce the test statistic

$$T_{F,G} = \left\{ \iint [\psi^s(x_1, x_2)]^2 dv_1 dv_2 \right\}^{1/2}$$

where $\psi^s(x_1, x_2) = \max\{D_F^s(x_1, x_2) - D_B^s(x_1, x_2), 0\}$. Of course, when the null is true, T is equal to zero.

In practice, this test involves estimating T and testing whether it is statistically different from zero. This process will involve estimating $\psi^s(x_1, x_2)$ over a set of grid points on the common support of the two distributions under consideration. The p-value of this test statistic is obtained via bootstrapping (see McCaig & Yatchew, 2007, for details).

To make such comparisons in a conditional manner, we use the fitted values from the non-parametric regressions considered above. The estimated joint density and distribution functions of these fitted values are shown in Figs. 5 and 6, respectively. We separate the observations by group and decade; that is, we consider six unique groupings (OECD and non-OECD members for the 1970s, OECD and non-OECD members for the 1980s and OECD and non-OECD members for the 1990s). As seen in Fig. 5, the distribution of bivariate conditional growth rates has become more concentrated within each group (OECD and non-OECD members) over time. Also, it is interesting to note that the (conditional) GDP per capita growth rates tend to be higher among OECD members, but that the (conditional) life expectancy growth rates tend to be higher among non-OECD members. However, these differences appear to be diminishing over time.

We now proceed to test for stochastic dominance of the fitted (conditional) bivariate growth rates between the two groups of countries under consideration: OECD members and non-OECD members. The values of the test statistics and their bootstrap p-values are presented in Table 5. As can be seen, we can strongly reject the null of first-order stochastic dominance of OECD members over non-OECD members (and vice-versa) in each of the three decades under consideration. We can also strongly reject the null of second-order stochastic dominance of OECD members over non-OECD members in each of the three decades, but not vice-versa. That is, we

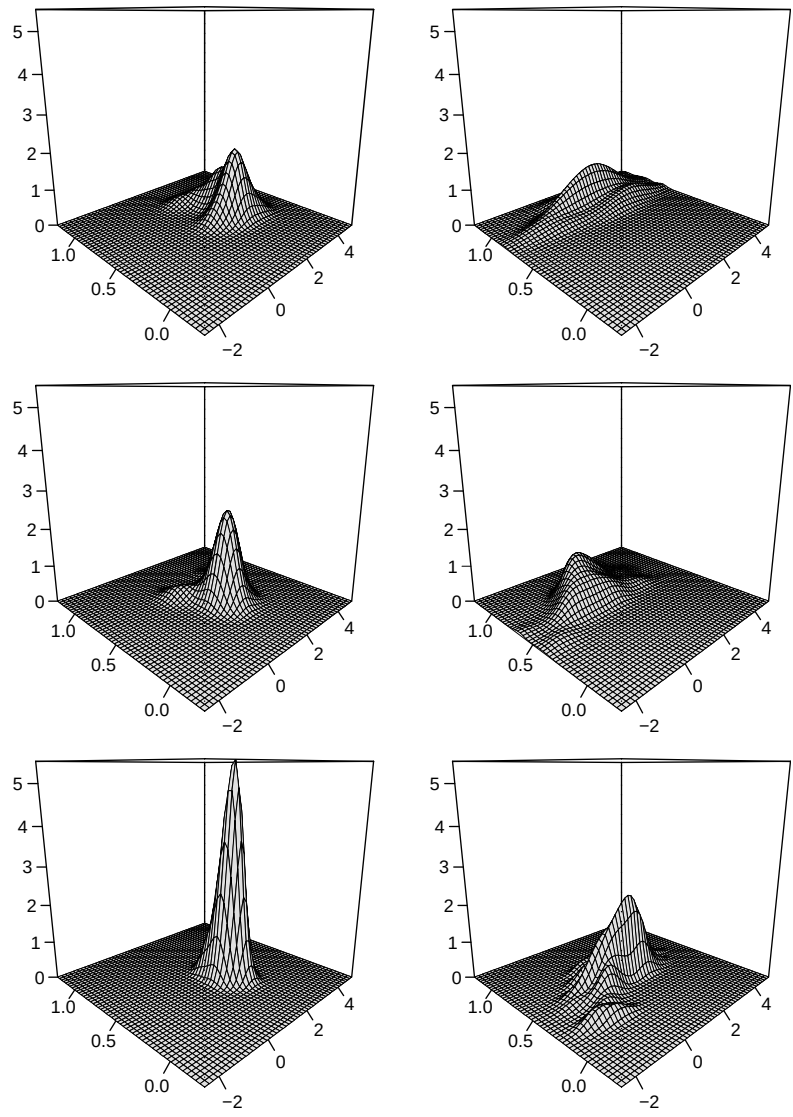


Fig. 5. Estimated Density Functions for the Fitted Values from the Non-Parametric Regressions. The Left Column is for OECD Countries, While the Right Column is for Non-OECD Countries. The First, Second and Third Rows are for the 1970s, 1980s and 1990s, Respectively. Within Each Plot, the Lower-Left Axis is for the Fitted Values from the Life Expectancy Growth Non-Parametric Regression, While the Lower-Right Axis is for the Fitted Values from the GDP Per Capita Growth Non-Parametric Regression.

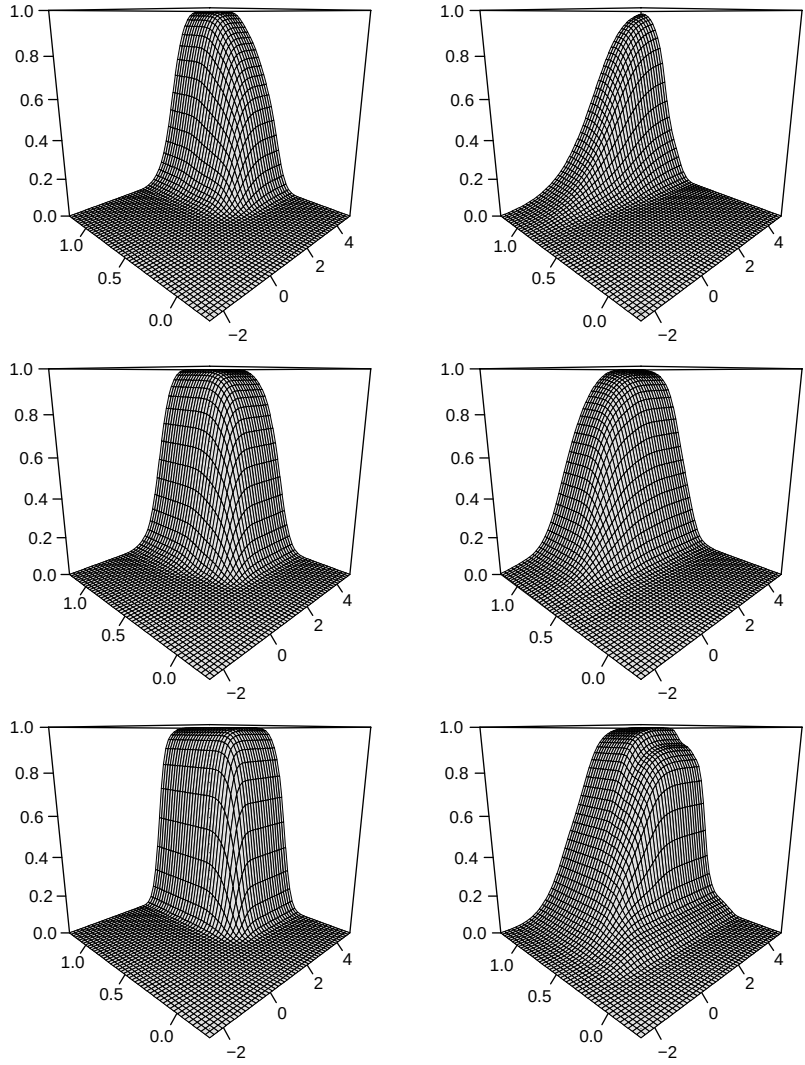


Fig. 6. Estimated Distribution Functions for the Fitted Values from the Non-Parametric Regressions. The Left Column is for OECD Countries, While the Right Column is for Non-OECD Countries. The First, Second and Third Rows are for the 1970s, 1980s and 1990s, Respectively. Within each Plot, the Lower-left Axis is for the Fitted Values from the Life Expectancy Growth Non-Parametric Regression, While the Lower-right Axis is for the Fitted Values from the GDP Per Capita Growth Non-Parametric Regression.

Table 5. Stochastic Dominance Tests: Between Groups.

	1970s	1980s	1990s
OECD $\succcurlyeq_1$ Non-OECD	4.2802	3.4902	1.8601
	0.0000	0.0000	0.0000
Non-OECD $\succcurlyeq_1$ OECD	1.0816	1.9978	2.1026
	0.0404	0.0000	0.0000
OECD $\succcurlyeq_2$ Non-OECD	27.8089	23.6515	10.0243
	0.0000	0.0000	0.0000
Non-OECD $\succcurlyeq_2$ OECD	0.0742	0.0513	1.761
	0.9495	0.9596	0.5051

Note: For each result, the first line is the value of the test statistic, while the second line is the bootstrap *p*-value.

Table 6. Stochastic Dominance Tests: Between Decades.

	OECD	Non-OECD
1970s $\succcurlyeq_1$ 1980s	0.0000	0.0000
	0.9393	0.9899
1980s $\succcurlyeq_1$ 1970s	2.0061	3.1569
	0.0000	0.0000
1980s $\succcurlyeq_1$ 1990s	0.4374	0.4291
	0.4545	0.2929
1990s $\succcurlyeq_1$ 1980s	1.3846	3.0810
	0.0000	0.0000
1970s $\succcurlyeq_1$ 1990s	0.0000	0.0000
	0.8788	0.9899
1990s $\succcurlyeq_1$ 1970s	2.9433	5.2524
	0.0000	0.000

Note: For each result, the first line is the value of the test statistic, while the second line is the bootstrap *p*-value.

are unable to reject the null of second-order stochastic dominance of non-OECD members over OECD members.

Next, we consider testing for first-order stochastic dominance of the same fitted values between the three decades under consideration: the 1970s, 1980s and 1990s. The values of the test statistics and their bootstrap *p*-values are presented in Table 6. For both the OECD and non-OECD groups, we are unable to reject the null of first-order stochastic dominance of the 1970s

over the 1980s, and the 1980s over the 1990s (and, of course, the 1970s over the 1990s).

These results somewhat agree with the findings of MRS, who show that the fitted (conditional) growth rates of per capita income have 'deteriorated' over time for OECD countries. However, we also want to point out that the MRS analysis is univariate, and as pointed out in WSS the overall results will underestimate substantially the degree of global inequality and poverty if one ignores the dependence between the two measures of welfare. Note, however, that the later analysis was conducted for the unconditional joint distribution of per capita income and life expectancy (levels), whereas here we analyze the conditional joint distribution of growth rates. The implication is that there was a more 'equal' joint distribution of growth rates in the earlier years than that in the later ones, not necessarily faster growth in the earlier years. Note that the interpretation of this result for growth rates is different from that for levels. For the case of the joint distribution of growth rates, the results suggest that in the earlier years 'convergence' between developing and more developed countries would be more difficult to achieve since countries in these groups would be growing more or less at equal rates. It is only in the later years that a more 'unequal' joint distribution of growth rates would allow for faster growing developing countries being able to catch up with slower growing developed countries. Hence, the results that we find are complementary to the ones found in WSS for levels, where the level of overall (unconditional) inequality in levels decreased over time. Overall, it seems that countries developed quite differently in the 1980s and 1990s with some jumping ahead and others falling behind. We leave it for future research to further explore the issue for subgroups of countries, such as OECD and non-OECD and especially African and non-African countries (see, e.g. Masanjala & Papageorgiou, 2008).

4. CONCLUSION

In this paper we have estimated the joint conditional distribution of health (life expectancy) and income growth and examined its evolution over time. The conditional distributions of these two variables is obtained by applying non-parametric methods to a bivariate non-parametric regression system of equations. Using a similar data set as WSS, we extend their analysis to go beyond unconditional distributions. Extending the MRS univariate framework we have looked at conditional distributions of a bivariate system of per capita income growth and life expectancy growth equations. Analyzing

the distributions of the non-parametric residuals from these models we establish that there is strong evidence of movement over time in the joint conditional bivariate densities of per capita growth and life expectancy. We also find strong evidence of first-order stochastic dominance of the earlier years over the later ones. Our results complement the findings of WSS who explored the unconditional behaviour of these joint distributions over time.

ACKNOWLEDGMENTS

The authors wish to thank the participants of the 7th Annual Advances in Econometrics Conference, November 14–16, Baton Rouge, LA, for their useful comments and questions. In particular, the authors are indebted to Jeff Racine for his insightful suggestions.

NOTES

1. Bourguignon and Morrisson (2002) provide data on income distribution for almost two centuries, the last three years being 1970, 1980 and 1992. We used their 1992 income data to represent 1990 in our data set (see also the next footnote). They provided data for very few individual countries but in most cases for geographic groups of countries (see their study for group definitions). Our study is based on country-level data. Therefore, where individual-country interval data were unavailable we used the corresponding geographic-group data.
2. Income interval data from the WDI are available only for selected years. When referring to data for 2000, we chose the year closest to 2000 with available data (in most cases the late 1990s). This practice is widely adopted in the literature as a practical matter because interval data are sparse. Many researchers acknowledge that it would not affect results much because income share data do not show wide fluctuations from year to year.

REFERENCES

Anand, S., & Ravallion, M. (1993). Human development in poor countries: On the role of private incomes and public services. *Journal of Economic Perspectives*, *7*, 133–150.

Barro, R., & Sala-i-Martin, X. (2004). *Economic growth* (2nd ed.). Cambridge, MA: MIT Press.

Bidani, B., & Ravallion, M. (1997). Decomposing social indicators using distributional data. *Journal of Econometrics*, *77*, 125–139.

Bourguignon, F., & Morrisson, C. (2002). Inequality among world citizens: 1820–1992. *American Economic Review*, *92*, 727–744.

Caldwell, J. C. (1986). Routes to low mortality in poor countries. *Population and Development Review*, *12*, 171–220.

Durlauf, S. N., & Quah, D. T. (1999). The new empirics of economic growth. In: J. B. Taylor & M. Woodford (Eds), *Handbook of Macroeconomics I* (pp. 235–308). Location: Amsterdam.

Filmer, D., & Pritchett, L. (1999). The impact of public spending on health: Does money matter? *Social Science and Medicine*, *49*, 1309–1323.

Hsiao, C., Li, Q., & Racine, J. S. (2008). A consistent model specification test with mixed categorical and continuous data. *Journal of Econometrics*, *140*, 802–826.

Liu, Z., & Stengos, T. (1999). Nonlinearities in cross country growth regressions: A semiparametric approach. *Journal of Applied Econometrics*, *14*, 527–538.

Maasoumi, E., Racine, J. S., & Stengos, T. (2007). Growth and convergence: A profile of distributional dynamics and mobility. *Journal of Econometrics*, *136*, 483–508.

Masanjala, W. H., & Papageorgiou, C. (2008). Rough and lonely road to prosperity: A reexamination of the sources of growth in Africa using Bayesian model averaging. *Journal of Applied Econometrics*, *23*, 671–682.

McCaig, B., & Yatchew, A. (2007). International welfare comparisons and nonparametric testing of multivariate stochastic dominance. *Journal of Applied Econometrics*, *22*, 951–969.

Musgrove, P. (1996). *Public and private roles in health*. World Bank, Discussion Paper No. 339.

Pritchett, L., & Summers, L. (1996). Wealthier is healthier. *Journal of Human Resources*, *31*, 841–868.

Quah, D. T. (1993). Empirical cross-section dynamics in economic growth. *European Economic Review*, *37*, 426–434.

Quah, D. T. (1997). Empirics for growth and distribution: Stratification, polarization and convergence clubs. *Journal of Economic Growth*, *2*, 27–59.

Racine, J. S., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*, 99–130.

Savvides, A., & Stengos, T. (2008). *Human capital and economic growth*. Stanford, CA: Stanford University Press.

Wu, X., Savvides, A., & Stengos, T. (2008). *The global joint distribution of income and health*. Department of Economics, University of Guelph, Discussion Paper No. 2008-7.

A NONPARAMETRIC QUANTILE ANALYSIS OF GROWTH AND GOVERNANCE

Kim P. Huynh and David T. Jacho-Chávez

ABSTRACT

Conventional wisdom dictates that there is a positive relationship between governance and growth. This article reexamines this empirical relationship using nonparametric quantile methods. We apply these methods on different levels of countries' growth and governance measures as defined in World Governance Indicators provided by the World Bank. We concentrate our analysis on three of the six measures: voice and accountability, political stability, and rule of law that were found to be significantly correlated with economic growth. To illustrate the nonparametric quantile analysis we use growth profile curves as a visual device. We find that the empirical relationship between voice and accountability, political stability, and growth are highly nonlinear at different quantiles. We also find heterogeneity in these effects across indicators, regions, time, and quantiles. These results are a cautionary tale to practitioners using parametric quantile methods.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 193–221

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025009

1. INTRODUCTION

Conventional wisdom dictates that countries with higher levels of governance also have higher growth.[1] This positive relationship has motivated policy makers to implement growth policies that target change in governance.[2] However, recent work by Rodrik (2006) has highlighted that increasing governance may not necessarily increase a country's growth level. For example, improving governance may divert resources from actual binding constraints. As a result, Hausmann, Rodrik, and Velasco (2008) advocate the need to perform growth diagnostics to ascertain the binding constraints on growth.

Given what is stake for development and aid policies, robust inference regarding the relationship between governance and growth is needed. Fig. 1 illustrates the economic growth patterns for the world in 2004. As expected, Western Europe and North America have low-to-moderate rates of growth while Russia, the Former Soviet Republics, and China are experiencing high rates of growth.

The study by Kaufmann and Kraay (2002) found that per capita incomes and the quality of governance are positively correlated across countries. They adopt an instrumental variable (IV) method in order to separate the correlation into: (i) a strong positive causal effect running from better governance to higher per capita incomes, and (ii) a weak and even negative causal effect running in the opposite direction from per capita incomes to governance. However, an illustration of Rule of Law (a measure of

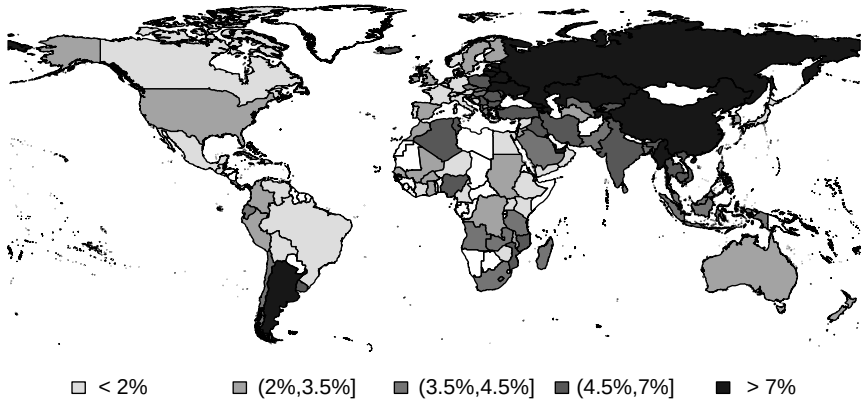


Fig. 1. Economic Growth Patterns, 2004. *Note:* Countries that are shaded in white do not have data for 2004.

governance) does not completely support this hypothesis; see Fig. 2. The Rule of Law patterns are reversed. Western Europe and North America have high rating of Rule of Law while for Russia, the Former Soviet Republics, and China the measures are extremely low. These graphs reveal that the correlation between governance and growth is not necessarily positive.

Huynh and Jacho-Chávez (2009) argue that these somehow controversial and contradictory findings can potentially be explained by the shortcomings of the parametric assumptions they rely on. The present work extends Huynh and Jacho-Chávez's (2009) framework to the nonparametric estimation of conditional quantiles functions. This is important because unlike conditional mean regression, nonparametric conditional quantiles model the relationship between governance measures at each level of growth a country might be. This provides a complete picture of the entire conditional distribution of this important relationship without imposing strict parametric restrictions. In particular the assumption of linearity, additivity, and no interaction among variables are relaxed when estimating the following object:

$$Q_{\text{growth}_{it}}[\tau|\text{REGION}_i,\ \text{DT}_t,\ \text{voice}_{it}, \text{stability}_{it}, \text{effectiveness}_{it}, \text{regulatory}_{it},\ \text{law}_{it},\ \text{corruption}_{it}] \quad (1)$$

where

$$Q_{y_{it}}[\tau|\mathbf{x}_{it}] \equiv \inf\{y_{it}|F(y_{it}|\mathbf{x}_{it}) \geq \tau\} = F^{-1}(\tau|\mathbf{x}_{it})$$

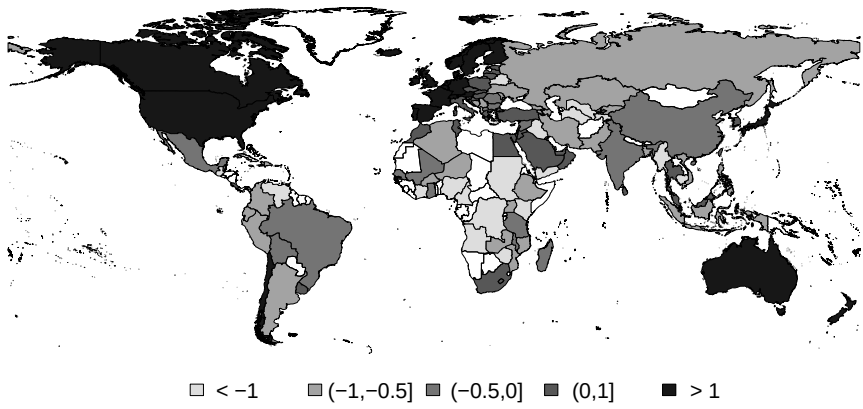


Fig. 2. Rule of Law Patterns, 2004. *Note:* Countries that are shaded in white do not have data for 2004.

represents the conditional τ-quantile of y_{it}, given $\mathbf{x}_{it}$; $F(\cdot|\cdot)$ denotes the conditional cumulative distribution function (CDF) of y_{it}, given $\mathbf{x}_{it}$; and $F^{-1}(\cdot|\cdot)$ is inverse. The conditioning variables REGION_i represents a categorical unordered variable indicating the region (1, 2, 3, 4, 5) to which country i belongs; DT_t is another ordered categorical variable indicating the year of measurement (1996, 1998, 2000, 2002, 2003, 2004, 2005, 2006); and the governance measures voice_{it}, stability_{it}, $\text{effectiveness}_{it}$, regulatory_{it}, law_{it}, and corruption_{it} are defined in Section 2.

We summarize our findings in the following two points:

- Parametric hypothesis testing indicates that the coefficients in a linear specification are the same across quantiles for all governance variables.
- Nonparametric conditional quantile estimation shows that the relationship between growth and governance is not necessarily positive and/or monotonic. The relationship exhibits heterogeneity across regions and time.

Finally, this article also demonstrates that fully nonparametric methods are not only useful, but they are also computationally feasible in a parallel computing environment. As suggested by Racine (2002), all numerical algorithms in this article use parallel computing[3] in the statistical environment Jacho-Chávez and Trivedi (2009) provide an overview of this important computational tool for empirical researchers. All the code and data for this article are available upon request from the authors.

The rest of this article is organized as follows. Section 2 briefly discusses the data used in the study. The empirical findings are described and discussed in Section 3 while Section 4 offers concluding remarks.

2. GOVERNANCE AND GROWTH DATA

The World Governance Indicators are provided by the World Bank and is updated annually with the most recent iteration by Kaufmann, Kraay, and Mastruzzi (2006). The six governance measures are:[4]

1. Voice and accountability (voice_{it}) measures the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.
2. Political stability and absence of violence (stability_{it}) measures the perceptions of the likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including domestic violence and terrorism.

3. Government effectiveness (effectiveness$_{it}$) measures the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.
4. Regulatory quality (regulatory$_{it}$) measures the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.
5. Rule of Law (law$_{it}$) measures the extent to which agents have confidence in and abide by the rules of society, in particular the quality of contract enforcement, the police, and the courts, as well as the likelihood of crime and violence.
6. Control of corruption (corruption$_{it}$) measures the extent to which public power is exploited for private gain, including petty and grand forms of corruption, as well as "capture" of the state by elites and private interests.

The data is provided for the period 1996–2006. Before 2002 the data was collected on a biannual basis. More details about these variables and their construction can be obtained by perusing the World Bank Governance Indicators URL.[5]

Data on economic growth is drawn from the Total Economy Database.[6] This database is provided by the *Conference Board and Groningen Growth and Development Centre*, and it is an extension of the *World Economy: Historical Statistics* provided by Angus Maddison. It extends the Maddison data from 2003 to 2006. We use this database since the Maddison data is widely used by researchers studying growth. Tables 1–5 list all countries and years under study. The growth rate is calculated as the two-year difference in logarithm of real GDP, and then converted to an annualized growth rate. The data consists of yearly observations of 125 countries classified in five regions. A total of 913 observations are used in this study.

As suggested by Huynh and Jacho-Chávez (2007), conditional density plots are constructed in lieu of descriptive statistics; see Fig. 3. The conditional density plots are computed for growth rates and three different measures (voice$_{it}$, stability$_{it}$, law$_{it}$) during three different years (1996, 2000, 2004). Unlike standard tables, these plots show a more complete picture of the underlying processes generating growth$_{it}$, voice$_{it}$, stability$_{it}$, and law$_{it}$ in all regions. For example, there is large dispersion at low levels of voice$_{it}$ in the relationship between growth and voice$_{it}$. The dispersion is more pronounced for stability$_{it}$ and law$_{it}$. Also, there is some evidence of bimodality in the year 2000 at low levels of governance. This *twin peaks* effect is

Table 1. Western Europe and Offshoots.

Country	Code	Data Coverage	Region
Australia	AUS	1996–2006	1
Austria	AUT	1996–2006	1
Belgium	BEL	1996–2006	1
Canada	CAN	1996–2006	1
Cyprus	CYP	1996–2006	1
Denmark	DNK	1996–2006	1
Finland	FIN	1996–2006	1
France	FRA	1996–2006	1
Germany	DEU	1996–2006	1
Greece	GRC	1996–2006	1
Iceland	ISL	1996–2006	1
Ireland	IRL	1996–2006	1
Italy	ITA	1996–2006	1
Luxembourg	LUX	1996–2006	1
Malta	MLT	1996–2006	1
Netherlands	NLD	1996–2006	1
New Zealand	NZL	1996–2006	1
Norway	NOR	1996–2006	1
Portugal	PRT	1996–2006	1
Spain	ESP	1996–2006	1
Sweden	SWE	1996–2006	1
Switzerland	CHE	1996–2006	1
United Kingdom	GBR	1996–2006	1
United States	USA	1996–2006	1

reminiscent of what previous research using nonparametric methods have found (see, e.g., Quah, 1993; Jones, 1997; Beaudry, Collard, & Green, 2005).

3. EMPIRICAL METHODOLOGY

This section describes the nonparametric empirical methodology utilized in this article. First, we will estimate a parametric specification and then move onto the nonparametric specification. The object of interest in this article is the conditional τ-quantile function (1). The estimation of function (1) is of great importance, because it measures how growth of country i in quantile τ, region REGION_i, at year DT_t is when its governance measures equal specific values of voice_{it}, stability_{it}, $\text{effectiveness}_{it}$, regulatory_{it}, law_{it}, and corruption_{it}. In other words, it provides a way to pin down the effect of governance in country's growth at $\tau = 25\%$, 50%, and 75%, for example. We now proceed to estimate various models for function (1).

Table 2. Eastern Europe and Offshoots.

Country	Code	Data Coverage	Region
Albania	ALB	1996–2005	2
Armenia	ARM	1996–2005	2
Azerbaijan	AZE	1996–2005	2
Belarus	BLR	1996–2005	2
Bosnia-Herzegovina	BIH	1996–2005	2
Bulgaria	BGR	1996–2006	2
Croatia	HRV	1996–2006	2
Czech Republic	CZE	1996–2006	2
Estonia	EST	1996–2006	2
Georgia	GEO	1996–2005	2
Hungary	HUN	1996–2006	2
Kazakhstan	KAZ	1996–2005	2
Kyrgyz Republic	KGZ	1996–2005	2
Latvia	LVA	1996–2006	2
Lithuania	LTU	1996–2006	2
Macedonia	MKD	1996–2005	2
Moldova	MDA	1996–2005	2
Poland	POL	1996–2006	2
Romania	ROM	1996–2006	2
Russia	RUS	1996–2005	2
Serbia and Montenegro	YUG	1996–2005	2
Slovakia	SVK	1996–2006	2
Slovenia	SVN	1996–2006	2
Tajikistan	TJK	1996–2005	2
Turkmenistan	TKM	1996–2005	2
Ukraine	UKR	1996–2005	2
Uzbekistan	UZB	1996–2005	2

3.1. Parametric Models

To provide a benchmark for the nonparametric approach, the following parametric model of 1 is estimated:

$$\begin{aligned} Q_{\text{growth}_{it}}[\tau|\text{REGION}_i,\ \text{DT}_t,\ \text{voice}_{it}, \text{stability}_{it},\ \text{effectiveness}_{it}, \text{regulatory}_{it},\\ \text{law}_{it},\ \text{corruption}_{it}] \\ = \beta_0 \text{REGION}_i + \sum_{t=1}^{8} \beta_t \text{DT}_t + \beta_9 \text{voice}_{it} + \beta_{10} \text{stability}_{it} \\ + \beta_{11} \text{effectiveness}_{it} + \beta_{12} \text{regulatory}_{it} + \beta_{13} \text{law}_{it} + \beta_{14} \text{corruption}_{it} \end{aligned} \tag{2}$$

Table 6 provides the estimates of βs in Eq. (2) at different quantile levels. The model is estimated using the ''check function'' approach for quantile

Table 3. Latin America & Caribbean.

Country	Code	Data Coverage	Region
Argentina	ARG	1996–2005	3
Barbados	BRB	1996–2005	3
Bolivia	BOL	1996–2005	3
Brazil	BRA	1996–2005	3
Chile	CHL	1996–2005	3
Colombia	COL	1996–2005	3
Costa Rica	CRI	1996–2005	3
Cuba	CUB	1996–2005	3
Dominican Republic	DOM	1996–2005	3
Ecuador	ECU	1996–2005	3
Guatemala	GTM	1996–2005	3
Jamaica	JAM	1996–2005	3
Mexico	MEX	1996–2006	3
Peru	PER	1996–2005	3
Puerto Rico	PRI	1998–2005	3
St. Lucia	LCA	1998–2005	3
Trinidad and Tobago	TTO	1996–2005	3
Uruguay	URY	1996–2005	3
Venezuela	VEN	1996–2005	3

regression as in Koenker and Bassett (1978). The results find that the variables: voice_{it}, law_{it}, and corruption_{it} are negatively related to growth. The significance varies across quantiles; voice_{it} is insignificant at $\tau = 0.25$, while corruption_{it} is not significant at $\tau = 0.75$. The other governance measures, stability_{it}, $\text{effectiveness}_{it}$, and regulatory_{it}, are positively related to growth. At $\tau = 0.25$, stability_{it} is insignificant while regulatory_{it} is only significant at $\tau = 0.50$. The time dummies are significant for various years for $\tau = 0.25$, 0.50, while for $\tau = 0.75$ only 2002 is significant. Across all quantiles the Eastern Europe and Offshoots and Asia dummy is positive and significant, while Africa is negative and significant for $\tau = 0.75$.

There are some differences between some of the governance measures across quantiles. To verify whether these differences are significant, we proceed to test[7] whether their associated coefficients are the same across different quantiles, that is,

$$H_0 : \beta_{l;0.25} = \beta_{l;0.50} = \beta_{l;0.75}$$

where $l = 9, \ldots, 14$ in (2). The results are summarized in Table 7.

The large p-values indicate that we fail to reject the null hypothesis that the coefficients are not different across different quantiles for each

Table 4. Asia.

Country	Code	Data Coverage	Region
Bahrain	BHR	1996–2005	4
Bangladesh	BGD	1996–2005	4
Cambodia	KHM	1996–2005	4
China	CHN	1996–2006	4
Hong Kong	HKG	1996–2005	4
India	IND	1996–2006	4
Indonesia	IDN	1996–2005	4
Iran	IRN	1996–2005	4
Iraq	IRQ	1996–2005	4
Israel	ISR	1996–2005	4
Japan	JPN	1996–2006	4
Jordan	JOR	1996–2005	4
Korea, South	KOR	1996–2006	4
Kuwait	KWT	1996–2005	4
Malaysia	MYS	1996–2005	4
Myanmar	MMR	1996–2005	4
Oman	OMN	1996–2005	4
Pakistan	PAK	1996–2005	4
Philippines	PHL	1996–2005	4
Qatar	QAT	1996–2005	4
Saudi Arabia	SAU	1996–2005	4
Singapore	SGP	1996–2005	4
Sri Lanka	LKA	1996–2005	4
Syria	SYR	1996–2005	4
Taiwan	TWN	1996–2005	4
Thailand	THA	1996–2005	4
Turkey	TUR	1996–2006	4
United Arab Emirates	ARE	1996–2005	4
Vietnam	VNM	1996–2005	4
Yemen	YEM	1996–2005	4

governance measure. Given these parametric results we turn our attention now to the nonparametric quantiles.

3.2. Nonparametric Models

We proceed to estimate object (1) as

$$\begin{aligned} Q_{\text{growth}_{it}}[\tau|\text{REGION}_i,\ \text{DT}_t,\ \text{voice}_{it},\ \text{stability}_{it},\ \text{effectiveness}_{it},\\ \text{regulatory}_{it},\ \text{law}_{it},\ \text{corruption}_{it}]\\ = q_\tau(\text{REGION}_i,\ \text{DT}_t,\ \text{voice}_{it}, \text{stability}_{it},\ \text{effectiveness}_{it},\\ \text{regulatory}_{it},\ \text{law}_{it},\ \text{corruption}_{it}), \end{aligned} \tag{3}$$

Table 5. Africa.

Country	Code	Data Coverage	Region
Algeria	DZA	1996–2005	5
Angola	AGO	1996–2005	5
Burkina Faso	BFA	1996–2005	5
Cameroon	CMR	1996–2005	5
Egypt	EGY	1996–2005	5
Ethiopia	ETN	1996–2005	5
Ghana	GHA	1996–2005	5
ivory Coast	CIV	1996–2005	5
Kenya	KEN	1996–2005	5
Madagascar	MDG	1996–2005	5
Malawi	MWI	1996–2005	5
Mali	MLI	1996–2005	5
Morocco	MAR	1996–2005	5
Mozambique	MOZ	1996–2005	5
Niger	NER	1996–2005	5
Nigeria	NGA	1996–2005	5
Senegal	SEN	1996–2005	5
South Africa	ZAF	1996–2005	5
Sudan	SDN	1996–2005	5
Tanzania	TZA	1996–2005	5
Tunisia	TUN	1996–2005	5
Uganda	UGA	1996–2005	5
Zaire	ZAR	1996–2005	5
Zambia	ZMB	1996–2005	5
Zimbabwe	ZWE	1996–2005	5

where $q_\tau(\cdot)$ is assumed to be a smooth continuous but otherwise unknown function. Nonparametric methods are more flexible since they require minimal assumptions on the function q_τ; see the appendix. Eq. (2) is a special case of Eq. (3); it will therefore capture both linear and nonlinear relationships automatically without the need of a model search.

We use the estimator proposed in[8] Li and Racine (2008) with bandwidths chosen as suggested[9] therein; see Li and Racine (2007, Section 6.5, pp. 193–196). The panel structure of the data is implicitly taken into account by this nonparametric estimator, because it works by averaging data points locally close to the point of interest. That is, it automatically gives larger weights to countries in the same region and/or year of measurement,[10] allowing for heterogenous time-varying effects across regions in the nonparametric sense akin to the inclusion of dummy

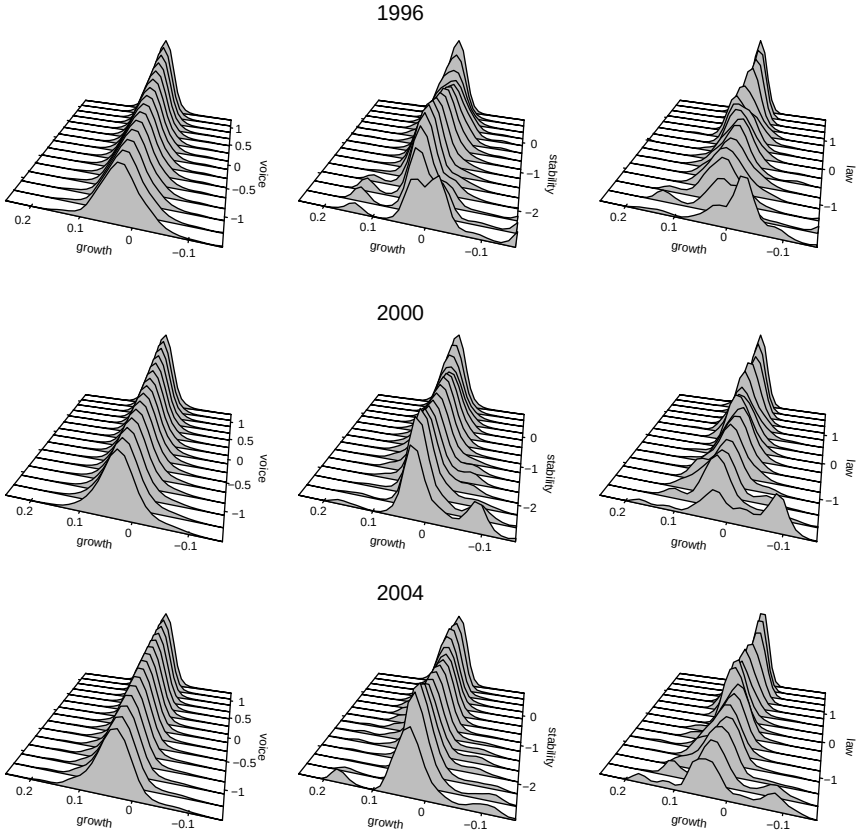


Fig. 3. Conditional Density Plots. *Note:* Bandwidths were chosen by maximum likelihood cross-validation; see Li and Racine (2007, Section 5.2.2, pp. 160–162). The resulting values are 0.0147 for growth$_{it}$, 0.2146 for REGION$_i$, 0.7787 for DT$_t$, 0.7517 for voice$_{it}$, 0.3402 for stability$_{it}$, and 0.2375 for law$_{it}$.

Table 6. Parametric Tests.

Variable	*p*-Value
voice$_{it}$	0.4364
stability$_{it}$	0.4691
effectiveness$_{it}$	0.1182
regulatory$_{it}$	0.5417
law$_{it}$	0.9787
corruption$_{it}$	0.6073

Table 7. Parametric Quantile Regression.

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$	
	Coef.	Std. Error	Coef.	Std. Error	Coef.	Std. Error
$REGION_i$						
Western Europe & Offshoots	–	–	–	–	–	–
Eastern Europe & Offshoots	0.0223	(0.0051)***	0.0240	(0.0039)***	0.0278	(0.0049)***
Latin America & Offshoots	−0.0028	(0.0051)	−0.0049	(0.0039)	−0.0047	(0.0052)
Asia	0.0108	(0.0048)**	0.0099	(0.0037)***	.0136	(0.0046)***
Africa	0.0042	(0.0053)	−0.0029	(0.0041)	−0.0101	(0.0054)*
DT_t						
1996	–	–	–	–	–	–
1998	0.0010	(0.0042)	−0.0012	(0.0032)	0.0002	(0.0041)
2000	0.0050	(0.0043)	0.0016	(0.0032)	−0.0003	(0.0041)
2002	−0.0044	(0.0043)	−0.0095	(0.0032)***	−0.0083	(0.0041)**
2003	−0.00004	(0.0042)	−0.0077	(0.0033)**	−0.0031	(0.0041)
2004	0.0102	(0.0043)**	0.0059	(0.0032)*	0.0059	(0.0041)
2005	0.0169	(0.0042)***	0.0120	(0.0033)***	0.0067	(0.0041)
2006	0.0138	(0.0061)**	0.0045	(0.0047)	0.0051	(0.0060)
$voice_{it}$	−0.0033	(0.0026)	−0.0068	(0.0021)***	−0.0056	(0.0026)**
$stability_{it}$	0.0034	(0.0023)	0.0068	(0.0018)***	0.0077	(0.0023)***
$effectiveness_{it}$	0.0285	(0.0048)***	0.0206	(0.0037)***	0.0148	(0.0047)***
$regulatory_{it}$	0.0031	(0.0035)	0.0051	(0.0024)**	0.0009	(0.0031)
law_{it}	−0.0173	(0.0057)***	−0.0182	(0.0044)***	−0.0173	(0.0055)***
$corruption_{it}$	−0.0095	(0.0044)**	−0.0065	(0.0035)*	−0.0038	(0.0044)
Constant	−0.0030	(0.0049)	0.0208	(0.0037)***	0.0384	(0.0047)***

Heteroskedasticity-robust standard errors are in parenthesis. (***) significant at 1%, (**) significant at 5%, and (*) significant at 10%.

variables would in the standard parametric set-up (see, e.g., Racine, 2008, Section 6.1, p. 59). More details concerning the estimating strategy can be found in the appendix.

Unfortunately, we have been unable to find a suitable nonparametric counterpart of these parametric tests performed above. We leave nonparametric quantile testing for future work. However, we draw upon our results in Huynh and Jacho-Chávez (2009) for the nonparametric conditional mean and focus on the same measures: $voice_{it}$, $stability_{it}$, and law_{it} from now onwards.

3.3. Growth Profile Curves

We illustrate the results using partial regression plots because non parametric methods do not yield scalar estimates of marginal effects. As in Huynh and Jacho-Chávez (2009), we call these partial regression plots – *growth profile curves* (GPC). As an illustrative example, a simple case is presented to give the reader some intuition. The top plot of Fig. 4 displays the expected growth of a country in Eastern Europe and Offshoots in 2002, and in the 50% quantile of the growth distribution, as a function of voice_{it} and stability_{it}. Once we condition on a specific value of voice_{it}, let's say, each black line on the surface represents a growth profile as a function of the remaining variables, in this case stability_{it}. The conditioning values are $\alpha = 25\%$, 50%, and 75% sample quantiles of each governance measure. These curves are put together into two-dimensional plots at the bottom of Fig. 4. These curves are informative about the growth path of a country in the 50% quantile with respect to a particular governance measure, once we condition the remaining variables to a prespecified value. We call these paths GPC. Intuitively, these curves are just slices of the fitted nonparametric hyperplane conditional on some variables.

These GPC can be generalized to multidimensional settings, that is, more than two conditioning variables, as it is implied by the empirical object of interest (Eq. 3). Figs. 5–7 show the results. Each plot in each figure displays a visualization of the estimated q_τ, i.e. $\hat{q}_\tau$, in 3 at $\tau = 25\%$ (first column), 50% (second column), and 75% (third column), and different conditioning variables. For example, the top row of plots in Fig. 5 shows

$$\begin{aligned}\hat{q}_\tau(&\text{REGION}_i = \text{Western Europe and Offshoots},\ \text{DT}_t,\ \text{voice}_{it},\\ &\text{stability}_{it} = Q_{\text{stability}_{it}}(0.5),\ \text{effectiveness}_{it} = Q_{\text{effectiveness}_{it}}(0.5),\\ &\text{regulatory}_{it} = Q_{\text{regulatory}_{it}}(0.5), \text{law}_{it} = Q_{\text{law}_{it}}(0.5),\ \text{corruption}_{it}\\ &= Q_{\text{corruption}_{it}}(0.5))\end{aligned}$$

as a function of voice_{it} for each value of DT_t, where $Q_{\mathbf{x}_{it}}(\alpha)$ represents the α-sample quantile of variable $\mathbf{x}_{it}$ across both i and t. Figs. 6 and 7 were constructed accordingly by resetting the varying variable to be stability_{it} and law_{it}, respectively.

Figs. 8–10 present a visualization of $\tau = 50\%$, but only when the remaining indicators are held at 0 for years 1996, 2000, and 2004,

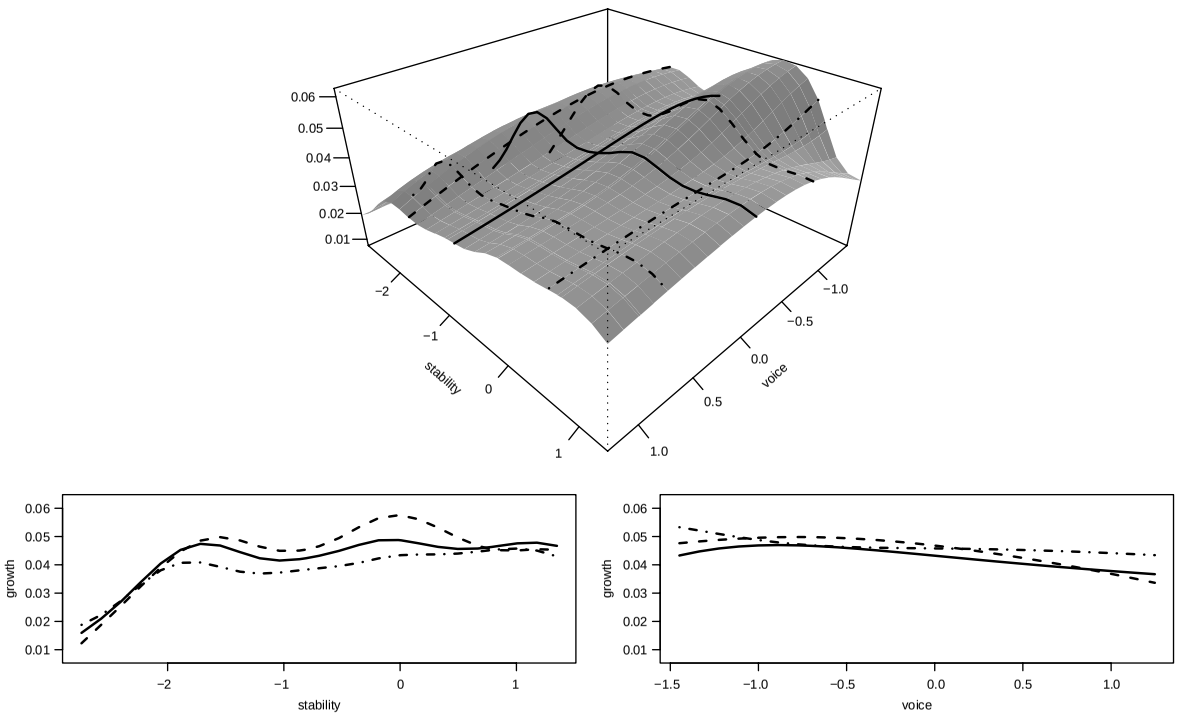


Fig. 4. 50% Quantile – Growth Profile Curves. *Note:* First plot is a three-dimensional surface representing $\hat{q}_{0.5}$ (Eastern Europe and Offshoots, 2002, voice_{it}, stability_{it}). Bottom plots show the corresponding *growth profile curves* highlighted in the three-dimensional surface. The conditioning variables are the 25% (dashed), 50% (solid), and 75% (dotted-dashed) estimated quantiles of the entire sample.

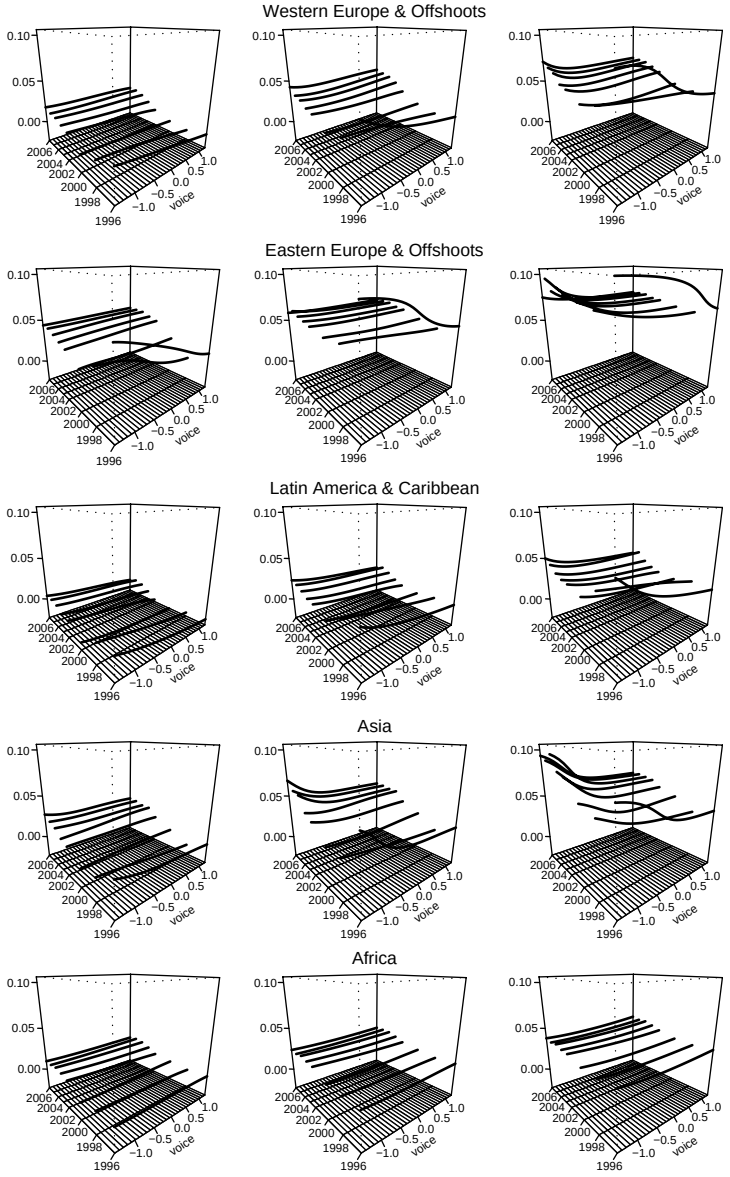


Fig. 5. Growth Profile Curves – Voice and Accountability. *Note:* Graphs represent *growth profile curves* at: $\tau = 0.25$ (first column), $\tau = 0.5$ (second column), and $\tau = 0.75$ (third column), when all continuous covariates but voice_{it} are kept constant at their respective sample median.

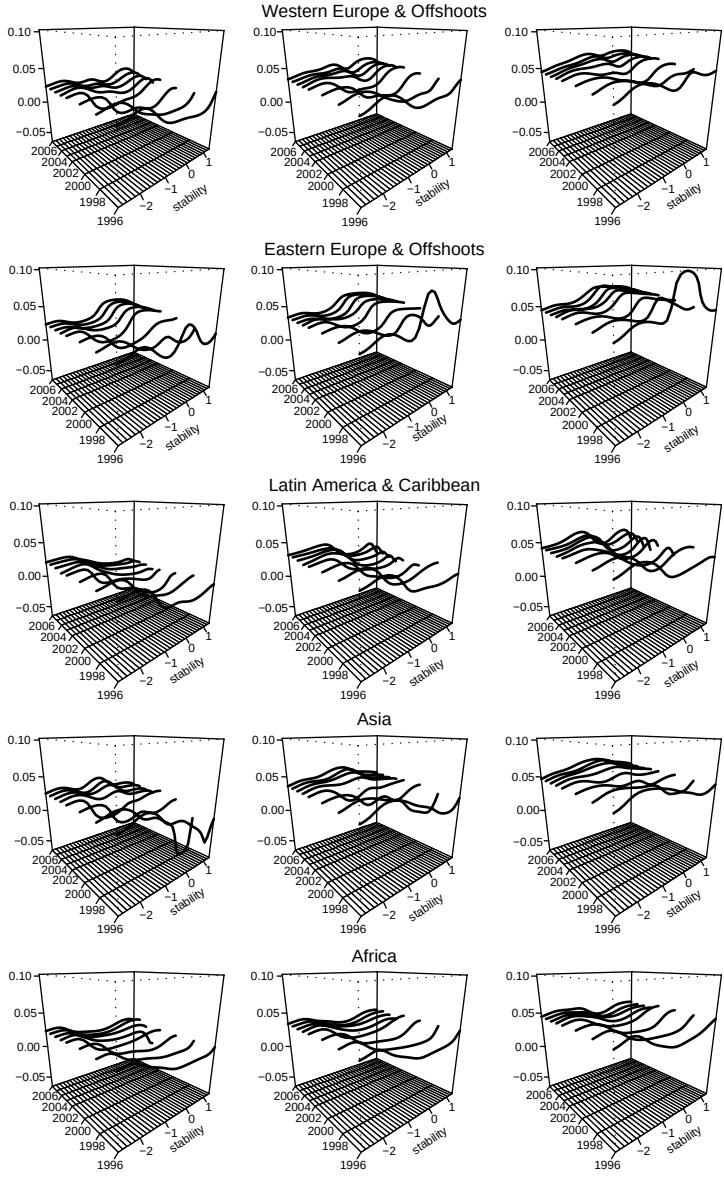


Fig. 6. Growth Profile Curves – Political Stability. *Note:* Graphs represent *growth profile curves* at: $\tau = 0.25$ (first column), $\tau = 0.5$ (second column), and $\tau = 0.75$ (third column), when all continuous covariates but stability_{it} are kept constant at their respective sample median.

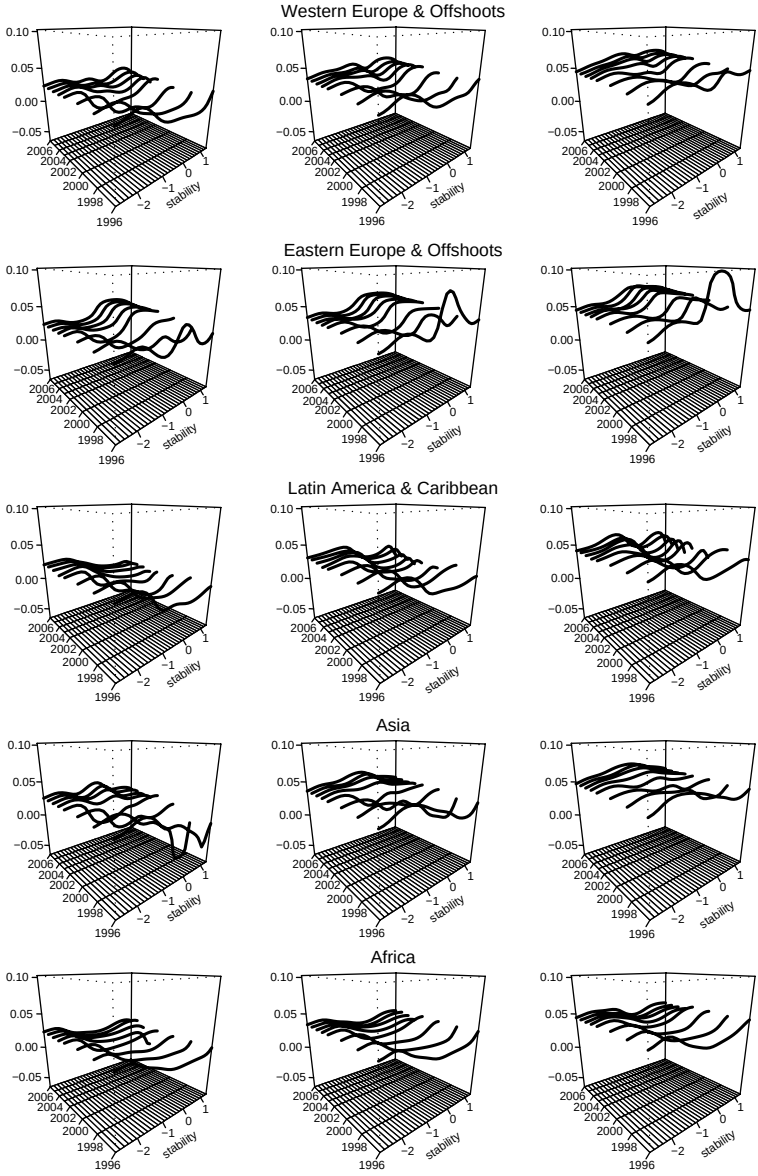


Fig. 7. Growth Profile Curves – Rule of Law. *Note:* Graphs represent *growth profile curves* at: $\tau = 0.25$ (first column), $\tau = 0.5$ (second column), and $\tau = 0.75$ (third column), when all continuous covariates but law_{it} are kept constant at their respective sample median.

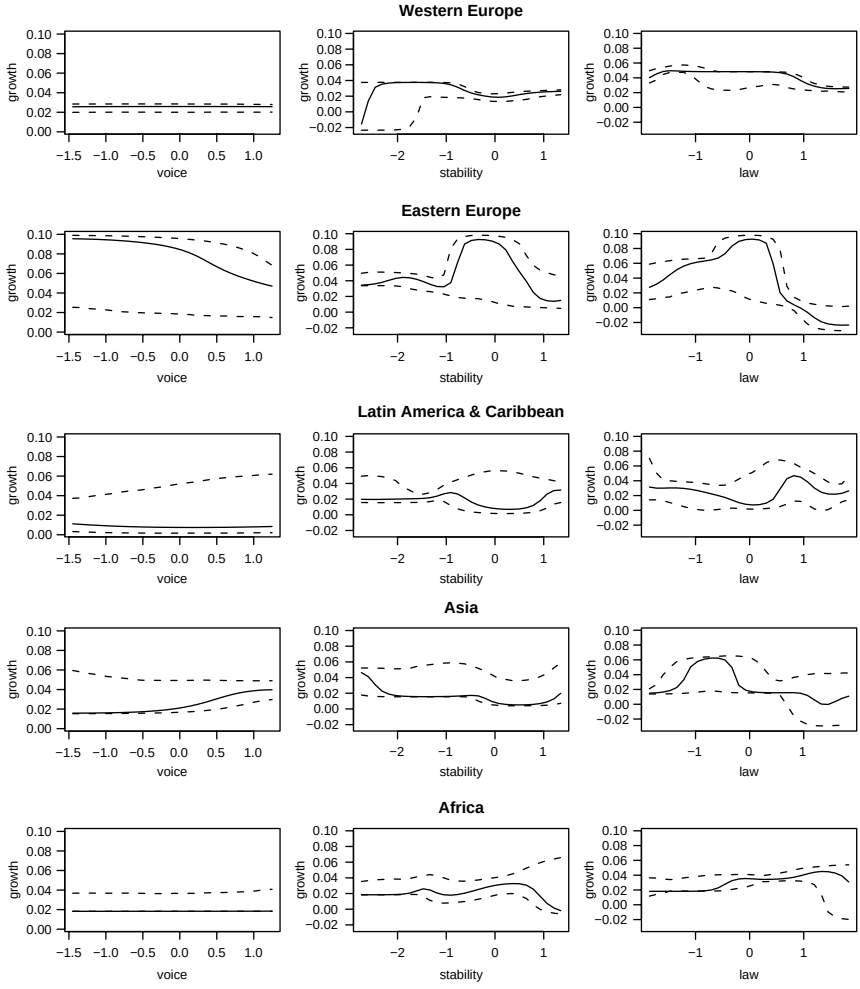


Fig. 8. 50% Quantile – Growth Profile Curves, 1996. *Note:* Dotted lines represent 90% bootstrap confidence intervals based on 499 bootstrap replications. They are not symmetric because they estimate stochastic variation of hyperplanes, and not of univariate functions.

respectively. They also present 90% bootstrap confidence interval based on 499 wild bootstrap replications. These conservative bootstrap confidence intervals are not symmetric in Figs. 8–10 because they estimate stochastic variation of hyperplanes, and not of univariate functions.

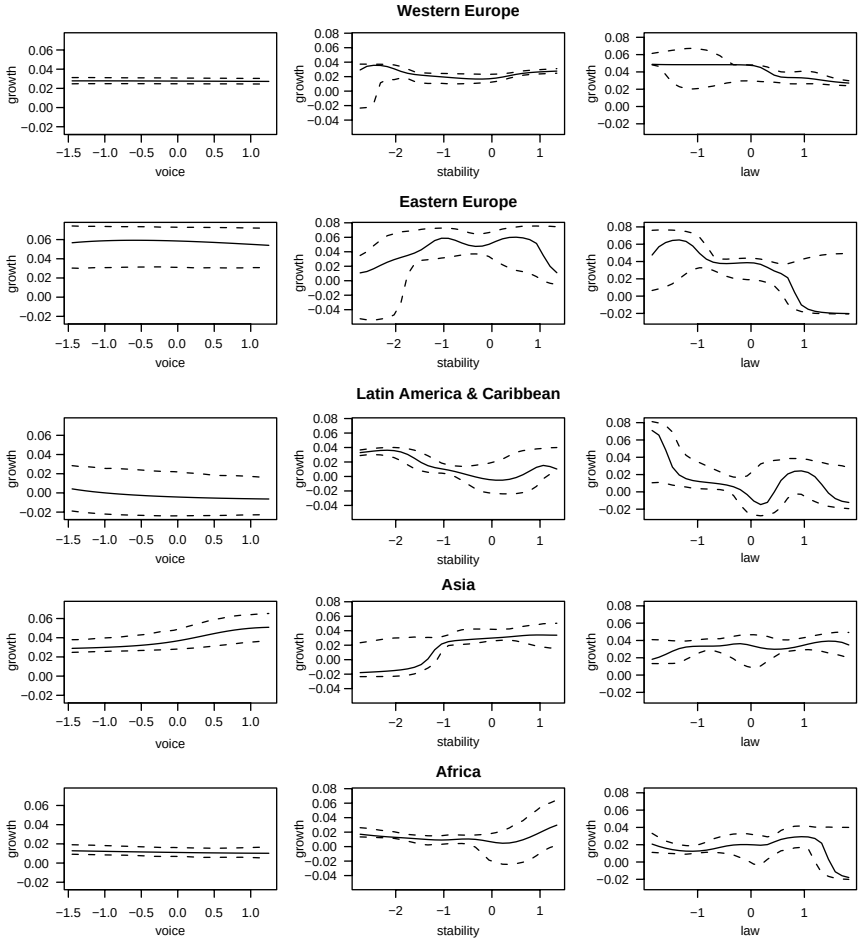


Fig. 9. 50% Quantile – Growth Profile Curves, 2000. *Note:* Dotted lines represent 90% bootstrap confidence intervals based on 499 bootstrap replications. They are not symmetric because they estimate stochastic variation of hyperplanes, and not of univariate functions.

3.4. Discussion

To illustrate the results of the nonparametric regression, GPC are constructed for the five regions of the world: Western Europe and Offshoots, Eastern Europe and Offshoots, Latin America and Caribbean,

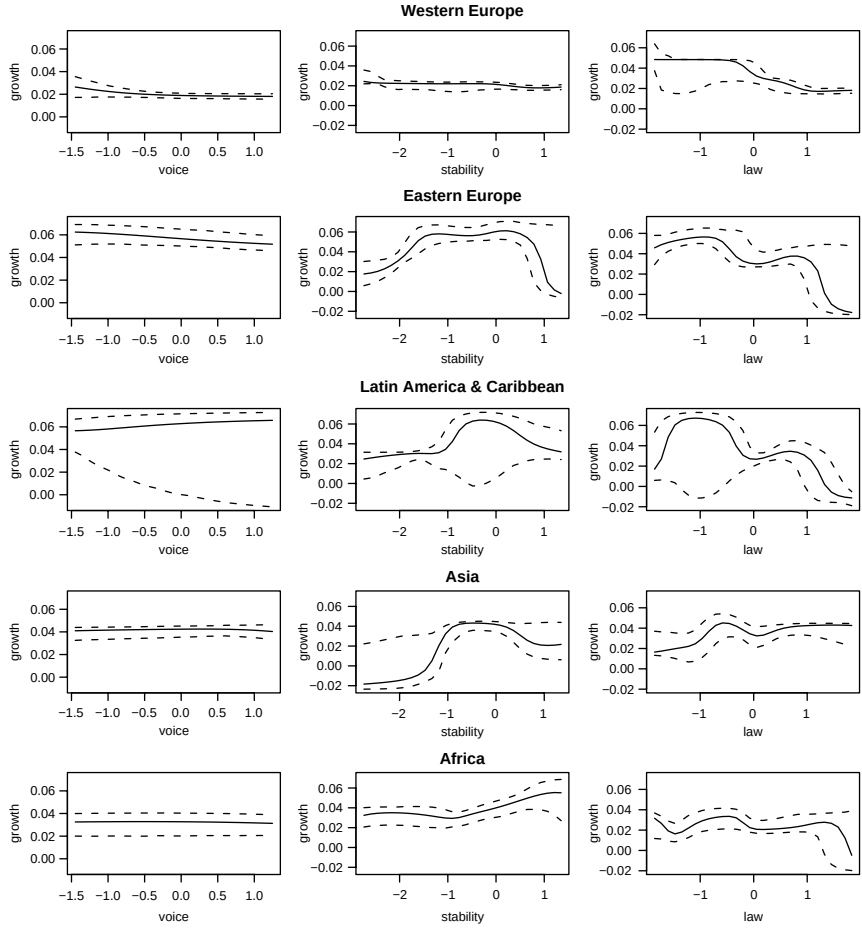


Fig. 10. 50% Quantile – Growth Profile Curves, 2004. *Note:* Dotted lines represent 90% bootstrap confidence intervals based on 499 bootstrap replications. They are not symmetric because they estimate stochastic variation of hyperplanes, and not of univariate functions.

Asia, and Africa. Each plot is conditioned on the year and governance measure for each of the three significant variables as found in Huynh and Jacho-Chávez (2009) (see, e.g., Alexeev, Huynh, & Jacho-Chávez, 2009). For brevity, we present the results for the quantiles ($\tau = 0.25, 0.50, 0.75$) conditioned on $\alpha = 0.50$. We have also computed the quantile graphs

conditioning on $\alpha = 0.25$, 0.75 quantile;, these extra results, data, R code, and full set of confidence intervals are available on request.

3.4.1. Voice and Accountability

Fig. 5 illustrates the results for voice_{it}. There are differences in GPC across τ-quantiles in terms of regions. For Western Europe and Offshoots the GPC is relatively flat for $\tau = 0.25$, 0.50 quantile, but in $\tau = 0.75$ there is some variation at the lower quantities of voice_{it}. This pattern is mirrored with Eastern Europe and Offshoots, Latin America and Caribbean, and Asia. For Asia the effect is most dramatic. However, for Africa the effect is uniformly flat across quantiles. From the parametric testing the quantile coefficients were deemed similar, but the GPC reveal interestingly that voice_{it} is variable across regions. The nonparametric quantile methods are able to capture the complex interactions between voice_{it}, region, and year effects without parameterizing interaction terms. Therefore, the attractiveness of nonparametric quantile methods comes through.

3.4.2. Political Stability

Fig. 6 illustrates the results for stability_{it}. The nonparametric conditional quantiles GPC are similar across quantiles for reach region. This result accords with the parametric quantile testing. However, across regions the GPC are different. The Western Europe and Offshoots, not surprisingly, have a relative smooth albeit nonmonotonic shape. Eastern Europe and Offshoots have more volatility in GPC especially for the earlier years to illustrate the immense structural changes in these countries. The GPC for Latin American and Caribbean and Asia are smooth for $\tau = 0.75$, but for the lower quantile there is much volatility in 1996 and 1998, which were the times of the various financial/banking crises in these regions. Africa's GPC are also smooth and display a positive relationship at low levels of governance. At higher governance measures, the relationship is negative.

3.4.3. Rule of Law

Fig. 7 illustrates the results for law_{it}. The patterns are stark, the variation in the GPC are amplified as we move from $\tau = 0.25$ to $\tau = 0.50$ quantile. In fact, the relationship between law_{it} and growth is negative (similar to the parametric model). However, the GPC show that there is considerable variation in the quantile function. There is heterogeneity in year and regions. In particular, Eastern Europe and Offshoots and Africa display large amounts of variation. Compared to the nonparametric conditional

mean results in Huynh and Jacho-Chávez (2009) the conditional quantiles for law_{it} show a clearer pattern.

3.5. Case Study: Latin America and Caribbean and Africa

We focus on Latin America and Caribbean and Africa in the year 2004 to illustrate the efficacy of nonparametric conditional quantile estimation. Both regions display interesting GPC for the variables stability_{it} and law_{it} at the 50% quantile that are worth discussing. Figs. 11 and 12 plot both the observed data and their respective GPC with 90% bootstrap confidence intervals.

For stability_{it}, Latin America and Caribbean's GPC are nonmonotonic but with confidence intervals, whereas in Africa the GPC is nonlinear with smaller uncertainty. With law_{it} the GPC curves for both regions are nonmonotonic with no discernable pattern. Again, Latin America and Caribbean's GPC are more variable than Africa's. This result may be indicative of the varying levels of development in Latin America and Caribbean, while in Africa as a continent it is similar as a whole.

These empirical results can be use to understand the tradeoffs between growth and governance in the context of growth diagnostics advocated by Rodrik (2006). Increasing governance may not necessarily lead to increase in growth because the binding constraint is not governance. In Hausmann et al. (2008) the growth diagnostics yield different policy recommendations for Brazil and the Dominican Republic. They argue that in Brazil a reform of the governance would not increase growth or that it is not a binding constraint. Instead they argue that the slow growth can be explained by Brazil's lack of access to external capital markets and low domestic savings. The Dominican Republic has been labeled an *unlikely success story* because of the low-level governance but high growth rates until a banking crisis occurred in 2002. The suggested cure for Dominican Republic need not require wholesale reforms but targeted reforms.

4. CONCLUDING REMARKS

This paper considers the growth and governance relationship through the lens of nonparametric quantile analysis. The analysis focuses on three

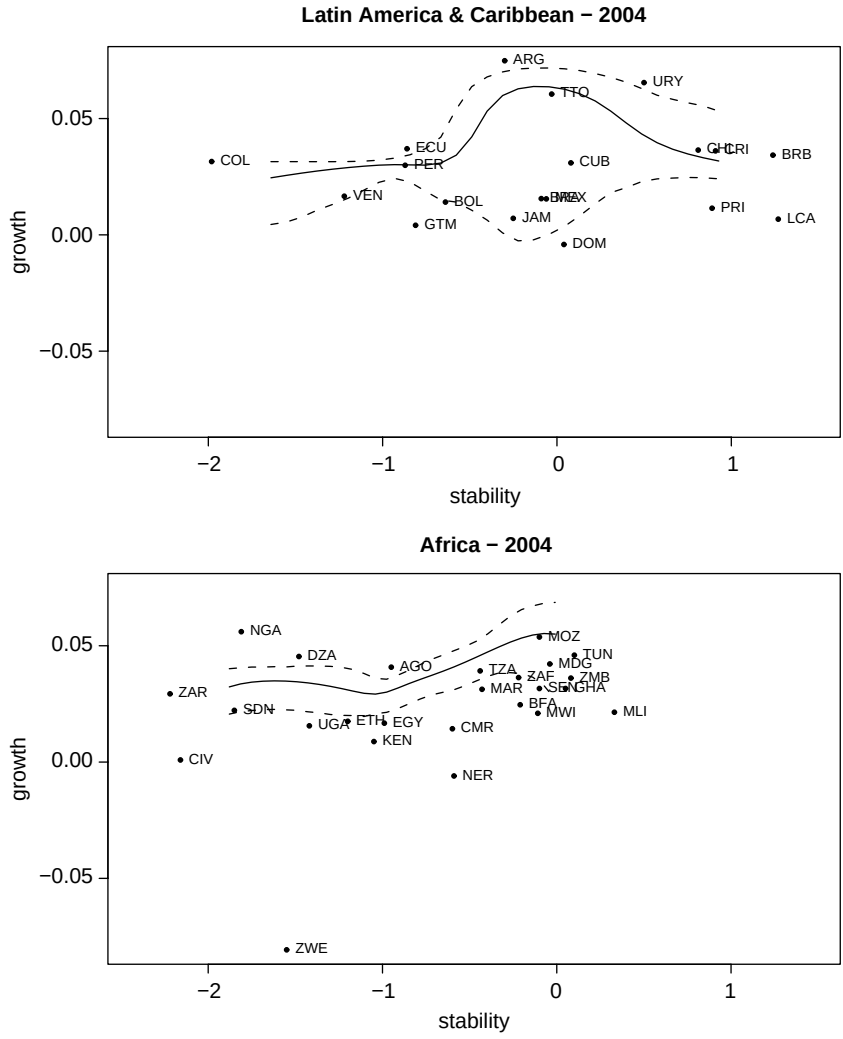


Fig. 11. 50% Quantile – Case Study – Political Stability. *Note:* Solid line in top graph displays $\hat{q}_{0.5}$ ($\text{REGION}_i = \text{Latin America \& Caribbean}$, $\text{DT}_t = 2004$, $\text{voice}_{it} = 0$, $\text{stability}_{it} = \text{stability}$, $\text{effectiveness}_{it} = 0$, $\text{regulatory}_{it} = 0$, $\text{law}_{it} = 0$, $\text{corruption}_{\text{it}} = 0$). Solid line in bottom graph displays $\hat{q}_{0.5}$ ($\text{REGION}_i = \text{Africa}$, $\text{DT}_t = 2004$, $\text{voice}_{it} = 0$, $\text{stability}_{it} = \text{stability}$, $\text{effectiveness}_{it} = 0$, $\text{regulatory}_{it} = 0$, $\text{law}_{it} = 0$, $\text{corruption}_{it} = 0$). Dotted lines represent 90% bootstrap confidence intervals. They are not symmetric because they estimate stochastic variation of hyperplanes, and not of univariate functions.

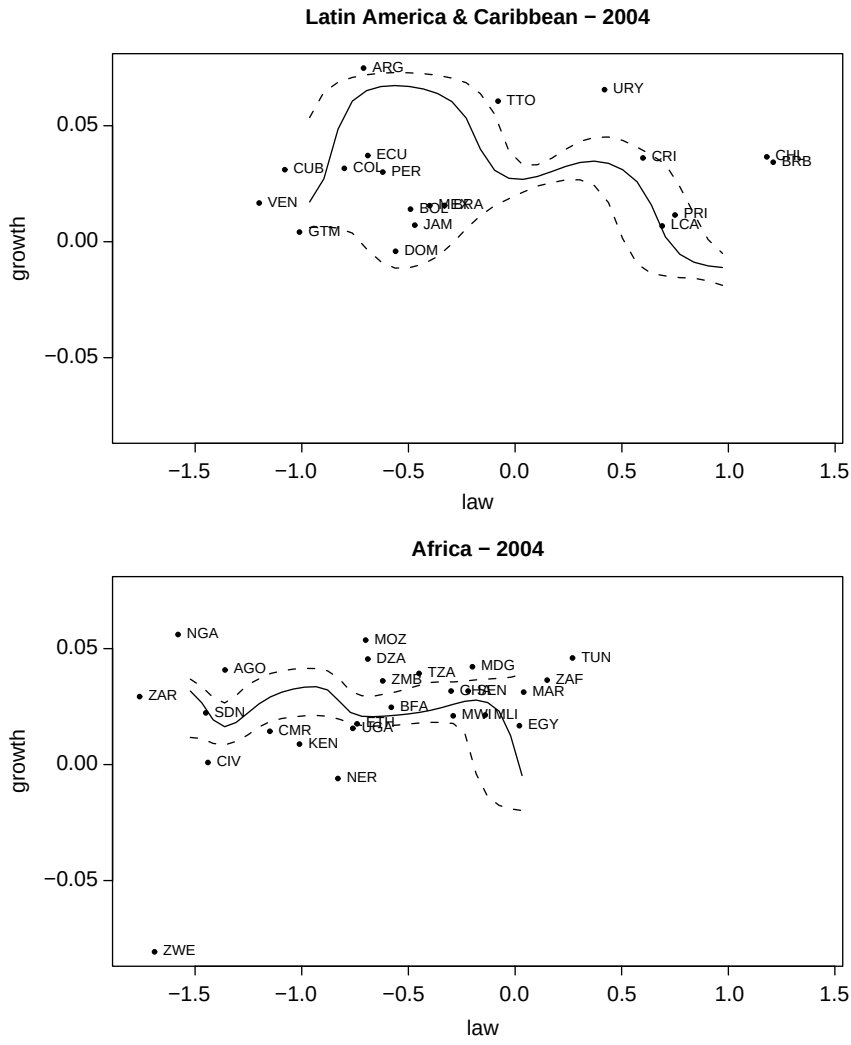


Fig. 12. 50% Quantile – Case Study – Rule of Law. *Note:* Solid line in top graph displays $\hat{q}_{0.5}$ ($\text{REGION}_i = \text{Latin America \& Caribbean}$, $\text{DT}_t = 2004$, $\text{voice}_{it} = 0$, $\text{stability}_{it} = 0$, $\text{effectiveness}_{it} = 0$, $\text{regulatory}_{it} = 0$, $\text{law}_{it} = \text{law}$, $\text{corruption}_{it} = 0$). Solid line in bottom graph displays $\hat{q}_{0.5}$ ($\text{REGION}_i = \text{Africa}$, $\text{DT}_t = 2004$, $\text{voice}_{it} = 0$, $\text{stability}_{it} = 0$, $\text{effectiveness}_{it} = 0$, $\text{regulatory}_{it} = 0$, $\text{law}_{it} = \text{law}$, $\text{corruption}_{it} = 0$). Dotted lines represent 90% bootstrap confidence intervals. They are not symmetric because they estimate stochastic variation of hyperplanes, and not of univariate functions.

governance measures: The relationship between growth and governance at each quantile is nonmonotonic across regions and year. Nonparametric quantiles reveal substantial heterogeneity that is not captured by parametric quantiles estimation. For example, without introducing interaction terms between variables and regions the nonparametric quantiles are able to capture these effects in the GPC. Nonparametric quantiles also demonstrate heterogeneity of results across different quantiles.

These nonmonotonicities and heterogeneity across quantiles highlight the importance of careful modeling of the growth and governance relationships. These empirical results lend credence to the arguments of Rodrik (2006) and Hausmann et al. (2008) that caution policy makers from applying policies uniformly across countries and years. Proper growth diagnostics are required to understand what are the bottlenecks and barriers to growth. Understanding the binding constraints can help policy makers to enact the relevant reforms.

Overall, these findings indicate that caution must be used when using parametric quantile models to analyze the relationship between World Governance Indicators and growth. However, there are some important omissions in this study. Most important is that this paper does not address the issue of causality or control for endogeneity in a regression framework. This could potentially be addressed adapting Horowitz and Lee's (2006) estimator to our framework, while using *European settler mortality rates* (see Acemoglu, Johnson, & Robinson, 2001) as valid instruments for example. Other important features to consider are the dynamics of these measures across time. Finally, little is known about misspecification tests applied to nonparametric quantiles. We leave these important considerations for future study.

NOTES

1. Examples of these conjectures can be found in North (1990), Mauro (1995), and Hall and Jones (1999).
2. The last two World Bank presidents (Paul Wolfowitz and Robert Zoellick) have made public statements regarding this relationship; see http://go.worldbank.org/ATJXPHZMH0 and http://blogs.iht.com/tribtalk/business/globalization/?p = 632
3. We would like to thank Jeffrey S. Racine for providing us with the necessary software to perform these computations at Indiana University's High Performance Clusters.
4. The definitions are taken from http://info.worldbank.org/governance/wgi2007/faq.htm

5. http://info.worldbank.org/governance/wgi2007/
6. http://www.ggdc.net/Dseries/totecon.html
7. See Koenker (2005, Section 3.3.2, pp. 76–77) for details. Although this test statistics assumes a random independent sample, no further modifications for time series were performed in this set-up.
8. We use a second-order Gaussian kernel for each continuous variable, that is, growth_{it}, voice_{it}, stability_{it}, government_{it}, regulatory_{it}, law_{it}, and corruption_{it}. The Aitchison and Aitken's (1976) kernel for unordered categorical variable was used for the regional indicator (REGION_i), and Wang and van Ryzin's (1981) kernel was used for the ordered categorical variable DT_t.
9. The resulting bandwidths are 0.2146, 0.7787, 0.7517, 0.3402, 0.1685, 0.4267, 0.2375, and 0.4686 for REGIONi, DT_t, voice_{it}, stability_{it}, government_{it}, regulatory_{it}, law_{it}, and corruption_{it}, respectively; and 0.1468 for growth_{it}.
10. Alternatively, we could also condition on a country-specific unordered categorical variable as well. We thank an anonymous referee for pointing this out.

ACKNOWLEDGMENTS

We thank the editors and two anonymous referees for their valuable suggestions that greatly improved the exposition and readability of the paper. We acknowledge the usage of the np package by Hayfield and Racine (2008). We also acknowledge Takuya Noguchi, UITS Center for Statistical and Mathematical Computing (Indiana University), for installing the necessary software in the Quarry High Performance Cluster at Indiana University where all the computations were performed. Abhijit Ramalingam provided excellent research assistance. Finally, we thank Gerhard Glomm for his constant and unconditional encouragement for this project.

REFERENCES

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, *91*(5), 1369–1401.

Aitchison, J., & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, *63*(3), 413–420.

Alexeev, M., Huynh, K. P., & Jacho-Chávez, D. T. (2009). *Robust nonparametric inference for growth and governance relationships*. Unpublished manuscript.

Beaudry, P., Collard, F., & Green, D. A. (2005). Changes in the world distribution of output per worker, 1960–1998: How a standard decomposition tells an unorthodox story. *The Review of Economics and Statistics*, *87*(4), 741–753.

Hall, R. E., & Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, *114*(1), 83–116.

Hausmann, R., Rodrik, D., & Velasco, A. (2008). Growth diagnostics. In: N. Serra & J. E. Stiglitz (Eds), *The Washington consensus reconsidered towards a new global governance* (pp. 324–355). New York, NY: Oxford University Press.

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*(5), 1–32.

Horowitz, J., & Lee, S. (2006). *Nonparametric instrumental variables estimation of a quantile regression model*. CeMMAP Working Paper CWP09/06, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Huynh, K. P., & Jacho-Chávez, D. T. (2007). Conditional density estimation: An application to the Ecuadorian manufacturing sector. *Economics Bulletin*, *3*(62), 1–6.

Huynh, K. P., & Jacho-Chávez, D. T. (2009). Growth and governance: A nonparametric analysis. *Journal of Comparative Economics*, *37*(1), 121–143.

Jacho-Chávez, D. T., & Trivedi, P. K. (2009). Computational considerations in empirical microeconometrics: Selected examples. In: T. C. Mills & K. Patterson (Eds), *Palgrave handbook of econometrics, Volume 2: Applied econometrics* (Chapter 15, pp. 775–817). Great Britain: Palgrave Macmillan.

Jones, C. I. (1997). On the evolution of the world income distribution. *Journal of Economic Perspectives*, *11*(3), 19–36.

Kaufmann, D., & Kraay, A. (2002). Growth without governance. World Bank Policy Research Working Paper No. 2928.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2006). Governance matters VI: Governance indicators for 1996–2006. World Bank Policy Research Working Paper No. 4280.

Koenker, R. (2005). *Quantile Regression, Econometric Society Monograph Series*. New York, NY: Cambridge University Press.

Koenker, R., & Bassett, G. J. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Li, Q., & Racine, J. S. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, *86*(2), 266–292.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Li, Q., & Racine, J. S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, *26*, 423–434.

Mauro, P. (1995). Corruption and growth. *The Quarterly Journal of Economics*, *110*(3), 681–712.

North, D. (1990). *Institutions, institutional change and economic performance*. Cambridge, UK: Cambridge University Press.

Quah, D. (1993). Empirical cross-section dynamics in economic growth. *European Economic Review*, *37*(2–3), 426–434.

Racine, J. S. (2002). Parallel distributed kernel estimation. *Computational Statistics and Data Analysis*, *40*(2), 293–302.

Racine, J. S. (2008). Nonparametric econometrics: A primer. *Foundations and Trends in Econometrics*, *3*(1), 1–88.

Rodrik, D. (2006). Goodbye Washington consensus, hello Washington confusion? A review of the World Bank's *Economic Growth in the 1990s: Learning from a Decade of Reform*. *Journal of Economic Literature*, *44*(4), 973–987.

Wang, M. C., & van Ryzin, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika*, *68*, 301–309.

A. TECHNICAL APPENDIX

Kernel Smoothing

Suppose we observed a sample $\{y_i, \mathbf{x}_i^\top\}$, $i = 1, \ldots, n$ from a random vector $[y, \mathbf{x}^\top]$ where $y \in \mathbb{R}$, and $\mathbf{x}$ is a mixture of continuous variables $\mathbf{x}^c = [x^1, \ldots, x^{q_1}] \in \mathbb{R}^{q_1}$ and discrete $\mathbf{x}^d = [x^{q_1+1}, \ldots, x^q]^\top \in S^d$ where S^d is the support of $\mathbf{x}^d$, and $q_2 = q - q_1$. For particular two points y_i, $\mathbf{x}_i = [\mathbf{x}_i^c, \mathbf{x}_i^d]$, and y_j, $\mathbf{x}_j = [\mathbf{x}_j^c, \mathbf{x}_j^d]$, let us define the functions

$$K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) = \prod_{l=1}^{q_1} \frac{1}{h_l} k\left(\frac{x_i^l - x_j^l}{h_l}\right) \tag{A.1}$$

$$L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda) = \prod_{l=1}^{q_2} l(x_i^l, x_j^l; \lambda_l) \tag{A.2}$$

$$G(y_i, y_j; h_y) = \int_{-\infty}^{(y_i - y_j/h_y)} k(t)dt \tag{A.3}$$

where i indexes the "estimation data" and j the "evaluation data," which are typically the same. The kernel function $k(\cdot)$ for continuous variables satisfies $\int k(u)du = 1$ and some other regularity conditions depending on its order p, and $h = [h_1, \ldots, h_{q_1}]^\top$ is a vector of smoothing parameters along with h_y satisfying $h_s \to 0$ as $n \to \infty$ for $s = 1, \ldots, q_1$, and y. Similarly the kernel function $l(\cdot)$ for discrete variables lies between 0 and 1, and $\lambda = [\lambda_1, \ldots, \lambda_{q_2}]^\top$ is a vector of smoothing parameters such that $\lambda_s \in [0,1]$, and $\lambda_s \to 0$ as $n \to \infty$ for $s = 1, \ldots, q_2$ (see, e.g., Li & Racine, 2003).

Conditional CDF Estimation

Let $\mathbb{I}(\cdot)$ be the indicator function that equals 1 if its argument is true, and 0 otherwise. Then, the conditional CDF of y_j given $\mathbf{x}_j$,

$$F(y_j|\mathbf{x}_j) = E[\mathbb{I}(Y \leq y_j)|\mathbf{X} = \mathbf{x}_j]$$

can be estimated consistently by

$$\hat{F}(y_j|\mathbf{x}_j) = \frac{\sum_{i=1,\, i\neq j}^n G(y_i, y_j; h_y) K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}{\sum_{i=1,\, i\neq j}^n K(\mathbf{x}_i^c, \mathbf{x}_j^c; h) L(\mathbf{x}_i^d, \mathbf{x}_j^d; \lambda)}$$

when $F(\cdot|\cdot)$ is at least twice continuously differentiable, such that $nh_1 \times \ldots \times h_{q_1} \to \infty$ as $n \to \infty$. This estimator is asymptotically normally distributed under further regularity conditions (see, e.g., Li & Racine, 2007, Theorem 6.5, p. 194).

Conditional Quantile Estimation

The conditional τ-quantile function of y given $\mathbf{x}_j$ can be estimated consistently by

$$\hat{q}_\tau(\mathbf{x}_j) = \arg\min_q |\tau - \hat{F}(q|\mathbf{x}_j)| \tag{A.4}$$

when $q_\tau(\cdot)$ is assumed to be at least twice continuously differentiable with respect to $\mathbf{x}^c$, such that $nh_1 \times \ldots \times h_{q_1} \to \infty$ as $n \to \infty$. This estimator has also been shown to be asymptotically normally distributed under certain regularity conditions (see, e.g., Li & Racine, 2007, Theorem 6.7, pp. 195–196).

NONPARAMETRIC ESTIMATION OF PRODUCTION RISK AND RISK PREFERENCE FUNCTIONS

Subal C. Kumbhakar and Efthymios G. Tsionas

ABSTRACT

This paper deals with estimation of risk and the risk preference function when producers face uncertainties in production (usually labeled as production risk) and output price. These uncertainties are modeled in the context of production theory where the objective of the producers is to maximize expected utility of normalized anticipated profit. Models are proposed to estimate risk preference of individual producers under (i) only production risk, (ii) only price risk, (iii) both production and price risks, (iv) production risk with technical inefficiency, (v) price risk with technical inefficiency, and (vi) both production and price risks with technical inefficiency. We discuss estimation of the production function, the output risk function, and the risk preference functions in some of these cases. Norwegian salmon farming data is used for an empirical application of some of the proposed models. We find that salmon farmers are, in general, risk averse. Labor is found to be risk decreasing while capital and feed are found to be risk increasing.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 223–260

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025010

1. INTRODUCTION

Risk in production theory is mostly analyzed under (i) output price uncertainty and (ii) production uncertainty (commonly known as production risk). Output price can be uncertain due to a variety of reasons. Perhaps the most important factor is the presence of a time lag between use of inputs and output produced. Moreover, produced output is often sold at a later date when output price is likely to be different from the date when the production plan was made. Uncertainty in output price makes profit uncertain. Profit can also be uncertain if the output is risky, which may be affected by input quantities. That is, input quantities not only determine the volume of output produced, but some of these inputs might also be affecting variability of output (often labeled as production risk). For example, fertilizer might be risk augmenting in the production of crop, while labor might decrease output risk. Here we address the implications of these risks in a framework where producers maximize expected utility of anticipated profit. In particular, we examine input allocation decisions in the presence of price uncertainty and production risk. Since input demand and output supply (as well as own and cross price elasticities, returns to scale, etc.) are affected by the presence of these uncertainties, it is desirable to accommodate uncertainty in production studies, especially in estimating the underlying production technology.

Although the theoretical work on risk in the production literature is quite extensive, there are relatively fewer empirical studies devoted to analyzing different sources of risk on production and input allocation. Most of these studies either looked at output price uncertainty (Appelbaum & Ullah, 1997; Kumbhakar, 2002; Sandmo, 1971; Chambers, 1983) or production risk along the Just–Pope framework (Tveteras, 1999, 2000; Asche & Tveteras, 1999; Kumbhakar & Tveteras, 2003). To examine producers' behavior under risk, some parametric forms of the utility function, production function, and output risk function along with specific distributional assumptions on the error term representing risk are considered in the existing literature (Love & Buccola, 1991; Saha, Shumway & Talpaz, 1994). Thus, the risk studies in the production literature have some or all of these features built in, viz., (i) parametric forms of the production and risk function, (ii) parametric form of the utility function, and (iii) distributional assumption(s) on the error term(s) representing either production risk or output price uncertainty or both.

In the present paper we estimate the production function, the risk function (output risk), and risk preference functions (associated with price

and production uncertainties). We derive estimates of risk preference functions that do not depend on specific functional form of the underlying utility function. In estimating these functions no distributional assumptions are made on the random terms associated with production and output uncertainties. Furthermore, we obtain estimates of producer-specific risk premium (RP).

The rest of the paper is organized as follows. The models with price uncertainty and production risk are presented in Section 2. Extensions of these models to accommodate technical inefficiency are considered in Section 3. Section 4 describes various parametric econometric models first without and then with technical inefficiency. Nonparametric versions of some of the models are considered in Section 5. The Norwegian salmon farming and the empirical results are presented in Section 6. Finally, Section 7 concludes the paper with a brief summary of results.

2. RISK MODELS WITH OUTPUT PRICE UNCERTAINTY AND PRODUCTION RISK

We assume that the production technology can be represented by a Just–Pope (1978) form, viz.,

$$y = f(X, Z) + h(X, Z)\varepsilon, \quad \varepsilon \sim (0, 1) \tag{1}$$

where y is output, X and Z are vectors of variable and quasi-fixed inputs, $f(X, Z)$ is the mean output function, and ε is a random variable that represents production uncertainty. Since output variance is represented by $h^2(X, Z)$, the $h(X, Z)$ function is labeled as the output risk function. In this framework an input j is said to be risk increasing (decreasing) if the partial derivative $h_j(X, Z) > (<) 0$.

2.1. Only Production Risk (Model I)

First we start with the case where output and input markets are competitive and their prices are known with certainty. Production is, however, uncertain. Assume that producers maximize expected utility of anticipated normalized profit $E[U(\pi^e/p)]$ to choose optimal input quantities, which in turn determines output supply.[1] Define anticipated profit π^e as

$$\pi^e = py - wX = pf(X, Z) - wX + ph(X, Z)\varepsilon \equiv \mu_\pi + ph(X, Z)\varepsilon \tag{2}$$

where $\mu_\pi = pf(X, Z) - wX$, p being the output price and w the price vector of the variable inputs. Note that we have not subtracted the cost of quasi-fixed inputs to define profit. That is, profit in Eq. (2) is defined as variable (restricted) profit. The concept of variable/restricted profit is appropriate here because by definition quasi-fixed inputs are not choice variables (in the optimization problem) in the short run. In other words, the variable inputs are choice variables in maximizing profit in the short run. Thus, for example, capital (which is often decided from a medium-/long-term perspective) in most of the studies is treated as quasi-fixed input. The advantage of doing so is that it is not necessary to construct price of capital (which is nontrivial).

The first-order conditions (FOCs) of expected utility of anticipated normalized profit $E[U(\pi^e/p)]$ maximization can be written as

$$E\left[U'\left(\frac{\pi^e}{p}\right)\{f_j(X, Z) - \tilde{w}_j + h_j(X, Z)\varepsilon\}\right] = 0 \tag{3}$$

where $U'(\pi^e/p)$ is the marginal utility of anticipated normalized profit, $f_j(X, Z)$ and $h_j(X, Z)$ are partial derivatives of $f(X, Z)$ and $h(X, Z)$ functions, respectively, with respect to input X_j. Finally, $\tilde{w}_j = w_j/p$.

We can rewrite the above FOCs as

$$f_j(X, Z) = \tilde{w}_j - h_j(X, Z)\theta_1(\cdot) \tag{4}$$

where

$$\theta_1(\cdot) \equiv \frac{E[U'(\pi^e/p)\varepsilon]}{E[U'(\pi^e/p)]} \tag{5}$$

The $\theta_1(\cdot)$ term in the FOCs in Eq. (4) is the risk preference function associated with production risk. If producers are risk averse, then $\theta_1(\cdot) < 0$ (i.e., an increase in ε (which can be viewed as a positive production/technological shock) increases π^e/p which in turn reduces $U'(\pi^e/p)$ since $U''(\pi^e/p) < 0$ (utility function being concave)). Similarly, $\theta_1(\cdot)$ is positive if producers are risk lovers and is zero for risk neutral producers.

If $h_j(X, Z) > 0$, then for risk averse producers the value of the (expected) marginal product of input X_j exceeds its price $p\, f_j(\cdot) > w_j$. Consequently, a risk averse producer will use the input less relative to a risk neutral producer $\theta_1(\cdot) = 0$. Similarly, if producer A is more risk averse than an otherwise identical producer B, producer A will use less of input X_j than producer B. Thus, input demand functions (the solution of X_j from Eq. (4)) will depend not only on observed prices but also on the risk preference functions. Consequently, anything that depends on the demand functions (e.g., own

and cross price elasticities, returns to scale, technical change, etc.) is likely to be affected by the presence of risk via $\theta_1(\cdot)$. Since input demand functions are affected, output supply will also be affected even if the producers share the same technology, and face the same input and output prices.

2.2. Only Output Price Uncertainty (Model II)

We now consider the case where output price is uncertain (Appelbaum & Ullah, 1997; Sandmo, 1971) and there is no production uncertainty ($h(X, Z)$ is constant). We describe output price uncertainty by postulating anticipated price p^e as pe^{η} with the assumption that $E(e^{\eta}) = 1$ (Zellner, Kmenta, & Dreze, 1966) so that the expected value of p^e is the same as the observed price p. Note that in this specification p^e is random (not p) because η is a random variable. The anticipated price differs from the observed price at a point in time because the production process is not always instantaneous, and the quantity of output cannot be perfectly predicted at the time production decisions are made.

Similar to Model I, we assume that producers maximize expected utility of anticipated normalized profit $E[U(\pi^e/p)]$ to determine optimal input quantities, which in turn determines output supply. The production function is the same as in Eq. (1). Define anticipated profit π^e as

$$\begin{aligned} \pi^e &= p^e y - wX = pf(X,Z) - wX + pf(X,Z)(e^{\eta} - 1) \\ &\Rightarrow \frac{\pi^e}{p} = f(X,Z) - \tilde{w}X + f(X,Z)(e^{\eta} - 1) = \mu_{\pi} + f(X,Z)z_1 \end{aligned} \tag{6}$$

where $z_1 = (e^{\eta} - 1)$ and $\tilde{w}_j = w_j/p$. Note that z_1 is a zero mean random variable since e^{η} is a random variable with mean zero.

The FOCs of expected utility of anticipated normalized profit $E[U(\pi^e/p)]$ maximization can be written as

$$E\left[U'\left(\frac{\pi^e}{p}\right)\{f_j(X,Z) - \tilde{w}_j + f_j(X)z_1\}\right] = 0 \tag{7}$$

We can rewrite Eq. (7) as

$$f_j(X,Z)(1 + \theta_2(\cdot)) = \tilde{w}_j \tag{8}$$

where

$$\theta_2(\cdot) \equiv \frac{E[U'(\pi^e/p)z_1]}{E[U'(\pi^e/p)]} \tag{9}$$

The $\theta_2(\cdot)$ term in the FOCs (Eq. (9)) is the risk preference function associated with output price uncertainty. If producers are risk averse, then $\theta_2(\cdot)<0$ (i.e., an increase in e^{η} increases π^e/p which in turn reduces $U'(\pi^e/p)$ since $U''(\pi^e/p)<0$ (utility function being concave)). Similarly, $\theta_2(\cdot)$ is positive if producers are risk lovers and is zero for risk neutral producers.

2.3. Both Production Risk and Output Price Uncertainty (Model III)

Now we consider the case where producers face both production risk and uncertainty in output price. Output price is assumed to be governed by the same process as in Model II, and the production function is given in Eq. (1). For simplicity we assume that ε is independent of η. Furthermore the variance of e^{η} is assumed to be constant.

With the presence of both types of uncertainties the anticipated normalized profit π^e/p can be written as

$$\begin{aligned}\frac{\pi^e}{p} &= e^{\eta}y - \tilde{w}X = f(X,Z) - \tilde{w}X + f(X,Z)(e^{\eta}-1) + h(X,Z)(e^{\eta}\varepsilon) \\ &\equiv \mu_{\pi} + f(X,Z)z_1 + h(X,Z)z_2\end{aligned} \tag{10}$$

where $z_1 = e^{\eta}-1$ and $z_2 = e^{\eta}\varepsilon$. The FOCs of expected utility of anticipated profit $E[U(\pi^e/p)]$ maximization can be written as

$$E\left[U'\left(\frac{\pi^e}{p}\right)\{f_j(X,Z) - \tilde{w}_j + f_j(X,Z)z_1 + h_j(X,Z)z_2\}\right] = 0 \tag{11}$$

where $U'(\pi^e/\mathrm{p})$, $f_j(\cdot)$, and $h_j(\cdot)$ are the same as before.

We can rewrite Eq. (11) as

$$f_j(X,Z)(1+\tilde{\theta}_2(\cdot)) = \tilde{w}_j - h_j(X,Z)\tilde{\theta}_1 \tag{12}$$

where

$$\tilde{\theta}_1(\cdot) \equiv \frac{E(U'(\pi^e/p)z_2)}{E(U'(\pi^e/p))} \tag{13}$$

and

$$\tilde{\theta}_2(\cdot) \equiv \frac{E(U'(\pi^e/p)z_1)}{E(U'(\pi^e/p))} \tag{14}$$

The $\tilde{\theta}_1(\cdot)$ and $\tilde{\theta}_2(\cdot)$ functions in Eqs. (13) and (14) are called risk preference functions associated with output price uncertainty and

production risk, respectively.[2] If producers are risk averse, then $\tilde{\theta}_2(\cdot)<0$. A similar reasoning shows that $\tilde{\theta}_2(\cdot)=0$ when producers are risk neutral (i.e., $U''(\pi^e/p)=0$, which implies that the utility function is linear), and if producers are risk loving, then $\tilde{\theta}_2(\cdot)>0$. (i.e., $U''(\pi^e/p)=0$, which means that the utility function is convex). Finally, it can be shown, using similar arguments, that $\tilde{\theta}_1(\cdot)$ is negative if producers are risk averse, positive for risk loving, and zero for risk neutral producers.

The model with only output price uncertainty can be obtained from the above model by assuming that there is no output risk (i.e., $h(X, Z)$ is a constant thereby meaning that $h_j(X, Z)=0$). This means that the $\tilde{\theta}_1(\cdot)$ function will disappear from the FOCs. Similarly, if there is only production risk and no uncertainty in output price, then $z_1=0$, and the $\tilde{\theta}_2(\cdot)$ function will disappear from the FOCs. Finally, if the producers are risk neutral, then both $\tilde{\theta}_1(\cdot)$ and $\tilde{\theta}_2(\cdot)$ will disappear from the FOCs in Eq. (12).

3. RISK MODELS WITH TECHNICAL EFFICIENCY

3.1. Only Production Risk (Model IV)

If the producers face production risk and are technically inefficient, the production function can be written as

$$Y=f(X,Z)+h(X,Z)\varepsilon-g(X,Z)u \quad h(X,Z)>0, \quad g(X,Z)>0, \quad u\geq 0 \tag{15}$$

In this specification, $u\geq 0$ represents technical inefficiency. For estimation purposes u is often assumed to be truncated (or half) normal. Furthermore, u and ε are assumed to be independent. This model in Eq. (15) is a generalization of the Battese, Rambaldi, and Wan (1997) model. If $h(X, Z)=g(X, Z)$, then the model reduces to the Battese et al. (1997) model.

We assume that producers maximize $E[U(\pi^e/p)]$ conditional on u. Anticipated profit π^e is

$$\pi^e=pY-wX \Rightarrow \frac{\pi^e}{p}=f(X,Z)+h(X,Z)\varepsilon-g(X,Z)u-\left(\frac{w}{p}\right)X$$

The FOCs of $E[U(\pi^e/p)]$ maximization, given u, are

$$\begin{aligned} &E[U'(.)\{f_j(X,Z)+h_j(X,Z)\varepsilon - g_j(X,Z)u-\tilde{w}_j\}]=0 \\ &\Rightarrow f_j(X,Z)-g_j(X,Z)u+h_j(X,Z)\frac{E[U'(.)\varepsilon]}{E[U'(.)]}-\tilde{w}_j=0 \\ &\Rightarrow f_j(X,Z)-\tilde{w}_j-g_j(X,Z)u+h_j(X,Z)\lambda(.)=0 \end{aligned} \tag{16}$$

where $\lambda_1(\cdot)=(E[U'(.)\varepsilon])/(E[U'(.)])$ is the risk preference function associated with production risk. The only difference between $\lambda_1(\cdot)$ and $\theta_1(\cdot)$ is that $\lambda_1(\cdot)$ depends on inefficiency as well through the utility function.

3.2. Only Output Price Uncertainty (Model V)

Now we introduce the presence of technical inefficiency into the model with only output price uncertainty. The production function is

$$Y=f(X,Z)+h_0\varepsilon-g(X,Z)u$$

where h_0 is a constant. This is basically a stochastic frontier model in which determinants of technical inefficiency are modeled through the scaling function $g(X, Z)$ (see Wang & Schmidt, 2002). Since we are considering an optimizing model and output price is uncertain, input choices will be affected by price uncertainty. Here we are interested in estimating the production function, determinants of technical inefficiency, and the risk preference function associated with output price uncertainty.

As before, we assume that producers choose X by maximizing $E[U(\pi^e/p)]$ where $\pi^e=pY-wX \Rightarrow \pi^e/p=e^{\eta}[f(X,Z)+h_0\varepsilon-g(X,Z)u-\tilde{w}X]$. We rewrite anticipated normalized profit as

$$\begin{aligned} &\frac{\pi^e}{p}=f(X,Z)-\tilde{w}X-g(X,Z)ue^{\eta}+h_0\varepsilon e^{\eta}+f(X,Z)(e^{\eta}-1) \\ &\Rightarrow \frac{\pi^e}{p}=f(X,Z)-\tilde{w}X-g(X,Z)\,(1+z_1)+h_0z_2+f(X,Z)z_1 \end{aligned} \tag{17}$$

The FOCs of maximization $E[U(\pi^e/p)]$ with respect to the elements of X (given u) are

$$\begin{aligned} &E[U'(.)\{f_j(X,Z)-\tilde{w}_j-g_j(X,Z)u(1+z_1)+f_j(X,Z)z_1\}]=0 \\ &\Rightarrow f_j(X,Z)-\tilde{w}_j-g_j(X,Z)u(1+\lambda_2(.))+f_j(X,Z)\lambda_2=0 \end{aligned} \tag{18}$$

where $\lambda_2(\cdot)=E[U'(.)z_1]/E[U'(.)]$ is the risk preference function associated with price risk.

3.3. Both Production Risk and Price Uncertainty (Model VI)

In this section we introduce both output price and production uncertainty into the analysis. The production function is the same as the one in Eq. (15), that is,

$$Y = f(X, Z) + h(X, Z)\varepsilon - g(X, Z)u$$

Output price uncertainty is modeled as before (in Model II), that is, $p^{\mathrm{e}} = pe^{\eta}$ such that $E[e^{\eta}] = 1$ and $V(e^{\eta}) = \beta^2 > 0$. Furthermore u, ε, and η are independent of each other. Here our objectives are to estimate (i) the production risk function $h(X, Z)$; (ii) technical inefficiency u and the determinants of technical inefficiency through the scaling function $g(X, Z)$; and (iii) the risk preference functions associated with production risk and output price uncertainty.

As before, we assume that producers choose X by maximizing $E[U(\pi^e/p)]$ where $\pi^e = pY - wX \Rightarrow \pi^e/p = e^{\eta}[f(X, Z) + h(X, Z)\varepsilon - g(X, Z)u] - \tilde{w}X$.

Now we rewrite anticipated profit as

$$\begin{aligned}\frac{\pi^e}{p} &= f(X, Z) - \tilde{w}X - g(X, Z)ue^{\eta} + h(X, Z)\varepsilon e^{\eta} + f(X, Z)(e^{\eta} - 1) \\ &= f(X, Z) - \tilde{w}X - g(X, Z)(1 + z_1) + h(X, Z)z_2 + f(X, Z)z_1\end{aligned} \tag{19}$$

The FOCs of maximization $E[U(\pi^e/p)]$ with respect to the elements of X (given u) are

$$\begin{aligned}&E[U'(.)\{f_j(X, Z) - \tilde{w}_j - h_j(X, Z)z_2 - g_j(X, Z)ue^{\eta} + f_j(X, Z)z_1\}] = 0 \\ &\Rightarrow f_j(X, Z) - \tilde{w}_j + h_j(X, Z)\tilde{\lambda}_2 - g_j(X, Z)u(1 + \tilde{\lambda}_1) + f_j(X, Z)\tilde{\lambda}_1 = 0\end{aligned} \tag{20}$$

where $\tilde{\lambda}_1 = E[U'(.)z_1]/E[U'(.)]$ and $\tilde{\lambda}_2 = E[U'(.)z_2]/E[U'(.)]$ are risk preference functions associated with price and production risks, respectively.

4. PARAMETRIC ECONOMETRIC MODELS OF RISK

Since our interest is to estimate the parameters of the mean output function, output risk function, and the risk preference function, the most important task is to derive an algebraic form of the risk preference function, which is easy to implement econometrically, and imposes minimum restrictions on the structure of risk preferences on the individual producers. Certain specific

forms of $U(.)$ together with some specific distributional assumptions on ε give an explicit closed form solution of $\theta_1(.)$ (Love & Buccola, 1991; Saha et al., 1994). However, estimation of these models is quite complex. It is, however, possible to derive an algebraic expression for the risk preference function without assuming any distribution on ε and without any specific functional form on $U(.)$ that imposes a priori restrictions on the structure of risk aversion.[3] In fact, our result would be very useful in empirical applications, especially if one is interested in estimating general forms of risk preferences without estimating a complicated system of equations (Chavas & Holt, 1996; Love & Buccola, 1991; Saha et al., 1994). Note that it is not even necessary to assume that $U(\cdot)$ is concave.

4.1. Specification and Estimation of Model I

If $U(\mu_\pi + ph(X, Z)\varepsilon)$ is continuous and differentiable, and we take a linear approximation of $U'(\mu_\pi + ph(X, Z)\varepsilon)$ at $\varepsilon = 0$, then the risk preference function in Model I takes the following form[4]:

$$\theta_1(\cdot) = -\mathrm{AR}(\mu_\pi)h(X, Z) \tag{21}$$

where $\mathrm{AR}(\mu_\pi) = -U''(\mu_\pi)/U'(\mu_\pi)$ is the Arrow–Pratt measure of absolute risk aversion.

Using the above result the FOC in Eq. (4) can be expressed as

$$f_j(X, Z) = \tilde{w}_j + h_j(X, Z)\mathrm{AR}(\mu_\pi)h(X, Z) \tag{22}$$

A close look at the FOC in Eq. (22) shows that the focus of the problem is now shifted from the utility function to the AR function. In addition to the mean production and risk functions, one needs to specify a functional form on AR, which will define a system of J equations in J variable inputs (X) in Eq. (22). It is worth noting here that any specification of the AR function will indirectly imply some underlying utility function, viz., $U = \int e^{-\mathrm{AR}} d\mu_\pi$. That is, the AR function gives all the information possessed by the utility function (Pratt, 1964). The main advantage of working with the AR function is that one doesn't have to worry about (i) the underlying utility function (which may not be always solvable analytically), (ii) the derivation of $\theta_1(\cdot)$ (which might not always give a closed form solution), and (iii) the solution $\theta_1(\cdot)$ (which, although solvable for some specific utility functions, might not be easy to work with empirically). Furthermore, one can assume a functional form on AR that is flexible enough to test whether producers are risk neutral ($\mathrm{AR} = 0$) or not. If risk neutrality does not exist, then we can

also test for constant absolute risk aversion (CARA), decreasing absolute risk aversion (DARA), and increasing absolute risk aversion (IARA) hypotheses.

AR can be parameterized to allow (test) for CARA, IARA, and DARA. For example, if $AR = \delta_1 + \delta_2\mu_\pi + 0.5\delta_3\mu_\pi^2$, then $CARA \Rightarrow \delta_2 = \delta_3 = 0$, $IARA \Rightarrow \delta_2 + \delta_3\mu_\pi > 0$, and $DARA \Rightarrow \delta_2 + \delta_3\mu_\pi < 0$. Furthermore, $\delta_1 = \delta_2 = \delta_3 = 0 \Rightarrow AR = 0 \Rightarrow \theta = 0$, that is, risk neutrality. These are all testable hypotheses. Some other nonlinear functions can also be used to parameterize and test different forms of risk preferences. Although a parametric form on AR indirectly implies some form of a utility function, it is not necessary to know the exact parametric form of the underlying utility function in specifying a functional form for AR. Note that although the specification of the models under the abovementioned null hypotheses are well defined, the models under the alternative hypotheses are not unique. That is, one can test a specific null hypothesis (e.g., CARA) by specifying many different AR functions. Since the tests used in the literature are always against some specific alternatives, it is worth mentioning that the test results might be inconsistent if the models under the alternatives are incorrect.

The model outlined above (Model I) can be estimated by estimating the system consisting of the production function in Eq. (1) along with the FOCs in Eq. (22) once parametric functional forms are chosen for $f(X, Z)$, $h(X, Z)$, and AR(.) functions, and classical error terms are added to each of the FOCs in Eq. (22). Two things are to be noted here. First, the system is highly complicated and nonlinear is parameters, and therefore a nonlinear system approach has to be used. Second, the endogenous variables are the variable inputs (X) and output (Y), which appear almost everywhere in the system. Thus, a nonlinear three-stage least squares or other instrumental variable approach (system GMM) has to be used. The exogenous variables (instruments) are the quasi-fixed inputs (Z) and prices (p and w).[5]

4.2. Specification and Estimation of Model II

A similar procedure can be used to estimate Model II that incorporates only output price risk discussed in Section 2.2. We use the following result to express the risk preference function in terms of the AR function.

If $U(\mu_\pi + f(X, Z)z_1 + z_2)$ is continuous and differentiable, and we take a linear approximation of $U'(\mu_\pi + f(X, Z)z_1 + z_2)$ at $z_1 = z_2 = 0$, then the risk preference function takes the following form[6]:

$\theta_2(\cdot) = -\mathrm{AR}(\mu_\pi).f(X, Z)$, where $\mathrm{AR} = -U''(\cdot)/\mathrm{U}'(\cdot)$ evaluated at μ_π.

Using this result we write the FOCs in Eq. (8) as

$$f_j(X,Z)[1-\mathrm{AR}(\mu_\pi)f(X,Z)] = \tilde{w}_j + v_j \tag{23}$$

where v_j can be viewed as an optimization error in choosing the jth variable input. Thus, the estimating model consists of the production function in Eq. (1) and the FOCs in Eq. (23) that can be estimated using a nonlinear system approach. This system is also heavily parametric and difficult to estimate.

4.3. Specification and Estimation of Model III

To estimate Model III that incorporates both production and output price risk discussed in Section 2.3, we express the risk preference functions (specified in Eqs. (13) and (14)) in terms of the AR function.

If $U(\mu_\pi+f(X, Z)\, z_1+h(X, Z)z_2)$ is continuous and differentiable, and we take a linear approximation of $U'(\mu_\pi+f(X, Z)\, z_1+h(X, Z)z_2)$ at $z_1 = z_2 = 0$, then the risk preference functions are

$$\tilde{\theta}_2(\cdot) = -\mathrm{AR}(\mu_\pi)f(X,Z), \quad \tilde{\theta}_1(\cdot) = -\mathrm{AR}(\mu_\pi)h(X,Z)$$

Using this result we write the FOCs in Eq. (12) as

$$f_j(X,Z)[1-\mathrm{AR}(\mu_\pi)f(X,Z)] = \tilde{w}_j + h_j(X,Z)h(X,Z)\mathrm{AR}(\mu_\pi) + v_j \tag{24}$$

where v_j can be viewed as an optimization error in choosing the jth variable input. Thus, the estimating model consists of the production function in Eq. (1) and the FOCs in Eq. (24) that can be estimated using a nonlinear system approach.

4.4. Specification and Estimation of Model IV

To derive an estimable expression of $\lambda_1(\cdot)$, we express it, as before, in terms of the AR$(\cdot)$ function. For this, first, we expand $U'(\pi^e/p)$ around $\varepsilon = 0$, that is,

$$U'\left(\frac{\pi^e}{p}\right) = U'(q(X,Z,u)) + U''(q(X,Z,u))h(X,Z)\varepsilon + \cdots$$

where $q(X,Z,u) = f(X,Z) - g(X,Z)u - \tilde{w}X$.

Thus,

$$\left.\begin{aligned} E[U'(.)] &= U'(q(X,Z,u)), \\ E[U'(.)\varepsilon] &= U''(q(X,Z,u))h(X,Z) \end{aligned}\right\} \text{ ignoring higher order terms}$$
$$\Rightarrow \lambda_1(\cdot) = \frac{U''(q(X,Z,u))h(X,Z)}{U'(q(X,Z,u))} = -\mathrm{AR}(X,Z,u)h(X,Z) \tag{25}$$

where $\mathrm{AR}(X,\ Z,\ u) = -U''(\cdot)/U'(\cdot)$ is the Arrow–Pratt absolute risk aversion function evaluated at $q(X,\ Z,\ u)$. For risk averse producers $\lambda_1(\cdot) < 0 \Rightarrow \mathrm{AR}(.) > 0$.

Using the above expression for $\lambda_1(\cdot)$, we write Eq. (16) as:

$$\begin{aligned} &f_j(X,Z) - \tilde{w}_j - g_j(X,Z)u + h_j(X,Z)\,[-\mathrm{AR}(X,Z,u)\,h(X,Z)] = v_j \\ &\quad\Rightarrow f_j(X,Z) - \tilde{w}_j - \mathrm{AR}(X,Z,u)\,h_j(X,Z)\,h(X,Z) = v_j + g_j(X,Z)u \end{aligned} \tag{26}$$

where the error term v_j in Eq. (26) can be viewed as optimizing error associated with the jth variable input.

Estimation of the above model can be done in either two steps or a single step.

4.4.1 Two-Step Procedure

Step 1. Use the maximum likelihood (ML) method to estimate the production function in Eq. (15) with the following distributional assumptions on u and ε:[7]

(i) $u \sim$ i.i.d. $N^+(\mu, \sigma_u^2)$,
(ii) $\varepsilon \sim$ i.i.d. $N(0, 1)$,
(iii) u and ε are independent.

In specifying the variance of ε to unity we assume that the $h(X,\ Z)$ function is proportional to a constant. Based on the above distributional assumptions, the likelihood function can be derived by making a few changes to the one derived in Battese et al. (1997).[8] By specifying parametric functional forms for $f(X,\ Z)$, $h(X,\ Z)$, and $g(X,\ Z)$, one can obtain estimates of the parameters in $f(X,\ Z)$, $h(X,\ Z)$, and $g(X,\ Z)$, as well as μ and σ_u^2.

These parameters can then be used to estimate u (for each observation) from either the mean or mode of $u|\varepsilon^*$ where $\varepsilon^* = h(X,\ Z)\varepsilon - g(X,\ Z)u$ (see the appendix). It is straightforward to show that the conditional distribution of u is truncated normal. Once u is estimated, technical

efficiency (TE) can be estimated from

$$\mathrm{TE} = \frac{E(Y|X,Z,u)}{E(Y|X,Z,u=0)} = 1 - \frac{g(X,Z)u}{f(X,Z)} \tag{27}$$

Step 2. Step 1 gives estimates of $f(X, Z)$, $g(X, Z)$, and $h_1(X, Z)$, as well as the estimates of u. These estimates can be used in Eq. (26) to compute $\lambda_1(.)$ and AR as follows:

$$\begin{aligned}
&\sum_j (f_j(X,Z) - \tilde{w}_j - g_j(X,Z)u) = \sum_j v_j - \lambda_1(X,Z,u)\sum_j h_j(X,Z) \\
&\Rightarrow \hat{\lambda}_1(X,Z,u) = -\frac{\sum_j (f_j(X,Z) - \tilde{w}_j - g_j(X,Z)u)}{\sum_j h_j(X,Z)} \\
&\Rightarrow \widehat{\mathrm{AR}}\,(X,Z,u) = -\frac{\hat{\lambda}_1(X,Z,u)}{h(X,Z)}
\end{aligned} \tag{28}$$

assuming that $\sum_j v_j = 0$. These estimates are observation specific. Thus, one can obtain estimates of risk preference (and absolute risk aversion) for each observation.

An alternative strategy is to assume a functional for AR and estimate the parameters of it from the FOCs in Eq. (26), which is rewritten as

$$\frac{[\hat{f}_j(X,Z) - \tilde{w}_j - \hat{g}_j(X,Z)u]}{[\hat{h}_j(X,Z)\,\hat{h}(X,Z)]} = \mathrm{AR}(X,Z,u) + v_j \quad j = 1,\ldots,J \tag{29}$$

where v_j is an error term.

For example, if the AR function is assumed to be linear, that is,

$$\mathrm{AR} = b_0 + b_1 q(X,Z,u) = b_0 + b_1(f(X,Z) - \tilde{w}X - g(X,Z)u) \tag{30}$$

one can substitute AR from Eq. (30) into Eq. (29) and estimate b_0 and b_1 parameters from the system of J equations in Eq. (29), using the estimated values of $f(X, Z)$, $g(X, Z)$, and u. It is to be noted that the X variables are endogenous variables. This means that one should use instruments for the X variables.

4.4.2 Single-Step Procedure

We write the FOCs in Eq. (29) as

$$\psi_j(X,Z) = m_j(X,Z)u + v_j \quad j = 1,\ldots,J$$

where $\psi_j(X,Z) = f_j(X,Z_i) - \tilde{w}_j - h(X,Z)h_j(X,Z)\,[b_0 + b_1((f(X,Z) - \tilde{w}X)]$ and $m_j(X,Z) = g_j(X,Z) - b_1 h_j(X,Z)h(X,Z)g(X,Z)$.

The above FOCs together with the production function in Eq. (15) constitute the full system of $J+1$ equations with $J+1$ endogenous variables, which is written compactly as

$$\begin{bmatrix} Y - f(X,Z) \\ \Psi_1(X,Z) \\ \Psi_2(X,Z) \\ \vdots \\ \Psi_J(X,Z) \end{bmatrix} = \begin{bmatrix} h(X,Z)\varepsilon \\ v_1 \\ v_2 \\ \vdots \\ v_J \end{bmatrix} - u \begin{bmatrix} g(X,Z) \\ -m_1(X,Z) \\ -m_2(X,Z) \\ \vdots \\ -m_J(X,Z) \end{bmatrix} \tag{31}$$

The problem of dealing with this system is that the likelihood function (based on the distributions on ε, v, and u) cannot be expressed in a closed form. This is because the Jacobian of the transformation will depend on u. Because of this problem we do not discuss the full ML method here.

4.5. Specification and Estimation of Model V

To derive an estimable expression for λ_2 we take a Taylor series expansion of U' at $z_1 = z_2 = 0$, given u. This gives

$$\begin{aligned} U'\left(\frac{\pi^e}{p}\right) &= U'(q(X,Z,u)) + U''(q(X,Z,u))\, h_0 z_2 \\ &\quad + U''(q(X,Z,u))\,[f(X,Z) - g(X,Z)u]z_1 \end{aligned}$$

where $q(X,Z,u) = f(X,Z) - g(X,Z)\,u - \tilde{w}X$. As before we assume that η and ε are independent.

Thus,

$$E[U'(.)] = U'(q(X,Z,u))$$

and

$$E[U'(.)z_1] = U''(q(X,Z,u))\,[f(X,Z) - g(X,Z)u]$$

$$\begin{aligned} \Rightarrow \lambda_2(\cdot) &= \frac{U''(q(X,Z,u))\,[f(X,Z) - g(X,Z)u]}{U'(q(X,Z,u))} \\ &= -\mathrm{AR}(X,Z,u)\,[f(X,Z) - g(X,Z)u] \end{aligned}$$

using the result $\mathrm{AR}(\cdot) = -(U''(.))/(U'(.))$ evaluated at $q(X, Z, u)$.

Using the above results, we rewrite the FOCs in Eq. (18) as

$$\begin{aligned} & f_j(X,Z) - \tilde{w}_j - g_j(X,Z)\,u \\ & \quad = \mathrm{AR}(\cdot)\,[\{f(X,Z) - g(X,Z)u\}\{f_j(X,Z) - g(X,Z)\,u\}] \end{aligned} \tag{32}$$

We write Eq. (32) more compactly as

$$\frac{\Psi_{1j}(X,Z,u)}{[m_{1j}(X,Z)\,u]} = \mathrm{AR}(\cdot) \quad j = 1,\ldots,J \tag{33}$$

when $\Psi_{1j}(X,Z,u) = f_j(X,Z) - \tilde{w}_j - g_j(X,Z)\,u$, and $m_{1j}(X,Z) = [f(X,Z) - g(X,Z)u]\,[f_j(X,Z) - g_j(X,Z)\,u]$.

Given the complexity of the model we suggest a two-step procedure.

Step 1. We estimate the production function in Eq. (15) following the procedure discussed in section 4.4.1. By specifying parametric functional forms for $f(X, Z)$ and $g(X, Z)$ together with the distributions on u and ε, one can obtain ML estimates of the parameters in $f(X, Z)$ and $g(X, Z)$, as well as μ, σ_u^2, and h_0. These estimators are consistent.

Step 2. Use the estimated/predicted values from Step 1 to compute Ψ_j and m_j.

Assume a functional form for AR, for example, $\mathrm{AR} = b_0 + b_1(f(X,Z) - \tilde{w}X - g(X,Z)\,u)$. Using this specification, we rewrite Eq. (33) as

$$\frac{\hat{\Psi}_{1j}(X,Z,u)}{[b_0 + b_1(\hat{f}(X,Z) - \tilde{w}X - \hat{g}(X,Z)\,\hat{u})]} = \hat{m}_{1j}(X,Z)\,u + \eta_j \quad j = 1,\ldots,J \tag{34}$$

where η_j is an error term appended to the jth FOC. The above nonlinear system of J equations can be used to estimate b_0 and b_1. The Z, w, and p variables can be used as instruments in estimating the above system. Once b_0 and b_1 are estimated $\mathrm{AR}(\cdot)$ can be computed for each observation.

4.6. Specification and Estimation of Model VI

As before, first we derive estimable expressions for $\tilde{\lambda}_1(\cdot)$ and $\tilde{\lambda}_2(\cdot)$ by taking a linear Taylor series expansion of $U'(\cdot)$ at $z_1 = z_2 = 0$, given u.

This gives

$$U'\left(\frac{\pi^e}{p}\right) = U'(q(X,Z,u)) + U''(q(X,Z,u))\, h(X,Z)z_2 + U''(q(X,Z,u))\,[f(X,Z) - g(X,Z)\,u]z_1$$

where $q(X,Z,u) = f(X,Z) - g(X,Z)\,u - \tilde{w}X$.

Thus,

$$E[U'(.)z_1] = U''(q(X,Z,u))\,[f(X,Z) - g(X,Z)\,u]$$

and

$$\begin{aligned} E[U'(.)z_2] &= U''(q(X,Z,u))\, h(X,Z) \\ &\Rightarrow \tilde{\lambda}_1(\cdot) = \frac{U''(q(X,Z,u))\,[f(X,Z) - g(X,Z)u]}{U'(q(X,Z,u))} \\ &= -\mathrm{AR}(X,Z,u)\,[f(X,Z) - g(X,Z)u] \\ &\Rightarrow \tilde{\lambda}_2(\cdot) = \frac{U''(q(X,Z,u))\, h(X,Z)}{U'(q(X,Z,u))} = -\mathrm{AR}(X,Z,u)\, h(X,Z) \end{aligned}$$

when $\mathrm{AR}(\cdot) = -(U''(.))/(U'(.))$ is evaluated at $q(X,Z,u) = f(X,Z) - g(X,Z)\,u - \tilde{w}X$.

Using the above results, we rewrite the FOCs in Eq. (20) as

$$\begin{aligned} & f_j(X,Z) - \tilde{w}_j - g_j(X,Z)u \\ &= \mathrm{AR}(\cdot)\,[f_j(X,Z)\{f(X,Z) - g(X,Z)u\} + h_j(X,Z)\, h(X,Z) \\ &\quad - g_j(X,Z)u\{f_j(X,Z) - g(X,Z)u\}] \\ &= \mathrm{AR}(\cdot)\,[\{(f(X,Z) - g(X,Z)u)\,(f_j(X,Z) - g_j(X,Z)u)\} + h_j(X,Z)\, h(X,Z)] \end{aligned} \tag{35}$$

We write Eq. (35) more compactly as

$$\frac{\Psi_{1j}(X,Z,u)}{[m_{1j}(X,Z)u + r_j]} = \mathrm{AR}(\cdot) \quad j = 1,\ldots,J \tag{36}$$

where $\Psi_{1j}(X,Z,u)$ and $m_{1j}(X,\ Z)$ are defined beneath Eq. (33). Finally, $r_j = h_j(X,\ Z)\ h(X,\ Z)$.

Given the complexity of the model we suggest a two-step procedure.

Step 1. We estimate the production function in Eq. (15) following the procedure discussed in the previous section. By specifying parametric functional forms for $f(X,\ Z)$, $h(X,\ Z)$, and $g(X,\ Z)$, together with the

distributions on u and ε, one can obtain ML estimates of the parameters in $f(X, Z)$, $h(X, Z)$, and $g(X, Z)$, as well as μ and σ_u^2. These estimators are consistent.

Step 2. Use the estimated/predicted values from Step 1 to compute Ψ_j and m_j.

Assume a functional form for AR, for example, $\mathrm{AR} = b_0 + b_1(f(X, Z) - \tilde{w}X - g(X, Z)u)$. Using this specification, we rewrite Eq. (31) as

$$\frac{\hat{\Psi}_j(X, Z, u)}{[\hat{m}_j(X, Z)u]} = [b_0 + b_1(\hat{f}(X, Z) - \tilde{w}X - \hat{g}(X, Z)\hat{u})] + \eta_j \quad j = 1, \ldots, J \tag{37}$$

where η_j is an error term appended to the jth FOC. The above nonlinear system of J equations can be used to estimate b_0 and b_1. The Z, w, and p variables can be used as instruments in estimating the above system. Once b_0 and b_1 are estimated, $\mathrm{AR}(\cdot)$, $\tilde{\lambda}_1(\cdot)$, and $\tilde{\lambda}_2(\cdot)$ can be computed for each observation.

Overall, it appears that estimation of the previously described systems in a parametric framework is highly complicated. Our computational experiences with some of these models (in unreported working papers) have been somewhat disappointing. Even estimating a production function of the form $y = f(x) + g(x)\varepsilon$ is, in some instances, a delicate matter that involves issues of convergence, stability of estimates, etc. The systems of FOCs are also ill-behaved in many instances and, as a result, the parametric approach is not only implausible in terms of assumptions but also highly unstable from the numerical point of view.

5. NONPARAMETRIC ESTIMATION OF MODELS I-III

5.1. Estimation of $f(X, Z)$ and $h(X, Z)$ Functions and Their Partial Derivatives

Suppose $\tilde{X} \in R^d$ is a vector of explanatory variables (that include both variable X and quasi-fixed inputs Z), and Y denotes output (the dependent variable). We assume that the production function is of the form

$$Y = f(\tilde{X}) + h(\tilde{X})\varepsilon \equiv f(\tilde{X}) + v \tag{38}$$

where f: $R^d \rightarrow R$ is an unspecified functional form, and v is an error term. Our objective is to obtain estimates of $f(\tilde{X})$ and $h(\tilde{X})$ as general is possible. So we do not consider separable specifications that are popular when dimensionality reductions are desired. We use the multivariate kernel method to obtain an estimate of $f(\tilde{X})$ at a particular point $f(\tilde{X})$ as follows. First, we estimate the density of $\tilde{X}(\tilde{p}(X))$ as

$$\tilde{p}(\tilde{X}) = (Nh)^{-1} \sum_{i=1}^{N} K_h(\tilde{X} - \tilde{X}_i) = (Nh)^{-1} \sum_{i=1}^{N} \prod_{j=1}^{d} K(Z_j - Z_i) \tag{39}$$

where $K_h(w) = \exp(-(1/2h^2)(w-w)'\tilde{\Sigma}_X^{-1}(w-w))$ is the d-dimensional normal kernel, $h>0$ is the bandwidth parameter, $K(w) = \exp(-(1/2)w^2)$ is the standard univariate normal kernel, $\tilde{\Sigma}_X$ is the sample covariance matrix of $\tilde{X}_i (i = 1, \ldots, d)$,

$$Z_i = \frac{A(\tilde{X}_i - \bar{\tilde{X}})}{\lambda}$$

$$A\tilde{\Sigma}_X A = I_d$$

$$\bar{\tilde{X}} = N^{-1} \sum_{i=1}^{N} \tilde{X}_i$$

and λ is a smoothing parameter. The optimal choices for h and λ are

$$h = \lambda^d |\tilde{\Sigma}_X|^{1/2}$$

$$\lambda = \left(\frac{4}{(2d+1)N} \right)^{d+4}$$

The unknown function is then estimated as

$$\tilde{f}(\tilde{X}) = (Nh)^{-1} \sum_{t=1}^{N} W_{hi}(\tilde{X}) Y_i \tag{40}$$

where

$$W_{hi}(\tilde{X}) \equiv \frac{K_h(\tilde{X} - \tilde{X}_i)}{\tilde{p}(\tilde{X})}$$

(see Hardle, 1990, pp. 33–34). The estimates are adjusted near the boundary using the procedures discussed in Rice (1984), Hardle (1990, pp. 130–132), and Pagan and Ullah (1999, Chapter 3).

First derivatives of $f(\tilde{X})$ with respect to X are obtained from

$$\frac{\partial \tilde{f}(\tilde{X})}{\partial X} = (Nh)^{-1} \sum_{i=1}^{N} \frac{\partial W_{hi}(\tilde{X}) Y_i}{\partial X}$$

More specifically,

$$\frac{\partial \tilde{f}(\tilde{X})}{\partial X_j} = -(Nh)^{-1} \frac{\left[\sum_{i=1}^{N} G_{ji} K_h(\tilde{X} - \tilde{X}_i) Y_i - \tilde{f}(\tilde{X}) \sum_{i=1}^{N} G_{ji} K_h(\tilde{X} - \tilde{X}_i)\right]}{\tilde{p}(\tilde{X})} \tag{41}$$

where

$$G_{ji} = \lambda^{-2} \sum_{k=1}^{d} \tilde{\sigma}_X^{jk} (\tilde{X}_k - \tilde{X}_{ki})$$

and

$$\tilde{\Sigma}_X = [\tilde{\sigma}_X^{jk}, j, k = 1, \ldots, d]$$

Given the estimate of $\tilde{f}(\tilde{X}_i)$ one can obtain the residuals e_i from $e_i = y_i - \tilde{f}(\tilde{X}_i)$. An estimate of the variance can then be obtained from

$$\tilde{\sigma}^2(\tilde{X}) = (Nh)^{-1} \sum_{i=1}^{N} W_{hi}(\tilde{X})\, e_i^2 \tag{42}$$

(see Hardle, 1990, p. 100; Pagan & Ullah, 1999, pp. 214–215). Since $g(\tilde{X}) = \tilde{\sigma}(\tilde{X})$, estimates of the $g(\tilde{X})$ function and its gradient $\partial g(\tilde{X})/\partial X$ can be obtained. Alternatively, $g(\tilde{X})$ can be obtained from a nonparametric regression of $|e_i|$ on X_i in a second step.[9] The gradient of $g(\tilde{X})$ could then be obtained by a procedure similar to the one used to obtain the gradient of $f(\tilde{X})$ in Eq. (41).

The asymptotic properties of this procedure are well established. However, the nonparametric procedure has not been used so far in applied studies, especially in agricultural economics where strong parametric and distributional assumptions are still in use. The main advantage of this approach is that the technology and risk properties can be recovered without strong and restrictive/questionable assumptions. Moreover, as we detail below, aspects of risk preference can be easily recovered in the following manner.

5.2. Estimation of Risk Preference Functions and Risk Premium

To estimate the risk preference function $\theta \equiv \theta(\tilde{X}, \tilde{w})$ in Model I we rewrite the relationship in Eq. (4) as

$$D_1 \equiv \frac{1}{J}\sum_j \left[\frac{\tilde{f}_j(\tilde{X}) - \tilde{w}_j}{-\tilde{g}_j(\tilde{X})}\right] \equiv \theta(\tilde{X}, \tilde{w}) \tag{43}$$

Note that although not stated explicitly the FOCs in Eq. (4) is allowed to have errors to capture optimization errors. Thus, the estimator of θ in Eq. (43) can be viewed as a minimum distance estimator.

Eq. (43) can be computed easily since all its components have been estimated. Therefore, fully nonparametric estimates of θ can be obtained at no cost.

In Model II the risk preference function can be expressed (using Eq. (9)) as

$$D_2 \equiv \frac{1}{J}\sum_j \left[\frac{\tilde{w}_j}{\tilde{f}_j(\tilde{X})} - 1\right] = \theta_2(\tilde{X}, \tilde{w}) \tag{44}$$

The above equation can be, again, easily computed under fully nonparametric conditions.

To estimate risk preference functions in Model III, we write the FOCs in Eq. (12) as

$$\begin{aligned}
\delta_j &\equiv \frac{\tilde{f}_j(\tilde{X})}{\tilde{f}_1(\tilde{X})} = \frac{[\tilde{w}_j - \tilde{g}_j(\tilde{X})\tilde{\theta}_1(\tilde{w}, \tilde{X})]}{[\tilde{w}_1 - \tilde{g}_1(\tilde{X})\tilde{\theta}_1(\tilde{w}, \tilde{X})]} \\
&\Rightarrow D_3 \equiv 1 + \sum_{j=2} \delta_j = \frac{1}{J}\sum_{j=1}\left[\frac{\tilde{f}_j(\tilde{X})}{\tilde{f}_1(\tilde{X})}\right] \\
&= \frac{1}{J}\sum_j \frac{[\tilde{w}_j - \tilde{g}_j(\tilde{X})\tilde{\theta}_1(\tilde{w}, \tilde{X})]}{[\tilde{w}_1 - \tilde{g}_1(\tilde{X})\tilde{\theta}_1(\tilde{w}, \tilde{X})]} \\
&\equiv \varphi(\tilde{w}, \tilde{X}) + \varsigma
\end{aligned} \tag{45}$$

where ς is an error term. Once the $\varphi(\cdot)$ function is estimated nonparametrically, we can recover $\tilde{\theta}_1(\tilde{w}, \tilde{X})$ from

$$\tilde{\theta}_1(\tilde{w}, \tilde{X}) = \frac{\sum_j [\tilde{w}_j - \tilde{\phi}(\tilde{w}, \tilde{X})\tilde{w}_1]}{\sum_j [g_j(\tilde{X}) - \tilde{\phi}(\tilde{w}, \tilde{X})\tilde{g}_1(\tilde{X})]}$$

The $\theta_2(\tilde{w}, \tilde{X})$ function can then be estimated from

$$\tilde{\theta}_2(\tilde{w}, \tilde{X}) = \frac{\sum_j [\tilde{w}_j - \tilde{g}_j(\tilde{X})\tilde{\theta}_1(\tilde{w}, \tilde{X})]}{\sum_f \tilde{f}_j(\tilde{X}) - 1}$$

One can estimate the AR functions from different specifications using the estimated values of θ_1 and θ_2.

6. APPLICATION TO NORWEGIAN SALMON FARMING

6.1. Data

Some of the models presented in the preceding sections are applied to Norwegian salmon farms. Norway, UK, and Chile are the largest producers of farmed Atlantic salmon (Bjorndal, 1990). Salmon farming is more risky than most other types of meat production due to the salmon's high susceptibility to the marine environment it is reared in. Biophysical factors such as fish diseases, sea temperatures, toxic algae, wave and wind conditions, and salmon fingerling quality are major sources of output risk.

It is believed that the effect of biophysical shocks on output risk can be influenced through the choice of input levels, although fish farmers cannot prevent occurrences of such exogenous shocks. The most important input in salmon farming is fish feed. Feed is expected to increase the level of output risk, ceteris paribus. Since salmons are not able to digest all the feed the residue is released into the environment as feed waste or feces. This organic waste consumes oxygen, and thus competes with the salmon for the limited amount of oxygen available in the cages. In addition, feed waste also leads to production of toxic by-products such as ammonia. Furthermore, production risk is expected to increase with the quantity of fish released into the cages, due to the increased consumption of oxygen and production of ammonia. We do not have any strong a priori presumptions on the risk effects of capital.

Since 1982 the Norwegian Directorate of Fisheries has compiled salmon farm production data. In the present study we use 2,447 observations on such farms observed during 1988–1992.[10] The output (y) is sales (in thousand kilograms) of salmon and the stock (in thousand kilograms) left at the pen at the end of the year. The input variables are feed (F),

labor (L), and capital (K). Feed is a composite measure of salmon feed measured in thousand kilograms. Labor is total hours of work (in thousand hours). Capital is the replacement value (in real terms) of pens, buildings, feeding equipment, etc. Price of salmon is the market price of salmon per kilogram in real Norwegian Kronors (NOK). The wage rate (in real NOK) is obtained by dividing labor cost by hours of labor. Price of feed is obtained by dividing the cost of feed by the quantity of feed.

In the present study we are treating labor and feed as variable inputs. Capital is treated as quasi-fixed input primarily because price data on it is not available. Moreover, since capital stock adjustment is not instantaneous, it is perhaps better to treat the capital variable as a quasi-fixed input, especially in the static model like the one used in the present study.

6.2. Results and Discussions

First, we report the estimated elasticities of the mean output function $f(X)$ with respect to labor, capital, and feed. We plot the empirical distribution of these elasticities for labor, capital, and feed in Fig. 1.[11] The mean values of these elasticities are: 0.029, 0.017, and 0.253, respectively. It can be seen that none of the distributions is symmetric. In fact they are all skewed to the right. Thus, the median values of these elasticities are less than their mean values (median elasticities of the mean output with respect to labor, capital, and feed are 0.017, 0.007, and 0.158, respectively). The standard deviations of these elasticities are: 0.078, 0.046, and 0.282, respectively. Although some of these elasticities are negative, this happens for a small proportion of salmon farmers. Alternatively, it is quite justifiable to do restricted estimation, and replace any negative elasticity for some farmer with its lowest allowable bound (zero), see Pagan and Ullah (1999, pp. 175–176).

Farm age is found to have a negative effect on mean output. The elasticity with respect to age is expected to be positive, especially when one associates age of the farmer with experience, knowledge, and learning. With an increase in experience and knowledge one would expect output to increase, ceteris paribus. However, salmon farm studies show that the marine environment around the farm tends to become more disease prone over time due to accumulation of organic sediments below the cages, leading to oxygen loss and increased risk of fish diseases. Hence, the farm age variable may capture both the positive learning effect and the negative disease proneness effect. According to our results, the negative disease proneness effect seems to dominate. The median (mean) value of age elasticity is

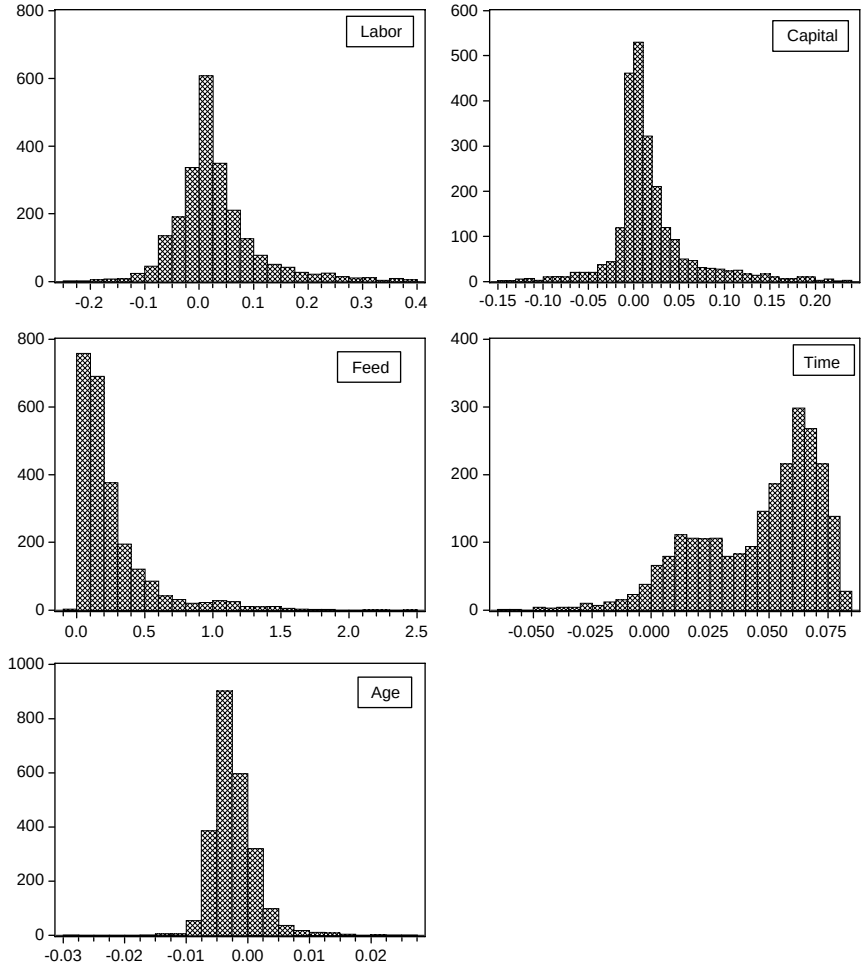


Fig. 1. Histograms of Elasticites of $f(X)$.

-0.003 (-0.002) with a standard deviation of 0.004. Similar result is found in parametric studies.

In production models the time variable is included to capture exogenous technical change (a shift in the production function, ceteris paribus). In the present model one can define technical progress in terms of the mean output function $f(X)$, that is, $\mathrm{TC} = \partial \ln f(X)/\partial t$). Based on this formula we find mean technical progress at the rate of 4.6% per year. The frequency

distribution of TC is given in Fig. 1. The distribution is skewed to the left. It seems that the average rate of TC for most of the farms is around 6%. The median value of TC is 5.3% with a standard deviation of 0.026. A notable feature of this distribution is that it is bimodal. The two modal values of TC are 2.5% and 7.5% per annum, respectively. Although the mean TC is around 6% per year, some farms experienced technical progress at the rate of 2.5% while other "leading" farms experienced a much higher rate.

For a risk neutral producer, the input elasticities (labor, feed, and capital) can be interpreted as the cost share of the input to the value of output (revenue). This is, however, not the case for a nonrisk neutral producer. It can be easily verified from the FOCs that the value of the marginal product of an input deviates from its price thereby meaning that cost share (in total revenue) of an input differs from its elasticity. For example, it can be seen from Eq. (4) that if a producer is risk averse, input elasticity exceeds the cost share for a risk augmenting input.

In farmed salmon production, risk plays an important part. Consequently, it is important to know which input(s) is (are) risk increasing (decreasing). For this we estimate the partial derivatives of the production risk, $g(X)$ function. Based on the estimates of the risk functions we find that labor is, in general, risk reducing. Labor plays a particularly important role in production risk management. Farm workers' main tasks are monitoring of the live fish in the pens, biophysical variables (sea temperature, salinity, oxygen concentration, algae concentrations, etc.), and the condition of the physical production equipment (pens, nets, feeding equipment, anchoring equipment, etc.). Thus, workers' ability to detect and diagnose abnormal fish behavior, detect changes in biophysical variables, and make prognoses on future development are crucial to mitigate adverse production condition and reduce production risk. We found (as expected) feed to increase the level of output risk, ceteris paribus.

In Fig. 2 we report the frequency distribution of elasticities of the risk function $g(x)$ with respect to labor, capital, feed, age, and time. The mean (median) values of these elasticities for labor, capital, feed, age, and time are -0.049 (-0.043), 0.016 (0.011), 0.085 (0.016), -0.001 (-0.001), and 0.002 (0.002), respectively. The risk part of the production technology seems to be quite insensitive to changes in the age (experience) of farmers. Similarly, no significant change in production risk has taken place over time.

Elasticities of the mean output and risk functions for each input are derived from the estimates of the $f(X)$ and the $g(X)$ functions and their partial derivatives. Since we used a multistep procedure in which the $f(X)$ and the $g(X)$ functions and their partial derivatives are estimated in the first

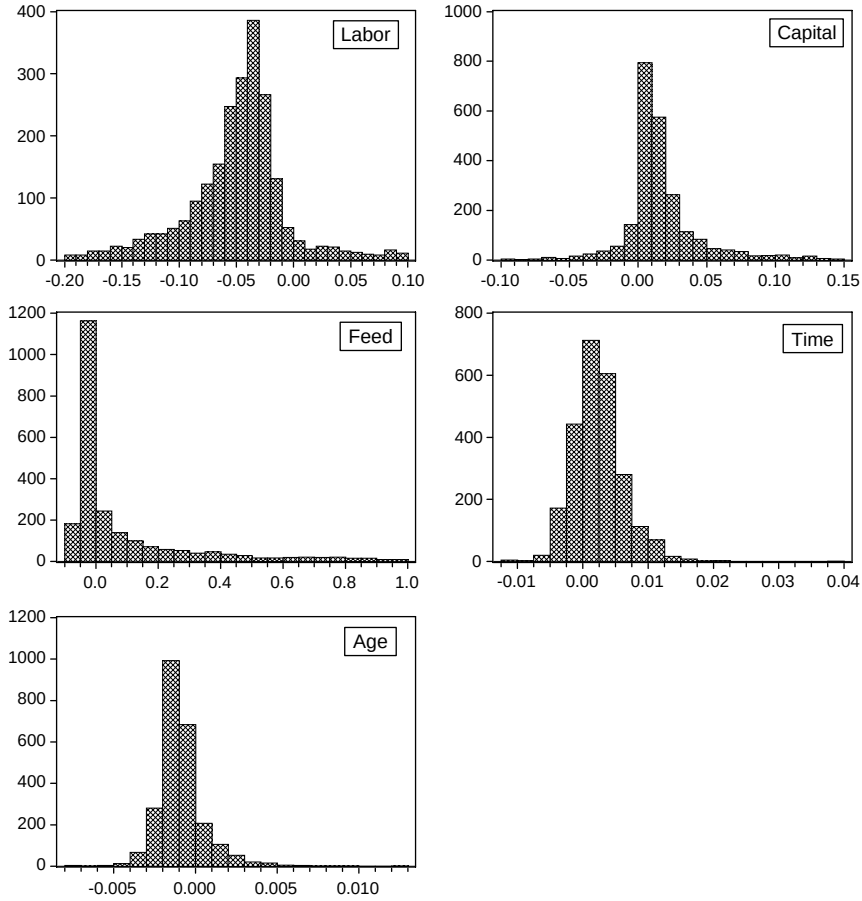


Fig. 2. Histograms of Elasticites of $g(X)$.

step, the estimated elasticities in Models I–III are the same. We use the estimated values of $f(X)$ and $g(X)$ and their partial derivatives to obtain estimates of the risk preference functions $\theta_2(\cdot)$ and $\theta_1(\cdot)$, and estimates of RP in the second step. The estimated values of $\theta_2(\cdot)$, $\theta_1(\cdot)$ (reported in Fig. 3), and RP depend on type of risk an individual farm faces. Two farms with different values of $\theta_2(\cdot)$ and $\theta_1(\cdot)$ are not directly comparable, unless both $\theta_2(\cdot)$ and $\theta_1(\cdot)$ for one farm is higher (lower) than the other. On the other hand, the RP measures among models with different sources of uncertainty and different values of $\theta_2(\cdot)$ and $\theta_1(\cdot)$ are directly comparable.

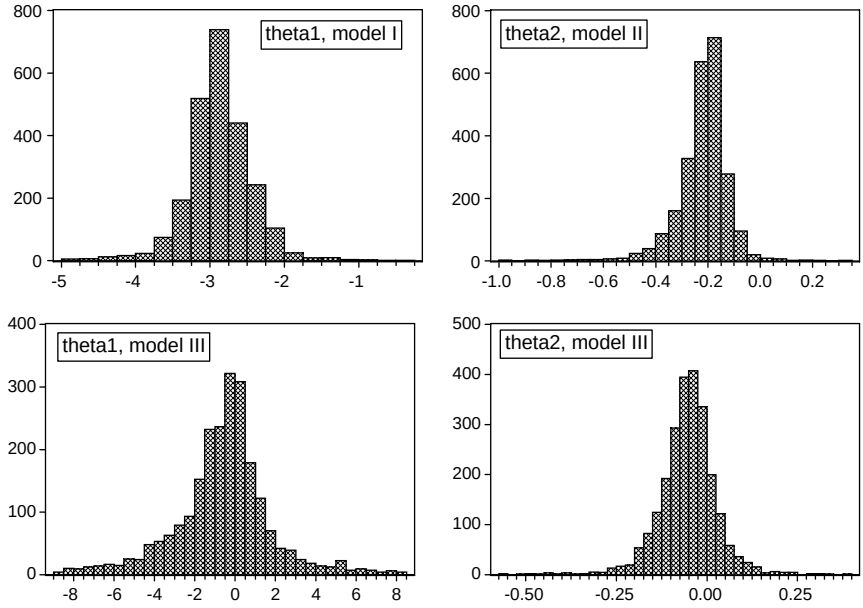


Fig. 3. Histograms of Risk Functions (θ) from Models I–III.

Since RP gives a direct and more readily interpretable result, reporting of RP is often preferred. Given that the RP measure is dependent on units of measurement, a relative measure of RP (defined as $\text{RRP} = \text{RP}/\mu_\pi$) is often reported. Relative risk premium (RRP) is independent of the units of measurement. RRP also takes farm heterogeneity into account by expressing RP in percentage terms.

The frequency distributions of RRP for Models I–III are reported in Fig. 4. These are all skewed to the right. Predicted values of RRP from Model III are much smaller for most of the farms. The mean (median) values of RRP associated with Models I–III are: 0.252 (0.224), 0.171 (0.145), and 0.087 (0.052), respectively. RP shows how much a risk averse farm is willing to pay to insure against uncertain profit due to production risk and/or output price uncertainty. The RRP, on the other hand, shows what percentage of mean profit a risk averse farm is willing to pay as insurance. The above results show that on average a farm is willing to pay 5.22% of the mean profit as an insurance against possible loss of profit due to both production risk and output price uncertainty (Model III).

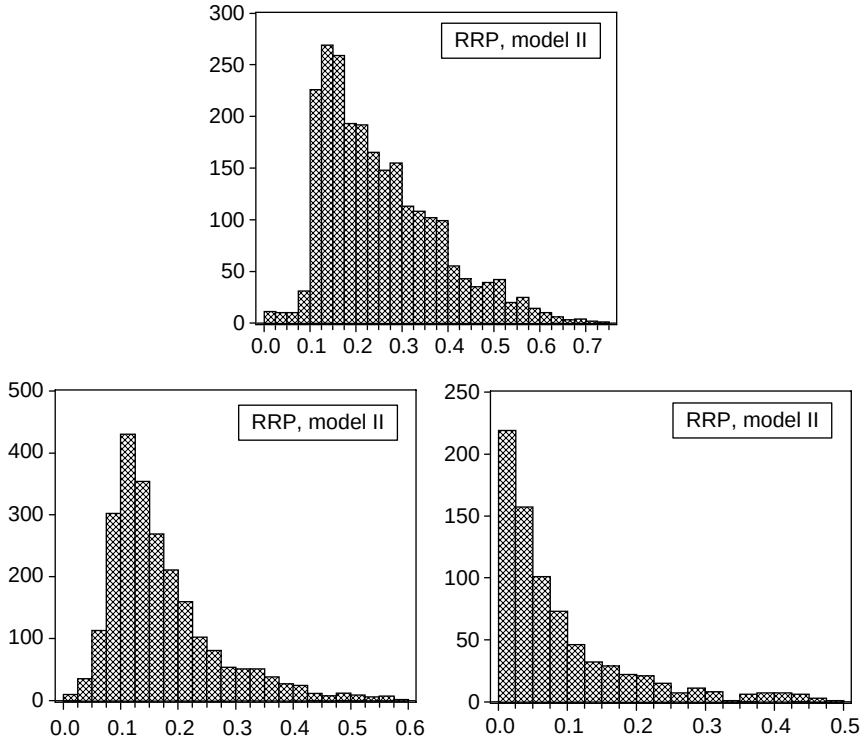


Fig. 4. Histograms of Relative Risk Premium from Models I–III.

Numerical values for the means and standard deviations of elasticities, θs, and RRPs are reported in Tables 1 and 2. In Table 2, also reported are 95% confidence intervals for θs and RRPs. These confidence intervals are somewhat wide, indicating the presence of considerable heterogeneity among salmon farmers regarding their attitudes toward risk.

In addition to reporting the standard errors in Tables 1 and 2, we also report confidence intervals of the elasticities (in terms of both the mean production and risk ($f(\cdot)$ and $g(\cdot)$) functions in Figs. 5 and 6. These figures plot the elasticities against the labor, capital, feed, and time associated with the $f(\cdot)$ and $g(\cdot)$ functions. It can be seen that the confidence intervals of these elasticities are quite wide, and the width does not change with larger values of labor, capital, feed, and time. Elasticities of mean output $f(X)$ with respect to labor, capital, and feed (in Fig. 4) tend to decline with an increase in these inputs. This is consistent with economic theoretic arguments.

Table 1. Elasticities of the Mean Production and Production Risk Functions.

	Mean	Median	Std. Deviation
$f(x)$ w.r.t.			
Labor	0.029	0.017	0.078
Capital	0.017	0.007	0.046
Feed	0.253	0.158	0.282
Time	0.046	0.053	0.026
Age	−0.002	−0.003	0.0036
$g(x)$ w.r.t.			
Labor	−0.0493	−0.0427	0.044
Capital	0.0163	0.0109	0.028
Feed	0.0851	0.0159	0.216
Time	0.0024	0.0021	0.0038
Age	−0.0009	−0.0011	0.0014

Table 2. Risk Preference Functions and Relative Risk Premium.

	Mean	Median	Std. Deviation	95% Confidence Interval	
Model I					
θ_1	−2.869	−2.888	0.435	−3.970	−2.810
RRP	0.252	0.224	0.124	0.122	0.592
Model II					
θ_2	−0.219	−0.205	0.097	−0.420	0.080
RRP	0.171	0.145	0.094	0.098	0.410
Model III					
θ_1	−0.577	−0.402	2.389	−5.240	4.150
θ_2	−0.053	−0.050	0.080	−0.231	0.212
RRP	0.087	0.052	0.096	0.0220	0.342

The positive sign with respect to time shows technical progress. It shows that technical change increased over time.

Fig. 6 shows that elasticities of risk $g(X)$ with respect to labor declined with an increase in labor, and thus labor is found to be risk reducing. On the other hand, feed and capital are found to be risk increasing. The last panel of Fig. 5 shows that production risk decreased over time. The confidence interval is quite similar for farms of all sizes (measured by the input levels).

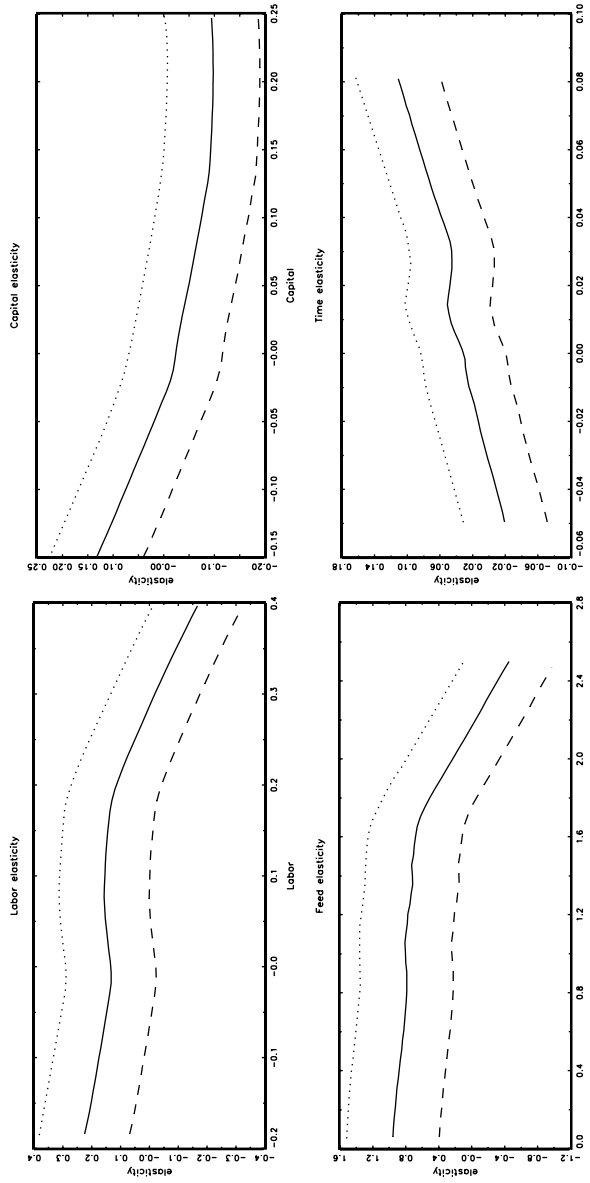


Fig. 5. Elasticites of $f(X)$ and the 95% Confidence Intervals.

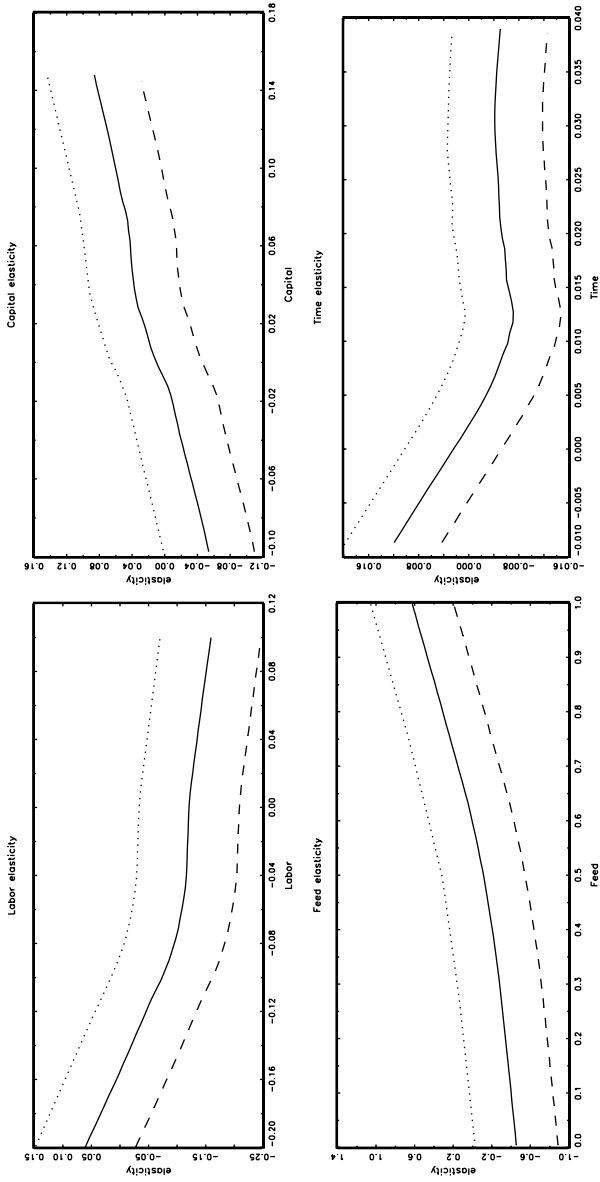


Fig. 6. Elasticites of $g(X)$ and the 95% Confidence Intervals.

Fig. 7 plots θ values for different models against wealth. In all the models we find evidence of an increase in risk averseness with an increase in wealth. The confidence interval is so wide that negative (θ_2) values (risk averseness associated with output price uncertainty) cannot be ruled out. That means almost none of the salmon farmers is risk averse (when it comes to price uncertainty). Finally, in Fig. 8 we plot RRP against wealth for various models. All the models show that RRP increases with wealth almost linearly. That is, these farmers are willing to pay more to protect from risk as their wealth increases.

7. SUMMARY AND CONCLUSIONS

In this paper we addressed modeling issues associated with risk and the risk preference function when producers face uncertainties related to production of output and output price. The modeling approach is based on the assumption that the objective of the producers is to maximize expected utility of normalized anticipated profit. Models are proposed to estimate risk preference of individual producers under (i) only production risk, (ii) only price risk, (iii) both production and price risks, (iv) production risk with technical inefficiency, (v) price risk with technical inefficiency, and (vi) both production and price risks with technical inefficiency. We discussed problems of parametric estimation of these models and discussed nonparametric approaches to some of these models, sometimes partial solutions of the problems (especially in the models with technical inefficiency). Additional theoretical work is necessary to implement some of the more complicated models. Norwegian salmon farming data is used for an empirical application of some of the proposed models. We find that salmon farmers are, in general, risk averse. Labor is found to be risk decreasing while capital and feed are found to be risk increasing.

Both the parametric and nonparametric models are quite challenging because of the complexities/nonlinearites involved in these model. The nonparametric models can relax the rigid functional form assumptions built into the system. However, more research is needed to estimate the nonparametric system models that involve cross-equational restrictions.

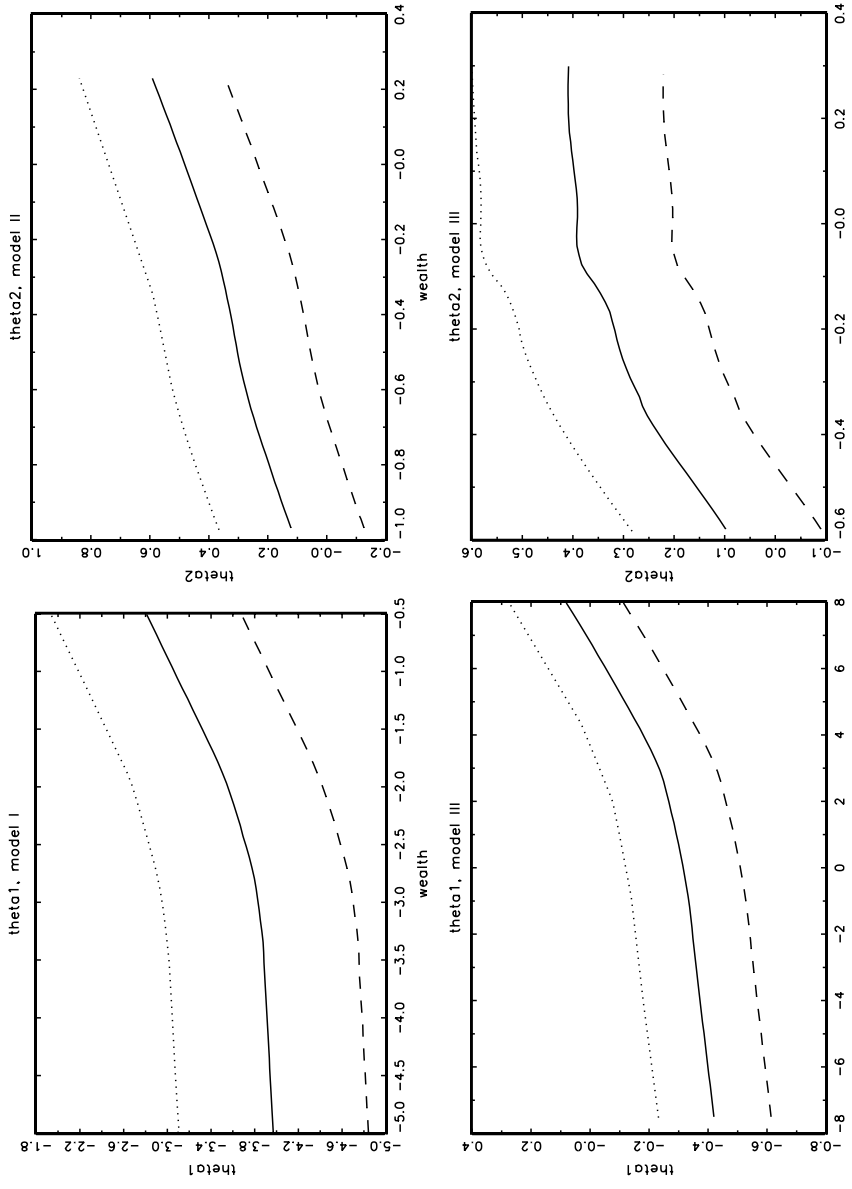


Fig. 7. Plots of $\theta(\cdot)$ and the 95% Confidence Interval against Wealth.

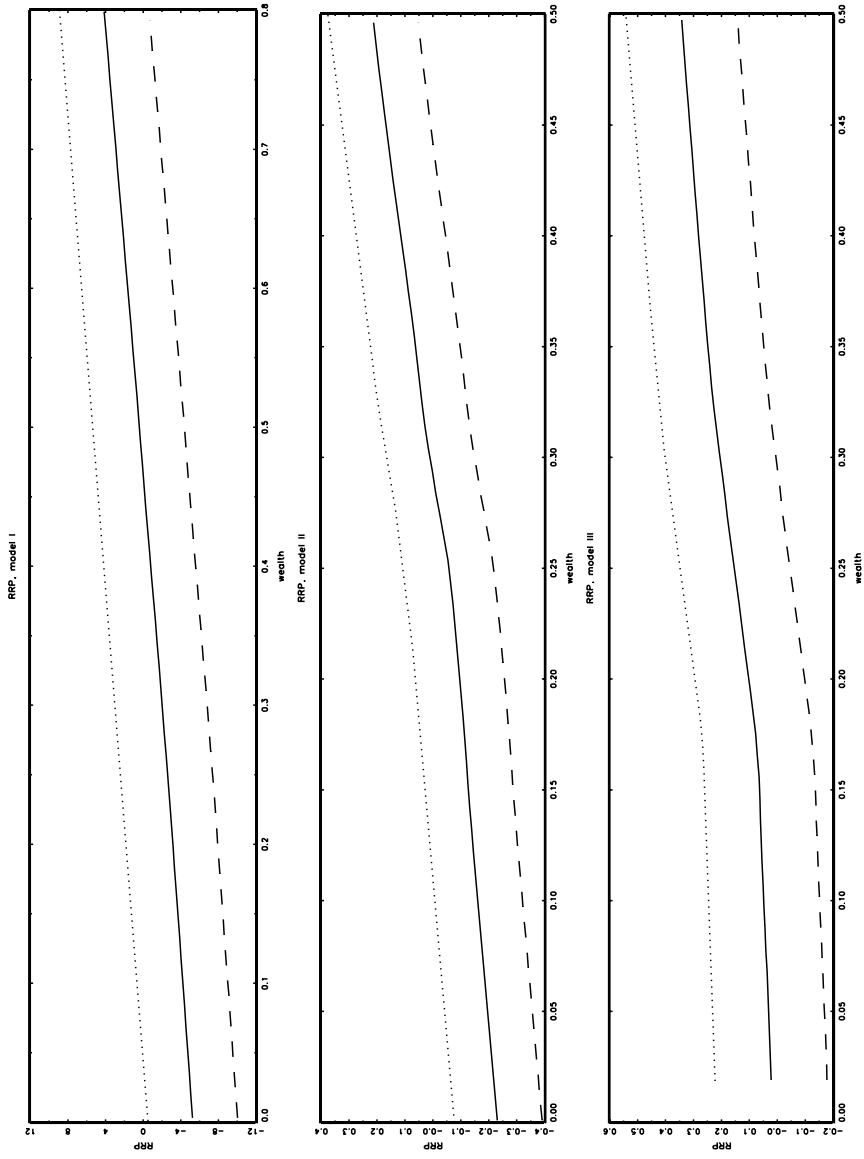


Fig. 8. Plots of RRP and the 95% Confidence Interval against Wealth.

NOTES

1. Since anticipated profit is homogeneous of degree 1 in output and input prices, it is customary to impose the homogeneity condition by normalizing anticipated profit in terms of either the output price (which is done here) or one of the input prices.
2. Note that π^e/p in Eq. (10) has two sources of randomness (η and ε) whereas the source of randomness in π^e in Model I (given in Eq. (2)) is ε. Consequently, the $\tilde{\theta}_1(\cdot)$ and $\tilde{\theta}_2(\cdot)$ functions in Eqs. (13) and (14) are not exactly the same as $\theta_1(\cdot)$ and $\theta_2(\cdot)$ in Eqs. (5) and (9), although we are interpreting them as risk functions associated with output price and production risk, respectively. In general, the $\tilde{\theta}_2(\cdot)$ and $\tilde{\theta}_1(\cdot)$ functions in Eqs. (13) and (14) will depend on the parameters of the distributions of both η and ε.
3. This is, for example, the case in Appelbaum (1991), where constant absolute risk aversion is assumed.
4. See Kumbhakar and Tveteras (2003) for a proof.
5. See Kumbhakar and Tveteras (2003) for details.
6. The proof is similar to Kumbhakar and Tveteras (2003).
7. Note that the production function (15) is more general than the one used by Battese et al. (1997).
8. The Battese et al. (1997) model can be obtained by imposing the restriction $h(X, Z) = g(X, Z)$, which is a testable hypothesis.
9. One anonymous referee suggested that we could use some alternative methods for conditional heteroskedastic models. One promising approach is to follow the procedure in Fan and Yao (1998) (also discussed in Li & Racine, 2007). This procedure has several advantages. It uses local linear estimation, which reduces the boundary bias of the local constant method. It also provides as a "by-product" the derivatives that we are interested in. We would like to pursue this approach in a separate paper.
10. We thank R. Tveteras for providing the data. Details on the sample and construction of the variables used here can be found in Tveteras (1997).
11. These elasticities are positive for most of the data points. There are, however, some farms for which the elasticities are negative, especially for capital. This type of violation of the properties of the underlying production technology (viz., positive marginal product) happens when one uses a flexible parametric production function such as the translog.

ACKNOWLEDGMENTS

We thank two anonymous referees for their helpful comments.

REFERENCES

Appelbaum, E. (1991). Uncertainty and the measurement of productivity. *Journal of Productivity Analysis*, *2*, 157–170.

Appelbaum, E., & Ullah, A. (1997). Estimation of moments and production decisions under uncertainty. *Review of Economics and Statistics*, *79*, 631–637.

Asche, F., & Tveteras, R. (1999). Modeling production risk with a two-step procedure. *Journal of Agricultural and Resource Economics*, *24*, 424–439.

Battese, G., Rambaldi, A., & Wan, G. (1997). A stochastic frontier production function with flexible risk properties. *Journal of Productivity Analysis*, *8*, 269–280.

Bjorndal, T. (1990). *The economics of salmon aquaculture*. London: Blackwell Scientific Publications Ltd.

Chavas, J.-P., & Holt, M. T. (1996). Economic behavior under uncertainty: A joint analysis of risk preferences and technology. *Review of Economics and Statistics*, 329–335.

Chambers, R. G. (1983). Scale and productivity measurement under risk. *American Economic Review*, *73*, 802–805.

Fan, J., & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrica*, *85*, 645–660.

Hardle, W. (1990). *Applied nonparametric regression*. Cambridge, MA: Cambridge University Press.

Just, R. E., & Pope, R. D. (1978). Stochastic specification of production functions and economic implications. *Journal of Econometrics*, *7*, 67–86.

Kumbhakar, S. C. (2002). Risk preference and productivity measurement under output price uncertainty. *Empirical Economics*, *27*, 461–472.

Kumbhakar, S. C., & Lovell, C. A. K. (2000). *Stochastic frontier analysis*. New York: Cambridge University Press.

Kumbhakar, S. C., & Tveteras, R. (2003). Production risk, risk preferences and firm-heterogeneity. *Scandinavian Journal of Economics*, *105*, 275–293.

Li, Q., & Racine, J. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Love, H. A., & Buccola, S. T. (1991). Joint risk preference-technology estimation with a primal system: Reply. *American Journal of Agricultural Economics*, *81*(February), 245–247.

Pagan, A., & Ullah, A. (1999). *Nonparametric econometrics*. Cambridge, MA: Cambridge University Press.

Pratt, J. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*, 122–137.

Rice, J. A. (1984). Boundary modification for kernel regression. *Communications in Statistics, Series A*, *13*, 893–900.

Saha, A., Shumway, C. R., & Talpaz, H. (1994). Joint estimation of risk preference structure and technology using expo-power utility. *American Journal of Agricultural Economics*, *76*, 173–184.

Sandmo, A. (1971). On the theory of competitive firm under price uncertainty. *American Economic Review*, *61*, 65–73.

Stevenson, R. E. (1980). Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics*, *13*(1), 57–66.

Tveteras, R. (1997). *Econometric modelling of production technology under risk: The case of Norwegian salmon aquaculture industry*. Ph.D. dissertation, Norwegian School Economics and Business Administration, Bergen, Norway.

Tveteras, R. (1999). Production risk and productivity growth: Some findings for Norwegian salmon aquaculture. *Journal of Productivity Analysis*, 161–179.

Tveteras, R. (2000). Flexible panel data models for risky production technologies with an application to salmon aquaculture. *Econometric Reviews*, *19*, 367–389.

Wang, H.-J., & Schmidt, P. (2002). One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis*, *18*, 129–144.

Zellner, A., Kmenta, J., & Dreze, J. (1966). Specification and estimation of Cobb-Douglas production function models. *Econometrica*, *34*, 784–795.

APPENDIX. ESTIMATION OF TECHNICAL INEFFICIENCY (MODEL IV)

In this appendix we derive estimators of technical inefficiency and technical efficiency (TE).

$$\text{TE} = \frac{E(Y|u)}{E(Y|u=0)} = \frac{f(X,Z) - g(X,Z)u}{f(X,Z)} = 1 - \frac{g(X,Z)}{f(X,Z)}u = 1 - \text{TI}$$

Production function: We write the production function as

$$y = f(X,Z) + h(X,Z)\varepsilon - g(X,Z)u \equiv f(X,Z) + v - u^A$$

where $v = h(X, Z)\varepsilon$ and $g(X, Z)\ u = u^A$.

Assume that

(i) $v \sim N(0, h^2(X,Z)) = N(0, \sigma_v^2)$,
(ii) $u^A \sim N^+(\mu g(X,Z),\ \sigma_u^2 g^2(X,Z)) = N^+(\mu_0, \sigma_0^2)$.

With these distributional assumptions the model is similar to the normal, truncated normal model proposed by Stevenson (1980). Following Kumbhakar and Lovell (2000, pp. 85–86) we get

$$u^A|\varepsilon^A \sim N^+(\tilde{\mu}, \sigma_*^2), \quad \varepsilon^A = v - u^A$$

$$\tilde{\mu} = \frac{-[\sigma_0^2\varepsilon^A + \mu_0\sigma_v^2]}{\sigma_0^2 + \sigma_v^2}, \quad \sigma_*^2 = \frac{\sigma_0^2\sigma_v^2}{\sigma_0^2 + \sigma_v^2}$$

which gives the following point estimators of inefficiency

$$E[u^A|\varepsilon^A] = \sigma_*\left[\frac{\tilde{\mu}}{\sigma_*} + \frac{\phi(\tilde{\mu}/\sigma_*)}{\Phi(\tilde{\mu}/\sigma_*)}\right]$$

$$M(u^A|\varepsilon^A) = \begin{cases} \tilde{\mu} & \text{if } \tilde{\mu} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where

$$\frac{\tilde{\mu}}{\sigma_*} = -\frac{[\sigma_0^2\varepsilon^A + \mu_0\sigma_v^2]}{[\sigma_0^2 + \sigma_v^2]}\frac{\sqrt{\sigma_0^2 + \sigma_v^2}}{\sigma_0\sigma_v} = -\frac{[\sigma_0^2\varepsilon^A + \mu_0\sigma_v^2]}{\sigma_0\sigma_v\sqrt{\sigma_0^2 + \sigma_v^2}}$$

Note that $\varepsilon^A = Y - f(X,Z)$, $\mu_0 = \mu g(X,Z)$, $\sigma_0^2 = \sigma_*^2 g^2(X,Z)$, and $\sigma_v^2 = h^2(X,Z)$. Estimates of all these functions can be obtained using the estimated parameters. Using the estimated values of u^A, one can obtain estimates of u for each observation from $u^A = g(x,z)u \Rightarrow E[u^A|\varepsilon^A] = g(x,z)\,E[u|\varepsilon^A] \Rightarrow E[u|\varepsilon^A] = E[u^A|\varepsilon^A]/g(x,z)$, and $\widehat{\text{TE}} = 1 - E[u^A|\varepsilon^A]/f(x,z)$.

PART IV
COPULA AND DENSITY ESTIMATION

EXPONENTIAL SERIES ESTIMATION OF EMPIRICAL COPULAS WITH APPLICATION TO FINANCIAL RETURNS

Chinman Chui and Ximing Wu

ABSTRACT

Knowledge of the dependence structure between financial assets is crucial to improve the performance in financial risk management. It is known that the copula completely summarizes the dependence structure among multiple variables. We propose a multivariate exponential series estimator (ESE) to estimate copula densities nonparametrically. The ESE has an appealing information-theoretic interpretation and attains the optimal rate of convergence for nonparametric density estimations in Stone (1982). More importantly, it overcomes the boundary bias of conventional nonparametric copula estimators. Our extensive Monte Carlo studies show the proposed estimator outperforms the kernel and the log-spline estimators in copula estimation. It also demonstrates that two-step density estimation through an ESE copula often outperforms direct estimation of joint densities. Finally, the ESE copula provides superior estimates of tail dependence compared to the empirical tail index coefficient. An empirical examination of the Asian financial markets using the proposed method is provided.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 263–290

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025011

1. INTRODUCTION

The modeling of multivariate distributions from multivariate outcomes is an essential task in economic model building. Two approaches are commonly used. The parametric approach assumes that the data come from a specific family. The maximum-likelihood estimators or the methods of moments are often used to estimate unknown parameters of the assumed parametric distributions. The multivariate normal distribution is a popular choice for multivariate density estimation. More generally, the elliptic distribution family is often used due to its appealing statistical properties. Although they are efficient when the distribution is correctly specified, the parametric estimators are generally inconsistent under misspecification. For example, the elliptic family is often inadequate to capture the pattern of empirical data. This is especially true when we estimate a multivariate asset return distribution or try to account for nonlinear dependence among several assets in financial econometrics (Embrechts, McNeil, & Straumann, 1999).

Alternatively, one can estimate densities using nonparametric methods. Popular nonparametric estimators include the kernel estimator and the series estimator. Because they do not impose functional form assumptions, nonparametric estimators are consistent under mild regularity conditions. However, this robustness against misspecification comes at the price of a slower convergence rate. In other words, nonparametric estimators typically require a larger sample than their appropriately specified parametric counterparts to achieve a comparable degree of accuracy. In addition, nonparametric estimations of multivariate outcomes suffer the "curse of dimensionality," the amount of data needed for the multivariate estimations to obtain a desirable accuracy grows exponentially.

In this study, we focus on a specific strategy of estimating multivariate densities: the copula approach. According to Sklar (1959), the joint density of a continuous multidimensional variable can be expressed uniquely as a product of the marginal densities and a copula function, which is a function of corresponding probability distribution functions of margins. Since the dependence structure among the variables is completely summarized by the copula, it provides an effective device for modeling dependence between random variables. It allows researchers to model each marginal distribution that best fits the sample, and to estimate a copula function with some desirable features separately. In practice, the joint distribution is often estimated with certain functional form restrictions on the specific margins and copula, respectively. For example, the t-distribution can capture the tail heaviness in the margins while the Clayton copula allows asymmetric

dependence. Extensive treatments and discussions on the properties of copulas can be found in Nelsen (2006) and Joe (1997).

There is a growing literature on the estimation of multivariate densities using copulas; see, for example, Sancetta and Satchell (2004), Chen, Fan, and Tsyrennikov (2006), Hall and Neumeyer (2006), Chen and Huang (2007), and Cai, Chen, Fan, and Wang (2008). Two approaches are commonly used. The two-step approach models the marginal distributions and the copula function sequentially, using the estimated marginal distributions as input in the second stage. Alternatively, one can estimate the margins and the copula function simultaneously. The one-step method is generally more efficient, but often computationally burdensome. For either approach, one can use parametric or nonparametric estimators for the margins and/or the copula function. Parametric copulas commonly used in the literature are parameterized by one or two coefficients, which sometimes are inadequate to capture the multivariate dependence structure. On contrary, nonparametric estimators for empirical copula densities are rather flexible, but might suffer boundary bias, especially the popular kernel estimator. The boundary bias problem is particularly severe in the estimation of copula densities, which are defined on the unit hypercube and often do not vanish at the boundaries.

In this paper, we propose to use an alternative nonparametric estimator: the exponential series estimator (ESE) in Wu (2007) for empirical copula density estimation. This estimator is based on the method of maximum entropy density subject to a given set of moment conditions. Compared with other nonparametric estimators, the effective number of nuisance parameters is largely reduced in the context of the ESE for a typical copula that is a smooth function. Furthermore, the ESE is free of boundary bias problem. Our Monte Carlo simulations demonstrate that the ESE provides an overall superior performance than some commonly used nonparametric estimators do in copula density estimations. The two-step density estimation through the ESE copula often outperforms direct estimation of multivariate densities. We also examine the estimation of the tail dependence index, an important risk measure in financial management. Our results suggest that estimations based on the ESE substantially outperform the empirical tail dependence index, especially for extreme tails and small samples.

The rest of the paper is organized as follows. Section 2 presents a brief review of copula, its estimation, and the tail dependence index, whose estimation is investigated in our simulations. Section 3 presents the ESE and discusses its merits as an empirical copula density estimator. Section 4 reports Monte Carlo simulations on the ESE estimations of copula

densities, multivariate densities, and tail dependence indices. Section 5 provides a financial application of the empirical copula density estimation. The last section concludes.

2. COPULA

In this section, we briefly review the literature on copula, its estimation, and the tail dependence index, which can be calculated from a copula function.

2.1. Background

Copula is introduced by Sklar (1959) and has been recognized as an effective device for modeling dependence among random variables. It allows researchers to model each marginal distribution that best fits the sample, and to estimate a copula function with some desirable features separately. The dependence structure among variables is completely summarized by the copula function.

According to Sklar's theorem (Sklar, 1959), the joint distribution function of a d-dimensional random variable $\boldsymbol{x}$ can be written as,

$$F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d))$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)$, F_i is the marginal distribution for x_i, $i = 1, \ldots, d$, and C: $[0, 1]^d \rightarrow [0,1]$ is the so-called copula function. If the joint distribution function is d-times differentiable, then taking the dth partial derivative with respect to $\boldsymbol{x}$ on both sides yields

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{\partial^d}{\partial x_1 \partial x_2 \cdots \partial x_d} F(\boldsymbol{x}) \\ &= \prod_{i=1}^{d} f_i(x_i) \frac{\partial^d}{\partial u_1 \partial u_2 \cdots \partial u_d} C(F_1(x_1), \ldots, F_d(x_d)) \\ &= \prod_{i=1}^{d} f_i(x_i) c(u_1, \ldots, u_d) \end{aligned} \tag{1}$$

where $f_i(\cdot)$ is the marginal density of x_i and $u_i = F_i(x_i), i = 1, \ldots, d$. In Eq. (1) we note that the multiplicative decomposition of the joint density into two parts. One describes the dependence structure among the random

variables in the copula function, and another describes the marginal behavior of each component.

There exists a unique copula function for a continuous multivariate variable. This copula function completely summarizes the dependence structure among variables. In addition, an appealing property of copula is that it is invariant under increasing transformation of the margins. This property is particularly useful in financial research. For example, the copula function of two asset returns does not change when the returns are transformed into a logarithm scale. In contrast, the commonly used linear correlation is only invariant under linear transformation of the margins.

2.2. Estimation

There is a growing literature on the estimation of multivariate densities using copulas. Both parametric and nonparametric estimators have been considered in the literature. Either method can take a two-step or a one-step approach. In the two-step approach, each margin is estimated first and the estimated marginal CDF's are used to estimate copulas in the second step. The estimated parameters (in the parametric case) are typically inefficient when estimated in two steps. In principle we can also estimate the joint density in one step, in which the margins and the copula are estimated simultaneously. Although the estimated parameters (in parametric case) are efficient in this case, the one-step approach is more computationally burdensome than the two-step approach. In empirical work, we sometimes have prior knowledge on the margins but not on the structure of the dependence structure among them. Consequently, the two-step approach may have an advantage over the one-step approach in terms of model specification, although the estimates may be less efficient.

In practice there is usually little guidance on how to choose the best combination of the margins and the copula in parametric estimations. Therefore, semiparametric and nonparametric estimations have become popular in the literature recently. The main advantage of these estimation methods is to let the data determine the copula function without restrictive functional assumptions. In semiparametric estimations, often a parametric form is specified for the copula but not for the margins. The parameters in the copula function are estimated by the maximum-likelihood estimator. See earlier application in Oakes (1986), Genest and Rivest (1993), Genest, Ghoudi, and Rivest (1995), and more recently in Liebscher (2005) and Chen et al. (2006).

Alternatively, nonparametric estimator does not assume parametric distributions for the margins or the copula function. In this way, nonparametric estimator offers a higher degree of flexibility, since the dependence structure of the copula is not directly observable. It also illustrates an approximate picture helpful to researchers in subsequent parametric estimation of the copula. In addition, the problem of misspecification in the copula can be avoided. The earliest nonparametric estimation of copulas is due to Deheuvels (1979), who estimated the copula density based on the empirical distribution. Estimators using kernel methods have been considered in Gijbels and Mielnicnuk (1990), Fermanian and Scaillet (2003) in a time series framework, and Chen and Huang (2007) with boundary corrections. Recently, Sancetta and Satchell (2004) use the Bernstein polynomials to approximate the Kimeldorf and Sampson copula. Hall and Neumeyer (2006) use wavelet estimators to approximate the copula density. Alternatively, Cai et al. (2008) use a mixture of parametric copulas to estimate unknown copula functions.

The kernel density estimator is one of the mostly popular methods in nonparametric estimations. Li and Racine (2007) provide a comprehensive review of this method. In spite of its popularity, there are several drawbacks in kernel estimation. If one uses a higher order kernel estimator in order to achieve a faster rate of convergence, it can result in negative density estimates. In addition, the support of data is often bounded with high concentration at or close to the boundaries in application. This boundary bias problem is well known in the univariate case, and can be more severe in the case of multivariate bounded support variables; see Muller (1991) and Jones (1993).[1]

The log-spline estimators have also drawn considerable attention in the literature and have been studied extensively by Stone (1990).[2] This estimator has been shown to perform well for density estimations. However, it suffers a saturation problem. If we denote s the order of the spline and the logarithm of the density defined on a bounded support has r square-integrable derivatives, the fastest convergence rate is achieved only if $s>r$. Like the kernel estimator, the log-spline estimator also faces a boundary bias problem. It is known that boundary bias exists if the tail has a nonvanishing kth order derivative, while the order the (local) polynomial at the tail is smaller than k. For example, suppose that the tails of a copula density can be represented as a K degree polynomial, where the coefficient for the Kth degree term is nonzero. If the order of the log-spline estimator is smaller than K, then the tails cannot be estimated consistently.

2.3. Tail Dependence Coefficient (TDC)

The copula facilitates study on the dependence structure among multiple variables. There are various measures of dependence. For example, the correlation is commonly used to capture *linear* dependence between two variables. However, it is known that two variables can be dependent while having a zero correlation. Moreover, the correlation is not invariant to nonlinear transformation of variables. A popular *nonlinear* dependence measure is Kendall's τ, which is invariant to increasing transformation of variables. Starting with two independent realizations (X_1, Y_1) and (X_2, Y_2) of the same pair of random variables X and Y, Kendall's τ gives the difference between the probability of concordance and the probability of discordance:

$$\tau(X, Y) = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

for $\tau \in [-1, 1]$. As is discussed above, the dependence structure between two variables can be completely summarized by their copula. In fact, Kendall's τ can be expressed as a function of the copula:

$$\tau(C) = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1$$

Although Kendall's τ offers some advantages over the correlation coefficient, it only captures certain features of the dependence structure. In financial industry, risk managers are often interested in the dependence between various asset returns of the extreme events (during the bear markets or market crashes). A useful dependence measure defined by copulas is the tail dependence. In the bivariate case, the tail dependence measures the dependence existing in the upper quadrant tail, or in the lower quadrant tail. By definition, the upper and lower TDCs are, respectively,

$$\lambda_U = \lim_{u \to 1} \Pr[X > F_X^{-1}(u) | Y > F_Y^{-1}(u)] = \lim_{u \to 1} \frac{1 - 2u + C(u, u)}{1 - u} \tag{2}$$

$$\lambda_L = \lim_{u \to 0} \Pr[X > F_X^{-1}(u) | Y > F_Y^{-1}(u)] = \lim_{u \to 0} \frac{C(u, u)}{u} \tag{3}$$

provided that these limits exist and λ_U and $\lambda_L \in [0, 1]$. The upper (lower) TDC quantifies the probability to observe a large (small) X, given that Y is large (small). In other words, suppose, Y is very large (small) (at the upper quantile of the distribution), the probability that X is very large (small)

at the same quantile defines the TDC $\lambda_U(\lambda_L)$. If $\lambda_U(\lambda_L)$ are positive, two random variables exhibit upper (lower) tail dependence. Eqs. (2) and (3) suggest that the TDC can be derived directly from the copula density. Furthermore, the tail dependence between X and Y is also invariant under strictly increasing transformation of X and Y.

A more useful interpretation of this concept in finance may be obtained if we rewrite the definition of λ_U as,

$$\lambda_U = \lim_{u \to 0^+} \Pr[X > \mathrm{VaR}_u(X) | Y > \mathrm{VaR}_u(Y)]$$

where $\mathrm{VaR}_u(X) = F_X^{-1}(1-u)$ is the Value at Risk (VaR). This notation implies that we have previously multiplied the return by -1. We treat the losses as positive values. Thus, λ_U captures the dependence related to stress periods. Many important applications of the TDC in finance and insurance concern the dependence modeling between extreme insurance claims and large default events in credit portfolios, and VaR considerations of asset portfolios.

3. EXPONENTIAL SERIES ESTIMATION

In this section, we present an alternative nonparametric density estimator based on the ESE in Wu (2007). We first briefly review the maximum entropy density, based on which the ESE is derived. We then discuss some features of the ESE that are particularly suitable for the estimation of empirical copula densities.

3.1. Maximum Entropy Density

Shannon's information entropy is a central concept of information theory. Given a density function f, its entropy is defined as,

$$W(f) = -\int f(x) \log f(x) dx \tag{4}$$

where W measures the randomness or uncertainty of a distribution. Suppose one is to infer a density from a given set of moments, the maximum entropy principle suggests choosing the density that maximizes Shannon's information entropy among all distributions that satisfy given moment conditions. Denote $f(x; \boldsymbol{\theta})$ the maximum entropy density function that maximizes

Eq. (4) subject to the following moment conditions:

$$\int f(x)dx = 1$$

$$\int \phi_k(x) f(x)dx = \mu_k, \quad k = 1, 2, \ldots, m$$

where μ_k is estimated by $\hat{\mu}_k = (1/N)\sum_{i=1}^{N} \phi_k(x_i) \xrightarrow{p} \mu_k$ for an i.i.d. sample $\{x_i\}_{i=1}^{N}$ and $\phi_k, k = 1, \ldots, m$, is a sequence of linearly independent functions. The first moment condition ensures that $f(x)$ is a proper density function. The resulting maximum entropy density takes the form

$$f(x; \boldsymbol{\theta}) = \exp\left(-\theta_0 - \sum_{k=1}^{m} \theta_k \phi_k(x)\right)$$

where $\boldsymbol{\theta}$ is the vector of Lagrange multipliers associated with given moment conditions. To ensure $f(x, \boldsymbol{\theta})$ a is proper density function, we set

$$\theta_0 = \log\left(\int \exp\left(-\sum_{k=1}^{m} \theta_k \phi(x)\right) dx\right)$$

Therefore,

$$f(x; \boldsymbol{\theta}) = \frac{\exp(-\sum_{k=1}^{m} \theta_k \phi(x))}{\int \exp(-\sum_{k=1}^{m} \theta_k \phi(x)) dx}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]$.

In general, analytical solutions for $\boldsymbol{\theta}$ cannot be obtained and nonlinear optimization is employed (see, Zellner & Highfield, 1988; Wu, 2003). To solve for $\boldsymbol{\theta}$, we use Newton's method to iteratively update $\boldsymbol{\theta}$ according to the following equation:

$$\boldsymbol{\theta}_{(t+1)} = \boldsymbol{\theta}_{(t)} - \boldsymbol{H}^{-1}\boldsymbol{b}$$

where $\boldsymbol{b} = [b_1, \ldots, b_m]$, $b_k = \int \phi_k(x) f(x; \boldsymbol{\theta}_t)dx - \mu_k$ and the Hessian matrix $\boldsymbol{H}$ takes the form

$$H_{ij} = \int \phi_i(x)\phi_j(x) f(x; \boldsymbol{\theta}_{(t)})dx$$

The maximum entropy problem and maximum-likelihood approach for exponential families can be considered as a duality problem (Golan, Judge, & Miller, 1996). The maximized entropy W is equivalent to the

sample average of the maximized negative log-likelihood function. This implies the estimated parameters $\hat{\boldsymbol{\theta}}$ are asymptotically normal and efficient.

The maximum entropy density is an effective method of density construction from a limited amount of information (moment conditions). Alternatively, one can use it as a nonparametric density estimator if the number of moment conditions is allowed to increase with the sample size at a suitable rate. We call this estimator the ESE, to distinguish it from the maximum entropy density, where the number of moment conditions is typically small and fixed. Barron and Sheu (1991) study the asymptotic properties of the ESE for a random variable x defined on a bounded support. A key concept used in their work is the relative entropy, or Kullback–Leibler distance. Given two densities f and g with a common support, the relative entropy is defined as:

$$D(f||g) \int f(x) \log \frac{f(x)}{g(x)} dx$$

The relative entropy measures the closeness, or the probability discrepancy, between two densities. Barron and Sheu (1991) show that if the logarithm of the density has r square-integrable derivatives, that is, $\int |D^r \log f(x)|^2 < \infty$, then the sequences of ESE density estimators $\hat{f}(x)$ converge to $f(x)$ in the sense of Kullback–Leibler distance $\int f \log(f/\hat{f}) dx$ at rate $O_p((1/m^{2r}) + (m/N))$ if $m \to \infty$ and $(m^3/N) \to 0$ as $N \to \infty$ where m is the degree of polynomial and N is the sample size. If $m = N^{1/(2r+1)}$, the optimal convergence rate becomes $O_p(N^{-2r/(2r+1)})$. Wu (2007) generalizes the results of Barron and Sheu to d-dimensional random variables and shows that under similar regularity conditions, the optimal convergence rate is $O_p(N^{-2r/(2r+d)})$ if we set $m = N^{1/(2r+d)}$. He further establishes the almost sure uniform convergence rate of the proposed estimator.

3.2. ESE for Copula Density

In this paper, we propose to use the multivariate ESE in Wu (2007) to estimate copula densities. In the context of entropic estimation, the ESE empirical copula can be understood as a minimum relative entropy density with a uniform reference density. Hence, it is most conservative in the sense that the estimated copula is as smooth as possible, as measured by the entropy, given the moment conditions.[3]

To ease exposition, we focus on bivariate case in this study. Generalization to higher dimensional cases is straightforward. As in the univariate

case, we denote $c(u, v; \boldsymbol{\theta})$ to be the copula density function. The objective is to maximize W, the entropy of the copula density

$$W = -\int_{[0,1]^2} c(u, v) \log c(u, v) du dv \tag{5}$$

subject to

$$\begin{aligned} &\int_{[0,1]^2} c(u, v) du dv = 1, \\ &\int_{[0,1]^2} \phi_{ij}(u, v) c(u, v) du dv = \mu_{ij} \end{aligned} \tag{6}$$

where $i = 0, \ldots, n$, $j = 0, \ldots, m$, $i + j > 0$, and $\phi_{ij}(u, v)$ are a sequence of linearly independent polynomials.[4] Given an i.i.d. sample $\{u_t, v_t\}_{t=1}^N$, the empirical moments are calculated as $\hat{\mu}_{ij} = (1/N)\sum_{t=1}^N \phi_{ij}(u_t, v_t) \xrightarrow{p} \mu_{ij}$, where $i+j>0$. The resulting copula density takes the form

$$c(u, v; \boldsymbol{\theta}) = \exp\left\{-\theta_0 - \sum_{i=0}^{n}\sum_{j=0}^{n} \theta_{ij}\phi_{ij}(u, v)\right\}, \quad i + j > 0$$

To ensure $c(u, v; \boldsymbol{\theta})$ is a proper density function, we set

$$\theta_0 = \log\left\{\int_{[0,1]^2} \exp\left(-\sum_{i=0}^{n}\sum_{j=0}^{m} \theta_{ij}\phi_{ij}(u, v) du dv\right)\right\}, \quad i + j > 0$$

Therefore,

$$c(u, v; \boldsymbol{\theta}) = \frac{\exp\{-\sum_{i+j>0, i\le n, j\le m} \theta_{ij}\phi_{ij}(u, v)\}}{\int_{[0,1]^2} \exp\{-\sum_{i+j>0, i\le n, j\le m} \theta_{ij}\phi_{ij}(u, v) du dv\}}$$

where $\{\theta_{ij}\}_{i+j>0, i\le n, j\le m}$. As in the univariate case, we solve for $\boldsymbol{\theta}$ using Newton's method.

In practice, one needs to specify the order of polynomial n and m for the ESE. The selection of the order, which is essentially the "bandwidth" of the nonparametric ESE, is crucial to the performance of the proposed estimator. In practice, the order can be chosen automatically based on the data. Given the close relation between the ESE and the MLE, the likelihood-based AIC and BIC are two natural candidates. Haughton (1988) shows that for a finite number of exponential families, the BIC chooses the correct family with probability tending to 1. On contrary, Shibata (1981) indicates that the AIC leads to an optimal convergence rate for infinite

dimensional models. Wu (2007) reports that these two criteria provide similarly good performance in the selection of the degree of polynomials for small and moderate sample sizes.

It is known that both the AIC and the BIC are derived under the implicit assumption that the estimated parameters of the models in question are asymptotically normal. This condition is typically satisfied by parametric models with a fixed number of parameters under mild conditions. However, this is not necessarily true for nonparametric estimations. Portnoy (1988) examines the behavior of the MLE of the exponential family when the number of parameters, K, tends to infinity. He shows that the condition to warrant the asymptotic normality of estimated parameters is that $K^2/N \to 0$ when $N \to \infty$. Under Assumption 3 of Wu (2007), $K^3/N \to 0$ when $N \to \infty$, which satisfies Portnoy's condition. This result confirms the validity of using the AIC and the BIC for model selection for the proposed nonparametric estimator.

We conclude this section by noting several appealing features of the ESE for copula estimation. First, the effective number of estimated parameters is often substantially smaller compared to the kernel or the log-spline estimators for a given sample size. Hence the ESE enjoys good small sample performance, which is confirmed by our Monte Carlo simulations in the next section. Second, it is known that the ESE may not be well defined when the underlying variable is defined on an unbounded support. Since the copula is defined on the hypercube $[0, 1]^d$, the ESE copula estimator is always well defined. In addition, this bounded support of the copula also frees the ESE from potential outlier problem often associated with higher order polynomials. Lastly, the most important advantage of the ESE is that it does not suffer the boundary bias problem. This is particularly important for copula estimation where the mass of the density is at tails. This boundary bias problem is quite severe for the kernel estimator, and to a lesser extent, for the log-spline estimator. As demonstrated in our simulations below, the more substantial the tails are, the better the ESE performs compared to other estimators.

4. MONTE CARLO SIMULATIONS

To investigate the finite sample performance of the proposed ESE copula estimator, we conduct an extensive Monte Carlo simulation study on estimating copula densities and joint densities of bivariate random variables. We also compare the performance of the ESE with empirical estimator on TDCs (lower or upper).

We consider a variety of margins and copulas in our simulations. For margin distributions, we consider the normal, the Student's t-distribution and two normal mixtures as studied in Marron and Wand (1992). The normal density is often used as a benchmark, while the t distribution is commonly used in financial econometrics since distributions of financial returns are usually fat tailed. The two normal mixtures considered in this study are "skewed unimodal" and "bimodal" distributions as characterized by Marron and Wand (1992). For simplicity, we assume two margins follow a same distribution.

The bivariate copulas used in this study include the Gaussian copula, the t-copula, the Frank copula, and the Clayton copula. Each copula is able to capture a certain dependence structure. In our experiment, the dependence parameter for each type of copula is set such that their corresponding Kendall's τ values 0.2, 0.4, and 0.6. A larger Kendall's τ indicates a higher degree of association between two margins. Fig. 1 displays the contours of various copulas considered in our simulations, with Kendall's $\tau = 0.6$. Note that all these copulas exhibit nonvanishing densities in either or both tails, which may cause severe boundary bias problems for a general nonparametric estimator (Bouezmarni & Rombouts, 2007).

We conduct three sets of simulations in this study. We first examine the performance of copula density estimation of various nonparametric estimators. We then investigate two different approaches of joint density estimation: direct estimation of the joint density and the two-step estimation via the copula. Lastly, we compare the tail index coefficient estimates based on the ESE copula to the empirical tail index coefficient. In all experiments, the order of exponential polynomial of the ESE's is chosen by the BIC. The kernel estimator uses the product Gaussian kernel with individual bandwidth of either dimension selected according to the least squares cross-validation. The log-spline estimator uses the cubic spline with the smoothing parameter chosen by the method of modified cross-validation and the number of knots is determined using the rule $\max(30, 10N(2/9))$, where N is the sample size (see Gu & Wang, 2003 for details). Each experiment is repeated 500 times.

4.1. Estimation of Copula Densities

Our first example concerns the estimation of the copula. For simplicity, we assume that the marginal distributions are known. We consider three sample sizes: 50, 100, and 500. Table 1 reports the average mean integrated squared

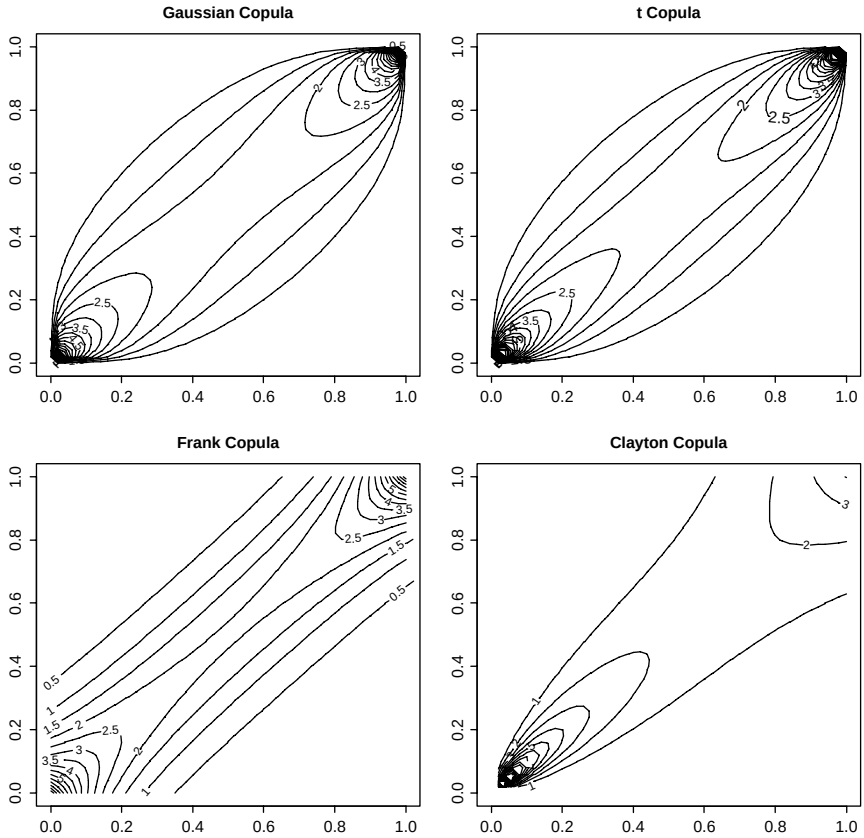


Fig. 1. Contour Plots of Parametric Copulas with Dependence Parameters Corresponding to Kendall's τ being 0.6.

errors (MISE) and their standard deviations of the three estimators for our experiments.

For all estimators, the performance improves with the sample size but reduces with the value of Kendall's τ. Intuitively, the larger is Kendall's τ, the higher is the dependency between the margins. Thus the copula is increasingly concentrated near the two tails along the diagonal, and the shape of the copula become more acute near the tails in Fig. 1. This makes the boundary bias problem more severe. We also note that the MISE decreases with sample size, but the decreasing rate is slower for a larger τ. For example, the MISE in the case of Gaussian copula decreases by 60%

Table 1. MISE of Copula Density Estimation.

Kendall's τ	Copula	n	ESE		Log-spline		Kernel	
0.2	Gaussian	50	0.164	(0.0071)	0.170	(0.0112)	0.233	(0.0135)
		100	0.107	(0.0014)	0.120	(0.0041)	0.171	(0.0024)
		500	0.065	(0.0001)	0.076	(0.0002)	0.099	(0.0002)
	t	50	0.210	(0.0194)	0.244	(0.0420)	0.270	(0.0165)
		100	0.139	(0.0014)	0.160	(0.0051)	0.201	(0.0032)
		500	0.098	(0.0001)	0.104	(0.0005)	0.127	(0.0006)
	Frank	50	0.172	(0.0117)	0.168	(0.0161)	0.221	(0.0114)
		100	0.103	(0.0020)	0.105	(0.0019)	0.170	(0.0046)
		500	0.058	(0.0001)	0.069	(0.0001)	0.104	(0.0001)
	Clayton	50	0.217	(0.0061)	0.236	(0.0194)	0.274	(0.0115)
		100	0.160	(0.0025)	0.188	(0.0189)	0.214	(0.0052)
		500	0.113	(0.0001)	0.118	(0.0008)	0.146	(0.0095)
0.4	Gaussian	50	0.240	(0.0080)	0.350	(0.0397)	0.381	(0.0223)
		100	0.180	(0.0024)	0.270	(0.0150)	0.293	(0.0070)
		500	0.131	(0.0001)	0.174	(0.0027)	0.199	(0.0059)
	t	50	0.345	(0.0139)	0.455	(0.0829)	0.480	(0.0415)
		100	0.263	(0.0024)	0.349	(0.0200)	0.371	(0.0059)
		500	0.217	(0.0005)	0.251	(0.0047)	0.288	(0.0007)
	Frank	50	0.215	(0.0164)	0.292	(0.0267)	0.329	(0.0187)
		100	0.142	(0.0031)	0.215	(0.0101)	0.235	(0.0048)
		500	0.090	(0.0002)	0.125	(0.0009)	0.157	(0.0009)
	Clayton	50	0.574	(0.0190)	0.664	(0.1138)	0.683	(0.0369)
		100	0.498	(0.0039)	0.555	(0.0669)	0.577	(0.0509)
		500	0.364	(0.0012)	0.410	(0.0130)	0.453	(0.0154)
0.6	Gaussian	50	0.484	(0.0155)	0.780	(0.0489)	0.805	(0.0545)
		100	0.401	(0.0047)	0.654	(0.0087)	0.645	(0.0166)
		500	0.328	(0.0014)	0.518	(0.0012)	0.532	(0.0005)
	t	50	0.721	(0.0315)	1.014	(0.0897)	1.072	(0.0812)
		100	0.629	(0.0087)	0.859	(0.0230)	0.865	(0.0209)
		500	0.451	(0.0039)	0.692	(0.0087)	0.721	(0.0070)
	Frank	50	0.302	(0.0202)	0.551	(0.0547)	0.585	(0.0648)
		100	0.212	(0.0051)	0.400	(0.0084)	0.408	(0.0123)
		500	0.142	(0.0003)	0.237	(0.0034)	0.281	(0.0004)
	Clayton	50	1.632	(0.0331)	1.917	(0.0872)	2.004	(0.0955)
		100	1.493	(0.0169)	1.695	(0.0543)	1.750	(0.0619)
		500	1.105	(0.0043)	1.409	(0.0040)	1.619	(0.0075)

Note: Standard deviations are in parentheses.

from $N = 50$ to 500 when $\tau = 0.2$; while its MISE decreases by 32% when $\tau = 0.6$. Among the copulas considered in the simulations, Gaussian and Frank copulas have smaller MISE values. For small τ, the ESE shows slightly better performance than the log-spline estimator does. As Kendall's

τ increases, the ESE outperforms the log-spline estimator more significantly. At the same time, the ESE and the log-spline estimator outperform the kernel estimator in almost all the cases. The better performance of the ESE can be explained by the fact that the kernel estimator allocates weight outside the boundary and underestimates the underlying copula density at the tails.[5] We also note that the log-spline estimator and the kernel estimator have substantially larger standard deviations in the MISE than does the ESE. Overall, the ESE outperforms the other two estimators considerably in our experiments.

4.2. Joint Density Estimation

We next compare the direct estimation of joint densities, without estimating a copula function, to that via the two-step copula method. We note that for two-step estimation, the convergence rate of the joint density is determined by the slower of two rates: convergence rate of the margins and that for the copula. When both are estimated nonparametrically with optimal smoothing parameters, since the later is asymptotically slower than the former (due to the curse of dimensionality), the convergence rate of the joint density estimation is of the same order as that of the copula density. This result implies that asymptotically, the performance of the joint density estimation is not affected by optimal estimation of the marginal densities. In our two-step estimation, we use the log-spline estimator for the margins, due to its good small sample performance for estimation of densities with unbounded supports. The results using the kernel estimator or the ESE for the marginal distributions are quantitatively similar and hence not reported.

Combining four margins and four copulas considered in study, we obtain 16 ($= 4 \times 4$) joint densities. In this experiment, we set the sample size to 50. The estimators we consider in the direct estimation are the ESE, the log-spline estimator and the kernel estimator. In the two-step estimation, we first estimate the margins by the log-spline estimator and then the copula density by the ESE. The MISE of estimated joint densities of various estimators are displayed in Table 2. Similar to the first experiment, the MISE increases with Kendall's τ. Comparing across different copulas, we note that Clayton copula has recorded the largest MISE as τ increases for all the margins. As shown in Fig. 1, Clayton copula has a relatively sharp tails near the boundary, which makes the estimation difficult. Another observation is that the Gaussian and the t margins exhibit smaller MISE in

Table 2. Ratio of MISE of the Direct Joint Density Estimation to the Two-step Copula Estimation.

Kendall's τ	Margin	Estimation Method	Copula			
			Gaussian	*t*	Frank	Clayton
0.2	Gaussian	Two-step Copula	0.464	0.518	0.544	0.635
		ESE	83.8%	77.8%	66.0%	169.3%
		Log-spline	119.8%	140.2%	110.8%	226.3%
		Kernel	261.0%	219.3%	212.5%	326.0%
	Skewed unimodal	Two-step Copula	0.513	0.468	0.481	0.483
		ESE	206.4%	218.2%	240.5%	258.2%
		Log-spline	173.7%	214.3%	209.8%	307.7%
		Kernel	310.5%	387.8%	356.1%	500.0%
	Bimodal	Two-step Copula	0.435	0.473	0.457	1.153
		ESE	323.9%	300.4%	290.4%	226.4%
		Log-spline	195.4%	195.6%	190.6%	104.4%
		Kernel	249.0%	257.5%	240.5%	209.8%
	t	Two-step Copula	0.268	0.244	0.245	0.278
		ESE	146.3%	175.8%	158.0%	139.6%
		Log-spline	173.5%	245.5%	234.3%	204.0%
		Kernel	384.3%	528.7%	449.0%	512.2%
0.4	Gaussian	Two-step Copula	0.836	0.784	0.704	1.07
		ESE	49.6%	59.6%	166.8%	143.5%
		Log-spline	98.4%	101.8%	184.2%	162.5%
		Kernel	178.5%	193.5%	216.2%	248.5%
	Skewed unimodal	Two-step Copula	1.028	1.09	1.071	1.213
		ESE	163.1%	134.8%	147.3%	125.3%
		Log-spline	105.6%	107.6%	124.6%	152.0%
		Kernel	177.1%	179.5%	196.8%	236.2%
	Bimodal	Two-step Copula	1.081	1.066	1.034	1.533
		ESE	152.5%	155.3%	163.1%	185.5%
		Log-spline	105.5%	107.8%	105.3%	104.8%
		Kernel	134.0%	134.3%	154.2%	188.7%
	t	Two-step Copula	0.626	0.712	0.571	0.872
		ESE	155.3%	149.4%	215.2%	128.0%
		Log-spline	112.3%	126.4%	151.7%	97.0%
		Kernel	214.4%	198.0%	282.3%	209.6%
0.6	Gaussian	Two-step Copula	1.201	1.302	1.08	2.662
		ESE	46.9%	53.3%	188.8%	78.4%
		Log-spline	112.5%	111.3%	218.4%	103.0%
		Kernel	224.8%	193.5%	283.1%	126.7%
	Skewed unimodal	Two-step Copula	1.786	1.88	1.428	2.205
		ESE	115.1%	121.7%	177.8%	101.4%
		Log-spline	103.5%	111.0%	255.1%	111.7%
		Kernel	157.8%	171.2%	293.1%	168.6%

Table 2. (*Continued*)

Kendall's τ	Margin	Estimation Method	Copula			
			Gaussian	t	Frank	Clayton
	Bimodal	Two-step Copula	1.703	1.901	1.569	3.183
		ESE	137.6%	135.9%	137.7%	112.0%
		Log-spline	91.5%	104.2%	104.4%	104.3%
		Kernel	120.0%	136.3%	142.0%	115.8%
	t	Two-step Copula	0.841	1.08	0.906	1.934
		ESE	180.7%	152.3%	181.5%	122.1%
		Log-spline	76.8%	81.7%	86.7%	95.3%
		Kernel	179.3%	176.1%	195.1%	134.3%

Note: The percentages in this table represent the MISE ratio of joint density estimation to the two-step copula estimation for each margin and copula specification.

all the copulas and τ values. This is expected since relatively simple shapes of the margins tend to reduce the estimation errors.

In the case of direct estimation, the patterns of MISE are similar to two-step copula estimation in terms of the margins and copulas under consideration. Except for the bimodal margin, the kernel estimator is dominated by the ESE and the log-spline estimator. Under the Gaussian margin, the ESE outperforms the log-spline estimator. The MISE under the ESE is 50% of that of the log-spline. In general, as the shapes of the margins become more complicated, the log-spline estimator dominates the ESE.

Further examination on the extent of improvement in the MISE under the two-step copula estimation shows that, except for the Gaussian margins, two-step copula estimator generally outperforms the other three estimators in almost all Kendall's τ and copulas under consideration.[6] Since a regular and consistent pattern cannot be observed for the difference of the MISE between two-step copula and other three estimators, we average the MISE across the margins and calculate percentage of the MISE of two-step copula to the MISE of the other estimators. More than 50% of improvement is found for small τ using two-step copula estimation, although the improvement decreases as the dependence between the variables increases. The improvements are in the order of Clayton > Frank > t > Gaussian in terms of copulas.

4.3. Tail Dependence Coefficient Estimation

In the last experiment, we compare the ESE with empirical estimator for the TDC. For the ESE estimator, the copula density function is estimated first

via the ESE, then the corresponding TDC is derived using the *last* part of TDC definitions in Eqs. (2) and (3). For the empirical estimator, the TDC is calculated using second equality given in Eqs. (2) and (3), where the population distributions are replaced by empirical distributions. One hundred observations are generated from the Frank copula with Kendall's $\tau = 0.2$, 0.4, and 0.6. Each experiment is repeated 500 times. The mean-squared error (MSE) and the variance of these two estimators are reported in Table 3. The larger Kendall's τ is, the larger is the MSE. The MSE decreases as the percentile increases for the upper tail and deceases for the lower tail. For all the τ and percentiles under consideration, the ESE gives a remarkably smaller MSE compared with the empirical estimator.

Table 3. Mean Square Error of the Tail Dependence for the Frank Copula ($n = 100$).

τ		0.6		0.4		0.2	
Percentile		MSE	Variance	MSE	Variance	MSE	Variance
95	Empirical	5.4311	5.4293	3.3131	3.3131	2.4778	2.4714
	ESE	1.0452	0.3560	0.3100	0.2413	0.1249	0.1208
97.5	Empirical	6.7480	6.7426	4.7859	4.7832	2.1218	2.1215
	ESE	0.4522	0.1131	0.1032	0.0770	0.0386	0.0376
99	Empirical	4.9867	4.9839	3.0999	3.0999	1.5585	1.5581
	ESE	0.1013	0.0204	0.0191	0.0132	0.0068	0.0066
99.5	Empirical	2.0725	2.0461	1.1080	1.0988	0.7665	0.7665
	ESE	0.0287	0.0060	0.0047	0.0034	0.0019	0.0019
99.75	Empirical	1.1412	1.1381	1.0308	1.0308	0.0030	0.0000
	ESE	0.0078	0.0013	0.0012	0.0008	0.0004	0.0004
99.9	Empirical	0.0062	0.0000	0.0018	0.0000	0.0005	0.0000
	ESE	0.0013	0.0002	0.0002	0.0001	0.0000	0.0000
5	Empirical	5.3932	5.3917	3.3348	3.3342	2.3995	2.3990
	ESE	1.0197	0.3622	0.2832	0.2375	0.1342	0.1341
2.5	Empirical	6.4563	6.4560	3.9032	3.8992	2.2174	2.2144
	ESE	0.4269	0.1124	0.0955	0.0762	0.0410	0.0410
1	Empirical	5.0478	5.0477	2.2984	2.2903	1.4224	1.4201
	ESE	0.0938	0.0221	0.0189	0.0149	0.0078	0.0077
0.5	Empirical	3.6785	3.6757	1.5357	1.5342	0.0119	0.0000
	ESE	0.0270	0.0058	0.0054	0.0044	0.0019	0.0019
0.025	Empirical	3.6762	3.6332	0.0109	0.0000	0.0030	0.0000
	ESE	0.0075	0.0017	0.0012	0.0008	0.0004	0.0004
0.01	Empirical	0.4810	0.4806	0.0018	0.0000	0.0005	0.0000
	ESE	0.0012	0.0002	0.0002	0.0002	0.0001	0.0001

Note: MSE and variance are multiplied by 100.

The ratios between them increase as τ decreases, except for very large or very small percentiles. The variances of the MSE decrease as τ decreases. The ESE shows a decreasing pattern for variance as the percentile increases for the upper tail and decreases for the lower tail. Finally, the ESE shows smaller variances compared with empirical estimator in general. Overall the ESE for TDC outperforms the empirical estimator in terms of the MSE and its variance.

5. EMPIRICAL APPLICATION

An important question in risk management is whether the financial markets become more interdependent during financial crises. The fact that international equity markets move together more in downturns than in the upturns has been documented in the literatures, for example, see Longin and Solnik (2001) and Forbes and Rigobon (2002). Hence, the concept of tail dependence plays an increasingly important role in measuring the financial contagion. If all stock prices tend to fall together as a crisis occurs, the value of diversification might be overstated by ignoring the increase in downside dependence (Ang & Chen, 2002). During the 1990s, several international financial crises occurred. Asian financial crisis is one of the crises that have been studied extensively in the literature. It started in Thailand with the financial collapse of Thai Baht on July 2, 1997. News of the devaluation dropped the value of the baht by as much as 20% – a record low. As the crisis spread, most of Southeast Asia saw slumping currencies, devalued stock markets and asset prices.

Early studies on the dependence structure between financial assets are mostly based on their correlations, which ignore potential nonlinear dependence structures. Some recent studies use parametric copulas to capture the nonlinear dependence. They derive the corresponding values of tail dependence based on the estimated copulas. The parametric approach may lack flexibility and the estimated dependence will be biased if the copula is misspecified. In this section, we model the dependence structure of the Asian stock markets returns using the ESE copula. No assumptions on the dependence structure in the data are imposed. We emphasize that the results in the section are presented as an illustration of the ESE copula estimation, rather than a detailed study of financial contagion in Asian financial crisis. Following Kim (2005) and Rodriguez (2007), we analyze the dependent structure for the Asian stock index returns by pairing all other countries with Thailand, the originator of the Asian financial crisis.

The data used in this study are daily returns consisting of the daily stock-market indices of six East Asian countries from the DATASTREAM. When these Asian countries experienced the Asian financial crisis in 1997, these data present an interesting case for the study of tail dependence as all these countries in the sample experienced a crisis of some severity during this period. Specifically, the data include Hong Kong Hang Seng (HK), Singapore Strait Times (SG), Malaysia Kuala Lumpur Composite (ML), Philippines Stock Exchange Composite (PH), Taiwan Stock Exchange Weighted (TW), and Thailand Bangkok S.E.T. (TH). The dataset covers the sample period from January 1994 to December 1998. We have altogether 1,305 daily observations. We take the log-difference of each stock index to calculate the stock index returns. Following the standard practice, we fit each return series by a GARCH(1,1) model using the maximum-likelihood estimation. Based on the fitted model, we calculate the implied standardized residuals from the GRACH model. The standardized residuals obtained in the first step estimation will be as the input for the copula density estimation in the second step.[7]

Table 4 gives summary statistics of the data. Standard deviations reveal that the Malaysia market is the most volatile, followed by Thailand. All the series exhibit excessive skewness and kurtosis relative to those of the Gaussian distribution. The Jarque–Bera test demonstrates the nonnormality of each series, which implies the violation of multivariate Gaussian distribution assumption. In fact, it is well known, according to Mandelbort (1963), that most financial time series are fat tailed. Existing studies often replace the assumption of normality with the fat-tailed t distribution. Notice that Malaysia has the highest risk in terms of volatility, skewness, kurtosis, and nonnormality; while Taiwan ranks last in terms of volatility, kurtosis, and nonnormality.

Table 4. Descriptive Statistics for the Stock Indices Returns.

	Hong Kong	Singapore	Malaysia	Philippines	Taiwan	Thailand
Mean	−0.0131	−0.0329	−0.1127	−0.0574	−0.0103	−0.1118
Median	0.0046	−0.0139	−0.0280	0.0000	0.0000	−0.0533
Maximum	17.2702	15.3587	23.2840	13.3087	6.3538	16.3520
Minimum	−14.7132	−10.6621	−37.0310	−11.0001	−11.3456	−15.8925
SD	1.9668	1.6484	2.7410	1.9546	1.5864	2.5656
Skewness	0.2615	0.6875	−1.2933	−0.0172	−0.4601	0.5974
Kurtosis	13.8503	16.1685	42.0452	8.1313	7.2913	9.4746
Jarque–Bera	6411.5	9524.6	83196.3	1430.7	1046.6	2355.2
P-value	<1.00e-04	<1.00e-04	<1.00e-04	<1.00e-04	<1.00e-04	<1.00e-04

Table 5. Correlation Matrix of the Stock Indices Returns.

	Hong Kong	Singapore	Malaysia	Philippines	Taiwan	Thailand
Hong Kong		0.6526	0.3797	0.3816	0.2488	0.3867
Singapore	0.3597		0.4104	0.2581	0.2797	0.5126
Malaysia	0.2816	0.4783		0.3020	0.1993	0.3862
Philippines	0.1954	0.5080	0.2069		0.1999	0.4133
Taiwan	0.1214	0.1312	0.0992	0.0979		0.1886
Thailand	0.2205	0.2900	0.2726	0.2069	0.0669	
Average dependence						
Linear	0.4099	0.4863	0.3491	0.3609	0.2233	0.3775
Kendall	0.2357	0.2899	0.2541	0.1930	0.1033	0.2114

Note: Upper triangle is the linear correlation and the lower triangle is the Kendall's τ.

To investigate the dependence between different markets, we calculate the linear correlation and Kendall's τ between Thailand and other countries. The estimated correlation and Kendall's τ are reported in Table 5. The patterns revealed by these two dependence measures are qualitatively similar. The linear correlations range from 0.19 (Taiwan and Thailand) to 0.65 (Singapore and Hong Kong) among the pairs we consider. Singapore has the highest average dependence with other countries, while Taiwan has the lowest average dependence. Although Thailand is suggested to play a trigger role in the Asian financial crisis, it only shows moderate dependence with other countries. The Kendall's τ ranges from 0.07 (Taiwan and Thailand) to 0.50 (Philippines and Singapore).

Table 6 reports empirical estimates of lower and upper TDCs in the bivariate equity index returns. The first cell in the table is 0.308, which indicates that the probability of returns of Hong Kong being lower than the 5th percentile given that the returns of Thailand is lower than the 5th percentile equals 0.308. While Singapore has the strongest lower dependence, Hong Kong has the strongest upper dependence with Thailand. Philippines and Malaysia show moderate tail dependence with Thailand, and Taiwan has the weakest tail dependence with Thailand. In general, the lower tail dependences are larger than the upper tail dependences. This fact is consistent with the literature that financial markets exhibit asymmetric tail dependence: they tend move together more in downturns than in upturns.

The next step is to estimate the copula density. Frahm, Junker, and Schmidt (2005) show that using misspecified parametric margins instead of nonparametric margin may lead to misleading interpretations of dependence structure. Instead of assuming parametric margins, we estimate the margins

Table 6. Estimated Tail Dependence for Bivariate Standardized Returns.

Percentile	Hong Kong		Singapore		Malaysia		Philippines		Taiwan	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Empirical										
5	0.308	0.242	0.354	0.167	0.262	0.242	0.231	0.182	0.169	0.106
4	0.288	0.283	0.346	0.132	0.269	0.151	0.250	0.170	0.154	0.113
3	0.282	0.225	0.359	0.125	0.179	0.100	0.179	0.150	0.128	0.150
2	0.346	0.185	0.346	0.148	0.077	0.000	0.192	0.074	0.115	0.111
1	0.231	0.143	0.231	0.214	0.000	0.000	0.000	0.000	0.000	0.143
ESE										
5	0.314	0.257	0.296	0.209	0.205	0.198	0.181	0.187	0.143	0.145
4	0.259	0.211	0.237	0.167	0.162	0.156	0.159	0.162	0.121	0.119
3	0.198	0.167	0.168	0.123	0.134	0.129	0.126	0.121	0.097	0.092
2	0.135	0.121	0.119	0.087	0.089	0.081	0.087	0.089	0.078	0.076
1	0.075	0.061	0.062	0.048	0.065	0.057	0.056	0.058	0.046	0.048

Note: Lower and upper represent the estimated lower and upper tail dependence coefficient, respectively.

by the log-spline estimator in the first step and the copula density by the ESE method in the second step. Different dependence structures can be visualized by plotting their estimated copula densities along the diagonal $u = v$. Fig. 2 shows the results. Notice that the scales in the graphs are different. Hong Kong, Singapore, and Malaysia show clearly asymmetric shapes with lower tail higher than upper tail; while Philippines and Taiwan have relatively symmetric shapes. In the case of Hong Kong, most of the mass is concentrated in the two tails, as suggested by the height of the estimated density with a small peak in the center of the density. Singapore and Malaysia also exhibit similar patterns with less mass concentrated on the two tails. On contrary, Philippines and Taiwan show a symmetric tail patterns and their densities are relatively flat compared to the previous three markets.

The lower and upper TDCs are then calculated from estimated ESE copula densities. The results are reported in Table 6. We note that for the ESE, the lower TDC increases and the upper TDC decreases monotonically in all the markets, while nonmonotonic patterns are observed in empirical TDC estimates. Asymmetric tail dependences are observed in Hong Kong, Singapore, and Malaysia, but not in Philippines and Taiwan. Compared with empirical tail dependence estimates, the ESE TDC tends to be smaller.

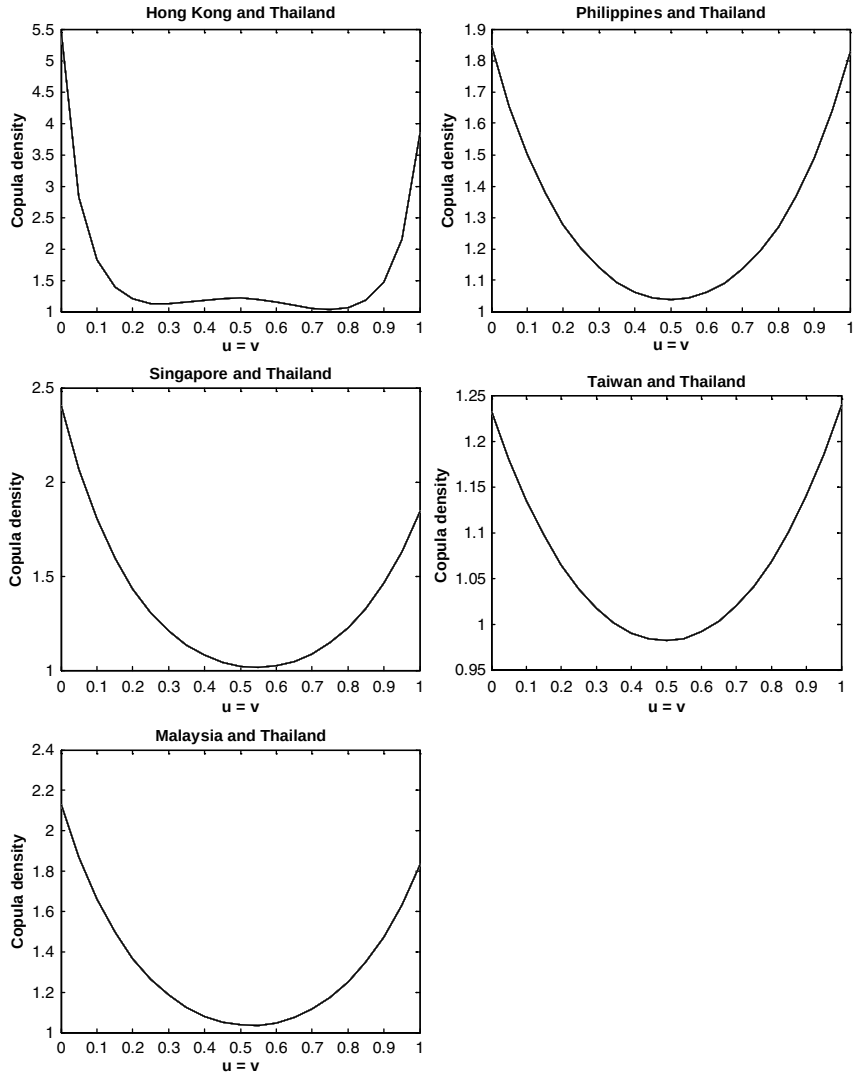


Fig. 2. Plots of Diagonals of Estimated Copula Densities between various Asian Countries and Thailand.

Again according to the ESE estimates, Hong Kong exhibits the strongest lower and upper tail dependence with Thailand and the lower tails dependence is stronger than the upper tail dependence.

Lastly, for the sake of comparison, we report the TDC estimates based on the Gaussian copula. The results are reported in the bottom panel of Table 6. As expected, the estimates are not able to capture the asymmetric dependence between the markets. In addition, the estimates are considerably smaller than the ESE-based and the empirical estimates. The comparison demonstrates that the risk associated with a misspecified parametric copula estimate can be quite substantial.

6. CONCLUSION

This paper proposes a nonparametric estimator for copula densities based on the ESE. The ESE has an appealing information-theoretic interpretation and attains the optimal rate of convergence for nonparametric densities in Stone (1982). More importantly, it overcomes the boundary bias in copula density estimation. We examine finite sample performance of the estimator in several simulations. The results show that the ESE outperforms the popular kernel and log-spline estimators in copula estimation. Estimating a joint density by first estimating the margins and the copula separately in a two-step approach often outperforms direct estimation of the joint density. In addition, the proposed estimator provides superior estimates to the tail dependence index compared to the empirical tail dependence index. We apply the ESE copula to estimate the joint distributions of stock returns of several Asian countries during the Asian financial crisis and examine their interdependence based on the estimated joint densities and copulas.

NOTES

1. Many methods have been proposed to resolve this boundary bias problem of the kernel estimator. These methods either adopt different functional forms of kernel beyond the Gaussian kernel (e.g., see Lejeune & Sarda, 1992; Jones, 1993; Jones & Foster, 1996) or transform data before applying the Gaussian kernel (Marron & Ruppert, 1994). Recent studies included Chen (1999), Bouezmarni and Rombouts (2007). These studies propose to use the gamma kernel or the local linear kernel estimators.

2. A closely related literature is the bivariate log-spline estimator studied by Stone (1994), Koo (1996), and Kooperberg (1998).

3. Alternatively, Miller and Liu (2002) use the mutual information that is defined as,

$$I(f : g) = \int \log\left[\frac{f(x_1, x_2)}{g_1(x_1)g_2(x_2)}\right] dF(x_1, x_2)$$

to measure the degree of association among the variables. Note that $I(f: g)$ is not invariant under increasing transformation of the margins.

4. In this paper, we choose $\phi_{ij}(u, v) = u^i v^j$.

5. As is pointed out by a referee, a more appropriate comparison with the kernel estimation shall be based on kernels that correct for the boundary bias. We leave this interesting comparison for future study.

6. The good performance of the ESE under Gaussian margins is expected because the ESE with two moment conditions is the Gaussian distribution.

7. This two-step procedure has been proposed by McNeil and Frey (2000). Jalal and Rockinger (2008) investigated the consequences of using GARCH filter on various misspecified processes. Their results show that two-step approach appears to provide very good tail-related risk measures.

ACKNOWLEDGMENTS

We thank helpful comments from David Bessler, James Richardson, Suojin Wang, and participants at the 7th Advances in Econometrics conference.

REFERENCES

Ang, A., & Chen, J. (2002). Asymmetric correlation of equity portfolios. *Journal of Financial Economics*, *63*, 294–442.

Barron, A., & Sheu, C. H. (1991). Approximation of density functions by sequences of exponential families. *Annals of Statistics*, *19*, 1347–1369.

Bouezmarni, T., & Rombouts, J. (2007). *Nonparametric density estimation for multivariate bounded data*. Unpublished manuscript, HEC Montreal.

Cai, Z., Chen, X., Fan, Y., & Wang, X. (2008). *Selection of copulas with applications in Finance*. Unpublished manuscript, University of North Carolina at Charlotte, Charlotte, NC.

Chen, S. (1999). A beta Kernel estimation for density functions. *Computational Statistics and Data Analysis*, *31*, 131–145.

Chen, S. X., & Huang, T. (2007). *Nonparametric estimation of copula functions for dependence modeling*. Unpublished manuscript, Department of Statistics, Iowa State University, Ames, IA.

Chen, X. H., Fan, Y. Q., & Tsyrennikov, V. (2006). Efficient estimation of semi-parametric multivariate copula models. *Journal of American Statistical Association*, *101*, 1228–1241.

Deheuvels, P. (1979). La Function de Dependance Empirique et Ses Proprietes. Un Test Non Paramtrique d'independence. *Academie Royale de Belgique, Bulletin de la Classe des Sciences*, *65*, 274–292.

Embrechts, P., McNeil, A., & Straumann, D. (1999). Correlation: Pitfalls and alternatives. *Risk*, *5*, 69–71.

Fermanian, J. D., & Scaillet, O. (2003). Nonparametric estimation of copulas for time series. *Journal of Risk*, *5*, 25–54.

Forbes, K., & Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market co-movements. *Journal of Finance*, *57*, 2223–2262.

Frahm, G., Junker, M., & Schmidt, R. (2005). Estimating the tail-dependence coefficient: Properties and pitfalls. *Insurance, Mathematics and Economics*, *37*, 80–100.

Genest, C., Ghoudi, K., & Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*, 534–552.

Genest, C., & Rivest, L. P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of American Statistical Association*, *88*, 1034–1043.

Gijbels, I., & Mielnicnuk, J. (1990). Estimating the density of a Copula function. *Communications in Statistics A*, *19*, 445–464.

Golan, A., Judge, G., & Miller, D. (1996). *Maximum entropy econometrics: Robust estimation with limited data*. New York: Wiley.

Gu, C., & Wang, J. (2003). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica*, *13*, 811–826.

Hall, P., & Neumeyer, N. (2006). Estimating a bivariate density when there are extra data on one or both components. *Biometrika*, *93*, 439–450.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics*, *16*, 342–355.

Jalal, A., & Rockinger, M. (2008). Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data. *Journal of Empirical Finance*, *15*, 868–877.

Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman Hall.

Jones, M. (1993). Simple boundary correction for kernel density estimation. *Statistical Computing*, *3*, 135–146.

Jones, M., & Foster, P. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica*, *6*, 1005–1013.

Kim, Y. (2005). *Dependence structure in international financial markets: Evidence from Asian stock markets*. Unpublished manuscript, Department of Economics, University of California at San Deigo, San Deigo, CA.

Koo, J. Y. (1996). Bivariate B-splines for tensor logspline density estimation. *Computational Statistics and Data Analysis*, *21*, 31–42.

Kooperberg, C. (1998). Bivariate density estimation with an application to survival analysis. *Journal of Computational and Graphical Statistics*, *7*, 322–341.

Lejeune, M., & Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics and Data Analysis*, *14*, 457–471.

Li, Q., & Racine, J. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Liebscher, E. (2005). Semiparametric density estimators using copulas. *Communications in Statistics A*, *34*, 59–71.

Longin, F., & Solnik, B. (2001). Extreme correlation of international equity markets. *Journal of Finance*, *56*, 69–676.

Mandelbort, B. (1963). New methods in statistical economies. *Journal of Political Economy*, *71*, 421–440.

Marron, J., & Ruppert, P. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society, Series B*, *56*, 653–671.

Marron, J., & Wand, P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, *20*, 712–736.

McNeil, A. J., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, *7*, 271–300.

Miller, D., & Liu, W. H. (2002). On the recovery of joint distributions from limited information. *Journal of Econometrics*, *107*, 259–274.

Muller, H. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika*, *78*, 521–530.

Nelsen, R. B. (2006). *An introduction to copulas* (2nd ed.). New York: Springer-Verlag.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika*, *73*, 353–361.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics*, *16*, 356–366.

Rodriguez, J. C. (2007). Measuring financial contagion: A Copula approach. *Journal of Empirical Finance*, *14*, 401–423.

Sancetta, A., & Satchell, S. (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory*, *20*, 535–562.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrica*, *68*, 45–54.

Sklar, A. (1959). Fonctions De Repartition a n Dimensionset Leurs Mrges. *Publications de l'Institut Statistique de l'Université de Paris*, *8*, 229–231.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, *10*, 1040–1053.

Stone, C. (1990). Large-sample inference for log-spline models. *The Annals of Statistics*, *18*, 717–741.

Stone, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, *22*, 118–184.

Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, *115*, 347–354.

Wu, X. (2007). *Exponential series estimator of multivariate density*. Unpublished manuscript, Department of Agricultural Economics, Texas A&M University, College Station, TX.

Zellner, A., & Highfield, R. A. (1988). Calculation of maximum entropy distribution and approximation of marginal posterior distributions. *Journal of Econometrics*, *37*, 195–209.

NONPARAMETRIC ESTIMATION OF MULTIVARIATE CDF WITH CATEGORICAL AND CONTINUOUS DATA

Gaosheng Ju, Rui Li and Zhongwen Liang

ABSTRACT

In this paper we construct a nonparametric kernel estimator to estimate the joint multivariate cumulative distribution function (CDF) of mixed discrete and continuous variables. We use a data-driven cross-validation method to choose optimal smoothing parameters which asymptotically minimize the mean integrated squared error (MISE). The asymptotic theory of the proposed estimator is derived, and the validity of the cross-validation method is proved. We provide sufficient and necessary conditions for the uniqueness of optimal smoothing parameters when the estimation of CDF degenerates to the case with only continuous variables, and provide a sufficient condition for the general mixed variables case.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 291–318

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025012

1. INTRODUCTION

As the rapid advancement of modern computer technology makes the computing of complicated problems feasible, nonparametric statistic methods become increasingly popular. Nonparametric methods have been applied in many economic contexts. The most striking advantage of nonparametric methods over parametric ones is that no prior assumptions, which often turn out to be inappropriate, about the unknown true distributions are taken. The joint distributions of multiple economic variables can give a direct illustration of the relationship among these variables and help researchers to infer the underlying causality. Consequently, the estimation of joint distributions is an important and fundamental issue in the nonparametric econometrics/statistics literature.

Traditionally, nonparametric methods focus on the estimation of either continuous variables or discrete variables (see, e.g., Grund, 1993; Grund & Hall, 1993; Hall, 1981). However, estimation and testing methods able to handle mixed data are quite desirable because most data sets contain both continuous and discrete variables. For instance, labor economists are usually interested in the relationships between the continuous income and discrete explanatory variables such as gender, race, education levels, locations, etc. Recently, Li and Racine (2003), Racine and Li (2004), and Li and Racine (2008) discussed nonparametric smoothing estimations of probability density functions, regression functions, and conditional cumulative distribution functions (CDF) and quantile functions (with mixed discrete and continuous variables). Their work is of great importance for enlarging the scope of the application of nonparametric methods to the context with both continuous and discrete variables. This paper contributes to this literature by investigating a nonparametric estimation of the unconditional joint CDF of mixed data types.

One difficulty in dealing with the estimation of discrete and continuous variables simultaneously is a lack of joint observations. Conventional approaches to handle the estimation of CDF of discrete variables are frequency based. Although we can directly combine it with the kernel estimator of continuous variables, the approach suffers because the number of observations for estimation of discrete variables by a frequency-based approach may be insufficient to ensure an accurate nonparametric estimation of marginal CDF for the remaining continuous variables. Aitchison and Aitken (1976) proposed a novel nonparametric smoothing method to estimate distribution functions defined over binary data. Their method can mitigate the problem of data insufficiency for finite-sample applications.

Their proposed smoothing method can reduce the estimation variance significantly, though it incurs some mild estimation bias. Li and Racine (2003) extended Aitchison and Aitken's method to a context with mixed discrete and continuous variables. In this paper, we adopt their ideas of smoothing both discrete and continuous variables to estimate an unconditional CDF which contains both discrete and continuous components.

It is well known that the selection of smoothing parameters is of crucial importance in nonparametric estimations. There exist several popular methods of smoothing parameter selections. Among them, the most popular ways are the plug-in method and the cross-validation method. There are many discussions about these methods (e.g., Härdle & Marron, 1985; Loader, 1999). However, there is no clear conclusion which method is better. In practice, the cross validation may be a preferred choice, especially in multivariate settings. This is because the cross-validation method is fully data driven. Rudemo (1982) and Bowman (1984) introduced the cross-validation selection of smoothing parameters for density estimation (see Wand & Jones, 1995, Chapter 3; Li & Racine, 2007, for a thorough discussion). Bowman, Hall, and Prvan (1998) presented a cross-validation bandwidth selection for the smoothing estimation of continuous distribution functions. In this paper, we propose to use the least squares cross-validation method to choose the smoothing parameters. We will show that the resultant smoothing parameters are optimal in the sense of minimizing the mean integrated squared error (MISE).

Another interesting problem is the uniqueness of the smoothing parameter vector in cross-validation methods. This was first tackled in Li and Zhou (2005) for the nonparametric kernel estimation of the PDF and regression function of continuous variables. We also discuss this problem in the paper. We give a sufficient and necessary condition for uniqueness when the estimation of CDF degenerates to a case with only continuous variables. For the case of mixed variables, we provide a sufficient condition.

The estimation of CDF is quite useful in econometrics and economics, especially for the econometric theory and economic applications of tests of stochastic dominance. Recently, there are some theories and applications about nonparametric tests of stochastic dominance. Among them are Barrett and Donald (2003) which provided some consistent tests of stochastic dominance for any pre-specified order, Anderson (1996) which gave a nonparametric test of stochastic dominance applied in income distributions, and Davidson and Duclos (2000) which showed some statistical inference and applications in poverty, inequality, and social welfare. Our estimation can be

readily used in the test of stochastic dominance under the circumstance of mixed data.

The paper is organized as follows. In Section 2, we propose an estimator of distribution function that admits mixed discrete and continuous variables. We derive the rates of convergence and establish the asymptotic normality of our estimator. In Section 3, we show that the smoothing parameters selected by the cross-validation method are optimal in the sense that they converge to the minimizer of MISE in probability. In Section 4, we give a sufficient and necessary condition for the uniqueness of the smoothing parameter vector when the estimation contains continuous variables only, and we give a sufficient condition for the mixed case. Section 5 provides an empirical application to examine the relationship between city size and unemployment rate. Section 6 concludes the paper.

2. ESTIMATION OF CDF WITH MIXED DISCRETE AND CONTINUOUS VARIABLES

We consider the case for which $\boldsymbol{x}$ is a vector containing a mix of discrete and continuous variables. Let $x = (x^{\mathrm{c}}, x^{\mathrm{d}})$, where $x^{\mathrm{c}} \in \mathbb{R}^q$ is a q-dimensional continuous random vector, and where x^{d} is an r-dimensional discrete random vector. Let $X_{is}^{\mathrm{d}}(x_s^{\mathrm{d}})$ denote the sth component of $X_i^{\mathrm{d}}(x^{\mathrm{d}})$, $s = 1, \ldots, r$, $i = 1, \ldots n$, where n is the sample size. We restrict the discrete components to a finite support. Without loss of generality, assume that the support of X_{is}^{d} is $\{0, 1, \ldots, c_s - 1\}$, hence the support of X_i^{d} is $S^{\mathrm{d}} = \Pi_{s=1}^{r}\{0, 1, \ldots, c_s - 1\}$. For discrete variables, we use the following kernel:

$$l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, \lambda_s) = \begin{cases} 1 - \lambda_s, & \text{if } X_{is}^{\mathrm{d}} = x_s^{\mathrm{d}} \\ \lambda_s/(c_s - 1), & \text{if } X_{is}^{\mathrm{d}} \neq x_s^{\mathrm{d}} \end{cases}$$

Note that λ_s is a bandwidth having the following properties: when $\lambda_s = 0$, $l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, 0)$ becomes an indicator function, and when $\lambda_s = (c_s - 1)/c_s$, $l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, (c_s - 1)/c_s) = 1/c_s$ becomes a uniform weight function. Thus, the range of λ_s is $[0, (c_s - 1)/c_s]$. The product kernel function is given by

$$L(X_i^{\mathrm{d}}, x^{\mathrm{d}}, \lambda) = \prod_{s=1}^{r} l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, \lambda_s)$$

We use $k(\cdot)$ to denote a univariate kernel function for a continuous variable. The product kernel function used for the continuous variables

is given by

$$K\left(\frac{X_i^{\mathrm{c}} - x^{\mathrm{c}}}{h}\right) = \prod_{j=1}^{q} k\left(\frac{X_{ij}^{\mathrm{c}} - x_j^{\mathrm{c}}}{h_j}\right)$$

where $X_{ij}^{\mathrm{c}}(x_j^{\mathrm{c}})$ denotes the jth component of $X_i^{\mathrm{c}}(x^{\mathrm{c}})$, $j = 1, \ldots, q$, $i = 1, \ldots, n$, and h_j is the bandwidth associated with x_j^{c}. We use $f(x)$ and $F(x)$ to denote the density function and CDF of X, respectively. Following Li and Racine (2003), the kernel estimator of density function $f(x)$ is given by

$$\hat{f}(x) = \hat{f}(x^{\mathrm{c}}, x^{\mathrm{d}}) = \frac{1}{nh_1h_2\cdots h_q}\sum_{i=1}^{n} K\left(\frac{X_i^{\mathrm{c}} - x^{\mathrm{c}}}{h}\right) L(X_i^{\mathrm{d}}, x_i^{\mathrm{d}}, \lambda)$$

Naturally, one can obtain a kernel estimator of $F(x)$ by integrating $\hat{f}(x)$, which is expressed as

$$\hat{F}(x) = \hat{F}(x^{\mathrm{c}}, x^{\mathrm{d}}) = \frac{1}{n}\sum_{i=1}^{n}\left[G\left(\frac{x^{\mathrm{c}} - X_i^{\mathrm{c}}}{h}\right)\left(\sum_{u\le x^{\mathrm{d}}} L(X_i^{\mathrm{d}}, u, \lambda)\right)\right] \tag{1}$$

where $G(x) = \int_{-\infty}^{x} k(v)dv$, and $G((x^{\mathrm{c}} - X_i^{\mathrm{c}})/h) = \prod_{j=1}^{\mathrm{q}} G((x_j^{\mathrm{c}} - X_{ij}^{\mathrm{c}})/h_j)$.

We introduce some notations before we state the main theorem of this section. Let $\mathbf{1}(\mathrm{A})$ denote an indicator function that takes the value 1 if A occurs and 0 otherwise. Define an indicator function $\mathbf{1}s(\cdot, \cdot)$ by

$$\mathbf{1}_s(z^{\mathrm{d}}, u) = \mathbf{1}(z_s^{\mathrm{d}} \neq u_s)\prod_{t\neq s}^{r} \mathbf{1}(z_t^{\mathrm{d}} = u_t) \tag{2}$$

We can see that $\mathbf{1}_{\mathrm{s}}(\cdot, \cdot)$ equals to one if and only if z^{d} and u differ only in the sth component. The following assumptions will be used in studying the asymptotic behavior of cross-validated smoothing parameters and in deriving the asymptotic distribution of our CDF estimator.

Condition (C1). The data $\{(X_i^{\mathrm{c}}, X_i^{\mathrm{d}})\}_{i=1}^{n}$ are independent and identically distributed as $(X^{\mathrm{c}}, X^{\mathrm{d}})$. $F(x^{\mathrm{c}}, x^{\mathrm{d}})$ has continuous third-order partial derivatives with respect to x^{c}.

Condition (C2). $k(\cdot)$ is a bounded and symmetric kernel density function with a compact support. $\int k(v)dv = 1$, $\int v^2k(v)dv = \kappa_2 < \infty$.

Condition (C3). As $n\to\infty$, $h_j\to 0$, $nh_j^6 \to 0$, for $j = 1, \ldots, q$ and $\lambda_s \to 0$, $n\lambda_s^4 \to 0$, for $s = 1, \ldots, r$.

Let $F_j^{(1)}(x^c, x^d) = (\partial F(x))/(\partial x_j^c)$, $F_{jj}^{(2)}(x^c, x^d) = (\partial^2 F(x))/(\partial x_j^c \partial x_j^c)$. The next theorem shows the rate of convergence in terms of MSE and MISE and the asymptotic normality of our estimator.

Theorem 1. Under condition (C1), (C2), and (C3), we have

(i) $$\mathrm{MSE}(\hat{F}(x^c, x^d)) = \frac{1}{n}F(x^c, x^d)(1 - F(x^c, x^d)) - \sum_{j=1}^{q} A_{1j}\frac{h_j}{n} + \sum_{s=1}^{r} A_{2s}\frac{\lambda_s}{n} + \left(\sum_{j=1}^{q} B_{1j}h_j^2 + \sum_{s=1}^{r} B_{2s}\lambda_s\right)^2 + O\left(\frac{1}{n}\sum_{j=1}^{q} h_j^2 + \frac{1}{n}\sum_{s=1}^{r}\lambda_s^2 + \sum_{j=1}^{q} h_j^6 + \sum_{s=1}^{r}\lambda_s^4\right)$$

where $\alpha_0 = 2\int vG(v)k(v)dv$, $A_{1j} = \alpha_0 F_j^{(1)}(x^c,\ x^d)$, $A_{2s} = 2/(c_s - 1)\sum_{u \le x^d}\sum_{v \le x^d,\ v \ne u}\mathbf{1}_s(u, v)F(x^c \mid u)p(u) - 2F(x^c, x^d) - 2F(x^c, x^d)B_{2s}$, $B_{1j} = (1/2)\kappa_2 F_{jj}^{(2)}(x^c, x^d)$, and $B_{2s} = 1/(c_s - 1)\sum_{z^d \in S^d}\sum_{u \le x^d}\mathbf{1}_s(z^d, u)F(x^c|x^d)p(x^d) - F(x^c, x^d)$.

(ii) $$\mathrm{MISE}(\hat{F}(x^c, x^d)) = Z^T\left(\sum_{x^d \in S^d}\int BB^T dx^c\right)Z + \frac{1}{n}\mathcal{A}^T\tilde{Z} + \frac{1}{n}\sum_{x^d \in S^d}\int F(x^c, x^d)(1 - F(x^c, x^d))dx^c + O\left(\frac{1}{n}\sum_{j=1}^{q} h_j^2 + \frac{1}{n}\sum_{s=1}^{r}\lambda_s^2 + \sum_{j=1}^{q} h_j^6 + \sum_{s=1}^{r}\lambda_s^4\right) \quad (3)$$

where $Z = (h_1^2,\ \ldots,\ h_q^2,\ \lambda_1,\ \ldots,\ \lambda_r)^T$, $\tilde{Z} = (h_1,\ \ldots,\ h_q,\ \lambda_1,\ \ldots,\ \lambda_r)^T$, $B = (B_{11},\ \ldots,\ B_{1q},\ B_{21},\ \ldots,\ B_{2r})^T$, and $\mathcal{A} = \sum_{x^d \in S^d}\int(-A_{11},\ \ldots,\ -A_{1q},\ A_{21},\ \ldots,\ A_{2r})^T dx^c$.

(iii) $$\sqrt{n}\left(\hat{F}(x^c, x^d) - F(x^c, x^d) - \sum_{j=1}^{q} B_{1j}h_j^2 - \sum_{s=1}^{r} B_{2s}\lambda_s\right) \xrightarrow{d} N(0,\ F(x^c, x^d)(1 - F(x^c, x^d))).$$

The proof of Theorem 1 is given in Appendix A.

We can see that the convergence rate of our CDF estimator is $\sqrt{n}$. Under the optimal convergence rates for h_j and λ_s, $j = 1,\ \ldots,\ q$, $s = 1,\ \ldots,\ r$

(i.e., $h_j \sim n^{-1/3}$ and $\lambda_s \sim n^{-2/3}$), the statement (iii) in Theorem 1 simplifies to $\sqrt{n}(\hat{F}(x^c, x^d) - F(x^c, x^d)) \xrightarrow{d} N(0,\ F(x^c, x^d)(1 - F(x^c, x^d)))$.

3. CROSS-VALIDATION BANDWIDTH SELECTION

In this section, we focus on how to choose the smoothing parameters when estimating $\hat{F}(\cdot)$. Theoretically, we may choose the optimal bandwidths by minimizing the leading term of MISE given by Eq. (3) in Theorem 1. Taking derivatives with respect to h_j and λ_s, one can easily see that optimal smoothing requires that $h_j \sim n^{-1/3}$, $j = 1, \ldots, q$ and $\lambda_s \sim n^{-2/3}$, $s = 1, \ldots, r$, as $q \geq 1$. However, we can see that the coefficients of these orders involve unknown functions. Therefore, this method is infeasible in practice. In practice one can compute plug-in bandwidths based on Eq. (3) by choosing some initial "pilot" bandwidths, the results may be sensitive to the choice of these pilots. Therefore, it is highly desirable to construct an automatic data-driven bandwidth selection procedure, which does not rely on some ad hoc pilot bandwidth values to estimate unknown functions.

Following Bowman et al. (1998), we suggest choosing the smoothing parameters $(h, \lambda) = (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ by minimizing the following cross-validation function:

$$\mathrm{CV}(h, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{x^d \in S^d} \int (I(x^c, X_i^c) I(x^d, X_i^d) - \hat{F}_{-i}(x^c, x^d))^2 dx^c \right]$$

where $\hat{F}_{-i}(x^c, x^d) = (1/(n-1)) \sum_{j \neq i} G((x^c - X_j^c)/h) \sum_{u \leq x^d} L(X_j^d, u, \lambda)$, $I(x^c, X_i^c) = \mathbf{1}(X_i^c \leq x^c)$, and $I(x^d, X_i^d) = \mathbf{1}(X_i^d \leq x^d)$.

Define $I_i \equiv I(x, X_i) = I(x^c, X_i^c) I(x^d, X_i^d)$ and a term unrelated to smoothing parameters

$$J_n = \sum_{x^d \in S^d} \int \{(F_n - F)^2 - E[(F_n - F)^2]\} dx^c$$
$$- \frac{1}{n} \sum_{i=1}^{n} \sum_{x^d \in S^d} \int [I(x, X_i) - F(x^c, x^d)]^2 dx^c$$

where $F_n(x^c, x^d) = (1/n) \sum_{i=1}^{n} I(x^c, X_i^c) I(x^d, X_i^d)$ is the empirical distribution function. In Theorem 2 below, we show that $H(h, \lambda) = \mathrm{CV}(h, \lambda) + J_n$ is a good approximation to MISE(h, λ).

Theorem 2. Define $H(h, \lambda) = \mathrm{CV}(h, \lambda) + J_n$, then under condition (C1) and (C2), we have for each δ, ε, $C > 0$,

$$H(h,\lambda) = \mathrm{MISE}(h,\lambda) + O_p\left(\left(n^{-3/2} + n^{-1}\sum_{j=1}^{q} h_j^q + n^{-1/2}\sum_{j=1}^{q} h_j^{q+2} + n^{-1/2}\sum_{j=1}^{q} h_j^4 + n^{-1/2}\sum_{s=1}^{r} \lambda_s^2\right) n^{\delta}\right)$$

with probability 1, uniformly in $0 \le h_j, \lambda_s \le Cn^{-\varepsilon}$ for $j = 1, \ldots, q$, $s = 1, \ldots, r$, as $n \to \infty$.

Essentially, Theorem 2 says that $\mathrm{CV}(h,\lambda) =$ (leading terms of $\mathrm{MISE}(h, \lambda)$)+ (terms unrelated to h, λ)+(small order terms). Therefore, minimizing cross-validation function is asymptotically equivalent to minimizing MISE (h, λ). Therefore, we immediately have the following corollary.

Corollary 1. Under the conditions (C1) and (C2), let $\hat{h}_j$, $\hat{\lambda}_s$, $j = 1, \ldots, q$, $s = 1, \ldots, r$ denote the smoothing parameters that minimizes the $\mathrm{CV}(h, \lambda)$ over the set $[0, Cn^{-\varepsilon}]^{q+r}$ for any $C > 0$ and any $0 < \varepsilon < 1/3$, let h_j^0, λ_s^0, $j = 1, \ldots, q$, $s = 1, \ldots, r$ denote the smoothing parameters that minimizes the $\mathrm{MISE}(h, \lambda)$, then we have

$$\frac{\hat{h}_j}{h_j^0} \to 1 \quad \text{and} \quad \frac{\hat{\lambda}_s}{\lambda_s^0} \to 1 \quad (\text{if } \lambda_s^0 \neq 0) \quad \text{or} \quad \hat{\lambda}_s \to 0 \quad (\text{if } \lambda_s^0 = 0)$$

in probability, for all $j = 1, \ldots, q$, and $s = 1, \ldots, r$.

The proof of Theorem 2 is given in the Appendix B.

4. UNIQUENESS OF SMOOTHING PARAMETER VECTOR

Section 3 has established the fact that minimizing cross-validation function is asymptotically equivalent to minimizing MISE. Hence, to investigate the asymptotic uniqueness of the cross-validated smoothing parameters, we only need to examine the uniqueness of parameters minimizing the leading terms of MISE. When there does not exist discrete variables,

our objective function is

$$\inf_{Z\in\mathbb{R}_+^q,\ ||Z||=1} Z^T MZ + \frac{1}{n}\mathcal{A}^T Z^{1/2} \tag{4}$$

where $Z=(h_1^2,\ \ldots,\ h_q^2)^T$, $Z^{1/2}=(h_1,\ \ldots,\ h_q)^T$, $M=\sum_{x^d\in S^d}\int BB^T dx^c$, and both $\mathcal{A}$ and B are of dimension $q\times 1$ (they are the first q elements of the general mixed variable case). Based on the previous discussion, the optimal rates for h_j and λ_s are $n^{-1/3}$ and $n^{-2/3}$, respectively. Let $h_j=a_j n^{-1/3}$, for $j=1,\ \ldots,\ q$. Substituting these parameters into Eq. (4), then minimize $Z^T MZ+(1/n)\mathcal{A}^T Z^{1/2}$ is equivalent to minimize $Z^T MZ+\mathcal{A}^T Z^{1/2}$, where we abuse notation a little bit, $Z=(a_1^1,\ \ldots,\ a_q^2)^T$ and $Z^{1/2}=(a_1,\ \ldots,\ a_q)^T$.

When the estimation of CDF degenerates to the case with only continuous variables, we give the necessary and sufficient condition in the following theorem.

Theorem 3. Assume that $r=0$, let $Z=(h_1^2,\ \ldots,\ h_q^2)^T$, define $\mu=\inf_{Z\in\mathbb{R}_+^q,\ ||Z||=1} Z^T MZ$. Then $\chi(Z)=Z^T MZ+\mathcal{A}^T Z^{1/2}$ has a unique minimizer $Z^*\in\mathbb{R}_+^q$, if and only if $\mu>0$.

Proof. Our proof follows similar arguments as in Li and Zhou (2005). First we prove the "only if" part. Suppose $\mu=0$ is attained at some $Z^{(0)}\in\mathbb{R}_+^q$ with $||Z^{(0)}||=1$. Then there exists at least one component $Z_i^{(0)}\neq 0$, that is, $Z_i^{(0)}>0$. So $\chi(tZ^{(0)})=t^2(Z^{(0)})^T MZ^{(0)}+\mathcal{A}^T\sqrt{t}(Z^{(0)})^{1/2}=\mathcal{A}^T\sqrt{t}(Z^{(0)})^{1/2}\to-\infty$, as $t\to+\infty$. Note that the components of $\mathcal{A}$ are negative, and $tZ^{(0)}\in\mathbb{R}_+^q$. This implies that χ has no minimizer.

Next we prove the "if" part. If $\mu>0$, for any $Z\in\mathbb{R}_+^q$, with $||Z||=1$, we have that $tZ\in\mathbb{R}_+^q, t>0$. Then $\chi(tZ)=t^2Z^T MZ+\sqrt{t}\mathcal{A}^T Z^{1/2}=(t\sqrt{Z^T MZ}-(1/(2\sqrt{Z^T MZ})))^2+(\sqrt{t}+((\mathcal{A}^T Z^{1/2})/2))^2-(1/4Z^T MZ)-((\mathcal{A}^T Z^{1/2})^2/4)\to+\infty$, as $t\to+\infty$. For $R>0$, denote $B_R=\{Z\in\mathbb{R}_+^q:\ ||Z||\le R\}$. Since χ is a continuous function on $\mathbb{R}_+^q$, B_R is a compact set and $\chi(tZ)\to+\infty$, as $t\to+\infty$, we have that there exists $R>0$ such that $\min_{Z\in\mathbb{R}_+^q}\chi(Z)\Leftrightarrow\min_{Z\in B_R}\chi(Z)$.

From $\chi(tZ)=t^2Z^T MZ+t^{1/2}\mathcal{A}^T Z^{1/2}$, we know that $\chi(tZ)$ attains its minimum at $t=(-(\mathcal{A}^T Z)/(4Z^T MZ))^{2/3}>0$. So 0 is not the minimizer of χ. Similarly, we get that $\chi(Z+t(0,\ \ldots,\ 1,\ \ldots,\ 0)^T)=Z^T MZ+2tZ^T M(0,\ \ldots,\ 1,\ \ldots,\ 0)^T+\mathcal{A}^T Z+t^2m_{ii}+A_i t^{1/2}$ cannot attain its

minimum at $t = 0$. So Z with $h_i^2 = 0$ cannot be the minimizer of χ, which means that χ can only attain its minimum in the interior of B_R.

The Hessian matrix $\mathcal{H}$ of χ is $\mathcal{H} = (\partial^2\chi/(\partial Z \partial Z^T)) = 2M + G$, where $G = (1/4)\,\mathrm{diag}\,(-c_1 z_1^{-3/2}, -c_2 z_2^{-3/2}, \ldots, -c_q z_q^{-3/2})$ is a diagonal matrix. Since $c_i < 0$, G is positive definite in the interior of B_R. Also, M is symmetric and positive semi-definite. So $\mathcal{H}$ is positive definite in the interior of B_R. Therefore, χ has a unique minimizer in the interior of B_R. This completes the proof.

In general, our objective function is

$$\inf_{Z \in \mathbb{R}_+^q,\ ||Z||=1} Z^T M Z + \frac{1}{n} \mathcal{A}^T \tilde{Z} \tag{5}$$

where $Z = (h_1^2, \ldots, h_q^2, \lambda_1, \ldots, \lambda_r)^T$, $\tilde{Z} = (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)^T$, $M = \sum_{x^d \in S^d} \int B B^T dx^c$, $B = (B_{11}, \ldots, B_{1q}, B_{21}, \ldots, B_{2r})^T$, and $\mathcal{A} = \sum_{x^d \in S^d} \int (-A_{11}, \ldots, -A_{1q}, A_{21}, \ldots, A_{2r})^T dx^c$ are defined in Theorem 1. Substituting $h_j = a_j n^{-1/3}$, for $j = 1, \ldots, q$, and $\lambda_s = b_s n^{-2/3}$, for $s = 1, \ldots, r$ into Eq. (5), we have that Eq. (5) is equivalent to minimize $Z^T M Z + \mathcal{A}^T \tilde{Z}$ with respect to $Z = (a_1^2, \ldots, a_q^2, b_1, \ldots, b_r)^T$ and $\tilde{Z} = (a_1, \ldots, a_q, b_1, \ldots, b_r)^T$. A sufficient condition for the estimation of the CDF of the mixed discrete and continuous variables is given as follows.

Theorem 4. Let $\mu = \inf_{Z \in \mathbb{R}_+^{q+r},\ ||Z||=1} Z^T M Z$. If $\mu > 0$, then χ has a minimizer $Z^* \in \mathbb{R}_+^{q+r}$. If M is positive definite, then Hessian matrix $\mathcal{H}$ of χ is positive definite at every point of $\mathbb{R}_+^{q+r}$. Thus, χ has a unique minimizer $Z^* \in \mathbb{R}_+^{q+r}$.

Proof. If $\mu > 0$, for any $Z \in \mathbb{R}_+^{q+r}$, with $||Z|| = 1$, we have that $tZ \in \mathbb{R}_+^{q+r}$, $t > 0$. Using the notation $Z_{(1)} = (a_1^2, \ldots, a_q^2)^T$, $Z_{(1)}^{1/2} = (a_1, \ldots, a_q)^T$ and $Z_{(2)} = (b_1, \ldots, b_r)^T$, we have $\chi(tZ) = t^2 Z^T M Z + \sqrt{t}\,\mathcal{A}_1^T Z_{(1)}^{1/2} + t\,\mathcal{A}_2^T Z_{(2)} = (t\sqrt{Z^T M Z} + ((\mathcal{A}_2^T Z_{(2)} - 1)/(2\sqrt{Z^T M Z})))^2 + (\sqrt{t} + \mathcal{A}_1^T Z_{(1)}^{1/2}/2)^2 - ((\mathcal{A}_2^T Z_{(2)} - 1)^2/(4 Z^T M Z)) - ((\mathcal{A}_1^T Z_{(1)}^{1/2})^2/4) \to +\infty$, as $t \to +\infty$, where $\mathcal{A}_1 = (c_1, \ldots, c_q)^T$, $\mathcal{A}_2 = (c_{q+1}, \ldots, c_{q+r})^T$. For $R > 0$, denote $B_R = \{Z \in \mathbb{R}_+^{q+r} : ||Z|| \le R\}$. Since χ is a continuous function on $\mathbb{R}_+^{q+r}$, B_R is a compact set, and $\chi(tZ) \to +\infty$, $t \to +\infty$, we have that there exists $R > 0$, such that $\min_{Z \in \mathbb{R}_+^{q+r}} \chi(Z) \Leftrightarrow \min_{Z \in B_R} \chi(Z)$. Therefore, χ has a minimizer $Z^* \in \mathbb{R}_+^{q+r}$.

The Hessian matrix $\mathcal{H}$ of χ is $\mathcal{H} = \partial^2\chi/(\partial Z \partial Z^T) = 2M + \begin{pmatrix} G & 0 \\ 0 & 0 \end{pmatrix}$. If M is positive definite, then $\mu > 0$, since $Z^T M Z > 0$ on the compact set

$\{Z : Z \in \mathbb{R}_+^{q+r}, ||Z|| = 1\}$. Also, $\mathcal{H}$ is positive definite at every point $Z \in \mathbb{R}_+^{q+r}$. Thus, χ has a unique minimizer $Z^* \in \mathbb{R}_+^{q+r}$. This completes the proof.

5. AN EMPIRICAL APPLICATION

Gan and Zhang (2006) presented a theory predicting that a large city tends to have smaller unemployment rate. Their empirical study applied US data on city population and average unemployment rate based upon a sample of 295 cities. The average unemployment rate, which is continuous, ranges from 2.4% to 19.6%. To get a categorical variable, we artificially stipulate that those with population of more than 200,000 are large cities, and the others are small cities. This classification gives 112 large cities and 183 small cities. In Fig. 1, we plot the conditional CDF of unemployment rate, which is calculated from our estimation of the joint CDF, for large and small cities. We use a Gaussian kernel for the unemployment rate. The cross-validated bandwidths for the continuous variable and categorical variable are 0.3470 and 0.0289, respectively.[1]

The conditional CDF estimate is consistent with the theory that large cities tend to have lower unemployment rates than small cities.

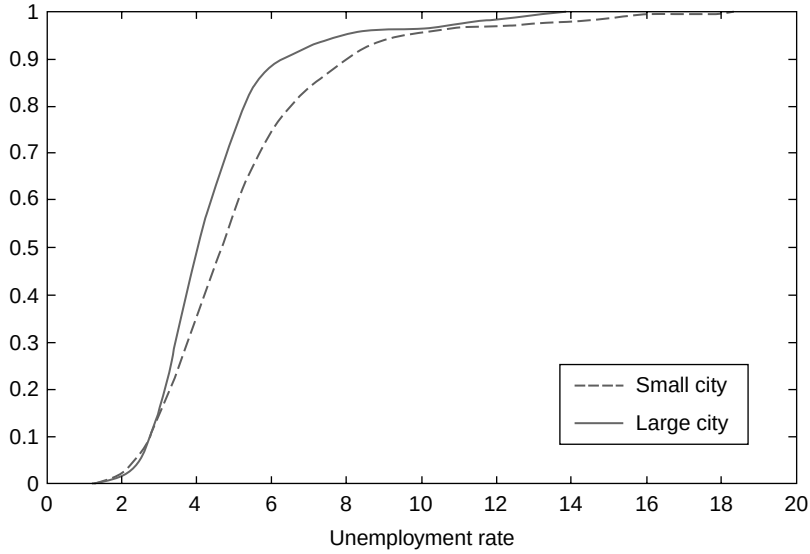


Fig. 1. CDF Estimate of Unemployment Rate of Large and Small Cities.

The conditional CDF curve for large cities is above that of the small cities for most part. Fig. 1 shows that, for most of the unemployment range, the distribution of unemployment rate for large cities stochastically dominates that of small cities.

6. CONCLUSION

We propose a consistent nonparametric kernel estimator of joint unconditional CDF defined over a mix of discrete and continuous variables. A data-driven cross-validation method for selecting the smoothing parameters is examined. We show that it is asymptotically equivalent to minimizing integrated MSE. The uniqueness condition of the cross-validation procedure is discussed. In view of the fact that many economic data sets involve both continuous and discrete variables, our proposed estimator should prove useful to applied researchers.

NOTE

1. For practical implementations of nonparametric econometrics, refer to Scott (1992) and Hayfield and Racine (2008).

ACKNOWLEDGMENTS

We thank the editors and two referees for their insightful comments which help us to improve our paper substantially. We also thank Dr. Qi Li who leads us to this fruitful field and for the intense discussion of this paper.

REFERENCES

Aitchison, J., & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, *63*(3), 413–420.

Anderson, G. (1996). Nonparametric tests of stochastic dominance in income distributions. *Econometrica*, *64*(5), 1183–1193.

Barrett, G. F., & Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, *71*(1), 71–104.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, *71*(2), 353–360.

Bowman, A., Hall, P., & Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, *85*(4), 799–808.

Davidson, R., & Duclos, J.-Y. (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica*, *68*(6), 1435–1464.

Gan, L., & Zhang, Q. (2006). The thick market effect on local unemployment rate fluctuations. *Journal of Econometrics*, *133*(1), 127–152.

Grund, B. (1993). Kernel estimators for cell probabilities. *Journal of Multivariate Analysis*, *46*(2), 283–308.

Grund, B., & Hall, P. (1993). On the performance of kernel estimators for high-dimensional, sparse binary data. *Journal of Multivariate Analysis*, *44*(2), 321–344.

Hall, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika*, *68*(1), 287–294.

Hall, P., & Heyde, C. C. (1980). *Martingale limit theory and its applications*. New York, NY: Academic Press.

Härdle, W., & Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics*, *13*(4), 1465–1481.

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*(5), 1–32.

Li, Q., & Racine, J. S. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, *86*(2), 266–292.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Li, Q., & Racine, J. S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, *26*(4), 423–434.

Li, Q., & Zhou, J. (2005). The uniqueness of cross-validation selected smoothing parameters in kernel estimation of nonparametric models. *Econometric Theory*, *21*(5), 1017–1025.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in?. *The Annals of Statistics*, *27*(2), 415–438.

Racine, J. S., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, *9*(2), 65–78.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York, NY: Wiley.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall.

APPENDIX A. PROOF OF THEOREM 1

Proof of Theorem 1

As we all know $\mathrm{MSE}(\hat{F}(x)) = E[\hat{F}(x) - F(x)]^2 = [\mathrm{bias}(\hat{F}(x))]^2 + \mathrm{var}(\hat{F}(x))$. We will evaluate the terms $\mathrm{bias}(\hat{F}(x))$ and $\mathrm{var}(\hat{F}(x))$ separately. For simplicity, we use dz and dv to denote $dz_1 \ldots dz_q$ and $dv_1 \ldots dv_q$, respectively, throughout the appendices.

For the continuous variables, using the change of variables, integration by parts, and Taylor expansion, we have

$$
\begin{aligned}
E\left[G\left(\frac{x^c - X_i^c}{h}\right)\right] &= E\left[\prod_{j=1}^{q} G\left(\frac{x_j^c - X_{ij}^c}{h_j}\right)\right] = \int \left[\prod_{j=1}^{q} G\left(\frac{x_j^c - z_j}{h_j}\right)\right] f(z_1, z_2, \ldots, z_q)dz \\
&= h_1 h_2 \cdots h_q \int \left[\prod_{j=1}^{q} G(v_j)\right] f(x_1^c - h_1 v_1, x_1^c - h_2 v_2, \ldots, x_q^c - h_q v_q)dv \\
&= \int \left[\prod_{j=1}^{q} k(v_j)\right] F(x_1^c - h_1 v_1, x_2^c - h_2 v_2, \ldots, x_q^c - h_q v_q)dv \\
&= \int \left[\prod_{j=1}^{q} k(v_j)\right] \left\{F(x^c) - \sum_{j=1}^{q} F_j^{(1)}(x^c) h_j v_j + \frac{1}{2}\sum_{i,j=1}^{q} F_{ij}^{(2)}(x^c) h_i h_j v_i v_j\right\} dv \\
&\quad + O\left(\sum_{j=1}^{q} h_j^3\right) = F(x^c) + \frac{\kappa_2}{2}\sum_{j=1}^{q} F_{jj}^{(2)}(x^c) h_j^2 + O\left(\sum_{j=1}^{q} h_j^3\right)
\end{aligned} \tag{A.1}
$$

where $\kappa_2 = \int v^2 k(v)dv$, and

$$
\begin{aligned}
E\left[G^2\left(\frac{x^c - X_i^c}{h}\right)\right] &= E\left[\prod_{j=1}^{q} G^2\left(\frac{x_j^c - X_{ij}^c}{h_j}\right)\right] = \int \left[\prod_{j=1}^{q} G^2\left(\frac{x_j^c - z_j}{h_j}\right)\right] f(z_1, z_2, \ldots, z_q)dz \\
&= h_1 h_2 \cdots h_q \int \left[\prod_{j=1}^{q} G^2(v_j)\right] f(x_1^c - h_1 v_1, x_2^c - h_2 v_2, \ldots, x_q^c - h_q v_q)dv \\
&= 2^q \int \left[\prod_{j=1}^{q} G(v_j)\right] \left[\prod_{j=1}^{q} k(v_j)\right] F(x_1^c - h_1 v_1, x_2^c - h_2 v_2, \ldots, x_q^c - h_q v_q)dv \\
&= 2^q \int \left[\prod_{j=1}^{q} G(v_j)\right] \left[\prod_{j=1}^{q} k(v_j)\right] \{F(x^c) - \sum_{j=1}^{q} F_j^{(1)}(x^c) h_j v_j\} dv + O\left(\sum_{j=1}^{q} h_j^2\right) \\
&= F(x^c) - \alpha_0 \sum_{j=1}^{q} F_j(x^c) h_j + O\left(\sum_{j=1}^{q} h_j^2\right)
\end{aligned} \tag{A.2}
$$

where $\alpha_0 = 2\int vG(v)k(v)dv$.

For the discrete variables, we have

$$
\begin{aligned}
L(z^{\mathrm{d}}, u, \lambda) &= \prod_{s=1}^{r} l(z_s^{\mathrm{d}}, u_s, \lambda) = \prod_{s=1}^{r} \left(\frac{\lambda_s}{c_s - 1}\right)^{\mathbf{1}(z_s^{\mathrm{d}} \neq u_s)} (1 - \lambda_s)^{\mathbf{1}(z_s^{\mathrm{d}} = u_s)} \\
&= \left(\prod_{s=1}^{r}(1 - \lambda_s)\right)\mathbf{1}(z^{\mathrm{d}} = u) + \sum_{s=1}^{r}\left(\frac{\lambda_s}{c_s - 1}\prod_{t \neq s}^{r}(1 - \lambda_t)\right)\mathbf{1}_s(z^{\mathrm{d}}, u) \\
&\quad + O\left(\sum_{s=1}^{r}\lambda_s^2\right) = \left(1 - \sum_{s=1}^{r}\lambda_s\right)\mathbf{1}(z^{\mathrm{d}} = u) \\
&\quad + \sum_{s=1}^{r}\frac{\lambda_s}{c_s - 1}\mathbf{1}_s(z^{\mathrm{d}}, u) + O\left(\sum_{s=1}^{r}\lambda_s^2\right) \qquad \text{(A.3)}
\end{aligned}
$$

where $\mathbf{1}(z^{\mathrm{d}} = u)$ and $\mathbf{1}_s(z^{\mathrm{d}}, u)$ are indicator functions. $\mathbf{1}_s(z^{\mathrm{d}}, u)$ denotes that z^{d} and u only differ in sth component. Note that if z^{d} and u differ in more than one component, $L(z^{\mathrm{d}}, u, \lambda) = O(\sum_{s=1}^{r}\lambda_s^2)$.

From (A.3), it is easy to obtain:

$$
\begin{aligned}
\left[\sum_{u \leq x^{\mathrm{d}}} L(z^{\mathrm{d}}, u, \lambda)\right]^2 &= \left[\sum_{u \leq x^{\mathrm{d}}}\left(1 - \sum_{s=1}^{r}\lambda_s\right)\mathbf{1}(z^{\mathrm{d}} = u)\right. \\
&\quad \left. + \sum_{u \leq x^{\mathrm{d}}}\sum_{s=1}^{r}\frac{\lambda_s}{c_s - 1}\mathbf{1}_s(z^{\mathrm{d}}, u) + O\left(\sum_{s=1}^{r}\lambda_s^2\right)\right]^2 \\
&= \left(1 - \sum_{s=1}^{r}\lambda_s\right)^2\left[\sum_{u \leq x^{\mathrm{d}}}\mathbf{1}(z^{\mathrm{d}} = u)\right] \\
&\quad + 2\sum_{s=1}^{r}\frac{\lambda_s}{c_s - 1}\left[\sum_{u,\ v \leq x^{\mathrm{d}}}\mathbf{1}(z^{\mathrm{d}} = u)\mathbf{1}_s(z^{d}, v)\right] + O\left(\sum_{s=1}^{r}\lambda_s^2\right) \\
&= \left(1 - 2\sum_{s=1}^{r}\lambda_s\right)\left[\sum_{u \leq x^{\mathrm{d}}}\mathbf{1}(z^{\mathrm{d}} = u)\right] \\
&\quad + 2\sum_{s=1}^{r}\frac{\lambda_s}{c_s - 1}\left[\sum_{u}\sum_{v \neq u}\mathbf{1}(z^{\mathrm{d}} = u)\mathbf{1}_s(u, v)\right] \\
&\quad + O\left(\sum_{s=1}^{r}\lambda_s^2\right) \qquad \text{(A.4)}
\end{aligned}
$$

Here and in the following, for any two vectors $x, y \in \mathbb{R}^r$, $x \leq y$ denotes $x_i \leq y_i$ for all $i = 1, \ldots, r$, where x_i and y_i are the ith component of x and y, respectively.

We use $f(x^c|x^d)$ and $F(x^c|x^d)$ to denote the conditional density function and conditional CDF of X, respectively. Then,

$$f(x^c, x^d) = f(x^c|x^d)p(x^d) \tag{A.5}$$

$$F(x^c, x^d) = \sum_{z^d \in S^d,\ z^d \le x^d} F(x^c|z^d)p(z^d) \tag{A.6}$$

With (A.5) and (A.6), we can calculate $E[G(\cdot)\sum L(\cdot)]$ by two steps. First, integrate the integrand with respect to x^c conditional on x^d and then take the summation with respect to x^d. Thus,

$$\begin{aligned}
E\left[\hat{F}(x^c, x^d)\right] &= E\left[G\left(\frac{x^c - X_i^c}{h}\right)\sum_{u \le x^d} L(X_i^d, u, \lambda)\right] \\
&= \sum_{z^d \in S^d}\int G\left(\frac{x^c - z^c}{h}\right)\sum_{u \le x^d} L(z^d, u, \lambda) f(z^c|z^d)p(z^d)dz^c \\
&= \sum_{z^d \in S^d}\left(\int G\left(\frac{x^c - z^c}{h}\right)f(z^c|z^d)dz^c\right)\left(\sum_{u \le x^d} L(z^d, u, \lambda)\right)p(z^d) \\
&= \sum_{z^d \in S^d}\left[\left(F(x^c|z^d) + \frac{\kappa_2}{2}\sum_{j=1}^{q} F_{jj}^{(2)}(x^c|z^d)h_j^2\right.\right. \\
&\quad \left. + O\left(\sum_{j=1}^{q} h_j^3\right)\right)\sum_{u \le x^d}\left(\left(1 - \sum_{s=1}^{r}\lambda_s\right)\mathbf{1}(z^d = u)\right. \\
&\quad \left.\left. + \sum_{s=1}^{r}\frac{\lambda_s}{c_s - 1}\mathbf{1}_s(z^d, u) + O\left(\sum_{s=1}^{r}\lambda_s^2\right)\right)\right]p(z^d) \\
&= F(x^c, x^d) + \sum_{s=1}^{r}\left[\frac{1}{c_s - 1}\sum_{z^d \in S^d}\sum_{u \le x^d}\mathbf{1}_s(z^d, u)F(x^c|x^d)p(x^d) - F(x^c, z^d)\right]\lambda_s \\
&\quad + \sum_{j=1}^{q}\left[\frac{\kappa_2}{2}F_{jj}^{(2)}(x^c, x^d)\right]h_j^2 + O\left(\sum_{j=1}^{q} h_j^3 + \sum_{s=1}^{r}\lambda_s^2\right) \\
&= F(x^c, x^d) + \sum_{j=1}^{q} B_{1j}h_j^2 + \sum_{s=1}^{r} B_{2s}\lambda_s + O\left(\sum_{j=1}^{q} h_j^3 + \sum_{s=1}^{r}\lambda_s^2\right)
\end{aligned} \tag{A.7}$$

where $B_{1j} = (\kappa_2/2)F_{jj}^{(2)}(x^c, x^d)$, $B_{2s} = (1/(c_s - 1))\sum_{z^d \in S^d}\sum_{u \le x^d}\mathbf{1}_s(z^d, u)F(x^c|x^d)$ $p(x^d) - F(x^c, x^d)$.

So we obtain

$$\text{bias}(\hat{F}(x^{\text{c}}, x^{\text{d}})) = \sum_{j=1}^{q} B_{1j} h_j^2 + \sum_{s=1}^{r} B_{2s} \lambda_s + O\left(\sum_{j=1}^{q} h_j^3 + \sum_{s=1}^{r} \lambda_s^2\right) \tag{A.8}$$

Similarly, combining (A.2) and (A.4), we have

$$\begin{aligned}
&E\left[G^2\left(\frac{x^{\text{c}} - X_i^{\text{c}}}{h}\right)\left[\sum_{u \le x^{\text{d}}} L(X_i^{\text{d}}, u, \lambda)\right]^2\right] \\
&= \sum_{z^{\text{d}} \in S^{\text{d}}} \int G^2\left(\frac{x^{\text{c}} - z^{\text{c}}}{h}\right)\left[\sum_{u \le x^{\text{d}}} L(z^{\text{d}}, u, \lambda)\right]^2 f(z^{\text{c}}|z^{\text{d}}) p(z^{\text{d}}) dz^{\text{c}} \\
&= \sum_{z^{\text{d}} \in S^{\text{d}}} \left[F(x^{\text{c}}|z^{\text{d}}) - \alpha_0 \sum_{j=1}^{q} F_j^{(1)}(x^{\text{c}}|z^{\text{d}}) h_j + O\left(\sum_{j=1}^{q} h_j^2\right)\right] \\
&\quad \times \left[\left(1 - 2\sum_{s=1}^{r} \lambda_s\right) \sum_{u \le x^{\text{d}}} \mathbf{1}(z^{\text{d}} = u)\right. \\
&\quad \left. + 2\sum_{s=1}^{r} \frac{\lambda_s}{c_s - 1} \sum_{u} \sum_{v \ne u} \mathbf{1}(z^{\text{d}} = u) \mathbf{1}_s(u, v) + O\left(\sum_{s=1}^{r} \lambda_s^2\right)\right] p(z^{\text{d}}) \\
&= F(x^{\text{c}}, x^{\text{d}}) - \alpha_0 \sum_{j=1}^{q} F_j^{(1)}(x^{\text{c}}, x^{\text{d}}) h_j \\
&\quad + \sum_{s=1}^{r}\left[\frac{2}{c_s - 1} \sum_{u} \sum_{v \ne u} \mathbf{1}_s(u, v) F(x^{\text{c}}|u) p(u) - 2F(x^{\text{c}}, x^{\text{d}})\right] \lambda_s \\
&\quad + O\left(\sum_{j=1}^{q} h_j^2 + \sum_{s=1}^{r} \lambda_s^2\right) \\
&= F(x^{\text{c}}, x^{\text{d}}) - \sum_{j=1}^{q} A_{1j} h_j + \sum_{s=1}^{r} C_{2s} \lambda_s + O\left(\sum_{j=1}^{q} h_j^2 + \sum_{s=1}^{r} \lambda_s^2\right)
\end{aligned}$$

where $A_{1j} = \alpha_0 F_j^{(1)}(x^{\text{c}}, x^{\text{d}})$, $C_{2s} = (2/(c_s - 1)) \sum_{u} \sum_{v \ne u} \mathbf{1}_s(u, v) F(x^{\text{c}}|u) p(u) - 2F(x^{\text{c}}, x^{\text{d}})$.

Hence,

$$
\begin{aligned}
\operatorname{var}[\hat{F}(x^{\mathrm{c}}, x^{\mathrm{d}})] &= \frac{1}{n}\operatorname{var}\left[G\left(\frac{x^{\mathrm{c}} - X_i^{\mathrm{c}}}{h}\right)\sum_{u\le x^{\mathrm{d}}} L(X_i^{\mathrm{d}}, u, \lambda)\right] \\
&= \frac{1}{n}\left[E\left[G^2\left(\frac{x^{\mathrm{c}} - X_i^{\mathrm{c}}}{h}\right)\left(\sum_{u\le x^{\mathrm{d}}} L(X_i^{\mathrm{d}}, u, \lambda)\right)^2\right]\right. \\
&\quad \left. - \left[E\left[G\left(\frac{x^{\mathrm{c}} - X_i^{\mathrm{c}}}{h}\right)\sum_{u\le x^{\mathrm{d}}} L(X_i^{\mathrm{d}}, u, \lambda)\right]\right]^2\right] \\
&= \frac{1}{n}\left[F(x^{\mathrm{c}}, x^{\mathrm{d}}) - \sum_{j=1}^{q} A_{1j}h_j + \sum_{s=1}^{r} C_{2s}\lambda_s + O\left(\sum_{j=1}^{q} h_j^2 + \sum_{s=1}^{r}\lambda_s^2\right)\right. \\
&\quad \left. -\left(F(x^{\mathrm{c}}, x^{\mathrm{d}}) + \sum_{j=1}^{q} B_{1j}h_j^2 + \sum_{s=1}^{r} B_{2s}\lambda_s + O\left(\sum_{j=1}^{q} h_j^3 + \sum_{s=1}^{r}\lambda_s^2\right)\right)^2\right] \\
&= \frac{1}{n}F(x^{\mathrm{c}}, x^{\mathrm{d}})(1 - F(x^{\mathrm{c}}, x^{\mathrm{d}})) - \sum_{j=1}^{q} A_{1j}\frac{h_j}{n} \\
&\quad + \sum_{s=1}^{r}(C_{2s} - 2F(x^{\mathrm{c}}, x^{\mathrm{d}})B_{2s})\frac{\lambda_s}{n} \\
&\quad + O\left(\frac{1}{n}\sum_{j=1}^{q} h_j^2 + \frac{1}{n}\sum_{s=1}^{r}\lambda_s^2\right) \\
&= \frac{1}{n}F(x^{\mathrm{c}}, x^{\mathrm{d}})(1 - F(x^{\mathrm{c}}, x^{\mathrm{d}})) - \sum_{j=1}^{q} A_{1j}\frac{h_j}{n} + \sum_{s=1}^{r} C_{2s}\frac{\lambda_s}{n} \\
&\quad + O\left(\frac{1}{n}\sum_{j=1}^{q} h_j^2 + \frac{1}{n}\sum_{s=1}^{r}\lambda_s^2\right) \qquad \text{(A.9)}
\end{aligned}
$$

where $A_{2s} = C_{2s} - 2F(x^{\mathrm{c}}, x^{\mathrm{d}})B_{2s}$.

Using (A.8) and (A.9), we have

$$\begin{aligned}\mathrm{MSE}(\hat{F}(x^{\mathrm{c}},x^{\mathrm{d}})) &= [\mathrm{bias}(\hat{F}(x^{\mathrm{c}},x^{\mathrm{d}}))]^2 + \mathrm{var}(\hat{F}(x^{\mathrm{c}},x^{\mathrm{d}}))\\ &= \left(\sum_{j=1}^{q} B_{1j}h_j^2 + \sum_{s=1}^{r} B_{2s}\lambda_s + O\left(\sum_{j=1}^{q} h_j^3 + \sum_{s=1}^{r}\lambda_s^2\right)\right)^2\\ &\quad + \frac{1}{n}F(x^{\mathrm{c}},x^{\mathrm{d}})(1-F(x^{\mathrm{c}},x^{\mathrm{d}})) - \sum_{j=1}^{q} A_{1j}\frac{h_j}{n} + \sum_{s=1}^{r} A_{2s}\frac{\lambda_s}{n}\\ &\quad + O\left(\frac{1}{n}\sum_{j=1}^{q} h_j^2 + \frac{1}{n}\sum_{s=1}^{r}\lambda_s^2\right) = \frac{1}{n}F(x^{\mathrm{c}},x^{\mathrm{d}})(1-F(x^{\mathrm{c}},x^{\mathrm{d}}))\\ &\quad - \sum_{j=1}^{q} A_{1j}\frac{h_j}{n} + \sum_{s=1}^{r} A_{2s}\frac{\lambda_s}{n} + \left(\sum_{j=1}^{q} B_{1j}h_j^2 + \sum_{s=1}^{r} B_{2s}\lambda_s\right)^2\\ &\quad + O\left(\frac{1}{n}\sum_{j=1}^{q} h_j^2 + \frac{1}{n}\sum_{s=1}^{r}\lambda_s^2 + \sum_{j=1}^{q} h_j^6 + \sum_{s=1}^{r}\lambda_s^4\right)\end{aligned}$$

Thus, we obtain

$$\begin{aligned}\mathrm{MISE}(\hat{F}(x^{\mathrm{c}},x^{\mathrm{d}})) &= \sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int \mathrm{MSE}(\hat{F}(x^{\mathrm{c}},x^{\mathrm{d}}))dx^{\mathrm{c}} = \sum_{x^{\mathrm{d}}\in s^{\mathrm{d}}}\int\left(\frac{1}{n}F(x^{\mathrm{c}},x^{\mathrm{d}})(1-F(x^{\mathrm{c}},x^{\mathrm{d}}))\right.\\ &\quad \left. - \sum_{j=1}^{q} A_{1j}\frac{h_j}{n} + \sum_{s=1}^{r} A_{2s}\frac{\lambda_s}{n} + \left(\sum_{j=1}^{q} B_{1j}h_j^2 + \sum_{s=1}^{r} B_{2s}\lambda_s\right)^2\right)dx^{\mathrm{c}}\\ &\quad + O\left(\frac{1}{n}\sum_{j=1}^{q} h_j^2 + \frac{1}{n}\sum_{s=1}^{r}\lambda_s^2 + \sum_{j=1}^{q} h_j^6 + \sum_{s=1}^{r}\lambda_s^4\right)\\ &= Z^T\left(\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int BB^T dx^{\mathrm{c}}\right)Z + \frac{1}{n}\mathcal{A}^T\tilde{Z}\\ &\quad + \frac{1}{n}\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int F(x^{\mathrm{c}},x^{\mathrm{d}})(1-F(x^{\mathrm{c}},x^{\mathrm{d}}))dx^{\mathrm{c}}\\ &\quad + O\left(\frac{1}{n}\sum_{j=1}^{q} h_j^2 + \frac{1}{n}\sum_{s=1}^{r}\lambda_s^2 + \sum_{j=1}^{q} h_j^6 + \sum_{s=1}^{r}\lambda_s^4\right)\end{aligned}$$

where $Z = (h_1^2, \ldots, h_q^2, \lambda_1, \ldots, \lambda_r)^T$, $\tilde{Z} = (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)^T$, $B = (B_{11}, \ldots, B_{1q}, B_{21}, \ldots, B_{2r})^T$ and $\mathcal{A} = \sum_{x^d \in S^d} \int (-A_{11}, \ldots, -A_{1q}, A_{21}, \ldots, A_{2r})^T dx^c$.

Let $W_i = G((x^c - X_i^c)/h) \sum_{u \le x^d} L(X_i^d, u, \lambda)$. From (A.8), (A.9), and condition (C3), we have

$$\begin{aligned}
&\sqrt{n}\left(\hat{F}(x^c, x^d) - F(x^c, x^d) - \sum_{j=1}^{q} B_{1j} h_j^2 - \sum_{s=1}^{r} B_{2s} \lambda_s\right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[W_i - F(x^c, x^d) - \sum_{j=1}^{q} B_{1j} h_j^2 - \sum_{s=1}^{r} B_{2s} \lambda_s \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [W_i - E(W_i)] + \sqrt{n} O_p\left(\sum_{j=1}^{q} h_j^3 + \sum_{s=1}^{r} \lambda_s^2\right) \\
&\xrightarrow{d} N(0, F(x^c, x^d)(1 - F(x^c, x^d)))
\end{aligned}$$

by Lyapunov's central limit theorem and $\mathrm{var}((1/\sqrt{n}) \sum_{i=1}^{n} [W_i - E(W_i)]) \to F(x^c, x^d)(1 - F(x^c, x^d))$. This completes the proof of Theorem 1.

APPENDIX B. PROOF OF THEOREM 2

Proof of Theorem 2

Recall that

$$\mathrm{CV}(h, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{x^d \in S^d} \int (I(x^c, X_i^c) I(x^d, X_i^d) - \hat{F}_{-i}(x^c, x^d))^2 dx^c \right]$$

where $\hat{F}_{-i}(x^c, x^d) = (1/(n-1)) \sum_{j \ne i} G((x^c - X_j^c)/h) \sum_{u \le x^d} L(X_j^d, u, \lambda)$, $I(x^c, X_i^c) = \mathbf{1}(X_i^c \le x^c)$, and $I(x^d, X_i^d) = \mathbf{1}(X_i^d \le x^d)$.

Let $I_i \equiv I(x, X_i) = I(x^c, X_i^c) I(x^d, X_i^d)$ and $\mathcal{H} = \mathrm{CV}(h, \lambda) - (1/n) \sum_{i=1}^{n} \sum_{x^d \in s^d} \int [I(x, X_i) - F(x^c, x^d)]^2 dx^c$. For simplicity, we use $\hat{F}_{-i}$ and F to denote $\hat{F}_{-i}(x^c, x^d)$ and $F(x^c, x^d)$, respectively, throughout this appendix.

Then we have

$$
\begin{aligned}
n\mathcal{H} &= \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int [(I_i - \hat{F}_{-i})^2 - (I_i - F)^2] dx^{\mathrm{c}} \\
&= \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int \{(\hat{F}_{-i} - F)^2 - 2(I_i - F)(\hat{F}_{-i} - F)\} dx^{\mathrm{c}} \\
&= \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F}_{-i} - F)^2 dx^{\mathrm{c}} - 2 \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (I_i - F)(\hat{F}_{-i} - F) dx^{\mathrm{c}} \\
&\equiv S_1 - 2S_2 \qquad \text{(B.1)}
\end{aligned}
$$

Let $D_i = G((x^{\mathrm{c}} - X_i^{\mathrm{c}})/h) \sum_{u \le x^{\mathrm{d}}} L(X_i^{\mathrm{d}}, u, \lambda) - F(x^{\mathrm{c}}, x^{\mathrm{d}}), \quad D_i^0 = I_i - F$, then

$$
\begin{aligned}
S_1 &= \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F}_{-i} - F)^2 dx^{\mathrm{c}} = \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int \left[\frac{n}{n-1} (\hat{F} - F) - \frac{1}{n-1} D_i \right]^2 dx^{\mathrm{c}} \\
&= \frac{n^3}{(n-1)^2} \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F)^2 dx^{\mathrm{c}} - \frac{2n}{(n-1)^2} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F) D_i dx^{\mathrm{c}} \\
&\quad + (n-1)^{-2} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int D_i^2 dx^{\mathrm{c}} \\
&= \frac{n^3 - 2n^2}{(n-1)^2} \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F)^2 dx^{\mathrm{c}} + \frac{1}{(n-1)^2} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int D_i^2 dx^{\mathrm{c}} \qquad \text{(B.2)}
\end{aligned}
$$

and

$$
\begin{aligned}
S_2 &= \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (I_i - F)(\hat{F}_{-i} - F) dx^{\mathrm{c}} \\
&= \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (I_i - F) \left[\frac{n}{n-1} (\hat{F} - F) - \frac{1}{n-1} D_i \right] dx^{\mathrm{c}} \\
&= \frac{n^2}{n-1} \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int \left[\frac{1}{n} \sum_i I_i - F \right] (\hat{F} - F) dx^{\mathrm{c}} - \frac{1}{n-1} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (I_i - F) D_i dx^{\mathrm{c}} \\
&= \frac{n^2}{n-1} \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (F_n - F)(\hat{F} - F) dx^{\mathrm{c}} - \frac{1}{n-1} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int D_i D_i^0 dx^{\mathrm{c}} \qquad \text{(B.3)}
\end{aligned}
$$

by noting that $F_n \equiv F_n(x^{\mathrm{c}}, x^{\mathrm{d}}) = (1/n) \sum_{i=1}^n I(x^{\mathrm{c}}, X_i^{\mathrm{c}}) I(x^d, X_i^d) \equiv (1/n) \sum_i I_i$.

Combining (B.1), (B.2), and (B.3), we have

$$\begin{aligned}\mathcal{H} &= \frac{1}{n}[S_1 - 2S_2] \\ &= \left(1 - \frac{1}{(n-1)^2}\right) \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F)^2 dx^{\mathrm{c}} + \frac{1}{n(n-1)^2} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int D_i^2 dx^{\mathrm{c}} \\ &\quad - 2\left(1 + \frac{1}{n-1}\right) \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (F_n - F)(\hat{F} - F) dx^{\mathrm{c}} \\ &\quad + \frac{2}{n(n-1)} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int D_i D_i^0 dx^{\mathrm{c}} \end{aligned} \tag{B.4}$$

Let $\mu(h, \lambda) = \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E(D_i D_i^0) dx^{\mathrm{c}}$. Using lemma (B.1) and (B.4), we have that

$$\begin{aligned}\mathcal{H} + \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (F_n - F)^2 dx^{\mathrm{c}} &= \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F_n)^2 dx^{\mathrm{c}} \\ &\quad - \frac{1}{(n-1)^2} \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F)^2 dx^{\mathrm{c}} - \frac{2}{n-1} \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (F_n - F)(\hat{F} - F) dx^{\mathrm{c}} \\ &\quad + \frac{1}{n(n-1)^2} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int D_i^2 dx^{\mathrm{c}} + \frac{2}{n(n-1)} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int D_i D_i^0 dx^{\mathrm{c}} \\ &= \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F_n)^2 dx^{\mathrm{c}} + \frac{2}{n-1} \mu(h, \lambda) \\ &\quad + O_p\left(n^{-3/2} + n^{-1} \sum_{j=1}^{q} h_j^4 + n^{-1} \sum_{s=1}^{r} \lambda_s^2\right) \end{aligned} \tag{B.5}$$

Recall that $W_i = G((x^{\mathrm{c}} - X_i^{\mathrm{c}})/h) \sum_{u \le x^{\mathrm{d}}} L(X_i^{\mathrm{d}}, u, \lambda)$, we have that

$$\begin{aligned}\sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F_n)^2 dx^{\mathrm{c}} &= \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int \left(\frac{1}{n} \sum_{i=1}^{n} W_i - \frac{1}{n} \sum_{i=1}^{n} I_i\right)^2 dx^{\mathrm{c}} \\ &= \frac{1}{n^2} \sum_{i \neq j} \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (W_i - I_i)(W_j - I_j) dx^{\mathrm{c}} \\ &\quad + \frac{1}{n^2} \sum_i \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (W_i - I_i)^2 dx^{\mathrm{c}} \\ &= \frac{1}{n^2} \sum_{i \neq j} g(X_i, X_j) + \frac{1}{n^2} \sum_{i=1}^{n} g(X_i, X_i) = S + T \end{aligned} \tag{B.6}$$

where the definitions of S and T are obvious, and $g(X_i, X_j) = \sum_{x^d \in S^d} \int (W_i - I_i)(W_j - I_j) dx^c$.

We can see that S is a second-order U-statistic. Define $g_1(x) = E[g(x, X_1)]$ and $g_0 = E[g_1(X_1)]$, then we have $g_1(X_i) = E[g(X_i, X_j)|X_i]$ and $g_1(X_j) = E[g(X_i, X_j)|X_j]$, if $i \neq j$. Using the Hoeffding decomposition, we have

$$\begin{aligned} S &= n^{-2} \sum\sum_{i \neq j} g(X_i, X_j) \\ &= n^{-2} \sum\sum_{i \neq j} \{g(X_i, X_j) - g_1(X_j) - g_1(X_j) + g_0\} \\ &\quad + 2\frac{1}{n}\left(1 - \frac{1}{n}\right) \sum_{i=1}^{n} \{g_1(X_i) - g_0\} + \left(1 - \frac{1}{n}\right) g_0 \\ &= S^{(1)} + S^{(2)} + \left(1 - \frac{1}{n}\right) g_0 \end{aligned} \tag{B.7}$$

where the definitions of $S^{(1)}$ and $S^{(2)}$ are obvious.

Then by the law of iterated expectations, we have

$$E(S^{(1)}) = n^{-2} \sum\sum_{i \neq j} [E(g(X_i, X_j)) - E(g_1(X_i)) - E(g_1(X_j)) + g_0] = 0 \tag{B.8}$$

$$E(S^{(2)}) = 2n^{-1}(1 - n^{-1}) \sum_{i=1}^{n} (E(g_1(X_i)) - g_0) = 0 \tag{B.9}$$

Also, it is easy to see that $E[S^{(1)}|X_i] = 0$ for all $i = 1, \ldots, n$ and $E[S^{(2)}|X_j] = 0$ for $j \neq i$, since X_i and X_j are independent. Thus, we have

$$\begin{aligned} E(S^{(1)})^2 &= E\left(n^{-2} \sum\sum_{i \neq j} (g(X_i, X_j) - g_1(X_i) - g_1(X_j) + g_0)\right)^2 \\ &= n^{-4} \sum\sum_{i \neq j} E(g(X_i, X_j) - g_1(X_i) - g_1(X_j) + g_0)^2 \end{aligned} \tag{B.10}$$

and

$$\begin{aligned} E(S^{(2)})^2 &= E\left[2n^{-1}(1 - n^{-1}) \sum_{i=1}^{n} (g_1(X_i) - g_0)\right]^2 \\ &= \frac{4(n-1)^2}{n^4} \sum_{i=1}^{n} E[g_1(X_i) - g_0]^2 \end{aligned} \tag{B.11}$$

From lemma B.2 and (B.10), (B.11), we have

$$E(S^{(1)})^2 = O\left(\frac{1}{n^4}(n^2-n)(E(g(X_i,X_j))^2 + E(g_1(X_1))^2 + g_0^2\right)$$
$$= O\left(n^{-2}\left(\sum_{j=1}^{q} h_j^{3q} + \sum_{j=1}^{q} h_j^{2q+4} + \sum_{j=1}^{q} h_j^8 + \sum_{s=1}^{r} \lambda_s^4\right)\right)$$

and

$$E(S^{(2)})^2 = \frac{4(n-1)^2}{n^4}\sum_{i=1}^{n} E[g_1(X_i) - g_0]^2$$
$$= O\left(n^{-1}\left(\sum_{j=1}^{q} h_j^{2q+4} + \sum_{j=1}^{q} h_j^8 + \sum_{s=1}^{r} \lambda_s^4\right)\right)$$

Also, $E[g(X_1,X_1)]^2 = E[\sum_{x^d\in S^d}\int(W_1 - I_1)^2 dx^c]^2 = O(1)$ implies $\mathrm{Var}(T) = \mathrm{Var}(n^{-2}\sum_i g(X_i,X_i)) = (1/n^3)\mathrm{Var}(g(X_i,X_i)) = O(n^{-3})$.

Combining (B.6) and (B.7), we have $\sum_{x^d\in S^d}\int(\hat{F} - F_n)^2 dx^c = S + T = S^{(1)} + S^{(2)} + (1-n^{-1})g_0 + T$.With (B.8) and (B.9), we have

$$E\left[\sum_{x^d\in S^d}\int(\hat{F} - F_n)^2 dx^c\right] = E(S^{(1)}) + E(S^{(2)}) + (1-n^{-1})g_0 + E(T) = (1-n^{-1})g_0 + E(T) \tag{B.12}$$

Using (B.10), (B.11), and (B.12), we can see that $E(\sum_{x^d\in S^d}\int(\hat{F} - F_n)^2 dx^c - (1-n^{-1})g_0 - E(T))^2 = E[S + T - (1-n^{-1})g_0 - E(T)]^2 = E[S^{(1)} + S^{(2)} + (T - E(T))]^2 = O(E(S^{(1)})^2 + E(S^{(2)})^2 + \mathrm{Var}(T)) = O(n^{-3} + n^{-2}\sum_{j=1}^{q} h_j^{3q} + n^{-1}\sum_{j=1}^{q} h_j^{2q+4} + n^{-1}\sum_{j=1}^{q} h_j^8 + n^{-1}\sum_{s=1}^{r}\lambda_s^4)$. Hence,

$$\sum_{x^d\in S^d}\int(\hat{F} - F_n)^2 dx^c = E(T) + (1-n^{-1})g_0 + O_p\left(n^{-3/2} + n^{-1}\sum_{j=1}^{q} h_j^{3q/2} + n^{-1/2}\sum_{j=1}^{q} h_j^{q+2} + n^{-1/2}\sum_{j=1}^{q} h_j^4 + n^{-1/2}\sum_{s=1}^{r}\lambda_s^2\right) \tag{B.13}$$

Combining (B.5), (B.12), and (B.13), we have

$$\begin{aligned}\mathcal{H} + \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (F_n - F)^2 dx^{\mathrm{c}} &= E(T) + (1 - n^{-1}) g_0 + 2(n-1)^{-1} \mu(h, \lambda) \\ &\quad + O_p\left(n^{-3/2} + n^{-1} \sum_{j=1}^{q} h_j^{3q/2} + n^{-1/2} \sum_{j=1}^{q} h_j^{q+2} + n^{-1/2} \sum_{j=1}^{q} h_j^4 + n^{-1/2} \sum_{s=1}^{r} \lambda_s^2 \right) \\ &= E\left[\sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F_n)^2 dx^{\mathrm{c}} \right] + 2(n-1)^{-1} \mu(h, \lambda) + O_p\left(n^{-3/2} + n^{-1} \sum_{j=1}^{q} h_j^{3q/2} \right. \\ &\quad \left. + n^{-1/2} \sum_{j=1}^{q} h_j^{q+2} + n^{-1/2} \sum_{j=1}^{q} h_j^4 + n^{-1/2} \sum_{s=1}^{r} \lambda_s^2 \right) \end{aligned} \tag{B.14}$$

It is easy to see that

$$\begin{aligned} E\left[\sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F_n)^2 dx^{\mathrm{c}} \right] &= \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E[(\hat{F} - F)^2] dx^{\mathrm{c}} + \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E\left[(F_n - F)^2\right] dx^{\mathrm{c}} \\ &\quad - 2E\left[\sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E(\hat{F} - F)(F_n - F) dx^{\mathrm{c}} \right] \\ &= \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E[(\hat{F} - F)^2] dx^{\mathrm{c}} + \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E[(F_n - F)^2] dx^{\mathrm{c}} \\ &\quad - 2\frac{1}{n} \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E[(W_i - F)(I_i - F)] dx^{\mathrm{c}} \\ &= \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E[(\hat{F} - F)^2] dx^{\mathrm{c}} \\ &\quad + \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E[(F_n - F)^2] dx^{\mathrm{c}} - \frac{2}{n} \mu(h, \lambda) \end{aligned}$$

Also, we have $\mu(h,\lambda) = E[\sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (D_i D_i^0) dx^{\mathrm{c}}] = \sum_{x_1^{\mathrm{d}} \in S^{\mathrm{d}}} \int \{\sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (G(v) \sum_{u \le x^{\mathrm{d}}} L(x_1^{\mathrm{d}}, u, \lambda) - F(x_1^{\mathrm{c}} + hv, x^{\mathrm{d}}))(I(x_1^{\mathrm{c}} + hv, x_1^{\mathrm{c}}) I(x^{\mathrm{d}}, x_1^{\mathrm{d}}) - F(x_1^{\mathrm{c}} + hv, x^{\mathrm{d}})) h dv\} f(x_1^{\mathrm{c}}, x_1^{\mathrm{d}}) dx_1^{\mathrm{c}} = O(\sum_{j=1}^{q} h_j^q)$.

Thus, we have

$$\begin{aligned} E\left[\sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int (\hat{F} - F_n)^2 dx^{\mathrm{c}} \right] &= \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E[(\hat{F} - F)^2] dx^{\mathrm{c}} \\ &\quad + \sum_{x^{\mathrm{d}} \in S^{\mathrm{d}}} \int E[(F_n - F)^2] dx^{\mathrm{c}} + O\left(n^{-1} \sum_{j=1}^{q} h_j^q \right) \end{aligned} \tag{B.15}$$

Combining (B.14) and (B.15), we obtain that

$$\mathcal{H}+\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int(F_n-F)^2dx^{\mathrm{c}}-\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int E[(F_n-F)^2]dx^{\mathrm{c}}=\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int E[(\hat{F}-F)^2]dx^{\mathrm{c}}$$
$$+O_p\left(n^{-3/2}+n^{-1}\sum_{j=1}^{q}h_j^q+n^{-1/2}\sum_{j=1}^{q}h_j^{q+2}+n^{-1/2}\sum_{j=1}^{q}h_j^4+n^{-1/2}\sum_{s=1}^{r}\lambda_s^2\right)$$

That is,

$$\mathrm{CV}(h,\lambda)+J_n=\mathrm{MISE}(h,\lambda)+O_p\left(n^{-3/2}+n^{-1}\sum_{j=1}^{q}h_j^q\right.$$
$$\left.+n^{-1/2}\sum_{j=1}^{q}h_j^{q+2}+n^{-1/2}\sum_{j=1}^{q}h_j^4+n^{-1/2}\sum_{s=1}^{r}\lambda_s^2\right)$$

Essentially, we have proved the upper bound of the second moment of $\mathrm{CV}(h,\lambda)+J_n-\mathrm{MISE}(h,\lambda)$. Using Markov's inequality to the left hand side of (B.16) and Rosenthal's inequality (see Hall & Heyde, 1980, p. 23) to $S^{(1)}$ in (B.7) and repeating the previous proof, we can give the upper bound of each order moment of $\mathrm{CV}(h,\lambda)+J_n-\mathrm{MISE}(h,\lambda)$. With the aid of n^{δ} and the differentiability of the kernel function, we can get

$$P\left\{\sup|\mathrm{CV}(h,\lambda)+J_n-\mathrm{MISE}(h,\lambda)|>\left(n^{-3/2}+n^{-1}\sum_{j=1}^{q}h_j^q\right.\right.$$
$$\left.\left.+n^{-1/2}\sum_{j=1}^{q}h_j^{q+2}+n^{-1/2}\sum_{j=1}^{q}h_j^4+n^{-1/2}\sum_{s=1}^{r}\lambda_s^2\right)n^{\delta}\right\}=O(n^{-\gamma}) \tag{B.16}$$

for arbitrarily large γ. Then by the Borel–Cantelli lemma, we obtain the uniform strong convergence.

This completes the proof of Theorem 2.

Lemma B.1.

(i) $$\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int(\hat{F}-F)^2dx^{\mathrm{c}}+\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int(F_n-F)(\hat{F}-F)dx^{\mathrm{c}}$$
$$=O_p\left(n^{-1}+\sum_{j=1}^{q}h_j^4+\sum_{s=1}^{r}\lambda_s^2\right)$$

(ii) $$n^{-3}\sum_i\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int D_i^2dx^{\mathrm{c}}=O_p(n^{-2})\ \text{ and }\ n^{-2}\sum_i\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int D_iD_i^0dx^{\mathrm{c}}$$
$$=n^{-1}\sum_{x^{\mathrm{d}}\in S^{\mathrm{d}}}\int E(D_iD_i^0)dx^{\mathrm{c}}+O(n^{-3/2}).$$

Proof. From (A.8) and (A.9), we have $\hat{F} - F = O_p(n^{-1/2} + \sum_{j=1}^{q} h_j^2 + \sum_{s=1}^{r} \lambda_s)$. So we have $\sum_{x^d \in S^d} \int (\hat{F} - F)^2 dx^c = O_p(n^{-1} + \sum_{j=1}^{q} h_j^4 + \sum_{s=1}^{r} \lambda_s^2)$. It is easy to see that $E[F_n(x^c, x^d)] = E[I(x, X_i)] = F(x^c, x^d)$ and $\mathrm{Var}(F_n(x^c, x^d)) = n^{-1}\{E[I(x, X_i)]^2 - (E[I(x, X_i)])^2\} = n^{-1}F(x^c, x^d)[1 - F(x^c, x^d)]$. Thus, we have $E[F_n(x^c, x^d) - F(x^c, x^d)]^2 = \mathrm{Var}[F_n(x^c, x^d)] = O(1/n)$, which implies $F_n(x^c, x^d) - F(x^c, x^d) = O_p(n^{-1/2})$ and $\sum_{x^d \in S^d} \int (F_n - F)(\hat{F} - F) dx^c = O(n^{-1} + \sum_{j=1}^{q} h_j^4 + \sum_{s=1}^{r} \lambda_s^2)$.

From the law of large numbers and the central limit theorem, we get that $n^{-1}\sum_i D_i^2 = O_p(1)$ and $n^{-1}\sum_i D_i D_i^0 = E(D_i D_i^0) + O_p(n^{-1/2})$. Therefore, $n^{-3}\sum_i \sum_{x^d \in S^d} \int D_i^2 dx^c = n^{-2}(1/n)\sum_i \sum_{x^d \in S^d} \int D_i^2 dx^c = O_p(n^{-2})$ and $n^{-2}\sum_i \sum_{x^d \in S^d} \int D_i D_i^0 dx^c = n^{-1}\sum_{x^d \in S^d} \int E(D_i D_i^0) dx^c + O_p(n^{-3/2})$.

This completes the proof of this lemma.

Lemma B.2. (i) $E[g(X_1, X_2)^2] = O(\sum_{j=1}^{q} h_j^{3q})$; (ii) $E(g_1(X_1))^2 = O(\sum_{j=1}^{q} h_j^{2q+4} + \sum_{s=1}^{r} \lambda_s^4)$; (iii) $g_0 = O(\sum_{j=1}^{q} h_j^4 + \sum_{s=1}^{r} \lambda_s^2)$.

Proof. Using the change of variables, we have

$$
\begin{aligned}
E[g(X_1, X_2)]^2 = & \sum_{x_1^d \in S^d} \sum_{x_2^d \in S^d} \int \left\{ \sum_{x^d \in S^d} \int \left[G\left(\frac{x^c - x_1^c}{h}\right) \sum_{u \le x^d} L(x_1^d, u, \lambda) \right.\right. \\
& \left. - I(x^c, x_1^c) I(x^d, x_1^d) \right] \\
& \left[G\left(\frac{x^c - x_2^c}{h}\right) \sum_{u \le x^d} L(x_2^d, u, \lambda) - I(x^c, x_2^c) I(x^d, x_2^d) \right] dx^c \Bigg\}^2 \\
& \times f(x_1^c, x_1^d, x_2^c, x_2^d) dx_1^c dx_2^c \\
= & \sum_{x_1^d \in S^d} \sum_{x_2^d \in S^d} \int \left\{ \sum_{x^d \in S^d} \int \left[G(v) \sum_{u \le x^d} L(x_1^d, u, \lambda) - I(x_1^c + hv, x_1^c) I(x^d, x_1^d) \right] \right. \\
& \left. \left[G\left(v + \frac{x_1^c - x_2^c}{h}\right) \sum_{u \le x^d} L(x_2^d, u, \lambda) - I(x_1^c + hv, x_2^c) I(x^d, x_2^d) \right] hdv \right\}^2 \\
& \times f(x_1^c, x_1^d, x_2^c, x_2^d) dx_1^c dx_2^c \\
= & \sum_{x_1^d \in S^d} \sum_{x_2^d \in S^d} \int \left\{ \sum_{x^d \in S^d} \int \left[G(v) \sum_{u \le x^d} L(x_1^d, u, \lambda) - I(hv, 0) I(x^d, x_1^d) \right] \right. \\
& \left. \left[G(v + w) \sum_{u \le x^d} L(x_2^d, u, \lambda) - I(h(v + w), 0) I(x^d, x_2^d) \right] hdv \right\}^2 \\
& \times f(x_2^c + hw, x_1^d, x_2^c, x_2^d) hdw dx_2^c = O\left(\sum_{j=1}^{q} h_j^{3q} \right) \qquad \text{(B.17)}
\end{aligned}
$$

From (A.7) and $E(I_i) = F(x^c, x^d)$, we obtain $E[W_1 - I_1] = O(\sum_{j=1}^{q} h_j^2 + \sum_{s=1}^{r} \lambda_s)$. Then we have

$$
\begin{aligned}
E(g_1(X_1))^2 &= E\left\{ \sum_{x^d \in S^d} \int (W_1 - I_1) E[W_1 - I_1] dx^c \right\}^2 \\
&= (E[W_1 - I_1])^2 \sum_{x_1^d \in S^d} \int \left\{ \sum_{x^d \in S^d} \int \left(G\left(\frac{x^c - x_1^c}{h} \right) \sum_{u \le x^d} L(x_1^d, u, \lambda) \right.\right. \\
&\quad \left.\left. - I(x^c, x_1^c) I(x^d, x_1^d) \right) dx^c \right\}^2 dx_1^c \\
&= O\left(\sum_{j=1}^{q} h_j^4 + \sum_{s=1}^{r} \lambda_s^2 \right) \sum_{x_1^d \in S^d} \int \left\{ \sum_{x^d \in S^d} \int \left(G(v) \sum_{u \le x^d} L(x_1^d, u, \lambda) \right.\right. \\
&\quad \left.\left. - I(x_1^c + hv, x_1^c) I(x^d, x_1^d) \right) h dv \right\}^2 dx_1^c \\
&= O\left(\sum_{j=1}^{q} h_j^{2q+4} + \sum_{s=1}^{r} \lambda_s^4 \right). \qquad \text{(B.18)}
\end{aligned}
$$

It is easy to see that $g_0 = E[g_1(X_1)] = (E[W_1 - I_1])^2 = O(\sum_{j=1}^{q} h_j^4 + \sum_{s=1}^{r} \lambda_s^2)$, which completes the proof.

HIGHER ORDER BIAS REDUCTION OF KERNEL DENSITY AND DENSITY DERIVATIVE ESTIMATION AT BOUNDARY POINTS

Peter Bearse and Paul Rilstone

ABSTRACT

A new, direct method is developed for reducing, to an arbitrary order, the boundary bias of kernel density and density derivative estimators. The basic asymptotic properties of the estimators are derived. Simple examples are provided. A number of simulations are reported, which demonstrate the viability and efficacy of the approach compared to several popular alternatives.

1. INTRODUCTION

Bias reduction in kernel estimation has received considerable attention in the statistics literature. As Jones and Foster (1993, 1996) and Foster (1995) survey, most of the suggestions in this regard can be seen as special cases of

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 319–331

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025013

generalized jackknifing in which linear combinations of kernels are constructed to reduce bias. Several authors have considered bias reduction in the context of a boundary problem. One way of viewing the boundary problem is that the effective support of the kernel becomes truncated so that the kernel neither integrates to one nor do its lower moments vanish as is usually required for bias reduction. Hall and Wehrly (1991) suggested "reflecting" data points around the boundary. This is somewhat ad hoc however and does not necessarily remove bias. Gasser and Müller (1979) have suggested various boundary kernels that mix the kernel with a polynomial constructed so that the mixture has vanishing lower moments. Contrasted with this, say, indirect approach, Rice (1984) suggested a direct method for eliminating the second-order bias using a linear combination of two estimators. However, this approach does not readily generalize for higher order bias reduction. Jones (1993) shows that a linear combination of a kernel and its derivative can also remove the second-order bias. However, he does not consider how to remove higher order bias.

In some situations one may wish to remove bias to a higher order. This may be the case in purely nonparametric estimation procedures when one wants a higher rate of convergence. Also, in some semiparametric problems, the kernel estimator of the nonparametric component of a model is assumed to have the higher order bias removed. For example, this is the case in Klein and Spady (1993) and Bearse, Canals, and Rilstone (2007). In many instances, such as with duration models, boundary problems are the norm rather than the exception.

In this paper we propose an alternative direct approach to higher order bias reduction. In the simulations we have conducted, we find that our approach has distinct advantages over other approaches.

The intuition of our approach is as follows. Let Y_i, $i = 1, \ldots, N$, be i.i.d. random variables whose common density f has support $[0, \infty)$. It is assumed that $f(0) > 0$. We focus on estimating at points close to zero. Right boundary problems and nonzero boundary problems can be dealt with in an analogous fashion. Let $\widehat{f}$ be a standard kernel density estimator:

$$\widehat{f}(y) = \frac{1}{N\gamma} \sum_{i=1}^{N} K\left(\frac{y - Y_i}{\gamma}\right) \tag{1}$$

where K has support $[-1, 1]$. Let $g^{(s)}$ denote the s-order derivative of a function g. Put $g^{|m|} = (g, g^{(1)}, g^{(2)}, \ldots, g^{(m)})^{\mathrm{T}}$, where superscript T indicates transposition. Under standard regularity conditions, it is straightforward

that the expected value of $\hat{f}(y)$ can be derived as

$$\begin{aligned} E[\hat{f}(y)] &= \frac{1}{\gamma} E\left[K\left(\frac{y - Y_i}{\gamma}\right)\right] \\ &= \int_{-1}^{y/\gamma} K(w) f(y - w\gamma) dw \\ &= f(y) \int_{-1}^{y/\gamma} K(w) dw - \gamma f^{(1)}(y) \int_{-1}^{y/\gamma} wK(w) dw \\ &\quad + \cdots + \gamma^s f^{(s)}(y) \int_{-1}^{y/\gamma} K(w) \frac{(-w)^s}{s!} dw + \gamma^{s+1} O(1) \end{aligned} \tag{2}$$

By inspection, the first-order bias can be removed by dividing the usual estimator by $\int_{-1}^{y/\gamma} K(w)dw$. Rice's (1984) proposal of taking a linear combination of two kernel estimators effectively provides a discrete approximation to $f^{(1)}(y)$. This approach can be extended to removing higher order bias, but the resulting estimator is somewhat unwieldy.

Our approach is as follows. By inspection, it is clear that any unbiased estimator of $f^{(1)}(y)$ can be used to remove the bias of $\hat{f}$ to order γ^2. However, the usual kernel density estimator of $f^{(1)}(y)$ is biased in the same manner that $\hat{f}$ is. In fact, the second-order bias of $\hat{f}^{(1)}$ depends on f, $f^{(1)}$, and $f^{(2)}$. More generally, it can be shown that the bias of, say, $\hat{f}^{(j)}$, $j \leq s$ depends on $f, f^{(1)}, \ldots, f^{(s)}$. In Section 2 we show how to construct a linear combination of $\hat{f}$ and its derivatives to obtain an estimator, unbiased to arbitrary order.

To illustrate this in the second-order case we have the following. Put $K^{|1|} = (K, K^{(1)})^T$. By standard manipulations we have

$$E\left[K^{|1|}\left(\frac{y - Y_i}{\gamma}\right)\right] = f(y)\gamma \int_{-1}^{y/\gamma} K^{|1|}(w)dw - \gamma^2 f^{(1)}(y) \int_{-1}^{y/\gamma} K^{|1|}(w)wdw + \gamma^3 O(1) \tag{3}$$

Let

$$Q_2\left(\frac{y}{\gamma}\right) = \int_{-1}^{y/\gamma} K^{|1|}(w)(1, -w)dw, \quad \Gamma_2 \begin{pmatrix} \gamma & 0 \\ 0 & \gamma^2 \end{pmatrix} \tag{4}$$

so that Q_2 is a matrix of incomplete moments of $K^{|1|}$. (Note that, for most kernels used in estimation, this is simply the identity matrix for $y \geq \gamma$.)

It is straightforward that

$$E\left[K^{|1|}\left(\frac{y-Y_i}{\gamma}\right)\right] = Q_2\left(\frac{y}{\gamma}\right)\Gamma_2 f^{|1|}(y) + \gamma^2 O(1) \tag{5}$$

Therefore, with

$$\tilde{f}^{|1|}(y) = \Gamma_2^{-1} Q_2\left(\frac{y}{\gamma}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} K^{|1|}\left(\frac{y-Y_i}{\gamma}\right) \tag{6}$$

we have

$$\begin{aligned} E[\tilde{f}^{|1|}(y)] &= \Gamma_2^{-1} Q_2\left(\frac{y}{\gamma}\right)^{-1} E\left[K^{|1|}\left(\frac{y-Y_i}{\gamma}\right)\right] \\ &= f^{|1|}(y) + \gamma^3 \Gamma^{-1} O(1) \end{aligned} \tag{7}$$

so that, using the first element of this to estimate $f(y)$, we have

$$E[\tilde{f}(y)] = f(y) + \gamma^2 O(1) \tag{8}$$

Also note that the second element of $\tilde{f}^{|1|}(y)$ provides an estimator of $f^{(1)}(y)$, which is also unbiased to order $O(\gamma^2)$.

In the next section we show how bias reduction can be done to arbitrary order. We also derive the pointwise variance and hence get a pointwise rate of convergence. In Section 3 we use a simple simulation to show how the procedure works in practice and compare its performance to unadjusted kernels and boundary kernels. Section 4 concludes the paper.

2. ASYMPTOTIC PROPERTIES AND EXAMPLE

Before stating the estimator, some additional notation is useful. Put

$$W_s(w) = \left(1, -w, \frac{(-w)^2}{2!}, \ldots, \frac{(-w)^{s-1}}{(s-1)!}\right) \tag{9}$$

and

$$\Gamma_s = \mathrm{Diag}(\gamma, \gamma^2, \ldots, \gamma^s) \tag{10}$$

Define an $s \times s$ matrix of partial moments for $K^{|s-1|}$ by

$$Q_s\left(\frac{y}{\gamma}\right) = \int_{-1}^{y/\gamma} K^{|s-1|}(w) W_s(w) dw \tag{11}$$

and define a $1 \times s$ row vector $\iota_0 = (1, 0, \ldots, 0)$. The estimator is thus given by

$$\tilde{f}(y) = \iota_0 \Gamma_s^{-1} Q_s^{-1}\left(\frac{y}{\gamma}\right) \frac{1}{N} \sum_{i=1}^{N} K^{|s-1|}\left(\frac{y - Y_i}{\gamma}\right) \tag{12}$$

We make standard assumptions about the kernel and window width as follows. K is bounded with support $[-1, 1]$; $\int K(w)dw = 1$; and $K(w)$ is s-times differentiable. $K(w)$ is an s-order kernel such that, for some $s \geq 1$, $\int w^m K(w)dw = 0$ for $m = 1, \ldots, s-1$ and $\int |w|^s |K(w)|dw < \infty$. The window width sequence satisfies $\lim_{N\to\infty}\gamma = 0$ and $\lim_{N\to\infty} N\gamma = \infty$; Q_s is nonsingular; and the elements of Q_s^{-1} are finite.

Proposition 1. Suppose that $f(y)$ is differentiable to order s, and these derivatives are uniformly bounded. Then, uniformly in $y \geq 0$,

$$E[\tilde{f}(y)] - f(y) = O(\gamma^s)$$

Proof. Using a change of variables and an sth-order Taylor series expansion of f we have,

$$\begin{aligned}
&E\left[K^{|s-1|}\left(\frac{y - Y_i}{\gamma}\right)\right] \\
&\quad = \int_0^\infty K^{|s-1|}\left(\frac{y - Y_i}{\gamma}\right) f(Y_i)dY_i \\
&\quad = \gamma \int_{-1}^{y/\gamma} K^{|s-1|}(w) f(y - w\gamma)dw \\
&\quad = \gamma \int_{-1}^{y/\gamma} K^{|s-1|}(w)\left[\sum_{k=0}^{s-1} f^{(k)}(y)\frac{(-w\gamma)^k}{k!} + f^{(s)}(\bar{y})\frac{(-w\gamma)^s}{s!}\right] dw \\
&\quad = Q_s\left(\frac{y}{\gamma}\right)\Gamma_s f^{|s-1|}(y) + \gamma^{s+1}\int_{-1}^{y/\gamma} K^{|s-1|}(w) f^{(s)}(\bar{y})\frac{(-w)^s}{s!}dw
\end{aligned} \tag{13}$$

where $\bar{y}$ is a mean value.[1] Since $Q_s^{-1}(y/\gamma)$, $K^{|s-1|}$, and $f^{(s)}$ are bounded, we have

$$\left| E\left[\iota_0 \Gamma_s^{-1} Q_s\left(\frac{y}{\gamma}\right)^{-1} K^{|s-1|}\left(\frac{y - Y_i}{\gamma}\right)\right] - f(y)\right| \leq \gamma^s C \tag{14}$$

uniformly in $y \geq 0$. QED

Put

$$M_N = Q_s\left(\frac{y}{\gamma}\right)^{-1} E\left[K^{|s-1|}\left(\frac{y-Y_i}{\gamma}\right)K^{|s-1|}\left(\frac{y-Y_i}{\gamma}\right)^T\right]\left(Q_s\left(\frac{y}{\gamma}\right)^{-1}\right)^T$$
$$= \gamma f(y) Q_s\left(\frac{y}{\gamma}\right)^{-1} \int_{-1}^{y/\gamma} \left[K^{|s-1|}(w)K^{|s-1|}(w)^T\right]dw \left(Q_s\left(\frac{y}{\gamma}\right)^{-1}\right)^T + o(\gamma) \quad (15)$$

The variance of the estimator is given by the following result.

Proposition 2. Suppose that $f(y)$ is differentiable to order s, and these derivatives are uniformly bounded. Then,

$$\mathrm{Var}[\tilde{f}(y)] = \frac{1}{N_\gamma} f(y) \int_{-1}^{1} K(w)^2 dw + o\left(\frac{1}{N\gamma}\right)$$

Proof. The result follows by standard change of variables as follows.

$$\mathrm{Var}[\tilde{f}^{[s-1]}(y)]$$
$$= \frac{1}{N}\Gamma_s^{-1} Q_s\left(\frac{y}{\gamma}\right)^{-1} \mathrm{Var}\left[K^{|s-1|}\left(\frac{y-Y_i}{\gamma}\right)\right]\left(Q_s\left(\frac{y}{\gamma}\right)^{-1}\right)^T \Gamma_s^{-1}$$
$$= \frac{1}{N}\Gamma_s^{-1} M_N \Gamma_s^{-1} + \Gamma_s^{-1} o\left(\frac{\gamma}{N}\right)\Gamma_s^{-1} \quad (16)$$

Note that $\iota_0\Gamma^{-1} = \gamma^{-1}\iota_0$. From the property that K is an sth-order kernel, the first row of $Q_s(1)$ is ι_0. $Q_s(1)^{-1}$ has the same property. (This is easily shown using the properties of partitioned matrices.) Hence, $\lim_{N\to\infty} \iota_0 Q_s(y/\gamma)^{-1} = \iota_0$ and

$$\mathrm{Var}[\iota_0 \tilde{f}^{|s-1|}(y)] = \frac{1}{N}\gamma^{-1} f(y)\iota_0 \int_{-1}^{1} [K^{|s-1|}(w)K^{|s-1|}(w)^T]dw\iota_0' + o\left(\frac{1}{N\gamma}\right)$$

and the result follows. QED

Remarks.

1. Note that the asymptotic variance is the same as the usual formula when there is no boundary issue. It may be possible to get a more accurate measure of dispersion by using $Q_s(y/\gamma)$ in the calculations.

2. By inspection, the bias and variance vanish as $N \to \infty$, and so $\tilde{f}(y)$ is consistent in mean squared error (MSE) and probability. Also by inspection, the rate of convergence in MSE is given by $\sqrt{\gamma^s + (N\gamma)^{-1}}$.
3. A biased reduced estimator of the jth derivative of $f(y)$ is provided by $\tilde{f}^{(j)}(y) = \iota_{j+1}\tilde{f}^{|s-1|}(y)$ where $\iota_{j+1} = (0, \ldots, 0, 1, 0, \ldots, 0)$ and the one is the $j+1$'th element of ι_{j+1}. It follows that the bias of $\tilde{f}^{(j)}(y)$ is of order $O(\gamma^s)$. It is also straightforward to confirm that $\mathrm{Var}[\tilde{f}^{(j)}(y)] = O((N\gamma^{1+2j})^{-1})$. The rate of convergence in MSE is given by $\sqrt{\gamma^s + (N\gamma^{1+2j})^{-1}}$.
4. Since each of the estimators are linear combinations of averages, it follows that the estimators, appropriately normalized (and with the appropriate conditions on the window width) are asymptotically normal in distribution.
5. One might be interested in estimation at a sequence of points $y_N \to c$ where, for example, c could be the boundary point. We note that our results are pointwise and stronger conditions may be necessary to derive the properties of, say, $\tilde{f}^{|s-1|}(y_N)$ in this case.

Consider a specific example using the Epanicheknikov kernel with $s = 1, 2$:

$$K(w) = \frac{3}{4}(1 - w^2)1[|w| < 1] \tag{17}$$

$$K^{|1|}(w) = \begin{pmatrix} \frac{3}{4}(1 - w^2) \\ -\frac{3}{2}w \end{pmatrix} 1[|w| \leq 1] \tag{18}$$

$$\begin{aligned} Q_1\left(\frac{y}{\gamma}\right) &= \int_{-1}^{y/\gamma} K^{(0)}(w)dw \\ &= \int_{-1}^{y/\gamma} \frac{3}{4}(1 - w^2)dw \\ &= \frac{3}{4}\left[\left(\frac{y}{\gamma} - \frac{1}{3}\left(\frac{y}{\gamma}\right)^3\right) - \left(-1 - \frac{1}{3}(-1)^3\right)\right] \\ &= \frac{3}{4}\left[\frac{2}{3} + \left(\frac{y}{\gamma} - \frac{1}{3}\left(\frac{y}{\gamma}\right)^3\right)\right] \end{aligned} \tag{19}$$

Note, for $y \geq \gamma$, $Q_1(1) = 1$.

$$
\begin{aligned}
Q_2\left(\frac{y}{\gamma}\right) &= \int_{-1}^{y/\gamma} \begin{pmatrix} K(w) \\ K^{(1)}(w) \end{pmatrix} (1-w)dw \\
&= \int_{-1}^{y/\gamma} \begin{pmatrix} \frac{3}{4}(1-w^2) & -\frac{3}{4}(w-w^3) \\ -\frac{3}{2}w & \frac{3}{2}w^2 \end{pmatrix} dw \\
&= \begin{pmatrix} \frac{3}{4}\left(w - \frac{w^3}{3}\right) & -\frac{3}{4}\left(\frac{w^2}{2} - \frac{w^4}{4}\right) \\ -\frac{3w^2}{4} & \frac{3}{6}w^3 \end{pmatrix} \Bigg|_{-1}^{y/\gamma} \\
&= \begin{pmatrix} \frac{3}{4}\left[\frac{2}{3} + \left(\frac{y}{\gamma} - \frac{1}{3}\left(\frac{y}{\gamma}\right)^3\right)\right] & \frac{3}{4}\left[\frac{1}{4} - \left(\frac{1}{2}\left(\frac{y}{\gamma}\right)^2 - \frac{1}{4}\left(\frac{y}{\gamma}\right)^4\right)\right] \\ \frac{3}{4}\left(1 - \left(\frac{y}{\gamma}\right)^2\right) & \frac{1}{2}\left(1 + \left(\frac{y}{\gamma}\right)^3\right) \end{pmatrix}
\end{aligned} \tag{20}
$$

Note, for $y \geq \gamma$, $Q_2(1)$ is simply the 2×2 identity matrix.

3. MONTE CARLO STUDY

Here we examine the performance of our bias reducing density estimation approach in the context of a small-scale Monte Carlo experiment. We construct the data Y_i, $i = 1, \ldots, N$, from an exponential (1) distribution implying a left boundary of zero. We evaluate the performance of each density estimator over a mesh of 101 equally spaced points in the boundary region $[0, \gamma]$, where $\gamma \equiv \gamma(N, K)$ is the smoothing parameter which is a function of both the sample size and the underlying kernel K. We use sample sizes $N = 50$, 100, 200, and 500. We consider two kernels:[2] the quadratic kernel

$$K_2(w) = \frac{3}{4}(1-w^2)I_{(-1,1)}(w) \tag{21}$$

and the quartic kernel

$$K_4(w) = \frac{15}{32}(3 - 10w^2 + 7w^4)I_{(-1,1)}(w) \tag{22}$$

where $I_{(-1,1)}(w)$ is an indicator taking the value 1 if $w \in (-1, 1)$, and zero otherwise. Each simulation is based on 500 replications (Tables 1 and 2).[3]

For a given kernel function K, we denote our bias reducing density estimator with order of bias reduction s by $\tilde{f}_s$. For the case of K_2 we consider $s = 1, 2, 3$ while for K_4 we consider $s = 1, 2, 3, 4, 5$.

For comparative purposes we also consider the typical fixed bandwidth density estimator

$$\hat{f}(y) = \frac{1}{N\gamma}\sum_{i=1}^{N} K\left(\frac{y - Y_i}{\gamma}\right) \tag{23}$$

and the adaptive density estimator

$$f_A(y) = \frac{1}{N\gamma}\sum_{i=1}^{N} \frac{1}{\lambda_i} K\left(\frac{y - Y_i}{\gamma\lambda_i}\right) \tag{24}$$

Table 1. Performance in the Boundary Region: Quadratic Kernel.

N	γ	$\hat{f}$	f_A	$\tilde{f}_1$	$\tilde{f}_2$	$\tilde{f}_3$	f_{GM}
Average bias							
50	0.9029	−0.0804	−0.0603	−0.0653	0.0102	0.0014	0.0183
100	0.7860	−0.0800	−0.0612	−0.0623	0.0084	0.0004	0.0138
200	0.6843	−0.0802	−0.0609	−0.0588	0.0065	−0.0008	0.0110
500	0.5697	−0.0786	−0.0587	−0.0514	0.0070	0.0015	0.0108
Average variance							
50	0.9029	0.0086	0.0197	0.0039	0.0116	0.0217	0.0139
100	0.7860	0.0056	0.0119	0.0027	0.0076	0.0145	0.0090
200	0.6843	0.0035	0.0074	0.0018	0.0053	0.0053	0.0053
500	0.5697	0.0019	0.00	0.0011	0.0026	0.0044	0.0029
Average MSE							
50	0.9029	0.0468	0.0580	0.0167	0.0124	0.0217	0.0145
100	0.7860	0.0434	0.0510	0.0132	0.0081	0.0145	0.0095
200	0.6843	0.0406	0.0452	0.0105	0.0049	0.0085	0.0056
500	0.5697	0.0380	0.0408	0.0075	0.0028	0.0044	0.0031

Table 2. Performance in the Boundary Region: Quartic Kernel.

N	γ	$\widehat{f}$	f_A	$\tilde{f}_1$	$\tilde{f}_2$	$\tilde{f}_3$	$\tilde{f}_4$	$\tilde{f}_5$	f_{GM}
Average bias									
50	2.2680	−0.0686	−0.0538	−0.0140	0.0056	0.0062	−0.0063	−0.0005	−0.0053
100	2.0999	−0.0700	−0.0507	−0.0139	0.0050	0.0057	−0.0048	−0.0003	−0.0047
200	1.9442	−0.0710	−0.0472	−0.0132	0.0047	0.0051	−0.0030	−0.0002	−0.0043
500	1.7560	−0.0725	−0.0437	−0.0125	0.0045	0.0045	−0.0021	−0.0002	−0.0024
Average variance									
50	2.2680	0.0010	0.0020	0.0015	0.0027	0.0040	0.0552	0.0116	0.0488
100	2.0999	0.0006	0.0013	0.0009	0.0017	0.0024	0.0638	0.0071	0.0298
200	1.9442	0.0004	0.0008	0.0006	0.0010	0.0014	0.0192	0.0038	0.0168
500	1.7560	0.0002	0.0005	0.0003	0.0005	0.0007	0.0082	0.0018	0.0070
Average MSE									
50	2.2680	0.0437	0.0356	0.0110	0.0043	0.0044	0.0563	0.0116	0.0502
100	2.0999	0.0431	0.0351	0.0097	0.0029	0.0027	0.0372	0.0071	0.0305
200	1.9442	0.0426	0.0354	0.0085	0.0021	0.0016	0.0194	0.0038	0.0173
500	1.7560	0.0416	0.0365	0.0070	0.0012	0.0008	0.0083	0.0018	0.0072

where λ_i is a local bandwidth factor given by[4]

$$\lambda_i = \left[\frac{\widehat{f}(Y_i)}{\exp\left(\frac{1}{N} \sum_{i=1}^{N} \log \widehat{f}(Y_i) \right)} \right]^{0.5} \tag{25}$$

Since $\widehat{f}$ is not designed to perform well in finite samples with bounded data, we also consider an alternative that was designed for this case. In particular, we consider the boundary kernel approach of Gasser and Müller (1979). Let K be a kth-order polynomial kernel with support $[-1, 1]$. In our context where the data has a left boundary of zero, the Gasser–Muller boundary kernel can then be written as

$$f_{GM}(y) = \frac{1}{N\gamma} \sum_{i=1}^{N} K_q\left(\frac{y - Y_i}{\gamma} \right) \quad y \in [0, \gamma] \tag{26}$$

where

$$K_q(w) = (c_{0,q} + c_{1,q} w + \cdots + c_{k-1,q} w^{k-1}) K(w) I_{(-1,q)}(w) \tag{27}$$

is the "boundary kernel"; $q(y/\gamma) = \min\{1, (y/\gamma)\}$; and $c_{0,q}, c_{1,q}, \ldots, c_{k-1,q}$ are chosen to ensure that

$$\int_{-1}^{q(y/\gamma)} K_q(w)dw = 1$$
$$\int_{-1}^{q(y/\gamma)} w^j K_q(w)dw = \begin{cases} 0 & j = 1, 2, \ldots, k-1 \\ C < \infty & j = k \end{cases} \tag{28}$$

at each point y in the boundary region $[0, \gamma]$ where the density is estimated. Thus, the boundary kernel approach adjusts the kernel weights to ensure that the weighting function used in the boundary region satisfies the same moment restrictions as the kth-order kernel. Note that when $y > \gamma$, $c_{0,q} = 1$ and $c_{1,q} = 0$, $c_{2,q} = 0$, ... $c_{k-1,g} = 0$ so that $f_{\mathrm{GM}}(y)$ reduces to $\widehat{f}(y)$ for all points outside the boundary.

For each sample size and each optimal kernel, we choose the bandwidth γ to minimize asymptotic mean integrated squared error of the fixed bandwidth kernel density estimator $\tilde{f}$ under exponential (1) data.[5] While we could consider choosing γ optimally for each density estimator, this would pose some problems for interpreting the results since the boundary region itself varies with γ.

The results for the third-order bias reduction are mixed. Our proposed estimator dominates in that case in bias, but not overall in MSE. However, with the fifth-order bias reduction our proposed estimator clearly dominates the others in terms of bias error and MSE.

4. CONCLUSION

A method has been developed for boundary bias reduction of a variety of kernel estimators. These estimators are simple to compute, their asymptotic properties are comparable to the usual kernel estimators outside the boundary region, and they performed well in the simulations we conducted. There are a number of possible modifications possible to the approach, such as varying the window width for derivative estimation and by using pointwise optimal bandwidths. Another alternative is Loader's (1996) local likelihood estimator. Preliminary results applying local likelihood to the models in Section 3 were not promising.[6] Variations to specialize local likelihood for derivative estimation may yield better results. We intend to explore these alternatives in future work.

NOTES

1. This is a slight abuse of notation. Since Eq. (13) actually represents a vector of Taylor series expansions, each of the remainder terms is evaluated at possibly different points or "mean values" between y and $y-w\gamma$.
2. See Gasser, Müller, and Mammitzsch (1985, p. 243); Table 1.
3. As summary descriptive statistics, for each estimator, we calculated its empirical (over the 500 replications) bias, variance, and MSE at each of the 101 grid points. Tables 1 and 2 report the average of these over the 101 grid points.
4. See Abramson (1982) and Silverman (1986). Klein and Spady (1993) use f_A as an alternative to explicit higher order bias reduction.
5. Note from Tables 1 and 2 that this can result in large boundary regions covering areas with substantial probabilities. This underscores the potential significance of the boundary issue. We thank a referee for pointing this out.
6. We thank J. Racine for this insight.

ACKNOWLEDGMENTS

Funding for Rilstone was provided by the Social Sciences and Humanities Research Council of Canada. The authors thank two anonymous referees for their comments.

REFERENCES

Abramson, I. S. (1982). On bandwidth variation in Kernel estimates: A square root law. *Annals of Statistics*, *10*, 1217–1223.

Bearse, P., Canals, J., & Rilstone, P. (2007). Efficient semiparametric estimation of duration models with unobserved heterogeneity. *Econometric Theory*, *23*, 281–308.

Foster, P. (1995). A comparative study of some bias correction techniques for kernel-based density estimators. *Journal of Statistical Computation and Simulation*, *51*, 137–152.

Gasser, T., & Müller, H. G. (1979). Kernel estimation of regression functions. In: T. Gasser & M. Rosenblatt (Eds), *Smoothing techniques for curve estimation* (pp. 23–68). Lecture Notes in Mathematics No. 757. Berlin: Springer.

Gasser, T., Müller, H. G., & Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, *47*, 238–252.

Hall, P., & Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, *86*, 665–672.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing*, *3*, 135–146.

Jones, M. C., & Foster, P. J. (1993). Generalized jackknifing and higher order kernels. *Nonparametric Statistics*, *3*, 81–94.

Jones, M. C., & Foster, P. J. (1996). A simple nonnegative boundary correction method for kernel density estimators. *Statistica Sinica*, *6*, 1005–1013.

Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, *61*, 387–422.

Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics*, *24*, 1602–1618.

Rice, J. (1984). Boundary modification for kernel regression. *Communications in Statistics, Part A Theory and Methods*, *13*, 893–900.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

PART V
COMPUTATION

NONPARAMETRIC AND SEMIPARAMETRIC METHODS IN R

Jeffrey S. Racine

ABSTRACT

The R environment for statistical computing and graphics (R Development Core Team, 2008) offers practitioners a rich set of statistical methods ranging from random number generation and optimization methods through regression, panel data, and time series methods, by way of illustration. The standard R distribution (base R) comes preloaded with a rich variety of functionality useful for applied econometricians. This functionality is enhanced by user-supplied packages made available via R servers that are mirrored around the world. Of interest in this chapter are methods for estimating nonparametric and semiparametric models. We summarize many of the facilities in R and consider some tools that might be of interest to those wishing to work with nonparametric methods who want to avoid resorting to programming in C or Fortran but need the speed of compiled code as opposed to interpreted code such as Gauss or Matlab by way of example. We encourage those working in the field to strongly consider implementing their methods in the R environment thereby making their work accessible to the widest possible audience via an open collaborative forum.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 335–375

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025014

1. INTRODUCTION

Unlike their more established parametric counterparts, many nonparametric and semiparametric methods that have received widespread theoretical treatment have not yet found their way into mainstream commercial packages. This has hindered their adoption by applied researchers, and it is safe to describe the availability of modern nonparametric methods as fragmented at best, which can be frustrating for users who wish to assess whether or not such methods can add value to their application. Thus, one frequently heard complaint about the state of nonparametric kernel methods concerns the lack of software along with the fact that implementations in interpreted[1] environments such as Gauss are orders of magnitude slower than compiled[2] implementations written in C or Fortran. Though many researchers may code their methods, often using interpreted environments such as Gauss, it is fair to characterize much of this code as neither designed nor suited as tools for general-purpose use as they are typically written solely to demonstrate "proof of concept." Even though many authors are more than happy to circulate such code (which is of course appreciated!), this often imposes certain hardships on the user including (1) having to purchase a (closed and proprietary) commercial software package and (2) having to modify the code substantially in order to use it for their application.

The R environment for statistical computing and graphics (R Development Core Team, 2008) offers practitioners a range of tools for estimating nonparametric, semiparametric, and of course parametric models. Unlike many commercial programs, which must first be purchased in order to evaluate them, you can adopt R with minimal effort and with no financial outlay required. Many nonparametric methods are well documented, tested, and are suitable for general use via a common interface[3] structure (such as the "formula" interface) making it easy for users familiar with R to deploy these tools for their particular application. Furthermore, one of the strengths of R is the ability to call compiled C or Fortran code via a common interface structure thereby delivering the speed of complied code in a flexible and easy-to-use environment. In addition, there exist a number of R "packages" (often called "libraries" or "modules" in other environments) that implement a variety of kernel methods, albeit with varying degrees of functionality (e.g., univariate vs. multivariate, the ability/inability to handle numerical and categorical data, and so forth). Finally, R delivers a rich framework for implementing and making code available to the community.

In this chapter, we outline many of the functions and packages available in R that might be of interest to practitioners, and consider some illustrative applications along with code fragments that might be of interest. Before proceeding further, we first begin with an introduction to the R environment itself.

2. THE R ENVIRONMENT

What is R? Perhaps, it is best to begin with the question "what is S"? S is a language and environment designed for statistical computing and graphics which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies). S has grown to become the de facto standard among econometricians and statisticians, and there are two main implementations, the commercial implementation called "S-PLUS" and the free, open-source implementation called "R." R delivers a rich array of statistical methods, and one of its strengths is the ease with which "packages" can be developed and made available to users for free. R is a mature open platform[4] that is ideally suited to the task of making one's method available to the widest possible user base free of charge.

In this section, we briefly describe a handful of resources available to those interested in using R, introduce the user to the R environment, and introduce the user to the `foreign` package that facilitates importation of data from packages such as SAS, SPSS, Stata, and Minitab, among others.

2.1. Web Sites

A number of sites are devoted to helping R users, and we briefly mention a few of them below:

http://www.R-project.org/: This is the R home page from which you can download the program itself and many R packages. There are also manuals, other links, and facilities for joining various R mailing lists.
http://CRAN.R-project.org/: This is the "Comprehensive R Archive Network," "a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for the R statistical package." Packages are only put on CRAN when they pass a rather stringent collection of quality assurance checks, and in particular are guaranteed to build and run on standard platforms.

http://cran.r-project.org/web/views/Econometrics.html: This is the CRAN "task view" for computational econometrics. "Base R ships with a lot of functionality useful for computational econometrics, in particular in the stats package. This functionality is complemented by many packages on CRAN, a brief overview is given below." This provides an excellent summary of both parametric and nonparametric packages that exist for the R environment.
http://pj.freefaculty.org/R/Rtips.html: This site provides a large and excellent collection of R tips.

2.2. Getting Started with R

A number of well-written manuals exist for R and can be located at the R web site. This section is clearly not intended to be a substitute for these resources. It simply provides a minimal set of commands which will aid those who have never used R before.

Having installed and run R, you will find yourself at the > prompt. To quit the program, simply type `q()`. To get help, you can either enter a command preceded by a question mark, as in `?help`, or type `help.start()` at the > prompt. The latter will spawn your web browser (it reads files from your hard drive, so you do not have to be connected to the Internet to use this feature).

You can enter commands interactively at the R prompt, or you can create a text file containing the commands and execute all commands in the file from the R prompt by typing `source(''commands.R'')`, where `commands.R` is the text file containing your commands. Many editors recognize the .R extension providing a useful interface for the development of R code. For example, GNU Emacs is a powerful editor that works well with R and also LaTeX (http://www.gnu.org/software/emacs/emacs.html).

When you quit by entering the `q()` command, you will be asked whether or not you wish to save the current session. If you enter Y, then the next time you run R *in the same directory* it will load all of the objects created in the previous session. If you do so, typing the command `ls()` will list all of the objects. For this reason, it is wise to use different directories for different projects. To remove objects that have been loaded, you can use the command `rm(objectname)` or `rm(list = ls())` which will remove all objects in memory.

2.3. Importing Data from Other Formats

The `foreign` package allows you to read data created by different popular programs. To load it, simply type `library (foreign)` from within `R`. Supported formats include:

read.arff: Read Data from ARFF Files
read.dbf: Read a DBF File
read.dta: Read Stata Binary Files
read.epiinfo: Read Epi Info Data Files
read.mtp: Read a Minitab Portable Worksheet
read.octave: Read Octave Text Data Files
read.S: Read an S3 Binary or data.dump File
read.spss: Read an SPSS Data File
read.ssd: Obtain a Data Frame from a SAS Permanent Dataset, via read.xport
read.systat: Obtain a Data Frame from a Systat File
read.xport: Read a SAS XPORT Format Library

The following code snippet reads the Stata file "mroz.dta" directly from one's working directory (Carter Hill, Griffiths, & Lim, 2008) and lists the names of variables in the data frame.

```
R>    library(foreign)
R>    Mydat <- read.dta(file = ''mroz.dta'')
R>    names(mydat)
```

```
[1] ''taxableinc''''federaltax''''hsiblings''   ''hfathereduc''''hmothereduc''
[6] ''siblings''  ''lfp''       ''hours''       ''kids16''     ''kids618''
[11]''age''       ''educ''      ''wage''        ''wage76''     ''hhours''
[16]''hage''      ''heduc''     ''hwage''       ''faminc''     ''mtr''
[21]''mothereduc''''fathereduc''''unemployment''''largecity''  ''exper''
```

Alternatively, you might wish to read your Stata file directly from the Internet, as in

```
R> Mydat <- read.dta(file = ''http://www.principlesofeconometrics.
     com/stata/mroz.dta'')
```

Clearly `R` makes it simple to migrate data from one environment to another.

Having installed `R` and having read in data from a text file or supported format such as a Stata binary file, you can then install packages via the

`install.packages()` command, as in `install.packages(''np'')` which will install the `np` package (Hayfield & Racine, 2008) that we discuss shortly.

3. BASIC PARAMETRIC ESTIMATION IN R

Before proceeding, we demonstrate some basic capabilities of R via three examples, namely multiple linear regression, logistic regression, and a simple Monte Carlo simulation.

By way of example, we consider Wooldridge's (2002) "wage1" dataset ($n = 526$) that is included in the np package and estimate an earnings equation.

Variables are defined as follows:

(1) "lwage" log (wage);
(2) "female" ("Female" if female, "Male" otherwise);
(3) "married" ("Married" if married, "Nonmarried" otherwise);
(4) "educ" years of education;
(5) "exper" years of potential experience; and
(6) "tenure" years with current employer.

```
R>  library(np)
Nonparametric Kernel Methods for Mixed Data types (version 0.30-3)
R>  data(wage1)
R>  model.lm <- lm(lwage~factor(female)+
+       factor(married)+
+       educ+
+       tenure+
+       exper+
+       expersq,
+       data = wage1)

R>  summary(model.lm)


Call:

lm(formula = lwage~factor(female)+factor(married)+educ+
  tenure+exper+expersq, data = wage1)


Residuals:

   Min             1Q         Median             3Q            Max
-1.8185        -0.2568        -0.0253         0.2475         1.1815
```

```
Coefficients:
                          Estimate       SE      t-value    Pr(>|t|)
(Intercept)              0.181161    0.107075   1.69       0.091.
factor (female)Male      0.291130    0.036283   8.02       6.9e-15***
factor (married)        −0.056449    0.040926  −1.38       0.168
  Notmarried
educ                     0.079832    0.006827   11.69      <2e-16***
tenure                   0.016074    0.002880   5.58       3.9e-08***
exper                    0.030100    0.005193   5.80       1.2e-08***
expersq                 −0.000601    0.000110  −5.47       7.0e-08***
---

Significant Codes: 0'***' 0.001'**' 0.01'*' 0.05'.' 0.1' ' 1

Residual SE: 0.401 on 519 Degrees of Freedom (df)

Multiple R^2: 0.436                    Adjusted R^2: 0.43
F-statistic: 66.9 on 6 and 519 df      p-value: <2e-16
```

For the next example, we use data on birthweights taken from the R MASS library (Venables & Ripley, 2002), and compute a parametric logit model. We also construct a confusion matrix[5] and assess the model's classification ability. The outcome is an indicator of low infant birthweight (0/1). This application has $n = 189$ and 7 regressors.

Variables are defined as follows:

(1) "low" indicator of birthweight less than 2.5 kg;
(2) "smoke" smoking status during pregnancy;
(3) "race" mother's race ("1" = white, "2" = black, "3" = other);
(4) "ht" history of hypertension;
(5) "ui" presence of uterine irritability;
(6) "ftv" number of physician visits during the first trimester;
(7) "age" mother's age in years; and
(8) "lwt" mother's weight in pounds at last menstrual period.

Note that all variables other than age and lwt are categorical in nature in this example:

```
R>  data(''birthwt'',package = ''MASS'')
R>  attach(birthwt)
R>  model.logit <- glm(low~factor(smoke)+
+       factor(race)+
+       factor(ht)+
```

```
+       factor(ui)+
+       ordered(ftv)+
+       age+
+       lwt,
+       family = binomial(link = logit))
R>  cm <- table (low, ifelse(fitted(model.logit) > 0.5, 1, 0))
R>  ccr <- sum(diag(cm))/sum(cm)
R>  summary(model.logit)

Call:

glm(formula = low ~ factor(smoke)+factor(race)+factor(ht)
  +factor(ui)+ordered(ftv)+age+lwt, family = binomial(link
  = logit))

Deviance Residuals:

  Min       1Q      Median      3Q      Max
−1.707    −0.843    −0.508    0.975    2.146

Coefficients:

                    Estimate        SE        z-value     Pr(>|z|)
(Intercept)        −1.64947    147.13066     −0.01        0.991
factor(smoke)1      1.00001      0.41072      2.43        0.015*
factor(race)2       1.26760      0.53357      2.38        0.018*
Factor(race)3       0.91040      0.45142      2.02        0.044*
factor(ht)1         1.79128      0.70990      2.52        0.012*
factor(ui)1         0.89534      0.45108      1.98        0.047*
ordered(ftv).L     −7.22342    527.54069     −0.01        0.989
ordered(ftv).Q     −7.16294    481.57657     −0.01        0.988
ordered(ftv).C     −5.15187    328.98002     −0.02        0.988
ordered(ftv)^4     −2.06949    166.82485     −0.01        0.990
ordered(ftv)^5     −0.27780     55.61212     −0.005       0.996
Age                −0.01683      0.03627     −0.46        0.643
Lwt                −0.01521      0.00717     −2.12        0.034*
---

Significant Codes: 0‘***’ 0.001‘**’ 0.01‘*’ 0.05‘.’ 0.1‘ ’ 1


(Dispersion parameter for binomial family taken to be 1)

  Null Deviance   : 234.67 on 188 df
Residual Deviance: 202.21 on 176 df
AIC: 228.2
```

```
Number of Fisher Scoring Iterations: 13

R>    cm

low   0                  1
0     119                11
1     34                 25

R>    detach(birthwt)
```

It can be seen that both the `lm()` and `glm()` functions support a common formula interface, and the np package that we introduce shortly strives to maintain this method of interacting with functions with minimal changes where necessary.

As a final illustration of the capabilities and ease of use of the R environment, we consider a simple Monte Carlo experiment where we examine the finite-sample distribution of the sample mean for samples of size $n = 5$ when the underlying distribution is χ^2 with 1 df. We then plot the empirical PDF versus the asymptotic PDF of the sample mean (Fig. 1). $M = 10{,}000$ replications are computed.

4. SOME NONPARAMETRIC AND SEMIPARAMETRIC ROUTINES AVAILABLE IN R

Table 1 summarizes some of the nonparametric and semiparametric routines available to users of R. As can be seen, there appears to be a rich range of nonparametric implementations available to the practitioner. However, upon closer inspection, many are limited in one way or another in ways that might frustrate applied econometricians. For instance, some nonparametric regression methods admit only one regressor, while others admit only numerical data types and cannot admit categorical data that is often found in applied settings. Table 1 is not intended to be exhaustive, rather it ought to serve to orient the reader to a subset of the rich array of nonparametric methods that currently exist in the R environment. To see a routine in action, you can type `example (''funcname,'' package = ''pkgname'')` where `funcname` is the name of a routine and `pkgname` is the associated package, and this will run an example contained in the help file for that function. For instance, `example (''npreg,'' package = ''np'')` will run a kernel regression example from the package `np`.

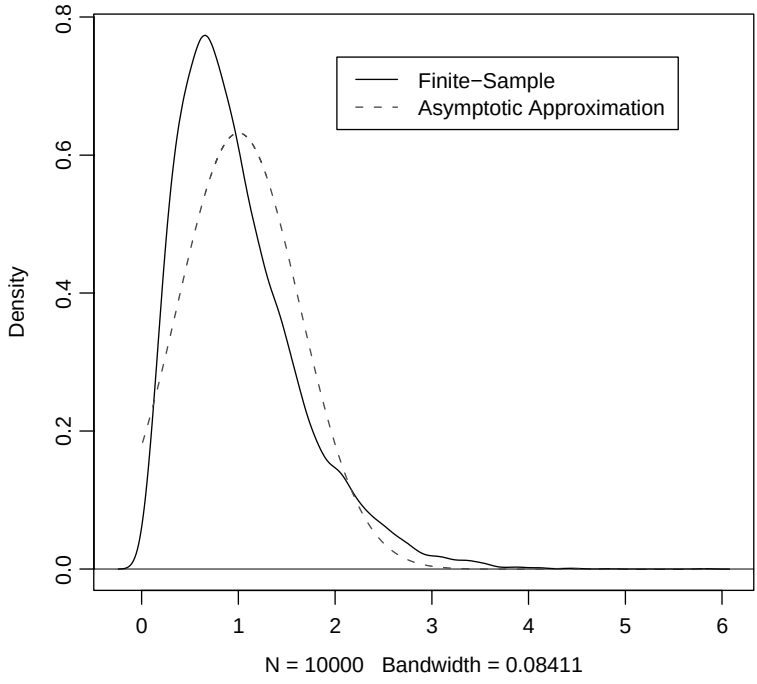


Fig. 1. Empirical Versus Asymptotic PDF.

4.1. Nonparametric Density Estimation in R

Univariate density estimation is one of the most popular exploratory nonparametric methods in use today. Readers will likely be familiar with two popular nonparametric estimators, namely the univariate histogram and kernel estimators. For an in-depth treatment of kernel density estimation, we direct the interested reader to the wonderful monographs by Silverman (1986) and Scott (1992), while for mixed data density estimation we direct the reader to Li and Racine (2003) and the references therein. We shall begin with an illustrative *parametric* example.

Consider any random variable X having probability density function $f(x)$, and let $f(\cdot)$ be the object of interest. Suppose one is presented with a series of independent and identically distributed draws from the unknown distribution and asked to model the density of the data, $f(x)$.

Table 1. An Illustrative Summary of R Packages that Implement Nonparametric Methods.

Package	Function	Description
Ash	*ash1*	Computes univariate averaged shifted histograms
	ash2	Computes bivariate averaged shifted histograms
car	*n. bins*	Computes number of bins for histograms with different rules
gam	*gam*	Computes generalized additive models using the method described in Hastie and Tibshirani (1990)
GenKern	*KernSec*	Computes univariate kernel density estimates
	KernSur	Computes bivariate kernel density estimates
Graphics (base)	*boxplot*	Produces box-and-whisker plot(s)
	nclass.Sturges	Computes the number of classes for a histogram
	nclass.scott	Computes the number of classes for a histogram
	nclass.FD	Computes the number of classes for a histogram
KernSmooth	*bkde*	Computes a univariate binned kernel density estimate using the fast Fourier transform as described in Silverman (1982)
	bkde2D	Compute a bivariate binned kernel density estimate as described in Wand (1994)
	dpik	Computes a bandwidth for a univariate kernel density estimate using the method described in Sheather and Jones (1991)
	dpill	Computes a bandwidth for univariate local linear regression using the method described in Ruppert, Sheather, and Wand (1995)
	locpoly	Computes a univariate probability density function, bivariate regression function or their derivatives using local polynomials
ks	*kde*	Computes a multivariate kernel density estimate for 1–6-dimensional numerical data
locfit	*locfit*	Computes univariate local regression and likelihood models
	sjpi	Computes a bandwidth via the plug-in Sheather and Jones (1991) method
	kdeb	Computes univariate kernel density estimate bandwidths
MASS	*bandwidth.nrd*	Computes Silverman's rule of thumb for choosing the bandwidth of a univariate Gaussian kernel density estimator
	hist.scott	Plot a histogram with automatic bin width selection (Scott)
	hist.FD	Plot a histogram with automatic bin width selection (Freedman–Diaconis)
	kde2d	Computes a bivariate kernel density estimate
	width.SJ	Computes the Sheather and Jones (1991) bandwidth for a univariate Gaussian kernel density estimator
	bcv	Computes biased cross-validation bandwidth selection for a univariate Gaussian kernel density estimator

Table 1. (*Continued*)

Package	Function	Description
	ucv	Computes unbiased cross-validation bandwidth selection for a univariate Gaussian kernel density estimator
np	*npcdens*	Computes a multivariate conditional density as described in Hall, Racine, and Li (2004)
	npcdist	Computes a multivariate conditional distribution as described in Li and Racine (2008)
	npcmstest	Conducts a parametric model specification test as described in Hsiao, Li, and Racine (2007)
	npconmode	Conducts multivariate modal regression
	npindex	Computes a multivariate single index model as described in Ichimura (1993), Klein and Spady (1993)
	npksum	Computes multivariate kernel sums with numeric and categorical data types
	npplot	Conducts general purpose plotting of nonparametric objects
	npplreg	Computes a multivariate partially linear model as described in Robinson (1988), Racine and Liu (2007)
	npqcmstest	Conducts a parametric quantile regression model specification test as described in Zheng (1998), Racine (2006)
	npqreg	Computes multivariate quantile regression as described in Li and Racine (2008)
	npreg	Computes multivariate regression as described in Racine and Li (2004), Li and Racine (2004)
	npscoef	Computes multivariate smooth coefficient models as described in Li and Racine (2007b)
	npsigtest	Computes the significance test as described in Racine (1997), Racine, Hart, and Li (2006)
	npudens	Computes multivariate density estimation as described in Parzen (1962), Rosenblatt (1956), Li and Racine (2003)
	npudist	Computes multivariate distribution functions as described in Parzen (1962), Rosenblatt (1956), Li and Racine (2003)
stats (base)	*bw.nrd*	Univariate bandwidth selectors for Gaussian windows in density
	density	Computes a univariate kernel density estimate
	hist	Computes a univariate histogram
	smooth.spline	Computes a univariate cubic smoothing spline as described in Chambers and Hastie (1991)
	ksmooth	Computes a univariate Nadaraya–Watson kernel regression estimate described in Wand and Jones (1995)
	loess	Computes a smooth curve fitted by the loess method described in Cleveland, Grosse, and Shyu (1992) (1–4 numeric predictors)

For this example, we shall simulate $n = 500$ draws but immediately discard knowledge of the true data generating process (DGP) pretending that we are unaware that the data is drawn from a mixture of normals ($N(-2, 0.25)$ and $N(3, 2.25)$ with equal probability). The following code snippet demonstrates one way to draw random samples from a mixture of normals.

```
R>  set.seed(123)
R>  n <- 250
R>  x <- sort(c(rnorm(n,mean = -2,sd = 0.5),rnorm(n,mean = 3,
      sd = 1.5)))
```

The following figure plots the true DGP evaluated on an equally spaced grid of 1,000 points (Fig. 2).

Suppose one naively presumed that the data is drawn from, say, the normal parametric family (not a mixture thereof), then tested this assumption using the Shapiro–Wilks test. The following code snippet demonstrates how this is done in R.

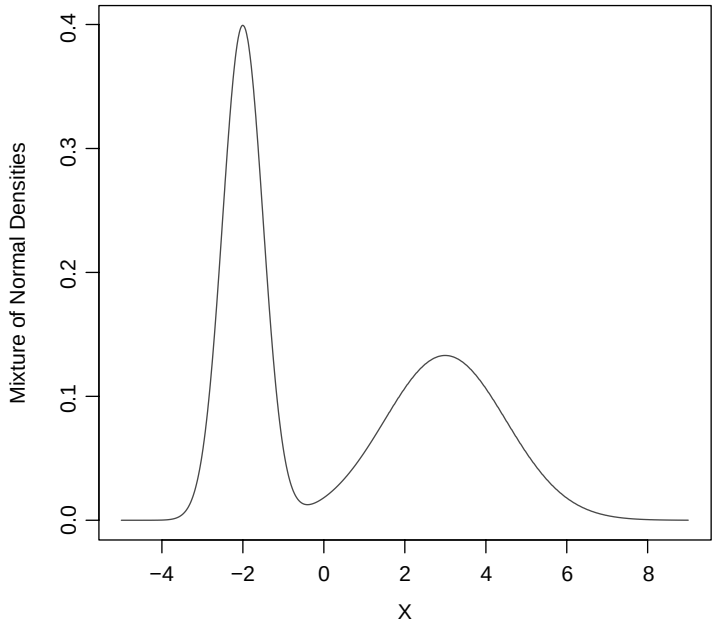


Fig. 2. True DGP.

```
R>    set.seed(123)
R>    M <- 10000
R>    n <- 5
R>    mean.vec <- numeric(length = M)
R>    for (i in 1:M) {
+       x <- rchisq(n,df = 1)
+       mean.vec [i] <- mean(x)
+     }
R>    mean.vec <- sort(mean.vec)
R>    plot(density(mean.vec),type = ''1'',lty = 1,main = '''')
R>    lines(mean.vec,dnorm(mean.vec,mean = mean(mean.vec),
        sd = sd(mean.vec)),
+         col = ''blue'',lty = 2)
R>    legend(2,0.75,
+         c(''Finite-Sample'',''Asymptotic Approximation''),
+         lty = c(1,2),col = c(''black'',''blue''))
R>    shapiro. test (x)
         Shapiro-Wilk normality test
R>    x.seq <- seq(-5,9,length = 1000)
R>    plot (x.seq,0.5*dnorm(x.seq,mean = -2,sd = 0.5)
        +0.5*dnorm(x.seq,mean = 3,sd = 1.5),
+       xlab = ''X'',
+       ylab = ''Mixture of Normal Densities''
+       type = ''1'',
+       main = '''',
+       col = ''blue'',
+       lty = 1)
data:   x
W = 0.87, p-value<2.2e-16
```

Given that this popular parametric model is flatly rejected by this dataset, we have two choices: (1) search for a more appropriate parametric model or (2) use more flexible estimators. For what follows, we shall presume that the readers have found themselves in just such a situation. That is, they have faithfully applied a parametric method and conducted a series of tests of model adequacy that indicate that the parametric model is not consistent with the underlying DGP. They then turn to more flexible methods of density estimation. Note that though we are considering density estimation at the moment, it could be virtually any parametric approach that we have been discussing, for instance, regression analysis and so forth.

If one wished to examine the histogram (Fig. 3) for this data one could use the following code snippet:

```
R>   hist(x,prob = TRUE,main = '''')
```

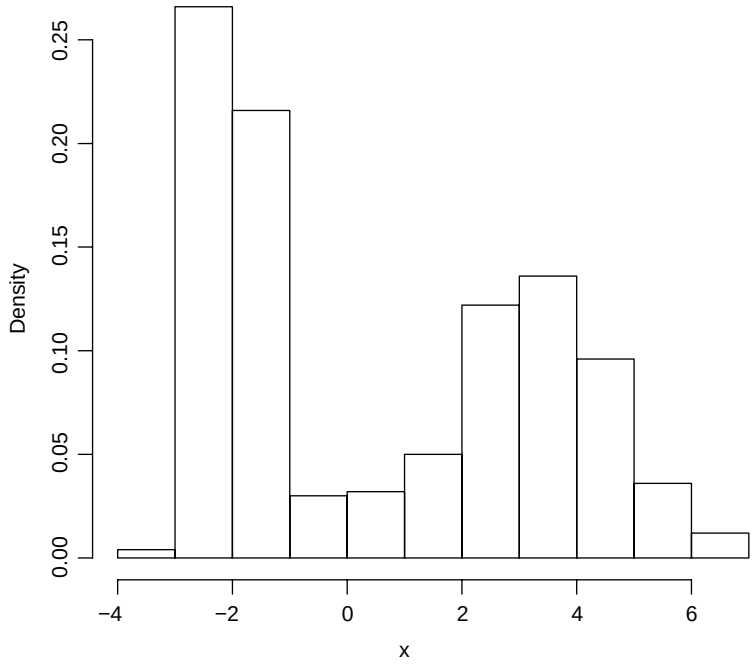


Fig. 3. Histogram.

Of course, though consistent, the histogram suffers from a number of drawbacks; hence, one might instead consider a smooth nonparametric density estimator such as the univariate Parzen kernel estimator (Parzen, 1962). A univariate kernel estimator can be obtained using the `density` command that is part of `R` base. This function supports a range of bandwidth methods (see `?bw.nrd` for details) and kernels (see `?density` for details). The default bandwidth method is Silverman's "rule of thumb" (Fig. 4) (Silverman, 1986, p. 48, Eq. (3.31)), and for this data we obtain the following:

```
R>  plot(density(x),main = '''')
```

The density function in `R` has a number of appealing features. It is extremely fast computationally speaking, as the algorithm disperses the mass of the empirical distribution function over a regular grid and then uses the fast Fourier transform to convolve this approximation with a discretized version of the kernel and then uses a linear approximation to evaluate the

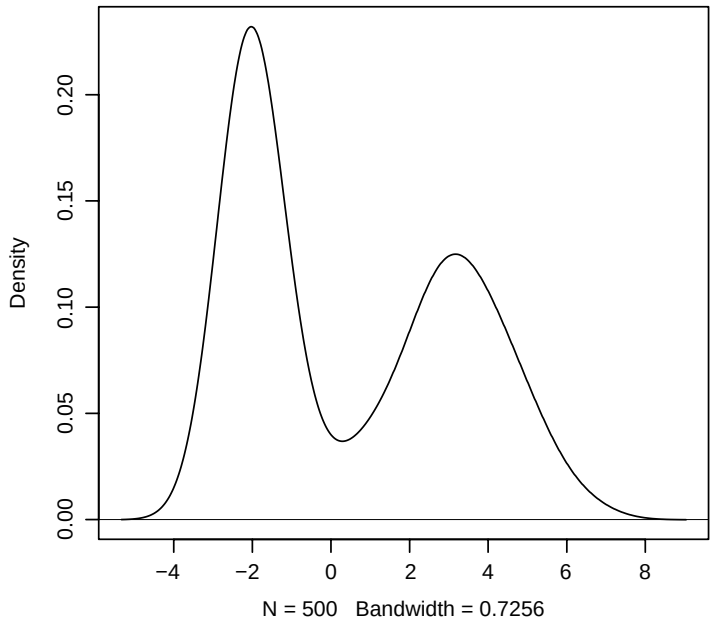


Fig. 4. PDF with Silverman's "Rule-of-Thumb" Bandwidth.

density at the specified points. If one wishes to obtain a univariate kernel estimate for a large sample of data, then this is definitely the function of choice. However, for a bivariate (or higher dimensional) density estimate, one would require alternative `R` routines. The function `bkd2dD` in the `KernSmooth` package can compute a two-dimensional density estimate as can `kde2d` in the MASS package and kde in the `ks` package though neither package implements a data-driven two-dimensional bandwidth selector. The `np` package, however, contains the function npudens that computes multivariate density estimates, is quite flexible, and admits data-driven bandwidth selection for an arbitrary number of dimensions and for both numeric and categorical data types. As the method does not rely on Fourier transforms and approximations, it is nowhere near as fast as the density function[6]; however, it is much more flexible. The default method of bandwidth selection is likelihood cross validation, and the following code snippet demonstrates this function using the "Old Faithful" dataset (Fig. 5). The Old Faithful Geyser is a tourist attraction located in Yellowstone National Park. This famous dataset containing $n = 272$ observations

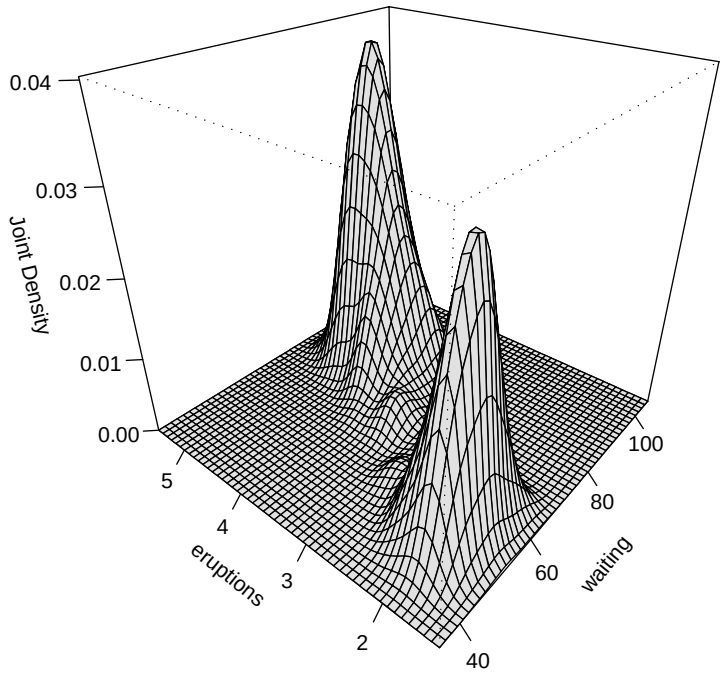


Fig. 5. PDF for Old Faithful Data.

consists of two variables, eruption duration (minutes) and waiting time until the next eruption (minutes).

```
R>  data(''faithful'',package = ''datasets'')
R>  Fhat <- npudens(~waiting+eruptions,data = faithful)
R>  plot(fhat,view = ''fixed'',xtrim = -0.1,theta = 310,
      phi = 30,main = '''')
```

For dimensions greater than two, one can plot "partial density surfaces" that plot one-dimensional slices of the density holding variables not on the axes constant at their median/modes (these can be changed by the user – see `?npplot` for details). One can also plot asymptotic and bootstrapped error surfaces, the CDF, and so forth as the following code snippet reveals (Fig. 6).

```
R> plot(fhat,cdf = TRUE,plot.errors.method = ''asymptotic'',
+    view = ''fixed'',xtrim = -0.1,theta = 310,phi = 30,main = '''')
```

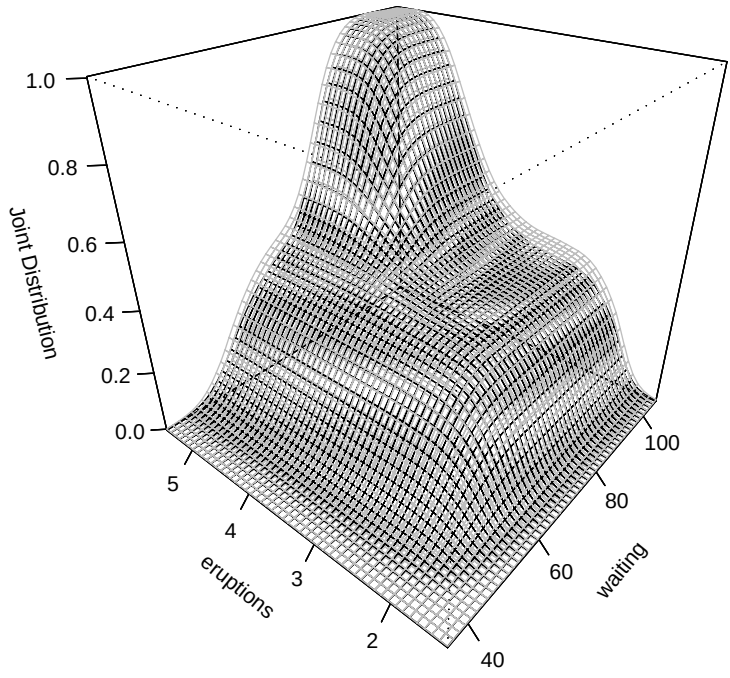


Fig. 6. CDF for Old Faithful Data.

4.2. Kernel Density Estimation with Numeric and Categorical Data

Suppose that we were facing a mix of categorical and numeric data and wanted to model the joint density[7] function. When facing a mix of categorical and numeric data, traditionally researchers using kernel methods resorted to a "frequency" approach. This approach involves breaking the numeric data into subsets according to the realizations of the categorical data (cells). This of course will produce consistent estimates. However, as the number of subsets increases, the amount of data in each cell falls leading to a "sparse data" problem. In such cases, there may be insufficient data in each subset to deliver sensible density estimates (the estimates will be highly variable). In what follows, we consider the method of Li and Racine (2003) that is implemented in the `np` package via the `npudens` function.

By way of example, we consider Wooldridge's (2002) "wage1" dataset ($n = 526$), and model the joint density of two variables (Fig. 7), one numeric (lwage) and one categorical (numdep). The "lwage" is the logarithm of

average hourly earnings for an individual and "numdep" the number of dependents (0, 1, …). We use likelihood cross validation to obtain the bandwidths. Note that this is indeed a case of "sparse" data, and the traditional approach would require estimation of a nonparametric univariate density function based upon only two observations for the last cell ($c = 6$) (Table 2).

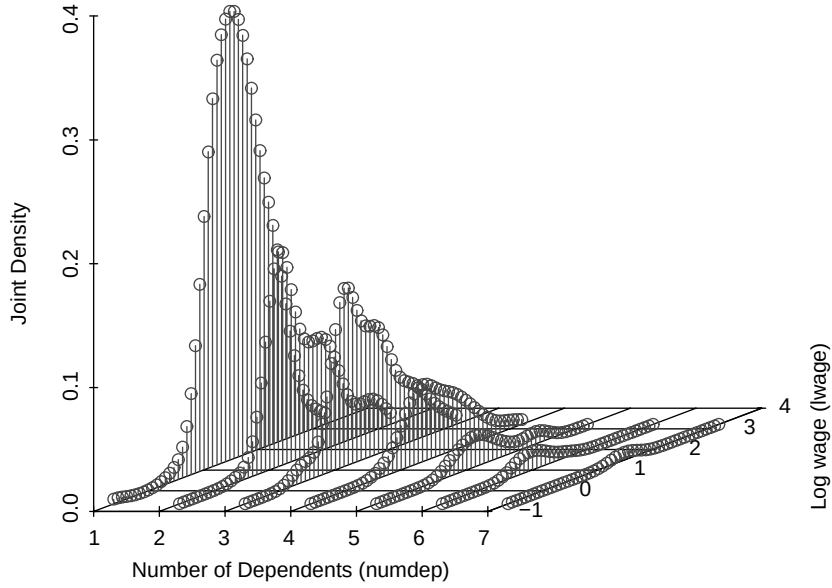


Fig. 7. Joint PDF for "lwage" and "numdep."

Table 2. Summary of numdep ($c = 0, 1, \ldots, 6$).

c	n_c
0	252
1	105
2	99
3	45
4	16
5	7
6	2

4.3. Conditional Density Estimation

Conditional density functions underlie many popular statistical objects of interest, though they are rarely modeled directly in parametric settings and have perhaps received even less attention in kernel settings. Nevertheless, as will be seen, they are extremely useful for a range of tasks, whether directly estimating the conditional density function, modeling count data (see Cameron & Trivedi, 1998) for a thorough treatment of count data models), or perhaps modeling conditional quantiles via estimation of a conditional CDF. And, of course, regression analysis (i.e., modeling conditional means) depends directly on the conditional density function, so this statistical object in fact implicitly forms the backbone of many popular statistical methods.

We consider Giovanni Baiocchi's Italian GDP growth panel for 21 regions covering the period 1951–1998 (millions of Lire, 1990 = base) (Fig. 8). There are 1,008 observations in total, and two variables, "gdp" and "year." We treat gdp as numeric and year as ordered.[8] The code snippet

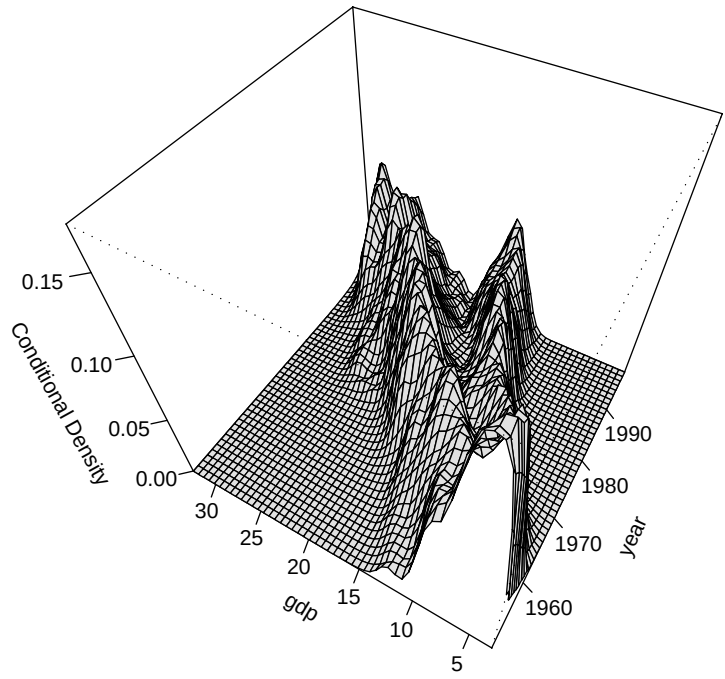


Fig. 8. Conditional PDF for Italy Panel.

below plots the estimated conditional density, $\hat{f}$(gdp|year) based upon likelihood cross-validated bandwidth selection.

It is clear that the distribution of income has evolved from a unimodal one in the early 1950s to a markedly bimodal one in the 1990s. This result is robust to bandwidth choice, and is observed whether using simple rules-of-thumb or data-driven methods such as least squares or likelihood cross validation. The kernel method readily reveals this evolution which might easily be missed if one were to use parametric models of the income distribution. For instance, the (unimodal) lognormal distribution is a popular parametric model for income distributions, but is incapable of revealing the multimodal structure present in this dataset.

```
R>  library(scatterplot3d)
R>  attach(wage1)
R>  bw <- npudensbw(~lwage+ordered(numdep),data = wage1)
R>  numdep.seq <- sort(unique(numdep))
R>  lwage.seq <- seq(min(lwage),max(lwage),length = 50)
R>  wage1.eval <- expand.grid(numdep = ordered(numdep.seq),
      lwage = lwage.seq)
R>  fhat <- fitted(npudens(bws = bw,newdata = wage1.eval))
R>  f <- matrix(fhat,length(unique(numdep)),50)
R>  scatterplot3d(wage1.eval[,1],wage1.eval[,2],fhat,
+       ylab = ''Log wage (lwage)'',
+       xlab = ''Number of Dependents (numdep)'',
+       zlab = ''Joint Density'',
+       angle = 15,box = FALSE,type = ''h'',grid = TRUE,
        color = ''blue'')
R>  detach(wage1)
R>  data(''Italy'')
R>  attach(Italy)
R>  fhat <- npcdens(gdp~year)
R>  plot(fhat,view = ''fixed'',main = '''',theta = 300,phi = 50)
```

4.4. Kernel Estimation of a Conditional Quantile

Estimating regression functions is a popular activity for applied economists. Sometimes, however, the regression function is not representative of the impact of the covariates on the dependent variable. For example, when the dependent variable is left (or right) censored, the relationship given by the regression function is distorted. In such cases, conditional quantiles above (or below) the censoring point are robust to the presence of censoring. Furthermore, the conditional quantile function provides a more

comprehensive picture of the conditional distribution of a dependent variable than the conditional mean function

We consider the method described in Li and Racine (2008) that is implemented in the npqreg function in the np package, which we briefly describe below.

The conditional αth quantile of a CDF $F(y|x)$ is defined as (Fig. 9)

$$q_\alpha(x) = \inf\{y : F(y|x) \geq \alpha\} = F^{-1}(\alpha|x)$$

where $\alpha \in (0, 1)$. In practice, we can estimate the conditional quantile function $q_\alpha(x)$ by inverting an estimated conditional CDF. Using a kernel estimator of $F(y|x)$, we would obtain

$$\hat{q}_\alpha(x) = \inf\{y : \hat{F}(y|x) \geq \alpha\} \equiv \hat{F}^{-1}(\alpha|x)$$

Because $\hat{F}(y|x)$ lies between zero and one and is monotone in Y, $\hat{q}_\alpha(x)$ always exists. In the example below, we compute the bandwidth object suitable for a conditional PDF and use this to estimate the conditional CDF and its conditional quantiles.

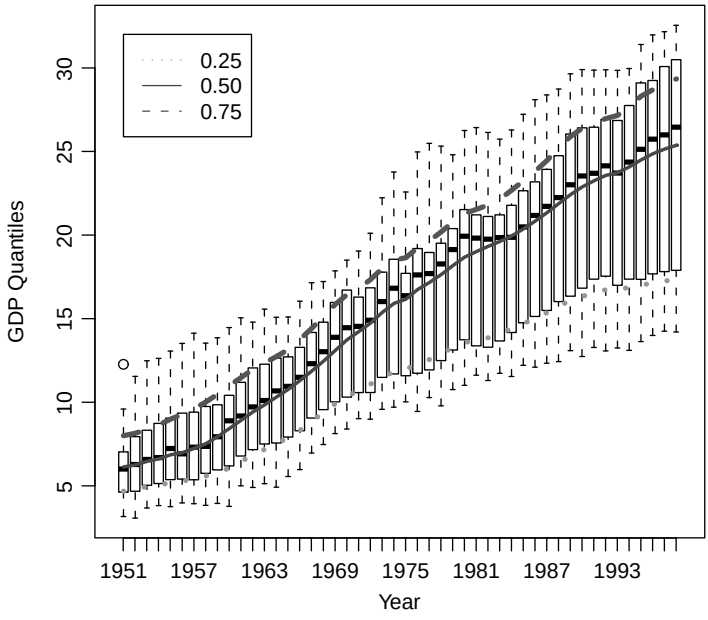


Fig. 9. Conditional Quantiles for Italy Panel.

The above plot, along with that for the conditional PDF, reveals that the distribution of income evolved from a unimodal one in the early 1950s to a markedly bimodal one in the 1990s.

4.5. Binary Choice and Count Data Models

We define a conditional mode by

$$m(x) = \max_y f(y|x)$$

In order to estimate a conditional mode $m(x)$, we need to model the conditional density. Let us call $\hat{m}(x)$ the estimated conditional mode, which is given by

$$\hat{m}(x) = \max_y \hat{f}(y|x)$$

where $\hat{f}(y|x)$ is the kernel estimator of $f(y|x)$. By way of example, we consider modeling low birthweights (a binary indicator) using this method.

For this example, we shall use the data on birthweights taken from the R MASS library (Venables & Ripley, 2002) that we used earlier in our introduction to parametric regression in R.

```
R>  data(''birthwt'',package = ''MASS'')
R>  attach(birthwt)
R>  bw <- npcdensbw(factor(low)~factor(smoke)+
+       factor(race)+
+       factor(ht)+
+       factor(ui)+
+       ordered(ftv)+
+       age+
+       lwt)
R>  model.np <- npconmode(bws = bw)
R>  model.np$confusion.matrix
R>  bw <- npcdensbw(gdp~ordered (year))
R>  model.q0.25 <- npqreg(bws = bw, tau = 0.25)
R>  model.q0.50 <- npqreg(bws = bw, tau = 0.50)
R>  model.q0.75 <- npqreg(bws = bw, tau = 0.75)
R>  plot(ordered(year), gdp,
+     main = '''',
+     xlab = ''Year'',
+     ylab = ''GDP Quantiles'')
```

```
R>   lines(ordered(year), model.q0.25$quant ile, col = ''green'',
       lty = 3, lwd = 3)
R>   lines(ordered(year), model.q0.50$quantile, col = ''blue'',
       lty = 1, lwd = 2)
R>   lines(ordered(year), model.q0.75$quant ile, col = ''red'',
       lty = 2,lwd = 3)
R>   legend(ordered(1951), 32, c(''0.25'', ''0.50'', ''0.75''),
+      lty = c(3, 1, 2), col = c(''green'', ''blue'', ''red''))
R>   detach(Italy)
```

```
              Predicted
Actual        0                  1
0             128                2
1             27                 32
R>            detach(birthwt)
```

4.6. Regression

One of the most popular methods for nonparametric kernel regression was proposed by Nadaraya (1965) and Watson (1964) and is known as the "Nadaraya–Watson" estimator (also known as the "local constant" estimator), though the "local polynomial" estimator (Fan, 1992) has emerged as a popular alternative; see Li and Racine (2007a, Chapter 2) for a detailed treatment of nonparametric regression.

For what follows, we consider an application taken from Wooldridge (2003, p. 226) that involves multiple regression analysis with both numeric and categorical data types.

We consider modeling an hourly wage equation using Wooldridge's (2002) "wage1" dataset that was outlined in Section 3. We use Hurvich, Simonoff, and Tsai's (1998) AIC_c approach for bandwidth selection in conjunction with local linear kernel regression (Fan, 1992). Note that the bandwidth object `bw.all` is precomputed and loaded when you load the `wage1` data, but we provide the code for its computation (commented out).

Note that the above figure displays "partial regression plots." A "partial regression plot" is simply a two-dimensional plot of the outcome y versus one covariate x_j when all other covariates are held constant at their respective medians/modes. The robust variability bounds are obtained by a nonparametric bootstrap.

4.7. Semiparametric Regression

Semiparametric methods constitute some of the more popular methods for flexible estimation. Semiparametric models are formed by combining parametric and nonparametric models in a particular manner. Such models are useful in settings where fully nonparametric models may not perform well, for instance, when the curse of dimensionality has led to highly variable estimates or when one wishes to use a parametric regression model but the functional form with respect to a subset of regressors or perhaps the density of the errors is not known. We might also envision situations in which some regressors may appear as a linear function (i.e., linear in variables) but the functional form of the parameters with respect to the other variables is not known, or perhaps where the regression function is nonparametric but the structure of the error process is of a parametric form (Fig. 10).

Semiparametric models such as the generalized additive model presented below can best be thought of as a compromise between fully nonparametric

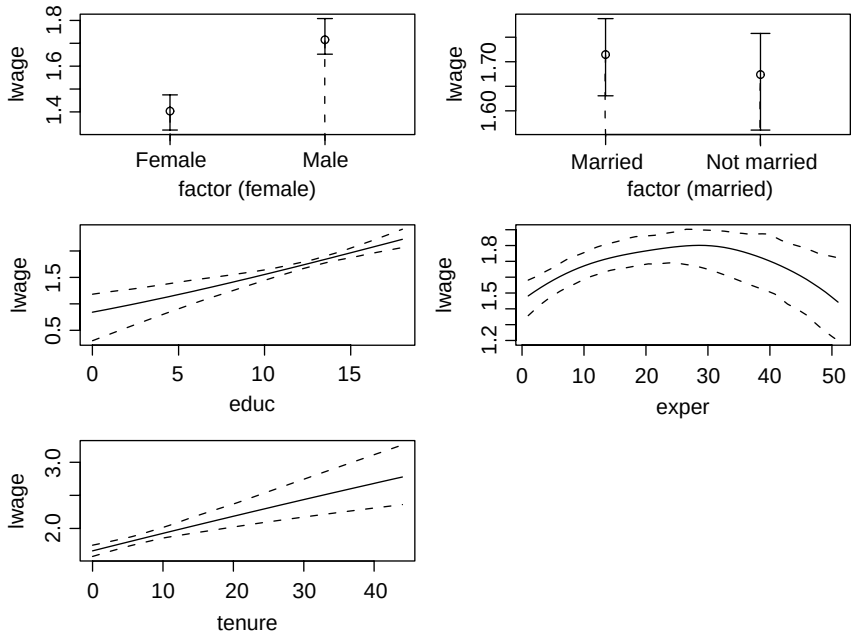


Fig. 10. Local Linear Wage Equation.

and fully parametric specifications. They rely on parametric assumptions and can therefore be misspecified and inconsistent, just like their parametric counterparts.

```
R>  attach (wage1)
R>  #bw.all <- npregbw(lwage~factor(female)+
R>  #    factor(married)+
R>  #    educ+
R>  #    exper+
R>  #    tenure,
R>  #    regtype = ''11'',
R>  #    bwmethod = ''cv.aic'',
R>  #    data = wage1)
R>
R>  model.np <- npreg(bws = bw.all)
R>  plot (model.np,
+     plot.errors.method = ''boot strap'',
+     plot.errors. boot.num = 100,
+     plot.errors.type = ''quantiles'',
+     plot.errors.style = ''band'',
+     common.scale = FALSE)
R>  detach(wage1)
```

4.8. Generalized Additive Models

Generalized additive models (see Hastie & Tibshirani, 1990) are popular in applied settings, though one drawback is that they do not support categorical variables (Fig. 11).

The semiparametric generalized additive model is given by

$$Y_i = c_0 + g_1(Z_{1i}) + g_2(Z_{2i}) + \cdots + g_q(Z_{qi}) + u_i, \quad i = 1, \ldots, n$$

where c_0 is a scalar parameter, the Z_{li}'s are all univariate continuous variables, and $g_l(\cdot)$ $(l = 1, \ldots, q)$ are unknown smooth functions.

The following code snippet considers the `wage1` dataset and uses three numeric regressors.

Note that the above figure again displays partial regression plots, but this time for the generalized additive model using only the continuous explanatory variables.

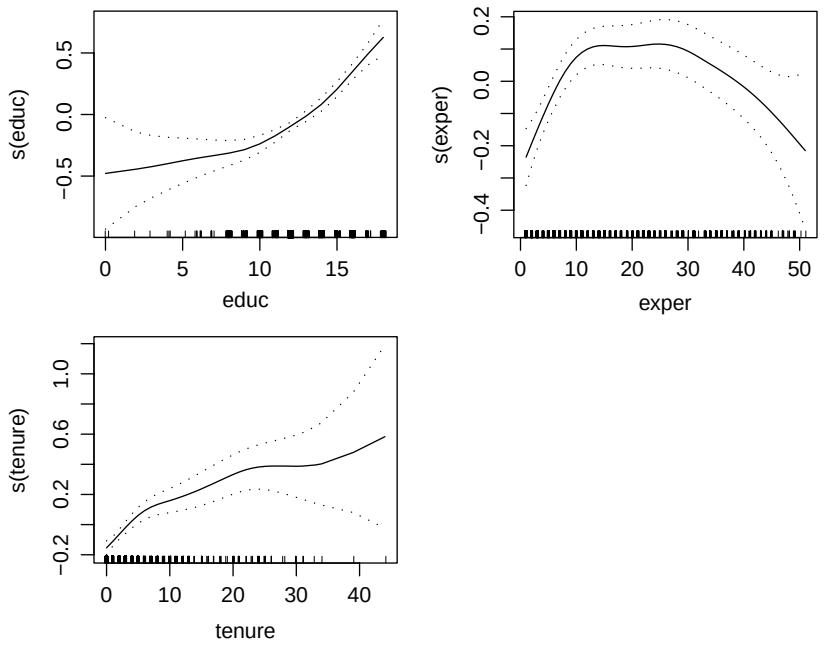


Fig. 11. Generalized Additive Wage Equation.

4.9. Partially Linear Models

The partially linear model is one of the simplest semiparametric models used in practice, and was proposed by Robinson (1988) while Racine and Liu (2007) extended the approach to handle the presence of categorical covariates.

A semiparametric partially linear model is given by

$$Y_i = X_i'\beta + g(Z_i) + u_i, \quad i = 1, \ldots, n$$

where X_i is a $p \times 1$ vector of regressors, β a $p \times 1$ vector of unknown parameters, and $Z_i \in \mathbb{R}^q$. The functional form of $g(\cdot)$ is not specified, and the finite dimensional parameter β constitutes the parametric part of the model and the unknown function $g(\cdot)$ the nonparametric part.

Suppose that we again consider the wage1 dataset from Wooldridge (2003, p. 222), but now assume that the researcher is unwilling to presume the nature of the relationship between exper and lwage, hence relegates

exper to the nonparametric part of a semiparametric partially linear model. The following code snippet considers a partially linear specification:

```
R>   bw <- npplregbw(lwage~factor(female)+
+        factor(married)+
+        educ+
+        tenure|exper,
+        data=wage1)
R>   model.pl <- npplreg(bw)
R>   summary(model.pl)
Partially Linear Model
Regression Data: 526 Training Points, in 5 Variable(s)
With 4 Linear Parametric Regressor(s), 1 Nonparametric Regressor(s)


                 y(z)
Bandwidth(s):    2.05

R>   options (SweaveHooks=list (multi fig=function ()
       par(mfrow=c(2,2))))
R>   library(gam)
R>   attach(wage1)
R>   model.gam <- gam(lwage~s(educ)+s(exper)+s(tenure))
R>   plot(model.gam,se=T)
R>   detach(wage1)

                 x(z)
Bandwidth(s):    4.19
                 1.35
                 3.16
                 5.24


                factor(female)  factor(married)    educ   tenure
Coefficient(s):         0.286         −0.0383      0.0788  0.0162

Kernel Regression Estimator: Local Constant
Bandwidth Type: Fixed

Residual SE: 0.154
R^2: 0.452

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

We can see from the above summary that the partially linear model yields coefficients for the explanatory variables entering the parametric part of the

model along with bandwidth from the nonparametric regression of Y on Z and each component of X on Z, where Y is the response, Z the explanatory variable entering the nonparametric component, and X the explanatory variables entering the parametric component.

4.10. Index Models

A semiparametric single index model is of the form

$$Y_i = g(X_i'\beta_0) + u_i, \quad i = 1, \ldots, n$$

where Y is the dependent variable, $X \in \mathbb{R}^q$ the vector of explanatory variables, β_0 the $q \times 1$ vector of unknown parameters, and u the error satisfying $E(u|X) = 0$. The term $x'\beta_0$ is called a "single index" because it is a scalar (a single index) even though $\boldsymbol{x}$ is a vector. The functional form of $g(\cdot)$ is unknown to the researcher. This model is semiparametric in nature since the functional form of the linear index is specified, while $g(\cdot)$ is left unspecified.

Ichimura (1993), Manski (1988), and Horowitz (1998, pp. 14–20) provide excellent intuitive explanations of the identifiability conditions underlying semiparametric single index models (i.e., the set of conditions under which the unknown parameter vector β_0 and the unknown function $g(\cdot)$ can be sensibly estimated), and we direct the reader to these references for details.

We consider applying Ichimura's (1993) single index method which is appropriate for numeric outcomes, unlike that of Klein and Spady (1993) outlined below. We again make use of the wage1 dataset found in Wooldridge (2003, p. 222).

```
R>  bw <- npindexbw(lwage~factor(female) +
+       factor(married)+
+       educ+
+       exper+
+       expersq+
+       tenure,
+       data = wage1)
R>  model <- npindex(bw)
R>  summary(model)

Single Index Model
Regression Data: 526 Training Points, in 6 Variable(s)
```

```
          factor      factor     educ    exper    expersq   tenure
         (female)    (married)
Beta:       1         −0.057    0.0427  0.0189  −0.000429  0.0101
Bandwidth: 0.0485
Kernel Regression Estimator: Local Constant

Residual SE: 0.151
R^2: 0.466

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

We again consider data on birthweights taken from the R MASS library (Venables & Ripley, 2002), and compute a single index model (the parametric logit model is outlined in Section 3). The outcome is an indicator of low infant birthweight (0/1) and so Klein and Spady's (1993) approach is appropriate. The confusion matrix is presented to facilitate a comparison of the index model and the logit model considered earlier.

```
R>  bw <- npindexbw(low~
+       factor(smoke)+
+       factor(race)+
+       factor(ht)+
+       factor(ui)+
+       ordered(ftv)+
+       age+
+       lwt,
+       method = ''kleinspady'',
+       data = birthwt)
R>  model.index <- npindex(bws = bw, gradients = TRUE)
R>  summary(model.index)

Single Index Model
Regression Data: 189 Training Points, in 7 Variable(s)


                  factor    factor  factor factor ordered   age
                  (smoke)   (race)   (ht)   (ui)   (ftv)
Beta:               1       0.051   0.364  0.184 −0.0506 −0.0159
         lwt
Beta: −0.00145
Bandwidth: 0.0159
Kernel Regression Estimator: Local Constant
```

```
Confusion Matrix
                Predicted
Actual            0        1
           0      119      11
           1      22       37


Overall Correct Classification Ratio: 0.825
Correct Classification Ratio By Outcome:
0             1
0.915         0.627

McFadden-Puig-Kerschner Performance Measure: 0.808

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

4.11. Smooth Coefficient (Varying Coefficient) Models

The smooth coefficient model is given by

$$\begin{aligned} Y_i &= \alpha(Z_i) + X_i'\beta(Z_i) + u_i \\ &= (1 + X_i')\begin{pmatrix} \alpha(Z_i) \\ \beta(Z_i) \end{pmatrix} + u_i \\ &= W_i'\gamma(Z_i) + u_i \end{aligned}$$

where X_i is a $k \times 1$ vector and where $\beta(z)$ is a vector of unspecified smooth functions of z.

Suppose that we once again consider the wage1 dataset from Wooldridge (2003, p. 222), but now assume that the researcher is unwilling to presume that the coefficients associated with the numeric variables do not vary with respect to the categorical variables female and married. The following code snippet presents a summary from the smooth coefficient specification.

```
R>   attach(wage1)
R>   bw <- npscoefbw(lwage~
+        educ+
+        tenure+
+        exper+
+        expersq/factor(female)+factor(married))
```

```
R>  model.scoef <- npscoef(bw,betas = TRUE)
R>  summary(model.scoef)
Smooth Coefficient Model
Regression Data: 526 Training Points, in 2 Variable(s)


                  factor(female)    factor(married)
Bandwidth(s):         0.00176            0.134

Bandwidth Type: Fixed

Residual SE: 0.147
R^2: 0.479

Unordered Categorical Kernel Type: Aitchison and Aitken
No. Unordered Categorical Explanatory Vars.: 2

R>  ## You could examine the matrix of smooth coefficients, or
      compute the average
R>  ## coefficient for each variable. One might then compare the
      average with the
R>  ## OLS model by way of example.
R>
R>  colMeans(coef(model.scoef))

Intercept        educ         tenure         exper         expersq
0.340213      0.078650      0.014296      0.030052     -0.000595

R>  coef (model.lm)

(Intercept)     factor(female)Male     factor(married)Notmarried
  0.181161           0.291130                  -0.056449
    educ              tenure                     exper
  0.079832           0.016074                  0.030100
expersq
-0.000601

R>  detach(wage1)
```

4.12. Panel Data Models

The nonparametric and semiparametric estimation of panel data models has received less attention than the estimation of standard regression models. Data panels are samples formed by drawing observations on N

cross-sectional units for T consecutive periods yielding a dataset of the form $\{Y_{it}, Z_{it}\}_{i=1,\,t=1}^{N,\,T}$. A panel is therefore simply a collection of N individual time series that may be short (small T) or long (large T).

The nonparametric estimation of time series models is itself an evolving field. However, when T is large and N is small then there exists a lengthy time series for each individual unit and in such cases one can avoid estimating a panel data model by simply estimating separate nonparametric models for each individual unit using the T individual time series available for each. If this situation applies, we direct the interested reader to Li and Racine (2007a, Chapter 18) for pointers to the literature on nonparametric methods for time series data.

When contemplating the nonparametric estimation of panel data models, one issue that immediately arises is that the standard (parametric) approaches that are often used for panel data models (such as first differencing to remove the presence of so-called "fixed effects") are no longer valid unless one is willing to presume additively separable effects, which for many defeats the purpose of using nonparametric methods in the first place.

A variety of approaches have been proposed in the literature, including Wang (2003), who proposed a novel method for estimating nonparametric panel data models that utilizes the information contained in the covariance structure of the model's disturbances; Wang, Carroll, and Lin (2005), who proposed a partially linear model with random effects; and Henderson, Carroll, and Li (2006), who consider profile likelihood methods for nonparametric estimation of additive fixed effect models which are removed via first differencing. In what follows, we consider direct nonparametric estimation of fixed effects models.

Consider the following nonparametric fixed effects panel data regression model:

$$Y_{it} = g(X_{it}) + u_{it}, \quad i = 1, 2 \ldots, N, \; t = 1, 2, \ldots, T$$

where $g(\cdot)$ is an unknown smooth function, $X_{it} = (X_{it,\,1}, \ldots, X_{it,q})$ is of dimension q, all other variables are scalars, and $E(u_{it}|X_{i1}, \ldots, X_{iT}) = 0$.

We say that panel data is "poolable" if one can "pool" the data, by in effect, ignoring the time series dimension, that is, by summing over both i and t without regard to the time dimension thereby effectively putting all data into the same pool then directly applying kernel regression methods. Of course, if the data is not poolable, this would obviously not be a wise choice.

However, to allow for the possibility that the data is in fact *potentially* poolable, one can introduce an unordered categorical variable, say $\delta_i = i$ for

$i = 1, 2, \ldots, N$, and estimate $E(Y_{it}|Z_{it}, \delta_i) = g(Z_{it}, \delta_i)$ nonparametrically using the mixed categorical and numeric kernel approach introduced earlier. Letting $\hat{\lambda}$ denote the cross-validated smoothing parameter associated with δ_i, then if $\hat{\lambda} = 1$, one gets $g(Z_{it}, \delta_i) = g(Z_{it})$ and the data is thereby pooled in the resulting estimate of $g(\cdot)$. If, on the other hand, $\hat{\lambda} = 0$ (or is close to 0), then this effectively estimates each $g_i(\cdot)$ using only the time series for the ith individual unit. Finally, if $0 < \hat{\lambda} < 1$, one might interpret this as a case in which the data is partially poolable.

We consider a panel of annual observations for six US airlines for the 15-year period, 1970–1984, taken from the Ecdat R package (Croissant, 2006) as detailed in Greene (2003, Table F7.1, p. 949). The variables in the panel are airline (airline), year (year), the logarithm of total cost in \$1,000 (lcost), the logarithm of an output index in revenue passenger miles (loutput), the logarithm of the price of fuel (lpf), and load factor, that is, the average capacity utilization of the fleet (lf). We treat "airline" as an unordered factor and "year" as an ordered factor and use a local linear estimator with Hurvich et al.'s (1998) AIC_c bandwidth approach.

```
R>  library(plm)
[1]  ''kinship is loaded''
R>  library(Ecdat)
R>  data(Airline)
R>  model.plm <- plm(log(cost) ~ log(output)+log(pf)+lf,
+       data = Airline,
+       model = ''within'',
+       index = c(''airline'',''year''))


[1]          90           3
R>           summary(model.plm)
Oneway (individual) effect Within Model


Call:
plm(formula = log(cost) ~ log(output)+log(pf)+lf, data = Airline,
  model = ''within'', index = c(''airline'', ''year''))

Balanced Panel: n = 6, T = 15, N = 90


Residuals:
  Min            1Q          Median          3Q           Max
−0.1560      −0.0352      −0.0093      0.0349      0.1660
```

```
Coefficients:
                Estimate       SE       t-value       Pr(>|t|)
log(output)       0.9193     0.0299      30.76        < 2e-16***
log(pf)           0.4175     0.0152      27.47        < 2e-16***
lf               -1.0704     0.2017      -5.31       0.00000011***
---

Significant codes: 0'***' 0.001'**' 0.01'*' 0.05'.' 0.1' ' 1

Total Sum of Squares: 39.4
Residual Sum of Squares: 0.293
F-statistic: 3604.81 on 3 and 81 df, p-value: <2e-16

R>   attach (Airline)
R>   lcost <- as.numeric(log(cost))
R>   loutput <- as.numeric(log(output))
R>   lpf <- as.numeric(log(pf))
R>   lf <- as.numeric(lf)
R>   bw <- npregbw(lcost~loutput +
+        lpf +
+        lf +
+        ordered(year) +
+        factor(airline),
+        regtype = ''11'',
+        bwmethod = ''cv.aic'',
+        ukertype = ''liracine'',
+        okertype = ''liracine'')
R>   summary(bw)
Regression Data (90 observations, 5 variable(s)):

Regression Type: Local Linear
Bandwidth Selection Method: Expected Kullback-Leibler Cross
  Validation
Formula: lcost~loutput+lpf+lf+ordered(year)+factor(airline)
Bandwidth Type: Fixed
Objective Function Value: -8.9e+15 (achieved on multistart 4)


Exp. Var. Name: loutput         Bandwidth: 1669084 Scale Factor: 2758857
Exp. Var. Name: lpf             Bandwidth: 0.0774  Scale Factor: 0.181
Exp. Var. Name: lf              Bandwidth: 0.0125  Scale Factor: 0.488
Exp. Var. Name: ordered(year)   Bandwidth: 0.167   Lambda Max:   1
Exp. Var. Name: factor(airline) Bandwidth: 0.0452  Lambda Max:   1


Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 3
```

```
Unordered Categorical Kernel Type: Li and Racine
No. Unordered Categorical Explanatory Vars.: 1

Ordered Categorical Kernel Type: Li and Racine
No. Ordered Categorical Explanatory Vars.: 1

R>   detach (Airline)
```

4.13. Rolling Your Own Functions

The np package contains the function `npksum` that computes kernel sums on evaluation data, given a set of training data, data to be weighted (optional), and a bandwidth specification (any bandwidth object).

The `npksum` exists so that you can create your own kernel objects with or without a variable to be weighted (default $Y = 1$). With the options available, you could create new nonparametric tests or even new kernel estimators. The convolution kernel option would allow you to create, say, the least squares cross-validation function for kernel density estimation.

The `npksum` uses highly optimized C code that strives to minimize its "memory footprint," while there is low overhead involved when using repeated calls to this function (see, by way of illustration, the example below that conducts leave-one-out cross validation for a local constant regression estimator via calls to the "R" function "nlm," and compares this to the "npregbw" function).

The `npksum` implements a variety of methods for computing multivariate kernel sums (p-variate) defined over a set of possibly numeric and/or categorical (unordered, ordered) data. The approach is based on Li and Racine (2003) who employ "generalized product kernels" that admit a mix of numeric and categorical data types.

Three classes of kernel estimators for the numeric data types are available: fixed, adaptive nearest neighbor, and generalized nearest neighbor. Adaptive nearest-neighbor bandwidths change with each sample realization in the set, x_i, when estimating the kernel sum at the point x. Generalized nearest-neighbor bandwidths change with the point at which the sum is computed, x. Fixed bandwidths are constant over the support of x. The `npksum` computes $\sum_j W_j' Y_j K(X_j)$, where A_j represents a row vector extracted from A. That is, it computes the kernel-weighted sum of the outer product of the rows of W and Y. In the examples from `?npksum`, the uses of such sums are illustrated.

The `npksum` may be invoked either with a formula-like symbolic description of variables on which the sum is to be performed or through a

simpler interface whereby data is passed directly to the function via the "txdat" and "tydat" parameters. Use of these two interfaces is mutually exclusive.

Data contained in the data frame "txdat" (and also "exdat") may be a mix of numeric (default), unordered categorical (to be specified in the data frame "txdat" using the "factor" command), and ordered categorical (to be specified in the data frame "txdat" using the "ordered" command). Data can be entered in an arbitrary order and data types will be detected automatically by the routine (see "np" for details).

A variety of kernels may be specified by the user. Kernels implemented for numeric data types include the second, fourth, sixth, and eighth order Gaussian and Epanechnikov kernels, and the uniform kernel. Unordered categorical data types use a variation on Aitchison and Aitken's (1976) kernel, while ordered data types use a variation of the Wang and van Ryzin (1981) kernel (see `?np` for details).

The following example implements leave-one-out cross validation for the local constant estimator using the `npksum` function and the `R nlm` function that carries out a minimization of a function using a Newton-type algorithm.

```
R>      n <- 100
R>      x1 <- runif(n)
R>      x2 <- rnorm(n)
R>      x3 <- runif(n)
R>      txdat <- data.frame(x1, x2, x3)
R>      tydat <- x1+sin(x2)+rnorm(n)
R>      ss <- function (h) {
+
+        if(min(h) <= 0) {
+
+          return(.Machine$double.xmax)
+
+        } else {
+
+          mean <- npksum(txdat,
+            tydat,
+            leave.one.out = TRUE,
+            bandwidth.divide = TRUE,
+            bws = h)$ksum/
+              npksum(txdat,
+                leave.one.out = TRUE,
+                bandwidth.divide = TRUE,
```

```
+                 bws = h)$ksum
+
+           return(sum((tydat-mean)^2)/length(tydat))
+
+         }
+
+       }
R>     nlm.return <- nlm(ss, runif(length(txdat)))
R>     bw <- npregbw(xdat = txdat, ydat = tydat)
R>     ## Bandwidths from nlm()
R>
R>     nlm.return$estimate
[1]       0.318           0.535     166.966
R>     ## Bandwidths from npregbw()
R>
R>     bw$bw
[1]       0.318           0.535     5851161.850
R>     ## Function value (minimum) from nlm()
R>
R>     nlm.return$minimum
[1]    1.02
R>     ## Function value (minimum) from npregbw()
R>
R>     bw$fval
[1]    1.02
```

5. SUMMARY

The R environment for statistical computing and graphics (R Development Core Team, 2008) offers practitioners a rich set of statistical methods ranging from random number generation and optimization methods through regression, panel data, and time series methods, by way of illustration. The standard R distribution (base R) comes preloaded with a rich variety of functionality useful for applied econometricians. This functionality is enhanced by user-supplied packages made available via R servers that are mirrored around the world. We hope that this chapter will encourage users to pursue the R environment should they wish to adopt nonparametric or semiparametric methods, and we wholeheartedly encourage those working in the field to strongly consider implementing their methods in the R environment thereby making their work accessible to the widest possible audience via an open collaborative forum.

NOTES

1. An interpreted programming language is one whose implementation is in the form of an interpreter. One often heard disadvantage of such languages is that when a program is interpreted, it tends to run slower than if it had been compiled.
2. A compiled language is one whose implementations are typically compilers (i.e., translators which generate "machine code" from "source code").
3. By "interface" we are simply referring to the way one interacts with the functions themselves. The np package that we discuss shortly supports the common "formula" interface which allows you to specify the list of covariates in a model in the same manner as you would any number of functions in the R environment (think of this as a "common look and feel" if you will).
4. An open software platform indicates that the source code and certain rights (those typically reserved for copyright holders) are provided under a license that meets the "open-source definition" or that is in the public domain.
5. A "confusion matrix" is simply a tabulation of the actual outcomes versus those predicted by a model. The diagonal elements contain correctly predicted outcomes while the off-diagonal ones contain incorrectly predicted (confused) outcomes.
6. To be specific, bandwidth selection is nowhere near as fast though computing the density itself is comparable once the bandwidth is supplied.
7. The term "density" is appropriate for distribution functions defined over mixed categorical and numeric variables. It is the measure defined on the categorical variables in the density function that matters.
8. It is good practice to classify your variables according to their data type in your data frame. This has already been done; hence, there is no need to write ordered (year).

ACKNOWLEDGMENTS

I would like to thank but not implicate the editors of this volume whose comments led to a much-improved version of this paper. All errors remain, naturally, my own. I would also like to gratefully acknowledge support from the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca) and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).

REFERENCES

Aitchison, J., & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, *63*(3), 413–420.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. New York: Cambridge University Press.

Carter Hill, R., Griffiths, W. E., & Lim, G. C. (2008). *Principles of econometrics* (3rd ed.). Hoboken, NJ: Wiley.

Chambers, J. M., & Hastie, T. (1991). *Statistical models in S*. London: Chapman and Hall.
Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. In: J. M. Chambers & T. J. Hastie (Eds), *Statistical models in S*. *Pacific Grove* (Chapter 8). CA: Wadsworth and Brooks/Cole.
Croissant, Y. (2006). *Ecdat: Data sets for econometrics*. R package version 0.1-5. URL: http://www.r-project.org
Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, *87*, 998–1004.
Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
Hall, P., Racine, J. S., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, *99*(468), 1015–1026.
Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman and Hall.
Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*(5), URL: http://www.jstatsoft.org/v27/i05/
Henderson, D., Carroll, R. J., & Li, Q. (2006). *Nonparametric estimation and testing of fixed effects panel data models*. Unpublished manuscript. Texas A&M University, Texas.
Horowitz, J. L. (1998). *Semiparametric methods in econometrics*. New York: Springer-Verlag.
Hsiao, C., Li, Q., & Racine, J. S. (2007). A consistent model specification test with mixed categorical and continuous data. *Journal of Econometrics*, *140*, 802–826.
Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B*, *60*, 271–293.
Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, *58*, 71–120.
Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, *61*, 387–421.
Li, Q., & Racine, J. (2007a). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.
Li, Q., & Racine, J. (2007b). *Smooth varying-coefficient nonparametric models for qualitative and quantitative data*. Unpublished manuscript. Department of Economics, Texas A&M University, Texas.
Li, Q., & Racine, J. S. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, *86*, 266–292.
Li, Q., & Racine, J. S. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, *14*(2), 485–512.
Li, Q., & Racine, J. S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, *26*, 423–434.
Manski, C. F. (1988). Identification of binary response models. *Journal of the American Statistical Association*, *83*(403), 729–738.
Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability*, *10*, 186–190.
Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*, 1065–1076.
Racine, J. S. (1997). Consistent significance testing for nonparametric regression. *Journal of Business and Economic Statistics*, *15*(3), 369–379.

Racine, J. S. (2006). *Consistent specification testing of heteroskedastic parametric regression quantile models with mixed data*. Unpublished manuscript. McMaster University, Hamilton, Ontario.

Racine, J. S., Hart, J. D., & Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, *25*, 523–544.

Racine, J. S., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130.

Racine, J. S., & Liu, L. (2007). *A partially linear kernel estimator for categorical data*. Unpublished manuscript. McMaster University, Hamilton, Ontario.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL: http://www.R-project.org

Robinson, P. M. (1988). Root-n consistent semiparametric regression. *Econometrica*, *56*, 931–954.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*, 832–837.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression (Corr: 96V91 p1380). *Journal of the American Statistical Association*, *90*, 1257–1270.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B, Methodological*, *53*, 683–690.

Silverman, B. W. (1982). Algorithm as 176: Kernel density estimation using the fast Fourier transform. *Applied Statistics*, *31*(1), 93–99.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Wand, M. P. (1994). Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics*, *3*(4), 433–445.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall.

Wang, M. C., & van Ryzin, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika*, *68*, 301–309.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, *90*, 43–52.

Wang, N., Carroll, R. J., & Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, *100*, 147–157.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, *26*(15), 359–372.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.

Wooldridge, J. M. (2003). *Introductory econometrics*. Mason, OH: South-Western (A division of Thompson Learning).

Zheng, J. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory*, *14*, 123–138.

PART VI
SURVEYS

SOME RECENT DEVELOPMENTS IN NONPARAMETRIC FINANCE

Zongwu Cai and Yongmiao Hong

ABSTRACT

This paper gives a selective review on some recent developments of nonparametric methods in both continuous and discrete time finance, particularly in the areas of nonparametric estimation and testing of diffusion processes, nonparametric testing of parametric diffusion models, nonparametric pricing of derivatives, nonparametric estimation and hypothesis testing for nonlinear pricing kernel, and nonparametric predictability of asset returns. For each financial context, the paper discusses the suitable statistical concepts, models, and modeling procedures, as well as some of their applications to financial data. Their relative strengths and weaknesses are discussed. Much theoretical and empirical research is needed in this area, and more importantly, the paper points to several aspects that deserve further investigation.

1. INTRODUCTION

Nonparametric modeling has become a core area in statistics and econometrics in the last two decades; see the books by Härdle (1990), Fan and Gijbels (1996), and Li and Racine (2007) for general statistical

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 379–432

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025015

methodology and theory as well as applications. It has been used successfully in various fields such as economics and finance due to its advantage of requiring little prior information on the data-generating process; see the books by Pagan and Ullah (1999), Mittelhammer, Judge, and Miller (2000), Tsay (2005), Taylor (2005), and Li and Racine (2007) for real examples in economics and finance. Recently, nonparametric techniques have been proved to be the most attractive way of conducting research and gaining economic intuition in certain core areas in finance, such as asset and derivative pricing, term structure theory, portfolio choice, risk management, and predictability of asset returns, particularly, in modeling both continuous and discrete financial time series models; see the books by Campbell, Lo, and MacKinlay (1997), Gourieroux and Jasiak (2001), Duffie (2001), Tsay (2005), and Taylor (2005).

Finance is characterized by time and uncertainty. Modeling both continuous and discrete financial time series has been a basic analytic tool in modern finance since the seminal papers by Sharpe (1964), Fama (1970), Black and Scholes (1973), and Merton (1973). The rationale behind it is that for most of time, news arrives at financial markets in both continuous and discrete manners. More importantly, derivative pricing in theoretical finance is generally much more convenient and elegant in a continuous-time framework than through binomial or other discrete approximations. However, statistical analysis based on continuous-time financial models has just emerged as a field in less than a decade, although it has been used for more than four decades for discrete financial time series. This is apparently due to the difficulty of estimating and testing continuous-time models using discretely observed data. The purpose of this survey is to review some recent developments of nonparametric methods used in both continuous and discrete time finance in recent years, and particularly in the areas of nonparametric estimation and testing of diffusion models, nonparametric derivative pricing and its tests, and predictability of asset returns based on nonparametric approaches. Financial time series data have some distinct important stylized facts, such as persistent volatility clusterings, heavy tails, strong serial dependence, and occasionally sudden but large jumps. In addition, financial modeling is often closely embedded in a financial theoretical framework. These features suggest that standard statistical theory may not be readily applicable to both continuous and discrete financial time series. This is a promising and fruitful area for both financial economists and statisticians to interact with.

Section 2 introduces various continuous-time diffusion processes and nonparametric estimation methods for diffusion processes. Section 3 reviews

the estimation and testing of a parametric diffusion model using nonparametric methods. Section 4 discusses nonparametric estimation and hypothesis testing of derivative and asset pricing, particularly the nonparametric estimation of risk neutral density (RND) functions and nonlinear pricing kernel models. Nonparametric predictability of asset returns is presented in Section 5. In Sections 2–5, we point out some open and interesting research problems, which might be useful for graduate students to review the important research papers in this field and to search for their own research topics, particularly dissertation topics for doctoral students. Finally, in Section 6, we highlight some important research areas that are not covered in this paper due to space limitation, say nonparametric volatility (conditional variance) and ARCH- or GARCH-type models and nonparametric methods in volatility for high-frequency data with/without microstructure noise. We plan to write a separate survey paper to discuss some of these omitted topics in the near future.

2. NONPARAMETRIC DIFFUSION MODELS

2.1. Diffusion Models

Modeling the dynamics of interest rates, stock prices, foreign exchange rates, and macroeconomic factors, inter alia, is one of the most important topics in asset pricing studies. The instantaneous risk-free interest rate or the so-called short rate is, for example, the state variable that determines the evolution of the yield curve in an important class of term structure models, such as Vasicek (1977) and Cox, Ingersoll, and Ross (1985, CIR). It is of fundamental importance for pricing fixed-income securities. Many theoretical models have been developed in mathematical finance to describe the short rate movement.[1]

In the theoretical term structure literature, the short rate or the underlying process of interest, $\{X_t,\ t \geq 0\}$, is often modeled as a time-homogeneous diffusion process, or stochastic differential equation:

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t \tag{1}$$

where $\{B_t,\ t \geq 0\}$ is a standard Brownian motion. The functions $\mu(\cdot)$ and $\sigma^2(\cdot)$ are, respectively, the drift (or instantaneous mean) and the diffusion (or instantaneous variance) of the process, which determine the dynamics of the short rate. Indeed, model (1) can be applied to many core areas in finance, such as options, derivative pricing, asset pricing, term structure of

interest rates, dynamic consumption and portfolio choice, default risk, stochastic volatility, exchange rate dynamics, and others.

There are two basic approaches to identifying $\mu(\cdot)$ and $\sigma(\cdot)$. The first is a parametric approach, which assumes some parametric forms of $\mu(\cdot, \theta)$ and $\sigma(\cdot, \theta)$, and estimates the unknown model parameters, say θ. Most existing models in the literature assume that the interest rate exhibits mean reversion and that the drift $\mu(\cdot)$ is a linear or quadratic function of the interest rate level. It is also often assumed that the diffusion $\sigma(\cdot)$ takes the form of $\sigma|X_t|^{\gamma}$, where γ measures the sensitivity of interest rate volatility to the interest rate level. In modeling interest rate dynamics, this specification captures the so-called "level effect," that is, the higher the interest rate level, the larger the volatility. With $\gamma = 0$ and 0.5, model (1) reduces to the well-known Vasicek and CIR models, respectively. The forms of $\mu(\cdot, \theta)$ and $\sigma(\cdot, \theta)$ are typically chosen due to theoretical wisdom or convenience. They may not be consistent with the data-generating process and there may be at risk of misspecification.

The second approach is a nonparametric one, which does not assume any restrictive functional form for $\mu(\cdot)$ and $\sigma(\cdot)$ beyond regularity conditions. In the last few years, great progress has been made in estimating and testing continuous-time models for the short-term interest rate using nonparametric methods.[2] Despite many studies, empirical analysis on the functional forms of the drift and diffusion is still not conclusive. For example, recent studies by Ait-Sahalia (1996b) and Stanton (1997) using nonparametric methods overwhelmingly reject all linear drift models for the short rate. They find that the drift of the short rate is a nonlinear function of the interest rate level. Both studies show that for the lower and middle ranges of the interest rate, the drift is almost zero, that is, the interest rate behaves like a random walk. But the short rate exhibits strong mean reversion when the interest rate level is high. These findings lead to the development of nonlinear term structure models such as those of Ahn and Gao (1999).

However, the evidence of nonlinear drift has been challenged by Pritsker (1998) and Chapman and Pearson (2000), who find that the nonparametric methods of Ait-Sahalia (1996b) and Stanton (1997) have severe finite sample problems, especially near the extreme observations. The finite sample problems with nonparametric methods cast doubt on the evidence of nonlinear drift. On the other hand, the findings in Ait-Sahalia (1996b) and Stanton (1997) that the drift is nearly flat for the middle range of the interest rate are not much affected by the small sample bias. The reason is that near the extreme observations, the nonparametric estimation might not be accurate due to the sparsity of data in this region. Also, this region is

close to the boundary point, so that the Nadaraya–Watson (NW) estimate suffers a boundary effect. Chapman and Pearson (2000) point out that this is a puzzling fact, since "there are strong theoretical reasons to believe that short rate cannot exhibit the asymptotically explosive behavior implied by a random walk model." They conclude that "time series methods alone are not capable of producing evidence of nonlinearity in the drift." Recently, to overcome the boundary effect, Fan and Zhang (2003) fit a nonparametric model using a local linear technique and apply the generalized likelihood ratio test of Cai, Fan, and Yao (2000) and Fan, Zhang, and Zhang (2001) to test whether the drift is linear. They support Chapman and Pearson's (2000) conclusion. However, the generalized likelihood ratio test is developed by Cai et al. (2000) for discrete time series and Fan et al. (2001) for independently and identically distributed (iid) samples, but it is still unknown whether it is valid for continuous time series contexts, which is warranted for a further investigation. Interest rate data are well known for persistent serial dependence. Pritsker (1998) uses Vasicek's (1977) model of interest rates to investigate the performance of a nonparametric density estimation in finite samples. He finds that asymptotic theory gives poor approximation even for a rather large sample size.

Controversies also exist on the diffusion $\sigma(\cdot)$. The specification of $\sigma(\cdot)$ is important, because it affects derivative pricing. Chan, Karolyi, Longstaff, and Sanders (1992) show that in a single factor model of the short rate, γ roughly equals to 1.5 and all the models with $\gamma \leq 1$ are rejected. Ait-Sahalia (1996b) finds that γ is close to 1; Stanton (1997) finds that in his semiparametric model γ is about 1.5; and Conley, Hansen, Luttmer, and Scheinkman (1997) show that their estimate of γ is between 1.5 and 2. However, Bliss and Smith (1998) argue that the result that γ equals to 1.5 depends on whether the data between October 1979 and September 1982 are included. From the foregoing discussions, it seems that the value of γ may change over time.

2.2. Nonparametric Estimation

Under some regularity conditions, see Jiang and Knight (1997) and Bandi and Nguyen (2000), the diffusion process in Eq. (1) is a one dimensional, regular, strong Markov process with continuous sample paths and time-invariant stationary transition density. The drift and diffusion are, respectively, the first two moments of the infinitesimal conditional distribution of X_t:

$$\mu(X_t) = \lim_{\Delta \to 0} \Delta^{-1} E[Y_t | X_t], \quad \text{and} \quad \sigma^2(X_t) = \lim_{\Delta \to 0} \Delta^{-1} E[Y_t^2 | X_t] \tag{2}$$

where $Y_t = X_{t+\Delta} - X_t$ (see, e.g., Øksendal, 1985; Karatzas & Shreve, 1988). The drift describes the movement of X_t due to time changes, whereas the diffusion term measures the magnitude of random fluctuations around the drift.

Using the Dynkin (infinitesimal) operator (see, e.g., Øksendal, 1985; Karatzas & Shreve, 1988), Stanton (1997) shows that the first-order approximation:

$$\mu(X_t)^{(1)} = \frac{1}{\Delta} E\{X_{t+\Delta} - X_t | X_t\} + O(\Delta)$$

the second-order approximation:

$$\mu(X_t)^{(2)} = \frac{1}{2\Delta}[4E\{Y_t|X_t\} - E\{X_{t+2\Delta} - X_t|X_t\}] + O(\Delta^2)$$

and the third-order approximation:

$$\mu(X_t)^{(3)} = \frac{1}{6\Delta}[18E\{Y_t|X_t\} - 9E\{X_{t+2\Delta} - X_t|X_t\} + 2E\{X_{t+3\Delta} - X_t|X_t\}] + O(\Delta^3)$$

etc. Fan and Zhang (2003) derive higher-order approximations. Similar formulas hold for the diffusion (see Stanton, 1997). Bandi and Nguyen (2000) argue that approximations to the drift and diffusion of any order display the same rate of convergence and limiting variance, so that asymptotic argument in conjunction with computational issues suggest simply using the first-order approximations in practice. As indicated by Stanton (1997), the higher the order of the approximations, the faster they will converge to the true drift and diffusion. However, as noted by Bandi and Nguyen (2000) and Fan and Zhang (2003), higher-order approximations can be detrimental to the efficiency of the estimation procedure in finite samples. In fact, the variance grows nearly exponentially fast as the order increases and they are much more volatile than their lower-order counterparts. For more discussions, see Bandi (2000), Bandi and Nguyen (2000), and Fan and Zhang (2003). The question arises is how to choose the order in application. As demonstrated in Fan and Zhang (2003), the first or second order may be enough in most applications.

Now suppose we observe X_t at $t = \tau\Delta$, $\tau = 1, \ldots, n$, in a fixed time interval $[0, T]$ with T. Denote the random sample as $\{X_{\tau\Delta}\}_{\tau=1}^{n}$. Then, it follows from Eq. (2) that the first-order approximations to $\mu(x)$ and $\sigma(x)$ lead to

$$\mu(x) \approx \frac{1}{\Delta} E[Y_\tau | X_{\tau\Delta} = x] \quad \text{and} \quad \sigma^2(x) \approx \frac{1}{\Delta} E[Y_\tau^2 | X_{\tau\Delta} = x] \tag{3}$$

for all $1 \le \tau \le n-1$, where $Y_\tau = X_{(\tau+1)\Delta} - X_{\tau\Delta}$. Both $\mu(x)$ and $\sigma^2(x)$ become classical nonparametric regressions and a nonparametric kernel smoothing approach can be applied to estimating them.

There are many nonparametric approaches to estimating conditional expectations. Most existing nonparametric methods in finance dwell mainly on the NW kernel estimator due to its simplicity. According to Ait-Sahalia (1996a, 1996b), Stanton (1997), Jiang and Knight (1997), and Chapman and Pearson (2000), the NW estimators of $\mu(x)$ and $\sigma^2(x)$ are given for any given grid point x, respectively, by

$$\hat{\mu}(x) \frac{1}{\Delta} \frac{\sum_{\tau=1}^{n-1} Y_\tau K_h(x - X_{\tau\Delta})}{\sum_{\tau=1}^{n-1} K_h(x - X_{\tau\Delta})}, \quad \text{and} \quad \hat{\sigma}^2(x) = \frac{1}{\Delta} \frac{\sum_{\tau=1}^{n-1} Y_\tau^2 K_h(x - X_{\tau\Delta})}{\sum_{\tau=1}^{n-1} K_h(x - X_{\tau\Delta})} \tag{4}$$

where $K_h(u) = K(u/h)/h, \quad h = h_n > 0$ is the bandwidth with $h \to 0$ and $nh \to \infty$ as $n \to \infty$, and $K(\cdot)$: $\mathbb{R} \to \mathbb{R}$ is a standard kernel. Jiang and Knight (1997) suggest first using Eq. (4) to estimate $\sigma^2(x)$. Observe that the drift

$$\mu(X_t) = \frac{1}{2\pi(X_t)} \frac{\partial[\sigma^2(X_t)\pi(X_t)]}{\partial X_t}$$

where $\pi(X_t)$ is the stationary density of $\{X_t\}$; see, for example, Ait-Sahalia (1996a), Jiang and Knight (1997), Stanton (1997), and Bandi and Nguyen (2000). Therefore, Jiang and Knight (1997) suggest estimating $\mu(x)$ by

$$\hat{\mu}(x) = \frac{1}{2\hat{\pi}(x)} \frac{\partial\{\hat{\sigma}^2(x)\hat{\pi}(x)\}}{\partial x}$$

where $\hat{\pi}(x)$ is a consistent estimator of $\pi(x)$, say, the classical kernel density estimator. The reason of doing so is based on the fact that in Eq. (1), the drift is of order dt and the diffusion is of order $\sqrt{dt}$, as $(dB_t)^2 = dt + O((dt)^2)$. That is, the diffusion has lower order than the drift for infinitesimal changes in time, and the local-time dynamics of the sampling path reflects more of the diffusion than those of the drift term. Therefore, when Δ is very small, identification becomes much easier for the diffusion term than the drift term.

It is well known that the NW estimator suffers from some disadvantages such as larger bias, boundary effects, and inferior minimax efficiency (see, e.g., Fan & Gijbels, 1996). To overcome these drawbacks, Fan and Zhang (2003) suggest using the local linear technique to estimate $\mu(x)$ as follows: When $X_{\tau\Delta}$ is in a neighborhood of the grid point x, by assuming that the second derivative of $\mu(\cdot)$ is continuous, $\mu(X_{\tau\Delta})$ can be approximated linearly as $\beta_0 + \beta_1(X_{\tau\Delta} - x)$, where $\beta_0 = \mu(x)$ and $\beta_1 = \mu'(x)$, the first

derivative of $\mu(x)$. Then, the locally weighted least square is given by

$$\sum_{\tau=1}^{n-1}\{\Delta^{-1}Y_\tau - \beta_0 - \beta_1(X_{\tau\Delta} - x)\}^2 K_h(X_{\tau\Delta} - x) \tag{5}$$

Minimizing the above with respect to β_0 and β_1 gives the local linear estimate of $\mu(x)$. Similarly, in view of Eq. (3), the local linear estimator of $\sigma^2(\cdot)$ can be obtained by changing $\Delta^{-1}Y_\tau$ in Eq. (5) into $\Delta^{-1}Y_\tau^2$. However, the local linear estimator of the diffusion $\sigma(\cdot)$ cannot be always nonnegative in finite samples. To attenuate this disadvantage of local polynomial method, a weighted NW method proposed by Cai (2001) can be used to estimate $\sigma(\cdot)$. Recently, Xu and Phillips (2007) study this approach and investigate its properties.

The asymptotic theory can be found in Jiang and Knight (1997) and Bandi and Nguyen (2000) for the NW estimator and in Fan and Zhang (2003) for the local linear estimator as well as Xu and Phillips (2007) for the weighted NW estimator. To implement kernel estimates, the bandwidth(s) must be chosen. In the iid setting, there are theoretically optimal bandwidth selections. There are no such results for diffusion processes available although there are many theoretical and empirical studies in the literature. As a rule of thumb, an easy way to choose a data-driven fashion bandwidth is to use the nonparametric version of the Akaike information criterion (see Cai & Tiwari, 2000).

One crucial assumption in the foregoing development is the stationarity of $\{X_t\}$. However, it might not hold for real financial time series data. If $\{X_t\}$ is not stationary, Bandi and Phillips (2003) propose using the following estimators to estimate $\mu(x)$ and $\sigma^2(x)$, respectively:

$$\hat{\mu}(x) = \frac{\sum_{\tau=1}^{n} K_h(x - X_{\tau\Delta})\tilde{\mu}(X_{\tau\Delta})}{\sum_{\tau=1}^{n} K_h(x - X_{\tau\Delta})}, \quad \text{and} \quad \hat{\sigma}^2(x) = \frac{\sum_{\tau=1}^{n} K_h(x - X_{\tau\Delta})\tilde{\sigma}^2(X_{\tau\Delta})}{\sum_{\tau=1}^{n} K_h(x - X_{\tau\Delta})}$$

where

$$\tilde{\mu}(x) = \frac{1}{\Delta}\frac{\sum_{\tau=1}^{n-1} I(|X_{\tau\Delta} - x| \le b) Y_\tau}{\sum_{\tau=1}^{n} I(|X_{\tau\Delta} - x| \le b)}, \quad \text{and} \quad \tilde{\sigma}^2(x) = \frac{1}{\Delta}\frac{\sum_{\tau=1}^{n-1} I(|X_{\tau\Delta} - x| \le b) Y_\tau^2}{\sum_{\tau=1}^{n} I(|X_{\tau\Delta} - x| \le b)}$$

See also Bandi and Nguyen (2000). Here, $b = b_n > 0$ is a bandwidth-like smoothing parameter that depends on the time span and on the sample size, which is called the spatial bandwidth in Bandi and Phillips (2003). This modeling approach is termed as the *chronological local time* estimation. Bandi and Phillips's approach can deal well with the situation that the series is not stationary. The reader is referred to the papers by

Bandi and Phillips (2003) and Bandi and Nguyen (2000) for more discussions and asymptotic theory.

Bandi and Phillips's (2003) estimator can be viewed as a double kernel smoothing method: The first step defines straight sample analogs to the values that drift and diffusion take at the sampled points and it can be regarded as a generalization of the moving average. Indeed, this step uses the smoothing technique (a linear estimator with the same weights) to obtain the raw estimates of the two functions $\tilde{\mu}(x)$ and $\tilde{\sigma}^2(x)$, respectively. This approach is different from classical two-step method in the literature (see Cai, 2002a, 2002b). The key is to figure out how important the first is to the second step. To implement this estimator, an empirical and theoretical study on the selection of two bandwidths b and h is needed.

2.3. Time-Dependent Diffusion Models

The time-homogeneous diffusion models in Eq. (1) have certain limitations. For example, they cannot capture the time effect, as addressed at the end of Section 2.1. A variety of time-dependent diffusion models have been proposed in the literature. A time-dependent diffusion process is formulated as

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t \tag{6}$$

Examples of Eq. (6) include Ho and Lee (HL) (1986), Hull and White (HW) (1990), Black, Derman, and Toy (BDT) (1990), and Black and Karasinski (BK) (1991), among others. They consider, respectively, the following models:

$$\begin{aligned}
\text{HL}: &\quad dX_t = \mu(t)dt + \sigma(t)dB_t \\
\text{HW}: &\quad dX_t = [\alpha_0 + \alpha_1(t)X_t]dt + \sigma(t)X_t^{\gamma}dB_t, \quad \gamma = 0 \quad \text{or} \quad 0.5 \\
\text{BDT}: &\quad dX_t = [\alpha_1(t)X_t + \alpha_2(t)X_t \log(X_t)]dt + \sigma(t)X_t dB_t \\
\text{BK}: &\quad dX_t = [\alpha_1(t)X_t + \alpha_2(t)X_t \log(X_t)]dt + \sigma(t)X_t dB_t
\end{aligned}$$

where $\alpha_2(t) = \sigma'(t)/\sigma(t)$. Similar to Eq. (2), one has

$$\mu(X_t,\ t) = \lim_{\Delta \to 0} \Delta^{-1} E\{Y_t | X_t\}, \quad \text{and} \quad \sigma^2(X_t,\ t) = \lim_{\Delta \to 0} \Delta^{-1} E\{Y_t^2 | X_t\}$$

where $Y_t = X_{t+\Delta} - X_t$, which provide a regression form for estimating $\mu(\cdot\,, t)$ and $\sigma^2(\cdot\,, t)$.

By assuming that the drift and diffusion functions are linear in X_t with time-varying coefficients, Fan, Jiang, Zhang, and Zhou (2003) consider the following time-varying coefficient single factor model:

$$dX_t = [\alpha_0(t) + \alpha_1(t)X_t]dt + \beta_0(t)X_t^{\beta_1(t)}dB_t \tag{7}$$

and use the local linear technique in Eq. (5) to estimate the coefficient functions $\{\alpha_j(\cdot)\}$ and $\{\beta_j(\cdot)\}$. Since the coefficients depend on time, $\{X_t\}$ might not be stationary. The asymptotic properties of the resulting estimators are still unknown. Indeed, the aforementioned models are a special case of the following more general time-varying coefficient multifactor diffusion model:

$$dX_t = \mu(X_t,\ t)dt + \sigma(X_t,\ t)dB_t \tag{8}$$

where

$$\mu(X_t,\ t) = \alpha_0(t) + \alpha_1(t)g(X_t) \quad \text{and} \quad (\sigma(X_t,\ t)\sigma(X_t,\ t)^\top)_{ij} = \beta_{0,ij}(t) + \beta_{1,ij}(t)^\top h_{ij}(X_t)$$

and $g(\cdot)$ and $\{h_{ij}(\cdot)\}$ are known functions. This is the time-dependent version of the multifactor affine model studied in Duffie, Pan, and Singleton (2000). It allows time-varying coefficients in multifactor affine models. A further theoretical and empirical study of the time-varying coefficient multifactor diffusion model in Eq. (8) is warranted. It is interesting to point out that the estimation approaches described above are still applicable to model (8) but the asymptotic theory is very challenging because of the nonstationarity of unknown structure of the underlying process $\{X_t\}$.

2.4. Jump-Diffusion Models

There has been a vast literature on the study of diffusion models with jumps.[3] The main purpose of adding jumps into diffusion models or stochastic volatility diffusion models is to accommodate impact of sudden and large shocks to financial markets, such as macroeconomic announcements, the Asian and Russian finance crisis, the US finance crisis, an unusually large unemployment announcement, and a dramatic interest rate cut by the Federal Reserve. For more discussions on why it is necessary to add jumps into diffusion models, see, for example, Lobo (1999), Bollerslev and Zhou (2002), Liu, Longstaff, and Pan (2002), and Johannes (2004), among others. Also, jumps can capture the heavy tail behavior of the distribution of the underlying process.

For the expositional purpose, we only consider a single factor diffusion model with jump:

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t + dJ_t \tag{9}$$

where J_t is a compensated jump process (zero conditional mean) with arrival rate (conditional probability) $\lambda_t = \lambda(X_t) \geq 0$, which is an instantaneous intensity function. There are several studies on specification of J_t. For example, a simple specification is to assume $J_t = \xi P_t$, where P_t is a Poisson process with an intensity $\lambda(X_t)$ or a binomial distribution with probability $\lambda(X_t)$, and the jump size, ξ, has a time-invariant distribution $\Pi(\cdot)$ with mean zero. $\Pi(\cdot)$ is commonly assumed to be either normally or uniformly distributed. If $\lambda(\cdot) = 0$ or $E(\xi^2) = 0$, the jump-diffusion model in Eq. (9) becomes the diffusion model in Eq. (1). More generally, Chernov, Gallant, Ghysels, and Tauchen (2003) consider a Lévy process for J_t. A simple jump-diffusion model proposed by Kou (2002) is discussed in Tsay (2005) by assuming that $J_t = \sum_{i=1}^{n_t}(L_i - 1)$, where n_t is a Poisson process with rate λ and $\{L_i\}$ a sequence of iid nonnegative random variables such that $\ln(L_i)$ has a double exponential distribution with probability density function $f(x) = \exp(-|x - \theta_1|/\theta_2)/2\theta_2$ for $0 < \theta_2 < 1$. This simple model enjoys several nice properties. The returns implied by the model are leptokurtic and asymmetric with respect to zero. In addition, the model can reproduce volatility smile and provide analytical formulas for the prices of many options.

In practice, $\lambda(\cdot)$ might be assumed to have a particular form. For example, Chernov et al. (2003) consider three different types of special forms, each having the appealing feature of yielding analytic option pricing formula for European-type contracts written on the stock price index. There are some open issues for the jump-diffusion model: (i) jumps are not observed and it is not possible to say surely if they exist; (ii) if they exist, a natural question arises is how to estimate a jump time τ, which is defined to be the discontinuous time at which $X_{\tau+} \neq X_{\tau-}$, and the jump size $\xi = X_{\tau+} - X_{\tau-}$. We conjecture that a wavelet method may be potentially useful here because a wavelet approach has an ability of capturing the discontinuity and removing the contaminated noise. For detailed discussion on how to use a wavelet method in this regard, the reader is referred to the paper by Fan and Wang (2007). Indeed, Fan and Wang (2007) propose using a wavelet method to cope with both jumps in the price and market microstructure noise in the observed data to estimate both integrated volatility and jump variation from the data sampled from jump-diffusion price processes, contaminated with the market microstructure noise.

Similar to Eq. (2), the first two conditional moments are given by

$$\mu_1(X_t) = \lim_{\Delta \downarrow 0} \Delta^{-1} E[Y_t | X_t] = \mu(X_t) + \lambda(X_t) E(\xi)$$

and

$$\mu_2(X_t) = \lim_{\Delta \downarrow 0} \Delta^{-1} E[Y_t^2 | X_t] = \sigma^2(X_t) + \lambda(X_t) E(\xi^2)$$

Clearly, $\mu_2(X_t)$ is much bigger than $\sigma^2(X_t)$ if there is a jump. This means that adding a jump into the model can capture the heavy tails. Also, it is easy to see that the first two moments are the same as those for a diffusion model by using a new drift coefficient $\tilde{\mu}(X_t) = \mu(X_t) + \lambda(X_t)E(\xi)$ and a new diffusion coefficient $\tilde{\sigma}^2(x) = \sigma^2(x) + \lambda(x)E(\xi^2)$. However, the fundamental difference between a diffusion model and a diffusion model with jumps relies on higher-order moments. Using the infinitesimal generator (Øksendal, 1985; Karatzas and Shreve, 1988) of X_t, we can compute, $j > 2$,

$$\mu_j(X_t) = \lim_{\Delta \to 0} \Delta^{-1} E[Y_t^j | X_t] = \lambda(X_t) E(\xi^j)$$

See Duffie et al. (2000) and Johannes (2004) for details. Obviously, jumps provide a simple and intuitive mechanism for capturing the heavy tail behavior of underlying process. In particular, the conditional skewness and kurtosis are, respectively, given by

$$s(X_t) \equiv \frac{\lambda(X_t)E(\xi^3)}{[\sigma^2(X_t) + \lambda(X_t)E(\xi^2)]^{3/2}}, \quad \text{and} \quad k(X_t) \equiv \frac{\lambda(X_t)E(\xi^4)}{[\sigma^2(X_t) + \lambda(X_t)E(\xi^2)]^2}$$

Note that $s(X_t) = 0$ if ξ is symmetric. By assuming $\xi \sim N(0, \sigma_\xi^2)$, Johannes (2004) uses the conditional kurtosis to measure the departures for the treasury bill data from normality and concludes that interest rates exchanges are extremely non-normal.

The NW estimation of $\mu_j(\cdot)$ is considered by Johannes (2004) and Bandi and Nguyen (2003). Moreover, Bandi and Nguyen (2003) provide a general asymptotic theory for the resulting estimators. Further, by specifying a particular form of $\Pi(\xi) = \Pi_0(\xi, \theta)$, say, $\xi \sim N(0, \sigma_\xi^2)$, Bandi and Nguyen (2003) propose consistent estimators of $\lambda(\cdot)$, σ_ξ^2, and $\sigma^2(\cdot)$ and derive their asymptotic properties.

A natural question arises is how to measure the departures from a pure diffusion model statistically. That is to test model (9) against model (1). It is equivalent to checking whether $\lambda(\cdot) \equiv 0$ or $\xi = 0$. Instead of using the conditional skewness or kurtosis, a test statistic can be constructed based on

the higher-order conditional moments. For example, one can construct the following nonparametric test statistics:

$$T_1 = \int \hat{\mu}_4(x)w(x)dx, \quad \text{or} \quad T_2 = \int \hat{\mu}_3^2(x)w(x)dx \tag{10}$$

where $w(\cdot)$ is a weighting function. The asymptotic theory for T_1 and T_2 is still unknown. It needs a further investigation theoretically and empirically. Based on a Monte Carlo simulation approach, Cai and Zhang (2008b) use the aforementioned testing statistics in an application, described as follows.

It is well known that prices fully reflect the available information in the efficient market. Thus, Cai and Zhang (2008b) consider the market information consisting of two components. The first is the anticipated information that drives market prices' daily normal fluctuation, and the second is the unanticipated information that determines prices to exceptional fluctuation, which can be characterized by a jump process. Therefore, Cai and Zhang (2008b) investigate the market information via a jump-diffusion process. The jump term in the dynamic of stock price or return rate reflects the sensitivity of unanticipated information for the related firms. This implies that the investigation of the jump parameters for firms with different sizes would help us to find the relationship between firm sizes and information sensitivity. With the nonparametric method as described above, Cai and Zhang (2008b) use the kernel estimation method, and reveal how the nonparametric estimation of the jump parameters (functions) reflect the so-called information effect. Also, they test the model based on the test statistic formulated in Eq. (10). Due to the lack of the relevant theory of the test statistics in Eq. (10), Cai and Zhang (2008b) use the Monte Carlo simulation, and find that a jump-diffusion process performs better to model with all market information, including anticipated and unanticipated information than the pure diffusion model. Empirically, Cai and Zhang (2008b) estimate the jump intensity and jump variance for portfolios with different firm sizes for data from both the US and Chinese markets, and find some evidences that there exists information effect among different firm sizes, from which we could get valuable references for investors' decision making. Finally, using a Monte Carlo simulation method, Cai and Zhang (2008a) examine the test statistics in Eq. (10) to see how the discontinuity of drift or diffusion function affects the performance of the test statistics. They find that the discontinuity of drift or diffusion function has an impact on the performance of the test statistics in Eq. (10).

More generally, given a discrete sample of a diffusion process, can one tell whether the underlying model that gave rise to the data was a diffusion, or should jumps be allowed into the model? To answer this question, Ait-Sahalia (2002b) proposes an approach to identifying the sufficient and necessary restriction on the transition densities of diffusions, at the sampling interval of the observed data. This restriction characterizes the continuity of the unobservable continuous sample path of the underlying process and is valid for every sampling interval including long ones. Let $\{X_t,\ t \geq 0\}$ be a Markovian process taking values in $D \subseteq \mathbb{R}$. Let $p(\Delta,\ y|x)$ denote the transition density function of the process over interval length Δ, that is, the conditional density of $X_{t+\Delta} = y$ given $X_t = x$, and it is assumed that the transition densities are time homogenous. Ait-Sahalia (2002b) shows that if the transition density $p(\Delta,\ y|x)$ is strictly positive and twice-continuously differentiable on $D \times D$ and the following condition:

$$\frac{\partial^2}{\partial x\, \partial y} \ln\ p(\Delta,\ y|x) > 0 \quad \text{for all } \Delta > 0 \quad \text{and} \quad (x,\ y) \in D \times D$$

(which is the so-called "diffusion criterion" in Ait-Sahalia, 2002b), is satisfied, then the underlying process is a diffusion. From a discretely sampled time series $\{X_{\tau\Delta}\}$, one could test nonparametrically the hypothesis that the data were generated by a continuous-time diffusion $\{X_t\}$. That is to test nonparametrically the null hypothesis

$$\mathbb{H}_0 : \frac{\partial^2}{\partial x\, \partial y} \ln\ p(\Delta,\ y|x) > 0 \quad \text{for all } x,\ y$$

versus the alternative

$$\mathbb{H}_a : \frac{\partial^2}{\partial x\, \partial y} \ln\ p(\Delta,\ y|x) \leq 0 \quad \text{for some } x,\ y$$

One could construct a test statistic based on checking whether the above "diffusion criterion" holds for a nonparametric estimator of $p(\Delta,\ y|x)$. This topic is still open. If the model has a specific form, say a parametric form, the diffusion criterion becomes a simple form, say, it becomes just a constraint for some parameters. Then, the testing problem becomes testing a constraint on parameters; see Ait-Sahalia (2002b) for some real applications.

2.5. Time-Dependent Jump-Diffusion Models

Duffie et al. (2000) consider the following time-dependent jump-diffusion model:

$$dX_t = \mu(X_t,\ t)dt + \sigma(X_t,\ t)dB_t + dJ_t \tag{11}$$

where J_t is a compensated jump process with the time-varying intensity $\lambda(X_t,\ t) = \lambda_0(t) + \lambda_1(t)X_T$; and Chernov et al. (2003) consider a more general stochastic volatility model with the time-varying stochastic intensity $\lambda(\xi_0,\ X_t,\ t) = \lambda_0(\xi_0,\ t) + \lambda_1(\xi_0,\ t)X_t$, where ξ_0 is the size of the previous jump. This specification yields a class of jump Lévy measures which combine the features of jump intensities depending on, say, volatility, as well as the size of the previous jump. Johannes, Kumar, and Polson (1999) also propose a class of jump-diffusion processes with a jump intensity depending on the past jump time and the absolute return. Moreover, as pointed out by Chernov et al. (2003), another potentially very useful specification of the intensity function would include the past duration, that is, the time since the last jump, say $\tau(t)$, which is the time that has elapsed between the last jump and t where $\tau(t)$ is a continuous function of t, such as

$$\lambda(\xi_0,\ X_t,\ \tau,\ t) = \{\lambda_0(t) + \lambda_1(t)X_t\}\lambda\{\tau(t)\}\exp\{G(\xi_0)\} \tag{12}$$

which can accommodate the increasing, decreasing, or hump-shaped hazard functions of the size of the previous jump, and the duration dependence of jump intensities. However, to the best of our knowledge, there have not been any attempts in the literature to discuss the estimation and test of the intensity function $\lambda(\cdot)$ nonparametrically in the above settings.

A natural question arises is how to generalize model (9) economically and statistically to a more general time-dependent jump-diffusion model given in Eq. (11) with the time-dependent intensity function $\lambda(\xi_0,\ X_t,\ \tau,\ t)$ without any specified form or with some nonparametric structure, say, like Eq. (12). Clearly, they include the aforementioned models as a special case, which are studied by Duffie et al. (2000), Johannes et al. (1999), and Chernov et al. (2003), among others. This is still an open problem.

3. NONPARAMETRIC INFERENCES OF PARAMETRIC DIFFUSION MODELS

3.1. Nonparametric Estimation

As is well known, derivative pricing in mathematical finance is generally much more tractable in a continuous-time modeling framework than through binomial or other discrete approximations. In the empirical literature, however, it is an usual practice to abandon continuous-time modeling when estimating derivative pricing models. This is mainly due to the difficulty that the transition density for most continuous-time models with discrete observations has no closed form and therefore the maximum likelihood estimation (MLE) is infeasible.

One major focus of the continuous-time literature is on developing econometric methods to estimate continuous-time models using discretely sampled data.[4] This is largely motivated by the fact that using the discrete version of a continuous-time model can result in inconsistent parameter estimates (see Lo, 1988). Available estimation procedures include the MLE method of Lo (1988); the simulated methods of moments of Duffie and Singleton (1993) and Gourieroux, Monfort, and Renault (1993); the generalized method of moments (GMM) of Hansen and Scheinkman (1995); the efficient method of moments (EMM) of Gallant and Tauchen (1996); the Markov chain Monte Carlo (MCMC) of Jacquier, Polson, and Rossi (1994), Eraker (1998), and Jones (1998); and the methods based on the empirical characteristic function of Jiang and Knight (2002) and Singleton (2001).

Below we focus on some nonparametric estimation methods of a parametric continuous-time model

$$dX_t = \mu(X_t, \theta)dt + \sigma(X_t, \theta)dB_t \tag{13}$$

where $\mu(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are known functions and θ an unknown parameter vector in an open bounded parameter space Θ. Ait-Sahalia (1996b) proposes a minimum distance estimator:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} n^{-1} \sum_{\tau=1}^{n} [\hat{\pi}_0(X_{\tau\Delta}) - \pi(X_{\tau\Delta}, \theta)]^2 \tag{14}$$

where

$$\hat{\pi}_0(x) = n^{-1} \sum_{\tau=1}^{n} K_h(x - X_{\tau\Delta})$$

is a kernel estimator for the stationary density of X_t, and

$$\pi(x,\ \theta) = \frac{c(\theta)}{\sigma^2(x,\ \theta)} \exp\left\{ \int_{x_0^*}^{x} \frac{2\mu(u,\ \theta)}{\sigma^2(u,\ \theta)} du \right\} \tag{15}$$

is the marginal density estimator implied by the diffusion model, where the standardization factor $c(\theta)$ ensures that $\pi(\cdot,\ \theta)$ integrates to 1 for every $\theta \in \Theta$, and x_0^* is the lower bound of the support of X_t. Because the marginal density cannot capture the full dynamics of the diffusion process, one can expect that $\hat{\theta}$ will not be asymptotically most efficient, although it is root-n consistent for θ_0 if the parametric model is correctly specified.

Next, we introduce the approximate maximum likelihood estimation (AMLE) approach, according to Ait-Sahalia (2002a). Let $p_x(\Delta,\ x|x_0\theta)$ be the conditional density function of $X_{\tau\Delta} = x$ given $X_{(\tau-1)\Delta} = x_0$ induced by model (13). The log-likelihood function of the model for the sample is

$$l_n(\theta) = \sum_{\tau=1}^{n} \ln\ p_X(\Delta,\ X_{\tau\Delta}|X_{(\tau-1)\Delta},\ \theta)$$

The MLE estimator that maximizes $l_n(\theta)$ would be asymptotically most efficient if the conditional density $p_x(\Delta,\ x|x_0,\ \theta)$ has a closed form. Unfortunately, except for some simple models, $p_x(\Delta,\ x|x_0,\ \theta)$ usually does not have a closed form.

Using the Hermite polynomial series, Ait-Sahalia (2002a) proposes a closed-form sequence $\{p_x^{(J)}(\Delta,\ x|x_0,\ \theta)\}$ to approximate $p_x(\Delta,\ x|x_0,\ \theta)$ and then obtains an estimator $\hat{\theta}_n^{(J)}$ that maximizes the approximated model likelihood. The estimator $\hat{\theta}_n^{(J)}$ enjoys the same asymptotic efficiency as the (infeasible) MLE as $J = J_n \to \infty$. More specifically, Ait-Sahalia (2002a) first considers a transformed process:

$$Y_t \equiv \gamma(X_t,\ \theta) = \int_{-\infty}^{X_t} \frac{1}{\sigma(u,\ \theta)} du$$

This transformed process obeys the following diffusion:

$$dY_t = \mu_y(Y_t,\ \theta)dt + dB_t$$

where

$$\mu_y(y,\ \theta) = \frac{\mu[\gamma^{-1}(y,\ \theta),\ \theta]}{\sigma[\gamma^{-1}(y,\ \theta),\ \theta]} - \frac{1}{2}\frac{\partial\sigma[\gamma^{-1}(y,\ \theta),\theta]}{\partial x}$$

The transform $X \to Y$ ensures that the tail of the transition density $p_y(\Delta, y|y_0, \theta)$ of Y_t will generally vanish exponentially fast so that Hermite series approximations will converge. However, $p_y(\Delta, y|y_0, \theta)$ may get peaked at y_0 when the sample frequency Δ gets smaller. To avoid this, Ait-Sahalia (2002a) considers a further transformation as

$$Z_t = \Delta^{-1/2}(Y_t - y_0)$$

and then approximates the transition density of Z_t by the Hermite polynomials:

$$p_z^{(J)}(z|z_0, \theta) = \phi(z) \sum_{j=0}^{J} \eta_z^{(j)}(z_0, \theta) H_j(z)$$

where $\phi(\cdot)$ is the $N(0, 1)$ density, and $\{H_j(z)\}$ is the Hermite polynomial series. The coefficients $\{\eta_z^{(j)}(z_0, \theta)\}$ are specific conditional moments of process Z_t, and can be explicitly computed using the Monte Carlo method or using a higher Taylor series expansion in Δ.

The approximated transition density of X_t is then given as follows:

$$\begin{aligned} p_x(x|x_0, \theta) &= \sigma(x, \theta)^{-1} p_y(\gamma(x, \theta)|\gamma(x, \theta), \theta) \\ &= \Delta^{-1/2} p_z(\Delta^{-1/2}(\gamma(x, \theta) - \gamma(x, \theta))|\gamma(x_0, \theta), \theta) \end{aligned}$$

Under suitable regularity conditions, particularly when $J = J_n \to \infty$ as $n \to \infty$, the estimator

$$\hat{\theta}_n^{(J)} = \arg \min_{\theta \in \Theta} \sum_{\tau=1}^{n} \ln p_x^{(J)}(X_{\tau\Delta}|X_{(\tau-1)\Delta}, \theta)$$

will be asymptotically equivalent to the infeasible MLE. Ait-Sahalia (1999) applies this method to estimate a variety of diffusion models for spot interest rates, and finds that $J = 2$ or 3 already gives accurate approximation for most financial diffusion models. Egorov, Li, and Xu (2003) extend this approach to stationary time-inhomogeneous diffusion models. Ait-Sahalia (2008) extends this method to general multivariate diffusion models and Ait-Sahalia and Kimmel (2007) to affine multifactor term structure models.

In contract to the AMLE in Ait-Sahalia (2002a), Jiang and Knight (2006) consider a more general Markov models where the transition density is unknown. The approach Jiang and Knight (2006) propose is based on the empirical characteristic function estimation procedure with an approximate optimal weight function. The approximate optimal weight function is obtained through an Edgeworth/Gram-Charlier expansion of the

logarithmic transition density of the Markovian process. They derive the estimating equations and demonstrate that they are equivalent to the AMLE as in Ait-Sahalia (2002a). However, in contrast to the common AMLE, their approach ensures the consistency of the estimator even in the presence of approximation error. When the approximation error of the optimal weight function is arbitrarily small, the estimator has MLE efficiency. For details, see Jiang and Knight (2006).

Finally, in a rather general continuous-time setup which allows for stationary multifactor diffusion models with partially observable state variables, Gallant and Tauchen (1996) propose an EMM estimator that also enjoys the asymptotic efficiency as the MLE. The basic idea of EMM is to first use a Hermite polynomial-based semi-nonparametric (SNP) density estimator to approximate the transition density of the observed state variables. This is called the auxiliary model and its score is called the score generator, which has expectation zero under the model-implied distribution when the parametric model is correctly specified. Then, given a parameter setting for the multifactor model, one may use simulation to evaluate the expectation of the score under the stationary density of the model and compute a χ^2 criterion function. A nonlinear optimizer is used to find the parameter values that minimize the proposed criterion.

Specifically, suppose $\{X_t\}$ is a stationary possibly vector valued process such that the true conditional density function $p_0(\Delta,\ X_{\tau\Delta}|X_{s\Delta},\ s \le \tau - 1) = p_0(\Delta,\ X_{\tau\Delta}|Y_{\tau\Delta})$ where $Y_{\tau\Delta} \equiv (X_{(\tau-1)\Delta},\ \ldots,\ X_{(\tau-d)\Delta})^\top$ for some fixed integer $d \ge 0$. This is a Markovian process of order d. To check the adequacy of a parametric model in Eq. (13), Gallant and Tauchen (1996) propose to check whether the following moment condition holds:

$$M(\beta_n,\ \theta) \equiv \int \frac{\partial\ \log f(\Delta,\ x,\ y;\ \beta_n)}{\partial \beta_n} p(\Delta,\ x,\ y;\ \theta) dx dy = 0, \quad \text{if } \theta = \theta_0 \in \Theta \tag{16}$$

where $p(\Delta, x, y; \theta)$ is the model-implied joint density for $(X_{\tau\Delta},\ Y_{\tau\Delta}^\top)^\top$, θ_0 the unknown true parameter value, and $f(\Delta, x, y; \beta_n)$ an auxiliary model for the conditional density of $(X_{\tau\Delta},\ Y_{\tau\Delta}^\top)^\top$. Note that β_n is the parameter vector in the SNP density model $f(\Delta, x, y; \beta_n)$ and generally does not nest the parametric parameter θ. By allowing the dimension of β_n to grow with the sample size n, the SNP density $f(\Delta, x, y; \beta_n)$ will eventually span the true density $p_0(\Delta, x, y)$ of $(X_{\tau\Delta},\ Y_{\tau\Delta}^\top)^\top$, and thus it is free of model misspecification asymptotically. Gallant and Tauchen (1996) use a Hermite polynomial approximation for $f(\Delta, x, y; \beta_n)$, with the dimension of β_n

determined by a model selection criterion such as the Baysian information criterion (BIC). The integration in Eq. (16) can be computed by simulating a large number of realizations under the distribution of the parametric model $p(\Delta, x, y; \theta)$.

The EMM estimator is defined as follows:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} M(\hat{\beta}_n, \theta)^\top \hat{I}^{-1}(\theta) M(\hat{\beta}_n, \theta)$$

where $\hat{\beta}$ is the quasi-MLE estimator for β_n, the coefficients in the Hermite polynomial expansion of the SNP density model $f(x, y, \beta_n)$, and the matrix $\hat{I}(\theta)$ is an estimate of the asymptotic variance of $\sqrt{n}\partial M_n(\hat{\beta}_n, \theta)/\partial\theta$ (Gallant & Tauchen, 2001). This estimator $\hat{\theta}$ is asymptotically as efficient as the (infeasible) MLE.

The EMM has been applied widely in financial applications. See, for example, Andersen and Lund (1997), Dai and Singleton (2000), and Ahn, Dittmar, and Gallant (2002) for interest rate applications; Liu (2000), Andersen, Benzoni, and Lund (2002), Chernov et al. (2003) for estimating stochastic volatility models for stock prices with such complications as long memory and jumps; Chung and Tauchen (2001) for estimating and testing target zero models of exchange rates; Jiang and van der Sluis (2000) for price option pricing; and Valderrama (2001) for a macroeconomic application. It would be interesting to compare the EMM method and Ait-Sahalia's (2002a) approximate MLE in finite sample performance and this topic is still open.

3.2. Nonparametric Testing

In financial applications, most continuous-time models are parametric. It is important to test whether a parametric diffusion model adequately captures the dynamics of the underlying process. Model misspecification generally renders inconsistent estimators of model parameters and their variance–covariance matrix, leading to misleading conclusions in inference and hypothesis testing. More importantly, a misspecified model can yield large errors in hedging, pricing, and risk management.

Unlike the vast literature of estimation of parametric diffusion models, there are relatively few test procedures for parametric diffusion models using discrete observations. Suppose $\{X_t\}$ follows a continuous-time diffusion process in Eq. (6). Often it is assumed that the drift and diffusion $\mu(\cdot, t)$ and $\sigma(\cdot, t)$ have some parametric forms $\mu(\cdot, t, \theta)$ and $\sigma(\cdot, t, \theta)$, where

$\theta \in \Theta$. We say that models $\mu(\cdot, t, \theta)$ and $\sigma(\cdot, t, \theta)$ are correctly specified for the drift and diffusion $\mu(\cdot, t)$ and $\sigma(\cdot, t)$, respectively, if

$$\mathbb{H}_0 : P[\mu(X_t, t, \theta_0) = \mu(X_t, t), \sigma(X_t, t, \theta_0) = \sigma(X_t, t)] = 1 \quad \text{for some } \theta_0 \in \Theta \tag{17}$$

As noted earlier, various methods have been developed to estimate θ_0, taking Eq. (17) as given. However, these methods generally cannot deliver consistent parameter estimates if $\mu(\cdot, t, \theta)$ or $\sigma(\cdot, t, \theta)$ is misspecified in the sense that

$$\mathbb{H}_a : P[\mu(X_t, t, \theta) = \mu(X_t, t), \sigma(X_t, t, \theta) = \sigma(X_t, t)] < 1 \quad \text{for all } \theta \in \Theta \tag{18}$$

Under $\mathbb{H}_a$ of Eq. (18), there exists no parameter value $\theta \in \Theta$ such that the drift model $\mu(\cdot, t, \theta)$ and the diffusion model $\sigma(\cdot, t, \theta)$ coincide with the true drift $\mu(\cdot, t)$ and the true diffusion $\sigma(\cdot, t)$, respectively.

There is a growing interest in testing whether a continuous-time model is correctly specified using a discrete sample $\{X_{\tau\Delta}\}_{\tau=1}^{n}$. Next we will present some test procedures for testing the continuous-time models. Ait-Sahalia (1996b) observes that for a stationary time-homogeneous diffusion process in Eq. (13), a pair of drift and diffusion models $\mu(\cdot, \theta)$ and $\sigma(\cdot, \theta)$ uniquely determines the stationary density $\pi(\cdot, \theta)$ in Eq. (15). Ait-Sahalia (1996b) compares a parametric marginal density estimator $\pi(\cdot, \hat{\theta})$ with a nonparametric density estimator $\hat{\pi}_0(\cdot)$ via the quadratic form:

$$M \equiv \int_{x_0^*}^{x_1^*} [\hat{\pi}_0(x) - \pi(x, \hat{\theta})]^2 \hat{\pi}_0(x) dx \tag{19}$$

where x_1^* is the upper bound for X_t, $\hat{\theta}$ the minimum distance estimator given by Eq. (14). The M statistic, after demeaning and scaling, is asymptotically normal under $\mathbb{H}_0$.

The M test makes no restrictive assumptions on the data-generating process and can detect a wide range of alternatives. This appealing power property is not shared by parametric approaches such as GMM tests (e.g., Conley et al., 1997). The latter has optimal power against certain alternatives (depending on the choice of moment functions) but may be completely silent against other alternatives. In an application to Euro-dollar interest rates, Ait-Sahalia (1996b) rejects all existing one-factor linear drift models using asymptotic theory and finds that "the principal source of rejection of existing models is the strong nonlinearity of the drift," which is further supported by Stanton (1997).

However, several limitations of this test may hinder its empirical applicability. First, as Ait-Sahalia (1996b) has pointed out, the marginal density cannot capture the full dynamics of $\{X_t\}$. It cannot distinguish two diffusion models that have the same marginal density but different transition densities.[5] Second, subject to some regularity conditions, the asymptotic distribution of the quadratic form M in Eq. (19) remains the same whether the sample $\{X_{\tau\Delta}\}_{\tau=1}^{n}$ is iid or highly persistently dependent (Ait-Sahalia, 1996b). This convenient asymptotic property unfortunately results in a substantial discrepancy between the asymptotic and finite sample distributions, particularly when the data display persistent dependence (Pritsker, 1998). This discrepancy and the slow convergence of kernel estimators are the main reasons identified by Pritsker (1998) for the poor finite sample performance of the M test. They cast some doubts on the applicability of first-order asymptotic theory of nonparametric methods in finance, since persistent serial dependence is a stylized fact for interest rates and many other high-frequency financial data. Third, a kernel density estimator produces biased estimates near the boundaries of the data (e.g., Härdle, 1990, and Fan & Gijbels, 1996). In the present context, the boundary bias can generate spurious nonlinear drifts, giving misleading conclusions on the dynamics of $\{X_t\}$.

Recently, Hong and Li (2005) have developed a nonparametric test for the model in Eq. (6) using the transition density, which can capture the full dynamics of $\{X_t\}$ in Eq. (13). Let $p_0(x, t|x_0, s)$ be the true transition density of the diffusion process X_t, that is, the conditional density of $X_t = x$ given $X_s = x_0$, $s<t$. For a given pair of drift and diffusion models $\mu(\cdot, t, \theta)$ and $\sigma(\cdot, t, \theta)$, a certain family of transition densities $\{p(x, t|x_0, s, \theta)\}$ is characterized. When (and only when) $\mathbb{H}_0$ in Eq. (17) holds, there exists some $\theta_0 \in \Theta$ such that $p(x, t|x_0, s, \theta_0) = p_0(x, t|x_0, s)$ almost everywhere for all $t>s$. Hence, the hypotheses of interest $\mathbb{H}_0$ in Eq. (17) versus $\mathbb{H}_a$ in Eq. (18) can be equivalently written as follows:

$$\mathbb{H}_0 : p(x, t|y, s, \theta_0) = p_0(x, t|y, s) \quad \text{almost everywhere for some } \theta_0 \in \Theta \tag{20}$$

versus the alternative hypothesis:

$$\mathbb{H}_a : p(x, t|y, s, \theta) \neq p_0(x, t|y, s) \quad \text{for some } t>s \quad \text{and for all } \theta \in \Theta \tag{21}$$

Clearly, to test $\mathbb{H}_0$ in Eq. (20) versus $\mathbb{H}_a$ in Eq. (21) would be to compare a model transition density estimator $p(x, t|x_0, s, \hat{\theta})$ with a nonparametric transition density estimator, say $\hat{p}_0(x, t|x_0, s)$. Instead of comparing

$p(x, t|x_0, s, \hat{\theta})$ and $\hat{p}_0(x, t|x_0, s)$ directly, Hong and Li (2005) first transform $\{X_{\tau\Delta}\}_{\tau=1}^n$ via a probability integral transformation. Define a discrete transformed sequence

$$Z_\tau(\theta) \equiv \int_{-\infty}^{X\tau\Delta} p[x, \tau\Delta|X_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta]dx, \quad \tau = 1, \ldots, n \tag{22}$$

Under (and only under) $\mathbb{H}_0$ in Eq. (20), there exists some $\theta_0 \in \Theta$ such that $p[x, \tau\Delta|X_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta_0] = p_0[x, \tau\Delta|X_{(\tau-1)\Delta}, (\tau-1)\Delta]$ almost surely for all $\Delta > 0$. Consequently, the transformed series $\{Z_\tau \equiv Z_\tau(\theta_0)\}_{\tau=1}^n$ is iid $U[0, 1]$ under $\mathbb{H}_0$ in Eq. (20). This result is first proven, in a simpler context, by Rosenblatt (1952), and is more recently used to evaluate out-of-sample density forecasts (e.g., Diebold, Gunther, & Tay, 1998) in a discrete-time context. Intuitively, we may call $\{Z_\tau(\theta)\}$ "generalized residuals" of the model $p(x, t|y, s, \theta)$.

To test $\mathbb{H}_0$ in Eq. (20), Hong and Li (2005) check whether $\{Z_\tau\}_{\tau=1}^n$ is both iid and $U[0, 1]$. They compare a kernel estimator $\hat{g}_j(z_1, z_2)$ defined in Eq. (23) below for the joint density of $\{Z_\tau, Z_{\tau-j}\}$ with unity, the product of two $U[0, 1]$ densities. This approach has at least three advantages. First, since there is no serial dependence in $\{Z_\tau\}$ under $\mathbb{H}_0$ in Eq. (20), nonparametric joint density estimators are expected to perform much better in finite samples. In particular, the finite sample distribution of the resulting tests is expected to be robust to persistent dependence in data. Second, there is no asymptotic bias for nonparametric density estimators under $\mathbb{H}_0$ in Eq. (20). Third, no matter whether $\{X_t\}$ is time inhomogeneous or even nonstationary, $\{Z_\tau\}$ is always iid $U[0, 1]$ under correct model specification.

Hong and Li (2005) employ the kernel joint density estimator:

$$\hat{g}_j(z_1, z_2) \equiv (n-j)^{-1} \sum_{\tau=j+1}^{n} K_h(z_1, \hat{Z}_\tau) K_h(z_2, \hat{Z}_{\tau-j}), \quad j > 0 \tag{23}$$

where $\hat{Z}_\tau = Z_\tau(\hat{\theta})$, $\hat{\theta}$ is any $\sqrt{n}$-consistent estimator for θ_0, and for $x \in [0, 1]$,

$$K_h(x, y) \equiv \begin{cases} h^{-1}k\left(\dfrac{x-y}{h}\right) \Big/ \int_{-(x/h)}^{1} k(u)du, & \text{if } x \in [0, h) \\ h^{-1}k\left(\dfrac{x-y}{h}\right), & \text{if } x \in [h, 1-h] \\ h^{-1}k\left(\dfrac{x-y}{h}\right) \Big/ \int_{-1}^{(1-x)/h} k(u)du, & \text{if } x \in (1-h, 1] \end{cases}$$

is the kernel with boundary correction (Rice, 1986) and $k(\cdot)$ is a standard kernel. This avoids the boundary bias problem, and has some advantages

over some alternative methods such as trimming and the use of the jackknife kernel.[6] To avoid the boundary bias problem, one might apply other kernel smoothing methods such as local polynomial (Fan & Gijbels, 1996) or weighted NW (Cai, 2001).

Hong and Li's (2005) test statistic is

$$\hat{Q}(j) \equiv \frac{\left[(n-j)h \int_0^1 \int_0^1 [\hat{g}_j(z_1, z_2) - 1]^2 dz_1 dz_2 - A_h^0\right]}{V_0^{1/2}}$$

where A_h^0 and V_0 are non-stochastic centering and scale factors which are functions of h and $k(\cdot)$.

In a simulation experiment mimicking the dynamics of US interest rates via the Vasicek model, Hong and Li (2005) find that $\hat{Q}(j)$ has rather reasonable sizes for $n = 500$ (i.e., about two years of daily data). This is a rather substantial improvement over Ait-Sahalia's (1996b) test, in lights of Pritsker's (1998) simulation evidence. Moreover, $\hat{Q}(j)$ has better power than the marginal density test. Hong and Li (2005) find extremely strong evidence against a variety of existing one-factor diffusion models for the spot interest rate and affine models for interest rate term structures. Egorov, Hong, and Li (2006) have recently extended Hong and Li (2005) to evaluate out of sample of density forecasts of a multivariate diffusion model possibly with jumps and partially unobservable state variables.

Because the transition density of a continuous-time model generally has no closed form, the probability integral transform $\{Z_\tau(\theta)\}$ in Eq. (22) is difficult to compute. However, one can approximate the model transition density using the simulation methods developed by Pedersen (1995), Brandt and Santa-Clara (2002), and Elerian, Chib, and Shephard (2001). Alternatively, we can use Ait-Sahalia's (2002a) Hermite expansion method to construct a closed-form approximation of the model transition density.

When a misspecified model is rejected, one may like to explore what are the possible sources for the rejection. For example, is the rejection due to misspecification in the drift, such as the ignorance of mean shifts or jumps? Is it due to the ignorance of GARCH effects or stochastic volatility? Or is it due to the ignorance of asymmetric behaviors (e.g., leverage effects)? Hong and Li (2005) consider to examine the autocorrelations in the various powers of $\{Z_\tau\}$, which are very informative about how well a model fits various dynamic aspects of the underlying process (e.g., conditional mean, variance, skewness, kurtosis, ARCH-in-mean effect, and leverage effect).

Gallant and Tauchen (1996) also propose an EMM-based minimum χ^2 specification test for stationary continuous-time models. They examine the simulation-based expectation of an auxiliary SNP score function under the model distribution, which is zero under correct model specification. The greatest appeal of the EMM approach is that it applies to a wide range of stationary continuous-time processes, including both one-factor and multifactor diffusion processes with partially observable state variables (e.g., stochastic volatility models). In addition to the minimum χ^2 test for generic model misspecifications, the EMM approach also provides a class of individual *t*-statistics that are informative in revealing possible sources of model misspecification. This is perhaps the most appealing strength of the EMM approach.

Another feature of the EMM tests is that all EMM test statistics avoid estimating long-run variance–covariances, thus resulting in reasonable finite sample size performance (cf. Andersen, Chung, & Sorensen, 1999). In practice, however, it may not be easy to find an adequate SNP density model for financial time series, as is shown in Hong and Lee (2003b). For example, Andersen and Lund (1997) find that an AR(1)-EGARCH model with a number of Hermite polynomials adequately captures the full dynamics of daily S&P 500 return series, using a BIC criterion. However, Hong and Lee (2003a) find that there still exists strong evidence on serial dependence in the standardized residuals of the model, indicating that the auxiliary SNP model is inadequate. This affects the validity of the EMM tests, because their asymptotic variance estimators have exploited the correct specification of the SNP density model.[7]

There has also been an interest in separately testing the drift model and the diffusion model in Eq. (13). For example, it has been controversial whether the drift of interest rates is linear. To test the linearity of the drift term, one can write it as a functional coefficient form (Cai et al., 2000) $\mu(X_t) = \alpha_0(X_t)+\alpha_1(X_t)X_t$. Then, the null hypothesis is $\mathbb{H}_0$: $\alpha_0(\cdot) \equiv \alpha_0$ and $\alpha_1(\cdot) \equiv \alpha_1$. Fan and Zhang (2003) apply the generalized likelihood ratio test developed by Cai et al. (2000) and Fan et al. (2001). They find that $\mathbb{H}_0$ is not rejected for the short-term interest rates. It is noted that the asymptotic theory for the generalized likelihood ratio test is developed for the iid samples, but it is still unknown whether it is valid for a time series context. One might follow the idea from Cai et al. (2000) to use the bootstrap or wild bootstrap method instead of the asymptotic theory for time series context. Fan and Zhang (2003) and Fan et al. (2003) conjecture that it would hold based on their simulations. On the other hand, Chen, Härdle, and Kleinow (2002) consider an empirical likelihood goodness-of-fit test for time series

regression model, and they apply the test to test a discrete drift model of a diffusion process.

There has also been interest in testing the diffusion model $\sigma(\cdot, \theta)$. The motivation comes from the fact that derivative pricing with an underlying equity process only depends on the diffusion $\sigma(\cdot)$, which is one of the most important features of Eq. (13) for derivative pricing. Kleinow (2002) recently proposes a nonparametric test for a diffusion model $\sigma(\cdot)$. More specifically, Kleinow (2002) compares a nonparametric diffusion estimator $\hat{\sigma}^2(\cdot)$ with a parametric diffusion estimator $\sigma^2(\cdot, \theta)$ via an asymptotically χ^2 test statistic

$$\hat{T}_\lambda = \sum_{t=1}^{k} [\hat{T}(x_t)]^2$$

where

$$\hat{T}(x) = [nh\hat{\pi}(x)]^{1/2} \left[\hat{\sigma}^2(x)/\hat{\sigma}^2(x, \hat{\theta}) - 1 \right]$$

$\hat{\theta}$ is an $\sqrt{n}$-consistent estimator for θ_0 and

$$\hat{\sigma}^2(x, \theta) = \frac{1}{nh\hat{\pi}(x)} \sum_{t=1}^{n} \sigma^2(x, \hat{\theta}) K_h \left[\frac{x - X_t}{h} \right]$$

is a smooth version of $\sigma^2(x, \theta)$. The use of $\hat{\sigma}^2(x, \hat{\theta})$ instead of $\sigma^2(x, \hat{\theta})$ directly reduces the kernel estimation bias in $\hat{T}(x)$, thus allowing the use of the optimal bandwidth h for $\hat{\sigma}^2(x)$. This device is also used in Härdle and Mammen (1993) in testing a parametric regression model. Kleinow (2002) finds that the empirical level of $\hat{T}_k$ is too large relative to the significance level in finite samples and then proposes a modified test statistic using the empirical likelihood approach, which endogenously studentizes conditional heteroscedasticity. As expected, the empirical level of the modified test improves in finite samples, though not necessarily for the power of the test.

Furthermore, Fan et al. (2003) test whether the coefficients in the time-varying coefficient single factor diffusion model of Eq. (7) are really time varying. Specially, they apply the generalized likelihood ratio test to check whether some or all of $\{\alpha_j(\cdot)\}$ and $\{\beta_j(\cdot)\}$ are constant. However, the validity of the generalized likelihood ratio test for nonstationary time series is still unknown and it needs a further investigation.

Finally, Kristensen (2008) considers an estimation method for two classes of semiparametric scalar diffusion models. In the first class, the diffusion term is parameterized and the drift is left unspecified, while in the second

class, only the drift term is specified. Under the assumption of stationarity, the unspecified term can be identified as a function of the parametric component and the stationary density. Given a discrete sample with a fixed time distance, the parametric component is then estimated by maximizing the associated likelihood with a preliminary estimator of the unspecified term plugged in. Kristensen (2008) shows that this pseudo-MLE (PMLE) is $\sqrt{n}$-consistent with an asymptotically normal distribution under regularity conditions, and demonstrates how the estimator can be used in specification testing not only of the semiparametric model itself but also of fully parametric ones. Since the likelihood function is not available on closed form, the practical implementation of the proposed estimator and tests will rely on simulated or approximate PMLE. Under regularity conditions, Kristensen (2008) verifies that the approximate/simulated version of the PMLE inherits the properties of the actual but infeasible estimator. Also, Kristensen (2007) proposes a nonparametric kernel estimator of the drift (diffusion) term in a diffusion model based on a preliminary parametric estimator of the diffusion (drift) term. Under regularity conditions, rates of convergence and asymptotic normality of the nonparametric estimators are established. Moreover, Kristensen (2007) develops misspecification tests of diffusion models based on the nonparametric estimators, and derives the asymptotic properties of the tests. Furthermore, Kristensen (2007) proposes a Markov bootstrap method for the test statistics to improve on the finite sample approximations.

4. NONPARAMETRIC PRICING KERNEL MODELS

In modern finance, the pricing of contingent claims is important given the phenomenal growth in turnover and volume of financial derivatives over the past decades. Derivative pricing formulas are highly nonlinear even when they are available in a closed form. Nonparametric techniques are expected to be very useful in this area. In a standard dynamic exchange economy, the equilibrium price of a security at date t with a single liquidating payoff $Y(C_T)$ at date T, which is a function of aggregate consumption C_T, is given by

$$P_t = E_t[Y(C_T)M_{t,T}] \tag{24}$$

where the conditional expectation is taken with respect to the information set available to the representative economic agent at time t, $M_{t,T} = \delta^{T-1}U'(C_T)/U'(C_t)$, the so-called stochastic discount factor (SDF), is the

marginal rate of substitution between dates t and T, δ the rate of time preference; and $U(\cdot)$ the utility function of the economic agent. This is the stochastic Euler equation, or the first-order condition of the intertemporal utility maximization of the economic agent with suitable budget constraints (e.g., Cochrane, 1996, 2001). It holds for all securities, including assets and various derivatives. All capital asset pricing (CAP) models and derivative pricing models can be embedded in this unified framework – each model can be viewed as a specific specification of $M_{t,T}$. See Cochrane (1996, 2001) for an excellent discussion.

There have been some parametric tests for CAP models (e.g., Hansen & Janaganan, 1997). To the best of our knowledge, there are only a few nonparametric tests available in the literature for testing CAP models based on the kernel method, see Wang (2002, 2003) and Cai, Kuan and Sun (2008a, 2008b), which will be elaborated in detail in Section 4.3 later. Also, all the tests for CAP models are formulated in terms of discrete-time frameworks. We focus on nonparametric derivative pricing in Section 4.2 and the nonparametric asset pricing will be discussed separately in Section 4.3.

4.1. Nonparametric Risk Neutral Density

Assuming that the conditional distribution of future consumption C_T has a density representation $f_t(\cdot)$, then the conditional expectation can be expressed as

$$E_t[Y(C_T)M_{t,T}] = \exp(-\tau r_t)\int Y(C_T)f_t^*(C_T)dC_T = \exp(-\tau r_t)E_t^*[Y(C_t)]$$

where r_t is the risk-free interest rate, $\tau = T-t$, and

$$f_t^*(C_T) = \frac{M_{t,T}f_t(C_T)}{\int M_{t,T}f_t(C_T)dC_T}$$

is called the RND function; see Taylor (2005, Chapter 16) for details about the definition and estimation methods. This function is also called the risk-neutral pricing probability (Cox & Ross, 1976), or equivalent martingale measure (Harrison & Kreps, 1979), or the state-price density (SPD). It contains rich information on the pricing and hedging of risky assets in an economy, and can be used to price other assets, or to recover the information about the market preferences and asset price dynamics (Bahra, 1997; Jackwerth, 1999). Obviously, the RND function differs from $f_t(C_T)$, the physical density function of C_T conditional on the information available at time t.

4.2. Nonparametric Derivative Pricing

In order to calculate an option price from Eq. (24), one has to make some assumption on the data-generating process of the underlying asset, $\{P_t\}$. For example, Black and Scholes (1973) assume that the underlying asset follows a geometric Brownian motion:

$$dP_t = \mu P_t dt + \sigma P_t dB_t$$

where μ and σ are two constants. Applying Ito's Lemma, one can show immediately that P_τ follows a lognormal distribution with parameter $(\mu - \frac{1}{2}\sigma^2)\tau$ and $\sigma\sqrt{\tau}$. Using a no-arbitrage argument, Black and Scholes (1973) show that options can be priced if investors are risk neutral by setting the expected rate of return in the underlying asset, μ, equal to the risk-free interest rate, r. Specifically, the European call option price is

$$\pi(K_t,\ P_t,\ r,\ \tau) = P_t\Phi(d_t) - e^{-r_t\tau}K_t\Phi(d_t - \sigma\sqrt{\tau}) \tag{25}$$

where K_t is the strike price, $\Phi(\cdot)$ the standard normal cumulative distribution function, and $d_t = \{\ln(P_t/K_t) + (r + \frac{1}{2}\sigma^2)\tau\}/(\sigma\sqrt{\tau})$. In Eq. (25), the only parameter that is not observable a time t is σ. This parameter, when multiplied with $\sqrt{\tau}$, is the underlying asset return volatility over the remaining life of the option. The knowledge of σ can be inferred from the prices of options traded in the markets: given an observed option price, one can solve an appropriate option pricing model for σ which is essentially a market estimate of the future volatility of the underlying asset returns. This estimate of σ is known as "implied volatility."

The most important implication of Black–Scholes option pricing model is that when the option is correctly priced, the implied volatility σ^2 should be the same across all exercise prices of options on the same underlying asset and with the same maturity date. However, the implied volatility observed in the market is usually a convex function of exercise price, which is often referred to as the "volatility smile." This indicates that market participants make more complicated assumptions than the geometric Brownian motion for the dynamics of the underlying asset. In particular, the convexity of "volatility smile" indicates the degree to which the market RND function has a heavier tail than a lognormal density. A great deal of effort has been made to use alternative models for the underlying asset to smooth out the volatility smile and so to achieve higher accuracy in pricing and hedging.

A more general approach to derivative pricing is to estimate the RND function directly from the observed option prices and then use it to price

derivatives or to extract market information. To obtain better estimation of the RND function, several econometric techniques have been introduced. These methods are all based on the following fundamental relation between option prices and RNDs: Suppose $G_t = G(K_t, P_t, r_t, \tau)$ is the option pricing formula, then there is a close relation between the second derivative of G_t with respect to the strike price K_t and the RND function:

$$\frac{\partial^2 G_t}{\partial K_t^2} = \exp(-\tau r_t) f_t^*(P_T) \tag{26}$$

This is first shown by Breeden and Litzenberger (1978) in a time-state preference framework.

Most commonly used estimation methods for RNDs are various parametric approaches. One of them is to assume that the underlying asset follows a parametric diffusion process, from which one can obtain the option pricing formula by a no-arbitrage argument, and then obtain the RND function from Eq. (26) (see, e.g., Bates, 1991, 2000; Anagnou, Bedendo, Hodges, & Tompkins, 2005). Another parametric approach is to directly impose some form for the RND function and then estimate unknown parameters by minimizing the distance between the observed option prices and those generated by the assumed RND function (e.g., Jackwerth & Rubinstein, 1996; Melick & Thomas, 1997; Rubinstein, 1994). A third parametric approach is to assume a parametric form for the call pricing function or the implied volatility smile curve and then apply Eq. (26) to get the RND function (Bates, 1991; Jarrow & Tudd, 1982; Longstaff, 1992, 1995; Shimko, 1993).

The aforementioned parametric approaches all impose certain restrictive assumptions, directly or indirectly, on the data-generating process as well as the SDF in some cases. The obtained RND function is not robust to the violation of these restrictions. To avoid this drawback, Ait-Sahalia and Lo (1998) use a nonparametric method to extract the RND function from option prices.

Given observed call option prices $\{G_t, K_t, \tau\}$, the price of the underlying asset $\{P_t\}$, and the risk-free rate of interest $\{r_t\}$, Ait-Sahalia and Lo (1998) construct a kernel estimator for $E(G_t|P_t, K_t, \tau, r_t)$. Under standard regularity conditions, Ait-Sahalia and Lo (1998) show that the RND estimator is consistent and asymptotically normal, and they provide explicit expressions for the asymptotic variance of the estimator.

Armed with the RND estimator, Ait-Sahalia and Lo (1998) apply it to the pricing and delta hedging of S&P 500 call and put options using daily data

obtained from the Chicago Board Options Exchange for the sample period from January 4, 1993 to December 31, 1993. The RND estimator exhibits negative skewness and excess kurtosis, a common feature of historical stock returns. Unlike many parametric option pricing models, the RND-generated option pricing formula is capable of capturing persistent "volatility smiles" and other empirical features of market prices. Ait-Sahalia and Lo (2000) use a nonparametric RND estimator to compute the economic value at risk, that is, the value at risk of the RND function.

The artificial neural network (ANN) has received much attention in economics and finance over the last decade. Hutchinson, Lo, and Poggio (1994), Anders, Korn, and Schmitt (1998), and Hanke (1999) have successfully applied the ANN models to estimate pricing formulas of financial derivatives. In particular, Hutchinson et al. (1994) use the ANN to address the following question: If option prices are truly determined by the Black–Scholes formula exactly, can ANN "learn" the Black–Scholes formula? In other words, can the Black–Scholes formula be estimated nonparametrically via learning networks with a sufficient degree of accuracy to be of practical use? Hutchinson et al. (1994) perform Monte Carlo simulation experiments in which various ANNs are trained on artificially generated Black–Scholes formula and then compare to the Black–Scholes formula both analytically and in out-of-sample hedging experiments. They begin by simulating a two-year sample of daily stock prices, and creating a cross-section of options each day according to the rules used by the Chicago Broad Options Exchange with prices given by the Black–Scholes formula. They find that, even with training sets of only six months of daily data, learning network pricing formulas can approximate the Black–Scholes formula with reasonable accuracy. The nonlinear models obtained from neural networks yield estimated option prices and deltas that are difficult to distinguish visually from the true Black–Scholes values.

Based on the economic theory of option pricing, the price of a call option should be a monotonically decreasing convex function of the strike price and the SPD proportional to the second derivative of the call function (see Eq. (26)). Hence, the SPD is a valid density function over future values of the underlying asset price and must be nonnegative and integrate to one. Therefore, Yatchew and Härdle (2006) combine shape restrictions with nonparametric regression to estimate the call price function and the SPD within a single least squares procedure. Constraints include smoothness of various order derivatives, monotonicity and convexity of the call function, and integration to one of the SPD. Confidence intervals and test procedures

are to be implemented using bootstrap methods. In addition, they apply the procedures to option data on the DAX index.

There are several directions of further research on nonparametric estimation and testing of RNDs for derivative pricing. First, how to evaluate the quality of an RND function estimated from option prices? In other words, how to judge how well an estimated RND function reflects the market expected uncertainty of the underlying asset? Because the RND function differs from the physical probability density function of the underlying asset, the valuation of the RND function is rather challenging. The method developed by Hong and Li (2005) cannot be applied directly. One possible way of evaluating the RND function is to assume a certain family of utility functions for the representative investor, as in Rubinstein (1994) and Anagnou et al. (2005). Based on this assumption, one can obtain the SDF and then the physical probability density function, to which Hong and Li's (2005) test can be applied. However, the utility function of the economic agent is not observable. Thus, when the test delivers a rejection, it may be due to either misspecification of the utility function or misspecification of the data-generating process, or both. More fundamentally, it is not clear whether the economy can be regarded as a proxy by a representative agent.

A practical issue in recovering the RND function is the limitation of option prices data with certain common characterizations. In other words, the sample size of option price data could be small in many applications. As a result, nonparametric methods should be carefully developed to fit the problems on hand.

Most econometric techniques to estimate the RND function is restricted to European options, while many of the more liquid exchange-traded options are often American. Rather complex extensions of the existing methods, including the nonparametric ones, are required in order to estimate the RND functions from the prices of American options. This is an interesting and practically important direction for further research.

4.3. Nonparametric Asset Pricing

The CAP model and the arbitrage asset pricing theory (APT) have been cornerstones in theoretical and empirical finance for decades. A classical CAP model usually assumes a simple and stable linear relationship between an asset's systematic risk and its expected return; see the books by Campbell et al. (1997) and Cochrane (2001) for details. However, this simple relationship assumption has been challenged and rejected by several

recent studies based on empirical evidences of time variation in betas and expected returns (as well as return volatilities). As with other models, one considers the conditional CAP models or nonlinear APT with time-varying betas to characterize the time variations in betas and risk premia. In particular, Fama and French (1992, 1993, 1995) use some instrumental variables such as book-to-market equity ratio and market equity as proxies for some unidentified risk factors to explain the time variation in returns. Although Ferson (1989), Harvey (1989), Ferson and Harvey (1991, 1993, 1998, 1999), Ferson and Korajczyk (1995), and Jagannathan and Wang (1996) conclude that beta and market risk premium vary over time, a static CAP model should incorporate time variations in beta in the model. Although there is a vast amount of empirical evidences on time variation in betas and risk premia, there is no theoretical guidance on how betas and risk premia vary with time or variables that represent conditioning information. Many recent studies focus on modeling the variation in betas using continuous approximation and the theoretical framework of the conditional CAP models; see Cochrane (1996), Jagannathan and Wang (1996, 2002), Wang (2002, 2003), Ang and Liu (2004), and the references therein. Recently, Ghysels (1998) discusses the problem in detail and stresses the impact of misspecification of beta risk dynamics on inference and estimation. Also, he argues that betas change through time very slowly and linear factor models like the conditional CAP model may have a tendency to overstate the time variation. Further, Ghysels (1998) shows that among several well-known time-varying beta models, a serious misspecification produces time variation in beta that is highly volatile and leads to large pricing errors. Finally, Ghysels (1998) concludes that it is better to use the static CAP model in pricing when we do not have a proper model to capture time variation in betas correctly.

It is well documented that large pricing errors could be due to the linear approach used in a nonlinear model, and treating a nonlinear relationship as a linear could lead to serious prediction problems in estimation. To overcome these problems, some nonlinear models have been considered in the recent literature. Following are some examples: Bansal, Hsieh, and Viswanathan (1993) and Bansal and Viswanathan (1993) advocate the idea of a flexible SDF model in empirical asset pricing, and they focus on nonlinear arbitrage pricing theory models by assuming that the SDF is a nonlinear function of a few state variables. Further, Akdeniz, Altay-Salih, and Caner (2003) test for the existence of significant evidence of nonlinearity in the time series relationship of industry returns with market returns using the heteroskedasticity consistent Lagrange multiplier test of Hansen (1996)

under the framework of the threshold model, and they find that there exists statistically significant nonlinearity in this relationship with respect to real interest rates. Wang (2002, 2003) explores a nonparametric form of the SDF model and conducted a test based on the nonparametric model. Parametric models for time-varying betas can be the most efficient if the underlying betas are correctly specified. However, a misspecification may cause serious bias, and model constraints may distort the betas in local area.

To follow the notions from Bansal et al. (1993), Bansal and Viswanathan (1993), Ghysels (1998), and Wang (2002, 2003), which are slightly different from those used in Eq. (24), a very simplified version of the SDF framework for asset pricing admits a basic pricing representation, which is a special case of model (24),

$$E[m_{t+1}r_{i,t+1}|\Omega_t] = 0 \tag{27}$$

where Ω_t denotes the information set at time t, m_{t+1} the SDF or the pricing kernel, and $r_{i,t+1}$ the excess return on the ith asset or portfolio. Here, $\varepsilon_{t+1} = m_{t+1}r_{i,t+1}$ is called the pricing error. In empirical finance, different models impose different constraints on the SDF. Particularly, the SDF is usually assumed to be a linear function of factors in various applications and then it becomes the well-known CAP model, see Jagannathan and Wang (2002) and Wang (2003). Indeed, Jagannathan and Wang (2002) give the detailed comparison of the SDF and CAP model representations. Further, when the SDF is fully parameterized such as linear form, the general method of moments (GMM) of Hansen (1982) can be used to estimate parameters and test the model; see Campbell et al. (1997) and Cochrane (2001) for details.

Recently, Bansal et al. (1993) and Bansal and Viswanathan (1993) assume that m_{t+1} is a nonlinear function of a few state variables. Since the exact form of the nonlinear pricing kernel is unknown, Bansal and Viswanathan (1993) suggest using the polynomial expansion to approximate it and then apply the GMM for estimating and testing. As pointed out by Wang (2003), although this approach is intuitive and general, one of the shortcomings is that it is difficult to obtain the distribution theory and the effective assessment of finite sample performance. To overcome this difficulty, instead of considering the nonlinear pricing kernel, Ghysels (1998) focuses on the nonlinear parametric model and uses a set of moment conditions suitable for GMM estimation of parameters involved. Wang (2003) studies the nonparametric conditional CAP model and gives an explicit expression for the pricing kernel m_{t+1}, that is, $m_{t+1} = 1 - b(Z_t)r_{p,t+1}$, where Z_t is a $k \times 1$

vector of conditioning variables from Ω_t, $b(Z_t) = E(r_{p,t+1}|Z_t)/E(r^2_{p,t+1}|Z_t)$ which is an unknown function, and $r_{p,t+1}$ is the return on the market portfolio in excess of the riskless rate. Since the functional form of $b(\cdot)$ is unknown, Wang (2003) suggests estimating $b(\cdot)$ by using the NW method to two regression functions $E(r_{p,t+1}|Z_t)$ and $E(r^2_{p,t+1}|Z_t)$. Also, he conducts a simple nonparametric test about the pricing error. Indeed, his test is the well-known F-test by running a multiple regression of the estimated pricing error $\hat{\varepsilon}_{t+1}$ versus a group of information variables; see Eq. (32) later for details. Further, Wang (2003) extends this setting to multifactor models by allowing $b(\cdot)$ to change over time, that is, $b(Z_t) = b(t)$. Finally, Bansal et al. (1993), Bansal and Viswanathan (1993), and Ghysels (1998) do not assume that m_{t+1} is a linear function of $r_{p,t+1}$ and instead they consider a parametric model by using the polynomial expansion.

To combine the models studied by Bansal et al. (1993), Bansal and Viswanathan (1993), Ghysels (1998), and Wang (2002, 2003), and some other models in the finance literature under a very general framework, Cai, Kuan, and Sun (2008a) assume that the nonlinear pricing kernel has the form of $m_{t+1} = 1 - m(Z_t)r_{p,t+1}$, where $m(\cdot)$ is unspecified and they focus on the following nonparametric APT model:

$$E[\{1 - m(Z_t)r_{p,t+1}\}r_{i,t+1}|\Omega_t] = 0 \tag{28}$$

where $m(\cdot)$ is an unknown function of Z_t which is a $k \times 1$ vector of conditioning variables from Ω_t. Indeed, Eq. (28) can be regarded as a moment (orthogonal) condition. The main interest of Eq. (28) is to identify and estimate the function $m(Z_t)$ as well as test whether the model is correctly specified.

Let I_t be a $q \times 1$ ($q \geq k$) vector of conditional variables from Ω_t, including Z_t, satisfying the following orthogonal condition:

$$E[\{1 - m(Z_t)r_{p,t+1}\}r_{i,t+1}|I_t] = 0 \tag{29}$$

which can be regarded as an approximation of Eq. (28). It follows from the orthogonality condition in Eq. (29) that for any vector function $Q(V_t) \equiv Q_t$ with a dimension d_q specified later,

$$E[Q_t\{1 - m(Z_t)r_{p,t+1}\}r_{i,t+1}|I_t] = 0$$

and its sample version is

$$\frac{1}{T}\sum_{t=1}^{T} Q_t\{1 - m(Z_t)r_{p,t+1}\}r_{i,t+1} = 0 \tag{30}$$

Therefore, Cai et al. (2008a) propose a new nonparametric estimation procedure to combine the orthogonality conditions given in Eq. (30) with the local linear fitting scheme of Fan and Gijbels (1996) to estimate the unknown function $m(\cdot)$. This nonparametric estimation approach is called by Cai et al. (2008a) as the nonparametric generalized method of moment (NPGMM).

For a given grid point z_0 and $\{Z_t\}$ in a neighborhood of z_0, the orthogonality conditions in Eq. (30) can be approximated by the following locally weighted orthogonality conditions:

$$\sum_{t=1}^{T} Q_t[1-(a-b^T(Z_t-z_0))r_{p,t+1}]r_{i,t+1}K_h(Z_t-z_0)=0 \tag{31}$$

where $K_h(\cdot)=h^{-k}K(\cdot/h)$, $K(\cdot)$ is a kernel function in $\mathbb{R}^k$ and $h=h_n>0$ a bandwidth, which controls the amount of smoothing used in the estimation. Eq. (31) can be viewed as a generalization of the nonparametric estimation equations in Cai (2003) and the locally weighted version of (9.2.29) in Hamilton (1994, p. 243). Therefore, solving the above equations leads to the NPGMM estimate of $m(z_0)$, denoted by $\hat{m}(z_0)$, which is $\hat{a}$, where $(\hat{a},\hat{b})$ is the minimizer of Eq. (31). Cai et al. (2008a) discuss how to choose Q_t and derive the asymptotic properties of the proposed nonparametric estimator.

Let $\hat{e}_{i,t+1}$ be the estimated pricing error, that is, $\hat{e}_{i,t+1}=\hat{m}_{t+1}r_{i,t+1}$, where $\hat{m}_{t+1}=1-\hat{m}(Z_t)r_{p,t+1}$. To test $E(e_{i,t+1}|\Omega_t)=0$, Wang (2002, 2003) considers a simple test as follows. First, to run a multiple regression

$$\hat{e}_{i,t+1}=V_t^T\delta_i+v_{i,t+1} \tag{32}$$

where V_t is a $q\times 1$ ($q\geq k$) vector of observed variables from Ω_t,[8] and then test if all the regression coefficients are zero, that is, $\mathbb{H}_0:\delta_1=\cdots=\delta_q=0$. By assuming that the distribution of $v_{i,t+1}$ is normal, Wang (2002, 2003) uses a conventional F-test. Also, Wang (2002) discusses two alternative test procedures. Indeed, the above model can be viewed as a linear approximation of $E[e_{i,t+1}|V_t]$. To examine the magnitude of pricing errors, Ghysels (1998) considers the mean square error (MSE) as a criterion to test if the conditional CAP model or APT model is misspecified relative to the unconditional one.

To check the misspecification of the model, Cai, Kuan, and Sun (2008b) consider the testing hypothesis $\mathbb{H}_0$,

$$\mathbb{H}_0: m(\cdot)=m_0(\cdot) \quad \text{versus} \quad \mathbb{H}_a: m(\cdot)\neq m_0(\cdot) \tag{33}$$

where $m_0(\cdot)$ has a particular form. For example, if $m_0(\cdot)=b(\cdot)$, where $b(\cdot)$ is given in Wang (2003), this test is about testing the mean-covariance

efficiency. If $m(\cdot)$ is a linear function, the test reduces to testing whether the linear pricing kernel is appropriate. Then, Cai et al. (2008b) construct a consistent nonparametric test based on a U-Statistics technique, described as follows. Since I_t is a $q \times 1$ $(q \geq k)$ vector of observed variables from Ω_t, similar to Wang (2003), I_t is taken to be Z_t. It is clear that $E(e_{i,t+1}|Z_t) = 0$, where $e_{i,t+1} = [1 - m_0(Z_t)r_{p,t+1}]r_{i,t+1}$, if and only if $[E(e_{i,t+1}|Z_t)]^2 f(Z_t) = 0$, and if and only if $E(e_{i,t+1}E(e_{i,t+1}|Z_t)f(Z_t) = 0$, where $f(\cdot)$ is the density of Z_t. Interestingly, the testing problem on conditional moment becomes unconditional. Obviously, the test statistic could be postulated as

$$U_T = \frac{1}{T}\sum_{t=1}^{T} e_{i,t+1}E(e_{i,t+1}|Z_t)f(Z_t) \tag{34}$$

if $e_{i,t+1}E(e_{i,t+1}|Z_t)f(Z_t)$ would be known. Since $E(e_{i,t+1}|Z_t)f(Z_t)$ is unknown, its leave-one-out Nadaraya–Watson estimator can be formulated as

$$\hat{E}(e_{i,t+1}|Z_t)f(Z_t) = \frac{1}{T-1}\sum_{s\neq t}^{T} e_{i,s+1}K_h(Z_s - Z_t) \tag{35}$$

Plugging Eq. (35) into Eq. (34) and replacing $e_{i,t+1}$ by its estimate $\hat{e}_{i,t+1} = \hat{e}_t$, one obtain the test statistic, denoted by $\hat{U}_T$, as

$$\hat{U}_T = \frac{1}{T(T-1)}\sum_{s\neq t} K_h(Z_s - Z_t)\hat{e}_s\hat{e}_t \tag{36}$$

which is indeed a second-order U-statistics. Finally, Cai et al. (2008b) show that this nonparametric test statistic is consistent. In addition, they apply the proposed testing procedure to test if either the CAP model or the Fama and French model, in the flexible nonparametric form, can explain the momentum profit which is the value-weighted portfolio of NYSE stocks as the market portfolio, using the dividend-price ratio, the default premium, the one-month Treasury bill rate, and the excess return on the NYSE equally weighted portfolio as the conditioning variables.

5. NONPARAMETRIC PREDICTIVE MODELS FOR ASSET RETURNS

The predictability of stock returns has been studied for the last two decades as a cornerstone research topic in economics and finance,[9] and it is now routinely used in studies of many financial applications such as mutual fund

performances, tests of the conditional CAP, and optimal asset allocations.[10] Tremendous empirical studies document the predictability of stock returns using various lagged financial variables, such as the log dividend-price ratio, the log earning-price ratio, the log book-to-market ratio, the dividend yield, the term spread and default premium, and the interest rates. Important questions are often asked about whether the returns are predictable and whether the predictability is stable over time. Since many of the predictive financial variables are highly persistent and even nonstationary, it is really challenging econometrically or statistically to answer these questions.

Predictability issues are generally assessed in the context of parametric predictive regression models in which rates of returns are regressed against the lagged values of stochastic explanatory variables (or state variables). Mankiw and Shapiro (1986) and Stambaugh (1986) were first to discern the econometric and statistical difficulties inherent in the estimation of predictive regressions through the structural predictive linear model as

$$y_t = \alpha_0 + \alpha_1 x_{t-1} + \varepsilon_t, \quad x_t = \rho x_{t-1} + u_t, \quad 1 \le t \le n \tag{37}$$

where y_t is the predictable variable, say excess stock return at time t; innovations $\{(\varepsilon_t, u_t)\}$ are iid bivariate normal $N(0, \Sigma)$ with $\sum = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & \sigma_u^2 \end{pmatrix}$; and x_{t-1} is the first lag of a financial variable such as the log dividend-price ratio, which is commonly modeled by an AR(1) model as the second equation in model (37).

There are several limitations to model (37) that should be seriously considered. First, note that the correlation between two innovations ε_t and u_t in Eq. (37) is $\phi = \sigma_{\varepsilon u}/\sigma_\varepsilon \sigma_u$, which is unfortunately non-zero for many empirical applications; see, for example, Table 4 in Campbell and Yogo (2006) and Table 1 in Torous, Valkanov, and Yan (2004) for some real applications. This creates the so-called "endogeneity" (x_{t-1} and ε_t may be correlated) problem which makes modeling difficult and produces biased estimation. Another difficulty comes from the parameter ρ, which is the unknown degree of persistence of the variable x_t. That is, x_t is stationary if $|\rho| < 1$ – see Viceira (1997), Amihud and Hurvich (2004), Paye and Timmermann (2006), and Dangl and Halling (2007); or it is unit root or integrated if $\rho = 1$, denoted by $I(1)$ – see Park and Hahn (1999), Chang and Martinez-Chombo (2003), and Cai, Li, and Park (2009b); or it is local to unity or nearly integrated if $\rho = 1 + c/n$ for some $c < 0$, denoted by NI(1) – see Elliott and Stock (1994), Cavanagh, Elliott, and Stock (1995), Torous et al. (2004), Campbell and Yogo (2006), Polk, Thompson, and

Vuolteenaho (2006), and Rossi (2007), among others. This means that the predictive variable x_t is highly persistent, and even nonstationary, which may cause troubles for econometric modeling.

The third difficulty is the instability issue of the return predictive model. In fact, in return predictive models based on financial instruments such as the dividend and earnings yield, short interest rates, term spreads, and default premium, and so on, there have been many evidences on the instability of prediction model, particularly based on the dividend and earnings yield and the sample from the second half of the 1990s. This leads to the conclusion that the coefficients should change over time; see, for example, Viceira (1997), Lettau and Ludvigsson (2001), Goyal and Welch (2003), Paye and Timmermann (2006), Ang and Bekaert (2007), and Dangl and Halling (2007). While the aforementioned studies found evidences of instability in return predictive models, they did not provide any guideline on how the coefficients change over the time and where the return models may have changed. It is well known that if return predictive models are unstable, one can only assess the economic significance of return predictability provided it can be determined how widespread such instability changes over time and the extent to which it affects the predictability of stock returns. Therefore, all of the foregoing difficulties about the classical predictive regression models motivate us to propose a new varying coefficient predictive regression model. The proposed model is not only interesting in its applications to finance and economics but also important in enriching the econometric theory.

As shown in Nelson and Kim (1993), because of the endogeneity, the ordinary least squares (OLS) estimate of the slope coefficient α_1 in Eq. (37) and its standard errors are substantially biased in finite samples if x_t is highly persistent, not really exogenous, and even nonstationary. Conventional tests based on standard t-statistics from OLS estimates tend to over reject the null of non-predictability in Monte Carlo simulations. Some improvements have been developed recently to deal with the bias issue. For example, the first-order bias-correction estimator is proposed by Stambaugh (1999) based on Kendall's (1954) analytical result for the bias expression of the least squares estimate of ρ, while Amihud and Hurvich (2004) propose a two-stage least squares estimator by using a linear projection of ε_t onto u_t. Finally, the conservative bias-adjusted estimator is proposed by Lewellen (2004) if ρ is very close to one for some predicting variables. Unfortunately, all of them still have not overcome the instability difficulty mentioned above. To deal with the instability problems, Paye and Timmermann (2006) analyze the excess returns on international equity indices related to state

variables such as the lagged dividend yield, short interest rate, term spread, and default premium, to investigate how widespread the evidence of structural breaks is and to what extent breaks affect the predictability of stock returns. Finally, Dangl and Halling (2007) consider equity return prediction model with random coefficients generated from a unit root process, related to 16 state variables.

Cai and Wang (2008a) consider a time-varying coefficient predictive regression model to allow the coefficients α_0 and α_1 in Eq. (37) to change over time (to be function of time), denoted by $\alpha_0(t)$ and $\alpha_1(t)$. They use a nonlinear projection of ε_t onto u_t, that is $\varepsilon_t = \alpha_2(t)\ u_t + v_t$, and then model (37) becomes the following time-varying coefficient predictive model:

$$y_t = \alpha_0(t) + \alpha_1(t)x_{t-1} + \alpha_2(t)u_t + v_t, \quad x_t = \rho x_{t-1} + u_t, \quad 1 \le t \le n \tag{38}$$

They apply the local linear method to find the nonparametric estimates for $\alpha_j(t)$ and derive the asymptotic properties for the proposed estimator. Also, they derive the limiting distribution of the proposed nonparametric estimator, which is a mixed normal with conditional variance being a function of integrations of an Ornstein–Uhlenbeck process (mean-reverting process). They also show that the convergence rates for the intercept function (the regular rate at $(nh)^{1/2}$) and the slope function (a faster rate at $(n^2h)^{1/2}$) are totally different due to the NI(1) property of the state variable, although the asymptotic bias, coming from the local linear approximation, is the same as the stationary covariate case. Therefore, to estimate the intercept function optimally, Cai and Wang (2008a) propose a two-stage optimal estimation procedure similar to the profile likelihood method; see, for example, Speckman (1988), Cai (2002a, 2002b), and Cai et al. (2009b), and they also show that the proposed two-stage estimator reaches indeed the optimality.

Cai and Wang (2008b) consider some consistent nonparametric tests for testing the null hypothesis of whether a parametric linear regression model is suitable or if there is no relationship between the dependent variable and predictors. Therefore, these testing problems can be postulated as the following general testing hypothesis:

$$\mathbb{H}_0 : \alpha_j(t) = \alpha_j(t,\ \theta_j) \tag{39}$$

where $\alpha_j(t,\ \theta_j)$ is a known function with unknown parameter θ_j. If $\alpha_j(t,\ \theta_j)$ is constant, Eq. (39) becomes to test if model (37) is appropriate. If $\alpha_1(t,\ \theta_1) = 0$, it is to test if there exists predictability. If $\alpha_j(t,\ \theta_j)$ is a piecewise constant function, it is to test whether there exits any structural change. Cai and Wang (2008b) propose a nonparametric test which is a U-statistic

type, similar to Eq. (36), and they also show that the proposed test statistic has different asymptotic behaviors depending on the stochastic properties of x_t. Specifically, Cai and Wang (2008b) address the following two scenarios: (a) x_t is nonstationary (either I(1) or NI(1)); (*b*) x_t contains both stationary and nonstationary components. Cai and Wang (2008a, 2008b) apply the estimation and testing procedures described above to consider the instability of predictability of some financial variables. Their test finds evidence for instability of predictability for the dividend-price and earnings-price ratios. They also find evidence for instability of predictability with the short rate and the long-short yield spread, for which the conventional test leads to valid inference.

For the linear projection used by Amihud and Hurvich (2004), it is implicitly assumed that the joint distribution of two innovations ε_t and u_t in model (37) is normal and this assumption might not hold for all applications. To relax this harsh assumption, Cai (2008) considers a nonlinear projection of ε_t onto x_{t-1} instead of u_t as $\varepsilon_t = \phi(x_{t-1}) + v_t$, so that $E(v_t|x_{t-1}) = 0$. Therefore, the endogeneity is removed. Then, model (37) becomes the following classical regression model with nonstationary predictors:

$$y_t = g(x_{t-1}) + v_t, \quad x_t = \rho x_{t-1} + u_t, \quad 1 \le t \le n \tag{40}$$

where $g(x_{t-1}) = \alpha_0 + \alpha_1 x_{t-1} + \phi(x_{t-1})$ and $E(v_t|x_{t-1}) = 0$. Now, for model (40), the testing predictability $\mathbb{H}_0$: $\alpha_1 = 0$ for model (37) as in Campbell and Yogo (2006) becomes the testing hypothesis $\mathbb{H}_0$: $g(x) = c$ for model (40), which is indeed more general. To estimate $g(\cdot)$ nonparametrically, Cai (2008) uses a local linear or local constant method and derives the limiting distribution of the nonparametric estimator when x_t is an I(1) process. It is interesting to note that the limiting distribution of the proposed nonparametric estimator is a mixed normal with a conditional variance associated with a local time of a standard Brownian motion and the convergence rate is $\sqrt{n^{1/2}h}$ instead of the conventional rate $\sqrt{nh}$. Furthermore, Cai (2008) proposes two test procedures. The first one is similar to the testing approach proposed in Sun, Cai, and Li (2008) when x_t is integrated and the second one is to use the generalized likelihood ratio type testing procedure as in Cai et al. (2000) and the bootstrap. Finally, Cai (2008) applies the aforementioned estimation and testing procedures to consider the predictability of some financial instruments. The tests find some strong evidences that the predictability exists for the log dividend-price ratio, log earnings-price ratio, the short rate, and the long-short yield spread.

6. CONCLUSION

Over the last several years, nonparametric methods for both continuous and discrete time have become an integral part of research in financial economics. The literature is already vast and continues to grow swiftly, involving a full spread of participants for both financial economists and statisticians and engaging a wide sweep of academic journals. The field has left indelible mark on almost all core areas in finance such as APT, consumption portfolio selection, derivatives, and risk analysis. The popularity of this field is also witnessed by the fact that the graduate students at both master and doctoral levels in economics, finance, mathematics, and statistics are expected to take courses in this discipline or alike and review the important research papers in this area to search for their own research interests, particularly dissertation topics for doctoral students. On the other hand, this area also has made an impact in the financial industry, as the sophisticated nonparametric techniques can be of practical assistance in the industry. We hope that this selective review has provided the reader a perspective on this important field in finance and statistics and some open research problems.

Finally, we would like to point out that the paper by Cai, Gu, and Li (2009a) gives a comprehensive survey on some recent developments in nonparametric econometrics, including nonparametric estimation and testing of regression functions with mixed discrete and continuous covariates, nonparametric estimation/testing with nonstationary data, nonparametric instrumental variable estimations, and nonparametric estimation of quantile regression models, which can be applied to financial studies. Other two promising lines of nonparametric finance are nonparametric volatility (conditional variance) and ARCH- or GARCH-type models and nonparametric methods in volatility for high-frequency data with/without microstructure noise. The reader interested in these areas of research should consult with the recent works, to name just a few, including Fan and Wang (2007), Long, Su, and Ullah (2009), and Mishra, Su, and Ullah (2009), and the references therein. Unfortunately, these topics are omitted in this paper due to too vast literature. However, we will write a separate survey paper on this important financial area, which is volatility models for both low-frequency and high-frequency data.

NOTES

1. Other theoretical models are studied by Brennan and Schwartz (1979), Constantinides (1992), Courtadon (1982), Cox, Ingersoll, and Ross (1980),

Dothan (1978), Duffie and Kan (1996), Longstaff and Schwartz (1992), Marsh and Rosenfeld (1983), and Merton (1973). Heath, Jarrow, and Morton (1992) consider another important class of term structure models which use the forward rate as the underlying state variable.

2. Empirical studies on the short rate include Ait-Sahalia (1996a, 1996b), Andersen and Lund (1997), Ang and Bekaert (2002a, 2002b), Brenner, Harjes, and Kroner (1996), Brown and Dybvig (1986), Chan et al. (1992), Chapman and Pearson (2000), Chapman, Long, and Pearson (1999), Conley et al. (1997), Gray (1996), and Stanton (1997).

3. See, to name just a few, Pan (1997), Duffie and Pan (2001), Bollerslev and Zhou (2002), Eraker, Johannes, and Polson (2003), Bates (2000), Duffie et al. (2000), Johannes (2004), Liu et al. (2002), Zhou (2001), Singleton (2001), Perron (2001), Chernov et al. (2003).

4. Sundaresan (2001) states that "perhaps the most significant development in the continuous-time field during the last decade has been the innovations in econometric theory and in the estimation techniques for models in continuous time." For other reviews of the recent literature, see Melino (1994), Tauchen (1997, 2001), and Campbell et al. (1997).

5. A simple example is the Vasicek model, where if we vary the speed of mean reversion and the scale of diffusion in the same proportion, the marginal density will remain unchanged, but the transition density will be different.

6. One could simply ignore the data in the boundary regions and only use the data in the interior region. Such a trimming procedure is simple, but in the present context, it would lead to the loss of significant amount of information. If $h = sn^{-\frac{1}{5}}$ where $s^2 = \text{Var}(X_t)$, for example, then about 23, 20, and 10 of a uniformly distributed sample will fall into the boundary regions when $n = 100$, 500, and 5,000, respectively. For financial time series, one may be particularly interested in the tail distribution of the underlying process, which is exactly contained in (and only in) the boundary regions.

Another solution is to use a kernel that adapts to the boundary regions and can effectively eliminate the boundary bias. One example is the so-called jackknife kernel, as used in Chapman and Pearson (2000). In the present context, the jackknife kernel, however, has some undesired features in finite samples. For example, it may generate negative density estimates in the boundary regions because the jackknife kernel can be negative in these regions. It also induces a relatively large variance for the kernel estimates in the boundary regions, adversely affecting the power of the test in finite samples.

7. Chen, Gao, and Tang (2008) consider kernel-based simultaneous specification testing for both mean and variance models in a discrete-time setup with dependent observations. The empirical likelihood principle is used to construct the test statistic. They apply the test to check adequacy of a discrete version of a continuous-time diffusion model.

8. Wang (2003) takes V_t to be Z_t in his empirical analysis.

9. See, for example, Fama and French (1988), Keim and Stambaugh (1986), Campbell and Shiller (1988), Cutler, Poterba, and Summers (1991), Balvers, Cosimano, and McDonald (1990), Schwert (1990), Fama (1990), and Kothari and Shanken (1997).

10. See, Christopherson, Ferson, and Glassman (1998), Ferson and Schadt (1996), Ferson and Harvey (1991), Ghysels (1998), Ait-Sahalia and Brandt (2001), Barberis (2000), Brandt (1999), Campbell and Viceira (1998), and Kandel and Stambaugh (1996).

ACKNOWLEDGMENTS

The authors thank two referees, Federico M. Bandi, Haitao Li, and Aman Ullah for their valuable and helpful comments, suggestions, and discussions. Also, the authors thank the participants at the seminars at University of Chicago, Columbia University, Academica Sinica and NYU, and the audiences at the 7th Annual Advances in Econometrics Conference (November 2008 at Louisiana State University) for their helpful comments. Cai's research was supported, in part, by the National Science Foundation grant DMS-0404954 and the National Science Foundation of China grant no. 70871003, and funds provided by the University of North Carolina at Charlotte, the Cheung Kong Scholarship from Chinese Ministry of Education, the Minjiang Scholarship from Fujian Province, China, and Xiamen University. Hong thanks financial support from the Overseas Outstanding Youth Grant from the National Science Foundation of China and the Cheung Kong Scholarship from Chinese Ministry of Education and Xiamen University.

REFERENCES

Ahn, D. H., Dittmar, R. F., & Gallant, A. R. (2002). Quadratic term structure models: Theory and evidence. *Review of Financial Studies*, *15*, 243–288.

Ahn, D. H., & Gao, B. (1999). A parametric nonlinear model of term structure dynamics. *Review of Financial Studies*, *12*, 721–762.

Ait-Sahalia, Y. (1996a). Nonparametric pricing of interest rate derivative securities. *Econometrica*, *64*, 527–560.

Ait-Sahalia, Y. (1996b). Testing continuous-time models of the spot interest rate. *Review of Financial Studies*, *9*, 385–426.

Ait-Sahalia, Y. (1999). Transition densities for interest rate and other nonlinear diffusions. *Journal of Finance*, *54*, 1361–1395.

Ait-Sahalia, Y. (2002a). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approach. *Econometrica*, *70*, 223–262.

Ait-Sahalia, Y. (2002b). Telling from discrete data whether the underlying continuous-time model is a diffusion. *Journal of Finance*, *57*, 2075–2112.

Ait-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusion. *Annals of Statistics*, *36*, 906–937.

Ait-Sahalia, Y., & Brandt, M. (2001). Variable selection for portfolio choice. *Journal of Finance*, *56*, 1297–1350.

Ait-Sahalia, Y., & Kimmel, R. (2007). Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics*, *83*, 413–452.

Ait-Sahalia, Y., & Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance*, *53*, 499–547.

Ait-Sahalia, Y., & Lo, A. W. (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, *94*, 9–51.

Akdeniz, L., Altay-Salih, A., & Caner, M. (2003). Time-varying betas help in asset pricing: The threshold CAPM. *Studies in Nonlinear Dynamics & Econometrics*, *6*(4), 1–16.

Amihud, Y., & Hurvich, C. (2004). Predictive regression: A reduced-bias estimation method. *Journal of Financial and Quantitative Analysis*, *39*, 813–841.

Anagnou, I., Bedendo, M., Hodges, S., & Tompkins, R. (2005). The relation between implied and realized probability density functions. *Review of Futures Markets*, *11*, 41–66.

Anders, U., Korn, O., & Schmitt, C. (1998). Improving the pricing of options: A neural network approach. *Journal of Forecasting*, *17*, 369–388.

Andersen, T. G., Benzoni, L., & Lund, J. (2002). Towards an empirical foundation for continuous-time equity return models. *Journal of Finance*, *57*, 1239–1284.

Andersen, T. G., Chung, H.-J., & Sorensen, B. E. (1999). Efficient method of moments estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Econometrics*, *91*, 61–87.

Andersen, T. G., & Lund, J. (1997). Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics*, *77*, 343–377.

Ang, A., & Bekaert, G. (2002a). Short rate nonlinearities and regime switches. *Journal of Economic Dynamics and Control*, *26*, 1243–1274.

Ang, A., & Bekaert, G. (2002b). Regime switches in interest rates. *Journal of Business and Economic Statistics*, *20*, 163–182.

Ang, A., & Bekaert, G. (2007). Stock return predictability: Is it there? *Review of Financial Studies*, *20*, 651–707.

Ang, A., & Liu, J. (2004). How to discount cashflows with time-varying expected return. *Journal of Finance*, *59*, 2745–2783.

Bahra, B. (1997). *Implied risk-neutral probability density functions from option prices: Theory and application*. Working Paper. Bank of England.

Balvers, R. J., Cosimano, T. F., & McDonald, B. (1990). Predicting stock returns in an efficient market. *Journal of Finance*, *45*, 1109–1128.

Bandi, F. (2000). *Nonparametric fixed income pricing: Theoretical issues*. Working Paper. Graduate School of Business, The University of Chicago, Chicago, IL.

Bandi, F., & Nguyen, T. H. (2000). *Fully nonparametric estimators for diffusions: A small sample analysis*. Working Paper. Graduate School of Business, The University of Chicago, Chicago, IL.

Bandi, F., & Nguyen, T. H. (2003). On the functional estimation of jump-diffusion models. *Journal of Econometrics*, *116*, 293–328.

Bandi, F., & Phillips, P. C. B. (2003). Fully nonparametric estimation of scalar diffusion models. *Econometrica*, *71*, 241–283.

Bansal, R., Hsieh, D. A., & Viswanathan, S. (1993). A new approach to international arbitrage pricing. *Journal of Finance*, *48*, 1719–1747.

Bansal, R., & Viswanathan, S. (1993). No arbitrage and arbitrage pricing: A new approach. *Journal of Finance*, *47*, 1231–1262.

Barberis, N. (2000). Investing for the long run when returns are predictable. *Journal of Finance*, *55*, 225–264.

Bates, D. S. (1991). The crash of '87: Was it expected? The evidence from options markets. *Journal of Finance*, *46*, 1009–1044.

Bates, D. S. (2000). Post-'87 crash fears in the S&P 500 futures option market. *Journal of Econometrics*, *94*, 181–238.

Black, F., Derman, E., & Toy, W. (1990). "A one-factor model of interest rates and its application to treasury bond options. *Financial Analysts Journal*, *46*, 33–39.

Black, F., & Karasinski, P. (1991). Bond and option pricing when short rates are log-normal. *Financial Analysts Journal*, *47*, 52–59.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *71*, 637–654.

Bliss, R. R., & Smith, D. (1998). The elasticity of interest rate volatility: Chan, Karolyi, Longstaff, and Sanders revisited. *Journal of Risk*, *1*, 21–46.

Bollerslev, T., & Zhou, H. (2002). Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *Journal of Econometrics*, *109*, 33–65.

Brandt, M. W. (1999). Estimating portfolio and consumption choice: A conditional Euler equations approach. *Journal of Finance*, *54*, 1609–1646.

Brandt, M. W., & Santa-Clara, P. (2002). Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. *Journal of Financial Economics*, *63*, 161–210.

Breeden, D. T., & Litzenberger, R. H. (1978). Prices of state contingent claims implicit in option prices. *Journal of Business*, *51*, 621–651.

Brennan, M. J., & Schwartz, E. (1979). A continuous time approach to the pricing of bonds. *Journal of Banking and Finance*, *3*, 133–155.

Brenner, R., Harjes, R., & Kroner, K. (1996). Another look at alternative models of the short-term interest rate. *Journal of Financial and Quantitative Analysis*, *31*, 85–107.

Brown, S. J., & Dybvig, P. H. (1986). The empirical implications of the Cox, Ingersoll, Ross theory of the term structure of interest rates. *Journal of Finance*, *41*, 617–630.

Cai, Z. (2001). Weighted Nadaraya–Watson regression estimation. *Statistics and Probability Letters*, *51*, 307–318.

Cai, Z. (2002a). Two-step likelihood estimation procedure for varying-coefficient models. *Journal of Multivariate Analysis*, *82*, 189–209.

Cai, Z. (2002b). A two-stage approach to additive time series models. *Statistica Neerlandica*, *56*, 415–433.

Cai, Z. (2003). Nonparametric estimation equations for time series data. *Statistics and Probability Letters*, *62*, 379–390.

Cai, Z. (2008). *Nonparametric predictive regression models for asset returns*. Working Paper. Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC.

Cai, Z., Fan, J., & Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, *95*, 941–956.

Cai, Z., Kuan, C. M., & Sun, L. (2008a). *Nonparametric pricing kernel models*. Working Paper. Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC.

Cai, Z., Kuan, C. M., & Sun, L. (2008b). *Nonparametric test for pricing kernel models*. Working Paper. Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC.

Cai, Z., Gu, J., & Li, Q. (2009a). Some recent developments on nonparametric econometrics. *Advances in Econometrics*, *25*, 495–549.

Cai, Z., Li, Q., & Park, J. Y. (2009b). Functional-coefficient models for nonstationary time series data. *Journal of Econometrics*, *148*, 101–113.

Cai, Z., & Tiwari, R. C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, *11*, 341–350.

Cai, Z., & Wang, Y. (2008a). *Instability of predictability of asset returns*. Working Paper. Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC.

Cai, Z., & Wang, Y. (2008b). *Testing stability of predictability of asset returns*. Working Paper. Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC.

Cai, Z., & Zhang, L. (2008a). *Testing for discontinuous diffusion models versus jump diffusion models*. Working Paper. The Wang Yanan Institute for Studies in Economics, Xiamen University, China.

Cai, Z., & Zhang, L. (2008b). *Information effect for different firm-sizes via the nonparametric jump-diffusion model*. Working Paper. The Wang Yanan Institute for Studies in Economics, Xiamen University, China.

Campbell, J., & Shiller, R. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, *1*, 195–227.

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.

Campbell, J., & Yogo, M. (2006). Efficient tests of stock return predictability. *Journal of Financial Economics*, *81*, 27–60.

Campbell, J. Y., & Viceira, L. (1998). Consumption and portfolio decisions when expected returns are time varying. *Quarterly Journal of Economics*, *114*, 433–495.

Cavanagh, C. L., Elliott, G., & Stock, J. H. (1995). Inference in models with nearly integrated regressors. *Econometric Theory*, *11*, 1131–1147.

Chan, K. C., Karolyi, G. A., Longstaff, F. A., & Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance*, *47*, 1209–1227.

Chang, Y., & Martinez-Chombo, E. (2003). *Electricity demand analysis using cointegration and error-correction models with time varying parameters: The Mexican case*. Working Paper. Department of Economics, Texas A&M University, Texas.

Chapman, D., Long, J., & Pearson, N. (1999). Using proxies for the short rate: When are three months like an instant. *Review of Financial Studies*, *12*, 763–807.

Chapman, D., & Pearson, N. (2000). Is the short rate drift actually nonlinear? *Journal of Finance*, *55*, 355–388.

Chen, S. X., Gao, J., & Tang, C. (2008). A test for model specification of diffusion processes. *Annals of Statistics*, *36*, 167–198.

Chen, S. X., Härdle, W., & Kleinow, T. (2002). An empirical likelihood goodness-of-fit test for time series. In: W. Härdle, T. Kleinow & G. Stahl (Eds), *Applied quantitative finance* (pp. 259–281). Berlin, Germany: Spring-Verlag.

Chernov, M., Gallant, A. R., Ghysels, E., & Tauchen, G. (2003). Alternative models of stock price dynamics. *Journal of Econometrics*, *116*, 225–257.

Christopherson, J. A., Ferson, W., & Glassman, D. A. (1998). Conditioning manager alphas on economic information: Another look at the persistence of performance. *Review of Financial Studies*, *11*, 111–142.

Chung, C. C., & Tauchen, G. (2001). Testing target zone models using efficient method of moments. *Journal of Business and Economic Statistics*, *19*, 255–277.

Cochrane, J. H. (1996). A cross-sectional test of an investment based asset pricing model. *Journal of Political Economy*, *104*, 572–621.

Cochrane, J. H. (2001). *Asset pricing*. New Jersey: Princeton University Press.

Conley, T. G., Hansen, L. P., Luttmer, E. G. J., & Scheinkman, J. A. (1997). Short-term interest rates as subordinated diffusions. *Review of Financial Studies*, *10*, 525–577.

Constantinides, G. M. (1992). A theory of the nominal term structure of interest rates. *Review of Financial Studies*, *5*, 531–552.

Courtadon, G. (1982). A more accurate finite difference approximation for the valuation of options. *Journal of Financial and Quantitative Analysis*, *17*, 697–703.

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1980). An analysis of variable rate loan contracts. *Journal of Finance*, *35*, 389–403.

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, *53*, 385–407.

Cox, J. C., & Ross, S. A. (1976). The volatility of option for alternative stochastic processes. *Journal of Financial Economics*, *3*, 145–166.

Cutler, D. M., Poterba, J. M., & Summers, L. H. (1991). Speculative dynamics. *Review of Economic Studies*, *58*, 529–546.

Dai, Q., & Singleton, K. J. (2000). Specification analysis of affine term structure models. *Journal of Finance*, *55*, 1943–1978.

Dangl, T., & Halling, M. (2007). *Predictive regressions with time-varying coefficients*. Working Paper. School of Business, University of Utah, Utah.

Diebold, F. X., Gunther, T., & Tay, A. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*, 863–883.

Dothan, M. U. (1978). On the term structure of interest rates. *Journal of Financial Economics*, *6*, 59–69.

Duffie, D. (2001). *Dynamic asset pricing theory* (3rd ed.). Princeton, NJ: Princeton University Press.

Duffie, D., & Kan, R. (1996). A yield factor model of interest rate. *Mathematical Finance*, *6*, 379–406.

Duffie, D., & Pan, J. (2001). Analytical value-at-risk with jumps and credit risk. *Finance and Stochastics*, *5*, 155–180.

Duffie, D., Pan, J., & Singleton, K. J. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, *68*, 1343–1376.

Duffie, D., & Singleton, K. J. (1993). Simulated moments estimation of Markov models of asset prices. *Econometrica*, *61*, 929–952.

Egorov, A., Hong, Y., & Li, H. (2006). Validating forecasts of the joint probability density of bond yields: Can affine models beat random walk? *Journal of Econometrics*, *135*, 255–284.

Egorov, A., Li, H., & Xu, Y. (2003). Maximum likelihood estimation of time-inhomogeneous diffusions. *Journal of Econometrics*, *114*, 107–139.

Elliott, G., & Stock, J. H. (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric Theory*, *10*, 672–700.

Elerian, O., Chib, S., & Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, *69*, 959–993.

Eraker, B. (1998). *Markov chain Monte Carlo analysis of diffusion models with application to finance*. HAE thesis, Norwegian School of Economics and Business Administration.

Eraker, B., Johannes, M. S., & Polson, N. G. (2003). The impact of jumps in volatility and returns. *Journal of Finance*, *58*, 1269–1300.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*, 383–417.

Fama, E. F. (1990). Stock returns, real returns, and economic activity. *Journal of Finance*, *45*, 1089–1108.

Fama, E. F., & French, K. R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics*, *22*, 3–26.

Fama, E., & French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, *47*, 427–466.

Fama, E., & French, K. R. (1993). Common risk factors in the returns on bonds and stocks. *Journal of Financial Economics*, *33*, 3–56.

Fama, E., & French, K. R. (1995). Size and book-to-market factors in earning and returns. *Journal of Finance*, *50*, 131–155.

Fan, J., & Gijbels, I. (1996). *Local polynomial modeling and its applications*. London: Chapman and Hall.

Fan, J., Jiang, J., Zhang, C., & Zhou, Z. (2003). Time-dependent diffusion models for term structure dynamics and the stock price volatility. *Statistica Sinica*, *13*, 965–992.

Fan, J., & Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, *102*, 1349–1362.

Fan, J., & Zhang, C. (2003). A re-examination of diffusion estimators with applications to financial model validation. *Journal of the American Statistical Association*, *98*, 118–134.

Fan, J., Zhang, C., & Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, *29*, 153–193.

Ferson, W. E. (1989). Changes in expected security returns, risk and the level of interest rates. *Journal of Finance*, *44*, 1191–1214.

Ferson, W. E., & Harvey, C. R. (1991). The variation of economic risk premiums. *Journal of Political Economy*, *99*, 385–415.

Ferson, W. E., & Harvey, C. R. (1993). The risk and predictability of international equity returns. *Journal of Financial Studies*, *6*, 527–566.

Ferson, W. E., & Harvey, C. R. (1998). Fundamental determinants of national equity market returns: A perspective on conditional asset pricing. *Journal of Banking and Finance*, *21*, 1625–1665.

Ferson, W. E., & Harvey, C. R. (1999). Conditional variables and the cross section of stock return. *Journal of Finance*, *54*, 1325–1360.

Ferson, W. E., & Korajczyk, R. A. (1995). Do arbitrage pricing models explain the predictability of stock returns? *Journal of Business*, *68*, 309–349.

Ferson, W. E., & Schadt, R. W. (1996). Measuring fund strategy and performance in changing economic conditions. *Journal of Finance*, *51*, 425–461.

Gallant, A. R., & Tauchen, G. (1996). Which moments to match? *Econometric Theory*, *12*, 657–681.

Gallant, A. R., & Tauchen, G. (2001). *Efficient method of moments*. Working Paper. Department of Economics, Duke University, Durham, NC.

Ghysels, E. (1998). On stable factor structures in the pricing of risk: Do time varying betas help or hurt?. *Journal of Finance*, *53*, 549–573.

Gourieroux, C., Monfort, A., & Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, *8*, 85–118.

Goyal, A., & Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, *49*, 639–654.

Gray, S. (1996). Modeling the conditional distribution of interest rates as a regime switching process. *Journal of Financial Economics*, *42*, 27–62.

Gourieroux, C., & Jasiak, J. (2001). *Financial econometrics: Problems, models, and methods*. Princeton, NJ: Princeton University Press.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Hanke, M. (1999). Neural networks versus Black–Scholes: An empirical comparison of the pricing accuracy of two fundamentally different option pricing methods. *Journal of Computational Finance*, *5*, 26–34.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, *64*, 413–430.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.

Hansen, L. P., & Janaganan, R. (1997). Assessing specification errors in stochastic discount factor models. *Journal of Finance*, *52*, 557–590.

Hansen, L. P., & Scheinkman, J. A. (1995). Back to the future: Generating moment implications for continuous time Markov processes. *Econometrica*, *63*, 767–804.

Härdle, W. (1990). *Applied nonparametric regression*. New York: Cambridge University Press.

Härdle, W., & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, *21*, 1926–1947.

Harrison, J. M., & Kreps, D. M. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, *20*, 381–408.

Harvey, C. R. (1989). Time-varying conditional covariances in tests of asset pricing models. *Journal of Financial Economics*, *24*, 289–317.

Heath, D. C., Jarrow, R. A., & Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claim valuation. *Econometrica*, *60*, 77–105.

Ho, T. S. Y., & Lee, S. B. (1986). Term structure movements and pricing interest rate contingent claims. *Journal of Finance*, *41*, 1011–1029.

Hong, Y., & Lee, T. H. (2003a). Inference and forecast of exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics*, *85*, 1048–1062.

Hong, Y., & Lee, T. H. (2003b). Diagnostic checking for nonlinear time series models. *Econometric Theory*, *19*, 1065–1121.

Hong, Y., & Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to interest rate term structures. *Review of Financial Studies*, *18*, 37–84.

Hull, J., & White, H. (1990). Pricing interest-rate derivative securities. *Review of Financial Studies*, *3*, 573–592.

Hutchinson, J., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *Journal of Finance*, *49*, 851–889.

Jackwerth, J. C. (1999). Option-implied risk-neutral distributions and implied binomial trees: A literature review. *Journal of Derivative*, *7*, 66–82.

Jackwerth, J. C., & Rubinstein, M. (1996). Recovering probability distributions from contemporary security prices. *Journal of Finance*, *51*, 1611–1631.

Jacquier, E., Polson, N. G., & Rossi, P. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, *12*, 371–389.

Jagannathan, R., & Wang, Z. (1996). The conditional CAPM and the cross-section of expected returns. *Journal of Finance*, *51*, 3–53.

Jagannathan, R., & Wang, Z. (2002). Empirical evaluation of asset pricing models: A comparison of the SDF and beta methods. *Journal of Finance*, *57*, 2337–2367.

Jarrow, R., & Tudd, A. (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of Financial Economics*, *10*, 347–369.

Jiang, G. J., & Knight, J. L. (1997). A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model. *Econometric Theory*, *13*, 615–645.

Jiang, G. J., & Knight, J. L. (2002). Estimation of continuous time processes via the empirical characteristic function. *Journal of Business and Economic Statistics*, *20*, 198–212.

Jiang, G. J., & Knight, J. L. (2006). *ECF estimation of Markov models where the transition density is unknown*. Working Paper. Department of Economics, University of Western Ontario, London, Ontario, Canada.

Jiang, G. J., & van der Sluis, P. J. (2000). Option pricing with the efficient method of moments. In: Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo & A. S. Weigend (Eds), *Computational finance*. Cambridge, MA: MIT Press.

Johannes, M. S. (2004). The economic and statistical role of jumps to interest rates. *Journal of Finance*, *59*, 227–260.

Johannes, M. S., Kumar, R., & Polson, N. G. (1999). *State dependent jump models: How do US equity indices jump?* Working Paper. Graduate School of Business, University of Chicago, Chicago, IL.

Jones, C. S. (1998). *Bayesian estimation of continuous-time finance models*. Working Paper. Simon School of Business, University of Rochester, Rochester, NY.

Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika*, *41*, 403–404.

Kandel, S., & Stambaugh, R. (1996). On the predictability of stock returns: An asset allocation perspective. *Journal of Finance*, *51*, 385–424.

Karatzas, I., & Shreve, S. E. (1988). *Brownian motion and stochastic calculus* (2nd ed.). New York: Spring-Verlag.

Keim, D. B., & Stambaugh, R. F. (1986). Predicting returns in the stock and bond markets. *Journal of Financial Economics*, *17*, 357–390.

Kleinow, T. (2002). *Testing the diffusion coefficients*. Working Paper. Institute of Statistics and Economics, Humboldt University of Berlin, Germany.

Kothari, S. P., & Shanken, J. (1997). Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics*, *44*, 169–203.

Kou, S. (2002). A jump diffusion model for option pricing. *Management Science*, *48*, 1086–1101.

Kristensen, D. (2007). *Nonparametric estimation and misspecification testing of diffusion models*. Working Paper. Department of Economics, Columbia University, New York, NY.

Kristensen, D. (2008). *Pseudo-maximum likelihood estimation in two classes of semiparametric diffusion models*. Working Paper. Department of Economics, Columbia University, New York, NY.

Lettau, M., & Ludvigsson, S. (2001). Consumption, aggregate wealth, and expected stock returns. *Journal of Finance*, *56*, 815–849.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics*, *74*, 209–235.

Li, Q., & Racine, J. (2007). *Nonparametric econometrics: Theory and applications*. New York: Princeton University Press.

Liu, M. (2000). Modeling long memory in stock market volatility. *Journal of Econometrics*, *99*, 139–171.

Liu, J., Longstaff, F. A., & Pan, J. (2002). Dynamic asset allocation with event risk. *Journal of Finance*, *58*, 231–259.

Lo, A. W. (1988). Maximum likelihood estimation of generalized Ito processes with discretely sampled data. *Econometric Theory*, *4*, 231–247.

Lobo, B. J. (1999). Jump risk in the U.S. stock market: Evidence using political information. *Review of Financial Economics*, *8*, 149–163.

Long, X., Su, L., & Ullah, A. (2009). *Estimation and forecasting of dynamic conditional covariance: A semiparametric multivariate model*. Working Paper. Department of Economics, Singapore Management University, Singapore.

Longstaff, F. A. (1992). Multiple equilibria and tern structure models. *Journal of Financial Economics*, *32*, 333–344.

Longstaff, F. A. (1995). Option pricing and the martingale restriction. *Review of Financial Studies*, *8*, 1091–1124.

Longstaff, F. A., & Schwartz, E. (1992). Interest rate volatility and the term structure: A two-factor general equilibrium model. *Journal of Finance*, *47*, 1259–1282.

Mankiw, N. G., & Shapiro, M. (1986). Do we reject too often? Small sample properties of tests of rational expectation models. *Economics Letters*, *20*, 139–145.

Marsh, T., & Rosenfeld, E. (1983). *Stochastic processes for interest rates and equilibrium bond prices, Journal of Finance*, *38*, 635–646.

Melick, W. R., & Thomas, C. P. (1997). Recovering an asset's implied PDF from option prices: An application to crude oil during the Gulf crisis. *Journal of Financial and Quantitative Analysis*, *32*, 91–115.

Melino, A. (1994). Estimation of continuous-time models in finance. In: C. Sims (Ed.), *Advances in econometrics: Sixth world congress* (Vol. 2). Cambridge: Cambridge University Press.

Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science*, *4*, 141–183.

Mishra, S., Su, L., & Ullah, A. (2009). *Semiparametric estimator of time series conditional variance*. Working Paper. Department of Economics, Singapore Management University, Singapore.

Mittelhammer, R. C., Judge, G. G., & Miller, D. J. (2000). *Econometrics foundation*. New York: Cambridge University Press.

Nelson, C. R., & Kim, M. J. (1993). Predictable stock returns: The role of small sample bias. *Journal of Finance*, *48*, 641–661.

Øksendal, B. (1985). *Stochastic differential equations: An introduction with applications* (3rd ed.). New York: Springer-Verlag.

Pagan, A., & Ullah, A. (1999). *Nonparametric econometrics*. New York: Cambridge University Press.

Pan, J. (1997). *Stochastic volatility with reset at jumps*. Working Paper. School of Management, MIT.

Park, J. Y., & Hahn, S. B. (1999). Cointegrating regressions with time varying coefficients. *Econometric Theory*, *15*, 664–703.

Paye, B. S., & Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance*, *13*, 274–315.

Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, *22*, 55–71.

Perron, B. (2001). *Jumps in the volatility of financial markets'*. Working Paper. Department of Economics, University of Montreal, Quebec, Canada.

Polk, C., Thompson, S., & Vuolteenaho, T. (2006). Cross-sectional forecasts of the equity premium. *Journal of Financial Economics*, *81*, 101–141.

Pritsker, M. (1998). Nonparametric density estimation and tests of continuous time interest rate models. *Review of Financial Studies*, *11*, 449–487.

Rice, J. (1986). Boundary modification for kernel regression. *Communications in Statistics*, *12*, 1215–1230.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, *23*, 470–472.

Rossi, B. (2007). Expectation hypothesis tests and predictive regressions at long horizons. *Econometrics Journal*, *10*, 1–26.

Rubinstein, M. (1994). Implied binomial trees. *Journal of Finance*, *49*, 771–818.

Schwert, G. W. (1990). Stock returns and real activity: A century of evidence. *Journal of Finance*, *45*, 1237–1257.

Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, *19*, 425–442.

Shimko, D. (1993). Bounds of probability. *Risk*, *6*, 33–37.

Singleton, K. J. (2001). Estimation of affine asset pricing models using the empirical characteristic function. *Journal of Econometrics*, *102*, 111–141.

Speckman, P. (1988). Kernel smoothing in partially linear models. *Journal of the Royal Statistical Society, Series B*, *50*, 413–436.

Sun, Y., Cai, Z., & Li, Q. (2008). *Consistent nonparametric test on parametric smooth coefficient model with nonstationary data*. Working Paper. Department of Economics, Texas A&M University, College Station, TX.

Sundaresan, S. (2001). Continuous-time methods in finance: A review and an assessment. *Journal of Finance*, *55*, 1569–1622.

Stambaugh, R. (1986). *Bias in regressions with lagged stochastic regressors*. Working Paper. University of Chicago, Chicago, IL.

Stambaugh, R. (1999). Predictive regressions. *Journal of Financial Economics*, *54*, 375–421.

Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance*, *52*, 1973–2002.

Torous, W., Valkanov, R., & Yan, S. (2004). On predicting stock returns with nearly integrated explanatory variables. *Journal of Business*, *77*, 937–966.

Tauchen, G. (1997). New minimum chi-square methods in empirical finance. In: D. M. Kreps & K. Wallis (Eds), *Advances in econometrics: Seventh world congress*. Cambridge, UK: Cambridge University Press.

Tauchen, G. (2001). Notes on financial econometrics. *Journal of Econometrics*, *100*, 57–64.
Taylor, S. (2005). *Asset price dynamics, volatility, and prediction*. Princeton, NJ: Princeton University Press.
Tsay, R. S. (2005). *Analysis of financial time series* (2nd ed.). New York: Wiley.
Valderrama, D. (2001). *Can a standard real business cycle model explain the nonlinearities in U.S. national accounts data?* Ph.D. thesis, Department of Economics, Duke University, Durham, NC.
Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, *5*, 177–188.
Viceira, L. M. (1997). Testing for structural change in the predictability of asset returns. Manuscript, Harvard University.
Wang, K. (2002). Nonparametric tests of conditional mean-variance efficiency of a benchmark portfolio. *Journal of Empirical Finance*, *9*, 133–169.
Wang, K. Q. (2003). Asset pricing with conditioning information: A new test. *Journal of Finance*, *58*, 161–196.
Yatchew, A., & Härdle, W. (2006). Nonparametric state price density estimation using constrained least squares and the bootstrap. *Journal of Econometrics*, *133*, 579–599.
Xu, K., & Phillips, P. B. C. (2007). *Tilted nonparametric estimation of volatility functions*. Cowles Foundation Discussion Paper no. 1612R. Department of Economics, Yale University, New Haven, CT.
Zhou, H. (2001). *Jump-diffusion term structure and Ito conditional moment generator*. Working Paper. Federal Reserve Board.

IMPOSING ECONOMIC CONSTRAINTS IN NONPARAMETRIC REGRESSION: SURVEY, IMPLEMENTATION, AND EXTENSION

Daniel J. Henderson and Christopher F. Parmeter

ABSTRACT

Economic conditions such as convexity, homogeneity, homotheticity, and monotonicity are all important assumptions or consequences of assumptions of economic functionals to be estimated. Recent research has seen a renewed interest in imposing constraints in nonparametric regression. We survey the available methods in the literature, discuss the challenges that present themselves when empirically implementing these methods, and extend an existing method to handle general nonlinear constraints. A heuristic discussion on the empirical implementation for methods that use sequential quadratic programming is provided for the reader, and simulated and empirical evidence on the distinction between constrained and unconstrained nonparametric regression surfaces is covered.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 433–469

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025016

1. INTRODUCTION

Nonparametric estimation methods are a desirable tool for applied researchers since economic theory rarely yields insights into a model's appropriate functional form. However, when paired with the specific smoothness constraints imposed by an economic theory, such as monotonicity of a cost function in all input prices, this often increases the complexity of the estimator in practice. Access to a constrained nonparametric estimator that can handle general, multiple smoothness conditions is desirable.[1] Fortunately, a rich literature on constrained estimation has taken shape, and a multitude of potential suitors have been proposed for various constrained problems. Given the potential need for constrained nonparametric estimators in applied economic research and the availability of a wide range of potential estimators, coupled with the dearth of detailed, simultaneous descriptions of these methods, a survey on the current state of the art is warranted.

Smoothness constraints present themselves in a variety of economic milieus. In empirical studies on games, such as auctions, monotonicity of player strategies is a key assumption used to derive the equilibrium solution. This monotonicity assumption thus carries over to the estimated equilibrium strategy. And while parametric models of auctions have monotonicity "built-in," their nonparametric counterparts impose no such condition. Thus, using a nonparametric estimator of auctions that allows monotonicity to be imposed is expected to be more competitive against parametric alternatives than an unconstrained estimator. Recently, Henderson, List, Millimet, Parmeter, and Price (2009) have shown that random samples from equilibrium bid distributions can produce nonmonotonic nonparametric estimates for small samples. This suggests that being able to construct an estimator that is monotonic from the onset is important for analyzing auction data.

Analogously, convexity is theoretically required for either a production or a cost function, and the ability to impose this constraint in a nonparametric setting is thus desirable given that very few models of production yield reduced form parametric solutions. Cost functions are concave in input prices and outputs, nondecreasing and homogeneous of degree 1 in input prices. Thus, estimating a cost function requires the imposition of three distinct economic conditions. To our knowledge, applied studies that nonparametrically estimate cost functions (Wheelock & Wilson, 2001) do not impose these conditions directly. Thus, at the very least there is a loss of efficiency since these constraints are not directly imposed on the estimator.

Moreover, since the constraints are not imposed, it is difficult to test whether these smoothness conditions are valid.

Before highlighting the potential methods available we aim to gauge the necessity of imposing smoothness constraints via a primitive example. Consider the univariate data generating process:

$$y_i = \ln(x_i) + \varepsilon_i, i = 1, 2, \ldots, n$$

This data generating process is monotonic and concave. If we generate random samples under a variety of sample sizes and distributional assumptions for the pair (x_i, ε_i), we can gain insight into the *need* for a constrained estimator. Tables 1 and 2 provide the proportion of times, out of 9,999 simulations, a local constant kernel estimator, in unconstrained form, provides an estimate that is either monotonic or concave uniformly over a grid of points on the interior of the range of x (0.75–1.25). We use three different bandwidths for our simulations. Generically, we use bandwidths of the form $h = c\sigma_x n^{-1/5}$ where c is a user-defined constant, σ_x is the standard deviation of the regressand and n is the sample size being used. A traditional rule-of-thumb bandwidth is obtained by setting $c = 1.06$. We also use $c = 0.53$ (lesser-smoothed) and $c = 2.12$ (greater-smoothed) to assess the impact the bandwidth has on the ability of the unconstrained estimator to satisfy the constraints without further manipulation.

We see that as the sample size is increased from 100 to 200 to 500, the proportion of trials where monotonicity is uniformly found over the grid of points approaches unity. However, concavity is violated much more often. There are many instances, especially when the bandwidth is relatively small,

Table 1. Likelihood of an Estimated Monotonic Regression (9,999 Trials).

	$\varepsilon \sim N(0, 0.1)$			$\varepsilon \sim N(0, 0.2)$		
	100	200	500	100	200	500
$x \sim U[0.5, 1.5]$						
$c = 0.53$	0.996	0.999	1.000	0.731	0.825	0.933
$c = 1.06$	1.000	1.000	1.000	0.999	1.000	1.000
$c = 2.12$	1.000	1.000	1.000	1.000	1.000	1.000
$x \sim N(1, 0.25)$						
$c = 0.53$	0.978	0.993	0.99	0.584	0.699	0.841
$c = 1.06$	1.000	1.000	1.000	0.997	1.000	1.000
$c = 2.12$	1.000	1.000	1.000	1.000	1.000	1.000

Table 2. Likelihood of an Estimated Concave Regression (9,999 Trials).

	$\varepsilon \sim N(0, 0.1)$			$\varepsilon \sim N(0, 0.2)$		
	100	200	500	100	200	500
$x \sim U[0.5, 1.5]$						
$c = 0.53$	0.000	0.000	0.000	0.000	0.000	0.000
$c = 1.06$	0.021	0.033	0.040	0.016	0.016	0.014
$c = 2.12$	0.016	0.004	0.008	0.022	0.007	0.003
$x \sim N(1, 0.25)$						
$c = 0.53$	0.000	0.000	0.000	0.000	0.000	0.000
$c = 1.06$	0.027	0.019	0.014	0.019	0.0190	0.003
$c = 2.12$	0.445	0.527	0.683	0.397	0.427	0.498

where there are no cases where concavity is found uniformly over the grid of points. This result may be unexpected to some given that we have nearly 10,000 replications. Further, we see that as we increase the error variance, this leads to large decreases in the number of cases of both monotonicity and concavity.

Even with these alarming results we note that larger scale factors (c) increase the incidence of concavity. Somewhat surprising is that we do not always see that increasing the sample size leads to higher incidences of concavity. While increasing n increases the number of cases of concavity when we have large bandwidths, we often find the opposite result when $c = 1.06$. This conflicting result likely occurs because of two competing forces. First, the increase in the number of observations leads to more points in the neighborhood of x. This should lead to more cases of concavity. The second effect counteracts the first because increasing the number of observations decreases the bandwidth as $h \propto n^{-1/5}$. Finally, we note that the design of the experiment also has a noticeable effect on the likelihood of observing monotonicity or concavity without resorting to a constrained estimator. For instance, generating the regressor from the Gaussian distribution as opposed to the uniform distribution brings about much larger proportions of concave estimates when the bandwidth is relatively large (likely due to more data in the interior of x).[2]

The results from these tables suggest that constrained estimators are necessary tools for nonparametric analysis, as in even very simple settings direct observation of an unrestricted estimator that satisfies the constraints is by no means expected. One can imagine that with multiple covariates, multiple bandwidths and a variety of constraints to be imposed simultaneously, the likelihood that the constraints are satisfied *de facto* is low.

In general, a wide variety of constrained nonparametric estimation strategies have been proposed to incorporate economic theory within an estimation procedure. While many of these estimators are designed myopically for a specific smoothness constraint, a small but burgeoning literature has focused on estimators which can handle many arbitrary economic constraints simultaneously. Of note are the recent contributions of Racine, Parmeter, and Du (2009) who developed a constrained kernel regression estimator and Beresteanu (2004) who developed a similar type of estimator but for use with spline-based estimators.[3] In addition to providing a survey of the current menu of available constrained nonparametric estimators, we also shed light on the quantitative aspects for empirical implementation regarding the constrained kernel estimator of Racine et al. (2009). While they mention the ability of their method to handle general constraints, their existence results and simulated and real examples all focus on linear (defined in the appropriate sense) restrictions. We augment their discussion by providing existence results as well as heuristic arguments on the implementation of the method. Simulated and empirical evidence targeting imposing concavity on a regression surface is provided to showcase the full generality of the method.

The rest of this paper proceeds as follows. Section 2 reviews the literature on constrained nonparametric regression. Section 3 discusses imposing general nonlinear constraints, specifically concavity, using constraint weighted bootstrapping and shows how it can be implemented computationally. Section 4 presents a small-scale simulation and an empirical discussion of estimation of an age-earnings profile. Section 5 presents several concluding remarks and directions for future research.

2. AVAILABLE CONSTRAINED ESTIMATORS

Consider the standard nonparametric regression model

$$y_i = m(x_i) + \sigma(x_i)\varepsilon_i, \quad \text{for } i = 1, \ldots, n \tag{1}$$

where y_i is the dependent variable, $m(\cdot)$ is the conditional mean function with argument x_i, x_i being a $k \times 1$ vector of covariates, $\sigma(\cdot)$ is the conditional volatility function, and ε_i is a random variable with zero mean and unit variance. Our goal is to estimate the unknown conditional mean subject to economic constraints (e.g., concavity) in a smooth framework.

Imposing arbitrary constraints on nonparametric regression surfaces, while not new to econometrics, has not received as much attention as other

aspects of nonparametric estimation, for instance bandwidth selection, at least not in the kernel regression framework. Indeed, one can divide the literature on imposing constraints in nonparametric estimation frameworks into two broad classes:

1. Developing a nonparametric estimator to satisfy a particular constraint. Here the class of monotonically restricted estimators is a prime example.
2. Developing a nonparametric estimator (either smooth or interpolated) that satisfies a class of constraints.

Our goal is to highlight the variety of existing methods and document the differences across the available techniques to guide the reader to an appropriate estimator for the problem at hand.

2.1. Isotonic Regression

The first constrained nonparametric estimators were nonsmooth and fell under the heading of "isotonic regression," initially proposed by Brunk (1955). Brunk's (1955) estimator was a minmax estimator that was designed to impose monotonicity on a regression function with a single covariate, while Hansen, Pledger, and Wright (1973) extended the estimator to two dimensions and provided results on consistency of the estimator. To explain Brunk's estimator, let $\mathcal{C}_B$ be the discrete cone of restrictions in R^n:

$$\{(z_1, z_2, \ldots, z_n) : z_1 \leq z_2 \leq \cdots \leq z_n\}$$

We let y_i^* be a solution to the minimization problem

$$\min_{(y_1^*, \ldots, y_n^*) \in \mathcal{C}_B} \sum_{i=1}^{n} (y_i - y_i^*)^2$$

This minimization problem has a unique solution that is expressed succinctly by a minmax formula.

Use $X_{(1)}, \ldots, X_{(n)}$ to denote the order statistics of X and $y_{[i]}$ the corresponding observation of $X_{(i)}$. Then our "isotonized" fitted values can be represented as

$$y_i^* = \min_{s \geq i} \max_{t \leq i} \sum_{j=s}^{t} \frac{y_{[j]}}{(t - s + 1)} \tag{2}$$

or

$$y_i^* = \min_{s \le i} \max_{t \ge i} \sum_{j=s}^{t} \frac{y_{[j]}}{(t-s+1)} \tag{3}$$

In Brunk's (1955) approach there is no attempt to smooth the estimation results to values of x between the observation points. A simple approach would be to extend flatly between the values of x_i, but this has been criticized for the presence of too many flat spots and a slow rate of convergence.[4]

Interestingly, Hildreth (1954) introduced a related method to that in Brunk (1955), but geared toward estimating a regression function that is restricted to be concave. His procedure amounts to conducting least squares subject to discretized concavity restrictions. Similar to Brunk (1955), let $\mathcal{C}_H$ be the discrete cone of restrictions in R^n:

$$\left\{(z_1, z_2, \ldots, z_n) : \frac{z_{i+1} - z_i}{x_{i+1} - x_i} \ge \frac{z_{i+2} - z_{i+1}}{x_{i+2} - x_{i+1}}, i = 1, \ldots, n-2\right\}$$

Then y_i^* is a solution of

$$\min_{(y_1^*, \ldots, y_n^*) \in \mathcal{C}_H} \sum_{i=1}^{n} (y_i - y_i^*)^2 \tag{4}$$

An iterative procedure is required to solve the minimization as no closed form solution exists. However, unlike the monotonically constrained estimator of Brunk (1955), the concave restricted estimator of Hildreth (1954) extends between observation points linearly, thus falling into the classification of a least-squares spline estimator.

While both of these estimators construct restricted regression estimates predicated on simple concepts, they are not "smooth" in the traditional sense. The classic isotonic regression estimator of Brunk (1955) was smoothed by Mukerjee (1988) and Mammen (1991a). An alternative way to characterize their estimators is to say that they forced the traditional Nadaraya–Watson regression smoother to satisfy a monotonicity constraint. The key insight was to use a two-step estimator that consisted of a smoothing step and an isotonizing step. Mukerjee (1988) proved that one could preserve the isotonization constructed in the first step by using a log-concave kernel to smooth in the second step. Thus, after one uses either Eq. (2) or (3) to isotonize the regressand, a smooth, nonparametric estimate

of the unknown conditional mean is constructed as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K((x - X_{(i)})/h) y_i^*}{\sum_{i=1}^{n} K((x - X_{(i)})/h)} \tag{5}$$

where h is the bandwidth.[5] One does not need to use a special kernel, however, as a second-order Gaussian kernel is log concave, thus making this method easy to implement. Mammen (1991a) proved that asymptotically the order of the steps is irrelevant. No equivalent estimator exists for the concave variant introduced by Hildreth (1954), and as such the generalizability of smoothing isotonic-type estimators is unknown. Moreover, multivariate extensions to the traditional isotonic regression estimator are difficult to implement and often not available in closed form solutions.

2.2. Constrained Spline/Series Estimation

Both spline- and series-based functions provide the researcher with a flexible set of basis functions with which to construct a regression model that is linear in parameters, which is intuitively appealing. Early methods using splines or series, designed to impose general economic constraints, include Gallant (1981, 1982) and Gallant and Golub (1984). This work introduced the Fourier flexible form (FFF) estimator, whose coefficients could be restricted to impose concavity, homotheticity, and heterogeneity in a nonparametric setting.[6] Constrained spline smoothers were proposed by Dierckx (1980), Holm and Frisen (1985), Ramsay (1988), and Mammen (1991b), to name a few early approaches.

In what follows we describe the basic setup for constrained least-squares spline estimation.[7] We define our spline space to be $\mathcal{S}$ which has dimension p.[8] Our least-squares spline estimate is a function m, which represents a linear combination of spline functions from $\mathcal{S}$ that solves:

$$\min_{s \in \mathcal{S}} \sum_{i=1}^{n} (y_i - m(x_i))^2 \tag{6}$$

To impose constraints we note that positivity of either the first or the second derivative at a given point $\tilde{x}$ of the function $m(\cdot)$ can be written equivalently as positivity of a linear combination of the associated parameters with respect to the chosen basis. Thus, monotonicity or concavity can be readily imposed

on a discretized grid of points where each point adds additional *linear* constraints on the spline coordinates with the associated basis. It is a natural step to include these linear constraints directly into the least-squares spline problem.

Similar to isotonic regression, the literature appears to have focused on concavity first (Dierckx, 1980) and then monotonicity (Ramsay, 1988). In what will seen to be a common theme in constrained nonparametric regression, Dierckx (1980) used a quadratic program to enforce *local* concavity or convexity of a spline function. His function estimate, using normalized B-splines (see Schumaker, 1981) with basis N_j, is

$$\hat{m}(x) = \sum_{j=-3}^{k} c_j^* N_j(x)$$

Here k denotes the total number of knots. The values c_j^* solve the quadratic program

$$\min_{\sum_{j=-3}^{k} d_{j,l} c_j e_j \leq 0} \sum_{i=1}^{n} \left(y_i - \sum_{j=-3}^{k} c_j N_j(x_i) \right)^2 \tag{7}$$

The e_j in Eq. (7) determines the type of constraint being imposed on the function locally. That is, $e_j = 1$ if the function is locally convex at knot ℓ, $e_j = 0$ if the function is unrestricted at the ℓth knot and $e_j = -1$ if the function is locally concave at knot ℓ. The numbers $d_{j,l}$ are derived from the second derivatives of the basis splines at each of the knots, and have a simple representation

$$d_{j,l} = 0 \qquad \text{if } j \leq l-4 \text{ or } j \geq 4$$

$$d_{l-3,l} = \frac{6}{(t_{l+1} - t_{l-2})(t_{l+1} - t_{l-1})}$$

$$d_{l-1,l} = \frac{6}{(t_{l+2} - t_{l-1})(t_{l+1} - t_{l-1})}$$

$$d_{l-2,l} = -(d_{l-3,l} + d_{l-1,l})$$

where t_l refers to the lth point under consideration. Ramsay (1988) developed a similar monotonically constrained spline estimator using I-splines. I-splines have a direct link to the B-splines used by Dierckx (1980). An I-spline of order

M is an indefinite integral of a corresponding B-spline of the same order. Ramsay (1988) used I-splines because he was able to establish that they had the property that each individual I-spline is monotonic and that any linear combination of I-splines with positive coefficients is also monotonic. This made it easy to construct the associated monotonic spline estimator. Both of the aforementioned estimators can also be placed in the smoothing spline domain as well.

Yatchew and Bos (1997) developed a series-based estimator that can handle general constraints. This estimator is constructed by minimizing the sum of squared errors of a nonparametric function relative to an appropriate Sobolev norm. The basis functions that make up the series estimation are determined from a set of differential equations that provide "representors." Representors of function evaluation consist of two functions spliced together, where each of these functions is a linear combination of trigonometric functions. In essence, one can "represent" any function in Sobolev space through this process (see Yatchew & Bos, 1997, Appendix 2). Let R be an $n \times n$ "representor" matrix whose columns (equivalently rows) equal the representors of the function, evaluated at the observations $x_1, \ldots, x_n$.[9] Then, arbitrary constrained estimation of a nonparametric function

$$\min_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^{n} (y_i - m(x_i))^2 \quad \text{s.t } ||m||^2_{Sob} \leq L$$

can be recast as

$$\begin{aligned} &\min_{c} n^{-1} \sum_{i=1}^{n} (y_i - R_c)^2 \\ &\text{s.t. } c'Rc \leq L, c'R^{(1)}c \leq L^{(1)}, c'R^{(2)}c \leq L^{(2)}, \ldots, c'R^{(k)}c \leq L^{(k)} \end{aligned} \tag{8}$$

Here L denotes the upper bound on the squared Sobolev norm of our constrained function, c is an $n \times 1$ vector of coefficients, and $\mathcal{F}$ is our constrained function space which we are searching over. Since we are interested in constraints that relate directly to the derivatives of the nonparametric function we are estimating, $R^{(1)}, \ldots, R^{(k)}$ represent the appropriate derivatives of the original representor matrix and $L^{(1)}, \ldots, L^{(k)}$ are the corresponding bounds. For example, if one wished to impose monotonicity, $L^{(1)} = 0$ and $R^{(1)}$ represents the representor matrix with each of the representors first-order differentiated with respect to the corresponding column's variable (i.e., the fifth column of $R^{(1)}$ corresponds to the fifth

covariate so the representors are first-order differentiated with respect to that variable). Again, this is a quadratic programming (QP) problem with a quadratic constraint.[10]

Beresteanu (2004) introduced a spline-based procedure that can handle multivariate data and impose multiple, general, derivative constraints. His estimator is solved via QP over an equidistant grid created on the covariate space. These points are then interpolated to create a globally constrained estimator. He employed his method to impose monotonicity and supermodularity of a cost function for the telephone industry. His estimation setup is similar to the approaches described above and involves setting up a set of appropriately defined constraint matrices for the shape constraint(s) desired and solving for a set of coefficients, then interpolating these points to construct the nonparametric function that satisfies the constraints over the appropriate interval. In essence, since Beresteanu (2004) is constructing his estimator first based on a grid of points and then interpolating, this estimation procedure can be viewed as a two-step series-based equivalent of the isotonic regression discussed earlier (Mukerjee, 1988).

2.3. The Matzkin Approach

The seminal work of Matzkin (1991, 1992, 1993, 1994, 1999) considered identification and estimation of general nonparametric problems with arbitrary economic constraints. One of her pioneering insights was that when nonparametric identification was not possible, imposing shape constraints tied to economic theory could provide nonparametric identification in certain estimation settings. Her work laid the foundations for a general operating theory of constrained nonparametric estimation. Her methods focused on standard economic constraints (monotonicity, concavity, homogeneity, etc.) but were capable of being facilitated in more general settings than regression. Primarily, her work focused on binary-threshold crossing models and polychotomous choice models, although her definition of subgradients equally carried over to a regression context. One can suitably recast her estimation method in the regression context as nonparametric constrained least squares.

For example, to impose concavity on a regression function she created "subgradients," T^j, which were defined for any convex function $m : X \to \mathbb{R}^k$, where $X \subset \mathbb{R}$ is a convex set and $x \in X$ for any vector $T \in \mathbb{R}^k$ such that $\forall_y \in X\ m(y) \geq m(x) + T(y-x)$.[11] We use the notation T^j to denote

that the subgradients are calculated for the observations. Matzkin (1994) showed how to use the subgradients to impose concavity and monotonicity simultaneously. Using the Hildreth (1954) constraints for concavity of a regression surface, Matzkin (1994) rewrites them as

$$m(x_i) \leq m(x_j) + T^j(x_i - x_j), \quad i,j = 1, \ldots, n$$

She solves the minimization problem in Eq. (4), but the minimization is over $m(x_i) \forall i$ and $T^j \; \forall j$. To impose monotonicity one would add the additional constraint that $T^j > 0 \forall j$. Algorithms to solve the constrained optimization problem were first developed for the regression setup by Dykstra (1983), Goldman and Ruud (1992), and Ruud (1995) and for general functions by Matzkin (1999), who used a random search routine regardless of the function to be minimized.

Implementation of these constrained methods is of the two-step variety (see Matzkin, 1999). First, for the specified constraints, a feasible solution consisting of a finite number of points is determined through optimization of some criterion function (in Matzkin's choice framework setups this is a pseudo-likelihood function). Second, the feasible points are interpolated or smoothed to construct the nonparametric surface that satisfies the constraints. These methods can be viewed in the same spirit as that of Mukerjee (1988), but for a more general class of problems.

2.4. Rearrangement

Recent work on imposing monotonicity on a nonparametric regression function, known as rearrangement, is detailed in Dette, Neumeyer, and Pilz (2006) and Chernozhukov, Fernandez-Val, and Galichon (2009). The estimator of Dette et al. (2006) combines density and regression techniques to construct a monotonic estimator. The appeal of "rearrangement" is that no constrained optimization is required to obtain a monotonically constrained estimator, making it computationally efficient compared to the previously described methods. Their estimator actually estimates the inverse of a monotonic function, which can then be inverted to obtain an estimate of the function of interest.

To derive this estimator let M denote a natural number that dictates the number of equi-spaced grid points to evaluate the function. Then, their

estimator is defined as

$$\hat{m}^{-1}(x) = \int_{-\infty}^{x} \frac{1}{Mh} \sum_{j=1}^{M} K\left(\frac{\hat{m}(j/M) - u}{h}\right) du \tag{9}$$

where $\hat{m}(x)$ is any unconstrained nonparametric regression function estimate (kernel smoothed, local polynomial, series, splines, neural network, etc.). The intuition behind this estimator is simple; the connection rests on the properties of transformed random variables.

Note that $m(x_i)$ is a transformation of the random variable x_i. The estimator

$$\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{m(x_i) - u}{h}\right)$$

represents the classical kernel density of the random variable $u = m(x_1)$, which has density

$$g(u) = f(x_1)|(m^{-1})'(x_1)|$$

The integration in Eq. (9) is that of a probability density function and as such a CDF is constructed, which is always *monotonically increasing*. The equi-spaced grid is used for the estimation since the evaluation points are then treated as though they came from a uniform density, making $f(j/M) = I[a, b]$, where a and b denote the lower and upper bounds of the support of X, respectively. Thus, the integration in this case amounts to integrating $|(m^{-1})'(x_1)|$ over its domain, which gives us $m^{-1}(x_1)$. Once this has been obtained, it is a simple matter to reflect this estimate across the $y = x$ line in Cartesian 2-space to obtain our monotonically restricted regression estimator. Chernozhukov et al. (2009) discuss implementation of this estimator in a multivariate setting and show that the constrained estimator *always* improves (reduces the estimation error) over an original estimate whenever the original estimate is not monotonic.

The name rearrangement comes from the fact that the point estimates are rearranged so that they are in increasing order (monotonic). This happens because the kernel density estimate of the first-stage regression estimates sorts the data from low to high to construct the density, which is then integrated. This sorting, or rearranging, is how the monotonic estimate is produced. It works because monotonicity as a property is nothing more than a special ordering, and the kernel density estimator is ''unaware'' that

the points it is smoothing over to construct a density are from an estimate of a regression function as opposed to raw data.

One issue with this estimator is that while it is intuitive, computationally simple, and easy to implement with existing software, it requires the selection of two "bandwidths."[12] Additionally, the intuition underlying the ease of implementation does not readily extend itself to general constraints on nonparametric regression surfaces. No such transformation is obtainable to impose concavity using the same insights, for example.

2.5. *Data Sharpening*

Data sharpening derives from the work of Friedman, Tukey, and Tukey (1980) and was later employed in Choi and Hall (1999). These methods are designed to admit a wide range of constraints and are closely linked to biased-bootstrap methods (Hall & Presnell, 1999). Data sharpening is inherently different than biased-bootstrapping and constraint weighted bootstrapping (to be discussed later) as it alters the data, but keeps the weights associated with each point fixed, whereas biased-bootstrapping and constraint weighted bootstrapping change the weights associated with each point, but keep the points fixed. Both of these methods, however, can be thought of as data tuning methods which in some sense alter the underlying empirical distribution to achieve the desired outcome. We discuss the method of Braun and Hall (2001) in what follows.

Let our original data be $\{x_1, \ldots, x_n\}$ and our sharpened data be $\{z_1, \ldots, z_n\}$. Define the distance between original and sharpened points as $D(x_i, z_i) \geq 0$. We choose $\mathcal{Z} = \{z_1, \ldots, z_n\}$, our set of sharpened data, to minimize

$$D(\mathcal{X}, \mathcal{Z}) = \sum_{i=1}^{n} D(x_i, z_i)$$

subject to our constraints of interest. Once the sharpened data have been obtained we apply our method of interest, in this setting nonparametric regression, to the sharpened data.

More formally, our kernel regression (local constant, say) estimator is

$$\hat{m}(x|\mathcal{X}, \mathcal{Y}) = \frac{\sum_{i=1}^{n} K((x_i - x)/h) y_i}{\sum_{i=1}^{n} K((x_i - x)/h)} = \sum_{i=1}^{n} A_i(x) y_i$$

We want to impose an arbitrary constraint on the function, monotonicity for example, by "sharpening" y. Thus, we minimize

$$D(\mathcal{Y}, \mathcal{Q}) = \sum_{i=1}^{n} D(y_i, q_i) \tag{10}$$

for a preselected distance function, subject to the constraints

$$\hat{m}(x|\mathcal{X}, \mathcal{Q}) = \sum_{i=1}^{n} A_i'(x) q_i > 0 \tag{11}$$

Notice the conditioning set for which the estimator is defined over has changed from $\mathcal{Y}$ to $\mathcal{Q}$ Thus, we *construct* our restricted estimator while simultaneously minimizing our criterion function. If one chose $D(r,\ t) = (r-t)^2$, we would have a standard QP problem, provided the constraints were linear (which they are in our monotonicity example). Compared to rearrangement, given the fact that the data is smoothed, even though the response variables are moved around, the corresponding constrained curve is as smooth as the unconstrained curve. The rearranged curve will have ambiguous low-order kinks where the nonmonotonic portion of the curve is "forced" to be monotonic resulting in a curve that is less smooth than its unconstrained counterpart.

2.6. Constraint Weighted Bootstrapping

Hall and Huang (2001) suggest an alternative smooth, monotonic nonparametric estimator that admits any number of covariates. Racine et al. (2009) have generalized the method to accommodate a variety of "linear" constraints simultaneously. Start again with the standard local constant least-squares estimator

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K((x_i - x)/h) y_i}{\sum_{i=1}^{n} K((x_i - x)/h)} = \frac{1}{n} \sum_{i=1}^{n} A_i(x) y_i \tag{12}$$

where $A_i(x) = nK((x_i - x)/h)/\hat{f}(x)$ and $\hat{f}(x) = \sum_{i=1}^{n} K((x_i - x)/h)$. Even though we are choosing to use the local constant least-squares framework, this setup can be immediately extended to other types of kernel and local polynomial estimation routines. As it stands, the regression estimator in

Eq. (12) is not guaranteed to produce a monotonic estimator. Hall and Huang's (2001) insight was to introduce observation-specific weights p_i instead of the $1/n$ that appears in Eq. (12). These weights can then be manipulated so that the estimator satisfies monotonicity. To be clear,

$$\hat{m}(x|p) = \sum_{j=1}^{n} p_i A_i(x) y_i$$

is the constraint weighted bootstrapping estimator. It is still not monotonic until we properly restrict the weights.

In the unconstrained setting we have $p = (p_1, \ldots, p_n) = (1/n, \ldots, 1/n)$, which represents weights drawn from a uniform distribution. If the bandwidth chosen produces an estimate that *is already* monotonic, the weights should be set equal to the uniform weights. However, if the function by itself is not monotonic, then the weights are diverted away from the uniform case to create a monotonic estimate. In order to decide how to manipulate the weights, a distance metric is introduced based on power divergence (Cressie & Read, 1984):

$$D_\rho(p) = \frac{1}{\rho(1-\rho)}\left[n - \sum_{i=1}^{n}(np_i)^\rho\right], -\infty < \rho < \infty \tag{13}$$

where $\rho \neq 0, 1$. One needs to take limits for $\rho = 0$ or 1. They are given as

$$D_0(p) = -\sum_{i=1}^{n}\log(np_i); \quad D_1(p) = \sum_{i=1}^{n} p_i \log(np_i)$$

This distance metric is quite general. If one uses $\rho = 1/2$, then this corresponds to Hellinger distance, whereas $nD_0(p)+n^2\log(n)$ is equivalent to Kullback–Leibler divergence ($-\sum_{i=1}^{n} n\log(p_i/n)$). This metric is minimized for a selected ρ subject to the constraint that

$$\hat{m}'(\cdot|p) = \sum_{j=1}^{n} p_i A_i'(\cdot) y_i \geq \varepsilon$$

on a grid of selected points. Here $\varepsilon \geq 0$ can be used to guarantee either weak or strict monotonicity. A nice feature of this estimator is that the kernel and bandwidth are chosen before the weights are selected. This means that the user can choose their desired kernel estimator and bandwidth selector to construct their nonparametric estimator and then constrain it to be monotonic. This leaves the door open to straightforward modification of the

estimator. In fact, there is nothing special about monotonicity for the method of Hall and Huang (2001) to work. Any constraint that is desired could, in principle, be imposed on the regression surface.

Note that the monotonic constraint imposed in Hall and Huang (2001) can be written in the more general form:

$$\sum_{i=1}^{n} p_i \left[\sum_{s \in S} \alpha_{\mathbf{s}} A_i^{(\mathbf{s})}(x) \right] y_i - c(x) \geq 0 \tag{14}$$

where the inner sum is taken over all vectors $\mathbf{S}$ that correspond to our constraints of interest (monotonicity, say), $\alpha_{\mathbf{s}}$ are a set of constants used to generate various constraints, and $c(x)$ is a known function. $\mathbf{S}$ indexes the order of the derivative associated with the kernel portion of the regression estimator. In our example of monotonicity, $\mathbf{s} = e_j$ is a k-vector (since we have $x \in \mathbb{R}^k$) with 1 in the jth position and 0s everywhere else, $\alpha_{\mathbf{s}} = 1 \forall \mathbf{s} \in \mathbf{S}$ and $c(x) = 0$.[13] Racine et al. (2009) provide existence and uniqueness for a set of weights for constraints of the form (14). They call these constraints linear since they are linear with respect to the weights $p_i \forall_i$. Additionally, to make the constrained optimization computationally simple, they use the L_2 norm with respect to the uniform weights $(1/n)$, as opposed to the power divergence metric. This condenses the problem into a standard QP problem, which can be solved using existing packages in almost all standard econometric software.

Note the subtle difference between the data sharpening methods discussed previously and the constraint weighted bootstrapping methods here. When one chooses to sharpen the data, the actual data values are being transformed while the weighting is held constant. Here, the exact opposite occurs: the data is held fixed while the weights are changed. At the end of the day however, the two estimators can be viewed as ''visually'' equivalent. That is, both estimators can be looked at as

$$\hat{m}(x) = \sum_{j=1}^{n} A_i(x) y_i^* \tag{15}$$

where y_i^* corresponds to either the sharpened values or $p_i y_i$ obtained from the constraint weighted bootstrapping approach. The difference between the methods is how y_i^* is arrived at.[14] Also, note that both constraint weighted bootstrapping and data sharpening are *vertically* moving the data, whereas rearrangement methods *horizontally* move the data.

2.7. Summary of Methods

While our discussion of existing methods has indicated a number of choices for the user, there does not exist one clear-cut method for imposing arbitrary constraints on a regression surface for every given situation. Each of the methods discussed has computational or theoretical drawbacks when considered against the set of all available methods. Additionally, several of the key differences across the methods focus on the choice of operating in a kernel, spline, or series-based framework, the selection of smoothing parameters, the smoothness of the estimator, the adaptability/generalizability of the method, whether to impose global or discrete constraints, and the ability to use the method to conduct inference on the constraints being imposed.

2.7.1. Spline, Series, and Kernels

Given that the constrained estimation methods discussed earlier use vastly differing nonparametric methods, this choice cannot be overlooked. Kelly and Rice (1990) mention that if the coefficients in the B-spline bases are nondecreasing, then so is the function (if one was imposing monotonicity), and Delecroix and Thomas-Agnan (2000) focus attention on the fact that splines are defined as the solution to a minimization problem and this, in general, lends support for their use in constrained settings. However, given the prevalence of discrete data in applied settings, the seminal work of Racine and Li (2004) highlighting the fact that smoothing categorical data can lead to substantial finite sample efficiency gains, lends support for adopting a kernel-based method. Alternatively, given the ease with which one may construct and employ series-based methods, it is easy to advocate that these constrained methods are computationally easy to employ.

Given the adaptability of the methods of Yatchew and Bos (1997) (which is series based), Beresteanu (2004) (which is spline based), and Racine et al. (2009) (which is kernel based), we cannot advocate for a particular type of nonparametric method based on imposing general smoothness constraints. Nor do we advocate on behalf of the particular type of nonparametric smoothing one should engage in. However, given the ease with which one can implement a constrained estimator, we remark that the easiest method for which a researcher can incorporate the constraints should be used. Additionally, if a researcher traditionally uses a type of nonparametric method (spline, say), then they may have more familiarity with employing one set of constrained methods over another, which is an obvious benefit.

2.7.2. Choice of Smoothing Parameter

As with all nonparametric estimation methods, the choice of smoothing parameter plays a crucial role to the performance of the estimator both in practice and theory, yet there was no mention of the appropriate level of smoothing in the aforementioned constrained methods. Few results exist suggesting how the optimal level of smoothing should be imposed. For many of the methods described previously, one could engage in cross-validation simultaneously with the constraint imposition. This may actually help in determination of the optimal smoothing parameter. The simulations of Delecroix and Thomas-Agnan (2000) show that the mean integrated square error (typically used in cross-validation) as a function of the smoothing parameter typically had a wider zone of stability around the optimal level of the smoothing parameter, suggesting it may be easier to determine the optimal level; it is well known that various forms of the cross-validation function are noisy, making determination of the optimal level difficult in certain settings.

However, engaging in cross-validation and constraint imposition simultaneously is unnecessary in particular methods. For example, the constraint weighted bootstrapping methods of Hall and Huang (2001) and Racine et al. (2009) show that the constrained kernel estimator should use a bandwidth of the standard, unconstrained optimal order. In this setting both the restricted and unrestricted smooths will have the same level of smoothing. Further tuning could be performed by cross-validation after the constraint weights have been found and simple checks to determine if the constraints were still satisfied (similar to that described above).

2.7.3. Method Complexity

The methods discussed earlier range from simple computation (rearrangement and univariate isotonic regression) to involving quadratic or nonlinear program solvers. These numerical methods may dissuade the user from adopting a specific approach, but we note that with the drastic reductions in computation time and the availability of solvers in most econometric software packages, these constraints will continue to lessen over time. Indeed, part of this survey discusses in detail the implementation of a sequential quadratic program to showcase its implementation in practice. Also, given the ease with which a quadratic program can be solved with linear constraints, the method of Racine et al. (2009) addresses the critique of Dette and Pilz (2006, p. 56) who note "[rearrangement offers] substantial computational advantages, because it does not rely on constrained optimization methods." We mention here that rearrangement requires slightly more sophistication when one migrates from a univariate to multivariate setting and so this concern is of limited use in applied work.

2.7.4. Numerical Comparisons

Very little theoretical work exists to showcase the performance of one method against a set of competitors. Indeed, even numerical comparisons are scant. The most comprehensive study between methods is that of Dette and Pilz (2006) who conducted a Monte Carlo comparison of smooth isotonic regression and rearrangement, and the method of Hall and Huang (2001) for the constraint of monotonicity, in the univariate setting for a bevy of DGPs. Their findings suggest that rearrangement has desirable/equivalent finite sample performance compared to the other methods across all of the DGPs considered.

3. IMPOSING NONLINEAR CONSTRAINTS

We discuss a further generalization of Racine et al. (2009) that can handle general nonlinear constraints and discuss in detail the computational method of sequential quadratic programming (SQP) required to implement nonparametric regression in this setting. Our choice for a deeper, prolonged discussion of this method hinges on the necessity of SQP methods in several of the methods mentioned previously. Very rarely are the methods to obtain a solution discussed at length, and given the use of these methods in both data sharpening and constraint weighted bootstrapping, we feel it requisite to highlight the implementation of this technique.

While we discuss general constrained estimation in the face of arbitrary nonlinear constraints, to cement our ideas we focus on the specific example of concavity. Concavity is a common assumption used in the characterization of production functions. Concavity of the production function implies diminishing marginal productivity of each input.[15] This assumption is widely agreed upon by economists, and failure to impose it may lead to conclusions that are economically infeasible.

In the case of a single factor, a twice continuously differentiable function $m(x)$ is said to be concave if $m''(x) \leq 0 \ \forall x \in \mathcal{S}(x)$. Extending this result to the case of multiple x is relatively straightforward. Concavity implies that the Hessian matrix

$$H(m(x)) = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ m_{21} & m_{22} & \cdots & m_{2k} \\ \vdots & & \ddots & \vdots \\ m_{k1} & m_{k2} & \cdots & m_{kk} \end{bmatrix}$$

where $m_{lk} \equiv (\partial^2 m(x))/(\partial x_l \partial x_k)$ must be negative semidefinite. In other words, all the lth ($l = 1, 2, \ldots, k$) order principal minors of H are less than or equal to zero if l is odd, and greater than or equal to zero if l is even (alternatively, all the eigenvalues of this matrix are negative). We could, instead, choose to impose concavity via the constraints given in Hildreth (1954); however, many formal definitions of concavity are linked to the Hessian and as such we enforce concavity using this.

Following Hall and Huang (2001), we have the following constrained nonlinear programming problem:

$$\begin{aligned} &\min D_\rho(p) \;\; s.t. \;\; H(m(x|p)) \text{ is negative semidefinite } \forall x \in \mathcal{S}(x), \\ &p_i \geq 0 \; \forall i, \;\; \text{and} \;\; \sum_{i=1}^{n} p_i = 1 \end{aligned} \tag{16}$$

To solve this or any other constrained optimization problem in the spirit of Hall and Huang (2001) we need to use SQP.

3.1. Sequential Quadratic Programming

Although the steps to construct a constrained nonparametric estimator seem straightforward, implementing these types of programs are often not discussed in detail in econometrics papers. In this subsection we outline SQP.

Consider the inequality constrained problem

$$\min D(z) \text{ subject to } r_i(z) = 0, \;\; i \in \mathcal{E}, \;\; \text{and} \;\; c_j(z) \geq 0, \;\; j \in \mathcal{I} \tag{17}$$

where $D : \mathbb{R}^{q_o} \to \mathbb{R}, r_i : \mathbb{R}^{q_o} \to \mathbb{R}^{q_1}$, and $c_i : \mathbb{R}^{q_o} \to \mathbb{R}^{q_2}$ can all be nonlinear, but we require that all the functions are smooth in the z argument. The idea behind SQP is to convert the nonlinear programming problem in Eq. (17) into a conventional QP problem. To do this we need to "linearize" our constraints and "quadracize" our objective function. Before doing this we introduce some additional concepts.

The Lagrangian of our problem is defined as

$$\mathcal{L}(z, \lambda_r, \lambda_c) = D(z) - \lambda_r' r_i(z) - \lambda_c' c_j(z) \tag{18}$$

Also, define $B_r(z)' = [\nabla r_1(z), \nabla r_2(z), \ldots, \nabla r_n(z)]$ and $B_c(z)' = [\nabla c_1(z), \nabla c_2(z), \ldots, \nabla c_n(z)]$. Now pick an initial z, z_0, and an initial set of vectors of Lagrange multipliers, $\lambda_{r,0}$ and $\lambda_{c,0}$. Lastly, define $\nabla^2 \mathcal{L}_{zz}(z, \lambda_r, \lambda_c) = \nabla^2 D(z) - \nabla B_r(z)' \lambda_r - \nabla B_c(z)' \lambda_c$. We are now ready to describe how to solve our SQP problem.

Our QP at step 0 is

$$\min\ D(z_0) + \nabla D(z_0)'q + \frac{1}{2} q' \nabla^2_{zz} \mathcal{L}(z_0, \lambda_{r,0},\ \lambda_{c,0}) q \tag{19}$$

subject to

$$B_r(z_0)q + r(z_0) = 0 \text{ and } B_c(z)q + c(z_0) \geq 0 \tag{20}$$

The solution of this standard quadratic program, q_0, $\ell_{r,0}$, and $\ell_{c,0}$, can be used to update z_0, $\lambda_{r,0}$, and $\lambda_{c,0}$ as follows: $z_1 = z_0 + q_0$, $\lambda_{r,1} = \ell_{r,0}$, and $\lambda_{c,1} = \ell_{c,0}$. These updated values can then be plugged back into the SQP to repeat the whole process until convergence. SQP requires nothing more than repeated evaluation of the levels, first- and second-order derivatives of the objective and constraint functions. It is a simple matter to determine these derivatives; thus, this simplification process requires nothing more than taking derivatives of a set of functions.

3.2. Existence and Uniqueness of a Solution

When the following assumptions hold:

1. the constraint Jacobians $B_r(z)$ and $B_c(z)$ have full row rank, and
2. the matrix $\nabla^2_{zz} \mathcal{L}(z, \lambda_r, \lambda_c)$ is positive definite on the tangent space of constraints

our SQP has a unique solution that satisfies the constraints. Essentially, this result comes from the fact that one could have used Newton's method to solve the constrained optimization, and the result here is obtained from the associated iterate from running Newton's method instead. These two assumptions are enough to guarantee that a unique solution holds if one were to use Newton's method instead of the one we outlined. However, Nocedal and Wright (2000, pp. 531–532) show that these two procedures, in this setting, are equivalent. For more on existence of a *local* solution we direct the interested reader to Robinson (1974).

Additionally, since we have converted our general nonlinear programming problem into a QP problem, the conditions required for existence of a solution in QP problems are exactly the conditions we need to hold, at each iteration, to guarantee a solution exists in this setting. Thus, the results established in Racine et al. (2009) carry over to our setting, provided our nonlinear constraints are first-order differentiable in p and satisfy our

assumptions listed above, which are easily checked. Moreover, if the forcing matrix $(\nabla^2_{zz}\mathcal{L}(q, \lambda_r, \lambda_c))$ in the quadratic portion of our "quadricized" objective function is positive semidefinite, and if our solution satisfies the set of linearized equality/inequality constraints, then our solution is the unique, global solution to the problem (Nocedal & Wright, 2000, Theorem 16.4). Positive semidefiniteness guarantees that our objective function is convex, which is what yields a global solution. We note that this only shows uniqueness but does not guarantee a solution will exist.

However, it should be noted that because the constraint weights are restricted to be nonnegative and sum to 1, this implies that it may be difficult to impose a constraint that is "far away" from being satisfied. In essence, the constraints imposed on the problem may be inconsistent if a nonnegative weight or a weight greater than 1 is *needed* to satisfy the constraints of interest. However, the conditions needed to determine how far away is "far away" are not investigated here. Our conjecture is that the distance from an observation and the underlying function is dependent on the error process that perturbs the data generating process.

In essence the weights act as vertical scaling factors, and if the amount of scaling is restricted, then it can be difficult to find a solution. Hall and Presnell (1999) note the difficulty in finding the appropriately sharpened points using essentially the same technique described here in roughly 10% of their simulations. They advocate for an approach similar to simulated annealing that was always able to arrive at a solution although that procedure was computationally more intensive than SQP. An alternative, not followed here, would be to dispense with the power divergence metric and all constraints on the weights if no solution is found in the SQP format. In this setting one could use the L_2 norm of Racine et al. (2009) and linearize (provided the nonlinear constraints are differentiable) the nonlinear constraints, again engaging in an iterative procedure to determine the optimal set of weights that can be shown to always exist in this setting.

3.3. SQP Imposing Concavity

If we use the power divergence measure of Cressie and Read (1984):

$$D_\rho(p) = \frac{1}{\rho(1-\rho)}\left\{n - \sum_{i=1}^{n}(np_i)^\rho\right\}$$

for $-\infty < \rho < \infty$ and $\rho \neq 0, 1$, as our objective function to minimize, then we have the following set of functions that need to be estimated prior to solving our QP at any iteration (ℓth):

(i) $D_\rho(p_\ell) \equiv \dfrac{1}{\rho(1-\rho)} \left\{ n - \sum_{i=1}^{n} (np_{i,\ell})^\rho \right\}$.

(ii) $\nabla D_\rho(p_\ell) = v\text{ec}\left[\dfrac{-n}{1-\rho} (np_{i,\ell})^{\rho-1} \right]$.

(iii) $\nabla^2 D_\rho(p_\ell) = \text{diag}[n^2 (np_{i,\ell})^{\rho-2}]$.

(iv) $r(z) \equiv \sum_{i=1}^{n} p_{i,\ell} - 1$.

(v) $B_r(p_\ell) = [1, 1, \ldots, 1]$, an n-vector of 1s.

(vi) $\nabla B_r(p_\ell)$ which is an $n \times n$ matrix of 0s.

Our objective function is defined in (i), whereas (ii) and (iii) are the first and second partial derivatives of our objective function, respectively. Our equality constrained function (ensuring the weights sum to 1) is defined in (iv) and the first and second partial derivatives of this function are given in (v) and (vi).

Additionally, we have to calculate our inequality constrained functions as well as their first and second partial derivatives, which can be broken into two pieces. First, we focus directly on the linear inequality constraints $p_i \geq 0 \forall_i$. For this we have

(i) $B_{c,1}(p_\ell) = [e_1, e_2, \ldots, e_n]$, where e_j is an n-vector of 0s with a 1 in the jth spot.

(ii) $\nabla B_{c,1}(p_\ell)$, which is an $n \times n$ matrix of 0s.

We also have to calculate the first and second derivatives of the determinants of the principal minors of our Hessian matrix *for each* point we wish to impose concavity. In a local constant setting, the Hessian matrix is calculated as follows. Assume that we have q continuous covariates and we are smoothing with a standard product kernel with second-order, individual Gaussian kernels. Then, we have

$$\frac{\partial K_i(x)}{\partial x_s} = -\left(\frac{x_s - x_{si}}{h_s^2} \right) K_i(x), K_i(x) = (2\pi)^{-q/2} \prod_{j=1}^{q} h_j^{-1} e^{-(x_j - x_{ji})^2 / 2h_j^2} \qquad (21)$$

and we can easily determine that

$$\frac{\partial^2 K_i(x)}{\partial x_s \partial x_r} = \left[\left(\frac{x_s - x_{si}}{h_s^2} \right) \left(\frac{x_r - x_{ri}}{h_r^2} \right) + \delta_{sr} \frac{1}{h_s^2} \right] K_i(x) \tag{22}$$

where $\delta_{sr} = 1$, when $s = r$ and is 0 otherwise.

Recalling that $A_i(x) = nK_i(x)/\sum_{i=1}^n K_i(x)$ we have

$$\begin{aligned} \frac{\partial A_i(x)}{\partial x_s} &= \frac{n((\partial K_i(x))/\partial x_s) \sum_{i=1}^n K_i(x) - nK_i(x) \sum_{i=1}^n (\partial K_i(x))/\partial x_s}{\left[\sum_{i=1}^n K_i(x) \right]^2} \\ &= A_i(x) \left[n^{-1} \sum_{i=1}^n D_i(x_s) A_i(x) - D_i(x_s) \right] = A_i(x) M_s(x) \end{aligned} \tag{23}$$

where $D_i(x_s) = (x_s - x_{si})/(h_s^2)$. Similar arguments show that

$$\begin{aligned} \frac{\partial^2 A_i(x)}{\partial x_s \partial x_r} &= \frac{\partial A_i(x)}{\partial x_r} M_s(x) + A_i(x) \frac{\partial M_s(x)}{\partial x_r} \\ &= A_i(x) M_s(x) M_r(x) + A_i(x) \left[M_r(x) n^{-1} \sum_{i=1}^n D_i(x_s) A_i(x) \right] \\ &= A_i(x) M_r(x) [2M_s(x) + D_i(x_s)] \end{aligned} \tag{24}$$

Our first-order partial derivatives of our local constant smoother are

$$\frac{\partial \hat{m}(x|p)}{\partial x_s} = \sum_{i=1}^n p_i y_i \frac{\partial A_i(x)}{\partial x_s} = \sum_{i=1}^n p_i y_i A_i(x) M_s(x) \tag{25}$$

Note that we cannot pull $M_s(x)$ through the summation since it has a $D_i(x_s)$ inside of it so that it depends on the counter. To determine the second-order partial derivatives of our smooth regression function we use our results from Eq. (24) to obtain

$$\begin{aligned} \frac{\partial^2 \hat{m}(x|p)}{\partial x_s \partial x_r} &= \sum_{i=1}^n p_i y_i \frac{\partial^2 A_i(x)}{\partial x_s \partial x_r} = \sum_{i=1}^n p_i y_i [A_i(x) M_r(x)(2M_s(x) + D_i(x_s))] \\ &= 2 \sum_{i=1}^n p_i y_i A_i(x) M_r(x) M_s(x) + \sum_{i=1}^n p_i y_i A_i(x) M_r(x) D_i(x_s) \end{aligned} \tag{26}$$

One can save computation time by noting that terms required for calculation of $M_s(x)$, $M_r(x)$, and $D_i(x_s)$ are all calculated when $A_i(x)$ is

calculated. We suggest using numerical techniques in the user's preferred software to calculate the first and second derivatives of the Hessian matrix to then pass to the SQP.[16] For k covariates, if one imposes concavity for each of the n points, then this requires construction of n $k \times k$ Hessian matrices. There are k determinants of principal minors (or k eigenvalues) to be calculated for each Hessian representation, resulting in nk constraints to go with the $n+1$ constraints placed on the weights. This results in a total of $n(k+1)+1$ total constraints.[17] As noted in the introduction, imposing concavity over the entire support of the data may be burdensome since it will be harder to enforce the constraints near the boundaries. However, using an interior hypercube of the data will lessen the burden on the SQP since concavity is less likely to be violated (assuming concavity holds in the limit) on the interior of the support.

4. DEMONSTRATION

4.1. Simulated Examples

This section uses Monte Carlo simulations to examine the finite sample performance of the nonlinearly constrained estimator described above. Following the focus on concavity, we choose to perform our simulations imposing concavity in models which should be concave. We consider the following data generating process used to motivate our problem in the introduction:

$$y = \ln(x) + u \tag{27}$$

where x is generated as uniform distribution from 0.5 to 1.5, and u is generated as normal with mean zero and variance equal to 0.1. Note that this data generating process produces a theoretically consistent concave function. However, both the unknown error and finite sample biases of the estimator itself may cause the kernel estimate to exhibit ranges of nonconcavities.

We consider samples of $n = 100$ and 500 for each of our 999 Monte Carlo replications. We present results using $\rho = 0.5$, but note that other choices for ρ do not significantly change the results. We use local-constant least-squares and a Gaussian kernel with $h = 1.06\sigma_x n^{-1/5}$. The weights ($p$) are found using the SQP routine SQPSolve in the programming language GAUSS 8.0. While our problem is not a QP problem, this type of solver uses

a modified quadratic program to find the step length for moving in the direction of a minimum.

The simulation results for Eq. (27) are given in Figs. 1 and 2 for $n = 100$ and 500, respectively. Each of the curves corresponds to the 95th percentile of the distance metric for each sample size.[18] The solid line in panel (a) of each figure is the corresponding unconstrained local constant least-squares estimator and the dashed line is the constrained local constant least-squares estimator. We note that in each case the constrained estimator deviates from the unconstrained estimator where the second derivative is positive. This difference is shown by positive values for the distance metric. Specifically, in Figs. 1 and 2 the values of the distance metric are 0.111 and 0.069, respectively. Note that the distance metric decreases with the sample size. It is easy to see that as the sample size increases the incidence of concavity increases, and the constrained and the unconstrained estimator appear to be more similar. Recall that the distance metric reaches its minimum of 0 when each weight is set equal to $1/n$, or, in other words, the estimated function is de facto concave. This is related to the general trend of increasing observance of concavity as the sample size grows.

In panel (b) of each figure is the corresponding set of weights. The unconstrained estimator sets each of the weights equal to $1/n$. It is obvious that the unconstrained estimators show regions where the second derivative is positive. Our constrained estimator corrects for these nonconcavities by changing the probability weights. Where the weights are larger than $1/n$, these points are given a greater influence in the construction of the estimate, and where the weights are less than $1/n$ these observations are given a lesser influence in the construction of the estimate.

4.2. Empirical Application

The seminal work of Jacob Mincer on human capital suggested that the logarithm of a worker's earnings is concave in her age (potential work experience). Concavity is consistent with the investment behavior implied by the optimal distribution of human capital investment over a worker's life cycle. A voluminous literature within labor economics has generally specified age-earnings profiles as quadratic, consistent with concavity. Murphy and Welch (1990) challenged the conventional empirical strategy of specifying a quadratic in age for an age-earnings profile. Their work suggests that a quadratic specification in age understates early career earnings growth by 30–50% and overstates midcareer earnings growth by

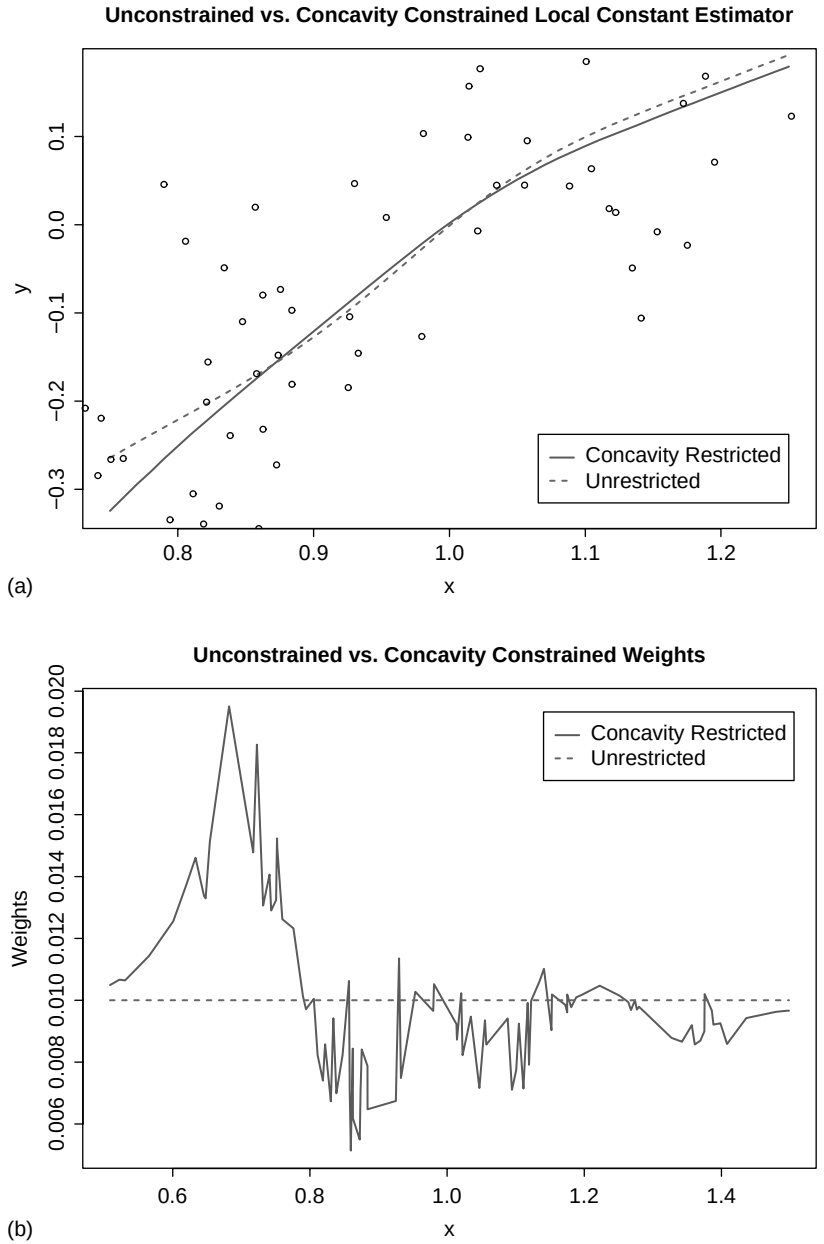


Fig. 1. Simulation for $n = 100$ Corresponding to 95th Percentile of $D_{1/2}(p)$ for 999 Simulations.

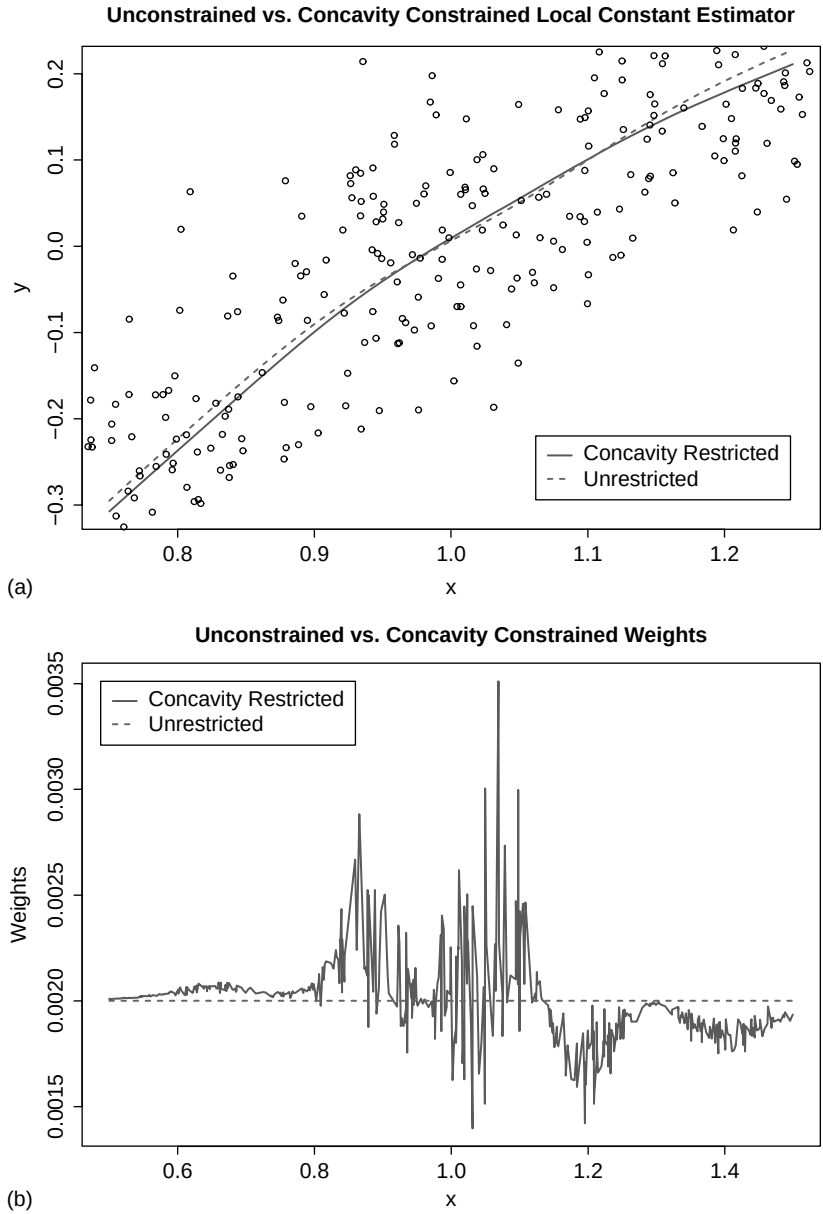


Fig. 2. Simulation for $n = 500$ Corresponding to 95th Percentile of $D_{1/2}(p)$ for 999 Simulations.

20–50%. An analysis of residual plots from their estimated quadratic relationships (as well as several statistical tests) reveals patterns suggesting that determinant differences from this specification exist. They advocate on behalf of a quartic age-earnings profile and find that this specification yields a substantial improvement in fit relative to the common quadratic relationship.

Given that the human capital theory of Mincer does not suggest a precise empirical relationship, Pagan and Ullah (1999, Section 3.14.2) considered the use of nonparametric regression techniques to shed light on the appropriate link between income and ages. They provided an example using the 1971 Canadian Census Public Use Tapes consisting of 205 individuals who had 13 years of education. Fitting a local constant kernel regression function (see Pagan & Ullah, 1999, Fig. 3.4) they found a visually substantial difference between the common quadratic specification and their nonparametric estimates. A "dip" in the age-earnings profile around age 40 suggested that the relationship was neither quadratic nor concave. Pagan and Ullah (1999) argue that this "dip" may occur because of generational effects present in the cross-section; specifically, pooling workers who have differing earnings trajectories.

Given the need to conform to theory in applied work, partnered with the findings of Murphy and Welch (1990) and Pagan and Ullah (1999), we fit a concavity-restricted age-earnings profile. This approach will adopt the theoretical restrictions but relax the functional form specifications primarily used in the empirical labor economics literature. Fig. 3(a) plots the unrestricted nonparametric regression estimator of Pagan and Ullah (1999) (using bandwidth $h = \hat{\sigma}_{\text{Age}} n^{-1/5}$), the concave-restricted estimator with identical bandwidth, and the common quadratic specification.[19] The corresponding weights are provided in Fig. 3(b).

We see that the concavity-restricted estimator still has a visually distinct difference from the quadratic specification around age 40 (as does the unrestricted nonparametric estimator), yet the concave-restricted estimator does not have the "dip" found in Pagan and Ullah (1999), consistent with the core interpretation of Mincer's human capital theory. Additionally, the unrestricted estimator appears to have a slight nonconcavity around age 25, further highlighting the need to impose concavity.

To focus on the importance of the bandwidth in examining this relationship, we plot the unrestricted estimator of Pagan and Ullah (1999) using their bandwidth as well as the optimal bandwidth found using least-squares cross-validation along with the corresponding concavity-restricted fits. These plots are provided in Fig. 4(a). The "dip" presented in Pagan and Ullah (1999) now takes on the appearance of a trough. Again,

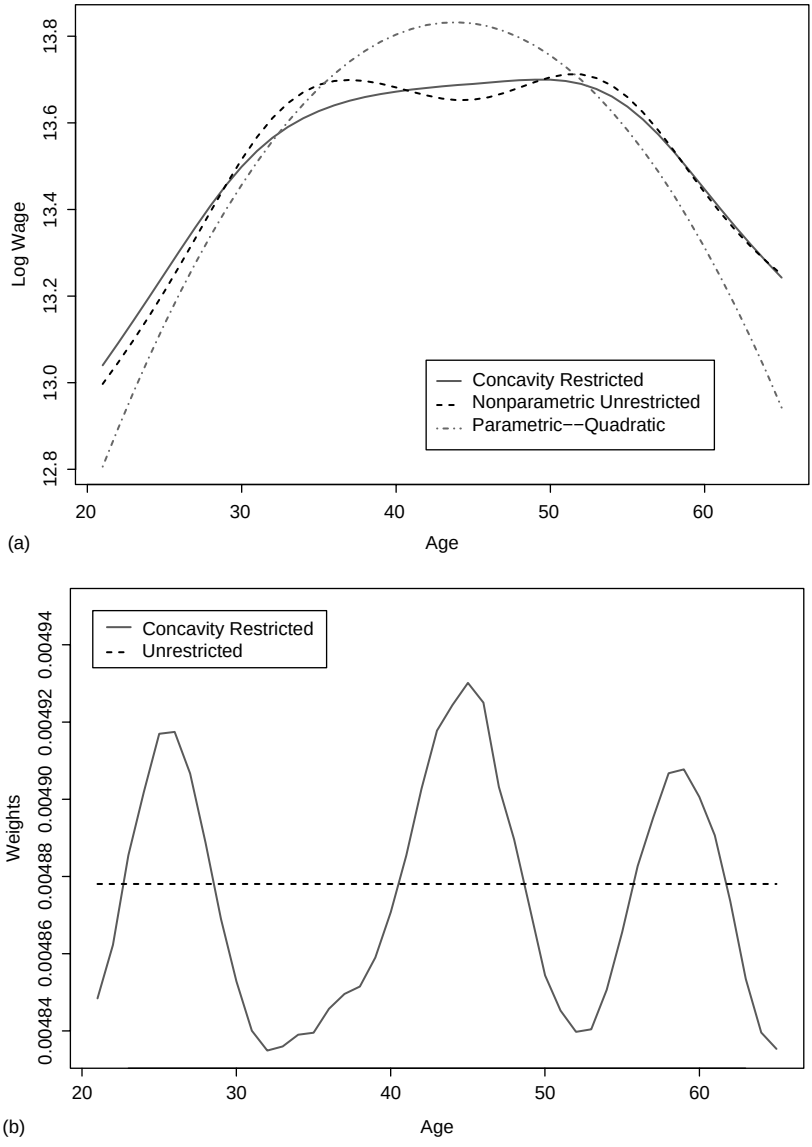


Fig. 3. Unrestricted, Restricted, and Quadratic Fits of the Age-Earnings Profile, CPS 1971 Data.

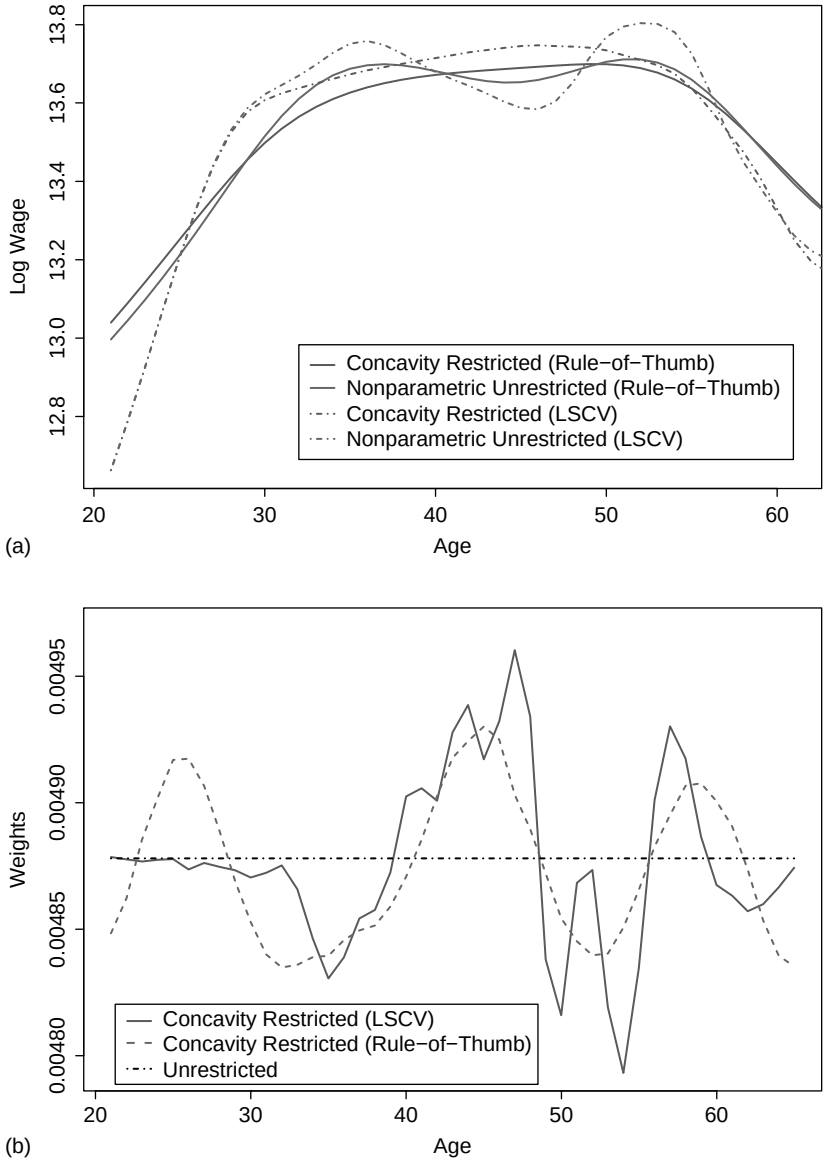


Fig. 4. Unrestricted and Restricted with Differing Bandwidths of the Age-Earnings Profile, CPS 1971 data.

both unconstrained estimators are nonconcave. The estimator using the cross-validated bandwidths produces a distance metric value of 0.005272, almost double of that found using the rule-of-thumb bandwidth. In addition to the nonconcave area around age 40, the cross-validated curve has a region of nonconcavity around age 33, which is more distinct than that for the curve of Pagan and Ullah (1999), which has a slight area of nonconcavity around age 25. The constraint weights, presented in Fig. 4(b), bear this out as well. An interesting feature of this comparison is that the constraint weights for the cross-validated curve appear to be rougher than those for the rule-of-thumb curve, whereas the cross-validated bandwidth is smaller than the rule-of-thumb bandwidth (1.89 vs. 4.22).

While we have not statistically tested for a difference between our concave-restricted nonparametric estimator and the unconstrained estimator, our example shows that we can think more soundly about the implementation of nonparametric estimators in the presence of economic smoothness conditions. We mention again that the ability to impose theoretically consistent smoothness constraints on an economic relationship, paired with the ability to relax restrictive functional form requirements, provides the researcher with a serious set of tools with which to investigate substantive economic questions.

5. CONCLUSION

This chapter has surveyed the existing literature on imposing constraints in nonparametric regression, described an array of methods and discussed computational implementation. This survey included recent research that has not been discussed previously in the literature. We also described a novel method to impose general nonlinear constraints in nonparametric regression that can be implemented using only a standard QP solver. We illustrated this method with a small simulated example focusing on concavity and a detailed example from the empirical labor economics literature. Our empirical results showcased that constrained nonparametric methods can still uncover detail in the data overlooked by rigid parametric models while maintaining theoretical consistency.

Overall future research should determine the relevant merits of each of the methods described here to narrow the set of potential methods down to a few, which can be easily and successfully used in applied nonparametric settings. Given the dearth of detailed simulation studies comparing the available methods highlighted here (notwithstanding Dette & Pilz, 2006), an interesting topic for future research would be to compare the varying

methods (kernel, spline, series) across various constraints to discover under what settings which methods perform the best. Additionally, we feel that our description of the available methods should help further research in extending these ideas to additional nonparametric settings, most notably in the estimation of quantile functions (Li & Racine, 2008), conditional densities, treatment effects (Li, Racine, & Wooldridge, 2008), and structural estimators (Henderson et al., 2009).

NOTES

1. An additional benefit of imposing constraints in a nonparametric framework is that it may provide nonparametric identification; see Matzkin (1994). Also, Mammen, Marron, Turlach, and Wand (2001) show that when one imposes smoothness constraints on derivatives higher than first order the rate of convergence is faster than had the constraints not been imposed.

2. We also looked at the proportion of times a single point on the interior of grid produced a monotonic or concave result. For example, when setting this value of x equal to the expected mean of each series, the incidence of both monotonicity and concavity increased. This percentage increase proved to be much larger for concavity. These results are available from the authors upon request.

3. We should also recognize Yatchew and Bos (1997) who also developed a general framework for constrained nonparametric estimation in a series-based setting. See also the recent application of their method in Yatchew and Härdle (2006).

4. Slower than conventional nonparametric rates.

5. This has connections with both data sharpening (Section 2.5) and constraint weighted bootstrapping (Section 2.6).

6. Monotonicity is not easily imposed in this setting.

7. For a more detailed treatment of either series- or spline-based estimation we refer the reader to Eubank (1988) and Li and Racine (2007, Chapter 15).

8. Unlike kernel smoothing where smoothing is dictated by a bandwidth, in series- and spline-based estimation, the smoothing is controlled by the dimension of the series or spline space.

9. For more on the construction of representor matrices, see Wahba (1990) or Yatchew and Bos (1997, Appendix 2).

10. See the work of Yatchew and Härdle (2006) for an empirical application of constrained nonparametric regression using the series-based method of Yatchew and Bos (1997). Yatchew and Härdle (2006) focus on nonparametric estimation of an option pricing model where the unknown function must satisfy monotonicity and convexity as well as the density of state prices being a true density (positivity and integrates to 1).

11. When $m(x)$ is differentiable at x the gradient of x is the *unique* subgradient of $m(\cdot)$ at x.

12. We use the word bandwidth loosely here as the first stage does not have to involve kernel regression. One could use series estimators in which case the selection

would be over the number of terms. Or, if one uses splines, then the number of knots would have to be selected in the first stage.

13. The notation $A^{(s)}$ refers to the order of the derivative of our weight function with respect to its argument.

14. An interesting topic for future research would be to compare the performance of these methods across a variety of constraints.

15. Quasi-concavity does not imply diminishing marginal productivity to factor inputs. However, under constant returns to scale, quasi-concavity does guarantee diminishing marginal products. This is because quasi-concavity combined with constant returns to scale yields concavity. That being said, a major issue with constant returns to scale is that it implies that both the average and marginal productivities of inputs are independent of the scale of production. In other words, they depend only on the relative proportion of inputs.

16. An alternative would be to solve analytically for all of these derivatives, perhaps with the assistance of a numerical software such as Maxima, Maple, or Mathematica.

17. If one can also assume monotonicity, then to impose concavity all one requires is that the second-order derivatives are negative; thus, only $2n$ constraints need to be imposed which is always fewer constraints than imposing concavity without monotonicity.

18. It should be noted that the number of times that the unconstrained estimator was concave over the grid was very small. Specifically, out of the 999 Monte Carlo simulations for each scenario, the unconstrained estimator was concave 20 and 37 times when $n = 100$ and 500 observations, respectively.

19. Our restricted estimator was calculated using $\rho = 1/2$, and at the optimum we had $D_{1/2}(\hat{p}) = 0.003806$.

ACKNOWLEDGMENTS

The research on this project has benefited from the comments of participants in seminars at Cornell University, the University of California, Merced, the University of California, Riverside, Drexel University, and the State University of New York at Albany, as well as participants at the 5th Annual Advances in Econometrics Conference held at Louisiana State University and the 3rd Annual New York Camp Econometrics. All GAUSS 8.0 code used in this paper is available from the authors upon request.

REFERENCES

Beresteanu, A. (2004). *Nonparametric estimation of regression functions under restrictions on partial derivatives*. Mimeo, Duke University.

Braun, W. J., & Hall, P. (2001). Data sharpening for nonparametric inference subject to constraints. *Journal of Computational and Graphical Statistics*, *10*, 786–806.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics*, *26*, 607–616.

Chernozhukov, V., Fernandez-Val, I., & Galichon, A. (2009). Improving point and interval estimates of monotone functions by rearrangement. *Biometrika*, *96*(3), 559–575.

Choi, E., & Hall, P. (1999). Data sharpening as a prelude to density estimation. *Biometrika*, *86*, 941–947.

Cressie, N. A. C., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, *46*, 440–464.

Delecroix, M., & Thomas-Agnan, C. (2000). Spline and kernel regression under shape restrictions. In: Schimek, M. G. (Ed.), *Smoothing and regression: Approaches, computation, and application*. Wiley series in probability and statistics (Chapter 5, pp. 109–133). Amsterdam: Wiley.

Dette, H., Neumeyer, N., & Pilz, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, *12*(3), 469–490.

Dette, H., & Pilz, K. F. (2006). A comparative study of monotone nonparametric kernel estimates. *Journal of Statistical Computation and Simulation*, *76*(1), 41–56.

Dierckx, P. (1980). An algorithm for cubic spline fitting with convexity constraints. *Computing*, *24*, 349–371.

Dykstra, R. (1983). An algorithm for restricted least squares. *Journal of the American Statistical Association*, *78*, 837–842.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. New York: Dekker.

Friedman, J., Tukey, J. W., & Tukey, P. (1980). Approaches to analysis of data that concentrate near intermediate-dimensional manifolds. In: E. Diday, et al. (Eds), *Data analysis and informatics*. Amsterdam: North-Holland.

Gallant, A. R. (1981). On the bias in flexible functional forms and an essential unbiased form: The Fourier flexible form. *Journal of Econometrics*, *15*, 211–245.

Gallant, A. R. (1982). Unbiased determination of production technologies. *Journal of Econometrics*, *20*, 285–323.

Gallant, A. R., & Golub, G. H. (1984). Imposing curvature restrictions on flexible functional forms. *Journal of Econometrics*, *26*, 295–321.

Goldman, S., & Ruud, P. (1992). *Nonparametric multivariate regression subject to constraint*. Technical report, Department of Economics, University of California, Berkeley, CA.

Hall, P., & Huang, H. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, *29*(3), 624–647.

Hall, P., & Presnell, B. (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society, Series B*, *61*, 143–158.

Hansen, D. L., Pledger, G., & Wright, F. T. (1973). On consistency in monotonic regression. *Annals of Statistics*, *1*(3), 401–421.

Henderson, D. J., List, J. L., Millimet, D. L., Parmeter, C. F., & Price, M. K. (2009). *Imposing monotonicity nonparametrically in first price auctions*. Virginia Tech AAEC Working Paper.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, *49*, 598–619.

Holm, S., & Frisen, M. (1985). *Nonparametric regression with simple curve characteristics*. Technical Report 4, Department of Statistics, University of Goteborg, Goteborg, Sweden.

Kelly, C., & Rice, J. (1990). Monotone smoothing with application to dose response curves and the assessment of synergism. *Biometrics*, *46*, 1071–1085.

Li, Q., & Racine, J. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Li, Q., & Racine, J. S. (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, *26*(4), 423–434.

Li, Q., Racine, J. S., & Wooldridge, J. M. (2008). Estimating average treatment effects with continuous and discrete covariates: The case of Swan–Ganz catherization. *American Economic Review*, *98*(2), 357–362.

Mammen, E. (1991a). Estimating a smooth monotone regression function. *Annals of Statistics*, *19*(2), 724–740.

Mammen, E. (1991b). Nonparametric regression under qualitative smoothness assumptions. *Annals of Statistics*, *19*(2), 741–759.

Mammen, E., Marron, J. S., Turlach, B. A., & Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statistical Science*, *16*(3), 232–248.

Matzkin, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica*, *59*, 1315–1327.

Matzkin, R. L. (1992). Nonparametric and distribution-free estimation of the binary choice and the threshold-crossing models. *Econometrica*, *60*, 239–270.

Matzkin, R. L. (1993). Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics*, *58*, 137–168.

Matzkin, R. L. (1994). Restrictions of economic theory in nonparametric methods. In: D. L. McFadden & R. F. Engle (Eds), *Handbook of econometrics* (Vol. 4). Amsterdam: North-Holland.

Matzkin, R. L. (1999). *Computation of nonparametric concavity restricted estimators. Mimeo*.

Mukerjee, H. (1988). Monotone nonparametric regression. *Annals of Statistics*, *16*, 741–750.

Murphy, K. M., & Welch, F. (1990). Empirical age-earnings profiles. *Journal of Labor Economics*, *8*(2), 202–229.

Nocedal, J., & Wright, S. J. (2000). *Numerical optimization* (2nd ed.). New York, NY: Springer.

Pagan, A., & Ullah, A. (1999). *Nonparametric econometrics*. New York: Cambridge University Press.

Racine, J. S., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130.

Racine, J. S., Parmeter, C. F., & Du, P. (2009). *Constrained nonparametric kernel regression: Estimation and inference*. Virginia Tech AAEC Working Paper.

Ramsay, J. O. (1988). Monotone regression splines in action (with comments). *Statistical Science*, *3*, 425–461.

Robinson, S. M. (1974). Perturbed Kuhn–Tucker points and rates of convergence for a class of nonlinear-programming algorithms. *Mathematical Programming*, *7*, 1–16.

Ruud, P. A. (1995). Restricted least squares subject to monotonicity and concavity restraints. Paper presented at the 7th World Congress of the Econometric Society.

Schumaker, L. (1981). *Spline functions: Basic theory*. New York: Wiley.

Wahba, G. (1990). *Spline models for observational data*. CBMS-NSF Conference Series in Applied Mathematics 59. Philadelphia, PA: SIAM.

Wheelock, D. C., & Wilson, P. W. (2001). New evidence on returns to scale and product mix among U.S. commercial banks. *Journal of Monetary Economics*, *47*(3), 653–674.

Yatchew, A., & Bos, L. (1997). Nonparametric regression and testing in economic models. *Journal of Quantitative Economics*, *13*, 81–131.

Yatchew, A., & Härdle, W. (2006). Nonparametric state price density estimation using constrained least squares and the bootstrap. *Journal of Econometrics*, *133*, 579–599.

FUNCTIONAL FORM OF THE ENVIRONMENTAL KUZNETS CURVE

Hector O. Zapata and Krishna P. Paudel

ABSTRACT

This is a survey paper of the recent literature on the application of semiparametric–econometric advances to testing for functional form of the environmental Kuznets curve (EKC). The EKC postulates that there is an inverted U-shaped relationship between economic growth (typically measured by income) and pollution; that is, as economic growth expands, pollution increases up to a maximum and then starts declining after a threshold level of income. This hypothesized relationship is simple to visualize but has eluded many empirical investigations. A typical application of the EKC uses panel data models, which allows for heterogeneity, serial correlation, heteroskedasticity, data pooling, and smooth coefficients. This vast literature is reviewed in the context of semiparametric model specification tests. Additionally, recent developments in semiparametric econometrics, such as Bayesian methods, generalized time-varying coefficient models, and nonstationary panels are discussed as fruitful areas of future research. The cited literature is fairly complete and should prove useful to applied researchers at large.

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 471–493

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025017

INTRODUCTION

This study provides a survey of the literature on the effect of economic growth on environmental quality using semiparametric and nonparametric techniques. The relationship between economic growth and environmental quality became increasingly important in economic development since the mid-1990s. Grossman and Krueger (1991, 1995) examined the relationship between economic growth and environmental quality during the North American Free Trade Agreement debate of the 1990s. Their major conclusion was that increased development initially led to environmental deterioration, but this deterioration started to decline (turning point) as some level of economic prosperity (income per capita) was obtained. While the location of the turning point varied with the indicator of pollution, the relative reduction in pollution started at income levels of less than $8,000 (in 1985 dollars). Given the similarity between the accepted relationship between income inequality and economic growth (typically referred to as the Kuznets curve, named after Simon Kuznets), this inverted-U relationship (where the level of pollution increased until some level of prosperity is obtained) has been labeled the environmental Kuznets curve (EKC). The first use of the term "environmental Kuznets curve" was by Panayatou (1993), while its first use in academic journals was by Seldon and Song (1994). Current development issues such as alternative sources of energy (biofuels, solar, wind) and global warming re-emphasize the importance of environmental quality in the pursuit of economic development, and thus, inquiries on the validity of the EKC will continue to emerge.

Literature on the subject is voluminous and continues to grow as do the controversial findings. One issue of controversy in the existing literature is the sensitivity of the relationship between economic growth and environmental quality to individual specific factors (de Bruyn, van den Bergh, & Opschoor, 1998). Different countries may experience different stages of development, and the point at which environmental quality begins to improve may vary accordingly. Similarly, some countries may have been slow to monitor environmental degradation, and data may not be available for a long enough period to reveal any significant relationship. From an econometric perspective, the complex problem of finding adequate model specifications for the EKC under the possibility of alternative data generating mechanisms provides a rich setting for the empirical implementation of model specification testing via more flexible nonparametric and semiparametric structures. Some of this nonparametric/semiparametric literature has started to emerge. Millimet, List, and Stengos (2003) and

Paudel, Zapata, and Susanto (2005), for example, provide empirical support for nonlinear effects between pollution and income for some pollutants but not for others, thus giving credence to the use of more flexible semiparametric functional forms of the EKC. Yet, it is difficult to generalize such findings without repeated samples in experimental or simulated data. Fortunately, the most recent econometric advances provide results that are useful for empirical modeling with small samples and under a much richer set of models.

This paper contributes to the literature through the following. First, it summarizes the existing literature on model specification tests with EKC research, and second, it provides a discussion of EKC research questions that can be addressed via recent advances in semiparametric econometric methods. The EKC specification problem has been the subject of extensive research in environmental economics, and since the specification issue is of continued research interest, the literature summarized in this paper may prove beneficial to a large empirical audience.

EKC MODELING BACKGROUND

The EKC literature is founded on the idea that an estimate of the quantity of air and water pollution (p_{it}) at place i at time t can be expressed as (e.g., Grossman & Krueger, 1995):

$$p_{it} = \beta_1 X_{it} + \beta_2 X_{it}^2 + \beta_3 X_{it}^3 + \beta_4 \bar{X}_{it-} + \beta_5 \bar{X}_{it-}^2 + \beta_6 \bar{X}_{it-}^3 + \beta_u Z'_{it} + \varepsilon_{it} \quad (1)$$

where X_{it} is the gross domestic product (GDP) per capita in the country where station i is located, $\bar{X}_{it-}$ is the average GDP per capita for the previous three years, Z_{it} is a vector of other covariates, and ε_{it} is an error term. This parametric specification is sufficiently flexible to allow for the hypothesized inverted-U formulation, but it also places several important restrictions on the estimated relationship. Intuitively, the inverted-U shape results because environmental quality is a luxury good. In the initial stages of development, each individual in the society is unwilling to pay the direct cost of reducing emissions (i.e., the marginal utility of income based on other goods is higher than the marginal utility of environmental quality). However, as income grows the marginal utility of income based on other goods falls as the marginal utility of environmental quality increases. Hence, the linear specification presented in Eq. (1) provides for a reduced form expression of these changes.

The most general specification of the EKC that appears in the literature is the two-way fixed effects panel data model:

$$p_{it} = \alpha_i + \phi_t + X_{it}\delta + Z_{it}\gamma + u_{it} \tag{2}$$

where p_{it} is concentration of a pollutant (e.g., SO_2 or NO_x) in state or county i in time t; α_i are specific state/country fixed effects that control location-specific factors that affect emission rates; ϕ_t are time effects such as the common effect of environmental or other policies; X_{it} is CPI-adjusted per capita income in state/county i in time t and is a vector containing polynomial effects up to order three on per capita income (i.e., $X_{it} = (x_{it}\ x_{it}^2\ x_{it}^3)$); δ is the associated vector of slope coefficients, and the Z_{it} includes other variables such as population density, lagged income, and dummy variables; and u_{it} is a contemporaneous error term. A variation of Eq. (2) is one where the polynomial income effect is replaced with a spline function of income based on a number of preselected knots K (e.g., Millimet et al., 2003; Schmalensee, Toker, & Judson, 1998). As articulated in List and Gallet (1999), Eq. (2) is a reduced form model that does not lend itself to the inclusion of endogenous characteristics of income or to causality inferences; its specification is general enough to allow for individual-specific effects (heterogeneous α and δ), thus avoiding heterogeneity bias; lastly, state-specific time trends can capture a number of implied effects related to technology, population changes, regulations, and pollution measurement.

The hypothesis of an inverted-U relationship between economic growth and environmental quality is by definition nonlinear in income. Implicitly, this nonlinearity can be approximated with a Taylor series expansion based on a low-order polynomial in income, but one question is whether these parametric restrictions adequately represent the nonlinearity of the EKC relationship. An alternative is to model the nonlinear effects using a nonparametric component on income while permitting fixed and time effects to enter through the error term in the following model:

$$p_{it} = \alpha_i + \phi_t + g(X_{it}) + f(\cdot) + u_{it} \tag{3}$$

where all previous definitions hold and $f(\cdot)$ represents other variables such as population density and other social and country characteristics; a nonparametric structure for income is indicated by $g(\cdot)$, which replaces the polynomial component of X_{it} in Eq. (2); and u_{it} is an error component, which can take different structures. The specification of error components can depend solely on the cross section to which the observation belongs or on both the cross section and time series. If the specification depends on the

cross section, then we have $u_{it} = v_i + \varepsilon_{it}$, and if the specification is assumed to be dependent on both cross section and time series, then the error components are modeled as $u_{it} = v_i + e_t + \varepsilon_{it}$. Here ε_{it} is assumed to be a classical error term with zero mean and a homoskedastic covariance matrix; v_i represents heterogeneity across individuals (region/country/state); and e_t represents the heterogeneity over time. The nature of the error structure leads to different estimation procedures, and this is also true in the parametric specification of Eq. (1).

SEMIPARAMETRIC ESTIMATION OF THE EKC

The interest of the present survey is to identify econometric advances in the estimation of the EKC that fall mainly into the subject of semiparametric modeling (a special issue of *Ecological Economics* (1998) provides a complete account of previous parametric EKC studies). The literature summary provided in Table 1 relates to EKC research that has employed semiparametric methods, and includes authors, journal, year of publication, type of model and specification tests as well as turning point findings.[1]

Table 1, column 1, makes it clear that the interest in the application of semiparametric methods to EKC research is recent and rising. Millimet et al. (2003) advance that the appropriateness of a parametric specification of the EKC should be based on the formulation of an alternative hypothesis of a semiparametric partial linear regression (PLR) model.[2] This idea is pursued using the same panel data as in List and Gallet (1999), and estimations are reported for sulfur dioxide and nitrogen oxide for the entire sample (1929–1994) and for a partial sample (1985–1994). Model specification tests of Zheng (1996) and Li and Wang (1998) were used to test parametric (Eq. (2)) and semiparametric (Eq. (3)) models of the EKC.[3] The parametric specification is a two-way fixed effects panel data model. The semiparametric model follows root-N consistent estimates (Robinson, 1988) of the intercepts and time effects in Eq. (3), conditional on the nonlinear income variables; the standard Gaussian density was used in local constant kernel estimation and cross-validation (CV) generated the smoothing parameters. As in List and Gallet (1999), individual-state EKCs were calculated for cubic parametric and semiparametric models. Convincing results were reported in favor of adopting model specification tests of the EKC to decipher whether the implications from parametric models were statistically different from those generated from semiparametric EKCs. The test statistic was labeled J_n, which has an asymptotic normal distribution under H_0.

Table 1. Existing Published Studies That Have Used Semiparametric Techniques in Environmental Kuznets Curve Estimation.

Authors, Journal, and Year of Publication	Types of Models Used	Parametric vs. Nonparametric Tests	EKC-Related Findings, Turning Points (TP)
Millimet et al., *Review of Economics and Statistics*, 2003	*Parametric*: Two-way fixed effects, cubic and spline *Semiparametric*: Partial linear model of Robinson	Zheng (1996) and Li and Wang (1998)	EKC existed for SO_2 and NO_x TP-SO_2: \$16,417 NO_x: \$8,657 LT, 16,417 ST
Van, *Applied Economics Letters*, 2003	*Parametric*: OLS *Semiparametric*: Additive partially linear model by Hastie and Tibshirani (1990)	Gain statistics developed by Hastie and Tibshirani (1990) is used in comparison	No EKC for protected areas
Roy and Van Kooten, *Economics Letters*, 2004	*Parametric*: Linear and cubic models *Semiparametric model*: Robinson (1988), Stock (1989), and Kniesner and Li (2002)	Li and Wang (1998)	EKC did exist for NO_x but not for CO and O_3 TP-NO_x: \$10,193
Bertinelli and Strobl, *Economics Letters*, 2005	*Parametric*: Quadratic fixed effects *Semiparametric*: Robinson (1988)	Ullah (1985)	No EKC existed for CO_2, SO_2
Paudel et al., *Environmental and Resource Economics*, 2005	*Parametric*: Fixed and random effects panel *Semiparametric:* Robinson (1988)	Hong and White (1995)	EKC existed for nitrogen but not for phosphorus and dissolved oxygen TP-N: \$12,993

Azomahou et al., *Journal of Public Economics*, 2006	*Parametric*: Within cubic panel estimation; *Nonparametric*: Wand and Jones (1995), Linton and Nielsen (1995)	Li and Wang (1998)	EKC existed for CO_2 in parametric model, but did not exist in a nonparametric model TP-CO2: $13,358
Van and Azomahou, *Journal of Development Economics*, 2007	*Parametric*: Fixed and random effects panel quadratic model *Semiparametric*: Smooth coefficient model by Li et al. (2002)	Li et al. (2002)	No EKC for deforestation
Criado, *Environmental and Resource Economics*, 2008	*Parametric*: Cubic panel fixed effects *Semiparametric*: Wood (2006) approach	V-test, Yatchew's pooling test (2003)	EKC existed for CH_4, CO, CO_2, NMVOC TP-CH_4: $17,300; CO: $16,800; CO_2: $16,400;NMVOC: $17,200
Luzzati and Orsini, *Energy*, 2009	*Parametric*: Fixed effects panel *Semiparametric*: Generalized additive model	No test done	EKC observed in energy consumption TP for energy consumption: LI, none; MI, $57,500; HI, $18,500; OC, $9,000

Note: TP, turning point; LI, low-income countries; MI, middle-income countries; HI, high-income countries; OC, oil-producing countries. Azomahou et al. found EKC within the model but not in the first difference model. Criado estimated cubic parametric models, but the turning points presented here are only the upper turning points. In Paudel et al., value shown is for the upper turning point.

Because of small sample skewness, bootstrapping of critical values is usually required. Millimet et al. provided results for the PLR and a spline model (for H_a) and the conclusion favors the semiparametric model of the EKC over the parametric one. State-specific EKCs are based on time series data; thus, Li and Stengos' test for first-order serial correlation in a PLR was estimated using a density-weighted version of Eq. (3) (this avoids the random denominator problem associated with nonparametric kernel estimation), and it was adapted to a panel data model (Li and Hsiao, 1998) for a I_n statistic. The results favored the null model of no serial correlation in this data set. A relevant policy finding of this study is that the location of the peak of the EKC is sensitive to modeling assumptions, a finding consistent with the heterogeneity results in List and Gallet (1999).

Van (2003) estimated a semiparametric additive partially linear model for protected areas for 89 countries to examine the EKC hypothesis. Van uses Hastie and Tibshirani's (1990) backfitting algorithm to estimate the semiparametric model. To compare the nonparametric function of a variable with the corresponding parametric function, he uses a "gain" statistic. He found that EKC did not exist for protected areas for the year 1996 for the set of countries included in the analysis.

Roy and van Kooten (2004) used a semiparametric model to examine the EKC for carbon monoxide (CO), ozone (O_3), and nitrogen oxide (NO_x). The estimation technique in this application adjusted the standard PLR to allow for heteroskedasticity (Robinson, 1988) and tested a quadratic parametric model against the semiparametric model using the Li and Wang (1998) test. As opposed to most previous applications, the variables are expressed as the natural log of pollutants and income. Because this is a panel data specification, a generalized local linear estimator (Henderson & Ullah, 2005) is used. Roy and van Kooten started the analysis by first considering linear, quadratic, and cubic models of income for each pollutant and analyzed the statistical significance of income; they found that income was significant in some models but not in others. This led to the specification of the semiparametric model as a more flexible functional form. The main result of this study is that the quadratic model is strongly rejected in favor of the semiparametric model, and similar results are obtained for estimates of the income elasticities.

Bertinelli and Strobl (2005) used Robinson's additive linear regression approach in estimating the relationship between pollution (SO_2 and CO_2 emission) and GDP. They used 1950–1990 observations of 108 and 122 countries for CO_2 and SO_2, respectively. Using the unit-root test in Im, Pesaran, and Shin (2003), they found the data to be stationary.

Using semiparametric regression, they found that the relationship between SO_2 and CO_2 to GDP is linear. The confidence interval is calculated at a 99% level using the approach suggested by Härdle (1990). The linearity of EKC was tested against semiparametric form using the approach suggested in Ullah (1985). The null hypothesis is linear in form where an alternative was a semiparametric form. They used a bootstrap procedure recommended by Lee and Ullah (2001) to obtain *p*-values. They were unable to reject the linearity of the relationship between pollution and GDP.

Nonpoint source water pollutants in Louisiana watersheds were studied in Paudel et al. (2005), and turning points were estimated for nitrogen (N), phosphorus (P), and dissolved oxygen (DO) at the watershed level for 53 parishes for the period 1985–1999 using data collected by the Department of Environmental Quality. Parametric and semiparametric models as in Eqs. (2) and (3) were estimated. The parametric model is similar to Eq. (2) with $f(\cdot)$ being a population density and a weighted income variable to represent spillover effect. One-way and two-way fixed and random effects models were estimated, and a Hausman test was used to evaluate the appropriateness of the model specifications. The best parametric model is set up as the null hypothesis and tested against a semiparametric model, that is,

$$\begin{aligned} H_0 &: p_{jit} = \alpha_i + \phi_t + X_{it}\delta + f(\cdot) + u_{jit} \\ H_a &: p_{jit} = \alpha + g(X_{it}) + f(\cdot) + u_{it} \end{aligned} \tag{4}$$

Paudel et al. used Hong and White's test and found that a semiparametric model better represented the Louisiana pollution–economic growth relationship for phosphorus. They also observed the existence of an EKC form for nitrogen but not for phosphorus and dissolved oxygen.

Deforestation can quickly deteriorate the quality of the environment, and in the process of economic development, most developing countries must confront local (loss in biodiversity) and global (carbon sequestration) dimensions of such environmental degradation. Van and Azomahou (2007) investigated nonlinearities and heterogeneity in the deforestation process with parametric and semiparametric EKCs, and their focus is on whether the EKC exists, and they identify the determinants of deforestation. The data set was a panel of 59 developing countries over the period 1972–1994. The EKC is first estimated as a quadratic parametric model with deforestation rate as the dependent variable and GDP per capita along with other variables as independent variables. *F*-tests of fixed time and country effects supported a fixed country effects model. A Hausman test supported the existence of a random effects model relative to a fixed effects specification;

however, the overall specification was insignificant. In order to check the robustness of the functional form between deforestation rate and GDP per capita, a semiparametric fixed effects model was estimated (as in Paudel et al.). The salient finding was the nonexistence of an EKC for the deforestation process. The analysis was extended to investigate whether other variables (e.g., population growth rate, trade ratio ((imports+ exports)/GDP), population density, the literacy rate, and political institutions) may be more relevant in the determination of deforestation, and a model similar to Eq. (2) was estimated. Contrary to the previous case, the data supported a fixed effects model and many of the new variables were significant, while a within estimator was preferred to a first difference estimator. A semiparametric model similar to Eq. (3) was specified, with GDP assumed to enter nonlinearly in the nonparametric function $g(\cdot)$. The method of Robinson (1988) was used to estimate a first difference representation of Eq. (3), but the results did not support the existence of an EKC. It was hypothesized that perhaps modeling bias could be reduced by specifying a smooth coefficient model (e.g., Li, Huang, Li, & Fu, 2002) that captures the influence of GDP on deforestation rates depending upon the state of development of each country. The smooth coefficient model can be written as:

$$p_{it} = \alpha(x_{it}) + z\beta(x_{it}) + u_{it} \tag{5}$$

where $\beta(x_{it})$ is a smooth function of x_{it}. Note that when $\beta(x_{it}) = \beta$ the model reduces to a standard PLR (similar to Eq. (3)). Having a nonparametric effect ($x_{it} = \text{GDP}$) on the deforestation rate and varying coefficients on other determinants of deforestation (z_{it}) allows the assumption that GDP per capita can have a direct effect and a nonneutral effect, respectively, on the deforestation rate. The model specification test (H_0 vs. H_a) in Li et al. (denoted as J_n) follows a standard normal distribution under H_a. One finding from smooth coefficients for the growth rate of GDP per capita was that for developing countries at a higher stage of economic development, the growth rate of GDP per capita accelerates the deforestation process and deteriorates environmental quality. The results from a J_n test supported the parametric over the semiparametric model at the 5% significance level; in fact, Eq. (2), with a quadratic polynomial in GDP, was preferred to all other models. Heterogeneity due to the economic development process, however, could not be ascertained with these data, and the authors suggested that further work is needed on this research question.

The question of whether a fixed effects panel data model (pooling) is appropriate has received limited attention in the EKC literature. Criado (2008) argues that in most applications, no formal tests of the homogeneity assumption is conducted for time (stability of the cross-sectional regressions over time) and space (stability of the cross-sectional regressions over individual units). Existing literature on the subject has generated mixed results. Criado tests poolability in the EKC by examining the adequacy of such an assumption on both dimensions via nonparametric tests robust to functional misspecification using models similar to those in Eqs. (2) and (3). The data set is a balanced panel of 48 Spanish provinces over the 1990–2002 period, and the pollutants include methane, carbon monoxide, carbon dioxide, nitrous oxide, ammonia, nonmethanic volatile organic compounds, and nitrogen and sulfur oxides. Poolability tests on the spatial dimension (spatial heterogeneity) reject it, particularly for nonparametric specifications. Time poolability (temporal homogeneity) results were mixed; it holds for three of four air pollutants in Spanish provinces, and the estimated pooled nonparametric functions reflected inverted U shapes. It was also pointed out that the parametric and nonparametric tests overwhelmingly rejected the null hypothesis of spatial homogeneity and fixed effects, and that failure to recognize this property of EKC panel data would lead to mixed findings. The work suggested that future EKC research should use advances in parametric and nonparametric quantile regression, random coefficient modeling, and panel heterogeneity. In similar research, Azomahou, Laisney, and Van (2006) use the local linear kernel regression to estimate $W(x_{it})$ with $x_{it} = (x_{it}\ x_{i,t-1})$. They claim that the local linear (polynomial of order 1) kernel estimator performs better than the local constant (polynomial of order 0) kernel estimator or Nadaraya–Watson estimator, since it is less affected by the bias resulting from data asymmetry, notably at the boundaries of the sample. They use standard univariate Gaussian kernel and marginal integration to estimate the nonparametric model. To select the bandwidth in the nonparametric regression, they used a least squares CV method. To develop the confidence interval of the estimated function they used a wild bootstrap method, and to test for the suitability of nonparametric versus parametric functional form, they used the specification test developed by Li and Wang (1998).

Luzzati and Orsini (2009) investigated the relationship between absolute energy consumption and GDP per capita for 113 countries over the period 1971–2004. They estimated both parametric fixed and random effects models and a semiparametric model. For the semiparametric model estimation, they used the approach presented by Wood (2006). Luzzati and

Orsini did not perform specification tests of parametric versus nonparametric functional form. However, they found that EKC existed for energy consumption for middle- and high-oil-producing countries.

The debate about the existence of an EKC in the empirical literature is likely to continue, and this presents an opportunity for the application of recent advances in semiparametric modeling and consistent specification testing that adds flexibility not only to model structures, but also that provides inference results for various dependent data structures with small samples. The summary of applications of the EKC presented in Table 1 is clearly a lagging indicator of the theoretical literature. Advances in econometrics are arriving at such a fast pace that a bridge is needed to connect the theory with the practice; this appears applicable to a number of applied fields. One example is the use of consistent specification tests in fixed effects panel data models that allow for continuous and discrete regressors (e.g., Racine & Li, 2004; Hsiao, Li, & Racine, 2007; Henderson, Carroll, & Li, 2008). A brief summary of recent advances on consistent specification tests that we feel would be relevant to applied researchers interested in EKC-related questions is provided in the next section. The section starts with a seminal paper by Li (1999), which provides a general framework for kernel-based tests (KBTs) for time series econometric models. Li and Racine (2006, Chapter 12) provide a rigorous theory of recent developments in a manner useful for applied researchers; they also provide proofs to many of the theorems related to these tests. For completeness, the most relevant literature is cited[4] in this paper, and the discussion of selected papers relevant to EKC research should be considered a complement to Li and Racine (2006, Chapter 12). It should be noted that emphasis is placed on the use of the "wild bootstrap," initially suggested in Härdle and Mammen (1993), because the existing literature on KBTs convincingly points to its use in the calculation of critical values in small samples, which are characteristic in EKC applications.

CONSISTENT SPECIFICATION TESTS

Consistent model specification tests, which are generalizations of those in Fan and Li (1996) and Lavergne and Vuong (1996), in the context of time series data, were introduced by Li (1999). Li develops the asymptotic normality theory of the proposed test statistics under similar regularity conditions as in the case of independent data, thus resolving previous conjectures about the validity of the tests. Using kernel methods to estimate

unknown functions, Li allows for the null model to be nonparametric or semiparametric, with the inclusion of a parametric model as one possible null model. At the cost of oversimplification, and consistent with previous work on the solution of the random denominator problem, the asymptotic results in Li can be summarized as follows. First, the asymptotic distribution of the test statistics, referred to as $nh^{d/2}J_n$, is normal with mean zero and variance σ_0^2, and a feasible test statistic is defined by an estimate of J_n (Li, Eq. (2), p. 105). Further, Li proves that under the null hypothesis of a nonparametric regression model (and under some regularity conditions), the statistic T_n^a converges to a standard normal distribution given a consistent estimator of the variance (Li, Theorem 3.1, p. 108). Because even in the independent data case this statistic has small sample bias, Li develops a new test, denoted V_n^a, with possible smaller finite sample bias and shows that it can be standardized to a $N(0,1)$ distribution (Li, p. 109). The above results for a nonparametric significance test were also applied to derive the asymptotic distribution for testing a partially linear model (H_0^b), with results equivalent to those above and leading to a standard normal distribution labeled T_n^b (Li, Corollary 4.2, p. 113). Li develops a Monte Carlo simulation and obtains the following general findings. The finite sample versions of the test statistic had much smaller estimated sizes than their feasible asymptotic counterpart (the $\hat{J}_n$ test). The $\hat{J}_n$ tests were less powerful than their finite sample versions, with power being sensitive to the relative smoothing parameter choices. It was found that smoothing should be carefully examined in light of data frequency: for low frequency data a relatively large smoothing parameter leads to a high power test; the opposite being true for high frequency data. Optimal methods for choosing such smoothing parameters were not explored. It was also suggested that parametric and nonparametric bootstrap methods with time series data to approximate the null distributions should be investigated. The finding that for low frequency data a relatively large smoothing coefficient improves power (at no size cost) is clearly relevant to studies of the EKC hypotheses that are based on annual data, whose frequency tends to be low. The combination of independent and weak dependent data in the context of the EKC would also suggest that application of the finite sample versions of null parametric model test statistics should be applicable in EKC testing problems. Clearly, the use of wild bootstrap methods should be a mandatory practice.

The literature on consistent model specification tests that appeared until the late 1990s used either nonparametric regression estimators (KBTs) or Bierens's (1982) tests (Integrated Conditional Moment, ICM tests). Fan and Li (2000) established the relationship between KBT and ICM tests and

provided results indicating that certain consistent KBTs with a fixed smoothing parameter (Härdle & Mammen's, 1993 T_n test and Li & Wang's, 1998 and Zheng's, 1996 I_n test) can be regarded as ICM tests of Bierens (1982) and Bierens and Ploberger (1997) with specific weight functions.[5] In the context of "singular" local alternatives, KBT can detect such alternatives converging in probability to the null model at a rate faster than $n^{-1/2}$. For the first time, it is shown that KBTs are a complement to ICM tests: KBT have higher power for high-frequency alternatives whereas ICM tests have higher power for low-frequency alternatives (Pitman type).[6] The relevance of these asymptotic results in finite samples was illustrated via a Monte Carlo experiment that compares the I_n KBT and the ICM tests under a variety of data generating processes and 5,000 replications. As in Li and Wang (1998), Fan and Li used the wild-bootstrap procedure to approximate the asymptotic null distributions of the test statistics with 1,000 replications and 1,000 wild bootstrap statistics for each replication. The Monte Carlo findings were in agreement with the theoretical results on local power properties of the KBT versus the ICM tests. The estimated sizes, based on the wild bootstrap for all the tests considered, were very close to the nominal sizes, suggesting a good approximation of the null distribution of the test statistics. Given the relative simplicity of the KBT, a feature appealing to applied researchers, applications of the wild-bootstrap tests (I_n type tests) in EKC work is warranted for two reasons. First, EKC models are estimated via panel data that combine independent and dependent data. Fan and Li (2000) are the first to extend such work in the context of weak dependent data. Second, EKC models are typically estimated with annual data of relatively short length, which would suggest that bootstrapping methods are recommended to obtain critical values for specification tests.

By the time the Fan and Li (2000) paper appeared, the literature on consistent specification tests using KBTs was growing fast, and there was a need to condense the available literature in a manner useful to practitioners. Lee and Ullah (2003) provide a comprehensive Monte Carlo study to analyze the size and power properties of two KBTs for neglected nonlinearity in time series models using bootstrap methods. The first test is a bootstrap version (Cai, Fan, & Yao, 2000) that compares the expected values of the squared errors under the null and alternative hypotheses (Ullah, 1985), referred to as a T-test, and the second test is a nonparametric conditional moment goodness of fit test (Li & Wang, 1998; Zheng, 1996), referred to as an L-test. Similar to other works, Lee and Ullah make use of existing asymptotic normality results to examine the bootstrap performance of the tests. One of the main conclusions is that the wild bootstrap L-test

worked well with conditionally heteroskedastic data; these tests had good size and power properties in the simulated DGPs. It was also found that the power of both tests was considerably influenced by the choice of nonlinear models and that no test (T or L) was uniformly superior. Since the L-test has an asymptotic normal distribution under the null, the bootstrap L-tests were found to be more accurate than their asymptotic counterpart, a finding consistent with previous work. It was stated, as in Li (1999) and others, that the choice of the bandwidth in the L-test is more important for time series processes than for independent processes. However, the effect of optimal bandwidth choice on the performance of the tests was not evaluated (see also the specification testing section in Ullah & Roy, 1998; Baltagi, 1995; Baltagi, Hidalgo, & Li, 1996). To our knowledge, specification of nonparametric EKC models in the presence of conditional heteroskedasticity has not been considered. Therefore, the bootstrap results of Lee and Ullah should serve as a useful guide in future work.

Discrete variables are often used in EKC regressions to capture a variety of effects that contribute to industrial activity and that lead to economic growth. These types of regressors are important and have been alluded to in the literature as ways of capturing scale-, composition-, and technique-related variables in EKC models (Grossman & Krueger, 1991; Copeland & Taylor, 2004; Kukla-Gryza, 2009). Examples include variables such as openness to trade (yes = 1, no = 0), democracy, and freedom ($X = 1$ for a democratic country and $X = 0$ otherwise). Another instance of need for dummy regressors relates to pooling countries of different income levels in an EKC model. The concern, such as whether it is accurate to have the same model for all the countries in one EKC, is raised by List and Gallet (1999). Criado (2008) proposed a nonparametric poolability test, but if a dummy for each country group can be included in the EKC regression, this concern can potentially be eliminated. In the situations described above, some recent econometric advances can be adopted.

Racine and Li (2004) propose an estimator for nonparametric regressions that admits continuous and discrete variables and which also allows for the discrete variables to have a natural order or not. Hsiao et al. (2007) expands Racine and Li's model by introducing nonparametric kernel-based consistent model specification tests.[7] Through smoothing both the continuous and discrete variables, and using least squares CV methods, they arrive at an asymptotically normal distribution under the null. Their approach has significant practical appeal because it avoids the "running out of observations" problem related to frequency-based nonparametric estimators that require sample splitting and associated efficiency losses.

It also provides new results on the use of CV methods in model specification testing and demonstrates its superior performance. Small samples are commonly used in the estimation of the EKC, a fact also related to modeling of economic time series with annual data. Hsiao et al. suggest using bootstrap methods as viable alternatives for approximating the finite-sample null distribution of the CV-based test statistics ($\hat{J}_n$), a statistic that is similar to that of Zheng and Li and Wang, given that the simulations showed a poor finite-sample performance of the asymptotic normal approximation of the CV test. They also illustrate the usefulness of the proposed test in testing for the correct specification for wage equations, an application whose specification issues parallel that of the EKC, and advocate that it may be useful in practice to consider the use of interaction terms that may better capture variation of a continuous dependent variable when the number of continuous regressors is insufficient. The usefulness of this new development cannot be overemphasized given that it is often the case that discrete variables are needed to capture a variety of indirect effects in EKC analyses.

DISCUSSION

This survey article emphasizes recent developments in semiparametric econometric methods and their application to the study of the pollution–economic growth tradeoff, commonly referred to as the EKC. The papers reviewed included the standard heterogeneous panel data model, which is the typical general structure used to represent the null model in semiparametric model specification evaluations. Variations of this parametric structure include the standard PLR of Robinson (1988) and extensions thereof, including a PLR with heterogeneity, serial correlation, and heteroskedasticity, poolability, and smooth coefficients.

Various advances in econometrics are absent in the EKC literature reviewed above. The functional form of Bayesian models, for example, provides a vehicle to introduce prior information around diffuse, independent, priors on the parametric component of the EKC and partially informative priors on the nonparametric function (e.g, Koop & Poirier, 2004; Huang & Lin, 2007). Another natural extension of future EKC research relates to the estimation of semiparametric models that contain continuous and discrete regressors. The nonparametric CV technique introduced by Hsiao et al. (2007) is applicable to the case where the EKC contains dummy variables; one appealing point of this estimator is that its

superior performance carries over to model specification tests (see also Racine & Li, 2004; Li & Racine, 2006).

Research using standard EKC parametric panel data models typically start by applying a Hausman test for fixed versus random effects. Subsequently, the best parametric structure is set up as the null model and a semiparametric model as the alternative (as in Eqs. (2) and (3), respectively). This model specification has been of recent interest in the econometrics literature. Henderson et al. (2008) introduce an iterative nonparametric kernel estimator for panel data with fixed effects that naturally carries into the typical panel data specification of the EKC. One of the specifications in Henderson et al. sets up the null hypothesis to be a parametric fixed effects model and the alternative a semiparametric model. The proposed test statistics converge to 0 under the null and to a positive constant under the alternative, and thus, it is argued that the proposed test can be used to detect the validity of the null model. The asymptotic normality results are left to future work, but it is conjectured that even if it were provided, the existing literature suggest that asymptotic theory does not provide a good approximation for nonparametric KBTs in finite samples, as summarized in the previous literature cited here. Henderson et al.'s approach would be a natural application of specification tests of the EKC in a way that is consistent with previous inquiries on random versus fixed effects, and on determining whether a semiparametric model is a more adequate specification. In light of the work by Racine and Li (2004) and Hsiao et al. (2007), an extension of this research to continuous and discrete panels is pending in the literature.[8]

The PLR smooth coefficient model of Li et al. (and other recent applications such as Henderson & Ullah, 2005; Lin & Carroll, 2006; Henderson et al., 2008) has been revisited by Sun and Carroll (2008), with the random effects and fixed effects as the null and alternative hypotheses, respectively (note that that in Li et al. the null hypothesis is a parametric smooth coefficient model whereas the alternative is a semiparametric smooth coefficient model). They propose an estimator that is consistent when there is an additive intercept term (case in which the conventional first difference model fails to generate a consistent estimator). They show the inconsistency of random effects estimators if the true model is one with fixed effects, and that fixed effects estimators are consistent under both random and fixed effects panel data models. It is concluded that estimation of a random effects model is appropriate only when the individual effect is independent of the regressors. They also introduce J_n-type statistics for the above hypotheses that, under asymptotic normality of the proposed

estimator, converges to a standard normal distribution. The test is one sided and rejects the random effects model for large values at some significance level. Sun and Carroll provide Monte Carlo evidence that supports the satisfactory finite sample performance of the estimator and test statistic and suggest bootstrapping critical values for future research. Given that the question of random effects often plays out in EKC applications (and is often rejected), the estimator and statistic introduced in Sun and Carroll should shed brighter light on heterogeneity properties of EKC panels with semiparametric varying coefficient models.

One of the most promising econometric advances, and an area that is still emerging, is the estimation of nonstationary semiparametric panel data models. There is considerable empirical evidence on the existence of unit roots in per capita pollutants and income variables (e.g., Romero-Avila, 2008; Liu, Skjerpen, Swensen, & Telle, 2006). This evidence points to the adequacy of vector autoregression and error correction models (ECM) for some nonstationary panels, and mixed results for others. The failure of many of these previous studies in finding an inverted U-shaped EKC in nonstationary panel data consistent with the data generation process led Romero-Avila to design a study that jointly controlled for structural breaks and cross-sectional dependence; the main finding was one of mixed unit roots for the emissions and income relationship of the EKC, putting to question findings that support ECM in world or specific country groups. Perhaps the most challenging case to model is that of mixed unit roots in panels and the ensuing interpretation of estimated parameters. Extensions of existing work (e.g., Ullah & Roy, 1998; Baltagi & Kao, 2000) to semiparametric nonstationary panels should enhance the empirical understanding of the tradeoff between pollution and growth in environmental economics and the practice of semiparametric econometrics in general.

NOTES

1. We thank an anonymous reviewer for suggesting this summary table.
2. As pointed out by an anonymous reviewer, consistency of estimates of a semiparametric model depends on the correct specification of the parametric component and no interaction among the variables of the semiparametric components.
3. A rigorous presentation of model specification tests in nonparametric regressions is found in Li and Racine (2006, Chapter 12).
4. To save space, the following is the list of papers on estimation and specification testing in parametric and nonparametric modeling that are related to this survey: Ramsey (1974), the pioneer paper by Hausman (1978), Breausch and Pagan (1980),

Davidson and MacKinnon (1981), White (1982), Bera, Jarque, and Lee (1984), Newey (1985), Tauchen (1985), Godfrey (1988), Ullah (1988), Robinson (1989), Bierens (1990), Scott (1992), Bera and Yoon (1993), Whang and Andrews (1993), Delgado and Stengos (1994), Li (1994), Härdle, Mammen, and Muller (1998), Silverman (1998), Härdle, Muller, Sperlich, and Werwatz (2004), Horowitz and Lee (2002), Li et al. (2002), Li and Stengos (1995, 1996, 2003), Li, Hsiao, and Zinn (2003). A century of history of parametric hypothesis testing, the reading of which motivated a larger initial version of this paper, can be found in Bera (2000).

5. The tests based on T_n and I_n are strictly asymptotically locally unbiased, that is, the conditional bias of the kernel regression estimator under H_0 has been removed.

6. The construction of consistent tests based on the estimation of unconditional moments results in what is referred to as nonsmoothing tests. As pointed out by an anonymous reviewer, this is a growing literature that may deserve further analysis, particularly in light of the simulation findings in Fan and Li (2000). An excellent up-to-date reading on this subject is Li and Racine (2006, Chapter 13) and the references therein.

7. Other useful references on consistent model specification tests are Yatchew (2003), Pagan and Ullah (1999), Ait-Sahalia, Bickel, and Stoker (2001), and references in Hsiao et al. (2007).

8. Recent advances in nonparametric econometrics have been implemented using the R package (Racine, 2008).

ACKNOWLEDGMENTS

Although the research in this article has been funded in part by the USDA Cooperative State Research, Education, Extension Service (Hatch Project LAB93787), it has not been subjected to USDA review and therefore does not necessarily reflect the views of the agency, and no official endorsement should be inferred. We want to thank two anonymous reviewers for their useful comments and suggestions, and also Elizabeth A. Dufour for her editorial suggestions.

REFERENCES

Ait-Sahalia, Y., Bickel, P. J., & Stoker, T. M. (2001). Goodness-of-fit tests for kernel regression with an application to option implies volatilities. *Journal of Econometrics*, *105*(2), 363–412.

Azomahou, T., Laisney, F., & Van, P. N. (2006). Economic development and CO_2 emissions: A nonparametric panel approach. *Journal of Public Economics*, *90*(6–7), 1347–1363.

Baltagi, B. H. (1995). *Econometric analysis of panel data.* New York: Wiley.

Baltagi, B. H., Hidalgo, J., & Li, Q. (1996). Nonparametric test for poolability using panel data. *Journal of Econometrics*, *75*(2), 345–367.

Baltagi, B. H., & Kao, C. (2000). *Nonstationary panels, cointegration in panels and dynamic panels: A survey*. Working Paper No. 16, Center for Policy Research, Maxwell School of Citizenship and Public Affairs, Syracuse, NY.

Bera, A. K. (2000). Hypothesis testing in the 20th century with a special reference to testing with misspecified models. In: C. R. Rao & G. J. Szekcly (Eds), *Statistics for the 21st century: Methodologies for applications of the future*. New York: Marcel Dekka.

Bera, A. K., Jarque, C. M., & Lee, L. F. (1984). Testing the normality assumption in limited dependent variable models. *International Economic Review*, *25*(3), 563–578.

Bera, A. K., & Yoon, M. J. (1993). Specification testing with locally misspecified alternatives. *Econometric Theory*, *9*(4), 649–658.

Bertinelli, L., & Strobl, E. (2005). The environmental Kuznets curve semi-parametrically revisited. *Economics Letters*, *88*(3), 350–357.

Bierens, H., & Ploberger, W. (1997). Asymptotic theory of integrated conditional moments. *Econometrica*, *65*(5), 1129–1151.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, *20*(1), 105–134.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica*, *58*(6), 1443–1458.

Breausch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies*, *47*(1), 239–253.

Cai, Z., Fan, J., & Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, *95*(452), 941–956.

Copeland, B. R., & Taylor, M. S. (2004). Trade, growth and the environment. *Journal of Economic Literature*, *42*(1), 7–71.

Criado, C. O. (2008). Temporal and spatial homogeneity in air pollutants panel EKC estimations: Two nonparametric tests applied to Spanish provinces. *Environmental Resource Economics*, *40*(2), 265–283.

Davidson, R., & MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, *49*(3), 781–793.

de Bruyn, S. M., van den Bergh, J. C. J. M., & Opschoor, J. B. (1998). Economic growth and emissions: Reconsidering the empirical basis of the environmental Kuznets curves. *Ecological Economics*, *25*(2), 161–175.

Delgado, M. A., & Stengos, T. (1994). Semiparametric specification testing of non-nested econometric models. *Review of Economic Studies*, *61*(2), 291–303.

Fan, Y., & Li, Q. (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica*, *64*(4), 865–890.

Fan, Y., & Li, Q. (2000). Consistent model specification tests: kernel-based test versus Bierens' ICM tests. *Econometric Theory*, *16*(6), 1016–1041.

Godfrey, L. G. (1988). *Misspecification tests in econometrics, the Lagrange multiplier principle and other approaches*. Cambridge: Cambridge University Press.

Grossman, G. M., & Krueger, A. B. (1991). *Environmental impacts of a North American free trade agreement*. NBER Working Paper No. 3914.

Grossman, G. M., & Krueger, A. B. (1995). Economic growth and the environment. *Quarterly Journal of Economics*, *110*(2), 353–377.

Härdle, W. (1990). *Applied nonparametric regression*. New York: Cambridge University Press.

Härdle, W., & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, *21*(4), 1926–1947.

Härdle, W., Mammen, E., & Muller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, *93*(444), 1461–1474.

Härdle, W., Muller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models, Springer Series in Statistics*. New York: Springer.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, *46*(6), 1251–1271.

Henderson, D. J., Carroll, R. J., & Li, Q. (2008). Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, *144*(1), 257–275.

Henderson, D. J., & Ullah, A. (2005). A nonparametric random effects estimator. *Economics Letters*, *88*(3), 403–407.

Hong, Y., & White, H. (1995). Consistent specification testing via nonparametric series regression. *Econometrica*, *63*(5), 1133–1159.

Horowitz, J. L., & Lee, S. (2002). Semiparametric methods in applied econometrics: Do the models fit the data?. *Statistical Modelling*, *2*(1), 3–22.

Hsiao, C., Li, Q., & Racine, J. S. (2007). A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics*, *140*(2), 802–826.

Huang, H.-C., & Lin, S.-C. (2007). Semiparametric Bayesian inference of the Kuznets hypothesis. *Journal of Development Economics*, *83*(2), 491–505.

Im, K. S., Pesaran, H., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, *115*(1), 53–74.

Kniesner, T., & Li, Q. (2002). Nonlinearity in dynamic adjustment: Semiparametric estimation of panel labor supply. *Empirical Economics*, *27*(1), 131–148.

Koop, G., & Poirier, D. (2004). Bayesian variants of some classical semiparametric regression techniques. *Journal of Econometrics*, *123*(2), 259–282.

Kukla-Gryza, A. (2009). Economic growth, international trade and air pollution: A decomposition analysis. *Ecological Economics*, *68*(5), 1329–1339.

Lavergne, P., & Vuong, Q. (1996). Nonparametric selection of regressors: The nested case. *Econometrica*, *64*, 207–219.

Lee, T., & Ullah, A. (2001). Nonparametric bootstrap tests for neglected nonlinearity in time series regression models. *Journal of Nonparametric Statistics*, *13*(1), 425–451.

Lee, T.-H., & Ullah, A. (2003). Nonparametric bootstrap specification testing in econometric models. In: D. E. A. Giles (Ed.), *Computer-aided econometrics* (pp. 451–477). New York: Marcel Dekker.

Li, D., & Stengos, T. (2003). Testing serial correlation in semiparametric time series models. *Journal of Time Series Analysis*, *24*(3), 311–335.

Li, Q. (1994). *Some simple consistent tests for a parametric regression function versus semiparametric or nonparametric alternatives*. Unpublished manuscript. Department of Economics, University of Guelph.

Li, Q. (1999). Consistent model specification tests for time series econometric models. *Journal of Econometrics*, *92*(1), 101–147.

Li, Q., & Hsiao, C. (1998). Testing serial correlation in semiparametric panel data models. *Journal of Econometrics*, *87*(2), 207–237.

Li, Q., Hsiao, C., & Zinn, J. (2003). Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics*, *112*(2), 295–325.

Li, Q., Huang, C. J., Li, D., & Fu, T. T. (2002). Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics*, *20*(3), 412–422.

Li, Q., & Racine, J. S. (2006). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Li, Q., & Stengos, T. (1995). A semi-parametric non-nested test in a dynamic panel data model. *Economics Letters*, *49*(1), 1–6.

Li, Q., & Stengos, T. (1996). Semiparametric estimation of partially linear panel data models. *Journal of Econometrics*, *71*(1–2), 289–397.

Li, Q., & Wang, S. (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics*, *87*(1), 145–165.

Lin, X., & Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B*, *68*(1), 69–88.

Linton, O., & Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, *82*(1), 93–100.

List, J. A., & Gallet, C. A. (1999). The environmental Kuznets curve: Does one size fit all? *Ecological Economics*, *31*(3), 409–423.

Liu, G., Skjerpen, T., Swensen, A. R., & Telle, K. (2006). *Unit roots, polynomial transformations, and the environmental Kuznets curve*. Discussion Paper No. 443, Statistics Norway, Research Department, Oslo, Norway.

Luzzati, T., & Orsini, M. (2009). Investigating the energy-environmental Kuznets curve. *Energy*, *34*(3), 291–300.

Millimet, D. L., List, J. A., & Stengos, T. (2003). The environmental Kuznets curve: Real progress or misspecified models. *Review of Economics and Statistics*, *85*(4), 1038–1047.

Newey, W. K. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica*, *53*(5), 1047–1070.

Pagan, A., & Ullah, A. (1999). *Nonparametric econometrics*. New York: Cambridge University Press.

Panayatou, T. (1993). *Empirical tests and policy analysis of environmental degradation at different stages of economic development*. Working paper WP238, Technology and Employment Programme, International Labor Office, Geneva.

Paudel, K. P., Zapata, H., & Susanto, D. (2005). An empirical test of environmental Kuznets curve for water pollution. *Environmental and Resource Economics*, *31*(3), 325–348.

Racine, J. (2008). Nonparametric econometrics using R. Paper presented at the 7th Annual Advances in Econometrics Conference: Nonparametric Econometric Methods, Louisiana State University, Baton Rouge, November 14–16.

Racine, J., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130.

Ramsey, J. B. (1974). Classical model selection through specification error tests. In: P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 13–47). New York: Academic Press.

Robinson, P. M. (1988). Root-*n*-consistent semiparametric regression. *Econometrica*, *56*(4), 931–934.

Robinson, P. M. (1989). Hypothesis testing in semiparametric and nonparametric models for economic time series. *Review of Economic Studies*, *56*(4), 511–534.

Romero-Avila, D. (2008). Questioning the empirical basis of the environmental Kuznets curve for CO_2: New evidence from a panel stationary test robust to multiple breaks and cross-dependence. *Ecological Economics*, *64*(3), 559–574.

Roy, N., & van Kooten, G. C. (2004). Another look at the income elasticity of non point source air pollutants: A semiparametric approach. *Economics Letters*, *85*(1), 17–22.

Schmalensee, R., Toker, T. M., & Judson, R. A. (1998). World carbon dioxide emissions: 1950–2050. *Review of Economics and Statistics*, *80*(1), 15–27.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley.

Seldon, T., & Song, D. (1994). Environmental quality and development: Is there a Kuznets curve for air pollution emissions?. *Journal of Environmental Economics and Management*, *27*(2), 147–162.

Silverman, B. W. (1998). *Density estimation for statistics and data analysis*. New York: Chapman and Hall/CRC.

Stock, J. H. (1989). Nonparametric policy analysis. *Journal of the American Statistical Association*, *84*(406), 567–575.

Sun, Y., & Carroll, R. J. (2008). Semiparametric estimation of fixed effects panel data varying coefficient models. Paper presented at the 7th Annual Advances in Econometrics Conference: Nonparametric Econometric Methods, Louisiana State University, Baton Rouge, November 14–16.

Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, *30*(1–2), 415–443.

Ullah, A. (1985). Specification analysis of econometric models. *Journal of Quantitative Economics*, *1*, 187–209.

Ullah, A. (1988). Nonparametric estimation and hypothesis testing in econometric models. *Empirical Economics*, *13*(3–4), 223–249.

Ullah, A., & Roy, N. (1998). Nonparametric and semiparametric econometrics of panel data. In: A. Ullah & D. E. A. Giles (Eds), *Handbook of applied economic statistics, A* (pp. 579–604). New York: Marcel Dekker.

Van, P. N. (2003). Semiparametric analysis of determinants of a protected area. *Applied Economics Letters*, *10*(10), 661–665.

Van, P. N., & Azomahou, T. (2007). Nonlinearities and heterogeneity in environmental quality: An empirical analysis of deforestation. *Journal of Development Economics*, *84*(1), 291–309.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall.

Whang, Y., & Andrews, D. (1993). Tests of specification for parametric and semiparametric models. *Journal of Econometrics*, *57*(1–3), 277–318.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

Wood, S. (2006). *Generalized additive models: An introduction with R: Texts in statistical sciences* (Vol. 67). New York: Chapman and Hall.

Yatchew, A. (2003). *Semiparametric regression for the applied econometrician*. New York: Cambridge University Press.

Zheng, J. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, *75*(2), 263–289.

SOME RECENT DEVELOPMENTS ON NONPARAMETRIC ECONOMETRICS

Zongwu Cai, Jingping Gu and Qi Li

In this paper, we survey some recent developments of nonparametric econometrics in the following areas: (i) nonparametric estimation of regression models with mixed discrete and continuous data; (ii) nonparametric models with nonstationary data; (iii) nonparametric models with instrumental variables; and (iv) nonparametric estimation of conditional quantile functions. In each of the above areas, we also point out some open research problems.

1. INTRODUCTION

There is a growing literature in nonparametric econometrics in the recent two decades. Given the space limitation, it is impossible to survey all the important recent developments in nonparametric econometrics. Therefore, we choose to limit our focus on the following areas. In Section 2, we review the recent developments of nonparametric estimation and testing of regression functions with mixed discrete and continuous covariates. We discuss nonparametric estimation and testing of econometric models for nonstationary data in Section 3. Section 4 is devoted to surveying the literature of nonparametric instrumental variable (IV) models. We review

Nonparametric Econometric Methods
Advances in Econometrics, Volume 25, 495–549

ISSN: 0731-9053/doi:10.1108/S0731-9053(2009)0000025018

nonparametric estimation of quantile regression models in Section 5. In Sections 2–5, we also point out some open research problems, which might be useful for graduate students to review the important research papers in this field and to search for their own research interests, particularly dissertation topics for doctoral students. Finally, in Section 6 we highlight some important research areas that are not covered in this paper due to space limitation. We plan to write a separate survey paper to discuss some of the omitted topics.

2. MODELS WITH DISCRETE AND CONTINUOUS COVARIATES

In this section, we mainly focus on analysis of nonparametric regression models with discrete and continuous data. We first discuss estimation of a nonparametric regression model with mixed discrete and continuous regressors, and then we focus on a consistent test for parametric regression functional forms against nonparametric alternatives.

2.1. Nonparametric Regression Models with Discrete and Continuous Covariates

We are interested in estimating the following nonparametric regression model:

$$Y_i = g(X_i) + u_i, \qquad (i = 1, \ldots, n) \tag{1}$$

where $X_i = (X_i^{\mathrm{c}}, X_i^{\mathrm{d}}), X_i^{\mathrm{c}} \in \Re^q$ is a continuous random variable of dimension q ($q \geq 1$), and X_i^{d} is a discrete random variable of dimension r ($r \geq 0$). We will only consider independent and identically distributed data case in Section 2. Let X_{is}^{d} denote the sth component of X_i^{d}. We consider two possibilities: X_{is}^{d} can be an ordered and unordered discrete variable. If X_{is}^{d} is unordered, $X_{is}^{\mathrm{d}} \in \mathcal{D}_s = \{a_1, a_2, \ldots, a_{c_s}\}$ with c_s taking distinct different values and $c_s \in \mathcal{N}$, where $\mathcal{N}$ denotes the set of positive integers. Here we allow for the possibility that $c_s = \infty$. If $c_s = \infty$, we need to add a condition that $\inf_{x_s^{\mathrm{d}} \neq x_{s'}^{\mathrm{d}}, x_s^{\mathrm{d}}, x_{s'}^{\mathrm{d}} \in \mathcal{D}_s} |x_s^{\mathrm{d}} - x_{s'}^{\mathrm{d}}| \geq \delta > 0$ so that x_s^{d} can take at most countably infinitely many different values, and there is only finite many distinct points of x_s^{d} in any bounded interval.

The conventional approach dealing with the discrete variable is to split the sample into many parts sorted by different discrete cells. Then one uses the data falling into a given discrete cell to estimate the conditional mean

function of Y given the remaining continuous variables. However, this sample splitting method may give unreliable estimation results or even become infeasible when the number of discrete cells is not small compared with the sample size. In a seminal paper, Aitchison and Aitken (1976) proposed a novel method of smoothing discrete variables in estimating a discrete probability function. Hall, Racine, and Li (2004), Racine and Li (2004), and Hall, Li, and Racine (2007) generalized Aitchison and Aitken's smoothing method to the problem of estimating a conditional density function or a conditional mean function. Their proposed smoothing method avoids the sample splitting problem and therefore remains a feasible estimation method when the number of discrete cells is comparable or even larger than the sample size. An additional advantage of smoothing the discrete variables is that, as shown by Hall et al. (2004, 2007), irrelevant covariates can be automatically smoothed out (i.e., removed) from a conditional density or a regression model.

We now introduce the kernel smoothing function for discrete variables. The kernel function associated with unordered discrete variable X_{is}^{d} is given by $l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, \lambda_s) = 1$ if $X_{is}^{\mathrm{d}} = x_s^{\mathrm{d}}$; and $l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, \lambda_s) = \lambda_s$ if $X_{is}^{\mathrm{d}} \neq x_s^{\mathrm{d}}$, where λ_s is the smoothing parameter. If X_{is}^{d} is an ordered discrete variable, we use the following kernel function: $l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, \lambda_s) = \lambda_s^{|X_{is}^{\mathrm{d}} - x_s^{\mathrm{d}}|}$. Whether x_s^{d} is either ordered or unordered, when $\lambda_s = 0$, the kernel function becomes an indicator function, that is, $l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, 0) = \mathbf{1}(X_{is}^{\mathrm{d}} = x_s^{\mathrm{d}})$, where $I(A)$ denotes an indicator function that takes value one if event A holds true, and zero otherwise. Also, when $\lambda_s = 1$, $l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, 1) \equiv 1$ is a constant function. The range of λ_s is [0,1] for all $s = 1, \ldots, r$. The product kernel for the discrete variables X^{d} is $L(X_i^{\mathrm{d}}, x^{\mathrm{d}}, \lambda) = \prod_{s=1}^{r} l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, \lambda_s)$. For the continuous variable $X^{\mathrm{c}} = (X_1^{\mathrm{c}}, \ldots, X_q^{\mathrm{c}})$, we use the product kernel given by $W_h(x^{\mathrm{c}}, X_i^{\mathrm{c}}) = \prod_{s=1}^{q} h_s^{-1} w((x_s^{\mathrm{c}} - X_{is}^{\mathrm{c}})/h_s)$, where $w(\cdot)$ is a symmetric and univariate density function, and $0 < h_s < \infty$ is the smoothing parameter for x_s^{c}.

The kernel function for the mixed regressor case $X = (X^{\mathrm{c}}, X^{\mathrm{d}})$ is simply the product of W and L, that is, $K(x, X_i) = W_h(x^{\mathrm{c}}, X_i^{\mathrm{c}}) L(x^{\mathrm{d}}, X_i^{\mathrm{d}}, \lambda)$. Thus we estimate $g(x) = E(Y|X = x)$ by the Nadaraya–Watson (NW) (local constant (LC)) method, defined as,

$$\widehat{g}(x) = \frac{\sum_{i=1}^{n} Y_i K(x, X_i)}{\sum_{i=1}^{n} K(x, X_i)} \tag{2}$$

It is easy to see that if $\lambda_s = 0$ for all $s = 1, \ldots, r$, then the discrete kernel function becomes an indicator function, that is, $L(X_i^{\mathrm{d}}, x^{\mathrm{d}}, 1) = 1(X_i^{\mathrm{d}} = x^{\mathrm{d}})$. $\widehat{g}(x)$defined in (2) reduces to the conventional frequency

estimator of $g(x)$. Also, if $\lambda_s = 1$ for some $s \in \{1, \ldots, r\}$, since $l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, 1) \equiv 1$, in this case $\widehat{g}(x)$ becomes unrelated to x_s^{d}, that is, the covariate x_s^{d} is completely removed from the regression model. Similarly, for the continuous variable x_s^{c}, if h_s is sufficiently large, x_s^{c} is effectively removed from the regression model, see Hall et al. (2007) on a more detailed discussion on removing irrelevant covariates by oversmoothing these variables.

It is well known that the smoothing parameters play an essential role in the trade-off between reducing bias and variance, so that their choice in a nonparametric approach is very critical. For the aforementioned setting, Hall et al. (2007) suggested choosing the smoothing parameters $(h, \lambda) = (h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q)$ by minimizing the following cross-validation (CV) function:

$$\mathrm{CV}(h, \lambda) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{g}_{-i}(X_i))^2 w_1(X_i) \tag{3}$$

where $\hat{g}_{-i}(X_i) = \sum_{j \neq i}^{n} Y_j K(X_i, X_j) / \sum_{j \neq i}^{n} K(X_i, X_j)$ is the leave-one-out kernel estimator of $g(X_i) \equiv E(Y_i | X_i)$, and $0 \leq w_1(\cdot) \leq 1$ is a weight function (which has a compact support) that serves to avoid difficulties caused by dividing by zero, or by the slower convergence rate arising when X_i lies near the boundary of the support of X. Although it is necessary to introduce the weight function $w_1(\cdot)$ from the theoretical point of view, in practice the use of the weight function may not be necessary. In applications, since the data range is always finite, one usually does not need to use any weight function, or equivalently one can use $w_1(X_i) \equiv 1$ for all $i = 1, \ldots, n$.

Now suppose that X_s^{d}, the sth component of X^{d}, is an irrelevant component, that is, $E(Y_i | X_i = x) = E(Y_i | X_i / X_{is}^{\mathrm{d}} = x / x_s^{\mathrm{d}})$ almost everywhere, where $X_i / X_{is}^{\mathrm{d}}$ denote the set of variables in X_i with X_{is}^{d} being removed. Let λ_s denote the smoothing parameter associated with irrelevant component X_s^{d}. Hall et al. (2007) showed that, when X_s^{d} is an irrelevant regressor, the cross-validated λ_s converges to 1 in probability. Recall that when $\lambda_s = 1$, the corresponding variable X_s^{d} is completely removed from the nonparametric kernel estimator $\widehat{g}(x)$. This means that all irrelevant discrete variables can be automatically removed (asymptotically) by the least squares CV method. Similar results hold true for the continuous covariates. Indeed, Hall et al. (2007) showed that, when X_s^{c} is an irrelevant covariate, then the cross-validated smoothing parameter h_s diverges to $+\infty$. In such a case, the corresponding kernel function $w((X_{is}^{\mathrm{c}} - x_s^{\mathrm{c}})/h_s) \to w(0)$ becomes a constant. Moreover, this constant is cancelled out from $\widehat{g}(x)$ because the same constant appears at both the numerator and the denominator of $\widehat{g}(x)$.

Hence, asymptotically all irrelevant covariates, either continuous or discrete, is smoothed out from the regression model by the CV method.

The nonparametric estimator $\widehat{g}(x)$ with the cross-validated smoothing parameters has the same asymptotic distribution of a kernel estimator of $g(x)$ that first removes the irrelevant covariates. Hall et al. (2007) defined the irrelevant variables as those regressors that are independent with both the dependent variable and the relevant regressors. However, the simulation results suggest that the CV method can still remove irrelevant variables as long as those irrelevant variables are independent with the dependent variable conditional on the relevant variables. However, it is still of theoretical interest if one can also relax the independent assumption to conditional independent assumption, and this remains an interesting open question.

Note that the above result was extended by Li and Racine (2009) to the case of estimating a varying-coefficient model and by Li, Ouyang, and Racine (2009) and Su, Chen, and Ullah (2009) to weakly dependent data case.

When all the covariates are discrete, the asymptotic analysis is quite different and cannot be obtained from the regression model with mixed discrete and continuous regressors as a special case (since the above result assumes that $q \geq 1$, where q is the number of continuous regressors). When all the regressors are discrete variables, irrelevant discrete covariates are smoothed out by the least squares CV method with a positive probability, say δ. Indeed, Ouyang, Li, and Racine (2009) concluded that $0.5 < \delta < 1$. More precisely, the simulation results reported in their paper suggest that $\delta \in [0.6,\ 0.65]$. In summary, when all the regressors are discrete, one can still remove the irrelevant regressors (by the CV method) with a positive probability, but this probability is strictly less than one, even as the sample size goes to $+\infty$.

Finally, various programs for implementing the CV method to estimate a regression model with mixed discrete and continuous covariates are available. For example, a R-package (np) is currently available at http://www.R-project.org for a free download and a Stata program will be available soon.

2.2. Consistent Model Specification Tests

It is well known that the selection of smoothing parameter is of crucial importance in nonparametric estimation. It is probably less well known (say, to applied econometricians) what important roles the smoothing parameters play in nonparametric model specification testing. In this subsection, we first consider a simple univariate regression model to

illustrate how the selection of smoothing parameter affects the performance of a nonparametric test. Toward this end, we consider the following nonparametric regression model

$$Y_i = g(X_i) + u_i$$

where X_i is a univariate continuous random variable and $g(\cdot)$ is a smooth function. We are interested in testing the null hypothesis $H_0 : E(Y_i|X_i) = \beta_0 + X_i\beta_1$ almost surely (a.s.). One can construct a test based on $I = E[u_i E(u_i|X_i)f(X_i)]$, where $u_i = Y_i - \beta_0 - X_i\beta_1$ and $f(\cdot)$ is the density function of X_i. This is because $I = E[(E(u_i|X_i))^2 f(X_i)] \geq 0$, and it equals to 0 if and only if the null hypothesis is true. Hence, I serves as a proper candidate for testing H_0. A feasible test statistic based on I is given by:

$$I_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i \widehat{E}_{-i}(u_i|X_i)\widehat{f}_{-i}(X_i) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \widehat{u}_i\widehat{u}_j K_{h,ij}$$

where $K_{h,ij} = K_h(X_i - X_j)$ and $K_h(v) = h^{-1}K(v/h)$. It can be shown that I_n converges to 0 under H_0 (indeed, $I_n = O_p((nh^{1/2})^{-1})$ under H_0), and that I_n goes to a positive constant if H_0 is false. A standardized test is given by where $T_n = nh^{1/2}I_n/\widehat{\sigma}_0$, where $\widehat{\sigma}_0^2 = 2[n(n-1)h]^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\widehat{u}_i^2\widehat{u}_i^2 K_{h,ij}^2$. One can show that T_n converges to a standard normal random variable under H_0, and it diverges to $+\infty$ at the rate of $nh^{1/2}$ if H_0 does not hold. In practice, some residual-based bootstrap methods (say, the wild bootstrap method) are recommended for a better approximation to the finite-sample null distribution of the test statistic T_n. The conditions on h are the usual ones: $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

Now the question is: How does the selection of h affect the performance of the T_n test? And how should we select h in practice? Given that residual-based bootstrap methods can give quite satisfactory estimated sizes for T_n, a sensible starting point seems to examine the power property of the test. For a given significance level for a test, one would prefer a test with a large power. To examine how h affects the power of the test, we need to know the behavior of $g(x) \equiv E(Y_i|X_i = x)$ when H_0 fails to hold. In this case, $g(x)$ is a nonlinear function of x. Let us consider a specific example. Suppose that $X \in [0,2]$ and $g(x) = \sin(m\pi x)$, where m is a positive constant. Now consider the case that m is small, say $m = 1/4$. Then $g(x)$ changes from $\sin(0) = 0$ to $\sin(\pi/2) = 1$ as x varies from 0 to 2. The function is monotonically increasing (slowly) over the domain of x. For such a slowly changing function (as x varies), intuitively it is not hard to imagine that the optimal smoothing should be relatively large. In contrast, if $m = 2$, then $m\pi x$

changes from 0 to 4π (as x moves from 0 to 2) and the function $\sin(m\pi x)$ completes two full periods, moving up and down several times as x varies in the domain. This function changes more rapidly compared to the case of $m = 1/4$, the optimal smoothing for this fast changing function should be much smaller compared to a slow changing function (the case of $m = 1/4$). We generate X_i's uniformly from [0, 2] and use the least squares CV method to select the smoothing parameters. For a sample size of $n = 100$ and over 1,000 simulations, the median value of $\widehat{h}$ (cross-validated h) is 0.172 for $m = 1/4$, and 0.068 for $m = 2$. If we use an ad hoc rule such as $h = x_{sd}n^{-1/5} = 0.230$ for $n = 100$, where x_{sd} is the sample standard error of $\{X_i\}_{i=1}^n$. We say that the optimal smoothing parameter (in estimation) can be quite different depending on the different shapes of the unknown regression functions.

How is the nonparametric estimation accuracy related to a power of a nonparametric test? In general, more accurate estimation of the unknown function is expected to lead to a better power of a test if the test is based on the difference between the null hypothesized linear model and the true unknown function.[1] For this reason, Hsiao, Li, and Racine (2007) suggested using the least squares CV method to select the smoothing parameters in a nonparametric smoothing test. Hsiao et al. (2007) considered the problem of testing a parametric regression functional form with mixed discrete and continuous covariates. We next describe their testing procedure.

For testing the null hypothesis that a parametric regression model is correctly specified, we state it as,

$$H_0 : P[E(Y_i|X_i) = m(X_i, \beta)] = 1 \text{ for some } \beta \in \mathcal{B} \tag{4}$$

where $m(\cdot, \cdot)$ is a known function with β being a $p \times 1$ vector of unknown parameters and $\mathcal{B}$ is a compact subset in $\Re^p$. The alternative hypothesis is the negation of H_0, that is,

$$H_1 : P[E(Y_i|X_i) = m(X_i, \beta)] < 1 \text{ for all } \beta \in \mathcal{B} \tag{5}$$

Hsiao et al. (2007) considered a test statistic that was independently proposed by Fan and Li (1996) and Zheng (1996).[2]

The test statistic is based on $I = E[u_i E(u_i|X_i)f(X_i)]$ as we discussed earlier. The sample analogue of I is given by:

$$\begin{aligned} I_n &= n^{-1}\sum_{i=1}^{n} \widehat{u}_i \widehat{E}_{-i}(u_i|X_i)\widehat{f}_{-i}(X_i) = n^{-1}\sum_{i=1}^{n} \widehat{u}_i \left\{ n^{-1} \sum_{j=1, j\neq i}^{n} \widehat{u}_j W_{h,ij} L_{\lambda,ij} \right\} \\ &= n^{-2}\sum_{i}\sum_{j\neq i} \widehat{u}_i \widehat{u}_j K_{\gamma,ij} \end{aligned} \tag{6}$$

where $K_{\gamma,ij} = W_{h,ij}L_{\lambda,ij}(\gamma = (h,\lambda)), \widehat{u}_i = Y_i - m(X_i, \widehat{\beta})$ is the residual obtained from estimating the parametric null model, $\widehat{\beta}$ is a $\sqrt{n}$-consistent estimator of β (under H_0), and $\widehat{E}_{-i}(u_i|X_i)\widehat{f}_{-i}(X_i)$ is a leave-one-out kernel estimator of $E(Y_i|X_i)f(X_i)$. In the case where we have only continuous regressors X_i^c and use a nonstochastic value of h_s ($h_s \to 0$ and $nh_1 \ldots h_q \to \infty$), the asymptotic null (normal) distribution of the I_n test was derived independently by Fan and Li (1996) and Zheng (1996).

For the I_n test with the mixed discrete and continuous covariates, Hsiao et al. (2007) advocated the use of CV methods for selecting the smoothing parameter vectors h and λ. We use $\widehat{I}_n$ to denote the test statistic with CV selected smoothing parameters, that is, $\widehat{I}_n$ is defined the same way as I_n given in (6) but with $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ replaced by the CV smoothing parameters $(\widehat{h}_1, \ldots, \widehat{h}_q, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_r)$. The asymptotic distribution of our CV-based test was derived by Hsiao et al. (2007):

$$\widehat{T}_n \equiv \frac{n(\widehat{h}_1 \ldots \widehat{h}^q)^{1/2}\hat{I}_n}{\sqrt{\widehat{\Omega}}} \xrightarrow{d} N(0,1)$$

under H_0, where "$\xrightarrow{d}$" denotes the convergence in distribution and $\widehat{\Omega} = [2(\widehat{h}_1 \ldots \widehat{h}_q)/n^2]\sum_{i=1}^n \sum_{j\neq 1}^n \widehat{u}_i^2 \widehat{u}_j^2 \widehat{W}_{h,ij}^2 \widehat{L}_{\lambda,ij}^2$.

Hsiao et al. (2007) also showed that the $\widehat{T}_n$ test diverges to $+\infty$ if H_0 is false; thus it is a consistent test. Hsiao et al. (2007) recommended the use of a residual-based wild bootstrap method to better approximate the null distribution of $\widehat{T}_n$. Specifically, one generates the wild bootstrap error u_i^* via a two point distribution $u_i^* = [(1-\sqrt{5})/2]\hat{u}_i$ with probability $(1+\sqrt{5})/[2\sqrt{5}]$, and $u_i^* = [(1+\sqrt{5})/2]\hat{u}_i$ with probability $(\sqrt{5}-1)/[2\sqrt{5}]$. Using $\{u_i^*\}_{i=1}^n$, one generates $Y_i^* = m(X_i, \widehat{\beta}) + u_i^*$ for $i = 1, \ldots, n$. $\{X_i, Y_i^*\}_{i=1}^n$ is called the "bootstrap sample," and one uses this bootstrap sample to obtain a nonlinear least squares estimator of β (a least squares estimator if $m(X_i,\beta) = X_i^{\mathrm{T}}\beta$). Let $\widehat{\beta}^*$ denote the resulting estimator. The bootstrap residual is given by $\widehat{u}_i^* = Y_i^* - m(X_i, \widehat{\beta}^*)$. The bootstrap test statistic $\widehat{T}_n^*$ is obtained the same way as $\widehat{T}_n$ with $\widehat{u}_i$ being replaced by $\widehat{u}_i^*$. Note that we use the same CV selected smoothing parameters $\widehat{h}$ and $\widehat{\lambda}$ when computing the bootstrap statistics. That is, there is no need to rerun CV with the bootstrap sample. Therefore, our bootstrap test is computationally quite simple. In practice, one repeats the above steps a large number of times, say $B = 1{,}000$ times, then, the original test statistic $\widehat{T}_n$ plus the B bootstrap test statistics give us the empirical distribution of the bootstrap statistics, which is then used to approximate the finite-sample null distribution of $\widehat{T}_n$.

By adopting the concept of "convergence in distribution in probability" (see e.g., Li, Hsiao, & Zinn, 2003) to study the asymptotic distribution of the

bootstrap statistic $\widehat{T}_n^*$, Hsiao et al. (2007) showed that the wild bootstrap method works by proving the following result:

$$\sup_{z\in\Re}|P(\widehat{T}_n^* \le z|\{X_i, Y_i\}_{i=1}^n) - \Phi(z)| = o_p(1) \tag{7}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. The simulation results reported in Hsiao et al. (2007) show that the proposed bootstrap procedure indeed works well in finite sample applications. See Hsiao et al. (2007) for details on this regard.

2.3. Testing Significance (Relevance) of Discrete Variables

When all the regressors are discrete variables, Ouyang et al. (2009) showed that while the irrelevant variables can be smoothed out with about 65% probability, there is a 35% probability that the cross-validated λ takes values strictly <1 even as $n\to\infty$. Therefore, sometimes the CV method may not be able to determine whether a given variable is irrelevant or not. In such cases, one can use the test statistic proposed by Racine, Hart, and Li (2006) to test whether a given discrete variable is relevant or not. The null hypothesis is,

$$H_0 : m(x,z) = E(Y|X=x, Z=z) = E(Y|X=x) \text{ almost everywhere (a.e.)} \tag{8}$$

where Z is a discrete variable and X can contain both discrete and continuous components. Under the null hypothesis, the discrete variable Z is an irrelevant regressor.

Assume that Z takes c different values, without loss of generality, say that $Z\in\{0,1,\ldots,c-1\}$. The null hypothesis H_0 is equivalent to: $m(X, Z=l) = m(X, Z=0)$ for $l=1,\ldots,c-1$ (for all X). Racine et al. (2006) suggested constructing a test statistic based on

$$I = \sum_{l=1}^{c-1} E\{[m(X,Z=l) - m(X,Z=0)]^2\} \tag{9}$$

Obviously, $I\ge 0$ and $I=0$ if and only if H_0 is true. Therefore, I serves as a proper measure for testing H_0. A feasible test statistic is given by:

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{c-1}[\widehat{m}(X_i, Z_i=l) - \widehat{m}(X_i, Z_i=0)]^2 \tag{10}$$

where $\widehat{m}(X_i, Z_i)$ is the kernel estimator of $m(X_i, Z_i)$.

Racine et al. (2006) recommended using the least squares CV method to select the smoothing parameters. Let $\widehat{\lambda}_z$ denote the smoothing parameter selected by the CV method. Since under H_0, $\widehat{\lambda}_z$ has a nondegenerate (complicated) limiting distribution, the null distribution of $\widehat{I}_n$ is unknown even as $n \to \infty$. Therefore, Racine et al. (2006) recommended using some bootstrap procedures to approximate the null distribution of the $\widehat{I}_n$ test, one of which is described below.

2.3.1. A Bootstrap Procedure

1. Randomly select Z_i^* from $\{Z_j\}_{j=1}^n$ with replacement, and call $\{Y_i, X_i, Z_i^*\}_{i=1}^n$ the bootstrap sample.
2. Use the bootstrap sample to compute the bootstrap statistic $\widehat{I}_n^*$, where $\widehat{I}_n^*$ is the same as $\widehat{I}_n$ except that Z_i is replaced by Z_i^* (using the same cross-validated smoothing parameters of h, $\widehat{\lambda}$, and $\widehat{\lambda}_z$ obtained earlier).
3. Repeat steps 1 and 2, a large number of times, say B times. Let $\{\widehat{I}_{n,j}^*\}_{j=1}^B$ be the ordered (in an ascending order) statistic of the B bootstrap statistics, and let $\widehat{I}_{n,(\alpha)}^*$ denote the $(1-\alpha)$th percentile of $\{\widehat{I}_{n,j}^*\}_{j=1}^B$. We reject H_0 if $\widehat{I}_n > \widehat{I}_{n,(\alpha)}^*$ at the level α.

The simulation results reported in Racine et al. (2006) show that the above bootstrap procedure works well in finite sample applications. See Racine et al. (2006) for details on empirical studies.

3. NONPARAMETRIC REGRESSION MODELS WITH NONSTATIONARY DATA

Phillips and Park (1998) were the first to study the asymptotic theory on nonparametric estimation of econometric models with nonstationary data. Recently, nonparametric estimation of regression functions has attracted many attentions among statisticians and econometricians. Juhl (2005) and Wang and Phillips (2008, 2009) considered nonparametric regression models with nonstationary regressors, while Cai, Li, and Park (2009) and Xiao (2009) considered semiparametric varying-coefficient models with some of the regressors being nonstationary. Gao, King, Lu, and Tjøstheim (2008) and Sun, Cai, and Li (2008a) considered nonparametric testing issues with nonstationary data. Finally, Karlsen, Myklebust, and Tjöstheim (2007) considered nonparametric estimation of a regression model for a more general type of nonstationary processes, a subclass of the class of null recurrent Markov chains. We summarize some of these works below.

3.1. Nonparametric Density and Regression Function Estimation

Phillips and Park (1998) considered a nonparametric autoregressive regression model with the true data generated by an unit root process:

$$Y_t = m(Y_{t-1}) + u_t \equiv Y_{t-1} + u_t$$

where u_t, for expositional simplicity, is assumed to be i.i.d. $(0, \sigma_u^2)$. Phillips and Park (1998) suggested using a LC method to estimate $m(\cdot)$ as,

$$\widehat{m}(x) = \frac{\sum_{t=1}^{n} Y_t K_h(Y_{t-1} - x)}{\sum_{t=1}^{n} K_h(Y_{t-1} - x)} \equiv \frac{(nh)^{-1} \sum_{t=1}^{n} Y_t K_h(Y_{t-1} - x)}{\widehat{f}_n(x)} \tag{11}$$

where $K_h(v) = h^{-1}K(v/h)$, h is the bandwidth, $K(\cdot)$ the kernel function, and $\widehat{f}_n(x) = (nh)^{-1} \sum_{t=1}^{n} K_h(Y_{t-1} - x)$, which would be regarded as an estimator of the density function if Y_t were stationary. Phillips and Park (1998) derived the asymptotic distributions for both $\widehat{m}(x)$ and $\widehat{f}_n(x)$.

It follows from Donsker's theorem that under some regularity conditions, for $0 \le r \le 1$, $Y_{[nr]}/\sqrt{n} \Rightarrow W_u(r)$, where $[\cdot]$ denotes the integer part of $\cdot$, $\Rightarrow$ denotes weak convergence, $W_u(\cdot)$ is a Brownian motion on $[0, 1]$, $\sigma_u^{-1} W_u(r)$ is a standard Brownian motion on [0, 1], and $\sigma_u^2 = E(u_t^2)$. Define the local time $L_W(t, x)$ for a Brownian motion $W(\cdot)$ as,

$$L_W(t, x) = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_0^t \mathbf{1}(|W(s) - x| \le \epsilon) ds \tag{12}$$

Under some regularity conditions including $h \to 0$ and $nh \to \infty$, as $n \to \infty$, Phillips and Park (1998) established the following result:

$$n^{1/4} h^{1/2} (\widehat{m}(x) - m(x)) \xrightarrow{d} \mathrm{MN}\left(0, \frac{\sigma_u^2 v_0(K)}{L_{W_u}(1, 0)}\right) \tag{13}$$

where $\mathrm{MN}(\mu, \Sigma)$ denotes a mixed normal distribution with mean μ and conditional variance Σ, and $v_0(K) = \int K^2(v) dv$. Note that there is no bias term in Eq. (13) because $m(x) = x$ is a linear function so that its derivatives with orders greater or equal to two all vanish.

Wang and Phillips (2009) considered the following nonlinear cointegration model:

$$Y_t = g(X_t) + u_t, \qquad t = 1, 2, \ldots, n$$

where $X_0 = 0$ and $X_t = X_{t-1} + \varepsilon_t$, both u_t and ε_t are mean zero stationary processes. Wang and Phillips (2009) considered the LC estimator for $g(x)$

given by:

$$\widehat{g}(x) = \frac{\sum_{t=1}^{n} Y_t K_h(X_t - x)}{\sum_{t=1}^{n} K_h(X_t - x)}$$

Under some regularity conditions including $nh \to \infty$ and $nh^3 \to 0$ (undersmoothing) as $n \to \infty$, Wang and Phillips (2009) showed that

$$\left(n^{-1/2} \sum_{t=1}^{n} K_h(X_t - x) \right)^{1/2} n^{1/4} h^{1/2} (\widehat{g}(x) - g(x)) \xrightarrow{d} N(0, \sigma_1^2) \qquad (14)$$

where $\sigma_1^2 = \sigma_u^2 \nu_0(K)$. When $X_t = Y_{t-1}$, Eq. (14) gives the asymptotic distribution of $\widehat{m}(x)$ defined in Eq. (11). This is because the asymptotic variances in Eqs. (13) and (14) are the same since it can be shown that $n^{-1} \sum_{t=1}^{n} K_h(X_t - x) \xrightarrow{p} L_W(1,0)/\sigma_\varepsilon$, where $W(\cdot)$ is a standard Brownian motion and $\sigma_\varepsilon^2 = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2} \sum_{t=1}^{n} \varepsilon_t)$ ($\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon_t)$ if ε_t is serially uncorrelated). Finally, Wang and Phillips (2008) extended the result of Wang and Phillips (2009) to allow for endogenous regressors.

3.2. Semiparametric Estimation of a Varying-Coefficient Model with Nonstationary Covariates

Cai et al. (2009) considered the following varying-coefficient model:

$$Y_t = X_t^{\mathrm{T}} \beta(Z_t) + u_t = X_{t1}^{\mathrm{T}} \beta_1(Z_t) + X_{t2}^{\mathrm{T}} \beta_2(Z_t) + u_t, \qquad t = 1, \ldots, n \qquad (15)$$

where A^{T} denotes the transpose of a matrix or vector $\mathbf{A}$, X_{t1}, Z_t, and u_t are stationary, X_{t2} is an $I(1)$ process, $\beta(Z_t) = (\beta_1(Z_t)^{\mathrm{T}}, \beta_2(Z_t)^{\mathrm{T}})^{\mathrm{T}}$, and $X_t = (X_{t1}^{\mathrm{T}}, X_{t2}^{\mathrm{T}})^{\mathrm{T}}$. Here X_{ti} is a $d_i \times 1$ vector, $i = 1, 2$, $d_1 + d_2 = d$, and the first component of X_{t1} is identically one. Also, Y_t, Z_t, and u_t are scalars, and $E(u_t) = 0$, $\sigma_u^2 = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2} \sum_{t=1}^{n} u_t)$ is finite, and u_t is assumed to be independent with (X_t, Z_t).[3] When there is no term $X_{t1}^{\mathrm{T}} \beta_1(Z_t)$, Eq. (15) reduces to the model investigated by Xiao (2009). Note that Y_t can be stationary or nonstationary. If Y_t is nonstationary, model (15) implies that Y_t and X_{t2} are cointegrated with a varying cointegration vector $\beta_2(Z_t)$. The reason why Cai et al. (2009) considered a following varying-coefficient model in Eq. (15) is that it might approximate a general nonparametric model well (see Eq. (36) for details).

It is easy to see that the local linear (LL) estimator for $\beta(z)$ and its derivative function $\beta^{(1)}(z) = d\beta(z)/dz$ is given by:

$$\begin{pmatrix} \widehat{\beta}(z) \\ \widehat{\beta}^{(1)}(z) \end{pmatrix} = \left[\sum_{t=1}^{n} \begin{pmatrix} X_t \\ (Z_t - z)X_t \end{pmatrix}^{\otimes 2} K_h(Z_t - z) \right]^{-1} \times \sum_{t=1}^{n} \begin{pmatrix} X_t \\ (Z_t - z)X_t \end{pmatrix} Y_t K_h(Z_t - z) \tag{16}$$

where $A^{\otimes 2} = AA^{\mathrm{T}}$ and $A^{\otimes 1} = A$.

We assume that X_{t2} can be written as $X_{t2} - X_{t-1,2} = \eta_t$, where η_t is a zero mean stationary process. Then under some standard regularity conditions, $X_{t2}/\sqrt{n} \Rightarrow W_{\eta 2}(r)$, where $W_{\eta 2}(\cdot)$ is a d_2-dimensional Brownian motion on [0, 1]. By the continuous mapping theorem, we know that, for $l = 1, 2$,

$$\frac{1}{n} \sum_{t=1}^{n} \left(\frac{X_{t2}}{\sqrt{n}} \right)^{\otimes l} \xrightarrow{d} \int_0^1 [W_{\eta 2}(r)]^{\otimes l} dr \equiv W_{\eta 2}^{(l)} \tag{17}$$

Let $f_z(z)$ be the marginal density of Z_t. Define $M_k(z) = E[X_{t1}^{\otimes k} | Z_t = z]$ for $1 \le k \le 2$. Further, let

$$S(z) = \begin{pmatrix} M_2(z) & M_1(z) W_{\eta 2}^{(1)^{\mathrm{T}}} \\ W_{\eta 2}^{(1)} M_1(z)^{\mathrm{T}} & W_{\eta 2}^{(2)} \end{pmatrix}$$

and $D_n = \mathrm{diag}\{I_{d_1}, \sqrt{n} I_{d_2}\}$. Then, Cai et al. (2009) showed that under some regularity conditions,

$$\sqrt{nh} D_n \left[\widehat{\beta}(z) - \beta(z) - \frac{1}{2} h^2 \mu_2(K) \beta^{(2)}(z) \right] \xrightarrow{d} MN(0, \underset{\beta}{\Sigma}(z)) \tag{18}$$

where $MN(0, \Sigma_\beta(z))$ is a mixed normal variable with mean zero and conditional covariance $\Sigma_\beta(z) = \sigma_u^2 v_0(K) S(z)^{-1} / f_z(z)$ and $\mu_2(K) = \int v^2 K(v) dv$.

Eq. (18) implies that $\widehat{\beta}_1(z) - \beta_1(z) = O_p(h^2 + (nh)^{-1/2})$ and $\widehat{\beta}_2(z) - \beta_2(z) = O_p(h^2 + (n^2 h)^{-1/2})$. Thus, the convergence rate for $\widehat{\beta}_2(z) - \beta_2(z)$ is faster than that of $\widehat{\beta}_1(z) - \beta_1(z)$. The bias term is $O(h^2)$ for both $\widehat{\beta}_1(z)$ and $\widehat{\beta}_2(z)$, and the variance of $\widehat{\beta}_1(z)$ is $O((nh)^{-1})$, while the variance of $\widehat{\beta}_2(z)$ is $O((n^2 h)^{-1/2})$. This is similar to the linear regression model case because $\sum_{t=1}^{n} X_{2t} X_{t2}^T = O_p(n^2)$ and $\sum_{t=1}^{n} X_{t1} X_{t1}^{\mathrm{T}} = O_p(n)$. The estimated coefficient for the $I(1)$ regressor is n-consistent, while the estimated coefficient for the $I(0)$ regressor has the standard $\sqrt{n}$ rate of convergence.

Cai et al. (2009) also considered the case that X_t is $I(0)$ but Z_t is $I(1)$. For such a case, Z_t can be expressed as $Z_t = Z_{t-1} + v_t = Z_0 + \sum_{s=1}^{t} v_s$, where $\{v_s\}$ is a stationary process with mean zero and $\sigma_v^2 = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2}\sum_{t=1}^{n} v_t) > 0$. Then, it follows from Donsker's theorem that under some regularity conditions, for $0 \le r \le 1$, $Z_{[nr]}/\sqrt{n} \Rightarrow W_v(r)$, where $W_v(\cdot)$ is a Brownian motion on [0, 1] and $\sigma_v^{-1} W_v(r)$ is a standard Brownian motion on [0, 1]. Cai et al. (2009) established the following asymptotic result:

$$\sqrt{n^{1/2}h}[\widehat{\beta}(z) - \beta(z) - h^2 B(z)] \xrightarrow{d} MN(0, \Sigma_1) \tag{19}$$

where $B(z) = \mu_2(K)\beta^{(2)}(z)/2$, $MN(0, \Sigma_1)$ is a mixed normal distribution with mean zero and conditional covariance $\Sigma_1 = \sigma_v \sigma_u^2 v_0(K)[E(X_t X_t^{\mathrm{T}}]L_W(1,0)]^{-1}$. Eq. (19) implies that $\widehat{\beta}(z) - \beta(z) = O_p(h^2 + (n^{1/4}h^{1/2})^{-1})$ so that the optimal smoothing h is proportional to $n^{-1/10}$. Thus, h should converge to 0 at a fairly slow rate at $n^{-1/10}$. This is because when Z_t is $I(1)$, it returns to the fixed interval $[z-h, z+h]$ less often compared to the case when Z_t is $I(0)$. Therefore, one needs to let h go to 0 slowly so as to balance the squared bias and the variance.

When $d = 1$ and $X_t = 1$, the varying-coefficient model reduces to a simple regression model $Y_t = \beta(Z_t) + u_t(Z_t \text{ is } I(1))$. The asymptotic variance in Eq. (19) simplifies to $\sigma_v \sigma_u^2 v_0(K) L_W(1,0)^{-1}$. It can be shown that $\widehat{f}(z) \equiv n^{-1/2}\sum_{t=1}^{n} K_h(Z_t - z)$ consistently estimates $L_W(1,0)/\sigma_v$; see Phillips and Park (1998). Hence, in this case Eq. (19) can be equivalently written as,

$$[\widehat{\sigma}_u^2 v_0(K)]^{-1/2}[\widehat{f}(z)]^{1/2}\sqrt{n^{1/2}h}[\widehat{\beta}(z) - \beta(z) - h^2 B(z)] \xrightarrow{d} N(0,1) \tag{20}$$

where $\widehat{\sigma}_u^2 = n^{-1}\sum_{i=1}^{n}[Y_t - \widehat{\beta}(Z_t)]^2$ is a consistent estimator for σ_u^2. As expected, Eq. (20) is the same as that in Wang and Phillips (2009) for a nonparametric regression model with an $I(1)$ regressor.

Bachmeier, Leelahanon, and Li (2006) considered the following semiparametric dynamic varying-coefficient model:

$$Y_t = \beta_1(Z_t) + Y_{t-1}\beta_2(Z_t) + u_t \tag{21}$$

where Y_t is the rate of inflation, and Z_t is an $I(1)$ variable "velocity of money supply." Bachmeier et al. (2006) applied the above model to forecast U.S. inflation rate and showed that the semiparametric varying-coefficient dynamic model (with a nonstationary covariate) has smaller forecast mean squared error compared with the conventional linear model, or some nonparametric model using only stationary covariates. For more examples in finance, the reader is referred to the paper by Cai and Hong (2009).

Park and Hahn (1999) considered the varying-coefficient model in Eq. (15) with Z_t being replaced by the time trend variable t, and established the asymptotic distribution of a series-based estimator for $\beta(t)$. Park and Hahn (1999) also proposed a test statistics for testing a parametric function form for $\beta(\cdot)$ and for testing cointegration in a time-varying coefficient model framework.

Cai and Wang (2009) considered a similar time-varying coefficient model as the one considered in Park and Hahn (1999) with nonstationary or nearly nonstationary (local to unit root) and endogenous regressors. Cai and Wang (2009) used a LL estimation method and derived the asymptotic distribution of their proposed estimators. Finally, Cai and Wang (2009) applied the above model to test the stability of the predictability of asset returns in finance. That is,

$$r_t = \beta_{0t} + \beta_{1t} x_{t-1} + u_t$$

where r_t is the asset return and x_{t-1} is the first lag of financial instrument, say the logarithm of the earnings-price ratio or the dividend-price ratio or other financial variables. But u_t and x_{t-1} is usually correlated and x_t is nonstationary like $I(1)$ or near $I(1)$ and highly persistent. For details about the theory and applications, we refer the reader to the paper by Cai and Wang (2009).

3.3. Data-Driven Method of Selecting Smoothing Parameter

Sun and Li (2009a) considered the problem of selecting the smoothing parameter h of model (15) by the least squares CV method. They proposed to choosing h by minimizing the following least squares CV objective function:

$$\mathrm{CV}(h) = n^{-1} \sum_{t=1}^{n} [Y_t - X_t^{\mathrm{T}} \widehat{\beta}_{-t}(Z_t)]^2 M(Z_t) \tag{22}$$

where $\widehat{\beta}_{-t}(Z_t)$ is a leave-one-out kernel estimator of $\beta(Z_t)$.

Sun and Li (2009a) first considered the case that X_t is $I(1)$ (there is no $I(0)$ components in X_t), Z_t and u_t are stationary processes. They found an interesting result that the LC and the LL estimation methods lead to very different asymptotic behaviors for $\hat{h}$ by the CV method selected smoothing parameter. Specifically, they showed that for the LC estimation

method (assuming X_t is a scalar to simplify the notation)

$$\sqrt{n}\widehat{h}_{\mathrm{LC-CV}} - \sqrt{\frac{c_{1n}\sigma_u^2 v_0 \int M(z)dz}{v_2 c_{2n} \int (\beta^{(1)}(z))^2 M(z)dz}} \xrightarrow{p} 0 \tag{23}$$

where $c_{1n} = n^{-2}\sum_{t=1}^{n} X_t^2$, $c_{2n} = n^{-3}\sum_t X_t^4$, and $v_j = \int v^j K^2(v)du$. For the LL estimation method the result is,

$$n^{2/5}\widehat{h}_{\mathrm{LL-CV}} - \left(\frac{4\sigma_u^2 v_0 \int M(z)dz}{c_{1n}\mu_2(K)E((\beta_t^{(2)})^2 M_t)}\right)^{1/5} \xrightarrow{p} 0 \tag{24}$$

One interesting implication of Eqs. (23) and (24) is that the CV selected h is stochastic even asymptotically. Also, comparing Eq. (23) with Eq. (24) we see that the CV selected h has different convergence rates. Both these results are in sharp contrast to the stationary data or independent data case where we know that the CV selected smoothing parameter is asymptotically nonstochastic and that the CV functions have the same probability order whether one uses the LC or the LL method. The reason for the different rates of convergence of $\widehat{h}$ is that $\mathrm{CV}_{\mathrm{LC}}(h) = O_p(h + (nh)^{-1})$, while $\mathrm{CV}_{\mathrm{LL}}(h) = O_p(nh^4 + (nh)^{-1})$. This also implies that $\mathrm{CV}_{\mathrm{LC}}(\widehat{h}) = O_p(n^{-1/2})$ and $\mathrm{CV}_{\mathrm{LL}}(\widehat{h}) = O_p(n^{-3/5})$. Hence, the LL method leads to more efficient estimation than the LC method.

Sun and Li (2009a) further provided asymptotic analysis for CV selected h for model (15) with X_t containing both $I(0)$ and $I(1)$ components.

3.4. Testing a Parametric Coefficient Functional Form

Sun et al. (2008a) considered the problem of testing the null hypothesis (H_0) that $P(\beta(Z) = \beta_0) = 1$ for some $\boldsymbol{d} \times 1$ vector of constant coefficient β_0 in the following semiparametric model:

$$Y_t = X^{\mathrm{T}}\beta(Z_t) + u_t = X_{1t}^{\mathrm{T}}\beta_1(Z_t) + X_{2t}^{\mathrm{T}}\beta_2(Z_t) + u_t$$

where X_{1t}, Z_t, and u_t are $I(0)$ variables, and X_{2t} is an $I(1)$ process. They proposed a test statistic based on the sample analogue of $\int ||D(\widehat{\beta}(z) - \widehat{\beta}_0(z))||^2 dz$, where $\widehat{\beta}(z)$ is the semiparametric estimator of $\beta(z)$, $\widehat{\beta}_0$ is the least squares estimator of β_0 and D is a positive definite weight matrix.

The test statistic proposed by Sun et al. (2008a) can be simplified to

$$\widehat{I}_n = \frac{1}{n^3}\sum_{t=1}^{n}\sum_{s\neq t}^{n} X_t^T X_s \widehat{u}_t \widehat{u}_s K_{h,ts} \tag{25}$$

where $\widehat{u}_t$ is the residual obtained from the parametric null model.

Sun et al. (2008a) showed that under some regularity conditions and under H_0,

$$\widehat{J}_n = \frac{n\sqrt{h}\widehat{I}_n}{\sqrt{\widehat{\sigma}_b^2}} \xrightarrow{d} N(0,1)$$

where $\widehat{\sigma}_b^2 = n^{-4}h\sum_{t=1}^{n}\sum_{s\neq t}^{n}\tilde{u}_t^2\tilde{u}_s^2[X_t^T X_s]^2 K_{h,ts}^2$, $\tilde{u}_t = Y_t - X_t^T\widehat{\beta}_{-t}(Z_t)$ is the nonparametric residual and $\widehat{\beta}_{-t}(Z_t)$ is the leave-one-out estimator of $\beta(Z_t)$.

The power of the test statistic J_n depends on whether $\beta_2(z) = \beta_{20}$ or not, where β_{20} is a vector of constant parameters. If $\beta_2(z)\neq\beta_{20}$ for some z in a set with positive measure, Sun et al. (2008a) showed that the $\widehat{J}_n$ test statistic diverges to $+\infty$ at the rate of n^2h. However, when $\beta_2(z) = \beta_{20}$ for all z, and $\beta_1(z)\neq\beta_{10}$ on a set with positive measure, $\widehat{J}_n$ diverges to $+\infty$ at the rate of $n\sqrt{h}$. Intuition behind this result is that, since $X_{2t}X_{2t}^{\mathrm{T}}$ is larger than $X_{1t}X_{1t}^{\mathrm{T}}$ by an order of n, hence, the test statistic diverges to $+\infty$ at a faster rate when $\beta_2(z)$, the coefficient of X_{2t}, is not a constant vector. We summarize the above results on power of the J_n test statistic as follow.

Sun et al. (2008a) showed that under some regularity conditions and H_1, the following two results hold.

(i) If $P[\beta_2(Z_t) = \beta_{20}]<1$ for any $\beta_{20}\in\mathcal{B}_2$, where $\mathcal{B}_2$ is a compact subset of $\mathcal{R}^{d_2}$, then $P[J_n > B_n]\to 1$ as $n\to\infty$ for any nonstochastic sequence $B_n = o(n^2\sqrt{h})$.

(ii) If $P[\beta_2(Z_t) = \beta_{20}] = 1$ for some $\beta_{20}\in\mathcal{B}_2$, and $P[\beta_1(Z_t) = \beta_{10}]<1$ for any $\beta_{10}\in\mathcal{B}_1$, where $\mathcal{B}_1$ is a compact subset of $\mathcal{R}^{d_1}$, then $P[J_n > B_n]\to 1$ as $n\to\infty$ for any nonstochastic sequence $B_n = o(n\sqrt{h})$.

The above results imply that under H_1, the test statistic J_n diverges to $+\infty$ at different rates depending on whether $\beta_2(z) = \beta_{20}$ (a constant vector) or not. Nevertheless, the test statistic J_n is consistent in both cases, and a larger sample size might be required for the power of the test statistic to approach one if $\beta_2(z) = \beta_{20}$, and only the coefficients associated with the $I(0)$ variables are nonconstant $(\beta_1(z)\neq\beta_{10})$.

Also, Sun et al. (2008a) showed that when $\beta_1(z) = \beta_{10}$ (a constant vector) for all z, and $\beta_2(z)\neq\beta_{20}$, then the least squares estimator $\widehat{\beta}_{10}$ diverges to $+\infty$

at the rate of $\sqrt{n}$. Therefore, a misspecified linear model not only leads to inconsistent estimation result but also overestimates the true parameter β_{10} by a different order of magnitude (the true $\beta_{10} = O(1)$ is finite, while $\widehat{\beta}_{10}$ diverges to ∞ at the rate of $\sqrt{n}$). Thus, one drastically overestimates β_{10} in such a case if one estimates a misspecified linear model in which one assumes that the model is linear in both X_{1t} and X_{2t}, while in fact the true model is only linear in stationary covariate X_{1t}, but the coefficient of the nonstationary variable X_{2t} is a smoothing function of the stationary covariate Z_t. This result suggests that it is very important to test if the model specification is correct when there are integrated regressors in the model.

3.5. *Testing Cointegration in Semiparametric Varying-Coefficient Models*

In this subsection, we discuss the problem of testing whether u_t is an $I(1)$ or an $I(0)$ process through a varying-coefficient model:

$$Y_t = X_t^{\mathrm{T}}\beta(Z_t) + u_t$$

where X_t is a $\boldsymbol{d} \times 1$ vector of $I(1)$ variables, Z_t is an $I(0)$ scalar process, and u_t follows an AR(1) process as,

$$u_t = \rho u_{t-1} + \varepsilon_t$$

where ε_t is a mean zero stationary process.

Xiao (2009) set the null hypothesis as H_0^a: u_t is an $I(0)$ process (i.e., $\rho = 0$) and the alternative is H_1^a: u_t is an $I(1)$ process ($\rho = 1$). It is easy to see that under H_0^a, $\mathrm{Var}(u_t) = \sigma_u^2$, a positive constant, while under H_1^a, $\mathrm{Var}(u_t) = a_0 + a_1 t$, where a_0 and a_1 are positive constants. Hence, Xiao (2009) suggested testing H_0^a by testing $a_1 = 0$. The test statistic is based on the following regression:

$$\widehat{u}_t^2 = a_0 + a_1 t + \text{error} \tag{26}$$

where $\widehat{u}_t = Y_t - X_t^T\widehat{\beta}(Z_t)$. Xiao (2009) showed that under H_0^a, $\widehat{t}_{a_1} = \widehat{a}_1/se(\widehat{a}_1) \xrightarrow{d} N(0,1)$, where $\widehat{a}_1$ is the OLS estimator of a_1 based on Eq. (26) and $\mathrm{SE}(\widehat{a}_1)$ is the estimated standard error of $\widehat{a}_1$.

However, Sun and Li (2009b) considered the case that under the null hypothesis, u_t is an $I(1)$ process. Therefore, the null hypothesis considered by Sun and Li (2009b) is H_0^b: u_t is an $I(1)$ process, and the alternative is H_1^b: u_t is an $I(0)$ process. Thus, the null hypothesis is $H_0^b : \rho = 1$ and the alternative hypothesis is H_1^b $|\rho| < 1$. We consider only the case that $\beta(z)$ is not a constant

function. Based on the well-established cointegration testing for linear models, one can test H_0 based on

$$\widehat{\rho} = \frac{\sum_t \widehat{u}_t \widehat{u}_{t-1}}{\sum_t \widehat{u}_{t-1}^2}$$

where $\widehat{u}_t$ is an estimator for $u_t = Y_t - X_t^{\mathrm{T}}\beta(Z_t)$ and the test statistic is $n(\widehat{\rho} - 1)$. Sun and Li (2009b) showed that the leading term of the test statistic depends on $\widehat{\beta}(Z_t)$ in a complicated way and the asymptotic distribution is not nuisance parameter free. Therefore, one needs to design some simulation (or bootstrap) methods to approximate the null distribution of $n(\widehat{\rho} - 1)$. It is still an open question as how to approximate the null distribution of the test statistic considered by Sun and Li (2009b).

3.6. *Varying-Coefficient Models with Time Trend Variables*

Gu and Hernandez-Verme (2009) and Liang and Li (2009) considered a varying-coefficient model with regressors containing a time trend:

$$Y_t = X_t^{\mathrm{T}}\beta(Z_t) + u_t \tag{27}$$

where $X_t^{\mathrm{T}} = (X_{1t}^{\mathrm{T}}, t)$ and X_{1t} is an $I(0)$ variable. Gu and Hernandez-Verme (2009) considered the LL estimation method and applied the method to evaluate the presence of credit rationing in the U.S. credit markets, while Liang and Li (2009) considered both the LC and local polynomial estimation methods.

3.7. *Varying-Coefficient Models with* $I(1)$ Error

Sun, Hsiao, and Li (2008b) consider the problem of estimating a varying-coefficient model

$$Y_t = X_t^{\mathrm{T}}\beta(Z_t) + u_t \tag{28}$$

when both X_t and the error term u_t are integrated $I(1)$ processes. They show that in this case, it is still possible to obtain consistent estimate of $\beta(\cdot)$, but the rate of convergence will be reduced to $O_p(h^2 + (nh)^{-1/2})$ rather than $O_p(h^2 + (n^2h)^{-1/2})$ as compared to the case when u_t is a stationary process.

4. NONPARAMETRIC INSTRUMENTAL VARIABLE ESTIMATION

There is a vast amount of papers available in the literature on parametric IVs estimation of econometric models in economics and finance. As with other economic models, one may consider nonparametric structural modeling to permit greater flexibility than tightly specified parametric models in describing such relationships. However, new problems arise for inference in nonparametric structural models that are not present in standard nonparametric regression; see Newey and Powell (2003). Estimation of such models depend on strong regularization and sometimes preclude the asymptotic distribution theory required for inference. To deal with these problems, Newey and Powell (1988) were the first to explore the nonparametric IV models and part of their result was later published in Newey and Powell (2003). Since then, some of the other papers in this area include Newey, Powell, and Vella (1999), Daroles, Florens, and Renault (2002), Blundell and Powell (2003), Das (2003, 2005), Ai and Chen (2003), Das, Newey, and Vella (2003), Newey and Powell (2003), Hall and Horowitz (2005), Cai, Das, Xiong, and Wu (2006) (CDXW, hereinafter), Horowitz (2007), and the references therein.

We describe the nonparametric model (with endogenous regressors) below. Suppose we have i.i.d. data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, and the data are generated by the following data generating process:

$$Y_i = g(X_i, Z_{i1}) + u_i \tag{29}$$

where $g(\cdot)$ is an unknown structural function of interest, Z_{i1} is a $\boldsymbol{d}_1 \times 1$ vector of exogenous variables, and the u_i's denote disturbances. The u_i's are correlated with the explanatory variables X_i and, in particular, $E(u_i|X_i) \neq 0$, so that $X_i \in \Re^{d_x}$ is an endogenous variable. Suppose, however, that for each i, we have available another observed data value, $Z_i = (Z_{i1}, Z_{i2})$, for which $E(u_i|Z_i) = 0$, where Z_{i2} is a $\boldsymbol{d}_2 \times 1$ vector of the so-called IVs. Clearly, the nonparametric IV model is different from the standard nonparametric model in the sense that because $E(u_i|X_i, Z_{i1}) \neq 0$, the structural function $g(\cdot)$ is not given by the regression $E(Y_i|X_i, Z_{i1})$.

Taking the conditional expectation of Eq. (29) yields the following integration equation:

$$\zeta(z) \equiv E[Y_i|Z_i = z] = E[g(X_i, z_1)|Z_i = z] = \int g(x, z_1) dF_{x|z}(x|z) \tag{30}$$

where $F_{x|z}(x|z)$ is the conditional distribution function of X_i given as $Z_i = z$. Although $\zeta(z)$ and $F_{x|z}(x|z)$ are estimable based on data $\{(X_i, Y_i, Z_i)\}$, estimation of $g(\cdot)$ is difficult because the relation that identifies $g(\cdot)$ is a Fredholm equation of the first kind, which leads to the difficulty called ill-posed inverse problem in the literature. That is, for nonparametric estimators $\widehat{\zeta}(z)$ and $\widehat{F}_{x|z}(x|z)$ obtained from preliminary nonparametric estimation,

$$\widehat{\zeta}(z) = \int g(x, z_1) d\widehat{F}_{x|z}(x|z)$$

may not exist a solution for $\hat{g}(\cdot)$. Even if it exists, it may not be computable and continuous in $\widehat{\zeta}(z)$ and $\widehat{F}_{x|z}(x|z)$. As pointed out by Newey and Powell (2003), noncontinuity of $\hat{g}(\cdot)$ is the biggest obstacle to overcome and the lack of continuity of $\hat{g}(\cdot)$ in $\widehat{\zeta}(\cdot)$ and $\widehat{F}_{x|z}(\cdot)$ means that a small change in $\widehat{\zeta}(\cdot)$ and $\widehat{F}_{x|z}(\cdot)$ may cause a huge error to $\hat{g}(\cdot)$. Therefore, the consistency of $\hat{g}(\cdot)$ may not exist even if both $\widehat{\zeta}(\cdot)$ and $\widehat{F}_{x|z}(\cdot)$ are consistent. To recover the structural function $g(\cdot)$ and to overcome these difficulties, in nowadays, several methods were proposed in the literature, described below.

4.1. Series Estimation

Newey and Powell (2003) suggested using the series method to approximate the unknown function $g(\cdot)$ as,

$$g(w) \approx \sum_{j=1}^{J} \gamma_j \varphi_j(w) \tag{31}$$

where $w = (x, z_1)$, $\{\varphi_j(\cdot)\}$ is a sequence of basis functions and $\{\gamma_j\}$ are the corresponding coefficients. Substitution of Eq. (31) into Eq. (30) leads to

$$\zeta(z) = E[Y_i|Z_i = z] \approx \sum_{j=1}^{J} \gamma_j E[\varphi_j(W_i)|Z_i = z] \equiv \sum_{j=1}^{J} \gamma_j p_j(z) = \gamma^T P(z)$$

where $p_j(z) = E[\varphi_j(W_i)|Z_i = z]$, $\gamma = (\gamma_1, \ldots, \gamma_J)^{\mathrm{T}}$ and $P(z) = (p_1(z), \ldots, p_J(z))^{\mathrm{T}}$. Now, to estimate $g(z)$, one can use a nonparametric two-stage approach. At the first stage, using a nonparametric method to obtain $\widehat{p}_j(z)$ and then at the second stage, using the least squares method to obtain $\widehat{\gamma}_j$ by a regression of Y_i on $\{\widehat{p}_j(Z_i)\}$. Finally, one obtains $\widehat{g}(w) = \sum_{j=1}^{J} \widehat{\gamma}_j \varphi_j(w)$. Under some regularity conditions, Newey and Powell (2003) derived the consistency of $\widehat{g}(w)$. But they did not obtain the asymptotic distribution of their estimator.

4.2. Functional Operator Approach

Hall and Horowitz (2005) considered a functional operator approach for estimating $g(\cdot)$. Taking an expectation of $\zeta(Z_i)f_{x,z}(v, Z_i)$ for any fixed v, we have

$$E[\zeta(Z_i)f_{x,z}(v, Z_i)] = \int \zeta(z)f_{z,z}(v, z)f_z(z)dz$$

where $f_{x,z}(x, z)$ and $f_z(z)$, respectively, denote the joint density of (Z_i, X_i) and the marginal density of Z_i. Substitution of Eq. (30) into the above equation yields

$$E[\zeta(Z_i)f_{x,z}(v, Z_i)] = \int \int g(x, z_1)f_{x,z}(x, z)f_{z,z}(v, z)dxdz$$

If one assumes that $g(x, z_1) = g(x)$; that is, $g(\cdot)$ depends only on the endogenous variable X_i but not on any exogenous variable, then,

$$E[Y_i f_{x,z}(v, Z_i)] = E[E(Y_i|Z_i)f_{x,z}(v, Z_i)] = \int g(x)t(x, v)dx \equiv Tg(v)$$

which defines a functional operator T, where

$$t(x, v) = \int f_{x,z}(x, z)f_{z,z}(v, z)dz$$

Clearly, T is a functional operator defined on the space of functions that are square integrable on $L_2(\Re^{d_x} \times \Re^{d_x})$. Assume that the functional operator T is nonsingular. Then, for each v, g(v) can be expressed as,

$$g(v) = E[Y_i(T^{-1}f_{x,z})(v, Z_i)] \tag{32}$$

and $g(v)$ could be estimated easily by,

$$\widehat{g}(v) = \frac{1}{2}\sum_{i=1}^{n} Y_i(T^{-1}f_{x,z})(v, Z_i)$$

if the operator T and $f_{x,z}(v, Z_i)$ were known. Clearly, $f_{x,z}(v, Z_i)$ can be estimated by a kernel method plus jackknife (leave-one-out) approach, given by,

$$\widehat{f}_{x,z}(v, Z_i) = \frac{1}{n}\sum_{j=1, j\neq i}^{n} K_h(X_j - v, Z_j - Z_i) \tag{33}$$

where $K(\cdot,\cdot)$ is a kernel in $\Re^{d_x+d_x}$. Hall and Horowitz (2005) proposed the following estimator:

$$\widehat{g}(v) = \frac{1}{2}\sum_{i=1}^{n} Y_i(\widehat{T}^{+}\widehat{f}_{x,z})(v, Z_i) \tag{34}$$

where $\widehat{T}^{+} = (\widehat{T} + a_n I)^{-1}$, which is a ridge type estimator and $a_n \to 0$ is a ridge parameter, and

$$\widehat{t}(x, v) = \int \widehat{f}_{x,z}(x, z)\widehat{f}_{z,z}(v, z)dz$$

where $\widehat{f}_{x,z}(x, z)$ is defined in Eq. (33). Alternatively, Hall and Horowitz (2005) suggested using a series method to estimate $f_{x,z}(x, z)$; see Hall and Horowitz (2005) for details. Finally, for a general form of $g(x, z_1)$, one can still define the functional operator T_{z1} for a fixed z_1 and then apply the same idea as above to define the nonparametric estimator for g(x, z_1); see Section 3 of Hall and Horowitz (2005) for the detailed discussions.

Remark 1. As addressed in Hall and Horowitz (2005) and Horowitz (2007), Eq. (32) is a Fredholm equation of the first kind. T^{-1} may not always exist and if not, it generates the so-called ill-posed inverse problem. This phenomenon happens if zero is a limit point of the eigenvalues of T, in particular, when $f_{x,z}(x, z)$ is a well-behaved density function. In that case, T^{-1} is not a bounded operator, and $g(\cdot)$ cannot be estimated consistently by replacing unknown population quantities on the right-hand side of Eq. (32) with consistent estimators. This problem is well known in the theory of integral equations. One way to deal with this problem is to modify T^{-1} to make it a continuous operator. Hall and Horowitz (2005) suggested using a ridge idea to replace T^{-1} for estimation purposes with $(T+a_n I)^{-1}$ (see Eq. (34) above), where I is the identity operator and $\{a_n\}$ is a sequence of positive constants that converge to 0 as $n \to \infty$.

Hall and Horowitz (2005) derived the asymptotic mean square error of their estimator and showed that for a certain class of distributions, the convergence rates are optimal in a minimax sense, while Horowitz (2007) obtained the asymptotic normality of $\widehat{g}(v)$.

Remark 2. For convenience of discussion, assume that $d_x = 1$ (X_i is univariate). Unfortunately, both papers by Hall and Horowitz (2005) and Horowitz (2007) did not discuss whether the convergence rate $(nh)^{-1/2}$ for ordinary nonparametric regression models can be achievable or not,

since the convergence rates in both papers depend on the smoothness conditions for the functions $f_{x,z}(\cdot)$ and $g(\cdot)$. To answer the aforementioned question, let us look at Theorem 4.1 of Hall and Horowitz (2005) or Theorem 1 of Horowitz (2007), from which, it follows that the asymptotic integrated mean squared errors (AIMSE) is of the order $O(n^{-(2\beta-1)/(2\beta+\alpha)})$ by using the same notation as in both papers. If it would achieve the optimal convergence rate for ordinary nonparametric regression models, $(2\beta-1)/(2\beta+\alpha)=4/5$ so that $\alpha=\beta/2-5/4$ that does not satisfy Assumption A3 in Hall and Horowitz (2005) or Assumption 3 in Horowitz (2007). Therefore, one might conclude that the optimal convergence rate for $\widehat{g}(v)$ cannot reach the optimal AIMSE rate $O(n^{-4/5})$ for ordinary nonparametric regression models. Finally, both papers mentioned above did not give an explicit expression for the asymptotic bias. Therefore, it is difficult to make the adaptive bandwidth selection feasibly implemented in practice. Now, a natural question arises is whether the optimal convergence rate $(nh)^{-1/2}$ is achievable for a nonparametric estimator under nonparametric IV settings. If possible, it would be interesting to investigate what the scenarios are. Also, it would be warranted to explore the asymptotic bias.

4.3. Projection Method

Newey et al. (1999) proposed using a projection method to estimate $g(\cdot)$. The reduced form of Eq. (29) can be expressed as,

$$X_i=\pi(Z_i)+\xi_i, \qquad E[\xi_i|Z_i]=0$$

where $\pi(Z_i)=E(X_i|Z_i)$. Further, using the new notation $W_i=(\xi_i,X_i,Z_{i1})\in\Re^{2d_x+d_1}$ and taking the conditional expectation of Eq. (29) conditional on (X_i, Z_i), we have,

$$\begin{aligned}E[Y_i|X_i,Z_i]&=g(X_i,Z_{i1})+E[u_i|X_i,Z_i]=g(X_i,Z_{i1})+E[u_i|\xi_i]\\&\equiv g(X_i,Z_{i1})+\lambda_0(\xi_i)\equiv h_0(W_i)\end{aligned}\tag{35}$$

by assuming that $E[u_i|X_i,Z_i]=E[u_i|\xi_i]$, where the definitions of $\lambda_0(\xi_i)$ and $h_0(W_i)$ should be apparent. Since $E[u_i]=0$, we have the following projection:

$$E[h_0(x,z_1,\xi_i)]=g(x,z_1)+E[\lambda_0(\xi_i)]=g(x,z_1)+E[u_i]=g(x,z_1)$$

Therefore, $g(x, z_1)$ can be estimated by a projection method as,

$$\widehat{g}_p(x, z_1) = n^{-1} \sum_{i=1}^{n} \widehat{h}_0(x, z_1, \xi_i)$$

if $\widehat{h}_0(x, z_1, \xi_i)$ and ξ_i would be known. To find a nonparametric estimate $\widehat{h}_0(x, z_1, \xi_i)$ in $\Re^{2d_x+d_1}$, one can use a kernel smoothing technique (say, LL fitting) as ordinary nonparametric regression by regressing Y_i on $(X_i, Z_{i1}, \widehat{\xi}_i)$, where $\widehat{\xi}_i$ is the nonparametric residual obtained from the reduced form as $\widehat{\xi}_i = X_i - \widehat{\pi}(Z_i)$, where $\widehat{\pi}(Z_i)$ is a nonparametric estimate of $\pi(Z_i)$. Therefore, the feasible estimate $\widehat{g}_p(x, z_1)$ is given by:

$$\widehat{g}_p(x, z_1) = \frac{1}{n} \sum_{i=1}^{n} \widehat{h}_0(x, z_1, \widehat{\xi}_i)$$

This method is termed as two-stage nonparametric fitting plus a projection. By following the steps in Masry and Tjøstheim (1997) and Cai and Masry (2000), recently, Su and Ullah (2008) derived the asymptotic properties of the estimator that are the exactly same as that for the ordinary nonparametric regression models. The main disadvantage of using this approach is that it suffers from the problem associated with the curse of dimensionality. Since the unknown function $g(x, z_1)$ is defined in $\Re^{d_x+d_1}$, the nonparametric model fitting has to be implemented in $\Re^{2d_x+d_1}$. This might be infeasible in applications when d_x is large.

Due to the computational convenience and high efficiency in imposing additivity, alternatively, Newey et al. (1999) suggested a series method as follows. At the first step, $\pi(Z_i)$ is estimated by:

$$\widehat{\pi}(Z_i) = \sum_{j=1}^{K_1} \widehat{\gamma}_j r_j(Z_i)$$

where $\{\widehat{\gamma}_j\}$ are obtained by a regression of X_i versus $\{r_j(Z_i)\}$, $\{r_j(Z_i)\}$ is a sequence of basis functions. Then, one obtains the residual $\widehat{\xi}_i = X_i - \widehat{\pi}(Z_i)$. At the second step, a series method is used again as follows. Use the series approximation again to approximate $g(x, z_1)$ and $\lambda_0(\xi)$, respectively, as,

$$g(x, z_1) \approx \sum_{l=1}^{K_2} \beta_{l1} \phi_l(x, z_1), \quad \text{and} \quad \lambda_0(\xi) \approx \sum_{m=1}^{K_3} \beta_{m2} \psi_m(\xi)$$

where $\{\phi_l(x, z_1)\}$ and $\{\psi_m(\xi)\}$ are basis functions, so that

$$h_0(w) \approx \sum_{l=1}^{K_2} \beta_{l1}\phi_l(x, z_1) + \sum_{m=1}^{K_3} \beta_{m2}\psi_m(\xi)$$

Then, $\{\beta_{l1}\}$ and $\{\beta_{m2}\}$ can be easily estimated by regressing Y_i versus $\{\phi_l(X_i, Z_{i1})\}$ and $\{\psi_m(\widehat{\xi}_i)\}$. Therefore, $g(x, z_1)$ can be estimated as,

$$\widehat{g}_s(x, z_1) = \sum_{l=1}^{K_2} \widehat{\beta}_{l1}\phi_l(x, z_1)$$

Newey et al. (1999) derived the consistency of $\hat{g}_s(x, z_1)$ with a convergence rate for consistency, but they did not derive the asymptotic distribution of their proposed estimator.

4.4. Functional-Coefficient Modeling

Das (2005) considered a nonparametric IV model with discrete endogenous variables. That is, X_i is a discrete variable. Without loss of generality, assume that $X_i = 0$ or 1. Then, $g(x, z_1)$ can be rewritten as,

$$g(x, z_1) = g(0, z_1)\mathbf{1}(x = 0) + g(1, z_1)\mathbf{1}(x = 1) = a_0(z_1) + a_1(z_1)x$$

where $a_0(z_1) = g(0, z_1)$ and $a_1(z_1) = g(1, z_1) - g(0, z_1)$. Therefore, $g(x, z_1)$ is linear in endogenous variable but nonlinear in exogenous variable, which is called a functional-coefficient model in the literature; see Cai, Fan, and Yao (2000), Li, Huang, Li, and Fu (2002), CDXW (2006), Juhl (2005), and Cai and Xu (2008). Assuming that $g(x, z_1)$ has a higher order partial derivative with respect to x, then applying Taylor expansion to $g(x, z_1)$ we obtain

$$g(x, z_1) = \sum_{j=1}^{\infty} \frac{\partial^j g(0, z_1)}{\partial x^j}\frac{x^j}{j!} \approx \sum_{j=0}^{d} a_j(z_1)x_j \tag{36}$$

for some d, where $a_j(z_1) = \partial^j g(0, z_1)/\partial x^j$ and $x_j = x^j/j!$. This implies that a functional-coefficient model might approximate a general nonparametric model well. Therefore, CDXW (2006) studied the following functional-coefficient IV model:

$$Y_i = \sum_{j=1}^{d} a_j(Z_{i1})^{\mathrm{T}} X_{ij} + u_i = a(Z_{i1})^{\mathrm{T}} X_i + u_i, \quad E[u_i|Z_i] = 0 \tag{37}$$

where Y_i is an observable scalar random variable, $\{a_j(\cdot)\}$ are the unknown structural functions of interest, $X_{i0} \equiv 1, X_i = (X_{i0}, X_{i1}, \ldots, X_{id})^{\mathrm{T}}$ is a $(\boldsymbol{d}+1)$-dimension vector consisting of $\boldsymbol{d}$ endogenous regressors, $a(Z_{i1}) = (a_0(Z_{i1}), \ldots, a_d(Z_{i1}))^{\mathrm{T}}$, and Z_i is a $(\boldsymbol{d}_1+\boldsymbol{d}_2)$-dimension vector consisting of a $\boldsymbol{d}_1$-dimension vector Z_{i1} of exogenous variables and a $\boldsymbol{d}_2$-dimension vector Z_{i2} of IVs.

Model (37) includes the following nonparametric IV model with binary endogenous variable D_i as a special case:

$$Y_i = a_0(Z_{i1}) + a_1(Z_{i1})D_i + \varepsilon_i$$

which, as noted above, is analyzed in Das (2005). Further, if $a_j(\cdot)$ is a threshold function such as,

$$a_j(z) = a_{j1}\mathbf{1}(z \le r_j) + a_{j2}\mathbf{1}(z > r_j)$$

for some r_j, then model (37) may describe a threshold IV regression model. Recently, a threshold model related to this with endogenous covariates has been considered in Caner and Hansen (2004). In this way, the class of models in Eq. (37) includes some interesting special cases that arise commonly in empirical research.

As elaborated by CDXW (2006), functional-coefficient models are appropriate for many applications in economics and finance, and in particular when additive separability of covariates is unsuitable for the problem at hand. For a specific example, CDXW (2006) considered a labor economics problem which is to establish an empirical relationship between marginal returns to education and the level of schooling (see Schultz, 1997). If work experience is also an attribute valued by employers, then the marginal returns to education should vary with experience. As suggested by Card (2001), if a wage model assumes the additive separability of education and experience, the returns to education can be understated at higher levels of education because the marginal return to education is plausibly increasing in work experience. This setting is, therefore, a natural one for a functional-coefficient model, which was further explored by CDXW (2006). Indeed, the marginal returns to education vary positively and nonlinearly with experience and these returns are themselves declining in experience for both low experienced and high experienced workers; see CDXW (2006) for details.

To estimate $\{a_j(z_1)\}$ nonparametrically, CDXW (2006) proposed a two-stage nonparametric method, described as follows. We begin with the first stage, where we obtain $\widehat{\pi}_j(Z_i)$, the fitted value for $\pi_j(Z_i) = E[X_{ij}|Z_i]$ $(1 \le j \le d; 1 \le i \le n)$. To this end, we apply the LL fitting technique and

the jackknife (leave-one-out) idea as follows. Assuming that $\{\pi_j(\cdot)\}$ has a continuous second-order derivative, when Z_k falls in a neighborhood of Z_i, a Taylor expansion approximates $\pi_j(Z_k)$ by,

$$\pi_j(Z_k) \approx \pi_j(Z_i) + (Z_k - Z_i)^{\mathrm{T}}\pi_j'(Z_i) = \alpha_{ij} + (Z_k - Z_i)^{\mathrm{T}}\beta_{ij}$$

The jackknife idea is to use the all observations except the ith observations in estimating $\pi_j(Z_i)$. Then, the least squares estimator with a local weight (i.e., locally weighted least squares) is given by,

$$\sum_{k\neq i}^{n}\{X_{kj} - \alpha_{ij} - (Z_k - Z_i)^{\mathrm{T}}\beta_{ij}\}^2 \quad K_{h_1}(Z_k - Z_i)$$

Minimizing the above locally weighted least squares with respect to α_{ij} and β_{ij} gives the LL estimate $\pi_j(Z_i)$ by $\widehat{\pi}_{j,-i}(Z_i) = \widehat{\alpha}_{ij}$. Now, we derive the LL estimator of $\{a_j(\cdot)\}$. The LL estimators $\widehat{b_j}$ and $\widehat{c_j}$ are defined as the minimizers of the sum of weighted least squares

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=0}^{d}\{b_j + (Z_{i1} - z_1)^T c_j\}\widehat{\pi}_{j,-i}(Z_i)\right]^2 L_{h_2}(Z_{i1} - z_1)$$

and $\widehat{a}_j(z_i) = \widehat{b_j}$, where $L(\cdot)$ is a kernel function at this step.

CDXW (2006) showed that under some regularity conditions,

$$\sqrt{nh_2^{d_1}}\left[\widehat{a}(z_1) - a(z_1) - \frac{h_2^2}{2}\mathrm{tr}\{\mu_2(L)a''(z_1)\} + o_p(h_2^2)\right] \xrightarrow{d} N(0, \Sigma(z_1)) \qquad (38)$$

where $\Sigma(z_1) = f_{z_1}^{-1}(z_1)v_0(L)\Omega_0^{-1}(z_1)\Omega_1(z_1)\Omega_0^{-1}(z_1)$, $f_{z_1}(z_1)$ is the marginal density of Z_{i1}, $\Omega_0(z_1) = E[\pi(Z_i)\pi(Z_i)^{\mathrm{T}}|Z_{i1} = z_1]$, and $\Omega_1(z_i) = \Omega_{\eta,1}(z_1) + \Omega_{\xi,1}(z_i) - 2\Omega_{\eta\xi,1}(z_1)$. The definitions of $\Omega_{\eta,1}(z_1)$, $\Omega_{\xi,1}(z_1)$, and $\Omega_{\eta\xi,1}(z_1)$ can be found in CDXW (2006) and they are omitted here due to too many notations.

One difference of the results in Eq. (38) compared with those in some other two-stage instrumental regressions (see Newey & Powell, 2003; Newey et al., 1999) is the asymptotic variance term. Here the asymptotic variance consists of three terms: the first addresses the variation of measurement error in the second step, the second term accounts for variability of the estimated reduced form, and the third term accounts correctly for the asymptotic covariance between the first and second steps. The presence of the covariance term is different from some other IV estimators (e.g., Newey et al., 1999), and arises because the second step does not condition on the first step dependent variables.

4.5. Bandwidth Selection

Selecting an optimal (data-driven) bandwidth is an important aspect in applications. Unfortunately, there is basically not an elegant approach to discuss theoretically and empirically how to adaptively select a bandwidth under nonparametric IV settings, when a nonparametric method is applied to estimate the structural regression function, except a rule-of-thumb bandwidth proposed by CDXW (2006) for the functional-coefficient IV models in Eq. (37). As mentioned in CDXW (2006), the second stage estimation is not sensitive to the choice of the first stage bandwidth so long as the bandwidth h_1 at the first stage is chosen small enough such that the bias in the first stage is not too large. This gives us an ad hoc method to choose h_1, similar to that discussed in Cai (2002a): use the CV or generalized CV criterion of Cai, Fan, and Li (2000) or others to select the bandwidth $\widehat{h}_{10}$, Then use $h_1 = A_0\widehat{h}_{10}(A_0 = 1/2$, say, or smaller) or choose a very small h_1 as the first-stage bandwidth. Alternatively, A_0 can be taken to be $A_0 = n^{-\alpha 1}$ with $\alpha_1 > l/(d_1 + 4)(d_1 + l + 4)$, as discussed in Cai (2002a), where d_1 is the dimension of the regressor z_1.

In implementation at the second stage, the choice of bandwidth can be carried out as in standard nonparametric regression. In that case, a number of methods could be used to select h_2, including CV (Stone, 1974), generalized CV (Cai et al., 2000), preasymptotic substitution method (Fan & Gijbels, 1996), the plug-in bandwidth selector (Ruppert, Sheather, & Wand, 1995), empirical bias method (Ruppert, 1997), nonparametric version of the Akaike information criterion (AIC) (see Eq. (66) later) (Hurvich, Simonoff, & Tsai, 1998; Cai & Tiwari, 2000) or the Schwarz-type information criterion (SIC), among others. However, there appears to be no results in the literature for a data-driven bandwidth selection with optimal properties (see Newey et al. (1999) for the related discussion) under nonparametric IV settings. It is an open question for future work and it would be very interesting to give a more precise result. Nevertheless, as recommended by CDXW (2006), the procedure suggested above is a useful one for practitioners.

4.6. Semiparametric IV Models

Finally, we would like to mention some recent developments on nonparametric IV models with a parametric part, so that they become semiparametric IV models. Due to the limitation of space, we only cite some

references here. First, we mention the paper by Ai and Chen (2003) which discussed a general framework for analyzing economic data $(X,\ Y)$ by assuming that the data satisfy some conditional moment restrictions such as,

$$E[\rho(Z,\theta,m(\cdot))|X]=0 \tag{39}$$

where $Z=(T^{\mathrm{T}},X_z^{\mathrm{T}})^{\mathrm{T}}, X_z$ is a subset of X, and $\rho(\cdot)$ is a vector of known (residual) functions. The true conditional distribution of Y given X is assumed unknown and the parameters of interest contain a vector of finite dimensional unknown parameters θ and possibly a vector of infinite dimensional unknown functions $m(\cdot)$. Clearly, if $(Z=(Y_1,Y_2^{\mathrm{T}},X_1^{\mathrm{T}},X_2^{\mathrm{T}}), X_z=X_1$ and $\rho(Z_i,\theta,m(\cdot))=Y_{i1}-\theta^{\mathrm{T}}X_{i1}-m(Y_{i2})$, model (39) reduces to a partially linear model

$$Y_{1i}=\beta^{\mathrm{T}}X_{i1}+m(Y_{i2})+u_i \tag{40}$$

where $E[u_i|X_i]=0$, which was studied by Newey et al. (1999) and Park (2003), while Pakes and Olley (1995) considered a semiparametric IV model with endogenous variables restricted only to the parametric part. Newey et al. (1999) used the series method to approximate $m(\cdot)$ and then to estimate both β and $m(\cdot)$ based on the nonparametric series method, whereas Pakes and Olley (1995) and Park (2003) applied the generalized method of moment estimation method to estimate β and $m(\cdot)$.

As argued by Ai and Chen (2003), model (39) covers many known nonparametric and semiparametric models as a special case. To estimate θ and $m(\cdot)$, Ai and Chen (2003) proposed to approximate $m(\cdot)$ by a sieve method and then to estimate θ and the sieve parameters jointly by applying the method of minimum distance. They showed that the sieve estimator of $m(\cdot)$ is consistent with a rate faster than $n^{-1/4}$ under certain metric and the estimator of θ is $\sqrt{n}$-consistent and asymptotically normally distributed. Finally, they addressed the efficiency by choosing the optimally weighted minimum distance to attain the semiparametric efficiency bound. But, they did not provide the asymptotic normality for the sieve estimator of $m(\cdot)$ (see Ai & Chen, 2003 for details).

To obtain the asymptotic normality of nonparametric part, Cai and Xiong (2006) considered a partially varying-coefficient IV model with the following form:

$$Y=g(X,Z_1)+\varepsilon=g_1(Z_{11})^{\mathrm{T}}Z_{12}+g_2(Z_{11})^{\mathrm{T}}X_1+\beta_1^{\mathrm{T}}Z_{13}+\beta_2^{\mathrm{T}}X_2+\varepsilon \tag{41}$$

where Y is an observable scalar random variable, $\boldsymbol{X}=(\boldsymbol{X}_1^{\mathrm{T}},\boldsymbol{X}_2^{\mathrm{T}})^{\mathrm{T}}$ is a vector of endogenous variables including l-dimension vector $\boldsymbol{X}_1$ and p-dimension

vector $\boldsymbol{X}_2$, $\boldsymbol{Z}_1 = (\boldsymbol{Z}_{11}^{\mathrm{T}}, \boldsymbol{Z}_{12}^{\mathrm{T}}, \mathrm{Z}_{13}^{\mathrm{T}})^{\mathrm{T}}$ is a vector of exogenous variables, consisting of d_{11}-dimension vector $\boldsymbol{Z}_{11}$, d_{12}-dimension vector $\boldsymbol{Z}_{12}$ with its first element being one, and d_{13}-dimension vector $\boldsymbol{Z}_{13}$, $\boldsymbol{Z} = (\boldsymbol{Z}_1^{\mathrm{T}}, \boldsymbol{Z}_2^{\mathrm{T}})^{\mathrm{T}}$ is a d_z-dimension vector with $\boldsymbol{Z}_2$ being a vector of IVs of dimension d_2, $d_z = d_{11}+d_{12}+d_{13}+d_2$, and $E(\varepsilon\,|Z) = 0$.

To estimate β and $g(\cdot)$ in (41), Cai and Xiong (2006) proposed a three-stage method, briefly described below. First, by regarding β as a function of Z_{11}; that is $\beta(Z_{11})$, then model (41) becomes (37). The nonparametric two stage proposed in CDXW (2006) can be applied here to estimate $g(\cdot)$ and $\beta(\cdot)$. Note that while β is a global parameter, the estimation of $\beta(\cdot)$ only involves the local data points in a neighborhood of Z_{11} so that the variance is too large. To reduce variance, the estimation of the constant coefficients requires using all data points. Cai and Xiong (2006) proposed using the (weighting) average method to obtain the estimator for β and they showed that the average estimator of β is $\sqrt{n}$-consistent. To address the efficiency of the constant parameter estimator, the weighted version estimator, similar to Ai and Chen (2003), can be used to gain the efficiency by choosing the optimal weighting function to minimize the asymptotic variance. See Cai and Xiong (2006) for the related discussions.

Alternatively, one may use the profile likelihood (least squares for normal likelihood) approach to estimate β_1 and β_2 in (41). It is well documented in the literature that for ordinary semiparametric models, profile likelihood is a useful approach and is semiparametrically efficient; see Speckman (1988), Cai (2002a, 2002c), and Fan and Huang (2005) for details. Now we discuss applying the profile likelihood approach to estimate β_1 and β_2 in (41). For given β_1 and β_2, model (41) becomes

$$Y^* = g_1(Z_{11})^{\mathrm{T}} Z_{12} + g_2(Z_{11})^{\mathrm{T}} X_1 + \varepsilon \tag{42}$$

where $Y^* = Y - \beta_1^{\mathrm{T}} Z_{13} - \beta_2^{\mathrm{T}} X_2$ is the partial residual. This transforms the partially varying-coefficient IV model (41) into the varying-coefficient IV model (37). The two-stage LL estimation technique proposed in CDXW (2006) can be applied to estimate the coefficient functions $g_1(\cdot)$ and $g_2(\cdot)$, denoted by $\widehat{g}_1(\cdot)$ and $\widehat{g}_2(\cdot)$, respectively. According to CDXW (2006), both $\widehat{g}_1(\cdot)$ and $\widehat{g}_2(\cdot)$ are linear estimators of Y^*. That is,

$$\widehat{\mathbf{M}} = \begin{pmatrix} \hat{g}_1(Z_{11,1})^{\mathrm{T}} Z_{12,1} + \hat{g}_2(Z_{11,1})^{\mathrm{T}} X_{1,1} \\ \vdots \\ \hat{g}_1(Z_{11,n})^{\mathrm{T}} Z_{12,n} + \hat{g}_2(Z_{11,n})^{\mathrm{T}} X_{1,n} \end{pmatrix} = \boldsymbol{S}\boldsymbol{Y}^* = \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{Z}_{13}\beta_1 - \boldsymbol{X}_2\beta_2)$$

where $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_n^*)^{\mathrm{T}}$. The matrix $\boldsymbol{S}$ is a smoothing matrix and depends only on the data $\{(Z_{11,i}, Z_{12,i}, X_{1,i}, \widehat{X}_{1,i}), i = 1, \ldots, n\}$ and the kernel function, where $\widehat{X}_{1,i}$ is obtained from the reduced equation by the jackknife least squares method; see CDXW (2006) for the explicit expression for $\boldsymbol{S}$ and $\widehat{X}_{1,i}$ (which depends on the data $\{(X_j, Z_j), j = 1, \ldots, i-1, i+1, \ldots, n\}$). Substituting $\widehat{\boldsymbol{M}}$ into Eq. (42), we obtain the following linear IV model

$$(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{S})[\boldsymbol{Z}_{13}\beta_1 + \boldsymbol{X}_2\beta_2] + \varepsilon \tag{43}$$

Applying the two-stage least squares to the linear model (43), we obtain the profile likelihood estimators of β_1 and β_2, respectively, termed as *profile two-stage least squares estimate*. Note that if there is no endogeneity in the model, Fan and Huang (2005) showed that the profile likelihood estimator is semiparametrically efficient. Therefore, we conjecture that the profile least squares estimate for β_2 described above should be $\sqrt{n}$-consistent and semiparametrically efficient. It is interesting to justify this result theoretically.

5. NONPARAMETRIC QUANTILE REGRESSION MODELS

Since quantile regression or conditional quantile was introduced by Koenker and Bassett (1978), it has been successfully and widely used in various disciplines, such as finance, economics, medicine, and biology. In nowadays, estimation of conditional quantiles is a common practice in risk management operations and many other financial applications. The literature on estimating quantile regression function is large but is still swiftly growing. Much of the study on quantile regression is based on linear parametric quantile regression models. But in recent years, nonparametric quantile regression models in both theory and applications have attracted a great deal of research attentions due to their greater flexibility than tightly specified parametric models. A nonexhaustive list of important recent contributions to this growing literature include (but not limited to) Chaudhuri (1991), Koenker, Portnoy, and Ng (1992), Fan, Hu, and Troung (1994), Koenker, Ng, and Portnoy (1994), Chaudhuri, Doksum, and Samarov (1997), He, Ng, and Portnoy (1998), Yu and Jones (1998), He and Ng (1999), He and Portnoy (2000), Honda (2000, 2004), Khindanova and Rachev (2000), Cai (2002b), Cai and Ould-Said (2003), De Gooijer and Zerom (2003), Yu and Lu (2004), Engle and Manganelli (2004), Horowitz and Lee (2005), Kim (2007), and Cai and Xu (2008) and references therein

for recent statistics and econometrics literature on nonparametric estimation of quantile regression models.

Let $\{X_t, Y_t\}_{t=1}^n$ be a stationary sequence and $F(y|x)$ denote the conditional distribution of Y_t given $X_t = x$, where X_t is a vector of covariates in $\Re^d$, including possibly exogenous variables and lagged variables, the conditional quantile function of Y_t given $X_t = x$ is defined as, for any $0 < \tau < 1$,

$$q_\tau(x) = \inf\{y \in \Re : F(y|x) \geq \tau\} = F^{-1}(\tau|x) \tag{44}$$

where $F^{-1}(\tau|x)$ is the inverse function of $F(y|x)$. Equivalently, $q_\tau(x)$ can be expressed as,

$$q_\tau(x) = \arg\min_{a \in \Re} E\{\rho_\tau(Y_t - a)|X_t = x\} \tag{45}$$

where $\rho_\tau(y) = y[\tau - I\{y < 0\}]$ with $y \in \Re$ is called the loss ("check") function and $I\{A\}$ is the indicator function of any set A. Function $q_\tau(x)$ is called as a conditional quantile function or regression quantile.

It is well documented that quantile regression has several important properties, described as follows. It does not require knowing the distribution of Y_t and symmetry of the distribution. When $\tau = 1/2$, it becomes the median or least absolute deviation regression, which is well known to posses the robustness. Therefore, it has a robust property. Also, it has an ability to model heterogeneous effects and to account for unobserved heterogeneity. To see the intuitive behind this property, we use the basic Skorohod representation to express the quantile regression model. Using this representation, the dependent variable Y_t, conditional on the exogenous variable of interest X_t, takes the form

$$Y_t = q(X_t, U_t), \qquad \text{where} \qquad U_t|X_t \sim U(0, 1)$$

where $q(x, u) = q_u(x)$ is the conditional uth quantile of Y_t given $X_t = x$ and U_t is the nonseparable error. Furthermore, it is convenient to use the conditional quantile for detecting conditional heteroskedasticity. To this end, we assume that Y_t is related to X_t through the model

$$Y_t = m(X_t) + \sigma(X_t)\varepsilon_t$$

where $m(\cdot)$ is the mean function, $\sigma^2(\cdot)$ is the variance function, and X_t and ε_t are independent. The conditional quantile of Y_t given X_t is

$$q_\tau(X_t) = m(X_t) + \sigma(X_t)F_{\varepsilon_t}^{-1}(\tau)$$

where $F_{\varepsilon_t}(\cdot)$ is the distribution of ε_t. An informal way to test conditional heteroskedasticity is to use a graph. That is, if the curves of $q_\tau(x)$ for

different values of τ are parallel, this indicates that $\sigma(\cdot)$ should be a constant. Moreover, regression quantiles can also be useful for the estimation of predictive intervals. For example, in predicting the response from a given covariate X_t, estimates of $q_{\alpha/2}(X_t)$ and $q_{1-\alpha/2}(X_t)$ can be used to obtain a $(1-\alpha)$ 100% nonparametric predictive interval. Finally, it is very useful in various applied fields. For example, in risk management, it can be used to compute the conditional value-at-risk (CVaR): the percentage loss in market value over a given time horizon that is exceeded with a certain probability, and the conditional expected shortfall (CES). Indeed, CVaR can be regarded as a special case of quantile regression. Of course, there are many methods available to model the CVaR. The CES can be expressed in terms of a regression quantile as,

$$E[Y_t | Y_t \le q_\tau(X_t), X_t] = \int_0^\tau \frac{q_u(X_t)du}{\tau}$$

For details, see Cai and Wang (2008).

Given observed data $\{X_t, Y_t\}_{t=1}^n$, the main interest is to estimate $q_\tau(x)$. If we assume that $q_\tau(x) = \beta_\tau^T x$, we obtain a linear quantile regression model, which is popular in the literature; see the book by Koenker (2005), and we can estimate easily the parameters (see Eq. (60) below). In some practical applications, a linear quantile regression model might not be flexible enough to capture the underlying complex dependence structure. For example, some components may be highly nonlinear or some covariates may be interactive. Therefore, to make quantile regression models more flexible, there is a swiftly growing literature on nonparametric quantile regression. Various smoothing techniques, such as kernel methods, splines, and their variants, have been used to estimate the nonparametric quantile regression for both independent and time series data. Some recent developments and detailed discussions on theory, methodologies, and applications can be found in the literature. For example, Chaudhuri (1991), Fan et al. (1994), Chaudhuri et al. (1997), Yu and Jones (1998), Honda (2000), Cai (2002b), and Cai and Ould-Said (2003) considered nonparametric kernel smoothing estimate of quantile function, while He et al. (1998), He and Ng (1999), and He and Portnoy (2000) used spline methods to obtain nonparametric estimate. However, a purely nonparametric quantile regression model may suffer from the so-called "curse of dimensionality" problem, the practical implementation might not be easy, and the visual display may not be useful for the exploratory purposes. To deal with the aforementioned problems, some dimension reduction modeling methods have been proposed in the literature. For example, De Gooijer and Zerom (2003), Yu and Lu (2004),

and Horowitz and Lee (2005) considered the additive quantile regression models for i.i.d. data, while Honda (2004) and Cai and Xu (2008) investigated the varying-coefficient quantile regression models for time series processes. Particularly, there has been some study on a time-varying coefficient quantile regression model, which is potentially useful to see whether the quantile regression changes over time and in a case with a practical interest is, for example, the analysis of the reference growth data by Cole (1994), Wei, Pere, Koenker, and He (2006), Wei and He (2006), and Kim (2007).

5.1. Direct Methods

A direct procedure is based on equation (44), described as follows. First, estimate the conditional distribution function using a nonparametric method such as the "double-kernel" LL technique of Yu and Jones (1998) and then to invert the conditional distribution estimator to produce an estimator of a conditional quantile. This estimator is called the Yu and Jones estimator (see $\widehat{q}_{\tau,\mathrm{LL}}(x)$ in (57) later); see Yu and Jones (1998) for details. As noticed by Cai (2002b) and Cai and Wang (2008), the key for a direct estimation method is to find a good estimator for conditional distribution function. Further, as demonstrated by Cai (2002b), although LL estimators of the Yu and Jones type have some attractive properties such as no boundary effects, design adaptation, and mathematical efficiency; see, for example, Fan and Gijbels (1996), they have the disadvantage of producing conditional distribution function estimators that are not constrained either to lie between zero and one or to be monotone increasing although some modifications in implementation were addressed by Yu and Jones (1998). In both these respects, the NW methods are superior, despite their rather large bias and boundary effects. The properties of positivity and monotonicity are particularly advantageous if the method of inverting the conditional distribution estimator is applied to produce an estimator of a conditional quantile.

To overcome these difficulties, Cai (2002b) and Cai and Wang (2008) proposed a weighted version of the NW (WNW) estimator and weighted double kernel (WDK) estimator, which are designed to possess the superior properties of LL methods such as bias reduction and no boundary effect and to preserve the property that the NW estimator is always a distribution function. Cai (2002b) and Cai and Wang (2008) established the asymptotic normality and weak consistency for both the WNW and WDK estimators of

conditional distribution for α-mixing under a set of weaker conditions at both boundary and interior points. It is therefore shown, to the first order, that the WNW method enjoys the same convergence rates as those of the LL "double-kernel" procedure of Yu and Jones (1998). More importantly, both the WNW and WDK estimators have desired sampling properties at both boundary and interior points of the support of the design density. Cai (2002b) and Cai and Wang (2008) also derived both the WNW and WDK estimators of the conditional quantile by inverting their estimated conditional distributions estimator and showed that both the WNW and WDK quantile estimators always exist as a result of both the WNW and WDK distributions being a distribution function in finite samples and that they inherit all advantages from the WNW and WDK estimators of conditional distribution.

For simplicity of notation, we consider the case of $d = 1$. We now turn to the estimation of the conditional distribution function $F(y|x)$. To this end, let $p_t(x)$, for $1 \le t \le n$, denote the weight functions of the data $X_1, \ldots, X_n$ and the design point x with the property that each $p_t(x) \ge 0, \sum_{t=1}^{n} p_t(x) = 1$ and

$$\sum_{t=1}^{n} (X_t - x) p_t(x) K_h(x - X_t) = 0 \tag{46}$$

where $K(\cdot)$ is a kernel function, $K_h(\cdot) = K(\cdot/h)/h$, and $h = h_n > 0$ is the bandwidth. Motivated by the property of LL estimator, the constraint (46) can be regarded as a discrete moment condition; see Fan and Gijbels (1996, p. 63) for details. Of course, $\{p_t(x)\}$ satisfying these conditions are not uniquely defined and we specify them by maximizing $\prod_{t=1}^{n} p_t(x)$ subject to the constraints. The weighted version of NW estimator of the conditional distribution $F(y|x)$ of Y_t given $X_t = x$ is defined

$$\widehat{F}_{\mathrm{WNW}}(y|x) = \frac{\sum_{t=1}^{n} p_t(x) K_h(x - X_t) \mathbf{1}(Y_t \le y)}{\sum_{t=1}^{n} p_t(x) K_h(x - X_t)}$$

Note that $0 \le \widehat{F}_{\mathrm{WNW}}(y|x) \le 1$ and it is monotone in y. Cai (2002b) showed that $\widehat{F}_{\mathrm{WNW}}(y|x)$ is first-order equivalent to a LL estimator (see $\widehat{F}_{\mathrm{LL}}(y|x)$ in Eq. (56) later). More importantly, that $\widehat{F}_{\mathrm{WNW}}(y|x)$ has automatic good behavior at boundaries. In contrast, $\widehat{F}_{\mathrm{LL}}(y|x)$ may not take values in [0,1] and it may not be monotone in y.

The natural question arises regarding how to choose the weights. Borrowing the idea is from the empirical likelihood, Cai (2002b) suggested

maximizing $\sum_{t=1}^{n} \log\{p_t(x)\}$ subject to the constraints $\sum_{t=1}^{n} p_t(x) = 1$ and Eq. (46) through the Lagrange multiplier method, the $\{p_t(x)\}$ are simplified to

$$p_t(x) = n^{-1}\{1 + \lambda(X_t - x)K_h(x - X_t)\}^{-1}$$

where λ, a function of data and x, is uniquely defined by Eq. (46), which ensues that $\sum_{t=1}^{n} p_t(x) = 1$. Equivalently, λ is chosen to maximize

$$L_n(\lambda) = \frac{1}{nh}\sum_{t=1}^{n} \log\{1 + \lambda(X_t - x)K_h(x - X_t)\} \tag{47}$$

In implementation, Cai (2002b) recommended using the Newton Raphson scheme to find the root of equation $L_n'(\lambda) = 0$.

Cai (2002b) showed that, under some regularity conditions including that $\{(X_t, Y_t)\}_{t=1}^{n}$ is an α-mixing sequence, then as $n \to \infty$,

$$\widehat{F}_{\mathrm{WNW}}(y|x) - F(y|x) = \frac{1}{2}h^2\mu_2(K)F^{2,0}(y|x) + o_p(h^2) + O_p((nh)^{-1/2}) \tag{48}$$

where $F^{a,b}(y|x) = \partial^{a+b}/\partial y^a \partial x^b F(y|x)$ and $\mu_j(K) = \int u^j K(u)du$. This, of course, implies that $\widehat{F}_{\mathrm{WNW}}(y|x) \to F(y|x)$ in probability with a rate. In addition, Cai (2002b) derived the asymptotic normality for $\widehat{F}_{\mathrm{WNW}}(y|x)$ as,

$$\sqrt{nh}[\widehat{F}_{\mathrm{WNW}}(y|x) - F(y|x) - B_f(y|x) + o_p(h^2)] \xrightarrow{d} N(0, \sigma_f^2(y|x)) \tag{49}$$

where the bias and variance are given, respectively, by:

$$B_f(y|x) = \frac{1}{2}h^2\mu^2(K)F^{2,0}(y|x), \quad \text{and} \quad \sigma_f^2(y|x) = v_0(K)F(y|x)\frac{[1 - F(y|x)]}{f_1(x)} \tag{50}$$

with $f_1(x)$ being the marginal density of X_t. This implies that to the first order, the WNW method enjoys the exactly same convergence rates as those of LL "double-kernel" procedure (see $\widehat{F}_{\mathrm{LL}}(y|x)$ in Eq. (56) later) of Yu and Jones (1998), under similar regularity conditions. However, Yu and Jones (1998) treated only the case of independent data.

Based on Eq. (44), we define the WNW type conditional quantile estimator $\widehat{q}_{\mathrm{WNW}}(x)$ to satisfy $\widehat{F}_{\mathrm{WNW}}(\widehat{q}_{\mathrm{WNW}}(x)|x) = \tau$ so that

$$\widehat{q}_{\mathrm{WNW}}(x) = \inf\{y \in \Re : \widehat{F}_{\mathrm{WNW}}(y|x) \geq \tau\} \equiv \widehat{F}_{\mathrm{WNW}}^{-1}(\tau|x) \tag{51}$$

Clearly, $\widehat{q}_{\mathrm{WNW}}(x)$ always exists since $\widehat{F}_{\mathrm{WNW}}(y|x)$ is between 0 and 1 and monotone in y, and it involves only one bandwidth so that it makes practical implementation more appealing. In contrast, the LL double-kernel

estimator of Yu and Jones (1998) has some difficulty of inverting the conditional distribution estimator due to lack of monotonicity and it requires choosing two bandwidths although the second bandwidth should not be very sensitive (see Remark 3 later). Furthermore, Cai (2002b) showed that the WNW estimator $\widehat{q}_{\tau,\mathrm{WNW}}(x)$ maintains the aforementioned advantages as $\widehat{F}_{\mathrm{WNW}}(y|x)$ does. Also, Cai (2002b) showed that under some regularity conditions, as $n \to \infty$,

$$\sqrt{nh}[\widehat{q}_{\tau,\mathrm{WNW}}(x) - q_\tau(x) - B_\tau(x) + o_p(h^2)] \xrightarrow{d} N(0, \sigma_\tau^2(x)) \tag{52}$$

where the bias and variance are given, respectively, by:

$$B_\tau(x) = -\frac{B_f(q_\tau(x)|x)}{f(q_\tau(x)|x)} \quad \text{and} \quad \sigma_\tau^2(x) = \frac{\sigma_f^2(q_\tau(x)|x)}{f^2(q_\tau(x)|x)} = \frac{v_0(K)p[1-p]}{f^2(q_\tau(x)|x)f_1(x)} \tag{53}$$

where $f(y|x)$ is the conditional density of $Y_t = y$ given $X_t = x$.

It is clear that for given x, $\widehat{F}_{\mathrm{WNW}}(y|x)$ is not a continuous function of y. It might cause the computational trouble when computing the estimated conditional quantile $\widehat{q}_{\tau,\mathrm{WNW}}(x)$ by Eq. (51). To overcome this shortcoming, Cai and Wang (2008) proposed a WDK estimator (see below), which indeed is differentiable with respect to y. Cai and Wang (2008) showed that the differentiability of the estimated conditional distribution function cannot only make the asymptotic analysis much easier for the nonparametric estimators of quantile regression, but also can reduce the asymptotic variance (or asymptotic mean squared error) in a higher order sense. The main idea of Cai and Wang (2008) is described as follows.

It is noted for a given symmetric kernel $g(\cdot)$, where $G(\cdot)$ is the distribution function of $g(\cdot)$, as $h_0 \to 0$,

$$E\{G_{h_0}(y - Y_t)|X_t = x\} = F(y|x) + \frac{h_0^2}{2}\mu_2(g)F^{0,2}(y|x) + o(h_0^2) \to F(y|x) \tag{54}$$

where $G_{h_0}(u) = G(u/h_0)/h_0$. The above convergence ignores the higher terms $o(h_0^2)$ since $h_0 = o(h)$, where h is the smoothing bandwidth in the x direction (see Eq. (55) below). We can see that $Y_t^*(y) = G_{h_0}(y - Y_t)$ can be regarded as an initial estimate of $F(y|x)$ smoothing in the y direction. Thus, the left-hand side of Eq. (54) can be regraded as a nonparametric regression of the observed variable $Y_t^*(y)$ versus X_t and the LL (or polynomial) fitting scheme can be applied here. This leads to the locally weighted least squares regression problem:

$$\sum_{t=1}^{n}\{Y_t^*(y) - a - b(X_t - x)\}^2 K_h(x - X_t) \tag{55}$$

Note that Eq. (55) involves two kernels $g(\cdot)$ and $K(\cdot)$ and two bandwidths h_0 and h. This is the reason for calling it "double kernel."

Minimizing Eq. (55) with respect to a and b, we obtain the locally weighted least squares estimator of $F(y|x)$, which is $\widehat{a}$. It is easy to see that this estimator can be reexpressed as a linear estimator as,

$$\widehat{F}_{\mathrm{LL}}(y|x) = \sum_{t=1}^{n} W_{\mathrm{LL},t}(x,h) G_{h_0}(y - Y_t) \tag{56}$$

where with $S_{n,j}(x) = \sum_{t=1}^{n} K_h(x - X_t)(X_t - x)^j$, the weights $\{W_{\mathrm{LL},t}(x,h)\}$ are given by,

$$W_{\mathrm{LL},t}(x,h) = [S_{n,2}(x) - (x - X_t)S_{n,1}(x)]K_h(x - X_t)[S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)]^{-1}$$

Clearly, $\{W_{\mathrm{LL},t}(x,\ h)\}$ satisfy the discrete moments conditions given in Eq. (46). $\widehat{F}_{\mathrm{LL}}(y|x)$ is the so-called Yu and Jones estimator. Yu and Jones (1998) studied the asymptotic properties of $\widehat{F}_{\mathrm{LL}}(y|x)$ for i.i.d. data, which are similar to those given in Eqs. (48) and (49) if $h_0 = o(h)$.

Remark 3. If the bandwidth at the initial step h_0 is not undersmoothed, say $h_0 = O(h)$, then there is an extra term in the asymptotic bias and it is given by $\mu_2(g)(h_0^2/2)F^{0,2}(y|x)$, which is carried over from the initial estimation.

Also, Yu and Jones (1998) considered the nonparametric estimate of $q_\tau(x)$ based on $\widehat{F}_{\mathrm{LL}}(y|x)$, which is defined as,

$$\widehat{q}_{\tau,\mathrm{LL}}(x) = \widehat{F}_{\mathrm{LL}}^{-1}(\tau|x) \tag{57}$$

and they derived the asymptotic properties of $\widehat{q}_{\tau,\mathrm{LL}}(x)$, which is the exactly same as that given in Eq. (52). Further, Yu and Jones (1998) proposed an ad hoc method to adaptively select the optimal bandwidths h_0 and h. Clearly, $\widehat{F}_{\mathrm{LL}}(y|x)$ may not be constrained either to lie between zero and one or monotone increasing. To overcome this difficulty, some modifications in implementation of $\widehat{q}_{\tau,\mathrm{LL}}(x)$ were addressed in Yu and Jones (1998).

To accommodate all of the above attractive properties (monotonicity, continuity, differentiability, lying between zero and one, design adaption, avoiding boundary effects, and mathematical efficiency) of both estimators $\widehat{F}_{\mathrm{LL}}(y|x)$ and $\hat{F}_{\mathrm{WNW}}(y|x)$ under a unified framework, Cai and Wang (2008) proposed the following nonparametric estimator for conditional distribution $F(y|x)$, termed as WDK estimation,

$$\widehat{F}_{\mathrm{WDK}}(y|x) = \sum_{t=1}^{n} W_{\mathrm{WDK},t}(x,h) G_{h_0}(y - Y_t) \tag{58}$$

where

$$W_{\mathrm{WDK},t}(x,h) = p_t(x)W_h(x-X_t)\left[\sum_{t=1}^{n} p_t(x)W_h(x-X_t)\right]^{-1}$$

and $\{p_t(x)\}$ is chosen to be $p_t(x) = n^{-1}\{1+\lambda(X_t-x)W_h(x-X_t)\}^{-1} \geq 0$ to satisfy Eq. (46). Here λ is a function of the data and x and is uniquely defined by Eq. (47). Cai and Wang (2008) showed that the asymptotic properties for $\widehat{F}_{\mathrm{WDK}}(y|x)$ are similar to those given in Eqs. (48) and (49) if $h_0 = o(h)$. Note that this undersmoothing at the initial step is needed (see Remark 3).

Moreover, Cai and Wang (2008) considered the nonparametric estimate of $q_\tau(x)$ based on $\widehat{F}_{\mathrm{WDK}}(y|x)$, which is defined as,

$$\widehat{q}_{\tau,\mathrm{WDK}}(x) = \widehat{F}^{-1}_{\mathrm{WDK}}(\tau|x) \tag{59}$$

Note that $\widehat{q}_{\tau,\mathrm{WDK}}(x)$ always exists in finite samples and is uniquely determined since $\widehat{F}_{\mathrm{WDK}}(y|x)$ is a continuous distribution function. Cai and Wang (2008) also showed that $\widehat{q}_{\tau,\mathrm{WDK}}(x)$ has the exactly same asymptotic behavior as that given in Eq. (52). In addition, Cai and Wang (2008) proposed an ad hoc data-driven bandwidth selection method based on the nonparametric version of the AIC.

Finally, Yu and Jones (1998), Cai (2002b) and Cai and Wang (2008) discussed the asymptotic behavior of their nonparametric estimators $\widehat{q}_{\tau,\mathrm{LL}}(x), \widehat{q}_{\tau,\mathrm{WNW}}(x)$ and $\widehat{q}_{\tau,\mathrm{WDK}}(x)$ at boundaries and the result shows that all estimators have the exactly same asymptotic bias and do not have boundary effect; see Yu and Jones (1998), Cai (2002b) and Cai and Wang (2008) for details.

Cai and Wang (2008) considered a real data set on Dow Jones Industrials (DJI) index returns and applied the proposed method to estimate the 5% CVaR and CES functions. Both the CVaR and CES estimates exhibit a U-shape, which corresponds to the so-called ''volatility smile.'' Therefore, the risk tends to be lower when the lagged log loss of DJI is close to the empirical average, and larger otherwise. We can also observe the curves are asymmetric. This may indicate that the DJI index is more likely to fall if there were a loss within the last day than if there was a same amount of positive return.

5.2. Loss Function Approaches

Based on Eq. (45), if $q_\tau(x) = \beta_\tau^{\mathrm{T}} x$ is linear in x, then, one can find the estimate of β_τ by,

$$\widehat{\beta}_\tau = \mathrm{argmin}_{\beta_\tau} \sum_{t=1}^{n} \rho_\tau(Y_t - \beta_\tau^{\mathrm{T}} x) \tag{60}$$

see Koenker and Bassett (1978, 1982) for details.

To compute $\widehat{\beta}_\tau$ in Eq. (60), it can be implemented by using the function *rq*( ) in the package *quantreg* in the computing language *R*, due to Koenker (2004).

If $q_\tau(x)$ is a nonparametric function, there are several methods proposed in the literature to estimate $q_\tau(x)$, we describe some of them below.

5.2.1. Local Polynomial Methods

If $q_\tau(x)$ is assumed to have continuous $(m+1)$th order partial derivative, for X_t in a neighborhood of x, $q_\tau(X_t)$ can be approximated by $\sum_{j=0}^{m} \theta_j (X_t - x)^j$ where $\theta_j = (1/j!)\partial^j q_\tau(x)/\partial x^j$ is the jth partial derivative of $q_\tau(x)$. Then, we can use the following locally weighted loss function, which is a locally weighted version of Eq. (60),

$$\widehat{\theta} = \mathrm{argmin}_\theta \sum_{t=1}^{n} \rho_\tau \left(Y_t - \sum_{j=0}^{m} \theta_j (X_t - x)^j \right) K_h(x - X_t) \tag{61}$$

to obtain the local polynomial estimation of quantile function. Clearly, $\widehat{q}_\tau(x) = \widehat{\theta}_0$ estimates the quantile function and $\widehat{q}_\tau^{(j)}(x) = j!\widehat{\theta}_j$ estimates the jth partial derivative. Note that formula (61) has been addressed (essentially) by Chaudhuri (1991), Fan et al. (1994), Koenker et al. (1992), Yu and Jones (1998) for i.i.d. sample and Honda (2000) and Cai and Ould-Said (2003) for time series.

To compute $\widehat{q}_\tau(x)$ and $\widehat{q}_\tau^{(j)}(x)$, one also can use the function *rq*( ) by setting covariates as $X_t - x, \ldots, (X_t - x)^m$, and the weight as $K_h(X_t - x)$. Alternatively, one can use the function *lprq*( ) in the same package.

By using the series expansion method, Chaudhuri (1991) was the first to obtain the local Bahadur type representation of parameter's estimators so that one can easily derive some asymptotic results. Honda (2000) generalized these results to the α-mixing process by using local polynomial fitting, and obtained the similar asymptotic results. To derive the asymptotic properties, Honda (2000) and Cai and Xu (2008) gave the local Bahadur

representation for $\widehat{q}_\tau(x)$ for univariate case ($d = 1$). That is, they showed that under some regular conditions, the LL ($m = 1$) quantile estimator $\widehat{q}_\tau(x)$ has the following representation,

$$\sqrt{nh}[\widehat{q}_\tau(x) - q_\tau(x)] = \frac{1}{f_{y|x}(q_\tau(x)|x) f_1(x)\sqrt{nh}} \sum_{t=1}^{n} \psi_\tau(Y_t^*) K\left(\frac{(X_t - x)}{h}\right) + o_p(1) \tag{62}$$

where $\psi_\tau(x) = \tau - I_{x<0}$ and $Y_t^* = Y_t - q_\tau(x) - q'_\tau(x)(X_t - x_0)$. Therefore, one can easily obtain the asymptotic normality as,

$$\sqrt{nh}\left[\widehat{q}_\tau(x) - q_\tau(x) - \frac{h^2}{2}\mu_2(K) q''_\tau(x) + o_p(h^2)\right] \xrightarrow{d} N(0, \sigma_\tau^2(x)) \tag{63}$$

where $\sigma_\tau^2(x)$ is given in Eq. (53). Clearly, a comparison of Eqs. (52) and (63) leads to conclusions that the LL quantile estimator $\widehat{q}_\tau(x)$ and three direct estimators share the exactly same asymptotic variance, but the biases are quite different. Indeed, the bias term in Eq. (52) (see also Eq. (53)), the quantity $-F^{2,0}(q_\tau(x)|x)/f(q_\tau(x)|x)$, involving the second derivative of the conditional distribution function, is replaced by $q''_\tau(x)$, the second derivative of the conditional quantile function itself. This is not surprising since for the direct methods, the approximation is applied to the conditional distribution function, while for LL quantile estimator $\hat{q}_\tau(x)$, the approximation is applied to the conditional quantile function itself.

5.2.2. Spline Approaches

In the 1990s, there were many research papers on nonparametric estimation of quantile regression using various splines methods such as smoothing splines and B-splines. For example, for a single covariate, He and Shi (1994) used quantile regression B-splines and considered the convergence with a rate of B-splines estimator, while Koenker et al. (1994) suggested quantile smoothing splines. In bivariate smoothing, He et al. (1998) considered bivariate quantile smoothing splines that belong to the space of bilinear tensor product splines, while Portnoy (1997) and He and Portnoy (2000) provided the asymptotic properties of these bivariate quantile splines estimators. The optimality properties of the splines provide justification for their use in nonparametric quantile function estimation, and the optimization problems can be solved efficiently as linear programs. He and Ng (1999) considered a general additive (several covariates) model with univariate linear splines capturing the main effects and bilinear tensor product splines capturing the second-order interactions. But all splines methods encounter

the same difficulties that it is not easy to derive the asymptotic properties like asymptotic normality and to make statistical inferences (see Remark 5 later for more discussions), although they might be attractive in applications.

We now begin by briefly reviewing the smoothing splines technique; see the aforementioned papers for details. For a univariate design variable X_t with observed response Y_t, the τth quantile smoothing spline function $q_\tau(x)$ minimizes over

$$\sum_{t=1}^{n} \rho_\tau(Y_t - q_\tau(X_t)) + \lambda V(q'_\tau) \tag{64}$$

where $V(h) = \sup \sum_{j=1}^{k} |h(x_j) - h(x_{j-1})|$ denotes the total variation of the function $h(\cdot)$ with the supremum being taken over all finite partitions $x_0 < x_1 < \cdots < x_k$ of the support of $h(\cdot)$. If $h(\cdot)$ is differentiable, it is easy to see that

$$V(h) = \int_0^1 |h'(x)| dx, \quad \text{if the support of } h(\cdot) \text{ is } [0, 1]$$

The optimal solution $\widehat{q}_\tau(x)$ estimates the τth conditional quantile function $q_\tau(x)$. The problem of quantile smoothing in expression (64) can be viewed as a special case ($p = 1$) of the following general form of quantile smoothing

$$\sum_{t=1}^{n} \rho_\tau(Y_t - q_\tau(X_t)) + \lambda \left(\int |q''_\tau(x)|^p dx \right)^{1/p} \tag{65}$$

for $p \geq 1$. If $p = 2$ in Eq. (65), the solution to expression (65) is a natural cubic smoothing spline with knots at the observed design points. Its computation is rather efficient as it simply amounts to solving a linear system. The solution to expression (64) is a linear smoothing spline with possible breaks in the derivative at the design points, and the computation can be performed by modern linear programming methods. See the forgoing papers for the computational issue. As for selecting the smoothing parameter λ, the SIC is commonly suggested in the smoothing spline literature; see Koenker et al. (1994) and He and Ng (1999) for details. But it is well known that the SIC is overfitting due to the heavy penalty (see Eq. (66) later) when the sample is large.

Remark 4. As commented by He et al. (1998), generalization of smoothing splines to bivariate or multivariate cases is not always straightforward. The form of the solution often depends on the roughness penalty used in the optimization process and it is quite complex. Due to

the complicated notation, we ignore the presentation of smoothing splines for multivariate case. Instead, we refer the reader to the papers by He and Shi (1994), He et al. (1998), He and Ng (1999), and He and Portnoy (2000) for the detailed discussions.

Remark 5. It is well known in the splines literature; see the previously mentioned papers, that the rate of convergence for the nonparametric estimates depends mainly on two aspects: the smoothness of the function being estimated and the dimensionality of the spline space or, equivalently, the number of knots. These issues are still valid for the conditional quantile smoothing splines estimates. The asymptotic behavior such as the rate of convergence for the quantile smoothing splines is rather difficult to analyze, especially when a data-driven smoothing parameter is used. In the univariate case when the smoothing parameter is not data-driven, Portnoy (1997) derived some local asymptotic properties of the quantile smoothing splines, while He and Ng (1999) and He and Portnoy (2000) presented the asymptotic mean square error for bivariate and multivariate cases. Unfortunately, the asymptotic normality of a quantile spline (smoothing spline or B-spline) estimator for the data-driven smoothing parameters is still open and it is warranted as a future research topic.

A B-spline approach can be formulated as follows. It is well known that a B-spline approach depends on the degree of smoothness of the true quantile function, which determines how well the quantile function can be approximated. Therefore, it is commonly assumed that the quantile function with a certain degree of smoothness r defined as follows. To this end, define a functional space $\mathcal{Q}_r$ to be the collection of all functions on a domain, say [0, 1] for which the mth order derivative satisfies the Hölder condition of order of γ with $r = m+\gamma$. That is, for each $h \in \mathcal{Q}_r, |h^{(m)}(s) - h^{(m)}(t)| \leq W_0|s-t|^{\gamma}$ for any $0 \leq s, t \leq 1$ and a positive finite constant W_0.

Here we first assume that the quantile regression function $q_\tau(x)$ is from $\mathcal{Q}_r$ and then, we can define B-splines of order $m+1$ used to approximate the quantile function $q_\tau(\cdot)$. We consider a sequence of positive integers $\{k_n\}$, $n \geq 1$, (the number of knots) and an extended partition of [0, 1] by k_n knots with equal or unequal length. Then, we can define the associated B-spline basis functions by $\{B_j(x)\}, 1 \leq j \leq k_n + m + 1$; see Schumaker (1981) for details. The proposed B-spline estimator of $q_\tau(x)$ is given by,

$$\widehat{q}_\tau(x) = \sum_{j=1}^{k_n+m+1} \widehat{\theta}_j B_j(x)$$

where $\widehat{\theta}_j$ solves the minimization problem

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - \sum_{j=1}^{k_n+m+1} \theta_j B_j(X_t) \right)$$

Clearly, when the B-spline basis is given, computations can be easily carried using standard quantile regression algorithms as in Eq. (60). As for selecting the order and knots for the splines, the SIC is commonly suggested in the B-spline literature; see He and Shi (1994) and Kim (2007).

5.2.3. Smoothing Parameter Selection

It is well known that the smoothing tuning parameter η ($\eta = h$ for kernel smoothing and $\eta = \lambda$ for smoothing spline) plays an essential role in the trade-off between reducing bias and variance. To the best of our knowledge, there has been very limited literature about selecting η in the context of estimating the quantile regression even though there is a rich amount of literature on this issue in the mean regression setting; see, for example, Cai et al. (2000) and Cai and Tiwari (2000). Indeed, Yu and Jones (1998) or Yu and Lu (2004) proposed a simple and convenient method for the nonparametric quantile estimation. Their approach assumes that the second derivatives of the quantile function are parallel. However, this assumption might not be valid for many applications due to (nonlinear) heteroscedasticity. Further, the mean regression approach cannot directly estimate the variance function. To attenuate these problems, Cai and Xu (2008) proposed a method of selecting bandwidth for the foregoing estimation procedure, based on the nonparametric version of the AIC, which can attend to the structure of time series data and the overfitting or underfitting tendency. The basic idea is motivated by its analogue of Cai and Tiwari (2000) for nonlinear mean regression for time series models and we briefly describe it below.

By recalling the classical AIC for linear models under the likelihood setting; that is the negative of twice of the maximized log likelihood plus twice of the number of estimated parameters, Cai and Xu (2008) proposed the following nonparametric version of the bias-corrected AIC; see Hurvich et al. (1998) and Cai and Tiwari (2000) for nonparametric regression models, to select η by minimizing

$$\mathrm{AIC}(\eta) = \log\{\widehat{\sigma}_\eta^2\} + \frac{2(p_\eta + 1)}{[n - (p_\eta + 2)]} \tag{66}$$

where $\widehat{\sigma}_\eta^2 = n^{-1}\sum_{t=1}^{n} \rho_\tau(Y_t - \widehat{q}_\tau(X_t))$ and p_η is the nonparametric version of degrees of freedom, called the effective number of parameters. This criterion

may be interpreted as the AIC for the local quantile smoothing problem and seems to perform well in some limited applications. Note that similar to Eq. (66), Koenker et al. (1994) considered the SIC with the second term on the right-hand side of Eq. (66) replayed by $2n^{-1}\, p_\lambda \log n$, where p_λ is the number of "active knots" for the smoothing spline quantile setting.

For different smoothing techniques, the choice of p_η might be different. For example, see Koenker et al. (1994) on how to choose $p_\eta = p_\lambda$ in quantile smoothing splines setting and Cai and Xu (2008) for how to determine $p_\eta = p_h$ under kernel smoothing framework.

5.2.4. Dimension Reduction Modeling

As mentioned earlier, a purely nonparametric quantile regression model may suffer from the so-called "curse of dimensionality" problem. To overcome this difficulty, some dimension reduction modeling methods have been proposed in the literature such as additive and varying-coefficient models, discussed next.

5.2.4.1. Additive Models. An additive quantile regression model takes a form as,

$$q_\tau(x) = \delta + \sum_{j=1}^{d} q_{\tau,j}(x_j) \tag{67}$$

which was studied by De Gooijer and Zerom (2003), Yu and Lu (2004), and Horowitz and Lee (2005). For ease of notation, assume that $d = 2$ in what follows. De Gooijer and Zerom (2003) used a two-stage approach to estimate each component in Eq. (67) as follows. First, estimate the d-dimensional quantile regression surface $g_\tau(x)$ using Eq. (51) to obtain $\widehat{q}_{\tau,\mathrm{WNW}}(x)$ and then use the projection method of Cai and Masry (2000) as,

$$\widehat{q}_{\tau,1}(x_1) = \frac{1}{n}\sum_{t=1}^{n} \widehat{q}_{\tau,\mathrm{WNW}}(x_1, X_{t2}) W(x_1, X_{t2})$$

where $W(\cdot)$ is a weighting function, which can be chosen based on minimizing the asymptotic variance as in Cai and Fan (2000) to achieve the optimality or to screen out outliers. Similarly, one can estimate $\widehat{q}_{\tau,2}(x_2)$. De Gooijer and Zerom (2003) also presented the asymptotic normality of the proposed estimator.

Later, Yu and Lu (2004) proposed using a backfitting algorithm equipped with a LL fitting as follows.

1. Step (1), initial estimation. Set

$$\hat{\delta} = \text{argmin}_{\delta} \sum_{t=1}^{n} \rho_{\tau}(Y_t - \delta)$$

and, for $j = 1$ and 2,

$$(\widehat{a}_j, \widehat{b}_j) = \text{argmin}_{a,b} \sum_{t=1}^{n} \rho_{\tau}(Y_t - \widehat{\delta} - a - b(X_{tj} - x_j))K_{hj}(X_{tj} - x_j)$$

Then, set $q_{\tau,j}^{(0)}(x_j) = \widehat{a}_j$, and take $q_{\tau,j}^{*(0)}(x_j)$ as $q_{\tau,j}^{(0)}(x_j)$ minus the τth sample quantile of $\{q_{\tau,j}^{(0)}(X_{tj})\}_{t=1}^{n}$.

2. Step (2), iteration. Set

$$\widehat{\delta}^{(l)} = \text{argmin}_{\delta} \sum_{t=1}^{n} \rho_{\tau}(Y_t - q_{\tau,1}^{*(l-1)}(X_{t1}) - q_{\tau,2}^{*(l-1)}(X_{t2}) - \delta)$$

and for $j = 1$ and 2 and $m = 3-j$,

$$(\widehat{a}_j, \widehat{b}_j) = \text{argmin}_{a,b} \sum_{t=1}^{n} \rho_{\tau}(Y_t - \widehat{\delta}^{(l)} - q_{\tau,m}^{*(l-1)}(X_{tm}) - a - b(X_{tj} - x_j))K_{hj}(X_{tj} - x_j)$$

then take $q_{\tau,j}^{(l)}(x_j) = \widehat{a}_j$, and take $q_{\tau,j}^{*(l)}(x_j)$ as $q_{\tau,j}^{(l)}(x_j)$ minus the τth sample quantile of $\{q_{\tau,j}^{(l)}(X_{t,j})\}_{t=1}^{n}$.

3. Step (3), keep cycling step (2) for $l = 1, 2, 3, \ldots$ until the value of $q_{\tau}^{*(l)} = (\widehat{\delta}^{(l)}, q_{\tau,1}^{*(l)}, q_{\tau,2}^{*(l)})$ has converged. Next, for $j = 1$ and 2, let $(\widehat{a}_j, \widehat{b}_j) = (q_{\tau,j}^{*(l)}(x_j), \widehat{b}_j)$. Then, $(\widehat{a}_j, \widehat{b}_j)$ gives the estimators of $q_{\tau,j}(x_j)$ and $q'_{\tau,j}(x_j)$, respectively.

Further, Yu and Lu (2004) investigated the large sample behavior of the proposed backfitting estimator.

Recently, Horowitz and Lee (2005) used a two-stage approach which is different from that in De Gooijer and Zerom (2003). At the first stage, use a series approximation to each component as $q_{\tau,j}(x_j) \approx \sum_{l=0}^{kj} \theta_{lj}\phi_{jl}(x_j)$, where $\{\phi_{jl}(\cdot)\}$ is a basis function, and then estimate θ_{lj} by,

$$\text{argmin}_{\delta,\theta} \sum_{t=1}^{n} \rho_{\tau}\left(Y_t - \delta \sum_{j=1}^{2} \sum_{l=0}^{kj} \theta_{lj}\phi_{jl}(x_j)\right)$$

denoted by $\widehat{\theta}_{lj}$, to obtain

$$\widehat{q}^{(0)}_{\tau,j}(x_j) = \sum_{l=0}^{kj} \widehat{\theta}_{lj}\phi_{jl}(x_j)$$

At the second stage, estimate $q_{\tau,j}(x_j)$ by first finding

$$(\widehat{a}_j, \widehat{b}_j) = \text{argmin}_{a,b} \sum_{t=1}^{n} \rho_\tau(Y_t - \widehat{\delta} - \widehat{q}^{(0)}_{\tau,m}(X_{tm}) - a - b(X_{tj} - x_j))K_{hj}(X_{tj} - x_j)$$

and then taking $\widehat{q}_{\tau,j}(x_j) = \widehat{a}_j$. Also, Horowitz and Lee (2005) derived the asymptotic properties for the proposed two-stage estimator.

5.2.4.2. Varying-Coefficient Models. A varying-coefficient quantile regression model takes a form as,

$$q_\tau(u, x) = \sum_{j=1}^{d} a_{\tau,j}(u)x_j = a_\tau(u)^{\mathrm{T}}x \tag{68}$$

which was studied by Honda (2004) for i.i.d. data, Cai and Xu (2008) for dynamic time series observations, and Kim (2007) for time-varying coefficients (u is time) for i.i.d. samples. For easy exposition, we assume that u is univariate below.

To estimate $\{a_k(\cdot)\}$ using the local polynomial method based on $\{U_t, X_t, Y_t)\}_{t=1}^n$, assume that the coefficient functions $\{a(\cdot)\}$ have the $(m+1)$th derivative ($m \geq 1$), so that for any given gird point $u \in \Re, a_k(\cdot)$ can be approximated by a polynomial function in a neighborhood of the given grid point u as $a(U_t) \approx \sum_{j=0}^{m} \beta_j(U_t - u)^j$, where $\beta_j = a^{(j)}(u)/j!$ and $a^{(j)}(u)$ is the jth derivative of $a(u)$, so that $q_\tau(U_t, X_t) \approx \sum_{j=0}^{m} X_t^T \beta_j(U_t - u)^j$. Then, the locally weighted loss function is

$$\sum_{t=1}^{n} \rho_\tau\left(Y_t - \sum_{j=0}^{m} X_t^T \beta_j(U_t - u)^j\right) K_h(U_t - u) \tag{69}$$

Solving the minimization problem in Eq. (69) gives $\widehat{a}(u) = \widehat{\beta}_0$, the local polynomial estimate of $a(u)$, and $\widehat{a}^{(j)}(u) = j!\widehat{\beta}_j (j \geq 1)$, the local polynomial estimate of the jth derivative $a^{(j)}(u)$. By moving u along with the real line, the estimate of the entire curve $\widehat{a}(u)$ is obtained.

Cai and Xu (2008) derived the asymptotic properties for $\widehat{a}(u)$. Under some regularity conditions, we have the following asymptotic normality

for m odd,

$$\sqrt{nh}\left[\widehat{a}(u) - a(u) - \frac{h^{m+1}}{(m+1)!}a^{(m+1)}(u)\mu_{m+1}(K) + o_p(h^{m+1})\right] \xrightarrow{d} N(0, \Sigma_a(u))$$

where $\Sigma_a(u) = \tau(1-\tau)\Sigma(u)$, $\Sigma(u) = [\Omega^*(u)]^{-1}\Omega(u)[\Omega^*(u)]^{-1}/f_u(u)$, $\Omega(u) = E[X_t X_t^T | U_t = u]$, $\Omega^*(u) = E[X_t X_t^T f_{y|u,x}(q_\tau(u, X_t))|U_t = u]$, $f_u(\cdot)$ is the marginal density of U_t, and $f_{y|u,x}(y)$ is the conditional density of Y_t given U_t and X_t. Also, Cai and Xu (2008) proposed an ad hoc bandwidth selection method that is similar to that described in Section 5.2.3.

Finally, Kim (2007) considered the time-varying coefficient quantile regression model as,

$$q_\tau(t, x) = \sum_{j=1}^{d} a_{\tau,j}(t)x_j = a_\tau(t)^{\mathrm{T}}x \tag{70}$$

and used a B-spline technique to estimate $a_\tau(t)$. Note that model (70) might be potentially useful to see whether the quantile regression changes over time and in a case with a practical interest is, for example, the analysis of the reference growth data by Cole (1994), Wei et al. (2006), and Wei and He (2006) for longitudinal data, and Kim (2007) for i.i.d. samples. Finally, it is worth to point out that model (70) might be very useful for a nonparametric testing for testing structural changes in regression quantiles as in Qu (2008).

Cai and Xu (2008) used model (68) and its modeling approaches to explore the possible nonlinearity feature, heteroscedasticity, and predictability of the exchange rate series of the Japanese Yen in terms of the U.S. dollar. Their empirical findings are that the quantile has a complex structure and that both heteroscedasticity and nonlinearity exist. This implies that the GARCH effects occur in the exchange rate time series. Finally, they considered the one-step ahead post-sample forecasting for the last 25 observations and constructed the 95% nonparametric prediction interval $(\widehat{q}_{0.025}(\cdot), \widehat{q}_{0.975}(\cdot))$ based on the past two lags. It turns out that 24 of 25 predictive intervals contain the corresponding true values. This means that under the dynamic smooth coefficient quantile regression model assumption, the prediction intervals based on the proposed method work reasonably well.

6. CONCLUSION

In this paper, we survey some recent developments in nonparametric econometrics, including (i) nonparametric estimation and testing of regression functions with mixed discrete and continuous covariates; (ii) nonparametric estimation/testing with nonstationary data; (iii) nonparametric IV estimations; and (iv) nonparametric estimation of quantile regression models.

In the paper by Cai and Hong (2009), they gave a survey on the recent developments of nonparametric estimation and testing of financial econometric models. Due to space limitation, we omit some of the important areas such as nonparametric/semiparametric with limited dependent variable models and nonparametric/semiparametric panel data models. Another promising line of research is to impose less restrictions on econometric models and hence parameters may not be point identified but are set identified. Readers interested in these areas of research should consult with the works by Manski (2003), Imbens and Manski (2004), Honore and Tamer (2006), and the references therein.

NOTES

1. This argument may not be always true as one can also choose a fixed value of h in testing problems, resulting in a non-smoothing test, see Chapter 13 of Li and Racine (2007) on more detailed discussions of non-smoothing tests.
2. Fan and Li (1996) proposed a nonparametric significance test. Gu, Li, and Liu (2007) showed that a residual-based bootstrap method can be used to better approximate the null distribution of Fan and Li's test.
3. This independence assumption can be relaxed to $E(u_t|X_t, Z_t) = 0$, which leads to some modification to the asymptotic theory.

ACKNOWLEDGMENTS

We thank the referees for their careful reading of the manuscript and for their helpful comments. We also thank the Econometric Seminar audiences at the University of Guelph, and the participants in the Advances in Econometrics conference for helpful comments on this paper. Cai's research was supported, in part, by the National Science Foundation grant no. DMS-0404954, funds provided by the University of North Carolina at Charlotte, the Cheung Kong Scholarship from Chinese Ministry of Education, the Minjiang Scholarship from Fujian Province, China and Xiamen University, and the National Nature Science Foundation of China grant no. 70871 003.

Li's research is partially supported by the National Nature Science Foundation of China grant no. 70773005.

REFERENCES

Ai, C., & Chen, X. (2003). Efficient estimation of models with conditional moment restrictions conditioning unknown functions. *Econometrica*, *71*, 1795–1843.

Aitchison, J., & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, *63*, 413–420.

Bachmeier, L., Leelahanon, S., & Li, Q. (2006). Money growth and inflation in the United States. *Macroeconomic Dynamics*, *11*, 113–127.

Blundell, R., & Powell, J. (2003). Endogeneity in nonparametric and semiparametric regression models. In: M. Dewatripont, L. P. Hansen & S. J. Turnovsky (Eds), *Advances in economics and econometrics: Theory and applications, eighth world congress* (Vol. II). Cambridge: Cambridge University Press.

Cai, Z. (2002a). Two-step likelihood estimation procedure for varying-coefficient models. *Journal of Multivariate Analysis*, *81*, 189–209.

Cai, Z. (2002b). Regression quantile for time series. *Econometric Theory*, *18*, 169–192.

Cai, Z. (2002c). A two-stage approach to additive time series models. *Statistica Neerlandica*, *56*, 415–433.

Cai, Z., Das, M., Xiong, H., & Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics*, *133*, 207–241.

Cai, Z., & Fan, J. (2000). Average regression surface for dependent data. *Journal of Multivariate Analysis*, *75*, 112–142.

Cai, Z., Fan, J., & Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, *95*, 888–902.

Cai, Z., Fan, J., & Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, *95*, 941–956.

Cai, Z., & Hong, Y. (2009). Some recent developments in nonparametric finance. In: Q. Li & J. Racine (Eds), *Advances in Econometrics* (pp. 379–432). Bingley, UK: Emerald.

Cai, Z., Li, Q., & Park, J. (2009). Functional-coefficient models for nonstationary time series data. *Journal of Econometrics*, *148*, 101–113.

Cai, Z., & Masry, E. (2000). Nonparametric estimation in nonlinear ARX time series models: Projection and linear fitting. *Econometric Theory*, *16*, 465–501.

Cai, Z., & Ould-Said, E. (2003). Local M-estimator for nonparametric time series. *Statistics and Probability Letters*, *65*, 433–449.

Cai, Z., & Tiwari, R. C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, *11*, 341–350.

Cai, Z., & Wang, X. (2008). Nonparametric methods for estimating conditional value-at-risk and expected shortfall. *Journal of Econometrics*, *147*, 120–130.

Cai, Z., & Wang, Y. (2009). *Instability of predictability of assets returns*. Working Paper no. 2009–11. Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC.

Cai, Z., & Xiong, H. (2006). *Partially varying coefficient instrumental variable models*. Working Paper no. 2006–9. Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC.

Cai, Z., & Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, *103*, 1595–1608.

Caner, M., & Hansen, B. E. (2004). Instrumental variable estimation of a threshold model. *Econometric Theory*, *20*, 813–843.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*, 1127–1160.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics*, *19*, 760–777.

Chaudhuri, P., Doksum, K., & Samarov, A. (1997). On average derivative quantile regression. *The Annals of Statistics*, *25*, 715–744.

Cole, T. J. (1994). Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine*, *13*, 2477–2492.

Daroles, S., Florens, J. -P., & Renault, E. (2002). *Nonparametric instrumental regression*. Working Paper, GREMAQ, University of Social Science, Toulouse, France.

Das, M. (2003). Identification and sequential estimation of nonparametric panel models with insufficient exclusion restrictions. *Journal of Econometrics*, *114*, 297–328.

Das, M. (2005). Instrumental variables estimators for nonparametric models with discrete endogenous regressors. *Journal of Econometrics*, *124*, 335–361.

Das, M., Newey, W., & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, *70*, 33–58.

De Gooijer, J., & Zerom, D. (2003). On additive conditional quantiles with high dimensional covariates. *Journal of the American Statistical Association*, *98*, 135–146.

Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantile. *Journal of Business and Economics Statistics*, *22*, 367–381.

Fan, J., & Gijbels, I. (1996). *Local polynomial modeling and its applications*. London: Chapman and Hall.

Fan, J., Hu, T.-C., & Troung, Y. K. (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics*, *21*, 433–446.

Fan, J., & Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, *11*, 1031–1057.

Fan, Y., & Li, Q. (1996). Consistent model specification tests: Omitted variables and semi-parametric functional forms. *Econometrica*, *64*, 865–890.

Gao, J., King, M., Lu, Z., & Tjøstheim, D. (2008). *Nonparametric specification testing for nonlinear time series with nonstationarity*. Working Paper. Department of Economics, University of Adelaide, SA, Australia.

Gu, J., & Hernandez-Verme, P. (2009). *An empirical evaluation of the presence of credit rationing in the U.S. credit markets*. Working Paper. Department of Economics, Texas A&M University, College Station, TX.

Gu, J., Li, D., & Liu, D. (2007). A bootstrap nonparametric significance test. *Journal of Nonparametric Statistics*, *19*, 215–230.

Hall, P., & Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, *33*, 2904–2929.

Hall, P., Li, Q., & Racine, J. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economic and Statistics*, *89*, 784–789.

Hall, P., Racine, J., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, *99*, 1015–1026.
He, X., & Ng, P. (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference*, *75*, 343–352.
He, X., Ng, P., & Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society, Series B*, *60*, 537–550.
He, X., & Portnoy, S. (2000). Some asymptotic results on bivariate quantile splines. *Journal of Statistical Planning and Inference*, *91*, 341–349.
He, X., & Shi, P. (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, *3*, 299–308.
Honda, T. (2000). Nonparametric estimation of a conditional quantile for α-mixing processes. *Annals of the Institute of Statistical Mathematics*, *52*, 459–470.
Honda, T. (2004). Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inferences*, *121*, 113–125.
Honore, B., & Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, *74*, 611–630.
Horowitz, J. L. (2007). Asymptotic normality of a nonparametric instrument variables estimator. *International Economic Review*, *48*, 1329–1349.
Horowitz, J. L., & Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, *100*, 1238–1249.
Hsiao, C., Li, Q., & Racine, J. (2007). Consistent model specification tests with mixed discrete and continuous variables. *Journal of Econometrics*, *140*, 802–826.
Hurvich, C. M., Simonoff, J. S., & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, *60*, 271–293.
Imbens, G., & Manski, C. (2004). Confidence intervals for partially identified parameters. *Econometrica*, *72*, 1845–1857.
Juhl, T. (2005). Functional coefficient models under unit root behavior. *Econometrics Journal*, *8*, 197–213.
Karlsen, H. A., Myklebust, T., & Tjöstheim, D. (2007). Nonparametric estimation in a nonlinear cointegration type model. *The Annals of Statistics*, *35*, 252–299.
Khindanova, I. N., & Rachev, S. T. (2000). Value at risk: Recent advances. In: *Handbook on analytic-computational methods in applied mathematics* (pp. 801–858). Boca Raton, FL: CRC Press LLC.
Kim, M.-O. (2007). Quantile regression with varying coefficients. *The Annals of Statistics*, *35*, 92–108.
Koenker, R. (2004). Quantreg: An R package for quantile regression and related methods, from http://cran.r-project.org.
Koenker, R. (2005). *Quantile regression*. *Econometric society monograph series*. New York, NY: Cambridge University Press.
Koenker, R., & Bassett, G. W. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.
Koenker, R., & Bassett, G. W. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, *50*, 43–61.
Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*, 673–680.
Koenker, R., Portnoy, S., & Ng, P. (1992). Nonparametric estimation of conditional quantile functions. In: Y. Dodge (Ed.), *L_1-statistical analysis and related methods* (pp. 217–229). Amsterdam: Elsevier.

Li, C., Ouyang, D., & Racine, J. (2009). Nonparametric regression with weakly dependent data: The discrete and continuous regressor case. *Journal of Nonparametric Statistics*, *21*, 697–711.

Li, Q., Hsiao, C., & Zinn, J. (2003). Consistent model specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics*, *112*, 295–325.

Li, Q., Huang, C., Li, D., & Fu, T. (2002). Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics*, *20*, 412–422.

Li, Q., & Racine, J. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University.

Li, Q., & Racine, J. (2009). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. Accepted by *Econometric Theory*.

Liang, Z., & Li, Q. (2009). *Functional coefficient regression model with time trend*. Working Paper. Department of Economics, Texas A&M University, College Station, TX.

Manski, C. (2003). *Partial identification of probability distributions*. New York: Springer-Verlag.

Masry, E., & Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, *13*, 214–252.

Newey, W. K., & Powell, J. L. (1988). *Instrumental variables estimation for nonparametric models*. Working Paper. Department of Economics, Princeton University, Princeton, NJ.

Newey, W. K., & Powell, J. L. (2003). Nonparametric instrumental variables estimation. *Econometrica*, *71*, 1565–1578.

Newey, W. K., Powell, J. L., & Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, *67*, 565–603.

Ouyang, D., Li, Q., & Racine, J. (2009). Nonparametric estimation of regression functions with discrete regressors. *Econometric Theory*, *25*, 1–42.

Pakes, A., & Olley, S. (1995). A limit theorem for a smooth class of semiparametric estimators. *Journal of Econometrics*, *65*, 295–332.

Park, J. Y., & Hahn, S. B. (1999). Cointegrating regressions with time varying coefficients. *Econometric Theory*, *15*, 664–703.

Park, S. (2003). Semiparametric instrumental variables estimation. *Journal of Econometrics*, *112*, 381–399.

Phillips, P. C. B., & Park, J. (1998). *Nonstationary density estimation and kernel autoregression*. Cowles Foundation Discussion Paper 1181, Department of Economics, Yale University, New Haven, CT.

Portnoy, S. (1997). Local asymptotics for quantile smoothing splines. *The Annals of Statistics*, *25*, 414–434.

Racine, J., Hart, J., & Li, Q. (2006). Testing the significance of categorical variables. *Econometric Reviews*, *25*, 523–544.

Racine, J., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*, 99–130.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, *92*, 1057–1062.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selection for local least squares regression. *Journal of the American Statistical Association*, *90*, 1257–1270.

Qu, Z. (2008). Testing for structural change in regression quantiles. *Journal of Econometrics*, *146*, 170–184.

Schultz, T.P. (1997). *Human capital, Schooling and health*. IUSSP, XXIII, General Population Conference, Yale University.

Schumaker, L. L. (1981). *Spline functions: Base theory*. New York: Wiley.

Speckman, P. (1988). Kernel smoothing partial linear models. *The Journal of Royal Statistical Society, Series B*, *50*, 413–426.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, *36*, 111–147.

Su, L., Chen, Y., & Ullah, A. (2009). Functional coefficient estimation with both categorical and continuous data. In: Q. Li & J. Racine (Eds), *Advances in Econometrics*. Bingley, UK: Emerald.

Su, L., & Ullah, A. (2008). Local polynomial estimation of nonparametric simultaneous equations models. *Journal of Econometrics*, *144*, 193–218.

Sun, Y., Cai, Z., & Li, Q. (2008a). *Consistent nonparametric test on parametric smooth coefficient model with nonstationary data*. Working Paper. Department of Economics, Texas A&M University, College Station, TX.

Sun, Y., Hsiao, C., & Li, Q. (2008b). *Volatility spillover effect: A semiparametric analysis of non-cointegrated processes*. Working Paper. Department of Economics, Texas A&M University, College Station, TX.

Sun, Y., & Li, Q. (2009a). *Data-driven method selecting smoothing parameters in semiparametric models with integrated time series data*. Working Paper. Department of Economics, Texas A&M University, College Station, TX.

Sun, Y., & Li, Q. (2009b). *Cointegration test on semiparametric smooth coefficient models*. Working Paper. Department of Economics, Texas A&M University, College Station, TX.

Wang, Q., & Phillips, P. C. B. (2008). *Structural nonparametric cointegrating regression*. Cowles Foundation Discussion Paper 1657, Department of Economics, Yale University, New Haven, CT.

Wang, Q., & Phillips, P. C. B. (2009). Asymptotic theory for local time density estimation and nonparametric cointegrating regression. *Econometric Theory*, *25*, 710–738.

Wei, Y., & He, X. (2006). Conditional growth charts (with discussion). *The Annals of Statistics*, *34*, 2069–2097.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, *25*, 1369–1382.

Xiao, Z. (2009). Functional coefficient co-integration models. *Journal of Econometrics*, *152*, 81–92.

Yu, K., & Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, *93*, 228–237.

Yu, K., & Lu, Z. (2004). Local linear additive quantile regression. *Scandinavian Journal of Statistics*, *31*, 333–346.

Zheng, J. X. (1996). A consistent test for functional form via nonparametric estimation techniques. *Journal of Econometrics*, *75*, 263–289.