# UC Riverside

## UC Riverside Electronic Theses and Dissertations

**Title**

Information and Inference in Econometrics: Estimation, Testing and Forecasting

**Permalink**

https://escholarship.org/uc/item/0601z077

**Author**

Tu, Yundong

**Publication Date**

2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

# Information and Inference in Econometrics: Estimation, Testing and Forecasting

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Yundong Tu

June 2012

Dissertation Committee:

Professor Tae-Hwy Lee, Co-Chairperson
Professor Aman Ullah, Co-Chairperson
Professor Marcelle Chauvet
Professor Robert Russell

Copyright by
Yundong Tu
2012

The Dissertation of Yundong Tu is approved:

---

---

---

Co-Chairperson

---

Co-Chairperson

University of California, Riverside

## Acknowledgments

I feel very grateful to my advisors, Professor Aman Ullah and Professor Tae-Hwy Lee, without whose help this thesis could not come into existence. I appreciate very much their valuable questions and advice provided during lectures, seminars, meetings, tea time and dinners. Their support in various ways encourages me to explore into the unknown areas and helps me to survive in the hard times. I benefited from the lectures and researches of Professor Robert Russell and Professor Marcelle Chauvet and I own thanks to them for serving in my committee. Special thanks go to Liangjun Su, whose kindly help and joint work is gratefully acknowledged.

I own a special debt to Xiaohong Chen and Joel Howrowitz, who inspired me on studies in Econometrics, Yixiao Sun and Hong Han, who encouraged me to achieve the best, and Yongmiao Hong, who offered me the opportunity to meet my advisors.

This thesis also benefited in numerious ways from discussions with Andress Santos, Raffaelle Giacomini, Halbert White, Jerry Hausman, Bruce Hansen, Mehmet Caner, Thanasis Stengos, Amos Golan, Victoria Zinde-Walsh, and Songxi Chen.

I thank the staffs at the Economics Dept. and the librarians at UC Riverside, especially those working at Inter-Library Loans, for their kind and efficient assistance.

I would like to thank my parents for their hardworking spirit that stimulates all my work. I thank my brother for his support to the family during the hard times that allows me to concentrate on my research. I also feel grateful to my sister, my wife and their family, and friends, for their lasting love, support and confidence in me.

Last, but not the least, I would like to express my deep gratitude to Hengfu Zou, whose devotion to the Chinese Economics education inspired me to switch my interest from Mathematics to Economics.

To my parents, Dazeng Tu and Cuifeng Wu.

To my brother, Yungui Tu.

To my sister, Yunmei Tu.

To my wife, Liyuang Wang.

# ABSTRACT OF THE DISSERTATION

Information and Inference in Econometrics: Estimation, Testing and Forecasting

by

Yundong Tu

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2012
Professor Tae-Hwy Lee, Co-Chairperson
Professor Aman Ullah, Co-Chairperson

Economic and Financial phenomena convey enormous information about the underlying structure of economic and policy interest. The first objective of the thesis is mainly concerned with how to make use of information efficiently, specifically, (1) how to separate noises from useful information in the presence of large dimensional data, (2) how to incorporate prior information (economic constraint), and (3) how to employ model structure, to conduct more informed inference, and thus to understand the economic structure wisely and draw sound policy conclusions.

The second dimension of information refers to the recent developments in the information theory that measure how much information content the observed data contains. The formalism of Maximum Entropy provides an information-theoretic approach to tackle economic problems, especially those with data observed in aggregate terms. Thus, the second objective of the thesis is to make use of this line of research and develop a new estimation method to measure quantities of economic interest when researchers are faced with model uncertainty.

# Contents

# List of Figures

# List of Tables

# Part I

# Forecasting

# Chapter 1

# Introduction

The thesis is consisted of 9 chapters, including this introduction chapter and a conclusion chapter in the end. It is divided into two parts, with Part I on Forecasting and Part II on Nonparametrics, Semiparametrics and Information Theoretic Econometrics.

Part I, Forecasting, includes Chapters 2 to 4. Chapter 2 and 3 investigate how economic information, presented in the form of economic constraints, would be adopted in the estimation of economic and financial structure and the forecasting of financial variables. We consider monotonicity constraint in Chapter 2 and positivity constraint in Chapter 3. The constraints are imposed via indicator functions in a nonparametric/semiparametric framework. We consider the smoothing of the indicator functions via bootstrap aggregating. We apply our methods to forecast equity premium and demonstrate the advantage of the proposed methods. Chapter 4 is on forecasting using supervised factor models with large dimensional data. We consider various ways of supervision in the computation of factors. Variable selection and factor computation are casted in the same framework and we show its advantage over the traditional forecasting methods.

Part II, Nonparametrics, Semiparametrics and Information Theoretic Econometrics, includes Chapter 5 to 8. Chapter 5 studies the estimation of nonparametric simultaneous equations models and proposes an oracle efficient estimator that makes use of the additive error structure. Chapter 6 investigates the estimation of marginal effect with large dimensional data. MAPLE estimators are proposed and shown to outperform other competitors available in the literature, in both Monte Carlo simulations and empirical application. Chapter 7 is on testing separable additivity of error term, the condition usually assumed, in structure models. We propose a test based on the partial derivative of the unknown structure with respect to the error term and a bootstrap procedure is developed to improve the performance of the test.

The abstracts of these chapters are presented as below.

Chapter 2 considers nonparametric and semiparametric regression models subject to monotonicity constraint. We use bagging as an alternative approach to Hall and Huang (2001). Asymptotic properties of our proposed estimators and forecasts are established. Monte Carlo simulation is conducted to show their finite sample performance. An application to predicting equity premium is taken to illustrate the merits of the proposed approach. We introduce a new forecasting evaluation criterion based on the second order stochastic dominance in the size of forecast errors, which enables us to compare the competing forecasting models over different sizes of forecast errors. We show that imposing monotonicity constraint can mitigate the chance of making large size forecast errors.

Economic theory frequently dictates constraints that should be met by statistical models that are used for quantitative analysis. For example, equity premium, which measure the difference between returns on risky assets and risk free assets, should be positive. This type of prior information could be used in equity premium modeling,

especially for the purpose of forecasting stock returns. Chapter 3 considers imposing such positiveness constraint in a mean model, with its extension in nonparametric kernel framework. Our constrained estimator is defined via an indicator function. A second step of smoothing the indicator function is carried out through bagging (Breiman, 1996). This bagging estimator is shown to have an explanation of model averaging, with weights determined by a transformation of the limiting random variable of the unconstrained estimator. Asymptotic properties of our proposed estimators and forecasts produced with these estimators are established. Monte Carlo simulations are conducted to show their finite sample performance. An application to predicting U.S. equity premium is taken to illustrate our proposed approach for imposing positiveness constraints.

Chapter 4 examines the theoretical and empirical properties of a supervised factor model based on combining forecasts using principal components (CFPC), in comparison with two other supervised factor models (partial least squares regression, PLS, and principal covariate regression, PCovR) and with the unsupervised principal component regression, PCR. The supervision refers to training the factors of predictors for a variable to forecast. We compare the performance of the three supervised factor models and the unsupervised factor model in forecasting of U.S. CPI inflation. The main finding is that the predictive ability of the supervised factor models is much better than the unsupervised factor model. Second, the computation of the factors can be doubly supervised together with variable selection, which can further improves the forecasting performance of the supervised factor models. Third, among the three supervised factor models, the CFPC best performs and is also most stable. While PCovR also performs well and is stable, the performance of PLS is not stable over different forecast horizons and out-of-sample forecasting periods. Fourth, the effect of supervision gets even larger as forecast horizon increases. Fifth, supervision helps to reduce the number of factors

and lags needed in modeling economic structure, achieving more parsimony.

Chapter 5 defines a new procedure to efficiently estimate nonparametric simultaneous equations models that have been explored by Newey et al (1999) and Su and Ullah (2008). The proposed estimation procedure exploits the additive structure and achieves oracle efficiency without the knowledge of unobserved error terms. Further, simulation results show that our new estimator outperforms that of Su and Ullah (2008) in terms of Mean Squared Error.

Chapter 6 studies the estimation of the marginal effect of one economic variable on another in the presence of large amount of other economic variables—a problem frequently faced by applied researchers. The estimation is motivated via model uncertainty so that random components should be included to describe the economy according to the state of the world. A condition named "Conditional Mean Independence" is shown to be sufficient to identify the partial effect parameter of interest. In the case that the parameter of interest can be identified in more than one approximating model, we propose two estimators for such a parameter: generalized-method-of-moment-based model averaging partial effect (gMAPLE) estimator and entropy-based model averaging partial effect (eMAPLE) estimator. Consistency and asymptotic normality of the MAPLE estimators are established under a suitable set of conditions. Thorough simulation studies reveal that MAPLE estimators outperform factor based, variable selection based and other model averaging estimators available in the literature. An economic example is taken to illustrate the use of MAPLE estimator to evaluate the effect of inherited control on firms' performance.

Chapter 7 considers testing additive error structure in nonparametric structural models, against the alternative hypothesis that the random error term enters the nonparametric model non-additively. We propose a test statistic under a set of iden-

tification conditions considered by Hoderlein, Su and White (2011), which require the existence of a control variable such that the regressor is independent with the error term given the control variable. The test statistic is motivated from the observation that, under the additive error structure, the partial derivative of the nonparametric structural function with respect to the error term is one under identification. The asymptotic distribution of the test is established and a bootstrap version is proposed to enhance its finite sample performance. Monte Carlo simulations show that the test has proper size and reasonable power in finite samples.

Chapter 8 considers instability of the parametric parameter in semiparametric partial linear model proposed by Robinson (1988), through the introduction of a threshold variable. The extended model, called partial threshold model, is estimated via a three-step procedure. Estimator of the threshold parameter is shown to have a nonstandard asymptotic distribution yet free of nuisance parameter, while estimators of the slope parameters are asymptotically normally distributed. The nonparametric component is consistently estimated and it achieves oracle efficiency, as if the threshold parameter is known. Testing for threshold effects and slope parameters are also considered. Monte Carlo experiments are carried out to compare the finite sample performance of the proposed method with direct nonparametric estimation and semiparametric partial linear models. Moreover, the proposed model is applied to study consumer demand and it shows the existence of a threshold in the fuel Engel curve.

Chapter 9 briefly concludes and comments on untouched issues.

# Chapter 2

# Nonparametric and Semiparametric Regressions Subject to Monotonicity Constraints: Estimation and Forecasting

## 2.1 Introduction

Linear models are frequently used for economic predictions. They are popular for their simplicity, computational efficiency, easy interpretation, and straightforwardness to impose prior known constraints. Campbell and Thompson (2008) consider applying sign restriction to the linear forecasting model of stock returns. The sign restriction (monotonicity constraint) is taken to alleviate parameter uncertainty and to

reconcile often contradicting in-sample and out-of-sample performance of predictors. They show that once a sensible restriction on the sign of a coefficient is imposed, the out-of-sample forecasting performance of many predictors can be improved and sometimes beat the historical average return forecast. Hillebrand et al (2009) incorporate the bagging (*b*ootstrap *agg*regat*ing*) approach of Gordon and Hall (2009) to smooth sign restrictions in linear forecasting models and show that the bagging sign restriction approach has more predictive power than the simple sign restriction of Campbell and Thompson (2008).

However, possible misspecification of a linear model can undermine its forecasts compared to those produced via nonlinear models. In this chapter we extend this literature by considering nonlinear models, in particular, nonparametric (NP) and semiparametric (SP) kernel regressions with imposing the local monotonicity constraints on the local coefficients of a predictor and with applying bagging to the constraints. Chen and Hong (2009) find that, in the prediction of asset returns, nonparametric kernel regression model has a better forecasting power than the historical mean, due to the higher signal-to-noise ratio resulted from nonparametric models. However, Chen and Hong (2009) do not consider the monotonicity restriction as well as bagging in their nonlinear forecasting exercise. This chapter is to consider nonlinear models subject to local monotonicity constraint with and without bagging.

Nonparametric kernel estimation with constraints has long history that dates back to the work of Brunk (1955). Recent work on imposing monotonicity on nonparametric regression function includes Hall and Huang (2001), Dette et al (2006) and Chernozhukov et al (2007), among others. Hall and Huang (2001) propose a novel method of imposing the monotonicity constraint on a class of nonparametric kernel estimations. Their estimator is constructed by re-weighting the kernel for each response

data point so that the impact of each observation on the estimated regression function can be controlled to satisfy a constraint. Their method is rooted in a conventional kernel framework and is extended by Racine et al (2009) and Henderson and Parmeter (2009) to allow for a broader class of conventional constraints and to develop tests for these constraints.

Our contributions are as given below. First, we consider NP and SP models to generalize the linear models considered in Goyal and Welch (2008), Campbell and Thompson (2008) and Hillebrand et al (2009). These NP/SP regressions can capture possibly neglected nonlinearity in linear models and could improve the predictive ability of the predictors, as demonstrated in our Monte Carlo simulation and application to the equity premium prediction. Second, we consider a new method of imposing the monotonicity constraint on the NP and SP regressions. This is to make the prediction more accurate as we employ more information than Chen and Hong (2009). Our monotonicity constraint is local restriction while it is global monotonicity in Campbell and Thompson (2008). Third, we use bagging to smooth the monotonicity constraint in NP and SP regressions as Hillebrand et al (2009) do in linear regressions. It has been shown in Bühlmann and Yu (2002) that bagging can reduce asymptotic mean squared error in linear regressions. We obtain the similar results that hold locally in NP and SP regressions. Fourth, we conduct simulation study to demonstrate how the asymptotic results work in finite sample. We also conduct an empirical study in predicting equity premium using the same data from Campbell and Thompson (2008) to demonstrate the practical merit of the bagging monotonicity constrained NP and SP regression models. Fifth, in our simulation and empirical application, we find that, despite its simplicity to implement, our bagging constrained NP regression almost always and clearly outperforms the constrained NP regression of Hall and Huang (2001). Sixth, we introduce a new

forecast evaluation measure based on the second order stochastic dominance (SOSD) of the squared forecast errors, by which we can compare forecasting models in entire predictive distribution of squared forecast errors rather than just in mean of squared forecast errors. The new SOSD criterion enables us to compare forecasting models over different parts of the predictive distributions of squared forecast errors, e.g., over small size errors vs big size errors, as demonstrated using our empirical results for the equity premium prediction application. We show that imposing sensible constraints reduces the chance of making the big size forecast errors and thereby improves the forecasting ability of models.

The chapter is organized as follows. Section 2 presents the NP and SP methods with local monotonicity constraints and with bagging. Sections 3, 4, 5 establish the asymptotic properties of each of parametric, nonparametric, semiparametric bagging constrained estimators and forecasts. Section 6 conducts Monte Carlo simulation to compare our proposed bagging constrained NP and SP forecasts with forecasts from linear models and from the constrained NP method of Hall and Huang (2001). Section 7 presents empirical results on the equity premium prediction. Section 8 concludes.

## 2.2 Estimation with Constraints

Many economic models try to establish a relationship between a variable of interest $y_t$ and a scalar or vector predictor variable $x_t$. For the maximum clarity of presentation, we consider the case that $x_t$ is a scalar. All the results in this chapter would follow when $x_t$ is a vector. In forecasting, the $h$-step ahead forecast of $y_{n+h}$ at time $t = n$ given that $x_n = x$ is defined as

$$m_{n,h}(x) = E(y_{n+h}|x_n = x). \tag{2.1}$$

Sometimes a priori constraints may be suggested from economic theory, often in the form of bounds. For example, the marginal propensity to consume is less than 1; production technology is concave; the regression function $m_{n,h}(x)$ is positive, monotonic, homogeneous, homothetic, and etc. To this end, estimators or forecasts may be subject to constraints. In this chapter, we focus on slope constraint (i.e., monotonicity) of a curve that relates $y$ and $x$, while constraints of other type like curvature may be possible as well within our framework.

### 2.2.1 Parametric Estimation with Constraints

First, consider a parametric linear model with a single regressor $x$:

$$m_{n,h}(x) = \alpha + \beta x \tag{2.2}$$

Goyal and Welch (2008) use the unconstrained ordinary least squares (OLS) estimators $\tilde{\alpha}, \tilde{\beta}$ in the prediction of stock returns using a predictor $x$. Note that $\tilde{\alpha}$ and $\tilde{\beta}$ satisfy

$$\tilde{\alpha} = \bar{y}_n - \tilde{\beta}\bar{x}_n \tag{2.3}$$

where $\bar{y}_n = \frac{1}{n}\sum_{t=1}^{n} y_t$ and $\bar{x}_n = \frac{1}{n}\sum_{t=1}^{n} x_t$.

If a monotonicity constraint (positive slope) is considered as sensible, one can estimate $\beta$ through thresholding using an indicator function as done by Campbell and Thompson (2008),

$$\begin{aligned} \bar{\beta} &= 1_{\{\tilde{\beta}>0\}} \cdot \tilde{\beta}, \\ \bar{\alpha} &= 1_{\{\tilde{\beta}>0\}} \cdot \tilde{\alpha} + 1_{\{\tilde{\beta}\leq 0\}} \cdot \bar{y}_n. \end{aligned} \tag{2.4}$$

Note that the relationship between $\bar{\alpha}$ and $\bar{\beta}$ remains as in (2.3)

$$\bar{\alpha} = \bar{y}_n - \bar{\beta}\bar{x}_n, \tag{2.5}$$

since $\bar{\alpha} = 1_{\{\tilde{\beta}>0\}} \cdot \tilde{\alpha} + 1_{\{\tilde{\beta}\leq 0\}} \cdot \bar{y}_n = 1_{\{\tilde{\beta}>0\}} \cdot \left(\bar{y}_n - \tilde{\beta}\bar{x}_n\right) + 1_{\{\tilde{\beta}\leq 0\}} \cdot \bar{y}_n = \bar{y}_n - 1_{\{\tilde{\beta}>0\}} \cdot \tilde{\beta}\bar{x}_n$.

As the constraint is imposed using a hard-thresholding indicator function, it can introduce significant bias and variance. Gordon and Hall (2009) propose a bagging constrained estimator

$$\hat{\beta} = \frac{1}{J}\sum_{j=1}^{J} \bar{\beta}^{*(j)} \equiv E^* \bar{\beta}^*, \tag{2.6}$$

where $\bar{\beta}^{*(j)} = 1_{\left\{\tilde{\beta}^{*(j)}>0\right\}} \cdot \tilde{\beta}^{*(j)}$ and here $\tilde{\beta}^{*(j)}$ is the unconstrained OLS estimator from the $j$th bootstrap sample ($j = 1, \ldots, J$). We use $E^*(\cdot)$ to denote the bootstrap average. It can be shown that

$$\hat{\alpha} \equiv \bar{y}_n - \hat{\beta}\bar{x}_n = E^* \bar{\alpha}^*. \tag{2.7}$$

Bühlmann and Yu (2002) show that this bagging constrained estimator can have smaller asymptotic mean squared error (AMSE), notwithstanding the larger asymptotic bias.

### 2.2.2 Nonparametric Estimation with Constraints

Despite its simplicity a parametric linear model like $y_t = \alpha + \beta x_t + u_t$ may be subject to misspecification because it may be that $E(u_t|x_t) \neq 0$ due to possible neglected nonlinearity. This is to be avoided via a nonparametric regression, $y_t = m(x_t) + u_t$, where $m(x_t) = E(y_t|x_t)$ and $u_t = y_t - E(y_t|x_t)$. Kernel estimators of $m(x_t)$ such as Nadaraya-Watson or local linear estimators are common practice in the nonparametric literature. Yet, in the face of information derived from economic theory, we may wish to impose some constraints (e.g., monotonicity, positivity) on the nonparametric kernel regression models. Hall and Huang (2001) propose a re-weighted kernel method to impose constraints on a general class of kernel estimators, which is followed by Racine, Parmeter and Du (2009) and Henderson and Parmeter (2009). Alternatively, we propose here to use bagging to impose constraints in nonparametric kernel regression models.

#### 2.2.2.1 Nonparametric Estimation with Constraints: Hall and Huang (2001)

Consider a general class of kernel estimator written as weighted average of $y$'s

$$\hat{m}_{n,h}(x) = \frac{1}{n-h}\sum_{t=1}^{n-h} A_t(x) y_{t+h}, \tag{2.8}$$

where $A_t(x)$ is the weighting function. Hall and Huang (2001) suggested an estimator

$$m_{n,h}(x;\mathbf{p}) = \sum_{t=1}^{n-h} p_t A_t(x) y_{t+h}, \tag{2.9}$$

where $\mathbf{p} = (p_1, \ldots, p_{n-h})'$. Note that (2.8) is a special case of (2.9) with the uniform weights $\mathbf{p}_0 = (\frac{1}{n-h}, \ldots, \frac{1}{n-h})'$. $\mathbf{p}$ is to be estimated by $\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} D(\mathbf{p})$ subject to some conditions such as $\sum_{t=1}^{n-h} p_t = 1$, with a distance function $D(\mathbf{p})$ between $\mathbf{p}$ and $\mathbf{p}_0$. For example, $D(\mathbf{p}) = (\mathbf{p}-\mathbf{p}_0)'(\mathbf{p}-\mathbf{p}_0)$, or $D(\mathbf{p}) = (\mathbf{p}^{1/2} - \mathbf{p}_0^{1/2})'(\mathbf{p}^{1/2} - \mathbf{p}_0^{1/2})$ if the elements of $\mathbf{p}$ and $\mathbf{p}_0$ are on the unit interval, e.g., probability weights.

#### 2.2.2.2 Nonparametric Estimation with Constraints: Bagging

Take the first order Taylor series expansion of $m(x_t)$ around $x$ so that

$$\begin{aligned} y_t &= m(x_t) + u_t = m(x) + (x_t - x)m^{(1)}(x) + v_t \\ &= \alpha(x) + x_t\beta(x) + v_t = X_t\delta(x) + v_t \end{aligned} \tag{2.10}$$

where $X_t = (1 \; x_t)$ and $\delta(x) = [\alpha(x) \; \beta(x)']'$ with $\alpha(x) = m(x) - x\beta(x)$ and $\beta(x) = m^{(1)}(x)$. The local linear least square (LLLS) estimator of $\delta(x)$ is then obtained by minimizing

$$\sum_{t=1}^{n} v_t^2 K_h(x_t, x) = \sum_{t=1}^{n} (y_t - X_t\delta(x))^2 K_h(x_t, x), \tag{2.11}$$

where $K_h(x_t, x) = K\left(\frac{x_t - x}{h}\right)$ is a decreasing function of the distance of the regressor $x_t$ from the evaluation point $x$, and $h \to 0$ as $n \to \infty$ is the bandwidth which determines how rapidly the weights decrease as the distance of $x_t$ from $x$ increases. The LLLS

estimator is given by

$$\tilde{\delta}(x) = (\mathbf{X}'\mathbf{K}(x)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(x)\mathbf{y}, \tag{2.12}$$

where $\mathbf{X}$ is an $n \times (k+1)$ matrix with the $t$th row $X_t$ $(t = 1, \ldots, n)$, $\mathbf{y}$ is an $n \times 1$ vector with elements $y_t$ $(t = 1, \ldots, n)$, and $\mathbf{K}(x)$ is the $n \times n$ diagonal matrix with the diagonal elements $K_h(x_t, x)$ $(t = 1, \ldots, n)$. Then we have LLLS estimators $\tilde{\alpha}(x) = (1 \;\; 0)\tilde{\delta}(x)$ and $\tilde{\beta}(x) = (0 \;\; 1)\tilde{\delta}(x)$.

Using the constrained LLLS estimator $\bar{\beta}(x)$

$$\bar{\beta}(x) = 1_{\{\tilde{\beta}(x)>0\}} \cdot \tilde{\beta}(x), \tag{2.13}$$

we get the bagging constrained LLLS estimator $\hat{\beta}(x)$

$$\hat{\beta}(x) = \frac{1}{J}\sum_{j=1}^{J} \bar{\beta}(x)^{*(j)} = E^*\bar{\beta}(x)^*. \tag{2.14}$$

Observing (2.3) and (2.5), the unconstrained LLLS estimator is

$$\tilde{\alpha}(x) = \bar{y}(x) - \tilde{\beta}(x)\bar{x}(x), \tag{2.15}$$

where

$$\begin{aligned} \bar{y}(x) &= \frac{\sum_{t=1}^{n} K_h(x_t, x)y_t}{\sum_{t=1}^{n} K_h(x_t, x)} = (\mathbf{i}'\mathbf{K}(x)\mathbf{i})^{-1}\mathbf{i}'\mathbf{K}(x)\mathbf{y}, \\ \bar{x}(x) &= \frac{\sum_{t=1}^{n} K_h(x_t, x)x_t}{\sum_{t=1}^{n} K_h(x_t, x)} = (\mathbf{i}'\mathbf{K}(x)\mathbf{i})^{-1}\mathbf{i}'\mathbf{K}(x)\mathbf{x}, \end{aligned} \tag{2.16}$$

with $\mathbf{i}$ being an $n \times 1$ vector of unit elements and $\mathbf{x}$ being an $n \times 1$ vector with elements $x_t$ $(t = 1, \ldots, n)$. Following the same steps as for $\bar{\beta}(x)$ and $\hat{\beta}(x)$, the two constrained LLLS estimators for $\alpha(x)$ are obtained as

$$\bar{\alpha}(x) = \bar{y}(x) - \bar{\beta}(x)\bar{x}(x), \tag{2.17}$$

$$\hat{\alpha}(x) = \bar{y}(x) - \hat{\beta}(x)\bar{x}(x), \tag{2.18}$$

or equivalently $\hat{\alpha}(x) = \frac{1}{J}\sum_{j=1}^{J} \bar{\alpha}(x)^{*(j)} = E^*\bar{\alpha}(x)^*$.

We derive explicit formula for the NP forecast from the above. Note that from (2.10) we have the unconstrained NP forecast,

$$\begin{aligned}\tilde{m}(x) &= \tilde{\alpha}(x) + x\tilde{\beta}(x) = \bar{y}(x) - \tilde{\beta}(x)\bar{x}(x) + x\tilde{\beta}(x) \\ &= \bar{y}(x) - \tilde{\beta}(x)\left[\bar{x}(x) - x\right]. \end{aligned} \tag{2.19}$$

Similarly, we get the constrained NP forecast

$$\bar{m}(x) = \bar{y}(x) - \bar{\beta}(x)\left[\bar{x}(x) - x\right], \tag{2.20}$$

and the bagged constrained NP forecast

$$\hat{m}(x) = \bar{y}(x) - \hat{\beta}(x)\left[\bar{x}(x) - x\right]. \tag{2.21}$$

### 2.2.3 Semiparametric Estimation with Constraints

Glad (1998) and Martins-Filho et al (2009) note that sensible parametrically guided SP models outperform NP models in that the significant bias reduction is achieved while maintaining the asymptotic variance. Therefore we consider the SP model

$$\begin{aligned} y &= \alpha + \beta x + E(u|x) + [u - E(u|x)] \\ &= \alpha + \beta x + E(u|x) + v = m(x) + v \end{aligned} \tag{2.22}$$

where $m(x) = \alpha + \beta x + E(u|x)$ and $v = u - E(u|x)$. To estimate $m(\cdot)$, first we run a regression of $y$ on $(1\ x)'$ to obtain the estimation of the linear component $\tilde{\alpha} + \tilde{\beta}x$. The second step involves an LLLS estimation of $E(u|x)$, which can be performed in an NP regression of $\tilde{u} = y - \tilde{\alpha} - \tilde{\beta}x$ on $x$. Let $\tilde{\xi}(x)$ be the intercept and $\tilde{\eta}(x)$ be the slope function of this regression. Thus the LLLS estimator can be represented by $\tilde{\xi}(x) - \tilde{\eta}(x)\left(\bar{x}(x) - x\right)$. This two-step algorithm leads to an unconstrained SP estimator

of $m(\cdot)$ as

$$\begin{aligned}\tilde{m}_{sp}(x) &= \tilde{\alpha} + \tilde{\beta}x + \tilde{\xi}(x) - \tilde{\eta}(x)(\bar{x}(x) - x) \\ &= \tilde{\alpha} + \tilde{\xi}(x) - \tilde{\eta}(x)\bar{x}(x) + \left\{\tilde{\beta} + \tilde{\eta}(x)\right\}x,\end{aligned} \tag{2.23}$$

the slope of which is estimated by

$$\tilde{\beta}(x) \equiv \tilde{\beta} + \tilde{\eta}(x). \tag{2.24}$$

To impose the local monotonicity constraint, we define our constrained SP estimator as

$$\bar{\beta}(x) = 1_{\left\{\tilde{\beta}(x)>0\right\}} \cdot \tilde{\beta}(x) \tag{2.25}$$

When $\tilde{\beta}(x) \leq 0$, the slope of the regression function is zero $\bar{\beta}(x) = 0$. In this case, instead of fitting a semiparametric model, local constant kernel estimator should be adopted. This observation leads to the following local monotonicity constrained SP forecast

$$\bar{m}_{sp}(x) = \tilde{m}_{sp}(x) \cdot 1_{\left\{\tilde{\beta}(x)>0\right\}} + \tilde{m}_{lc}(x) \cdot 1_{\left\{\tilde{\beta}(x)\leq 0\right\}}, \tag{2.26}$$

where $\tilde{m}_{lc}(x) = \bar{y}(x)$ is the local constant kernel estimator of $m(x)$ as in (2.16).

With having $\bar{m}_{sp}(x)$ obtained, similarly to (8.12), we get the bagging constrained SP forecast from

$$\hat{m}_{sp}(x) = \frac{1}{J}\sum_{j=1}^{J} \bar{m}_{sp}^{*(j)}(x) = E^{*}\bar{m}_{sp}^{*}(x), \tag{2.27}$$

with $\bar{m}_{sp}^{*(j)}(x)$ obtained from the $j$th bootstrap sample.

## 2.3 Sampling Properties of Parametric Estimators

Sampling properties of parametric estimators, including constrained parametric estimator and bagging constrained estimator, are presented in this section, while NP

and SP estimators are treated in the two subsequent sections. Sampling properties of constrained parametric estimator have been established by Judge and Yancey (1986) under normality distribution. With general distribution condition of the unconstrained estimator, we prove the superiority of the constrained estimator (in terms of MSE) if the constraint is correctly specified. We also present the local asymptotic theory for the constrained estimator and its bagging version.

### 2.3.1 Constrained Parametric Estimator

**Theorem 1.** Let the unconstrained estimator $\tilde{\beta}$ of $\beta$ have a cumulative density function denoted by $F_{\tilde{\beta}}(\cdot)$. Then we have the following for the constrained estimator $\bar{\beta} = \max\{\tilde{\beta}, \beta_1\}$, for some given constant $\beta_1$, (1) $F_{\bar{\beta}}(z) = F_{\tilde{\beta}}(z) \cdot 1_{\{z \geq \beta_1\}}$. (2) $bias\bar{\beta} \geq bias\tilde{\beta}$. (3) $Var\left(\bar{\beta}\right) \leq Var\left(\tilde{\beta}\right)$ if $bias\tilde{\beta} \geq 0$ and $\beta_1 \leq \beta$ and (4) $MSE\left(\bar{\beta}\right) \leq MSE\left(\tilde{\beta}\right)$ if $\beta_1 \leq \beta$.

**Remark 1.** Theorem 1 establishes that the constrained estimator, $\bar{\beta}$, has a condensed density and it is biased upward, compared to its unconstrained counterpart, $\tilde{\beta}$. Part 1 depicts its CDF in terms of that of $F_{\tilde{\beta}}(\cdot)$. The indicator function compresses all the mass for $\tilde{\beta}$ that lie below $\beta_1$ to $\beta_1$. Part 2 states that $\bar{\beta}$ is biased upward compared to $\tilde{\beta}$. This upward bias is due to the max operator in its definition. If the constraint is an upper bound instead of a lower bound, then the min operator will incur downward bias. Part 3 shows that $\bar{\beta}$ would have smaller variance, provided that the constraint is correctly specified and $\tilde{\beta}$ is biased upward, while part 4 dictates the superiority of $\bar{\beta}$ in terms of MSE when the constraint is correct. It's yet clear that, even if the constraint is wrongly specified, there could still be reduction in MSE and variance for $\bar{\beta}$. However, this will require further conditions on $F_{\tilde{\beta}}(\cdot)$. These conditions are not informative, therefore we do not proceed in that direction.

Judge and Yancey (1986) consider the case in which $\tilde{\beta}$ has a normal distribution. They (p. 50) depict a figure showing that, the performance of $\bar{\beta}$ relative to that of $\tilde{\beta}$ depends on $\delta \equiv \beta_1 - \beta$. The constrained estimator is superior for a large range values of $\delta$, and when $\delta \to \infty$, $MSE\left(\bar{\beta}\right)$ is equal to the mean squared error of an equality constrained estimator, i.e. $\bar{\beta} = \beta_1$. Under the normality assumption, $Var\left(\bar{\beta}\right) \leq Var\left(\tilde{\beta}\right)$ over the whole range of parameter space and the former will approach the variance of the equality constrained estimator as $\delta \to \infty$. □

Next, we consider asymptotic distribution of $\bar{\beta}$ under suitable assumptions as stated in the following theorem.

**Theorem 2.** Consider an unconstrained parametric estimator $\tilde{\beta}$ of $\beta$ with

$$\gamma(n)\,\sigma_\beta^{-1}\left(\tilde{\beta} - \beta\right) \xrightarrow{d} Z \tag{2.28}$$

and $Z$ is a random variable with CDF $F(\cdot)$, where $\sigma_\beta$ is the asymptotic standard deviation of $\tilde{\beta}$ and $\lim_{n\to\infty} \gamma(n) = \infty$. Then the constrained estimator defined as $\bar{\beta} = \max\{\tilde{\beta}, \beta_1\}$, for some given constant $\beta_1$, has the following properties,

1. when $\beta > \beta_1$, $\gamma(n)\,\sigma_\beta^{-1}\left(\bar{\beta} - \beta\right) \xrightarrow{d} Z$ .
2. when $\beta = \beta_1$, $\Pr\left(\gamma(n)\,\sigma_\beta^{-1}\left(\bar{\beta} - \beta\right) < z\right) \to F(z) \cdot 1_{\{z \geq 0\}}$.

If we further assume that

$$\gamma(n)\,\sigma_\beta^{-1}(\beta - \beta_1) = b, \tag{2.29}$$

for some constant $b$, and that $F$ is standard normal CDF $\Phi$ (with its PDF $\varphi$) and $Z_b = Z + b$, then

3. $\lim_{n\to\infty} \gamma(n)\,\sigma_\beta^{-1}\left(\bar{\beta} - \beta\right) = Z_b 1_{\{Z_b > 0\}} - b$.

4. $\lim_{n\to\infty} \gamma(n)\,\sigma_\beta^{-1} E\left(\bar{\beta}-\beta\right) = \varphi(b) + b\Phi(b) - b.$

5. $\lim_{n\to\infty} Var\left[\left(\gamma(n)\,\sigma_\beta^{-1}\right)^{1/2}\bar{\beta}\right] = \Phi(b)+b\varphi(b)-\varphi^2(b)-2b\varphi(b)\,\Phi(b)+b^2\Phi(b)\left[1-\Phi(b)\right].$

**Remark 2(a).** Theorem 2 stated the limiting distribution of $\bar{\beta}$. Parts 1 and 2 present the usual asymptotic distribution when the constraint is strict (i.e., $\beta > \beta_1$) and when the parameter is on the boundary (i.e., $\beta = \beta_1$). Part 1 confirms the intuition that, as long as the constraint is strict, it will not be violated by the unconstrained estimator $\tilde{\beta}$ when the sample size is large enough. This leads to the conclusion that $\bar{\beta}$ would be asymptotically equivalent to $\tilde{\beta}$ in this case. On the other hand, when $\beta$ is on the boundary, the limiting CDF compresses all the mass of negative values at 0. Part 3 establishes the local asymptotic distribution of $\bar{\beta}$ that depends on the drift parameter $b$ with asymptotic bias and variance given in parts 4 and 5. It is easy to see that, if $b$ is allowed to grow as $n$, $Z_b 1_{\{Z_b>0\}} - b$ will collapse to $Z$, and result in part 3 becomes that in part 1. Similarly, 2 is reproduced with part 3 when $b = 0$.

**Remark 2(b).** Theorem 2 only requires $\tilde{\beta}$ satisfy some limiting theorem with asymptotic standard deviation $\sigma_\beta$. This is a very week condition that is met by a large class of estimators. We do not specify the convergence rate $\gamma(n)$ but simply let it explode as $n$ increases. This general setting accommodates both estimators with standard convergence rate $\sqrt{n}$ and estimators with nonstandard convergence rate, e.g., $n^{1/3}$ or $n^{3/2}$. The condition $\gamma(n)\,\sigma_\beta^{-1}(\beta-\beta_1) = b$ can be stated alternatively as $\beta = \beta_1 + \gamma^{-1}(n)\,\sigma_\beta b$ for some constant $b$. It dictates that the true parameter $\beta$ is a Pitman type drift to the specified bound $\beta_1$, with a drift parameter $b$. The local drift rate is the same as the convergence rate of $\tilde{\beta}$. Extensions to higher or lower rate than this convergence rate $(\gamma^{-1}(n))$ can be made by letting $b = b_n$ go to either infinity or zero as $n$ increases. We do not explore this issue here. □

### 2.3.2 Bagged Constrained Parametric Estimator

**Theorem 3.** Let an unconstrained estimator $\tilde{\beta}$ of $\beta$ and its bootstrap version $\tilde{\beta}^*$ have the following asymptotics,

$$\begin{aligned} \gamma(n)\,\sigma_{\beta}^{-1}\left(\tilde{\beta}-\beta\right) &\xrightarrow{d} Z, \\ \gamma(n)\,\sigma_{\beta}^{-1}\left(\tilde{\beta}^*-\tilde{\beta}\right) &\xrightarrow{d} Z, \end{aligned} \tag{2.30}$$

with $Z$ being a standard normal random variable, where $\sigma_\beta$ is the asymptotic standard deviation of $\tilde{\beta}$ and $\lim_{n\to\infty}\gamma(n)=\infty$. Further the constrained estimator is $\bar{\beta}=\max\left\{\tilde{\beta},\beta_1\right\}$, where $\beta_1$ satisfies

$$\gamma(n)\,\sigma_{\beta}^{-1}(\beta-\beta_1)=b, \tag{2.31}$$

for some constant $b$ and denote $Z_b=Z+b$. Then, for the bagged version of $\bar{\beta}$, $\hat{\beta}\equiv E^*\bar{\beta}^*$, we have

1. $\gamma(n)\,\sigma_{\beta}^{-1}\left(\hat{\beta}-\beta\right)\xrightarrow{d} Z-Z_b\Phi(-b-Z)+\varphi(-b-Z)$.
2. $\lim_{n\to\infty}\gamma(n)\,\sigma_{\beta}^{-1}E\left(\hat{\beta}-\beta\right)=2\varphi*\varphi(-b)-b\Phi*\varphi(-b)$.
3. $\lim_{n\to\infty}Var\left[\left(\gamma(n)\,\sigma_{\beta}^{-1}\right)^{1/2}\hat{\beta}\right]=1+\Phi^2*\varphi''(-b)+\Phi^2*\varphi(-b)-2b\Phi^2*\varphi'(-b)+b^2\Phi^2*\varphi(-b)+\varphi^2*\varphi(-b)-2\Phi*\varphi''(-b)-2\Phi*\varphi(-b)+2b\Phi*\varphi'(-b)-2\varphi*\varphi'(-b)+2(\Phi\cdot\varphi)*\varphi'(-b)-2b(\Phi\cdot\varphi)*\varphi(-b)-[2\varphi*\varphi(-b)-b\Phi*\varphi(-b)]^2$.

**Remark 3(a).** We adopted the notation $f*g$ to denote the convolution of two functions $f$ and $g$, which is defined as $f*g(s)=\int f(t)\times g(s-t)\,ds$.

**Remark 3(b).** It is clear from part 2 of Theorem 3 that both bias and variance of the bagging constrained estimator depend on the parameter $b$, which measures how accurate $\beta_1$, the lower bound of $\beta$, is. We compare the AMSE of bagging constrained estimator

with that without bagging, and numerical calculation reveals the superiority of bagging when $b > 0.392$. Figure 1 plots the asymptotic variance, asymptotic squared bias and asymptotic mean squared error of $\hat{\beta}$ together with those of $\bar{\beta}$, against values of $b$ in the range of $[-1, 5]$. It is seen that our bagging estimator enjoys a large reduction in asymptotic mean squared error for values of $b \in [1, 3]$.

**Remark 3(c).** (2.30) requires that bootstrap work for $\tilde{\beta}$, i.e., $\tilde{\beta}^*$ has the same asymptotic distribution as $\tilde{\beta}$. The necessary and sufficient conditions for this bootstrap consistency can be found in Mammen (1992). We emphasize that we do not require that bootstrap work for $\bar{\beta}$. In fact, the bootstrap fails for $\bar{\beta}$ as noted in Andrews (2000, p. 401). Theorem 3 shows that the asymptotic distribution of $\hat{\beta} \equiv E^* \bar{\beta}^*$ is different from the asymptotic distribution of $\bar{\beta}$ which is shown in Theorem 2. The difference is depicted in Figure 1. It is this bootstrap failure for $\bar{\beta}$ that leads to Theorem 3. □

Figure 1 About Here

## 2.4 Sampling Properties of Nonparametric Estimators

We consider sampling properties of NP estimators under constraint and its bagging version.

### 2.4.1 Constrained Nonparametric Estimator

**Theorem 4.** Let the nonparametric estimator $\tilde{\beta}(x)$ of $\beta(x)$ with

$$\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}(x) - \beta(x)\right) \xrightarrow{d} Z, \tag{2.32}$$

$$\gamma_2(n,h)\,\sigma_m^{-1}(x)\left(\tilde{m}(x) - m(x) - B_m(x)\right) \xrightarrow{d} Z,$$

where $\lim_{n\to\infty}\gamma_i(n,h)=\infty, i=1,2$, $h$ is the bandwidth satisfying $h=cn^{\tau}$ for some $c>0$, $\tau<0$, $Z$ is a standard normal random variable, $\sigma_\beta(x)$ is the asymptotic standard deviation of $\tilde{\beta}(x)$, $\sigma_m(x)$ is the asymptotic standard deviation of $\tilde{m}(x)$, $B_m(x)=\frac{1}{2}h^2m^{(2)}(x)\int v^2k(v)\,dv+o_p\left(h^2\right)$ is the asymptotic bias $\tilde{m}(x)$. Then the following limiting statements hold for the constrained estimator $\bar{\beta}(x)=\max\{\tilde{\beta}(x),\beta_1(x)\}$, for some given $\beta_1(x)$,

1. when $\beta(x)>\beta_1(x)$, $\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\bar{\beta}(x)-\beta(x)\right)\xrightarrow{d}Z$.
2. when $\beta(x)=\beta_1(x)$, $\Pr\left(\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\bar{\beta}(x)-\beta(x)\right)<z\right)\to\Phi(z)\cdot 1_{\{z\geq 0\}}$.
3. when $\beta(x)>\beta_1(x)$, $\gamma_2(n,h)\,\sigma_m^{-1}(x)\left[\bar{m}(x)-m(x)-B_m(x)\right]\xrightarrow{d}Z$.

If we further assume that $\gamma_1(n,h)\,\sigma_\beta^{-1}(\beta(x)-\beta_1(x))=b(x)$, for some real function $b(x)$, and denote $Z_{b(x)}=Z+b(x)$, then

4. $\lim_{n\to\infty}\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left[\bar{\beta}(x)-\beta(x)\right]=Z_{b(x)}1_{\left\{Z_{b(x)}>0\right\}}-b(x)$.
5. $\lim_{n\to\infty}\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\,E\left[\bar{\beta}(x)-\beta(x)\right]=\varphi(b(x))+b(x)\Phi(b(x))-b(x)$.
6. $\lim_{n\to\infty}Var\left[\left(\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\right)^{1/2}\bar{\beta}(x))\right]=\Phi(b(x))+b(x)\varphi(b(x))-\varphi^2(b(x))-2b(x)\varphi(b(x))\,\Phi(b(x))+b^2(x)\Phi(b(x))\left[1-\Phi(b(x))\right]$.

**Remark 4(a).** The above theorem shows the results for NP estimators with constraints. The implications are similar to the previous theorem on constrained parametric estimators. Note that the constraint bound $\beta_1(x)$ can vary for different values of $x$. As a special case in which $\beta_1(x)=\beta_1$, a constant, it is efficient to adopt the restriction if it is correctly specified via the constrained estimator.

**Remark 4(b).** The constrained estimator of $m(x)$, $\bar{m}(x)$, has the asymptotic property as the unconstrained nonparametric estimator, when the constraint is strict, as

established in part 3 of Theorem 4. The implication for bandwidth selection for the constrained estimator $\bar{m}(x)$ is that the classical cross-validation approach shall apply. The bias term $B_m(x)$ goes to zero if $\gamma_2(n,h)\,h^2$ tends to zero as $n$ tends to infinity.

### 2.4.2 Bagged Constrained Nonparametric Estimator

**Theorem 5.** Let an estimator $\tilde{\beta}(x)$ of $\beta(x)$ and its bootstrap version $\tilde{\beta}^*(x)$ have the following asymptotic,

$$\begin{aligned}
&\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}(x)-\beta(x)\right)\xrightarrow{d} Z,\\
&\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}^*(x)-\tilde{\beta}(x)\right)\xrightarrow{d} Z,
\end{aligned}\tag{2.33}$$

where $Z$ is a standard normal random variable, $\lim_{n\to\infty}\gamma_1(n,h)=\infty$, $h$ is the bandwidth satisfying $h=cn^{\tau}$ for some $c>0$, $\tau<0$, $\sigma_\beta(x)$ is the asymptotic standard deviation of $\tilde{\beta}(x)$. Define $\bar{\beta}(x)=\max\left\{\tilde{\beta}(x),\beta_1(x)\right\}$, with some known $\beta_1(x)<\beta(x)$ that satisfies

$$\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\beta(x)-\beta_1(x)\right)=b(x),\tag{2.34}$$

where $b(\cdot)$ is some real function and denote $Z_{b(x)}=Z+b(x)$. For the bagged version of $\bar{\beta}(x)$, $\hat{\beta}(x)\equiv E^*\bar{\beta}^*(x)$, we have

1. $\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\hat{\beta}(x)-\beta(x)\right)\xrightarrow{d} Z-Z_{b(x)}\Phi(-b(x)-Z)+\varphi(-b(x)-Z)$.

2. $\lim_{n\to\infty}\gamma_1(n,h)\,\sigma_\beta^{-1}E\left[\hat{\beta}(x)-\beta(x)\right]=2\varphi*\varphi(-b(x))-b(x)\,\Phi*\varphi(-b(x))$.

3. $\lim_{n\to\infty}Var\left[\left(\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\right)^{1/2}\hat{\beta}(x)\right]=1+\Phi^2*\varphi''(-b(x))+\Phi^2*\varphi(-b(x))-2b\Phi^2*\varphi'(-b(x))+b^2(x)\,\Phi^2*\varphi(-b(x))+\varphi^2*\varphi(-b(x))-2\Phi*\varphi''(-b(x))-2\Phi*\varphi(-b(x))+2b(x)\,\Phi*\varphi'(-b(x))-2\varphi*\varphi'(-b(x))+2(\Phi\cdot\varphi)*\varphi'(-b(x))-2b(x)\,(\Phi\cdot\varphi)*\varphi(-b(x))-[2\varphi*\varphi(-b(x))-b(x)\,\Phi*\varphi(-b(x))]^2$.

4. If $\gamma_2(n,h)\,\sigma_m^{-1}(x)\,(\tilde{m}(x)-m(x)-B_m(x)) \xrightarrow{d} Z$,where $B_m(x)=\frac{1}{2}h^2m^{(2)}(x)\int v^2k(v)\,dv+o_p\left(h^2\right)$ is the asymptotic bias $\tilde{m}(x)$, $\sigma_m(x)$ is the asymptotic standard deviation of $\tilde{m}(x)$, and $\gamma_2(n,h)$ follows similar conditions as $\gamma_1(n,h)$ ,then

$$\gamma_2(n,h)\,\sigma_m^{-1}(x)\,[\hat{m}(x)-m(x)-B_m(x)] \xrightarrow{d} Z. \tag{2.35}$$

**Remark 5.** When $b(\cdot)$ admits a constant function, the limiting distribution in part 1 is the same as in the parametric case. That is, for all possible values of $x$, $\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\hat{\beta}(x)-\beta(x)\right)$ converges to the same random variable as $\gamma_1(n)\,\sigma_\beta^{-1}\left(\hat{\beta}-\beta\right)$ does in the parametric case.

□

## 2.5 Sampling Properties of Semiparametric Estimators

SP estimators and its bagging version are considered in this section. We present, in sequence, their sampling properties in the following two theorems.

### 2.5.1 Constrained Semiparametric Estimator

**Theorem 6.** Consider an estimator $\tilde{\beta}(x)$ of $\beta(x)$ with

$$\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}(x)-\beta(x)\right) \xrightarrow{d} Z, \tag{2.36}$$

where $Z$ is a standard normal random variable, $\sigma_\beta(x)$ is the asymptotic standard deviation of $\tilde{\beta}(x)$, $\lim_{n\to\infty}\gamma_1(n,h)=\infty$, $h$ is the bandwidth satisfying $h=cn^\tau$ for some $c>0$, $\tau<0$. Then the constrained estimators $\bar{\beta}(x)$ and $\bar{m}_{sp}(x)$ as defined earlier, for some given constant $\beta_1(x)$ satisfying $\beta(x)\geq\beta_1(x)$, have the following properties,

1. when $\beta(x)>\beta_1(x)$, $\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\bar{\beta}(x)-\beta(x)\right) \xrightarrow{d} Z$.

2. when $\beta(x)=\beta_1(x)$, $\Pr\left(\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\bar{\beta}(x)-\beta(x)\right)<z\right)\to\Phi(z)\cdot 1_{\{z\geq 0\}}$.

3. when $\beta(x) \geq \beta_1(x)$, the semiparametric estimator has

$$\gamma_2(n,h)\,\sigma_m^{-1}(x)\left[\bar{m}_{sp}(x) - m(x) - B_m(x)\right] \xrightarrow{d} Z, \tag{2.37}$$

for some $\gamma_2(n,h)$ with similar properties as that in Theorem 4 and $\sigma_m(x) > 0$, where

$$B_m(x) = \frac{1}{2}h^2 m^{(2)}(x) \int v^2 k(v)\,dv + o_p\left(h^2\right), \tag{2.38}$$

the same as the asymptotic bias of $\tilde{m}_{sp}(x)$.

If we further assume that $\gamma_1(n,h)\,\sigma_\beta^{-1}\left(\beta(x) - \beta_1(x)\right) = b(x)$, for some real function $b(x)$, and denote $Z_{b(x)} = Z + b(x)$, then

4. $\lim_{n\to\infty} \gamma_1(n,h)\,\sigma_\beta^{-1}\left[\bar{\beta}(x) - \beta(x)\right] = Z_{b(x)} 1_{\left[Z_{b(x)}>0\right]} - b(x)$.

5. $\lim_{n\to\infty} \gamma_1(n,h)\,\sigma_\beta^{-1} E\left[\bar{\beta}(x) - \beta(x)\right] = \varphi(b(x)) + b(x)\Phi(b(x)) - b(x)$.

6. $\lim_{n\to\infty} Var\left[\left(\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\right)^{1/2}\bar{\beta}(x))\right] = \Phi(b(x)) + b(x)\varphi(b(x)) - \varphi^2(b(x)) - 2b(x)\varphi(b(x))\,\Phi(b(x)) + b^2(x)\Phi(b(x))\left[1 - \Phi(b(x))\right]$.

**Remark 6.** The powerful result shows that the estimation of $m(x)$ via the SP method is always a consistent estimator of the true function $m(x)$, independent of the specification of the model. While the NP estimator possesses this property as well, the parametric estimator considered in Theorem 2 does not enjoy this nice property. Parts 1 and 2 establish the asymptotic properties of the constrained slope estimator when the constraint is strict and when the equality constraint holds. Part 3 shows the asymptotic equivalence between constrained SP estimator and unconstrained SP estimator. The result for unconstrained estimator is first proved by Martins-Filho et al (2007). Part 4 considers the local asymptotics for the constrained slope estimator, with asymptotic bias and variance given in Parts 5 and 6.

### 2.5.2 Bagged Constrained Semiparametric Estimator

**Theorem 7.** Let an unconstrained estimator $\tilde{\beta}(x)$ of $\beta(x)$ and its bootstrap version $\tilde{\beta}^*(x)$ have the following asymptotic,

$$\begin{aligned} \gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}(x)-\beta(x)\right) &\xrightarrow{d} Z, \\ \gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}^*(x)-\tilde{\beta}(x)\right) &\xrightarrow{d} Z, \end{aligned} \tag{2.39}$$

where $Z$ is a standard normal random variable, $\lim_{n\to\infty}\gamma_1(n,h)=\infty$, $h$ is the bandwidth satisfying $h=cn^{\tau}$ for some $c>0$, $\tau<0$. Let $\beta_1(x)$ satisfy

$$\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\beta(x)-\beta_1(x)\right)=b(x), \tag{2.40}$$

where $b(\cdot)$ is some real function. For the bagged version of $\bar{\beta}(x)$, $\hat{\beta}(x)\equiv E^*\bar{\beta}^*(x)$, as defined earlier we have

1. $\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\hat{\beta}(x)-\beta(x)\right)\xrightarrow{d} Z\left[1-\Phi\left(-b(x)-Z\right)\right]+\varphi\left(-b(x)-Z\right).$

2. $\lim_{n\to\infty}\gamma_1(n,h)\,\sigma_\beta^{-1}E\left[\hat{\beta}(x)-\beta(x)\right]=2\varphi*\varphi\left(-b(x)\right)-b(x)\,\Phi*\varphi\left(-b(x)\right).$

3. $\lim_{n\to\infty} Var\left[\left(\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\right)^{1/2}\hat{\beta}(x)\right]=1+\Phi^2*\varphi''\left(-b(x)\right)+\Phi^2*\varphi\left(-b(x)\right)-2b\Phi^2*\varphi'\left(-b(x)\right)+b^2(x)\,\Phi^2*\varphi\left(-b(x)\right)+\varphi^2*\varphi\left(-b(x)\right)-2\Phi*\varphi''\left(-b(x)\right)-2\Phi*\varphi\left(-b(x)\right)+2b(x)\,\Phi*\varphi'\left(-b(x)\right)-2\varphi*\varphi'\left(-b(x)\right)+2\left(\Phi\cdot\varphi\right)*\varphi'\left(-b(x)\right)-2b(x)\left(\Phi\cdot\varphi\right)*\varphi\left(-b(x)\right)-\left[2\varphi*\varphi\left(-b(x)\right)-b(x)\,\Phi*\varphi\left(-b(x)\right)\right]^2.$

4. If $\gamma_2(n,h)\,\tilde{\sigma}_m^{-1}(x)\left(\tilde{m}_{sp}(x)-m(x)-B_m(x)\right)\xrightarrow{d} Z$, where

$$B_m(x)=\frac{1}{2}h^2m^{(2)}(x)\int v^2k(v)\,dv+o_p\left(h^2\right) \tag{2.41}$$

is the asymptotic bias $\tilde{m}_{sp}(x)$, and $\gamma_2(n,h)$ follows similar conditions as $\gamma_1(n,h)$, then

$$\gamma_2(n,h)\,\sigma_m^{-1}(x)\left[\hat{m}_{sp}(x)-m(x)-B_m(x)\right]\xrightarrow{d} Z. \tag{2.42}$$

**Remark 7.** Theorem 7 shows that the bagging constrained semiparametric estimator $\hat{m}_{sp}(x)$ is asymptotically equivalent to its unconstrained counterpart. The dependence of the asymptotic distribution on the drift function $b(\cdot)$ remains the same as those in Theorem 5. Thus Remark 5 applies here, which we do not intend to repeat.

## 2.6 Simulation

We perform Monte Carlo simulation to examine the finite sample properties of our proposed bagging NP and SP estimators. We consider the following data generating process (DGP) that features monotonicity in the conditional mean of $y_t$ given $x_t$

$$\text{DGP}: \ y_{t+1} = ax_t^3 + e_{t+1}, \tag{2.43}$$

where $e_t \sim$ i.i.d.$\mathcal{N}(0,1)$, $x_t \sim$ i.i.d.$\mathcal{N}\left(\frac{1}{2}, \sigma_x^2\right)$, with $\sigma_x^2 = 2, 3, 4, 5$ and $a = 0.0128$. We replicate the process for 100 times, with $J = 100$ bootstrap samples taken for bagging in each replication. We take $n = 200$ observations for in-sample estimation. The 1000 out-of-sample forecast values of $\hat{y}$ from the various forecasting models presented in the next subsection are computed over the 1000 equidistant evaluation points on the realized support of $\{x_t\}_{t=1}^n$ generated from $\mathcal{N}\left(\frac{1}{2}, \sigma_x^2\right)$. For the NP and SP estimators, we use cross-validation to select a bandwidth that minimizes the integrated mean squared error and use this same bandwidth for the 100 bootstrap samples generated within each replication.

Consider a forecasting model

$$\text{Model}: \ y_{t+h} = m(x_t) + u_{t+h}. \tag{2.44}$$

For a given evaluation predictor value $x$, we are interested in forming a forecast $\hat{y}_{n+h} = m_{n,h}(x|I_n)$, where $I_n = \{x_{n_0}, ..., x_n, y_{n_0}, ..., y_n\}$ is used to estimate a model. In the

Monte Carlo simulation of this section, $h = 1$ and we fix both $n_0 = 1$ and $n = 200$, and estimate various models using the $R \equiv n - n_0 + 1$ observations. Then we take 1000 equidistant fixed evaluation points $\{x\}_1^{1000}$ on a range of $\mathcal{N}\left(\frac{1}{2}, \sigma_x^2\right)$. The same 1000 equidistant evaluation points are used for all 100 Monte Carlo replications. In each Monte Carlo replication $i$ ($i = 1, \ldots, 100$), 1000 values of $\left\{\hat{m}^{(i)}(x)\right\}$ are computed at each of 1000 $x$ values, and also 1000 values of $\left\{\hat{u}^{(i)}(x) \equiv 0.0128x^3 - \hat{m}^{(i)}(x)\right\}$ are computed in each replication $i$. We compute the Monte Carlo average of the squared $\hat{u}^{(i)}(x)$ over $i$ for each evaluation point $x$, $\frac{1}{100}\sum_{i=1}^{100} \hat{u}^{(i)2}(x) \equiv \hat{u}^2(x)$. Then we use the 1000 values of the squared forecast errors $\left\{\hat{u}^2(x)\right\}_1^{1000}$ to compute the evaluation criteria discussed later in Section 2.6.2. The number of observations for in-sample estimation is $R \equiv n - n_0 + 1 = 200$, and the number of the out-of-sample evaluation points is $P = 1000$.

In the empirical application of Section 2.7, $h = 1, 6, 12$, and we move the time $t = n$ at which a pseudo out-of-sample forecast is made. We use a rolling window of fixed size $R = 120$ months from $t = n_0$ ($\equiv n - R + 1$) to $t = n$ for in-sample estimation of a model. We then compute $h$-month ahead forecasts of the equity premium $y_{n+h}$, with $n$ moving forward from 1960M1 to 2005M12, resulting in the total of $P = (552 - h)$ evaluation points over the 46 years. Once $\hat{y}_{n+h}$ is obtained, we define the forecast error $\hat{u}_{n+h} \equiv y_{n+h} - \hat{y}_{n+h}$. We use the $(552 - h)$ squared forecast errors $\left\{\hat{u}_{n+h}^2\right\}_{n=1960M1}^{2005M12-h}$ to compute the evaluation measures discussed later. The number of observations for in-sample estimation is $R \equiv n - n_0 + 1 = 120$, and the number of the out-of-sample evaluation points is $P = 552 - h$.

### 2.6.1 Forecasting Models

We consider the historical mean model (HM) as a benchmark

$$m_{n,h}^{\text{HM}}(x|I_n) = \frac{1}{R}\sum_{t=n_0}^{n} y_t.$$

and three linear regression models denoted as L, L-P, and L-P-B:

$$\begin{aligned} m_{n,h}^{\text{L}}(x|I_n) &= \tilde{\alpha} + \tilde{\beta}x, \\ m_{n,h}^{\text{L-P}}(x|I_n) &= \bar{\alpha} + \bar{\beta}x, \\ m_{n,h}^{\text{L-P-B}}(x|I_n) &= \hat{\alpha} + \hat{\beta}x, \end{aligned}$$

where $\left(\tilde{\alpha}, \tilde{\beta}\right)$ is the unconstrained OLS estimators, $\bar{\beta} = \max\left(\tilde{\beta}, 0\right)$, $\bar{\alpha} = \bar{y}_n - \bar{\beta}\bar{x}_n$, $\hat{\beta} = \frac{1}{J}\sum_{j=1}^{J}\bar{\beta}^{*(j)}$ with $\bar{\beta}^* = \max\left(\tilde{\beta}^*, 0\right)$, and $\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n$. Nonparametric models include LLLS forecast (NP), LLLS forecast with positive slope constraint (NP-P), the bagged LLLS forecast with positive slope constraint (NP-P-B)

$$\begin{aligned} m_{n,h}^{\text{NP}}(x|I_n) &= \bar{y}(x) - \tilde{\beta}(x)\left[\bar{x}(x) - x\right], \\ m_{n,h}^{\text{NP-P}}(x|I_n) &= \bar{y}(x) - \bar{\beta}(x)\left[\bar{x}(x) - x\right], \\ m_{n,h}^{\text{NP-P-B}}(x|I_n) &= \bar{y}(x) - \hat{\beta}(x)\left[\bar{x}(x) - x\right], \end{aligned}$$

and the monotonicity-constrained NP model proposed by Hall and Huang (2001) (NP-HH)

$$m_{n,h}^{\text{NP-HH}}(x|I_n) = \sum_{t=1}^{n-h}\hat{p}_t A_t(x) y_{t+h}.$$

Semiparametric models include SP, SP-P, and SP-P-B

$$\begin{aligned} m_{n,h}^{\text{SP}}(x|I_n) &= \tilde{m}_{sp}(x) \text{ as defined in (2.23)}, \\ m_{n,h}^{\text{SP-P}}(x|I_n) &= \bar{m}_{sp}(x) \text{ as defined in (2.26)}, \\ m_{n,h}^{\text{SP-P-B}}(x|I_n) &= \hat{m}_{sp}(x) \text{ as defined in (2.27)}. \end{aligned}$$

### 2.6.2 Evaluation Criteria

As discussed earlier, the Monte Carlo mean (averaged over 100 replications) of squared errors $\left\{\hat{u}^2(x)\right\}_1^P$ for each of $P$ evaluation points will be used to compute the evaluation criteria. We consider two such criteria. The first criterion is based on the mean of the squared errors (averaged over $P = 1000$ evaluating $x$ points) of model M

$$MSE_{\mathrm{M}} = \frac{1}{P}\sum_{\forall x}\hat{u}^2(x). \tag{2.45}$$

Further we compute the percentage reduction in the MSE of a model M ($MSE_{\mathrm{M}}$) relative to that of the historical mean model ($MSE_{\mathrm{HM}}$) by the following formula,

$$100R^2 = 100 \times \left(1 - \frac{MSE_{\mathrm{M}}}{MSE_{\mathrm{HM}}}\right). \tag{2.46}$$

This is the out-of-sample $R^2$ (multiplied by 100) as reported in Campbell and Thompson (2008).

The second criterion is new. It provides a better view of the whole predictive distribution of the squared forecast errors $\left\{\hat{u}^2(x)\right\}_1^P$. Statistical criteria such as MSE, $R^2$ and likelihood values are based on a summary statistic (e.g., mean) of $\left\{\hat{u}^2(x)\right\}_1^P$. Instead, as suggested in Granger (1999), a more desirable procedure is to associate an economic value with $\left\{\hat{u}^2(x)\right\}_1^P$ rather than just a summary statistic. The economic value of a model can be associated with a cost or a utility, which can then be compared using the second order stochastic dominance (SOSD) of the predictive distributions of $\left\{\hat{u}^2(x)\right\}_1^P$ for competing models. Denote the CDF of squared forecast errors $\left\{\hat{u}^2(x)\right\}_1^P$ from Model M as $F^{\mathrm{M}}(\cdot)$. We define the SOSD criterion as

$$SOSD^{\mathrm{M}}(r) = \int_0^r \left[F^{\mathrm{M}}(s) - F^{\mathrm{HM}}(s)\right] ds, \qquad r > 0, \tag{2.47}$$

where HM is taken as the benchmark model and the CDFs are estimated by their empirical distributions $F(s) = \frac{1}{P}\sum_{\forall x} 1_{\{\hat{u}^2(x)\leq s\}}$.

We can show (not presented here for space but available from the authors) that, if $SOSD^{\mathrm{M}}(r) > 0$ for all $r > 0$, then $E\left(\hat{u}_{\mathrm{M}}^2\right) < E\left(\hat{u}_{\mathrm{HM}}^2\right)$. Therefore, the second-order-stochastic dominance implies the mean-squared-error dominance (but not vice versa). Hence SOSD would also imply the dominance in $100R^2$.

Compared to $100R^2$ which measures the percentage gain in the mean of squared forecast errors, $SOSD^{\mathrm{M}}(r)$ delivers more information on the entire distribution of the squared forecast errors from Model M. For example, when $SOSD^{\mathrm{M}}(r)$ is positive for all positive $r$, it implies that Model M produces squared forecast errors that are relatively smaller than those of the benchmark model. The role of $SOSD(r)$ becomes more significant when $100R^2$ cannot differentiate the relative performances of the models under comparison. Following McFadden (1989), Granger (1999), and Linton et al (2005), we report the average (avg) and the maximum (max) of $SOSD(r)$ over $r$ in Table 1 and Table 2, in addition to $100R^2$.

While we have compared the empirical distribution of *squared* forecast errors $\left\{\hat{u}^2(x)\right\}_1^P$, the SOSD measure will be identical if we compare the empirical distributions of the *absolute* forecast errors $\{|\hat{u}(x)|\}_1^P$. We can also show that, if $SOSD^{\mathrm{M}}(r) > 0$ for all $r > 0$, then $E(|\hat{u}_{\mathrm{M}}|) < E(|\hat{u}_{\mathrm{HM}}|)$. Therefore, the second-order-stochastic dominance implies the mean-absolute-error dominance (but not vice versa). In fact, we can show that, if $SOSD^{\mathrm{M}}(r) > 0$ for all $r > 0$, then $E(c(\hat{u}_{\mathrm{M}})) < E(c(\hat{u}_{\mathrm{HM}}))$ for any symmetric convex function $c(\cdot)$. We will demonstrate the use of our new forecasting evaluation criterion using "the SOSD plots" (as shown Figure 2) in the empirical application in Section 2.7.

### 2.6.3 Simulation Results

The simulation results are presented in Table 1. We summarize the findings as follows:

First, note the varying slope of the cubic curve in the DGP in (2.43). A larger value of $\sigma_x^2$ would expand the range of the evaluation points $\{x\}$ to the steeper area of the cubic curve. When $\sigma_x^2 = 2$ (small), the evaluation points will be mostly in the flat area of the cubic curve. That corresponds to the area with small values of $b$ near zero in Figure 1c. The reduction in AMSE (hence the gain in $100R^2$) would be large as shown in Figure 1c. Table 1 confirms this by showing that the gains from imposing the monotonicity constraints and from bagging is large in this case. $100R^2$ is 43.7, 53.5, 55.7 for each of SP, SP-P, SP-P-B. The increase of these values is substantial. Similar observation can be made for $\text{avg}_r SOSD(r)$ and $\max_r SOSD(r)$. When $\sigma_x^2 = 5$ (large), the evaluation points will be in a wider range of the cubic curve including the areas with steeper slope. That corresponds to the area with large values of $b$ in Figure 1c, where the reduction in AMSE (hence the gain in $100R^2$) is small. Table 1 again confirms that by showing the small gains from imposing the monotonicity constraints and from bagging. For example, $100R^2$ is 94.6, 94.7, 94.8 for each of SP, SP-P, SP-P-B. The increase of these values is negligible. Similar observation can be made for $\text{avg}_r SOSD(r)$ and $\max_r SOSD(r)$. The same pattern is observed for NP, NP-P, NP-P-B when they are compared with small and large values of $\sigma_x^2$.

Second, note also the varying curvature of the cubic curve in DGP, which exhibits stronger nonlinearity as we move further away from the inflection point. Therefore the nonlinearity is stronger with a larger value of $\sigma_x^2$. When the range of the evaluation $x$ points expands to the stronger nonlinear part of the cubic curve, there is larger

gains by using nonlinear models (NP and SP) over the linear model (L). When $\sigma_x^2 = 5$ (large), $100R^2$ is about 64 for L, while it is much larger, nearly 95 for NP and SP. Similar observation can be made for $\text{avg}_r SOSD(r)$ and $\max_r SOSD(r)$. When $\sigma_x^2 = 2$ (small), the evaluation points will be near the flat part of the curve where nonlinearity is weak. And there, L is even better than the nonlinear NP/SP forecasts in all three criteria, $100R^2$, $\text{avg}_r SOSD(r)$ and $\max_r SOSD(r)$. Interestingly though, as remarked in the previous paragraph, the improvement by imposing the monotonicity constraint and by using bagging is much stronger for the nonlinear NP/SP models than for the linear model. There is little gain from L to L-P to L-P-B, while the gains are substantial from NP to NP-P to NP-P-B and also from SP to SP-P to SP-P-B.

Third, the constraint helps with NP and SP models, as seen that $R^2$ gets larger in NP-P, NP-HH and SP-P. This improvement in $R^2$ is due to the accuracy gain in estimation that is achieved at points where monotonicity is violated. At points where monotonicity is met, constrained model and unconstrained model perform the same since the constraint is not binding. The extent of the improvement from imposing the constraint depends on (i) the frequency of points where violations of constraints occur and (ii) the magnitude of the violations at these points. Monotonicity is satisfied in the estimated linear models (when $\sigma_x^2$ is not too small) so that L and L-P perform more or less the same.

Fourth, the simple monotonicity constrained NP-P model is generally better than NP-HH of Hall and Huang (2001). When bagging is added, NP-P-B is always better than NP-HH.

Fifth, bagging enhances the performance of the constrained NP/SP models with no exception. This confirms the theory that bagging reduces variance (Figure 1a), while incurring a small bias (Figure 1b). It is also found that, with bagging, our

constrained models, NP-P-B and SP-P-B, outperform NP-HH. Note that bagging does not improve for the linear model as much, because the monotonicity constraint is more likely to be met for L and because the constraint is less likely to be violated globally than locally.

Sixth, a positive value of $100R^2$ for a model indicates that the benchmark HM is dominated by the model. It is clear that all models are better than HM for all values of $\sigma_x^2$. However, this may be due to the design in our simulation. In empirical application to predicting equity premium in the next section, it will be shown (Table 2) that HM is indeed very hard to be beaten by a linear model even with the monotonicity constraint and bagging. This is reflected in the paper title of Campbell and Thompson (2008), and is a reason that HM has been taken as a benchmark in the vast literature on financial return predictability. Nevertheless, we will see in the next section that NP and SP can easily beat the HM, and even more easily with the monotonicity constraint and bagging.

Seventh, the nonlinear models, NP and SP, are substantially better than L when $\sigma_x^2$ is not too small. This signals the serious nonlinearity in the DGP. NP and SP are quite competing, with NP often slightly better than SP, due to the fact that the linear guide for SP is not present in the DGP. However, it is interesting to see that, once the monotonicity constraint is imposed, SP-P is always better than NP-P and also SP-P-B is always better than NP-P-B. It seems the constraint and bagging help SP more than NP.

Eighth, the role of SOSD is expected to be more significant when $100R^2$ cannot distinguish the relative performance of models under comparison because the SOSD looks at the entire predictive distribution of the squared forecast errors rather than just their mean. However, we do not see such a case yet from using the current simulation design. In Table 1, SOSD generally tends to convey the same signal about the

forecasting models as $100R^2$ does. We will be able to discuss the advantage of the distribution measure (SOSD) over the mean measure ($100R^2$) using Figure 2 for our empirical application in the next section.

Table 1 About Here

All of the above simulation results are consistent with the asymptotic results of Sections 3, 4, 5. It would be interesting to examine how the theory applies in practice in actual economic data application where the DGP is not known. In the next section, we examine this in forecasting the U.S. equity premium.

## 2.7 Application: Predicting the Equity Premium

As noted by Fama and French (2002), equity premium (the difference between the expected return on the market portfolio of common stocks and the rate of return on risk-free assets such as short term T-bills) plays an important role in decisions of portfolio allocation, in estimating the cost of capital, in debate of investing social security funds in stocks, and in many other economic and financial applications. However, the predictability of equity premium has been an unsettled issue in the financial literature as reviewed by Campbell and Thompson (2008) and many references therein.

Goyal and Welch (2008) examine various predictors that have been suggested as good instruments in the equity premium prediction literature but report their poor performance in both in-sample and out-of-sample forecasts relative to the historical mean of stock returns. Campbell and Thompson (2008) introduce a perspective of a real-world investor who would use a prior belief on the regression slope coefficient such that it must satisfy the expected sign. This simple but sensible sign constraint leads to a better out-

of-sample performance of predictors that have significant in-sample forecasting power. Chen and Hong (2009) went further to argue that such sign restriction imposed by Campbell and Thompson (2008) is a form of nonlinearity and suggest to use NP methods instead of linear models to form forecast of stock returns. They confirm the conclusion of Campbell and Thompson (2008).

As an alternative to these approaches, we impose the sign restriction on the local slope coefficients in estimation of the NP and SP forecast models. In that sense, we combine the two ideas of Campbell and Thompson (2008) and Chen and Hong (2009), imposing monotonicity on NP/SP models. We compare linear models of Goyal and Welch (2008), Campbell and Thompson (2008), Hillebrand et al (2009), with our proposed NP and SP models with constraints imposed and also with bagging implemented. The out-of-sample forecasting comparison is based on $100R^2$ and SOSD, relative to the historical mean return forecast. John Campbell and Sam Thompson kindly share their data in our study. We consider using one predictor at a time and impose their sign restrictions on the slope parameters, but locally for the NP and SP models. For details on data description and the sign restrictions, we refer to Campbell and Thompson (2008).

Our dependent variable $y$ to be forecast is the annualized (compounded for 12 months) equity premium on the S&P500 returns over the short term T-Bill rate, and the predictor variable $x$ is one of the following four predictors: smoothed earning-price ratio ($se/p$), yields on 3-Month Treasury Bill on the secondary market ($t\text{-}bill$), long term yields on U.S. government bonds ($lty$), and default spread ($ds$). Both $y$ and $x$ series are in monthly frequency.

As discussed earlier in Section 2.6, the in-sample estimation starts from 1950M1 and the first forecast begins in 1960M1. We keep a fixed window of in-sample size of 120 observations and roll the in-sample estimation window forward till the last available

observation on 2005M12. To evaluate various HM/L/NP/SP models considered in this paper, we report out-of-sample $100R^2$ together with $\text{avg}_r SOSD(r)$ and $\max_r SOSD(r)$ measures defined in Section 2.6.2. In Table 2 and Figure 2, we only present the results for $h = 1$ as the results for $h = 6, 12$ (available upon request) show the same patterns with respect to nonlinearity and monotonicity.

### 2.7.1 Empirical Results

We summarize the findings from Table 2 as follows:

First, a salient feature of the results is the nonlinear predictability of the equity premium, which confirms earlier results of Chen and Hong (2009). For all four predictors, NP and SP models perform much better than L (and better than HM too!), with an improvement in $R^2$ over 10% achieved by SP-P-B. The only exception is NP-HH, which is worse than linear models for *se/p* and *ds*. The impressive performance of parametrically guided SP models confirms the earlier conclusion by Martins-Filho et al (2007). Except with *se/p*, linear models are worse than HM, even though imposing constraint enhances their performance.

Second, another salient feature is the monotonicity, which improves the forecasting ability of NP and SP models although the improvement is sometimes small. This small improvement is due to mainly two facts: (1) the computed evaluation criteria $100R^2$ and SOSD, are aggregated (global) measures such that some significant local improvement may be averaged down, and (2) inherent uncertainty in the noise component of a model dominates the parameter estimation uncertainty in the signal component of the model in order of $\gamma(n, h)$ as presented in Theorems 2-7. The first fact is that, at many of $P$ out-of-sample months, the monotonicity constraint is locally met (i.e., not binding) and thus no improvement will be achieved by imposing such a

constraint for those months. It is at these (possibly many) data points that the improvement of forecasts made over other data points is offset, because our evaluation criteria are the averages over all $P$ points. The second fact dictates that parameter estimation error vanishes at a rate $\gamma(n,h)$ as sample size increases but innovation uncertainty will not. The constraint and bagging can reduce the parameter estimation error and improve forecasts for a finite sample size, but their contribution vanishes as the sample size increases.

Third, bagging improves the constrained NP and SP forecasts. The improvement of $R^2$ is around 1-2%. For example, for *ds,* NP-P-B improves $100R^2$ by more than 2.1% compared to NP-P. Bagging makes all constrained SP models work better. For one case with *se/p* where bagging makes little improvement, it may be due to the fact that we are outside of the range of local drift parameter $b(x)$ where bagging can improve when the constraint is met, as depicted for the linear case in Figure 1c in Section 2.3.

Fourth, the average SOSD and maximum SOSD measures in Table 2 are consistent with $100R^2$. SOSD also favors constrained models over unconstrained ones and shows that bagging helps to improve the forecasting performance of constrained models.

Table 2 About Here

We summarize the findings from Figure 2 as follows:

Figure 2 shows plots of $SOSD(r)$ as a function of squared forecasting errors $r$, and thus will be called "the SOSD plot". The $x$-axis is $r$ for the squared forecast error while the $y$-axis is $SOSD(r)$ as defined in (2.47). The SOSD plots show *where* the forecast gains are achieved for different sizes of forecast errors. The size of forecast error is measured in square in Figure 2, while it can be measured in any norm such as modulus.

Figure 2 reports the SOSD plots for *lty*. The SOSD plots for the other three predictors are similar in pattern and in ranking and so are not presented here. Figure 2 shows that SP-P-B produces many more moderately sized forecast errors than other models because $SOSD(r)$ increases steeply over the moderate values of $r$ (between 0.05 and 0.1) and then flattened for large values of $r$ (large size forecast errors). In other word, the SOSD plot reveals that constrained models perform better by reducing the magnitude of forecasting errors. Hence, the sensible constraints would help avoiding big mistakes.

The SOSD plots in Figure 2 show that $SOSD(r) > 0$ for all $r > 0$ for all NP and SP models. That means, for *lty*, these models stochastically dominate the HM model in any symmetric convex cost (loss) functions. To the contrary, $SOSD(r) < 0$ for all $r > 0$ for all three L models even with the monotonicity constraint and bagging. That means, for *lty,* the L models are stochastically dominated by HM in any symmetric convex cost functions. Interestingly, for NP-HH, Figure 2 shows that $SOSD(r)$ crosses zero once from below and stay above zero for large value of $r$ ($> 0.07$). This indicates that NP-HH is worse than HM when the forecast error size can be small (likely when the stock market is calm), but NP-HH becomes better than HM when the squared forecast errors are large (likely when the stock market are volatile). With this in mind, looking at the SOSD plots again for the linear models (L, L-P, L-P-B), we note that, for all sizes of the forecast errors, whether small or large, the linear models using *lty* make poorer forecasts than HM.

Figure 2 About Here

This type of forecast evaluation and comparison is not possible with the mean-based measure like $100R^2$. The novelty of the SOSD plots is that we can examine the

entire predictive distribution of the squared forecast errors, through which we are enabled to see how/when models are performing in forecasting over the different magnitude of the forecast errors and over different levels of market volatility.

## 2.8 Conclusions

Incorporating valuable economic information in economic modeling and forecasting deserves more attention in both theoretical and applied research. This paper considers nonparametric and semiparametric regression models with imposing such economic constraints as monotonicity. Our approach is an alternative approach to Hall and Huang (2001), Racine, Parmeter and Du (2009), and Henderson and Parmeter (2009). It is based on bagging, as in Hillebrand et al (2009), that improves the simple constrained linear regression model considered in Campbell and Thompson (2008). It is based on nonparametric models so that possible model misspecification of neglecting nonlinearity may be avoided. It reduces the computational time by eschewing the issue of solving weights to training units through the optimization problem considered in Hall and Huang (2001). Asymptotic properties of our bagging constrained NP and SP estimators and forecasts are established. Monte Carlo simulations are conducted to show their finite sample performance which demonstrates the practical merits of using our proposed methods.

We introduce a new forecasting evaluation criterion based on the second order stochastic dominance in the size of forecast errors, which enables us to compare the competing forecasting models over different sizes of forecast errors. The size of forecast errors may be measured in square, in modulus, or in any norm. The new SOSD criterion can compare forecasting models via the entire predictive distributions of a norm of

the forecast errors, e.g., over small size errors, moderate size errors, or big size errors, as demonstrated using our empirical results for the equity premium prediction application. With the use of new forecasting evaluation criterion, it is seen that imposing monotonicity constraints can mitigate the chance of making the large size forecast errors.

We apply the proposed approach for imposing economic constraints to predict the U.S. equity premium and show its usefulness likely under high market volatility. Although the predictability of equity premium has been an unsettled issue, our work together with those of Campbell and Thompson (2008) and Hillebrand et al (2009) reveal the value of constraints in economic modeling and forecasts.

Our results also confirm Chen and Hong (2009) that SP models usually outperform NP models, and thus should incite the applications of the SP models in future economic and financial research.

# Appendix

## A. Proof of Main Theorems

**Proof of Theorem 1.** (1) By the definition of $\bar{\beta}$, it is clear that it cannot take values less than $\beta_1$, which implies that $F_{\bar{\beta}}(z) = 0$ if $z < \beta_1$. When $z = \beta_1$, we have $F_{\bar{\beta}}(z) = \Pr(\bar{\beta} < \beta_1) + \Pr(\bar{\beta} = \beta_1) = \Pr\left(\tilde{\beta} \leq \beta_1\right) = F_{\tilde{\beta}}(\beta_1) = F_{\tilde{\beta}}(z)$. When $z > \beta_1$, $F_{\bar{\beta}}(z) = \Pr(\bar{\beta} \leq z) = \Pr(\bar{\beta} < \beta_1) + \Pr(\bar{\beta} = \beta_1) + \Pr(\beta_1 < \bar{\beta} \leq z) = F_{\tilde{\beta}}(\beta_1) + \Pr\left(\beta_1 < \tilde{\beta} \leq z\right) = F_{\tilde{\beta}}(z)$.

(2) Note that

$$\begin{aligned} E\bar{\beta} &= \int_{-\infty}^{\infty} z dF_{\bar{\beta}}(z) = \int_{-\infty}^{\beta_1} z dF_{\bar{\beta}}(z) + \int_{\beta_1}^{\infty} z dF_{\bar{\beta}}(z) \\ &= \beta_1 F_{\bar{\beta}}(\beta_1) + \int_{\beta_1}^{\infty} z dF_{\tilde{\beta}}(z) = \beta_1 F_{\tilde{\beta}}(\beta_1) + \int_{\beta_1}^{\infty} z dF_{\tilde{\beta}}(z) \\ &\geq \int_{-\infty}^{\beta_1} z dF_{\tilde{\beta}}(z) + \int_{\beta_1}^{\infty} z dF_{\tilde{\beta}}(z) = E\tilde{\beta}, \end{aligned}$$

where the third equality makes use of the property of $F_{\bar{\beta}}(z)$ established in (1).

(3) Note that for $\ddot{\beta} = \bar{\beta}$ or $\tilde{\beta}$, we have $Var\left(\ddot{\beta}\right) = MSE\left(\ddot{\beta}\right) - \left[bias\left(\ddot{\beta}\right)\right]^2$.It is known from (1) that $bias\left(\bar{\beta}\right) \geq bias\left(\tilde{\beta}\right) \geq 0$, if $E\tilde{\beta} \geq \beta$. $Var\left(\bar{\beta}\right) \leq Var\left(\tilde{\beta}\right)$ will be implied from the fact which is stated in (4).

(4) The proof is parallel to that in (2). By definition,

$$\begin{aligned} MSE\left(\bar{\beta}\right) &= \int_{-\infty}^{\infty} (z-\beta)^2 dF_{\bar{\beta}}(z) = \int_{-\infty}^{\beta_1} (z-\beta)^2 dF_{\bar{\beta}}(z) + \int_{\beta_1}^{\infty} (z-\beta)^2 dF_{\bar{\beta}}(z) \\ &= (\beta_1-\beta)^2 F_{\bar{\beta}}(\beta_1) + \int_{\beta_1}^{\infty} (z-\beta)^2 dF_{\tilde{\beta}}(z) = (\beta_1-\beta)^2 F_{\tilde{\beta}}(\beta_1) + \int_{\beta_1}^{\infty} (z-\beta)^2 dF_{\tilde{\beta}}(z) \\ &\leq \int_{-\infty}^{\beta_1} (z-\beta)^2 dF_{\tilde{\beta}}(z) + \int_{\beta_1}^{\infty} (z-\beta)^2 dF_{\tilde{\beta}}(z) = MSE\left(\tilde{\beta}\right), \end{aligned}$$

where the inequality follows from $\beta \geq \beta_1$. □

**Proof of Theorem 2.** For any $z \in R$,

$$\begin{aligned} &\Pr\left(\gamma(n)\left(\bar{\beta}-\beta\right) < z\right) = \Pr\left(\gamma(n)\left(\max\{\tilde{\beta},\beta_1\}-\beta\right) < z\right) \\ = \;& \Pr\left(\gamma(n)\left(\max\{\tilde{\beta},\beta_1\}-\beta\right) < z | \tilde{\beta} < \beta_1\right) \times \Pr\left(\tilde{\beta} < \beta_1\right) \\ &+ \Pr\left(\gamma(n)\left(\max\{\tilde{\beta},\beta_1\}-\beta\right) < z | \tilde{\beta} \geq \beta_1\right) \times \Pr\left(\tilde{\beta} \geq \beta_1\right) \\ = \;& \Pr\left(\gamma(n)\left(\beta_1-\beta\right) < z\right) \times \Pr\left(\tilde{\beta} < \beta_1\right) + \\ &\Pr\left(\gamma(n)\left(\tilde{\beta}-\beta\right) < z | \tilde{\beta} \geq \beta_1\right) \times \Pr\left(\tilde{\beta} \geq \beta_1\right) \end{aligned}$$

in which, (i) when $\beta > \beta_1$,

$$\Pr\left(\gamma(n)\left(\beta_1-\beta\right) < z\right) \to \Pr\left(-\infty < z\right) = 1,$$

since $\lim_{n\to\infty}\gamma(n)=\infty$, and when $\beta=\beta_1$,

$$\Pr\left(\gamma(n)(\beta_1-\beta)<z\right)=\begin{cases}1, & \text{if } z>0\\ 0, & \text{if } z\leq 0\end{cases}$$

(ii)

$$\begin{aligned}&\Pr\left(\tilde{\beta}<\beta_1\right)=\Pr\left(\gamma(n)\left(\tilde{\beta}-\beta\right)<\gamma(n)(\beta_1-\beta)\right)\\ \to&\begin{cases}\Pr(Z<-\infty)=0, & \text{if } \beta>\beta_1\\ \Pr(Z<0)=F(0), & \text{if } \beta=\beta_1\end{cases}\end{aligned}$$

(iii)

$$\begin{aligned}&\Pr\left(\gamma(n)\left(\tilde{\beta}-\beta\right)<z|\tilde{\beta}\geq\beta_1\right)\\ =&\frac{\Pr\left(\gamma(n)\left(\tilde{\beta}-\beta\right)<z,\gamma(n)\left(\tilde{\beta}-\beta_1\right)\geq 0\right)}{\Pr\left(\gamma(n)\left(\tilde{\beta}-\beta_1\right)\geq 0\right)}\\ =&\frac{\Pr\left(\gamma(n)\left(\tilde{\beta}-\beta\right)<z,\gamma(n)\left(\tilde{\beta}-\beta\right)\geq\gamma(n)(\beta_1-\beta)\right)}{\Pr\left(\gamma(n)\left(\tilde{\beta}-\beta\right)\geq\gamma(n)(\beta_1-\beta)\right)}\\ =&\begin{cases}\frac{F(z)-F(0)}{1-F(0)}, & \text{if } z>0;\\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$

and (iv)

$$\begin{aligned}\Pr\left(\tilde{\beta}\geq\beta_1\right)&=1-\Pr\left(\tilde{\beta}<\beta_1\right)\\ &=1-\Pr\left(\gamma(n)\left(\tilde{\beta}-\beta\right)<\gamma(n)(\beta_1-\beta)\right)\\ &\to\begin{cases}1-\Pr(Z<-\infty)=1, & \text{if } \beta>\beta_1\\ 1-\Pr(Z<0)=1-F(0), & \text{if } \beta=\beta_1\end{cases}\end{aligned}$$

Therefore, combining (i)-(iv) leads to, (1) when $\beta>\beta_1$,

$$\Pr\left(\gamma(n)\left(\bar{\beta}-\beta\right)<z\right)\to 1\times 0+F(z)\times 1=F(z)$$

and (2) when $\beta=\beta_1$, for $z>0$,

$$\Pr\left(\gamma(n)\left(\bar{\beta}-\beta\right)<z\right)\to 1\times F(0)+\frac{F(z)-F(0)}{1-F(0)}\times(1-F(0))=F(z)$$

and for $z = 0$,

$$\Pr\left(\gamma(n)\left(\bar{\beta}-\beta\right) < z\right) \to 1 \times F(0) + 0 \times (1 - F(0)) = F(0).$$

When $z < 0$,

$$\Pr\left(\gamma(n)\left(\bar{\beta}-\beta\right) < z\right) \to 0.$$

To prove (3), note that

$$\gamma(n)\left(\bar{\beta}-\beta\right) = \gamma(n)(\beta_1-\beta) + \gamma(n)\left(\tilde{\beta}-\beta_1\right)1_{\left\{\gamma(n)(\tilde{\beta}-\beta_1)>0\right\}} \xrightarrow{d} Z_b 1_{\{Z_b>0\}} - b.$$

Therefore, we have (4)

$$E\left[Z_b 1_{\{Z_b>0\}} - b\right] = EZ1_{\{Z_b>0\}} + bE1_{\{Z_b>0\}} - b = \phi(b) + b\Phi(b) - b,$$

by Lemma 1, and (5)

$$Var\left[Z_b 1_{\{Z_b>0\}} - b\right] = Var\left[Z_b 1_{\{Z_b>0\}}\right] = E\left\{\left[Z_b 1_{\{Z_b>0\}}\right]^2\right\} - \left\{E\left[Z_b 1_{\{Z_b>0\}}\right]\right\}^2.$$

We need to find

$$\begin{aligned} E\left\{\left[Z_b 1_{\{Z_b>0\}}\right]^2\right\} &= E\left\{\left[(Z+b)1_{\{Z_b>0\}}\right]^2\right\} \\ &= EZ^2 1_{\{Z_b>0\}} + b^2 E1_{\{Z_b>0\}} + 2bE\left[Z1_{\{Z_b>0\}}\right] \\ &= \Phi(b) - b\phi(b) + b^2\Phi(b) + 2b\phi(b) \\ &= \Phi(b) + b\phi(b) + b^2\Phi(b), \end{aligned}$$

where in the third equality we used (i) $E1_{\{Z_b>0\}} = \Phi(b)$ and (ii) $E\left[Z1_{\{Z_b>0\}}\right] = \phi(b)$ and (iii) $EZ^2 1_{\{Z_b>0\}} = -b\phi(b) + \Phi(b)$ by Lemma 1. Combining the results leads to (5).

□

**Proof of Theorem 3.** (1) Note that we can write

$$\begin{aligned} \hat{\beta} &= E^*\bar{\beta}^* = E^*\left[\max\left\{\tilde{\beta}^*, \beta_1\right\}\right] = E^*\left[\tilde{\beta}^* 1_{\left\{\tilde{\beta}^*\geq\beta_1\right\}} + \beta_1 1_{\left\{\tilde{\beta}^*<\beta_1\right\}}\right] \\ &= E^*\left[\tilde{\beta}^* 1_{\left\{\tilde{\beta}^*\geq\beta_1\right\}}\right] + \beta_1 E^*\left[1_{\left\{\tilde{\beta}^*<\beta_1\right\}}\right]. \end{aligned}$$

(i) The first term can be decomposed as,

$$\begin{aligned} E^*\left[\tilde{\beta}^* 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] &= E^*\left[\left(\tilde{\beta}^* - \tilde{\beta} + \tilde{\beta}\right) 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] \\ &= E^*\left[\left(\tilde{\beta}^* - \tilde{\beta}\right) 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] + \tilde{\beta} E^*\left[1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] \end{aligned}$$

Note that we have

$$\begin{aligned} E^*\left[\left(\tilde{\beta}^* - \tilde{\beta}\right) 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] &= \frac{\sigma}{\gamma(n)} E^*\left[\gamma(n)\sigma^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right) 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] \\ &= O\left(\frac{1}{\gamma(n)}\right) E^*\left[\gamma(n)\sigma^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right) 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] = O\left(\frac{1}{\gamma(n)}\right), \end{aligned}$$

since

$$\begin{aligned} &\left|E^*\left[\gamma(n)\sigma^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right) 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right]\right| \leq E^*\left|\gamma(n)\sigma^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right) 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right| \\ \leq\ & E^*\left[\left|\gamma(n)\sigma^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right)\right| \left|1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right|\right] \leq E^*\left|\gamma(n)\sigma^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right)\right| = O(1). \end{aligned}$$

And note that,

$$\begin{aligned} E^*\left[1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] &= E^*\left[1_{\left\{\gamma(n)\left(\tilde{\beta}^* - \tilde{\beta}\right) \geq \gamma(n)\left(\beta_1 - \tilde{\beta}\right)\right\}}\right] = E^*\left[1_{\left\{\gamma(n)\sigma^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right) \geq \gamma(n)\sigma^{-1}\left(\beta_1 - \beta\right) + \gamma(n)\sigma^{-1}\left(\beta - \tilde{\beta}\right)\right\}}\right] \\ &= 1 - E^*\left[1_{\left\{\gamma(n)\sigma^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right) < \gamma(n)\sigma^{-1}\beta_1 - \beta) + \gamma(n)\sigma^{-1}\left(\beta - \tilde{\beta}\right)\right\}}\right] = 1 - \Phi(-b - Z) + O\left(\frac{1}{n}\right). \end{aligned}$$

That is, we have $\tilde{\beta} E^*\left[1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] = \tilde{\beta}\left[1 - \Phi(-b - Z)\right] + O\left(\frac{1}{n}\right)$.

(ii) The second term,

$$\beta_1 E^*\left[1_{\left\{\tilde{\beta}^* < \beta_1\right\}}\right] = \beta_1\left(1 - E^*\left[1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right]\right) = \beta_1 \Phi(-b - Z) + O\left(\frac{1}{n}\right).$$

Combining (i) and (ii) leads to

$$\begin{aligned} \hat{\beta} &= E^*\left[\left(\tilde{\beta}^* - \tilde{\beta}\right) 1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] + \tilde{\beta} E^*\left[1_{\left\{\tilde{\beta}^* \geq \beta_1\right\}}\right] \\ &= \tilde{\beta}\left[1 - \Phi(-b - Z)\right] + \beta_1 \Phi(-b - Z) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{\gamma(n)}\right) \\ &= \tilde{\beta} + \left(\beta_1 - \tilde{\beta}\right) \Phi(-b - Z) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{\gamma(n)}\right). \end{aligned}$$

(2) Write

$$\begin{aligned}\gamma(n)\sigma^{-1}\left(\hat{\beta}-\beta\right) &= \gamma(n)\sigma^{-1}\left(E^*\left[\tilde{\beta}^* 1_{\left\{\tilde{\beta}^*\geq\beta_1\right\}}\right]+\beta_1 E^*\left[1_{\left\{\tilde{\beta}^*<\beta_1\right\}}\right]-\beta\right)\\ &= \gamma(n)\sigma^{-1}\left(E^*\left[\left(\tilde{\beta}^*-\beta\right)1_{\left\{\tilde{\beta}^*\geq\beta_1\right\}}\right]+(\beta_1-\beta)E^*\left[1_{\left\{\tilde{\beta}^*<\beta_1\right\}}\right]\right).\end{aligned}$$

We have (i)

$$\begin{aligned}&\gamma(n)\sigma^{-1}\left(E^*\left[\left(\tilde{\beta}^*-\beta\right)1_{\left\{\tilde{\beta}^*\geq\beta_1\right\}}\right]\right)=E^*\left[\gamma(n)\sigma^{-1}\left(\tilde{\beta}^*-\tilde{\beta}+\tilde{\beta}-\beta\right)1_{\left\{\tilde{\beta}^*\geq\beta_1\right\}}\right]\\ =\;& E^*\left[\gamma(n)\sigma^{-1}\left(\tilde{\beta}^*-\tilde{\beta}+\tilde{\beta}-\beta\right)1_{\left\{\gamma(n)\sigma^{-1}\left(\tilde{\beta}^*-\tilde{\beta}\right)\geq\gamma(n)\sigma^{-1}(\beta_1-\beta)+\gamma(n)\sigma^{-1}(\beta-\tilde{\beta})\right\}}\right]\\ &\xrightarrow{d} E_W\left[W1_{\{W\geq -b\}}|Z\right],\end{aligned}$$

where $W\sim N(Z,1)$.

$$\begin{aligned}&E_W\left[W1_{\{W\geq -b\}}|Z\right]=E_W[W]-E_W\left[W1_{\{W<-b\}}|Z\right]\\ =\;& Z-\int_{-\infty}^{-b}w\varphi(w-Z)\,dw=Z-\int_{-\infty}^{-b-Z}(s+Z)\varphi(s)\,ds\\ =\;& Z-Z\Phi(-b-Z)-\int_{-\infty}^{-b-Z}s\varphi(s)\,ds=Z-Z\Phi(-b-Z)+\varphi(-b-Z).\end{aligned}$$

Similarly, we get $\gamma(n)\sigma^{-1}(\beta_1-\beta)E^*\left[1_{\left\{\tilde{\beta}^*<\beta_1\right\}}\right]\xrightarrow{p}-b\Phi(-b-Z)$, by Slutsky's theorem. Putting together (i) and (ii) gives the result in (2).

(3) From (1), we can derive,

$$\begin{aligned}E\hat{\beta} &= E\left\{\tilde{\beta}\left[1-\Phi(-b-Z)\right]+\beta_1\Phi(-b-Z)\right\}+o(1)\\ &= \left(1-E\Phi(-b-Z)\right)E\tilde{\beta}+\beta_1 E\Phi(-b-Z)+o(1)\\ &= \left(1-\Phi*\varphi(-b)\right)\beta+\beta_1\Phi*\varphi(-b)+o(1)\\ &= \beta+(\beta_1-\beta)\Phi*\varphi(-b)+o(1)\\ &= \beta+O\left(\frac{1}{\gamma(n)}\right)\;\to\beta\end{aligned}$$

which is (3a). Note that

$$\begin{aligned} Var\left(\hat{\beta}\right) &= Var\left\{\tilde{\beta}\left[1-\Phi\left(-b-Z\right)\right]+\beta_1\Phi\left(-b-Z\right)+o\left(1\right)\right\} \\ &= E\left\{\tilde{\beta}\left[1-\Phi\left(-b-Z\right)\right]+\beta_1\Phi\left(-b-Z\right)\right\}^2 \\ &\quad -\left[E\left\{\tilde{\beta}\left[1-\Phi\left(-b-Z\right)\right]+\beta_1\Phi\left(-b-Z\right)\right\}\right]^2+o\left(1\right). \end{aligned}$$

The first term can be reduced as,

$$\begin{aligned} & E\left\{\tilde{\beta}\left[1-\Phi\left(-b-Z\right)\right]+\beta_1\Phi\left(-b-Z\right)\right\}^2 \\ = \; & E\left\{\tilde{\beta}\left[1-\Phi\left(-b-Z\right)\right]\right\}^2+2E\left\{\tilde{\beta}\left[1-\Phi\left(-b-Z\right)\right]\beta_1\Phi\left(-b-Z\right)\right\}+\beta_1^2E\left[\Phi\left(-b-Z\right)^2\right] \\ = \; & \beta^2\left\{1-2\Phi*\varphi\left(-b\right)+\Phi^2*\varphi\left(-b\right)\right\}+2\beta_1\beta\left\{\Phi*\varphi\left(-b\right)-\Phi^2*\varphi\left(-b\right)\right\}+\beta_1^2\Phi^2*\varphi\left(-b\right) \\ = \; & \beta^2+2\beta\left\{\Phi*\varphi\left(-b\right)-\Phi^2*\varphi\left(-b\right)\right\}\left(\beta_1-\beta\right)+\left\{\Phi^2*\varphi\left(-b\right)\right\}\left(\beta_1+\beta\right)\left(\beta_1-\beta\right) \\ = \; & \beta^2+O\left(\frac{1}{\gamma\left(n\right)}\right)=\beta^2+o\left(1\right). \end{aligned}$$

Hence, we have (3b).

(4) From (2), we get

$$\begin{aligned} \lim_{n\to\infty}E\left[\gamma\left(n\right)\sigma^{-1}\left(\hat{\beta}-\beta\right)\right] &= E\left\{Z-Z_b\Phi\left(-b-Z\right)+\varphi\left(-b-Z\right)\right\} \\ &= EZ-E\left[Z\Phi\left(-b-Z\right)\right]-bE\left[\Phi\left(-b-Z\right)\right]+E\varphi\left(-b-Z\right) \\ &= 0-\left[-\varphi*\varphi\left(-b\right)\right]-b\Phi*\varphi\left(-b\right)+\varphi*\varphi\left(-b\right) \\ &= 2\varphi*\varphi\left(-b\right)-b\Phi*\varphi\left(-b\right) \end{aligned}$$

where we used Lemma 2. Thus we have 4(a).

For 4(b), we need prove that

$$\begin{aligned}
&\lim_{n\to\infty} E\left[\gamma(n)\sigma^{-1}\left(\hat{\beta}-\beta\right)\right]^2 = E\left[Z - Z_b\Phi(-b-Z) + \varphi(-b-Z)\right]^2 \\
= \;& EZ^2 + E\left[Z_b^2\Phi^2(-b-Z)\right] + E\left[\varphi^2(-b-Z)\right] \\
& -2E\left[ZZ_b\Phi(-b-Z)\right] + 2E\left[Z\varphi(-b-Z)\right] - 2E\left[Z_b\Phi(-b-Z)\varphi(-b-Z)\right] \\
= \;& 1 + \Phi^2 * \varphi''(-b) + \Phi^2 * \varphi(-b) - 2b\Phi^2 * \varphi'(-b) + b^2\Phi^2 * \varphi(-b) \\
& +\varphi^2 * \varphi(-b) - 2\Phi * \varphi''(-b) - 2\Phi * \varphi(-b) + 2b\Phi * \varphi'(-b) \\
& -2\varphi * \varphi'(-b) + 2(\Phi\cdot\varphi) * \varphi'(-b) - 2b(\Phi\cdot\varphi) * \varphi(-b)
\end{aligned}$$

where we used Lemma 2. □

**Proof of Theorem 4.** The proofs for part (1) and (2), (4), (5) and (6) follows that in Theorem 2. We prove part (3). Note that $\bar{m}(x) = \tilde{m}_{LC}(x)\cdot 1_{\left\{\tilde{\beta}(x)\le\beta_1(x)\right\}} + \tilde{m}_{LL}(x)\cdot 1_{\left\{\tilde{\beta}(x)>\beta_1(x)\right\}}$.

$$\begin{aligned}
& \gamma_2(n,h)\sigma_m^{-1}(x)\left[\bar{m}(x) - m(x) - B_m(x)\right] \\
= \;& \gamma_2(n,h)\sigma_m^{-1}(x)\left[\tilde{m}_{LC}(x) - m(x) - B_m(x)\right]\cdot 1_{\left\{\tilde{\beta}(x)\le\beta_1(x)\right\}} \\
& +\gamma_2(n,h)\sigma_m^{-1}(x)\left[\tilde{m}_{LL}(x) - m(x) - B_m(x)\right]\cdot 1_{\left\{\tilde{\beta}(x)>\beta_1(x)\right\}} \\
\equiv \;& l_1\cdot 1_{\left\{\tilde{\beta}(x)\le\beta_1(x)\right\}} + l_2\cdot 1_{\left\{\tilde{\beta}(x)>\beta_1(x)\right\}},
\end{aligned}$$

where,

$$l_1 = \gamma_2(n,h)\sigma_m^{-1}(x)\left[\tilde{m}_{LC}(x) - m(x) - B_m(x)\right] = O_p(1),$$

and

$$l_2 = \gamma_2(n,h)\sigma_m^{-1}(x)\left[\tilde{m}_{LL}(x) - m(x) - B_m(x)\right] \xrightarrow{d} Z.$$

Note that

$$1_{\left\{\tilde{\beta}(x)\le\beta_1(x)\right\}} = 1_{\left\{\gamma_1(n,h)\sigma_\beta^{-1}(x)\left[\tilde{\beta}(x)-\beta(x)\right]\le\gamma_1(n,h)\sigma_\beta^{-1}(x)\left[\beta_1(x)-\beta(x)\right]\right\}} \to 1_{\{Z\le-\infty\}} = o_p(1)$$

Similarly, we can show that $1_{\left\{\tilde{\beta}(x)>\beta_1(x)\right\}} = 1 - 1_{\left\{\tilde{\beta}(x)\leq\beta_1(x)\right\}} \xrightarrow{p} 1$. Combining these results leads to $\gamma_2(n,h)\,\sigma_m^{-1}(x)\,[\bar{m}(x) - m(x) - B_m(x)] \xrightarrow{d} Z$. □

**Proof of Theorem 5.** The proofs for parts (1-2) parallel that in Theorem 3. We prove part (3). Note that

$$\begin{aligned}\hat{m}(x) &= \bar{y}(x) - \hat{\beta}(x)\,[\bar{x}(x) - x] \\ \tilde{m}(x) &= \bar{y}(x) - \tilde{\beta}(x)\,[\bar{x}(x) - x].\end{aligned}$$

Therefore, we have

$$\begin{aligned}\hat{m}(x) - \tilde{m}(x) &= \left[\tilde{\beta}(x) - \hat{\beta}(x)\right] \times [\bar{x}(x) - x] \\ &= \left\{\left[\tilde{\beta}(x) - \beta(x)\right] + \left[\beta(x) - \hat{\beta}(x)\right]\right\} \times [\bar{x}(x) - x] \\ &= O\left(\frac{1}{\gamma(n,h)\,h}\right) \times O\left(\frac{1}{\sqrt{nh}}\right) = o\left(\frac{1}{\gamma(n,h)}\right), \text{ as } \sqrt{nh^3} \to \infty.\end{aligned}$$

Therefore, we have the equivalence of $\hat{m}(x)$ and $\tilde{m}(x)$ asymptotically. □

**Proof of Theorem 6.** We only need prove part (3) of the theorem. Since

$$\begin{aligned}\bar{m}_{sp}(x) &= \bar{\alpha} + \bar{\xi}(x) + \bar{\eta}(x)\,\bar{x}(x) + \bar{\beta}(x)\,x, \\ \tilde{m}_{sp}(x) &= \tilde{\alpha} + \tilde{\xi}(x) + \tilde{\eta}(x)\,\bar{x}(x) + \tilde{\beta}(x)\,x,\end{aligned}$$

we know that

$$\begin{aligned}\bar{m}_{sp}(x) - \tilde{m}_{sp}(x) &= [\bar{\alpha} - \tilde{\alpha}] + \left[\tilde{\xi}(x) - \bar{\xi}(x)\right] + [\tilde{\eta}(x) - \bar{\eta}(x)]\,\bar{x}(x) + \left[\tilde{\beta}(x) - \bar{\beta}(x)\right] \\ &= 1_{\left\{\tilde{\beta}(x)<\beta_1(x)\right\}}\,[\tilde{m}_{lc}(x) - \tilde{m}_{sp}(x)] \\ &= 1_{\left\{\tilde{\beta}(x)<\beta_1(x)\right\}}\,[\tilde{m}_{lc}(x) - m(x) + m(x) - \tilde{m}_{sp}(x)] \\ &= 1_{\left\{\tilde{\beta}(x)<\beta_1(x)\right\}} O_p\left(\frac{1}{\gamma_2(n,h)}\right) = o_p(1) \times O_p\left(\frac{1}{\gamma_2(n,h)}\right) = o_p\left(\frac{1}{\gamma_2(n,h)}\right).\end{aligned}$$

that is, $\bar{m}_{sp}(x)$ and $\tilde{m}_{sp}(x)$ share the same asymptotic distribution. It is implied from Theorem 3 of Martins-Filho *et al* (2007) that $\gamma_2(n,h)\,\bar{\sigma}_m^{-1}(x)\,[\tilde{m}_{sp}(x) - m(x) - B_m(x)] \xrightarrow{d} Z \sim N(0,1)$. Combining the results completes the proof. □

**Proof of Theorem 7.** Part (1) follows in steps similar to part (2) of Theorem 5. We prove part (2). Note that

$$\begin{aligned}\bar{m}_{sp}(x) &= \bar{\alpha}+\bar{\xi}(x)+\bar{\eta}(x)\bar{x}(x)+\bar{\beta}(x)x,\\ \hat{m}_{sp}(x) &= \hat{\alpha}+\hat{\xi}(x)+\hat{\eta}(x)\bar{x}(x)+\hat{\beta}(x)x,\end{aligned}$$

Therefore, we have

$$\begin{aligned}\hat{m}_{sp}(x)-\bar{m}_{sp}(x) &= E^*\bar{m}^*_{sp}(x)-\bar{m}_{sp}(x)\\ &= E^*\left[\tilde{m}^*_{sp}(x)-\tilde{m}_{sp}(x)\right]\\ &\quad +E^*\left\{1_{\left\{\tilde{\beta}(x)<\beta_1(x)\right\}}\left[\tilde{m}^*_{lc}(x)-\tilde{m}_{lc}(x)+\tilde{m}^*_{sp}(x)-\tilde{m}_{sp}(x)\right]\right\}\\ &= o_p\left(\frac{1}{\gamma_2(n,h)}\right)+o_p(1)\times o_p\left(\frac{1}{\gamma_2(n,h)}\right)=o_p\left(\frac{1}{\gamma_2(n,h)}\right)\end{aligned}$$

Therefore, we have the equivalence of $\hat{m}_{sp}(x)$ and $\bar{m}_{sp}(x)$ asymptotically, which completes the proof. □

## Lemmas

We collect useful lemmas that are used in the proof of the main theorems. We use $Z$ to denote a standard normal random variable with CDF $\Phi(\cdot)$ and PDF $\varphi(\cdot)$, $b$ to denote some constant, and $1_{\{\cdot\}}$ an indicator function. Define $Z_b = Z + b$.

**Lemma 1.** (**a**) $E1_{\{Z_b>0\}} = \Phi(b)$. (**b**) $E\left[Z1_{\{Z_b>0\}}\right] = \varphi(b)$. (**c**) $E\left[Z^2 1_{\{Z_b>0\}}\right] = -b\varphi(b)+\Phi(b)$. (**d**) $E\left[Z_b 1_{\{Z_b>0\}}\right] = \varphi(b)+b\Phi(b)$. (**e**) $E\left[Z_b^2 1_{\{Z_b>0\}}\right] = \Phi(b)+b\varphi(b)+b^2\Phi(b)$.

**Proof of Lemma 1.** (**a**) $E1_{\{Z_b>0\}} = E1_{\{Z>-b\}} = \int_{-b}^{\infty} d\Phi(z) = 1-\Phi(-b) = \Phi(b)$. (**b**) $EZ1_{\{Z_b>0\}} = \int_{-b}^{\infty} z\varphi(z)\,dz = -\int_{-b}^{\infty}\varphi'(z)\,dz = -\varphi(z)|_{-b}^{\infty} = \varphi(b)$. (**c**) $EZ^2 1_{\{Z_b>0\}} = \int_{-b}^{\infty} z^2\varphi(z)\,dz = -\int_{-b}^{\infty} z\varphi'(z)\,dz = -z\varphi(z)|_{-b}^{\infty}+\int_{-b}^{\infty}\varphi(z)\,dz = -b\varphi(b)+\Phi(b)$. (**d**) $E\left[Z_b 1_{\{Z_b>0\}}\right] = EZ1_{\{Z_b>0\}}+bE1_{\{Z_b>0\}} = \varphi(b)+b\Phi(b)$. (**e**) $E\left[Z_b^2 1_{\{Z_b>0\}}\right] =$

$E\left[(Z+b)^2 1_{\{Z_b>0\}}\right] = EZ^2 1_{\{Z_b>0\}} + b^2 E1_{\{Z_b>0\}} + 2bE\left[Z1_{\{Z_b>0\}}\right] = \Phi(b) - b\varphi(b) + b^2\Phi(b) + 2b\varphi(b) = \Phi(b) + b\varphi(b) + b^2\Phi(b)$.

**Lemma 2.** (**a**) $E\varphi(-Z_b) = \varphi * \varphi(-b)$. (**b**) $E\varphi^2(-Z_b) = \varphi^2 * \varphi(-b)$. (**c**) $E[Z\varphi(-Z_b)] = -\varphi * \varphi'(-b)$. (**d**) $E[Z\Phi(-Z_b)] = -\varphi * \varphi(b)$. (**e**) $E\left[Z^2\Phi(-Z_b)\right] = \Phi * \varphi''(-b) + \Phi * \varphi(-b)$. (**f**) $E\left[Z^2\Phi^2(-Z_b)\right] = \Phi^2 * \varphi''(-b) + \Phi^2 * \varphi(-b)$. (**g**) $E[Z\Phi(-Z_b)\varphi(-Z_b)] = -(\Phi \cdot \varphi) * \varphi'(-b)$.

**Proof of Lemma 2.**

**(a)** $E\varphi(-Z_b) = E\varphi(-b-Z) = \int_{-\infty}^{\infty} \varphi(-b-z)\varphi(z)\,dz = \varphi * \varphi(-b)$.

**(b)** $E\varphi^2(-Z_b) = E\varphi^2(-b-Z) = \int_{-\infty}^{\infty} \varphi^2(-b-z)\varphi(z)\,dz = \varphi^2 * \varphi(-b)$.

**(c)** $E[Z\varphi(-Z_b)] = E[Z\varphi(-b-Z)] = \int_{-\infty}^{\infty} z\varphi(-b-z)\varphi(z)\,dz = -\int_{-\infty}^{\infty} \varphi(-b-z)\varphi'(z)\,dz = -\varphi * \varphi'(-b)$.

**(d)** $E[Z\Phi(-Z_b)] = E[Z\Phi(-b-Z)] = \int_{-\infty}^{\infty} z\Phi(-b-z)\varphi(z)\,dz = -\int_{-\infty}^{\infty} \Phi(-b-z)\varphi'(z)\,dz$
$= -\left\{\Phi(-b-z)\varphi(z)|_{z=-\infty}^{\infty} - \int_{-\infty}^{\infty} -\varphi(-b-z)\varphi(z)\,dz\right\} = -\varphi * \varphi(-b)$.

**(e)** $E\left[Z^2\Phi(-Z_b)\right] = E\left[Z^2\Phi(-b-Z)\right] = \int_{-\infty}^{\infty} z^2\Phi(-b-z)\varphi(z)\,dz$
$= \int_{-\infty}^{\infty} \Phi(-b-z)[\varphi(z) + \varphi''(z)]\,dz = \Phi * \varphi''(-b) + \Phi * \varphi(-b)$.

**(f)** $E\left[Z^2\Phi^2(-Z_b)\right] = E\left[Z^2\Phi^2(-b-Z)\right] = \int_{-\infty}^{\infty} z^2\Phi^2(-b-z)\varphi(z)\,dz$
$= \int_{-\infty}^{\infty} \Phi^2(-b-z)[\varphi(z) + \varphi''(z)]\,dz = \Phi^2 * \varphi''(-b) + \Phi^2 * \varphi(-b)$.

**(g)** $E[Z\Phi(-Z_b)\varphi(-Z_b)] = E[Z\Phi(-b-Z)\varphi(-b-Z)] = \int_{-\infty}^{\infty} z\Phi(-b-z)\varphi(-b-Z)\varphi(z)\,dz$
$= -\int_{-\infty}^{\infty} \Phi(-b-z)\varphi(-b-Z)\varphi'(z)\,dz = -(\Phi \cdot \varphi) * \varphi'(-b)$.

## References


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

Brunk, H. D. (1955), "Maximum likelihood estimates of monotone parameters", *Annals of Mathematical Statistics* 26, 607-616.

Bühlmann, P., and B. Yu (2002), "Analyzing Bagging," *The Annals of Statistics* 30, 927–961.

Campbell, J.Y. and S. Thompson (2008), "Predicting the equity premium out of sample: Can anything beat the historical average?," *Review of Financial Studies* 21(4), 1511-1531.

Chen, Q. and Y. Hong (2009), "Predictability of equity returns over different time horizons: A nonparametric approach," Working paper, Cornell University.

Chernozhukov, V., I. Fernandez-Val and A. Galichon (2009), "Improving point and interval estimators of monotone functions by rearrangement," *Biometrika* 96(3): 559-575.

Dette, H., N. Neumeyer and K.F. Pilz (2006), "A simple nonparametric estimator of a strictly monotone regression function", *Bernoulli* 12(3), 469-490.

Fama, E. F., and K. R. French, (2002), "The equity premium," *The Journal of Finance* 57(2), 637-659.

Glad, I. (1998), "Parametrically guided non-parametric regression," *Scandinavian Journal of Statistics* 25, 649-668.

Gordon, I., and P. Hall (2009), "Estimating a parameter when it is known that the parameter exceeds a given value," *Australian and New Zealand Journal of Statistics* 51(4), 449-460.

Goyal, A., and I. Welch (2008), "A comprehensive look at the empirical performance of equity premium prediction," *Review of Financial Studies* 21(4), 1455-1508.

Granger, C.W.J. (1999), *Empirical Modeling in Economics: Specification and Evaluation,* Cambridge University Press.

Hall, P. and H. Huang (2001), "Nonparametric kernel regression subject to monotonicity constraints," *The Annals of Statistics* 29(3), 624-647.

Henderson, D. J. and C. F., Parmeter (2009), "Imposing economic constraints in nonparametric regression: survey, implementation and extension," *Advances in Econometrics* 25, 433–469.

Hillebrand, E., T.-H., Lee, and M. Medeiros (2009), "Bagging constrained forecasts with application to forecasting equity premium," *JSM Proceedings for Business and Economic Statistics.*

Judge, G.G. and T.A. Yancey (1986), *Improved Methods of Inference in Econometrics*, Vol. 34, in Studies in Mathematical and Managerial Economics, edited by Henry Theil and Herbert Glejser, North-Holland.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005), "Consistent testing for stochastic dominance under general sampling schemes," *Review of Economic Studies* 72,735-765.

Mammen, E. (1992), *When Does Bootstrap Work? Asymptotic Results and Simulations,* Lecture Note in Statistics 77. Springer-Verlag, New York.

Martins-Filho, C., S. Mishra and A. Ullah (2007), "A class of improved parametrically guided nonparametric regression estimators," *Econometric Reviews* 27, 542-573.

McFadden, D. (1989), "Testing for stochastic dominance," in Part II of T. Fomby and T.K. Seo (eds.) *Studies in the Economics of Uncertainty* (in honor of J. Hadar), Springer-Verlag.

Racine, J. S., C. F. Parmeter and P. Du (2009), "Constrained nonparametric kernel regression: estimation and inference," Department of Economics, McMaster University.

Figure 2.1: Asymptotic variance, Asymptotic squared bias, and Asymptotic mean squared error of constrained estimator (CE) and bagging constrained estimator (BCE)

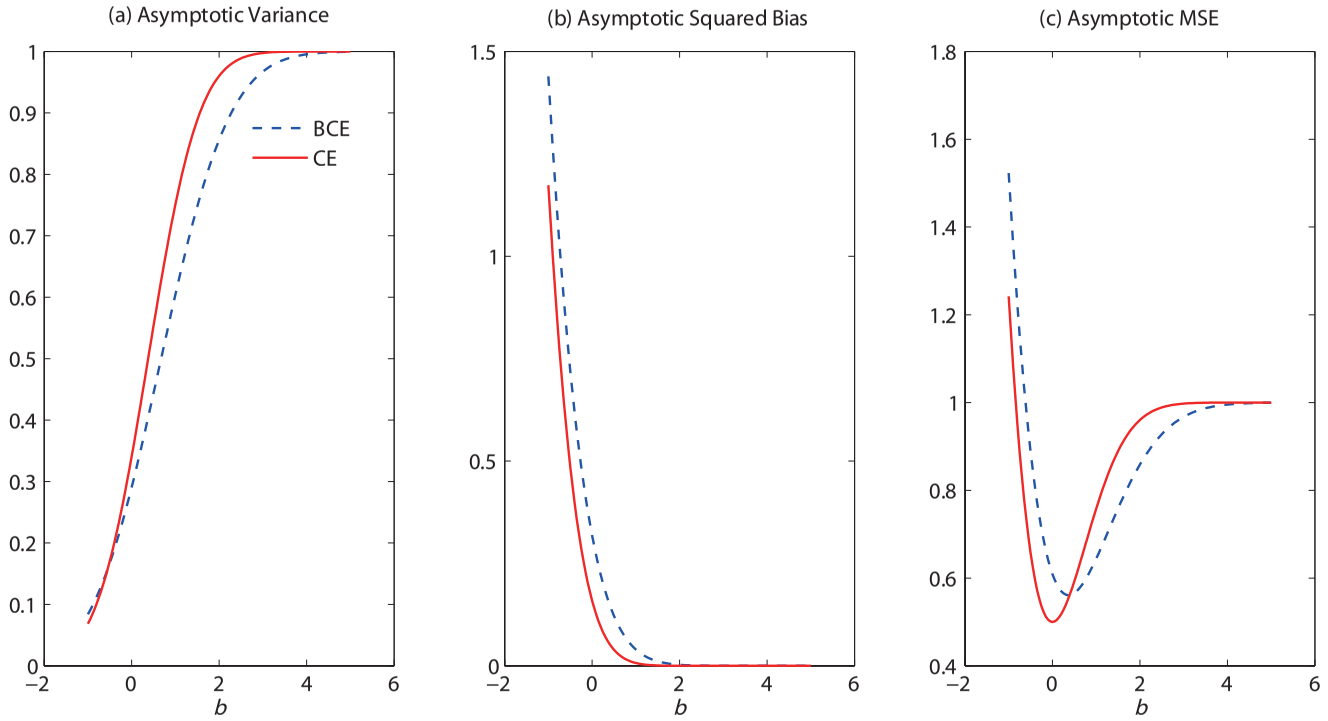


Table 2.1: Simulation Results: $R^2$ and SOSD

| | $\sigma_x^2 = 2$ | | | $\sigma_x^2 = 3$ | | | $\sigma_x^2 = 4$ | | | $\sigma_x^2 = 5$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $100R^2$ | avg[a] | max | $100R^2$ | avg | max | $100R^2$ | avg | max | $100R^2$ | avg | max |
| L | 55.567 | 34.529 | 52.117 | 61.544 | 39.522 | 57.335 | 63.398 | 41.624 | 59.566 | 64.261 | 42.803 | 60.864 |
| L-P | 55.686 | 34.641 | 52.247 | 61.544 | 39.522 | 57.335 | 63.398 | 41.624 | 59.566 | 64.261 | 42.803 | 60.864 |
| L-P-B | 57.077 | 35.683 | 53.568 | 61.512 | 39.491 | 57.305 | 63.280 | 41.524 | 59.444 | 64.112 | 42.690 | 60.721 |
| NP | 43.909 | 23.379 | 41.225 | 80.431 | 53.289 | 74.977 | 90.854 | 63.150 | 85.312 | 94.643 | 67.314 | 89.617 |
| NP-P | 53.569 | 31.060 | 50.273 | 81.850 | 54.554 | 76.295 | 91.162 | 63.434 | 85.603 | 94.694 | 67.362 | 89.665 |
| NP-P-B | 54.985 | 32.124 | 51.610 | 82.304 | 54.952 | 76.727 | 91.647 | 63.872 | 86.062 | 94.750 | 67.424 | 89.728 |
| NP-HH | 48.800 | 27.959 | 45.806 | 77.498 | 50.838 | 72.241 | 91.188 | 63.453 | 85.623 | 94.656 | 67.312 | 89.615 |
| SP | 43.662 | 23.226 | 40.990 | 80.623 | 53.462 | 75.152 | 90.847 | 63.145 | 85.306 | 94.638 | 67.313 | 89.616 |
| SP-P | 53.539 | 31.053 | 50.238 | 81.918 | 54.619 | 76.358 | 91.179 | 63.451 | 85.621 | 94.705 | 67.374 | 89.678 |
| SP-P-B | 55.713 | 32.662 | 52.289 | 82.887 | 55.442 | 77.263 | 91.653 | 63.879 | 86.069 | 94.782 | 67.450 | 89.755 |

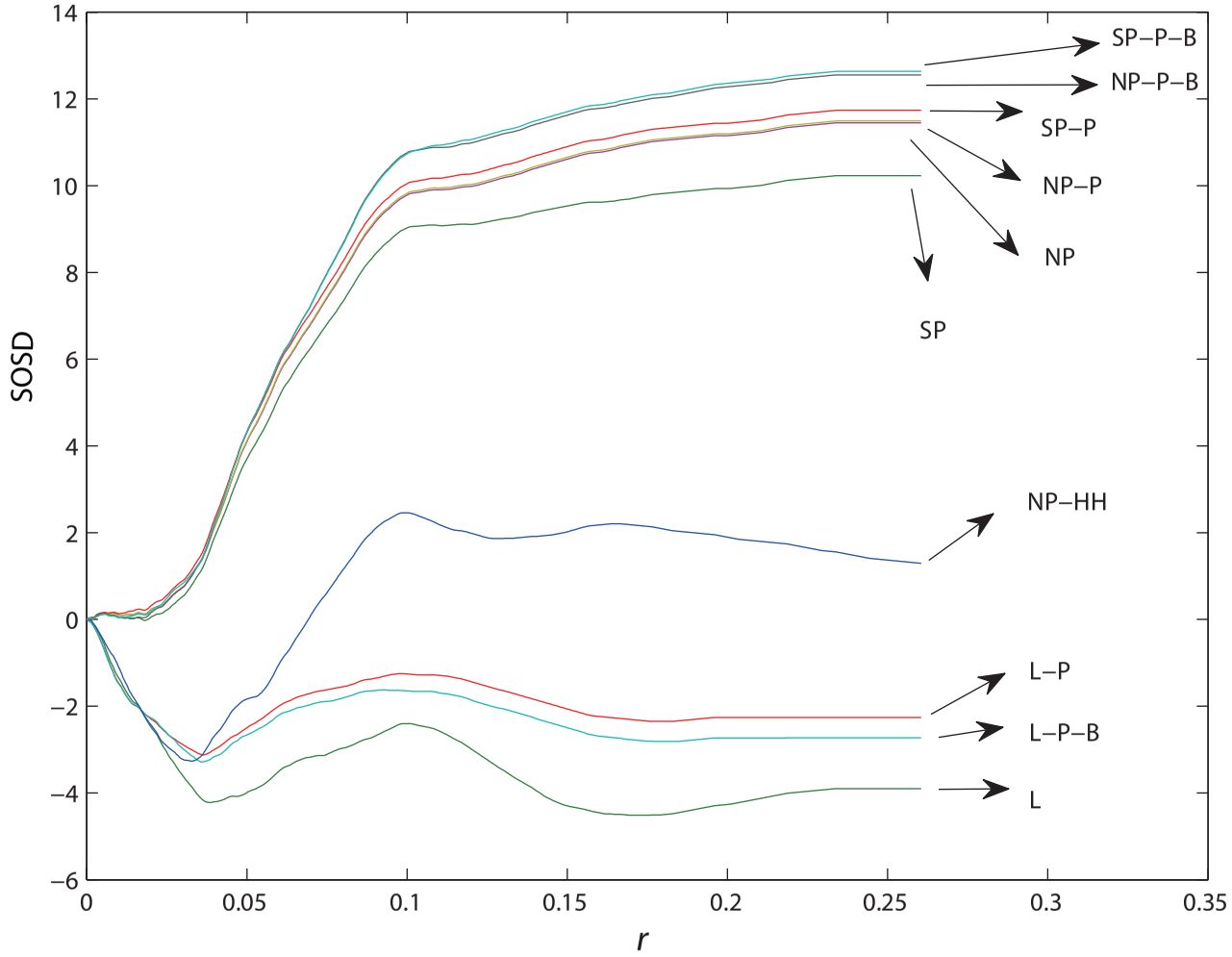
[a]avg is short for avg(SOSD), while max is max(SOSD)

Table 2.2: Equity Premium Forecasting Results: $R^2$ and SOSD

| | *se/p* | | | *t-bill* | | | *lty* | | | *ds* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $100R^2$ | avg[a] | max | $100R^2$ | avg | max | $100R^2$ | avg | max | $100R^2$ | avg | max |
| L | 2.559 | 0.773 | 1.987 | -5.478 | -3.841 | 0.000 | -4.186 | -3.521 | 0.000 | -0.240 | -0.214 | 0.662 |
| L-P | 2.567 | 0.838 | 2.002 | -2.927 | -2.532 | 0.000 | -2.432 | -2.008 | 0.000 | -0.046 | -0.445 | 0.336 |
| L-P-B | 2.637 | 0.887 | 2.079 | -2.946 | -2.531 | 0.002 | -2.918 | -2.349 | 0.002 | -0.157 | -0.523 | 0.323 |
| NP | 11.450 | 8.506 | 10.497 | 5.991 | 5.851 | 8.139 | 12.283 | 8.287 | 11.449 | 3.485 | 2.477 | 4.274 |
| NP-P | 11.472 | 8.524 | 10.514 | 5.932 | 5.762 | 8.043 | 12.312 | 8.321 | 11.492 | 3.529 | 2.509 | 4.309 |
| NP-P-B | 11.310 | 8.308 | 10.287 | 6.732 | 5.991 | 8.044 | 13.479 | 9.048 | 12.555 | 5.698 | 3.452 | 5.200 |
| NP-HH | 0.296 | 0.156 | 1.059 | 2.110 | 2.525 | 4.261 | 1.395 | 0.855 | 2.458 | -6.649 | -5.396 | -0.002 |
| SP | 16.684 | 11.384 | 13.636 | 6.497 | 6.229 | 8.677 | 10.994 | 7.452 | 10.231 | 5.124 | 3.862 | 5.799 |
| SP-P | 16.735 | 11.426 | 13.680 | 6.636 | 6.264 | 8.699 | 12.584 | 8.533 | 11.738 | 4.111 | 3.058 | 4.893 |
| SP-P-B | 17.009 | 11.555 | 13.885 | 6.807 | 6.047 | 8.272 | 13.568 | 9.098 | 12.636 | 5.985 | 3.888 | 5.482 |

[a]avg is short for avg(SOSD), while max is max(SOSD)

Figure 2.2: SOSD for *lty*



# Chapter 3

# Improving Historical Mean Forecast of Equity Premium by Constrained Nonparametric Approach

## 3.1 Introduction

Historical Mean (HM) forecast has been used as a bench mark in the forecast of Equity Premium in the literature to improve upon. The recent work of Goyal and Welch (2008) extensively studied the predictability of stock return using various forecasting specifications. The negative conclusion from their paper suggest that HM forecasts is unbeatable. However, this finding was contrasted by that of Campbell and Thompson (2008) that, once sensible sign restrictions are imposed on the forecasts, the out-of-sample forecasting performance of many predictors can be improved and sometimes beat

the historical average return forecast. Hillebrand et al (2009) incorporate the bagging (*b*ootstrap *agg*regat*ing*) approach of Gordon and Hall (2009) to smooth sign restrictions in HM forecasting models and show that the bagging sign restriction approach has more predictive power than the simple sign restriction of Campbell and Thompson (2008).

However, possible misspecification of a HM model can undermine its forecasts compared to those produced via nonlinear models. In this chapter we extend this literature by considering nonlinear models, in particular, nonparametric (NP) and semiparametric (SP) kernel regressions with imposing the sign constraints on the forecast and with applying bagging to the constraints. Chen and Hong (2009) find that, in the prediction of asset returns, nonparametric kernel regression model has a better forecasting power than the historical mean, due to the higher signal-to-noise ratio resulted from nonparametric models. However, Chen and Hong (2009) do not consider the sign restriction as well as bagging in their nonlinear forecasting exercise. This chapter is to consider nonlinear models subject to positiveness constraint with and without bagging. The previous chapter is concerned with monotonicity constraint.

Nonparametric kernel estimation with constraints has long history that dates back to the work of Brunk (1955). Recent work on imposing monotonicity on nonparametric regression function includes Hall and Huang (2001), Dette et al (2006) and Chernozhukov et al (2007), among others. Hall and Huang (2001) propose a novel method of imposing the monotonicity constraint on a class of nonparametric kernel estimations. Their estimator is constructed by re-weighting the kernel for each response data point so that the impact of each observation on the estimated regression function can be controlled to satisfy a constraint. Their method is rooted in a conventional kernel framework and is extended by Racine et al (2009) and Henderson and Parmeter (2009)

to allow for a broader class of conventional constraints and to develop tests for these constraints.

Our contributions are as given below. First, we consider NP models to generalize the HM models considered in Goyal and Welch (2008), Campbell and Thompson (2008) and Hillebrand et al (2009). These NP regressions can capture possibly neglected nonlinearity in HM models and could improve the predictive ability of the predictors, as demonstrated in our Monte Carlo simulation and application to the equity premium prediction. Second, we consider a new method of imposing the positivity constraint on the NP and regressions. This is to make the prediction more accurate as we employ more information than Chen and Hong (2009). Our positivity constraint is local restriction while it is global positivity in Campbell and Thompson (2008). Third, we use bagging to smooth the positivity constraint in NP regressions as Hillebrand et al (2009) do in linear regressions. It has been shown in Bühlmann and Yu (2002) that bagging can reduce asymptotic mean squared error in linear regressions. We obtain the similar results that hold locally in NP regressions. Fourth, we conduct simulation study to demonstrate how the asymptotic results work in finite sample. We also conduct an empirical study in predicting equity premium using the same data from Campbell and Thompson (2008) to demonstrate the practical merit of the bagging sign constrained NP regression models. Fifth, in our simulation and empirical application, we find that, despite its simplicity to implement, our bagging constrained NP regression almost always and clearly outperforms the constrained NP regression of Hall and Huang (2001).

The rest of the chapter is organized as follows. Section 2 presents nonparametric methods under constraints to produce forecasts. Section 3 and 4 establish the finite and asymptotic properties of the proposed estimators. Section 5 conducts Monte Carlo simulations to compare our proposed bagging nonparametric forecasts with constraints

with other forecasts, including linear parametric forecasts with constraints and non-parametric forecasts with constraints, etc. We evaluate different forecasting schemes considered in this chapter via the prediction of equity premium in Section 6 and we conclude in Section 7 with remarks.

## 3.2 Estimation with Constraints

Let $\{X_i, y_i\}_{i=1}^N$ be an observed sample, where $y$ is the variable of interest to a forecaster and $X$ is a $p \times 1$ vector of predictors. We assume that the mean of $y$, $\mu_y$, exceeds some known lower bound, $\alpha_1$. That is,

$$\mu_y > \alpha_1. \tag{3.1}$$

This information is known as a prior to a forecastor. We're concerned with the problem of how to adopt this information in the forecasting practice. We will treat in sequence parametric mean model and nonparametric local constant kernel model.

### 3.2.1 Constrained Parametric Forecast

We consider first the case that $X$ only contains a constant, i.e., the mean model

$$y_i = \alpha + \varepsilon_i,$$

where $\varepsilon_i$ is a disturbance term such that $E(\varepsilon_i) = 0, i = 1, ..., n$. Least square estimator of $\alpha$,

$$\tilde{\alpha} = \frac{1}{N} \sum_{i=1}^{N} y_i, \tag{3.2}$$

is the unconstrained estimator of $\alpha$. Note that $\tilde{\alpha}$ is a random variable which is asymptotically normal with mean

To impose constraint (3.1), we may set the forecast to be $\alpha_1$, whenever $\tilde{\alpha}$ computed via (3.2) is less than this lower bound. That is, our constrained estimator is

$$\bar{\alpha} = \tilde{\alpha} 1_{[\tilde{\alpha} > \alpha_1]} + \alpha_1 1_{[\tilde{\alpha} \leq \alpha_1]}. \tag{3.3}$$

The above estimator involves an indicator which is not stable in the sense of Buhlmann and Yu (2002). Bagging (Breiman, 1996) is the device that works to smooth this unstable estimator. To define our bagging estimator for $\alpha$, we construct a bootstrap sample $\{y_i^*\}_{i=1}^N$ which is used to derive a bootstrap constrained estimator via (3.3) using the plug-in principle. The bagging predictor is an expectation of this bootstrapped estimator over the bootstrapped samples. To be precise, denote $\tilde{\alpha}^{*(j)}$ as the unconstrained estimator of $\alpha$ computed from the $j$-th bootstrapped sample $\left\{y_i^{*(j)}\right\}_{i=1}^N$, $j = 1, ..., J$. Then $\bar{\alpha}^{*(j)} = \tilde{\alpha}^{*(j)} 1_{\left[\tilde{\alpha}^{*(j)} > \alpha_1\right]} + \alpha_1 1_{\left[\tilde{\alpha}^{*(j)} \leq \alpha_1\right]}$. Our bagging estimator is

$$\hat{\alpha} = E\bar{\alpha}^* = \frac{1}{J} \sum_{j=1}^{j} \bar{\alpha}^{*(j)}. \tag{3.4}$$

### 3.2.2 Constrained Nonparametric Forecast

We consider next that $X$ only contains one regressor, i.e., the local mean model

$$y_i = \alpha(x) + \varepsilon_i,$$

with

$$\alpha(x) = E[y|x]$$

where $\varepsilon_i$ is a disturbance term such that $E(\varepsilon_i|x) = 0$ by construction, $i = 1, ..., n$. Local constant kernel estiamtor of $\alpha(x)$,

$$\tilde{\alpha}(x) = \frac{\sum_{i=1}^N k\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^N k\left(\frac{x_i - x}{h}\right)}, \tag{3.5}$$

where $h$ is the bandwidth and $k(\cdot)$ is a kernel function. $\tilde{\alpha}(x)$ is shown to be asymptotically normal with mean $\alpha(x)$, c.f. Pagan and Ullah (1994).

The constraint in (3.1) is now assumed to be

$$\mu_y(x) > \alpha_1(x). \tag{3.6}$$

To impose this constraint in the estimation of $\alpha(x)$, we define our constrained estimator as

$$\bar{\alpha}(x) = \tilde{\alpha}(x) 1_{[\tilde{\alpha}(x)>\alpha_1]} + \alpha_1(x) 1_{[\tilde{\alpha}(x)\leq\alpha_1(x)]}. \tag{3.7}$$

Following similar procedures as in parametric estimation, we define our bagging estimator as

$$\hat{\alpha}(x) = E\bar{\alpha}^*(x) = \frac{1}{J}\sum_{j=1}^{j} \bar{\alpha}^{*(j)}(x),$$

where $\bar{\alpha}^{*(j)}(x)$ is the constrained estimator obtained via plug-in principle for the $j$-th bootstrap sample $\left\{x_i^{*(j)}, y_i^{*(j)}\right\}_{i=1}^{N}$, $j = 1, ..., J$.

## 3.3 Sampling Properties of Parametric Estimators

Sampling properties of constrained parametric estimator and its bagging version are established in this section.

### 3.3.1 Constrained Parametric Estimator

We start with some assumptions.

**Assumption A**

**(A.1)** $\gamma(n)\sigma^{-1}(\tilde{\alpha} - \alpha) \xrightarrow{d} Z$, where $\lim_{n\to\infty}\gamma(n) = \infty$, $\sigma > 0$ and $Z$ is a random variable with CDF $F_Z(\cdot)$.

**(A.2)** $\alpha = \alpha_1 + \gamma^{-1}(n)\sigma b$ for some $b \in \mathbb{R}$.

Assumption A.1 requires $\tilde{\alpha}$ satisfy some limiting theorem with asymptotic standard deviation $\sigma$. This is a very week condition that is met by a large class of estimators.

We do not specify the convergence rate $\gamma(n)$ but simply let it explode as $n$ increases. This general setting accommodates both estimators with standard convergence rate $\sqrt{n}$ and estimators with nonstandard convergence rate, e.g., $n^{1/3}$ or $n^{3/2}$. A.2 can be stated alternatively as $\gamma(n)\sigma^{-1}(\alpha - \alpha_1) = b$. It dictates that the true parameter $\alpha$ is a Pitman type drift to the specified bound $\alpha_1$, with a drift parameter $b$. The local drift rate is the same as the convergence rate of $\tilde{\alpha}$. Extension to higher or lower rate than this convergence rate is to be made possible by letting $b = b_n$ go to either infinity or zero as $n$ increases.

**Theorem 1.** *(i) Under assumption A.1,*

(a) *when* $\alpha > \alpha_1$, $\gamma(n)\sigma^{-1}(\bar{\alpha} - \alpha) \xrightarrow{d} Z$ .

(b) *when* $\alpha = \alpha_1$, $\Pr\left(\gamma(n)\sigma^{-1}(\bar{\alpha} - \alpha) < z\right) \to F_Z(z) \cdot 1_{\{z \geq 0\}}$.

*(ii) If we further assume A.2, then* $\lim_{n\to\infty} \gamma(n)\sigma^{-1}[\bar{\alpha} - \alpha] = Z_b 1_{[Z_b > 0]} - b$, *where* $Z_b = Z + b$.

**Remark 1.** Theorem 1 stated the limiting distribution of $\bar{\alpha}$. Part *(i)* presents the usual asymptotic distribution when the constraint is strict and when the parameter is on the boundary. The result confirm the intuition that, as long as the constraint is strict, it will be met by the unconstrained estimator $\tilde{\alpha}$ when the sample size is large enough. This leads to the conclusion that $\bar{\alpha}$ would be asymptotically equivalent to $\tilde{\alpha}$. When $\alpha$ is on the boundary, the limiting CDF compresses all the mass of negative values at 0. Part *(ii)* establishes the local asymptotic distribution of $\bar{\alpha}$ that depends on the drift parameter $b$. It is easy to see that, if $b$ is allowed to grow as $n$, $Z_b 1_{[Z_b > 0]} - b$ will collapse to $Z$, and result in *(ii)* becomes that in *(i.a)*. Similarly, *(i.b)* reproduces the result of (2) when $b = 0$. □

It is informative that when $Z$ is a standard normal random variable, the asymptotic bias and vriance of $\bar{\alpha}$ can be established as functions of the normal CDF and PDF, together with the drift parameter $b$. This is stated in the Corollary 1.

**Assumption Z**

$Z$ is standard normal with CDF $\Phi(\cdot)$ and PDF $\varphi(\cdot)$

**Corollary 1.** *If, in addition to A.1-A.2, assumption Z holds, then*

(a) $\lim_{n\to\infty} \gamma(n)\sigma^{-1}E[\bar{\alpha}-\alpha] = \varphi(b) + b\Phi(b) - b.$

(b) $\lim_{n\to\infty} Var\left[\left(\gamma(n)\sigma^{-1}\right)^{1/2}\bar{\alpha}\right] = \Phi(b)+b\varphi(b)-\varphi^2(b)-2b\varphi(b)\Phi(b)+b^2\Phi(b)[1-\Phi(b)].$

### 3.3.2 Bagged Constrained Parametric Estimator

To study the asymptotic properties of bagging constrained estimators, we start with the following assumptions:

**Assumption A**

**(A.3)** $\gamma(n)\sigma^{-1}(\tilde{\alpha}^* - \tilde{\alpha}) \xrightarrow{d} Z.$

Assumption A.3 requires that bootstrap work for $\tilde{\alpha}$. Lower level assumptions that lead to this condition can be refereed to, e.g., Freedman (1981), Hall (1994) and Horowitz (2001) among others. We empahsize that we don't require bootstrap work for $\bar{\alpha}$. Since we have $\bar{\alpha}$ on the boudary of the parameter space if $\alpha > \alpha_1$ holds, bootstrap fails for such kind of estimator (Andrews, 2000). Actually, it is the fact that bootstrap's failure for $\bar{\alpha}$ leads to the following theorem.

**Theorem 2.** *Under Assumption A.1-A.3,*

$$\gamma(n)\sigma^{-1}(\hat{\alpha}-\alpha) \xrightarrow{d} Z - Z_b F_Z(-Z_b) + E_W\left[W 1_{[W\leq -Z_b]}|Z\right], \text{where the CDF of } W \text{ is } F_W(\cdot) = F_Z(\cdot).$$

The limiting random variable contains truncated expectation that involves $W$, which has the same distribution as $Z$. When $Z$ is normal, we derive the following corollary. We adopt the notation $f * g$ to denote the convolution of two functions $f$ and $g$, which is defined as $f * g(s) = \int f(t) \times g(s-t)\, ds$.

**Corollary 2.** *If, in addition to A.1-A.3, assumption Z holds, then*

1. $\gamma(n)\sigma^{-1}(\hat{\alpha}-\alpha) \xrightarrow{d} Z - Z_b\Phi(-Z_b) + \varphi(-Z_b)$.

2. (a) $\lim_{n\to\infty}\gamma(n)\sigma^{-1}E[\hat{\alpha}-\alpha] = 2\varphi*\varphi(-b) - b\Phi*\varphi(-b)$.

   (b) $\lim_{n\to\infty} Var\left[\left(\gamma(n)\sigma^{-1}\right)^{1/2}\hat{\alpha}\right] = 1 + \Phi^2*\varphi''(-b) + \Phi^2*\varphi(-b) - 2b\Phi^2*\varphi'(-b) + b^2\Phi^2*\varphi(-b) + \varphi^2*\varphi(-b) - 2\Phi*\varphi''(-b) - 2\Phi*\varphi(-b) + 2b\Phi*\varphi'(-b) - 2\varphi*\varphi'(-b) + 2(\Phi\cdot\varphi)*\varphi'(-b) - 2b(\Phi\cdot\varphi)*\varphi(-b) - [2\varphi*\varphi(-b) - b\Phi*\varphi(-b)]^2$.

**Remark 2.** Theorem 2 stated the limiting distribution of $\hat{\alpha}$. Under the assumption that $Z$ is standard normal, corollary 2 shows the simple expression for the limiting random variable, which is a transformation of standard normal random variable via standard normal CDF and PDF. The dependence of the limiting distribution on the drift parameter is explicit through $Z_b$. In order to compare the performance of bagging constrained estimator $\hat{\alpha}$ and constrained estimator $\bar{\alpha}$ without bagging, we plot asymptotic variance, squared bias and MSE against the drift parameter $b$ in figure 1. We notice from the figure that there is a trade off using bagging, which reduce asymptotic variance while

incurring some additional bias. Overall, it is clear that for a large range of values of $b$ ($\geq 0.391$), bagging estimator enjoys a reduction in asymptotic MSE.

Figure 1 About Here

The following theorem shows that bagging estimator is a model averaging type estimator.

**Theorem 3.** *Under Assumption A.1-A.3,*

$$\hat{\alpha} = \tilde{\alpha}\Phi\left(Z_b\right) + \alpha_1\Phi\left(-Z_b\right) + O_p\left(\tfrac{1}{\gamma(n)}\right).$$

**Remark 3.** Theorem 3 stated that the bagging estimator is an average estimator that assigns a weight $\Phi\left(Z_b\right)$ to the unconstrained estimator $\tilde{\alpha}$ and a weight $\Phi\left(-Z_b\right)$ to the lower bound $\alpha_1$, up to order $O_p\left(\tfrac{1}{\gamma(n)}\right)$.

## 3.4 Sampling Properties of Nonparametric Estimators

We consider sampling properties of nonparametric estimators under constraint and its bagging version.

### 3.4.1 Constrained Nonparametric Estimator

**Assumption B**

**(B.1)** $(i)$ $\lim_{n\to\infty}\gamma\left(n,h\right)=\infty; (ii) h\to 0$ as $n\to\infty$.

**(B.2)** $\gamma\left(n,h\right)\sigma_\alpha^{-1}\left(x\right)\left(\tilde{\alpha}\left(x\right)-\alpha\left(x\right)-B_m\left(x\right)\right)\xrightarrow{d} Z$, where , $\sigma_\alpha\left(x\right)>0$, $Z$ is a standard normal random variable and $B_m\left(x\right)=\frac{1}{2}h^2m^{(2)}\left(x\right)\int v^2k\left(v\right)dv+o_p\left(h^2\right)$.

**(B.3)** $\alpha\left(x\right)=\alpha_1\left(x\right)+\gamma^{-1}\left(n,h\right)\sigma_\alpha\left(x\right)b\left(x\right)$ for some real function $b\left(\cdot\right)$.

**Theorem 4.** *(i) Under B.1-B.2, we have the following for the constrained estimator* $\bar{\alpha}(x)$,

(a) *when* $\alpha(x) > \alpha_1(x)$, $\gamma_1(n,h)\,\sigma_\alpha^{-1}(x)\,(\bar{\alpha}(x) - \alpha(x)) \xrightarrow{d} Z$.

(b) *when* $\alpha(x) = \alpha_1(x)$, $\Pr\left(\gamma_1(n,h)\,\sigma_\alpha^{-1}(x)\,(\bar{\alpha}(x) - \alpha(x)) < z\right) \to \Phi(z) \cdot 1_{\{z \geq 0\}}$.

*(ii) If we further assume that B.3, and denote* $Z_{b(x)} = Z + b(x)$, *then*

$$\lim_{n\to\infty} \gamma_1(n,h)\,\sigma_\beta^{-1}\left[\bar{\alpha}(x) - \alpha(x)\right] = Z_{b(x)} 1_{\left[Z_{b(x)} > 0\right]} - b(x).$$

**Corollary 3.** *Under B.1-B.3,*

(a) $\lim_{n\to\infty} \gamma_1(n,h)\,\sigma_\alpha^{-1} E\left[\bar{\alpha}(x) - \alpha(x)\right] = \varphi(b(x)) + b(x)\Phi(b(x)) - b(x)$.

(b) $\lim_{n\to\infty} Var\left[\left(\gamma_1(n,h)\,\sigma_\alpha^{-1}(x)\right)^{1/2} \bar{\alpha}(x)\right] = \Phi(b(x)) + b(x)\varphi(b(x)) - \varphi^2(b(x)) - 2b(x)\varphi(b(x))\,\Phi(b(x)) + b^2(x)\Phi(b(x))\left[1 - \Phi(b(x))\right]$.

**Remark 4.** The above theorem shows the counterpart results for nonparametric estimators with constraints. The implications are similar to the previous theorem on constrained parametric estimators. Note that the constraint bound $\alpha_1(x)$ can vary for different values of $x$. As a special case in which $\alpha_1(x) = \alpha_1$, a constant, it is efficient to adopt the restriction if it is correctly specified via the constrained estimator. The constrained estimator of $\alpha(x)$, $\bar{\alpha}(x)$, have the asymptotic property as the usual unconstrained nonparametric estimator as established in theorem 2.

### 3.4.2 Bagged Constrained Nonparametric Estimator

#### Assumption B

**(B.4)** $\gamma(n,h)\,\sigma_\alpha^{-1}\left(\tilde{\alpha}^*(x) - \tilde{\alpha}(x) - B_m(x)\right) \xrightarrow{d} Z$.

**Theorem 5.** *Under B.1-B.4, we have*

$$\gamma(n,h)\,\sigma_\alpha(x)^{-1}\left(\hat{\alpha}(x)-\alpha(x)-B_m(x)\right)\xrightarrow{d} Z\left[1-\Phi\left(-b(x)-Z\right)\right]+\varphi\left(-b(x)-Z\right).$$

**Corollary 4.** *If B.1-B.4 hold, then*

(a) $\lim_{n\to\infty}\gamma(n,h)\,\sigma_\beta^{-1}E\left[\hat{\alpha}(x)-\alpha(x)-B_m(x)\right]=2\varphi*\varphi\left(-b(x)\right)-b(x)\,\Phi*\varphi\left(-b(x)\right).$

(b) $\lim_{n\to\infty} Var\left[\left(\gamma(n,h)\,\sigma_\alpha^{-1}(x)\right)^{1/2}\hat{\alpha}(x)\right]=1+\Phi^2*\varphi''\left(-b(x)\right)+\Phi^2*\varphi\left(-b(x)\right)-$
$2b\Phi^2*\varphi'\left(-b(x)\right)+b^2(x)\,\Phi^2*\varphi\left(-b(x)\right)+\varphi^2*\varphi\left(-b(x)\right)-2\Phi*\varphi''\left(-b(x)\right)-$
$2\Phi*\varphi\left(-b(x)\right)+2b(x)\,\Phi*\varphi'\left(-b(x)\right)-2\varphi*\varphi'\left(-b(x)\right)+2\left(\Phi\cdot\varphi\right)*\varphi'\left(-b(x)\right)-$
$2b(x)\left(\Phi\cdot\varphi\right)*\varphi\left(-b(x)\right)-\left[2\varphi*\varphi\left(-b(x)\right)-b(x)\,\Phi*\varphi\left(-b(x)\right)\right]^2.$

**Remark 5.** When $b(\cdot)$ admits a constant function, the limiting distribution in theorem is the same as in parametric case. That is, for all possible values of $x$, $\gamma(n,h)\,\sigma_\alpha(x)^{-1}\left(\hat{\alpha}(x)-\alpha(x)-B_m\right.$ converges to the same random variable as
$\gamma(n)\,\sigma_\alpha^{-1}\left(\hat{\alpha}-\alpha\right)$ does in parametric case.

**Theorem 6.** *Under B.1-B.4, we have*

$$\hat{\alpha}(x)=\tilde{\alpha}(x)\,\Phi\left(Z_{b(x)}\right)+\alpha_1(x)\,\Phi\left(-Z_{b(x)}\right)+O_p\left(\tfrac{1}{\gamma(n,h)}\right).$$

**Remark 6.** Theorem 6 establish that the bagging estimator is a model averaging type estimator with a weight $\Phi(Z_b)$ assigned to the unconstrained estimator $\tilde{\alpha}(x)$ and a weight $\Phi\left(-Z_{b(x)}\right)$ to the lower bound $\alpha_1(x)$, up to order $O_p\left(\tfrac{1}{\gamma(n,h)}\right)$.

## 3.5 Simulation

In this section, we study the finite sample performance of our constrained estimator and its bagging version. We consider the following Data Generating Process.

$$y_i=a\left(4x_i-2\right)^3+e_i,$$

where $x_i$ is generated independently from a normal distribution with mean 1 and standard deviation 0.5. $e_i$ is *i.i.d.* standard normal disturbance. we consider $a$ to be a value taken from [0.001,0.05,0.01,0.05,0.1,0.5] that control the distance between $m(x)$ and 0. We evaluate the estimators at $x = 1$ and 1.5. We compute the mean of squared errors out of 200 replications. In each replication, we experiment with sample size $n = 50, 100, 200$, and bootstrap sample size 100. The relative mean squared errors are reported in Table 1 and 2.

We summarize the main findings as follows. At $x = 1$, the constrained estimator works better than unconstrained estimator for small values of $a$ in all sample sizes. The gain in mean squared errors can be as big as 34%. When $a$ gets larger, the constraint becomes non-binding and thus constrained estimator performs the same as the unconstrained. Bagging does not tend to work for sample size $n = 50$ for all values of $a$ considered here. When $n$ gets larger, bagging improves upon the constrained estimator when $a = 0.005$ and 0.01 with a maximum gain in mean squared error as large as 2%. This is consistent with the theory that bagging estimator works better than the constrained estimator when the sample size and the level of the funtion are of suitable proportion. For large values of $a$, the relative mean squared errors that are larger than 1 are due to sampling errors incurred in the bootstrap procedure.

The results become more appealing when the estimators are evaluated at $x = 1.5$. Again, the role of constraint becomes less important as $a$ gets larger, as can be seen from Table 2 that the relative mean squared error is increasing as $a$ increases. Bagging's role become more salient in this case, with a maximum gain in MSE as large as 10% when $a = 0.001$ and $n = 200$. As Figure 1 shows, the AMSE of bagging estimator can be over 10% smaller than constrained estimator. So the result we find is congruent with

Table 3.1: Relative Mean Squared Error: evaluated at $x = 1$

| | $n = 50$ | | $n = 100$ | | $n = 200$ | |
|---|---|---|---|---|---|---|
| $a$ | NPP/NP | NPPB/NP | NPP/NP | NPPB/NP | NPP/NP | NPPB/NP |
| 0.001 | 0.655 | 0.694 | 0.603 | 0.630 | 0.665 | 0.689 |
| 0.005 | 0.919 | 0.924 | 0.922 | 0.915 | 0.922 | 0.912 |
| 0.010 | 0.992 | 0.995 | 0.992 | 0.990 | 0.963 | 0.945 |
| 0.050 | 1.000 | 1.005 | 1.000 | 1.008 | 1.000 | 1.007 |
| 0.100 | 1.000 | 1.006 | 1.000 | 1.007 | 1.000 | 1.010 |
| 0.500 | 1.000 | 1.005 | 1.000 | 1.007 | 1.000 | 1.014 |

the asymptotic theory.

## 3.6 Application: Forecasting Equity Premium

To put our proposed estimators in practice, we consider to forecast U.S. Equity Premium. We adopt the data set used by Campbell and Thompson (2008). Equity premium is defined as the difference between the total rate of return on the stock market and the prevailing short-term interest rate. We consider use 11 predictors including dividend price ratio ($d/p$), earning price ratio ($e/p$), smooth earning price ratio ($se/p$), book to market ratio ($b/m$), return on equity ($roe$), treasure bill ($t\text{-}bill$), long term yield ($lty$), term spread ($ts$), default spread ($ds$), inflation ($inf$) and net equity issuance ($nei$). We follow Campbell and Thompson (2008) to impose the constraint that the equity premium should be positive. We consider the both annually and monthly forecasts starting from 1960 and 1980 and rolling till the end of 2005. The in-sample size for

Table 3.2: Relative Mean Squared Error: evaluated at $x = 1.5$

| | $n = 50$ | | $n = 100$ | | $n = 200$ | |
|---|---|---|---|---|---|---|
| $a$ | NPP/NP | NPPB/NP | NPP/NP | NPPB/NP | NPP/NP | NPPB/NP |
| 0.001 | 0.387 | 0.346 | 0.507 | 0.433 | 0.603 | 0.497 |
| 0.005 | 0.916 | 0.851 | 0.993 | 0.977 | 1.000 | 1.000 |
| 0.010 | 0.999 | 0.997 | 1.000 | 1.006 | 1.000 | 1.003 |
| 0.050 | 1.000 | 1.020 | 1.000 | 1.008 | 1.000 | 1.003 |
| 0.100 | 1.000 | 1.031 | 1.000 | 1.008 | 1.000 | 1.003 |
| 0.500 | 1.000 | 1.032 | 1.000 | 1.008 | 1.000 | 1.002 |

model estimation is kept fixed as 120. We report the results for Mean Squared Forecast Errors (MSFE) relative to the historical mean (M) forecast in Table 3 and 4.

In Table 3, we first see that nonparametric forecasts generally outperform the historical mean, except for the predictor *ts* starting from 1960 and *inf* starting from 1980. For these two predictors, however, we observe that imposing the positivity constraint reduces the MSFE, which is further reduced after the bagging procedure. Second, positive constraint works for predictors including *roe*, *t-bill*, *ldy*, *ts*, *ds* and *inf* to further reduce the MSFE of the nonparametric forecast. The maximum reduction in MSFE is achieved by *ds*, which is about 6%. Third, bagging does not always work for annual forecasts. The maximum gain that is achieved by bagging is about 2%, for the predictor *lty*.

For monthly forecasts, we hardly see much gain using unconstrained nonparametric methods. In Table 4, the only case that nonparametric MSFE gains, with 0.2% reduction, is for the predictor *d/p* when forecasts start from 1960. However, imposing

Table 3.3: Relative Mean Squared Forecast Error: Annually Results

| | 1960 | | | 1980 | | |
|---|---|---|---|---|---|---|
| | NP/M | NPP/M | NPPB/M | NP/M | NPP/M | NPPB/M |
| *d/p* | 0.690 | 0.714 | 0.721 | 0.624 | 0.648 | 0.655 |
| *e/p* | 0.860 | 0.859 | 0.865 | 0.833 | 0.821 | 0.820 |
| *se/p* | 0.885 | 0.886 | 0.898 | 0.771 | 0.771 | 0.773 |
| *b/m* | 0.751 | 0.772 | 0.770 | 0.945 | 0.947 | 0.949 |
| *roe* | 0.911 | 0.890 | 0.883 | 0.866 | 0.850 | 0.849 |
| *t-bill* | 0.934 | 0.903 | 0.895 | 0.916 | 0.915 | 0.908 |
| *lty* | 0.926 | 0.923 | 0.900 | 0.915 | 0.889 | 0.871 |
| *ts* | 1.071 | 1.031 | 1.018 | 0.923 | 0.929 | 0.932 |
| *ds* | 0.965 | 0.907 | 0.899 | 0.907 | 0.879 | 0.869 |
| *inf* | 0.990 | 0.975 | 0.979 | 1.044 | 1.013 | 1.013 |
| *nei* | 0.963 | 0.962 | 0.963 | 0.850 | 0.852 | 0.848 |

the positivity constraint always improves MSFE. This further confirms that the positivity constraint is consistent with the equity premium data. We further observe that bagging works most of the time. Especially for the predictors *d/p*, *e/p*, *b/m*, *roe*, *lty* and *inf*, bagging even help the nonparametric forecast to beat the "unbeatable" historical mean. This gain is as large as 1.1% for *lty*, which is significant according to Campbell and Thompson (2008).

Table 3.4: Relative Mean Squared Forecast Error: Monthly Results

| | 1960 | | | 1980 | | |
|---|---|---|---|---|---|---|
| | NP/M | NPP/M | NPPB/M | NP/M | NPP/M | NPPB/M |
| *d/p* | 0.998 | 0.996 | 0.995 | 1.024 | 1.010 | 0.995 |
| *e/p* | 1.011 | 0.998 | 0.996 | 1.013 | 1.010 | 0.994 |
| *se/p* | 1.050 | 1.004 | 1.003 | 1.149 | 1.081 | 1.054 |
| *b/m* | 1.001 | 1.000 | 0.995 | 1.036 | 1.033 | 1.029 |
| *roe* | 1.000 | 0.998 | 0.996 | 1.030 | 1.010 | 1.003 |
| *t-bill* | 1.144 | 1.048 | 1.038 | 1.067 | 1.019 | 1.012 |
| *lty* | 1.049 | 0.996 | 0.985 | 1.021 | 1.006 | 1.003 |
| *ts* | 1.026 | 1.011 | 1.024 | 1.090 | 1.054 | 1.046 |
| *ds* | 1.012 | 1.005 | 1.009 | 1.034 | 1.021 | 1.022 |
| *inf* | 1.011 | 0.999 | 0.997 | 1.034 | 1.025 | 1.023 |
| *nei* | 1.030 | 1.012 | 1.007 | 1.048 | 1.023 | 1.008 |

## 3.7 Conclusion

In this chapter, we presented the HM forecast with sign restriction. We established the asymptotic properties of the constrained parametric and nonparametric forecasts, and those of their bagging versions. Then we show the advantages of these forecasts over the unconstrained counterparts in both simulation and the forecasting of U.S. Equity Premium.

# Appendix

## Proof of Main Results

**Proof of Theorem 1.**

We first prove (**i**). For any $z \in R$,

$$\begin{aligned}
& \Pr\left(\gamma(n)\,\sigma^{-1}(\bar{\alpha}-\alpha)<z\right) \\
= \; & \Pr\left(\gamma(n)\,\sigma^{-1}(\max\{\tilde{\alpha},\alpha_1\}-\alpha)<z\right) \\
= \; & \Pr\left(\gamma(n)\,\sigma^{-1}(\max\{\tilde{\alpha},\alpha_1\}-\alpha)<z|\tilde{\alpha}<\alpha_1\right)\times\Pr(\tilde{\alpha}<\alpha_1) \\
& +\Pr\left(\gamma(n)\,\sigma^{-1}(\max\{\tilde{\alpha},\alpha_1\}-\alpha)<z|\tilde{\alpha}\geq\alpha_1\right)\times\Pr(\tilde{\alpha}\geq\alpha_1) \\
= \; & \Pr\left(\gamma(n)\,\sigma^{-1}(\alpha_1-\alpha)<z\right)\times\Pr(\tilde{\alpha}<\alpha_1)+ \\
& \Pr\left(\gamma(n)\,\sigma^{-1}(\tilde{\alpha}-\alpha)<z|\tilde{\alpha}\geq\alpha_1\right)\times\Pr(\tilde{\alpha}\geq\alpha_1)
\end{aligned}$$

in which,

(1) when $\alpha > \alpha_1$,

$$\Pr\left(\gamma(n)\,\sigma^{-1}(\alpha_1-\alpha)<z\right)\to\Pr(-\infty<z)=1,$$

since $\lim_{n\to\infty}\gamma(n)=\infty$, and when $\alpha=\alpha_1$,

$$\Pr\left(\gamma(n)\,\sigma^{-1}(\alpha_1-\alpha)<z\right)=\begin{cases}1, \text{ if } z>0\\ 0, \text{ if } z\leq 0\end{cases}$$

(2)

$$\begin{aligned}
& \Pr\left(\tilde{\alpha} < \alpha_1\right) \\
= \; & \Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha_1\right) < 0\right) \\
= \; & \Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha + \alpha - \alpha_1\right) < 0\right) \\
= \; & \Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha\right) < \gamma\left(n\right)\sigma^{-1}\left(\alpha_1 - \alpha\right)\right) \\
\rightarrow \; & \begin{cases} \Pr\left(Z < -\infty\right) = 0, & \text{if } \alpha > \alpha_1 \\ \Pr\left(Z < 0\right) = F\left(0\right), & \text{if } \alpha = \alpha_1 \end{cases}
\end{aligned}$$

(3)

$$\begin{aligned}
& \Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha\right) < z | \tilde{\alpha} \geq \alpha_1\right) \\
= \; & \frac{\Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha\right) < z, \gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha_1\right) \geq 0\right)}{\Pr\left(\gamma\left(n\right)\left(\tilde{\alpha} - \alpha_1\right) \geq 0\right)} \\
= \; & \frac{\Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha\right) < z, \gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha\right) \geq \gamma\left(n\right)\left(\alpha_1 - \alpha\right)\right)}{\Pr\left(\gamma\left(n\right)\left(\tilde{\alpha} - \alpha\right) \geq \gamma\left(n\right)\left(\alpha_1 - \alpha\right)\right)} \\
= \; & \begin{cases} \frac{F_Z(z) - F_Z(0)}{1 - F_Z(0)}, & \text{if } z > 0; \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

and (4)

$$\begin{aligned}
\Pr\left(\tilde{\alpha} \geq \alpha_1\right) &= 1 - \Pr\left(\tilde{\alpha} < \alpha_1\right) \\
&= 1 - \Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\tilde{\alpha} - \alpha\right) < \gamma\left(n\right)\left(\alpha_1 - \alpha\right)\right) \\
&\rightarrow \begin{cases} 1 - \Pr\left(Z < -\infty\right) = 1, & \text{if } \alpha > \alpha_1 \\ 1 - \Pr\left(Z < 0\right) = 1 - F\left(0\right), & \text{if } \alpha = \alpha_1 \end{cases}
\end{aligned}$$

Therefore, combining (1)-(4) leads to, (**i.a**) when $\alpha > \alpha_1$,

$$\Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\bar{\alpha} - \alpha\right) < z\right) \rightarrow 1 \times 0 + F_Z\left(z\right) \times 1 = F_Z\left(z\right)$$

and (**i.b**) when $\alpha = \alpha_1$, for $z > 0$,

$$\Pr\left(\gamma\left(n\right)\sigma^{-1}\left(\bar{\alpha} - \alpha\right) < z\right) \rightarrow 1 \times F_Z\left(0\right) + \frac{F_Z\left(z\right) - F_Z\left(0\right)}{1 - F_Z\left(0\right)} \times \left(1 - F_Z\left(0\right)\right) = F_Z\left(z\right)$$

and for $z = 0$,

$$\Pr\left(\gamma(n)\sigma^{-1}(\bar{\alpha}-\alpha) < z\right) \to 1 \times F_Z(0) + 0 \times (1 - F_Z(0)) = F_Z(0).$$

When $z < 0$,

$$\Pr\left(\gamma(n)\sigma^{-1}(\bar{\alpha}-\alpha) < z\right) \to 0.$$

Written compactly, we have

$$\Pr\left(\gamma(n)\sigma^{-1}(\bar{\alpha}-\alpha) < z\right) = F_Z(z)\,1_{\{z>0\}}.$$

To prove (**ii**), note that

$$\begin{aligned}
\gamma(n)\sigma^{-1}(\bar{\alpha}-\alpha) &= \gamma(n)\sigma^{-1}(\alpha_1-\alpha) + \gamma(n)\sigma^{-1}(\tilde{\alpha}-\alpha_1)\,1_{[\gamma(n)\sigma^{-1}(\tilde{\alpha}-\alpha_1)>0]} \\
&= \gamma(n)\sigma^{-1}(\alpha_1-\alpha) + \\
&\quad \gamma(n)\sigma^{-1}(\tilde{\alpha}-\alpha+\alpha-\alpha_1)\,1_{[\gamma(n)\sigma^{-1}(\tilde{\alpha}-\alpha+\alpha-\alpha_1)>0]} \\
&\to {}^{d} Z_b 1_{[Z_b>0]} - b.
\end{aligned}$$

by Assumption A.1 and A.2.

**Proof of Corollary 1.**

The results are directly applications of Lemma 1,

$$\begin{aligned}
E\left[Z_b 1_{[Z_b>0]} - b\right] &= EZ1_{[Z_b>0]} + bE1_{[Z_b>0]} - b \\
&= \phi(b) + b\Phi(b) - b,
\end{aligned}$$

and

$$\begin{aligned}
Var\left[Z_b 1_{[Z_b>0]} - b\right] &= Var\left[Z_b 1_{[Z_b>0]}\right] \\
&= E\left\{\left[Z_b 1_{[Z_b>0]}\right]^2\right\} - \left\{E\left[Z_b 1_{[Z_b>0]}\right]\right\}^2 \\
&= \Phi(b) + b\phi(b) + b^2\Phi(b) - \left[\phi(b) + b\Phi(b)\right]^2.
\end{aligned}$$

**Proof of Theorem 2.**

Write

$$\begin{aligned}\gamma(n)\sigma^{-1}(\hat{\alpha}-\alpha) &= \gamma(n)\sigma^{-1}\left(E^*\left[\tilde{\alpha}^* 1_{[\tilde{\alpha}^*\geq\alpha_1]}\right]+\alpha_1 E^*\left[1_{[\tilde{\alpha}^*<\alpha_1]}\right]-\alpha\right)\\ &= \gamma(n)\sigma^{-1}\left(E^*\left[(\tilde{\alpha}^*-\alpha)1_{[\tilde{\alpha}^*\geq\alpha_1]}\right]+(\alpha_1-\alpha)E^*\left[1_{[\tilde{\alpha}^*<\alpha_1]}\right]\right).\end{aligned}$$

Note that (1)

$$\begin{aligned}&\gamma(n)\sigma^{-1}\left(E^*\left[(\tilde{\alpha}^*-\alpha)1_{[\tilde{\alpha}^*\geq\alpha_1]}\right]\right)\\ =\;& E^*\left[\gamma(n)\sigma^{-1}(\tilde{\alpha}^*-\tilde{\alpha}+\tilde{\alpha}-\alpha)1_{[\tilde{\alpha}^*\geq\alpha_1]}\right]\\ =\;& E^*\left[\gamma(n)\sigma^{-1}(\tilde{\alpha}^*-\tilde{\alpha}+\tilde{\alpha}-\alpha)1_{[\gamma(n)\sigma^{-1}(\tilde{\alpha}^*-\tilde{\alpha})\geq\gamma(n)\sigma^{-1}(\alpha_1-\alpha)+\gamma(n)\sigma^{-1}(\alpha-\tilde{\alpha})]}\right]\\ &\xrightarrow{d} E_W\left[W1_{[W\geq -b]}|Z\right],\end{aligned}$$

where $W\sim N(Z,1)$.

Similarly, we get (2)

$$\gamma(n)\sigma^{-1}(\alpha_1-\alpha)E^*\left[1_{[\tilde{\alpha}^*<\alpha_1]}\right]\xrightarrow{p}-bF_Z(-Z_b),$$

by Slutsky's theorem.

Putting together (1) and (2) gives the desired result,

$$\gamma(n)\sigma^{-1}(\hat{\alpha}-\alpha)\xrightarrow{d}E_W\left[W1_{[W\geq -b]}|Z\right]-bF_Z(-Z_b).$$

**Proof of Corollary 2.**

(**i**) When Assumption Z holds,

$$
\begin{aligned}
& E_W\left[W1_{[W\geq -b]}|Z\right] \\
= \; & E_W\left[W\right] - E_W\left[W1_{[W<-b]}|Z\right] \\
= \; & Z - \int_{-\infty}^{-b} w\varphi\left(w-Z\right)dw \\
= \; & Z - \int_{-\infty}^{-b-Z} \left(s+Z\right)\varphi\left(s\right)ds \\
= \; & Z - Z\Phi\left(-b-Z\right) - \int_{-\infty}^{-b-Z} s\varphi\left(s\right)ds \\
= \; & Z - Z\Phi\left(-b-Z\right) + \varphi\left(-b-Z\right),
\end{aligned}
$$

together with results in Theorem 2, we have

$$
\gamma\left(n\right)\sigma^{-1}\left(\hat{\alpha}-\alpha\right) \xrightarrow{d} Z - Z_b\Phi\left(-Z_b\right) + \varphi\left(-Z_b\right).
$$

(**ii**) Repeated application of Lemma 2 leads to results in (**ii.a**) and (**ii.b**).

**Proof of Theorem 3.**

By definition,

$$
\begin{aligned}
\hat{\alpha} &= E^*\bar{\alpha}^* \\
&= E^*\left[\tilde{\alpha}^*1_{[\tilde{\alpha}^*\geq\alpha_1]}\right] + E^*\alpha_1\left[1_{[\tilde{\alpha}^*<\alpha_1]}\right] \\
&\equiv A_1 + A_2,
\end{aligned}
$$

where

$$
\begin{aligned}
A_2 &= E^*\alpha_1\left[1_{[\tilde{\alpha}^*<\alpha_1]}\right] \\
&= \alpha_1 E^*1_{[\tilde{\alpha}^*-\tilde{\alpha}\geq\alpha_1-\tilde{\alpha}]} \\
&= \alpha_1 E_W 1_{\{W\leq -Z_b\}} + O_p\left(\frac{1}{\gamma\left(n\right)}\right) \\
&= \alpha_1\Phi\left(-Z_b\right) + O_p\left(\frac{1}{\gamma\left(n\right)}\right)
\end{aligned}
$$

$$\begin{aligned} A_1 &= E^*\left[\tilde{\alpha}^* 1_{[\tilde{\alpha}^* \geq \alpha_1]}\right] \\ &= E^*\left\{\left[\tilde{\alpha}^* - \tilde{\alpha}\right] 1_{[\tilde{\alpha}^* - \tilde{\alpha} \geq \alpha_1 - \tilde{\alpha}]}\right\} + E^* \tilde{\alpha} 1_{[\tilde{\alpha}^* - \tilde{\alpha} \geq \alpha_1 - \tilde{\alpha}]} \\ &\equiv A_{11} + A_{12}, \end{aligned}$$

with

$$\begin{aligned} A_{11} &= E^*\left\{\left[\tilde{\alpha}^* - \tilde{\alpha}\right] 1_{[\tilde{\alpha}^* - \tilde{\alpha} \geq \alpha_1 - \tilde{\alpha}]}\right\} \\ &= E^*\left\{\left[\tilde{\alpha}^* - \tilde{\alpha}\right] 1_{[\tilde{\alpha}^* - \tilde{\alpha} \geq \alpha_1 - \alpha + \alpha - \tilde{\alpha}]}\right\} \\ &= \frac{1}{\gamma(n)} E^*\left\{\gamma(n)\left[\tilde{\alpha}^* - \tilde{\alpha}\right] 1_{[\gamma(n)(\tilde{\alpha}^* - \tilde{\alpha}) \geq \gamma(n)(\alpha_1 - \alpha + \alpha - \tilde{\alpha})]}\right\} \\ &= \frac{1}{\gamma(n)} E_W\left\{W 1_{[W \geq \gamma(n)(\alpha_1 - \alpha + \alpha - \tilde{\alpha})]}\right\} + o_p\left(\frac{1}{\gamma(n)}\right) \\ &= O_p\left(\frac{1}{\gamma(n)}\right), \end{aligned}$$

and

$$\begin{aligned} A_{12} &= \tilde{\alpha} E^* 1_{[\tilde{\alpha}^* - \tilde{\alpha} \geq \alpha_1 - \tilde{\alpha}]} \\ &= \tilde{\alpha} E_W 1_{\{W > -Z_b\}} \\ &= \tilde{\alpha}\Phi(Z_b) + O_p\left(\frac{1}{\gamma(n)}\right). \end{aligned}$$

Combining the results completes the proof, i.e.,

$$\hat{\alpha} = \tilde{\alpha}\Phi(Z_b) + \alpha_1\Phi(-Z_b) + O_p\left(\frac{1}{\gamma(n)}\right).$$

**Proof of Theorem 4, 5, 6, Corollary 3 and**

The proofs follow similar to those for Theorem 1, 2, 3 and Corollary 1 and 2, thus are omitted.

## Lemmas

We collect useful lemmas that are used in the proof of the main theorems. We use $Z$ to denote a standard normal random variable with CDF $\Phi(\cdot)$ and PDF $\varphi(\cdot)$, $b$

to denote some constant, and $1_{\{.\}}$ an indicator function. Define $Z_b = Z + b$.

**Lemma 1.** (**a**) $E1_{\{Z_b>0\}} = \Phi(b)$. (**b**) $E\left[Z1_{\{Z_b>0\}}\right] = \varphi(b)$. (**c**) $E\left[Z^2 1_{\{Z_b>0\}}\right] = -b\varphi(b) + \Phi(b)$. (**d**) $E\left[Z_b 1_{\{Z_b>0\}}\right] = \varphi(b) + b\Phi(b)$. (**e**) $E\left[Z_b^2 1_{\{Z_b>0\}}\right] = \Phi(b) + b\varphi(b) + b^2\Phi(b)$.

**Proof of Lemma 1.** (**a**) $E1_{\{Z_b>0\}} = E1_{\{Z>-b\}} = \int_{-b}^{\infty} d\Phi(z) = 1 - \Phi(-b) = \Phi(b)$. (**b**) $EZ1_{\{Z_b>0\}} = \int_{-b}^{\infty} z\varphi(z)\,dz = -\int_{-b}^{\infty} \varphi'(z)\,dz = -\varphi(z)|_{-b}^{\infty} = \varphi(b)$. (**c**) $EZ^2 1_{\{Z_b>0\}} = \int_{-b}^{\infty} z^2\varphi(z)\,dz = -\int_{-b}^{\infty} z\varphi'(z)\,dz = -z\varphi(z)|_{-b}^{\infty} + \int_{-b}^{\infty} \varphi(z)\,dz = -b\varphi(b) + \Phi(b)$. (**d**) $E\left[Z_b 1_{\{Z_b>0\}}\right] = EZ1_{\{Z_b>0\}} + bE1_{\{Z_b>0\}} = \varphi(b) + b\Phi(b)$. (**e**) $E\left[Z_b^2 1_{\{Z_b>0\}}\right] = E\left[(Z+b)^2 1_{\{Z_b>0\}}\right] = EZ^2 1_{\{Z_b>0\}} + b^2 E1_{\{Z_b>0\}} + 2bE\left[Z1_{\{Z_b>0\}}\right] = \Phi(b) - b\varphi(b) + b^2\Phi(b) + 2b\varphi(b) = \Phi(b) + b\varphi(b) + b^2\Phi(b)$.

**Lemma 2.** (**a**) $E\varphi(-Z_b) = \varphi * \varphi(-b)$. (**b**) $E\varphi^2(-Z_b) = \varphi^2 * \varphi(-b)$. (**c**) $E[Z\varphi(-Z_b)] = -\varphi * \varphi'(-b)$. (**d**) $E[Z\Phi(-Z_b)] = -\varphi * \varphi(b)$. (**e**) $E\left[Z^2\Phi(-Z_b)\right] = \Phi * \varphi''(-b) + \Phi * \varphi(-b)$. (**f**) $E\left[Z^2\Phi^2(-Z_b)\right] = \Phi^2 * \varphi''(-b) + \Phi^2 * \varphi(-b)$. (**g**) $E[Z\Phi(-Z_b)\varphi(-Z_b)] = -(\Phi\cdot\varphi) * \varphi'(-b)$.

**Proof of Lemma 2.**

**(a)** $E\varphi(-Z_b) = E\varphi(-b-Z) = \int_{-\infty}^{\infty} \varphi(-b-z)\varphi(z)\,dz = \varphi * \varphi(-b)$.

**(b)** $E\varphi^2(-Z_b) = E\varphi^2(-b-Z) = \int_{-\infty}^{\infty} \varphi^2(-b-z)\varphi(z)\,dz = \varphi^2 * \varphi(-b)$.

**(c)** $E[Z\varphi(-Z_b)] = E[Z\varphi(-b-Z)] = \int_{-\infty}^{\infty} z\varphi(-b-z)\varphi(z)\,dz = -\int_{-\infty}^{\infty} \varphi(-b-z)\varphi'(z)\,dz = -\varphi * \varphi'(-b)$.

**(d)** $E[Z\Phi(-Z_b)] = E[Z\Phi(-b-Z)] = \int_{-\infty}^{\infty} z\Phi(-b-z)\varphi(z)\,dz = -\int_{-\infty}^{\infty} \Phi(-b-z)\varphi'(z)\,dz$
$= -\left\{\Phi(-b-z)\varphi(z)|_{z=-\infty}^{\infty} - \int_{-\infty}^{\infty} -\varphi(-b-z)\varphi(z)\,dz\right\} = -\varphi * \varphi(-b)$.

**(e)** $E\left[Z^2\Phi(-Z_b)\right] = E\left[Z^2\Phi(-b-Z)\right] = \int_{-\infty}^{\infty} z^2\Phi(-b-z)\varphi(z)\,dz$
$= \int_{-\infty}^{\infty} \Phi(-b-z)[\varphi(z) + \varphi''(z)]\,dz = \Phi * \varphi''(-b) + \Phi * \varphi(-b)$.

**(f)** $E\left[Z^2\Phi^2\left(-Z_b\right)\right] = E\left[Z^2\Phi^2\left(-b-Z\right)\right] = \int_{-\infty}^{\infty} z^2\Phi^2\left(-b-z\right)\varphi\left(z\right)dz$

$= \int_{-\infty}^{\infty} \Phi^2\left(-b-z\right)\left[\varphi\left(z\right)+\varphi''\left(z\right)\right]dz = \Phi^2 * \varphi''\left(-b\right) + \Phi^2 * \varphi\left(-b\right).$

**(g)** $E\left[Z\Phi\left(-Z_b\right)\varphi\left(-Z_b\right)\right] = E\left[Z\Phi\left(-b-Z\right)\varphi\left(-b-Z\right)\right] = \int_{-\infty}^{\infty} z\Phi\left(-b-z\right)\varphi\left(-b-Z\right)\varphi\left(z\right)dz$

$= -\int_{-\infty}^{\infty} \Phi\left(-b-z\right)\varphi\left(-b-Z\right)\varphi'\left(z\right)dz = -\left(\Phi\cdot\varphi\right) * \varphi'\left(-b\right).$

## References


Breiman, L. (1996): "Bagging Predictors," *Machine Learning*, 36, 105–139.

Brunk, H. D. (1955), "Maximum likelihood estimates of monotone parameters", *Annals of Mathematical Statistics* 26, 607-616.

Bühlmann, P., and B. Yu (2002): "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961.

Campbell, J., and S. Thompson (2008), "Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies*, forthcoming.

Chen, Q. and Y. Hong (2009), "Predictability of Equity Returns over Different Time Horizons: A Nonparametric Approach," Working paper, Cornell University.

Chernozhukov, V., I., Fernandez-Val and A., Galichon (2007), Improving estimates of monotone functions by rearrangement. Mimeo.

Dette, H., N., Neumeyer and K. F., Pilz (2006), "A simple nonparametric estimator of a strictly monotone regression function", *Bernoulli* 12(3), 469–490.

Fama, E. F., and K. R. French, (1988), "Dividend yields and expected stock returns," *Journal of Financial Economics*, 22(1):3–25.

Fama, E. F., and K. R. French, (2002), "The Equity Premium," *The Journal of Finance* vol. 57(2), 637-659.

Gordon, I., and P. Hall (2008): "Estimating a Parameter When It Is Known that the Parameter Exceeds a Given Value," Working paper, University of Melbourne.

Goyal, A., and I. Welch (2008), "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21-4, 1455-1508.

Hall, P. and Huang, H. (2001), "Nonparametric kernel regression subject to monotonicity constraints," *The Annals of Statistics* 29(3), 624–647.

Henderson, D. J. and C. F., Parmeter (2009), "Imposing economic constraints in nonparametric regression: Survey, implementation and extension," Virginia Tech Working Paper 07/09.

Hillebrand, E., T.-H., Lee, M. Medeiros (2009), "Bagging Constrained Forecasts with Application to Forecasting Equity Premium," *American Statistical Association*, JSM Proceedings for Business and Economic Statistics

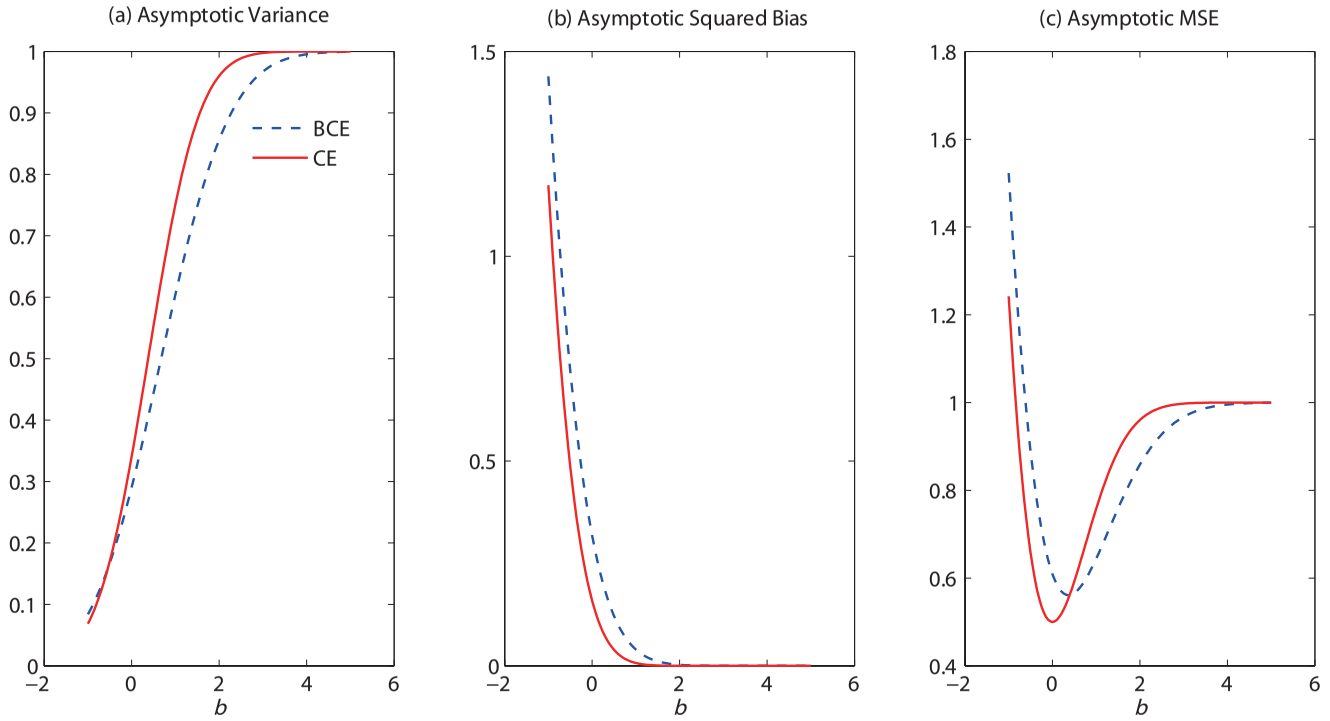
Lee, T.-H., and Y. Yang (2006): "Bagging Binary and Quantile Predictors for Time Series," *Journal of Econometrics*, 135, 465–497.

Li, Q. and J., Racine (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Pagan, A. and A., Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press, Cambridge.

Racine, J.S., C.F., Parameter and P., Du (2009), "Constrained nonparametric kernel regression: estimation and inference," Working chapter

Figure 3.1: Asymptotic variance, Asymptotic squared bias, and Asymptotic mean squared error of constrained estimator (CE) and bagging constrained estimator (BCE)



# Chapter 4

# Forecasting Using Supervised Factor Models

## 4.1 Introduction

High dimensional information in the presence of many predictors brings opportunities to improve the *efficiency* of a forecast by using much richer information than conventionally used and to enhance the *robustness* of a forecast against structural instability which can plague low dimensional forecasting. However, these opportunities come with the challenges. One notable challenge is that the availability of overwhelming information complicates the way we process it to make relevant instruments. For example. a large number of predictors makes the ordinary least square (OLS) estimation inadmissible.

Two directions can be marched to approach this challenge. The first direction focuses on variable selection. Variable selection or subset selection refers to selecting variables that are most predictive for a target variable of interest. Various variable selection methods have been proposed such as forward and backward selection, stepwise

regression, as presented in e.g. Miller (2002) and Hastie et al (2009). Recently the literature is crowded by more sophisticated methods such as LASSO (Tibshirani 1996, Zou 2006), Elastic Net (Zou and Hastie 2005, Zou and Zhang 2009), SCAD (Fan and Li 2001), Bridge (Huang, Horowitz and Ma 2008), Least Angle Regression (Efron et al 2004) and so on. All these methods seek to rank the variables and select a subset of variables based on their ranks. The second direction assumes the existence of a low dimensional latent factors in the high dimensional predictors. This approach includes Principal Component Regression (PCR), Partial Least Square (PLS) Regression (de Jong 1993, Garthwait 1994), Principal Covariate Regression (PCovR) (de Jong 1992), and Combining Forecast Principal Components (CFPC) (Huang and Lee 2010).

Considering the two directions, a natural question arises. Is it worthwhile to supervise the selection of predictors for a forecast target variable? Bai and Ng (2008) raised this question. They reported that after variable selection (by either hard-threshold method or soft-threshold method) PCR performs much better, reducing the mean squared forecast error (MSFE) to a large extent.

However, the PCR accounts only for the variation of the selected predictors, but does not directly employ the information about the forecast target. That is, no matter which variable to forecast (whether it is output growth, unemployment, stock returns, bond yields, housing price, interest rate, or inflation), the PCR uses the latent factors of the predictors only. Hence, the next question arises whether we can take a particular forecast target into the computation of latent factors. This chapter addresses this question, by considering the three supervised factor models (PLS, PCovR, CFPC). The question is whether the supervised factors from these supervised factor models are more efficient and more robust in out-of-sample forecasting than the unsupervised factors from PCR. We examine the properties of these factor models and compare their empirical

performance with supervision on the variable section and on the factor computation. The evidence is very clear. These supervisions do substantially improve the prediction. The predictive ability of the three supervised factor models is much better than the unsupervised PCR model. Interestingly, we find that the effect of supervision gets even larger as forecast horizon increases and that the supervision helps a model achieving more parsimonious structure. Among the three supervised factor models, the CFPC performs best and is most stable. While PCovR is nearly as efficient and robust as CFPC, PLS is neither as good nor stable as CFPC and PCovR. The performance of PLS is not robust over different out-of-sample forecasting periods and over the different forecast horizons.

The rest of the chapter is organized as follows. Section 4.2 introduces the basic forecasting setup and preliminary material that is needed for the understanding of factor models. Section 4.3 presents the unsupervised factor model, PCR. Section 4.4 examines the supervised factor models, PLS, PCovR and CFPC. Section 4.5 looks into ways to supervise the factor computation together with variable selection. In Section 4.6, forecasting exercises are carried out to compare the performance of these forecasting models for monthly CPI inflation in U.S. Section 4.7 concludes.

## 4.2 General Framework: Linear Factor Model

Consider the linear regression model,

$$y = X\beta + e, \tag{4.1}$$

where $y$ is a $T \times 1$ vector, $X$ is a $T \times N$ matrix of explanatory variables and $\beta$ is the true but unknown parameter. In case of $N \gg T$, or when columns of $X$ are highly correlated, the OLS estimation of the regression coefficient $\beta$ is not feasible. Hence, for

the purpose of forecasting, we consider the following factor model,

$$F = XR, \tag{4.2}$$

$$XB = FP' + E, \tag{4.3}$$

$$y = UQ' + G. \tag{4.4}$$

Here $F$ is a $T \times r$ factor matrix. Equation (4.2) says that the factor is linear in $X$. Each column of $F$ is a factor, which is a linear combination of rows of $X$. The $N \times r$ matrix $R$ is the weight matrix attaching to $X$. $U$ is the factor matrix for $y$, which is usually assumed to be the same as $F$. However, the estimation of $U$ varies as we take different estimation approaches and it can be far different from $F$ as in PLS. $P$ and $Q$ are corresponding factor loading for $X$ and $y$. The $N \times N$ matrix $B$ is called "supervision matrix". Note that the factor structure (4.3) contains that of Stock and Watson (2002a) and Bai (2003) as a special case, with $B$ being the identity matrix. As it is formulated, (4.4) is a linear factor model due to the linearity in both the construction of $F$ in (4.2) and the prediction equation (4.4). $E$ and $G$ are the error terms.

In the case that the number of factors used in (4.4) is less than or equal to the number of observations, $T$, the coefficient $Q$ can be estimated by OLS estimator $\hat{Q}$, with $U$ being estimated by $\hat{U}$. The forecast is formed as

$$\hat{y} = \hat{U}\hat{Q}'. \tag{4.5}$$

The factor models, PCR, PLS, PCovR and CFPC, that we consider in this chapter all fall into this general framework of (4.2), (4.3) and (4.4), with different ways of specifying $R$, $U$ and $B$. For example, as will be seen in the next section, PCR takes $B$ as the identity matrix, and then forms the weight matrix $R$ to be the matrix of eigenvectors of $X'X$, with $U$ being $F$.

The choice of the weight matrix and number of factors is the focus of factor modelling. To choose the number of factors, the usual information criterion such as AIC or BIC can be used. In the empirical section (Section 4.6), we will look into this aspect in further details. We focus on the choice of weight matrix in the next two sections. Section 4.3 will present a popular (unsupervised) factor model, PCR, which has been extensively used in economic forecasting as well as in other social sciences. See Stock and Watson (2002a). PCR is unsupervised and methods of supervising it will be presented in Section 4.4.

## 4.3 Principal Component Regression Model

In this section we review how PCR can be used in forecasting. First we begin by using the eigenvalue decomposition, and then in Section 4.4.1 we show PCR in an alternative framework for the principal component analysis. The purpose of presenting these two alternative framework is that we will use the former to introduce a supervised factor model called CFPC in Section 4.4.3 and we will use the latter to introduce another supervised factor model called PLS in Section 4.4.1.

Note that, PCR is when $P = R, B = I$ and $U = F$ in the framework in Section 4.2, namely:

$$\begin{aligned} F &= XR, \\ X &= FR' + E, \\ y &= FQ' + G, \end{aligned}$$

where $R$ is the matrix of eigenvectors of $X'X$. Stock and Watson (2002a) considered the case when $y$ is one variable with $(X, y)$ admitting the factor representation of (4.3) and (4.4). Equation (4.4) specifies the forecast equation while (4.3) gives the factor

structure. The factor $F$ in (4.2) is estimated using principal components and then it is used to form the prediction from (4.4) for $y$.

**Proposition 1.** *Let the $N \times r$ ($r \leq \min(T, N)$) matrix $R_1$ be the first $r$ eigenvectors, corresponding to the largest $r$ eigenvalues $\Lambda_1 = diag(\lambda_1, \ldots, \lambda_r)$ of $X'X$. Then, (i) the PCR estimator of $F$ is*

$$\hat{F} = XR_1,$$

*(ii) the OLS estimator of the factor loading for $y$ in (4.4) is*

$$Q' = \Lambda_1^{-1} R_1' X' y,$$

*and (iii) the PCR prediction of $y$ is*

$$\hat{y}_{\text{PCR}} = XR_1\Lambda_1^{-1}R_1'X'y. \tag{4.6}$$

□

The proof is in Appendix C, where the notation used in Proposition 1 is also established. For example, $R_1$ denotes a matrix containing the first $r$ columns of $R$, and similarly for $\Lambda_1$. Note that $XR_1$ is called the first $r$ principal component of $X$. Hence, the factors $F$ is estimated by the principal components of $X$.

The main criticism on PCR goes as follows. In the choice of the weight matrix $R$, PCR imposes only the factor structure for $X$. This is naive since it does not take into account the dependent variable $y$. That is, no matter what $y$ to forecast, PCR use the same fixed combination of $X$ to form the prediction equation. Ignoring the target information of $y$ in the computation of the factors leads to inefficient forecast of the $y$. Therefore, a supervision on the choice of weight matrix and thus supervised factor models will be called for to make more efficient predictions. This issue is to be addressed in the next section.

## 4.4 Supervised Factor Models

In this section we consider three supervised factor models, the partial least square (PLS), principal covariate regression (PCovR), and the combining forecast-principal component (CFPC). The analysis here is based on the factor framework in Section 4.2. The three models are generalization of the PCR in different ways to supervise the factors for the forecast target $y$.

### 4.4.1 Partial Least Square Regression

Although originally proposed by Wold (1966) in the field of econometrics, the partial least square (PLS) regression has rarely been used in economics but rather popular in chemometrics. Empirical results in chemistry show that PLS is a good alternative to multiple linear regression and PCR methods. See Wold et al (1984), Otto and Wegscheider (1985), and Garthwait (1994) for more details. Since PLS also supervises the factor computation process, it raises the possibility that it can outperform PCR, which is a reason that we include the PLS in this chapter.

There have been several algorithms designed for PLS, among which NIPALS and SIMPLS are most notable ones. de Jong (1993) has shown that results obtained with NIPALS and SIMPLS algorithms turn out to be the same for the univariate dependent variable case. The two algorithms give slightly different results for the case of multivariate dependent variable (due to the difference in the deflation of information matrix). In the next subsection we review the NIPALS algorithm briefly to show PLS in the general framework of factor models, (4.2), (4.3) and (4.4), in Section 4.2. The purpose of the next subsection is to show that PLS can be viewed as a generalization of PCR.

#### 4.4.1.1 NIPALS algorithm for PCR and PLS

Alternative to the eigenvalue decomposition used in Proposition 1 for PCR, we can use the Nonlinear Iterative PArtial Least Square (NIPALS) algorithm developed by Wold (1966, 1975) to perform the principal component analysis, which decomposes matrix $X$ of rank $r$ as a sum of $r$ matrices of rank 1 as

$$\begin{aligned} X &= M_1 + M_2 + \ldots + M_N \\ &= f_1 p_1' + f_2 p_2' + \ldots + f_r p_r' + f_{r+1} p_{r+1}' + \ldots + f_N p_N' \\ &\equiv FP' + E_r, \end{aligned} \tag{4.7}$$

where the second line uses the fact that the rank 1 matrices $M_h$ can be written as outer products of two vectors, $f_h$ (score) and $p_h'$ (loading), and $F = [f_1, f_2, \ldots, f_r]$, $P' = [p_1', p_2', \ldots, p_r']$. NIPALS does not compute all the principal components $F$ at once. But it calculates $f_1$ and $p_1'$ from $X$, then the outer product $f_1 p_1'$ is subtracted from $X$, and the residual $E_1$ is calculated. This residual is used to calculate $f_2$ and $p_2'$, and so on. The formal NIPALS algorithm for PCR is stated in Appendix A, where it is shown that, on convergence, the NIPALS algorithm gives the same principal components as derived by the eigenvalue decomposition of Proposition 1. The algorithm does converge in practice.

Now, to see how this algorithm can be extended from PCR to PLS, let us turn back to the regression problem (4.4). The NIPALS algorithm can work for both $X$ and $y$ separately to extract factors as in (4.7). That is,

$$\begin{aligned} X &= FP' + E_r = \sum_{h=1}^{r} f_h p_h' + E_r, \\ y &= UQ' + G_r = \sum_{h=1}^{r} u_h q_h' + G_r. \end{aligned} \tag{4.8}$$

Thus we can form an inner relationship between $x$-score, $f$, and $y$-score, $u$ as

$$u_h = b_h f_h + \epsilon_h, \tag{4.9}$$

for each pair of components. OLS estimation can be used for (4.9) thus we could use (4.8) to form a prediction with $x$-scores, $f$, extracted with newly observed $x$.

However, note that the decomposition process in (4.8) still does not incorporate the valuable information of $y$ when forming the $x$-scores. Thus we consider the modification of the decomposition of $X$ and $y$, using NIPALS, as stated in Appendix B. Note that in the special case of $y = X$, $x$-factors extracted by NIPALS gives exactly the principal components of $X$ as one might have already conjectured. Thus in this case, NIPALS for PLS is the same as NIPALS for PCR. See Geladi and Kowalski (1986) and Mardia et al (1980) for an excellent discussion for NIPALS algorithm and its adaptations for PCR and PLS.

### 4.4.2 Principal Covariate Regression

Principal Covariate Regression (PCovR) is a novel prediction method proposed by de Jong and Kiers (1992). "Covariate" was termed to stress that, apart from PCR, the components should vary with the dependent variable $y$. The attractiveness of PCovR features its combination of PCR on $X$ and a regression on $y$ by minimizing an appropriately defined least square loss function as follows,

$$l\left(\alpha_1, \alpha_2, R, P, Q\right) \equiv \alpha_1 \parallel X - XRP' \parallel^2 + \alpha_2 \parallel y - XRQ' \parallel^2 , \tag{4.10}$$

where $\alpha_1$ and $\alpha_2$ are the (non-negative) weights attached to PCR on $X$ and regression of $y$, respectively. That is, the choice of the factor weight matrix $R$ depends not only on the PCR of $X$, but also on the regression equation (4.4). Then the factor is computed from $F = XR$ as in (4.2).

Some special cases of PCovR needs to be pointed out here. For $\alpha_1 = 0$, the (4.10) emphasizes completely on fitting $y$. Specifically, when dependent variable is univariate, the first component $t_1$, can be chosen as the component being maximally correlated with $y$. And remaining components, which are irrelevant, are the principal components of the part of $X$ that is orthogonal to $y$. Another extreme is when $\alpha_2 = 0$. In this case, (4.10) emphasizes completely on the principal component analysis on $X$ or PCR as described in Section 4.3.

Note that the minimization of (4.10) is nonlinear in nature due to the product terms $RP$ and $RQ$. An algorithm for the estimation of the unknown parameters $(R, P, Q)$ is given in de Jong (1993). Or see Heij, Groenen and van Dijk (2007) for an explicit SVD based algorithm.[1]

Although supervision is incorporated in PCovR by allocating weight to the regression (4.4), there is no guidance regarding the optimal choice of the weight attached. Thus choice of $\alpha_1$ and $\alpha_2$ can only be done on rather arbitrary grounds. In practice, one might need consider a set of specifications for $\alpha_1$ and $\alpha_2$, as did in Heij, Groenen and van Dijk (2007).

For prediction purpose, we propose an estimation of optimal weights by a grid searching algorithm, with the exploit of information available. Note that only the relative weights attached matter here. We consider a normalization of the weights,

$$\alpha_1 = w / \parallel X \parallel^2 , \quad \text{and } \alpha_2 = (1 - w) / \parallel y \parallel^2 .$$

Therefore, we need to consider a choice of $w$ instead of choices of $\alpha_1$ and $\alpha_2$ simultaneously. In Section 4.6, we choose the value of $w$ from $\{10^{-6}, 10^{-4}, 0.1, 0.5, 0.9\}$ that minimizes model selection criterion, such as BIC.

[1]We would like to thank Christiaan Heij and Dick van Dijk for kindly sharing their Matlab code for PCovR.

### 4.4.3 CFPC

This subsection discusses another form of supervision on the computation of factors. This is a method quite different from those examined earlier in this section. The two previous supervised models directly compute the factors, while CFPC first computes forecasts and then computes the principal components of the forecasts as a tool to combining forecasts.

**Proposition 2.** *Define $\hat{y}_i = b_i x_i$ using $x_i$ from the $i$-th column of $X = (x_1, x_2, \ldots, x_N)$, and let $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N) \equiv XB$ with $B = diag(b)$. Let $L_1$ be the $N \times r$ eigenvectors corresponding to the $r$ largest eigenvalues $\Omega_1 = diag(\omega_1, \omega_2, \ldots, \omega_r)$ of $\hat{Y}'\hat{Y} = B'X'XB$. Then (i) the CFPC estimator of the factor $F$ of $XB$ is*

$$\hat{F} = \hat{Y}L_1 = XBL_1$$

*i.e., $R$ is estimated by $BL_1$, (ii) the OLS estimator of the factor loading for $y$ in (4.4) is*

$$\hat{Q} = \Omega_1^{-1} L_1' \hat{Y}' y = \Omega_1^{-1} L_1' B X' y,$$

*and (iii) the CFPC prediction of $y$ is*

$$\hat{y}_{\text{CFPC}} = \hat{Y} L_1 \Omega_1^{-1} L_1' \hat{Y}' y = XBL_1 \Omega_1^{-1} L_1' B X' y. \tag{4.11}$$

□

The proof is in Appendix C, where further notation used above in Proposition 2 is also established.

**Remark 1** (Combining forecasts with many forecasts)**:** Although Proposition 2 is explicitly stated for $\hat{Y} = XB$, the result is useful when we observe only $\hat{Y}$ but not $X$

(e.g., survey of professional forecasters). The CFPC forecast, $\hat{y}_{\text{CFPC}} = \hat{Y} L_1 \Omega_1^{-1} L_1' \hat{Y}' y$, would then produce a method of combining $N$ forecasts in $\hat{Y}$ when $N \to \infty$. □

**Remark 2**: The biggest difference between CFPC and PCR lies in the set of variables we use to extract the principal components. In PCR, the principal components are computed from $x$'s directly, without accounting for their relationship with the forecast target variable $y$. This problem with PCR leads Bai and Ng (2008) to consider first selecting a subset of predictors ("targeted predictors") of $x$'s that are informative in forecasting $y$, then using the subset to extract factors. In contrast, since CFPC combines forecasts not the predictors, the principal components in CFPC are computed from the set of individual forecasts $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ that contain both information on $x$'s and on all past values of $y$. This actually provides us further intuitions on why CFPC may be more successful than PCR. □

**Remark 3:** Chan, Stock and Watson (1999) and Stock and Watson (2004) choose the factor analytic structure where the set of individual forecasts permits one single factor. The specifications for individual forecasts in CFPC, however, differ from those in Chan, Stock and Watson (1999) and Stock and Watson (2004) in that individual forecasting models considered here use different and non-overlapping information sets, not a common total information set as assumed in Chan, Stock and Watson (1999) and Stock and Watson (2004). □

**Remark 4** (Comparison of PCR and CFPC when $X$ has full column rank): Instead of using original predictors $X$ to form principal components, CFPC uses the predicted matrix of $y$, $\hat{Y}$. This is where supervision is incorporated. It is interesting to note that there are cases that PCR and CFPC give the same prediction. Note that in case of $N \leq T$ and when $X$ has full column rank, and each column of $X$ is predictive for $y$

($b_i \neq 0$ for all $i = 1, \ldots, N$), we could exhaust all principal components of $X$ and those of $\hat{Y}$. Thus we have, from (4.25),

$$R_1 \Lambda_1^{-1} R_1' = \left(X'X\right)^{-1}. \tag{4.12}$$

And also

$$BL_1 \Omega_1^{-1} L_1' B = B\left(\hat{Y}'\hat{Y}\right)^{-1} B = B\left(BX'XB\right)^{-1} B = \left(X'X\right)^{-1}, \tag{4.13}$$

where the last equality follows from the fact that $B$ is also a full rank diagonal matrix. Thus, combining (4.6), (4.11), (4.12) and (4.13) gives

$$\hat{y}_{\text{PCR}} = \hat{y}_{\text{CFPC}}.$$

Therefore, PCR and CFPC are equivalent in this case when $X$ has a full column rank. When $X$ does not have a full column rank, the principal components of the forecasts in CFPC and the principal components of predictors in PCR will differ from each other, because the linear combinations maximizing covariances of forecasts (for which the supervision operates for the relationship between $y$ and $X$) and the linear combinations maximizing the covariances of predictors (for which there is no supervision) will be different. □

**Remark 5** (Regression one-at-a-time): CFPC described here employs the regression of $y$ on $x_i$ one-at-a-time to formulate the prediction matrix $\hat{Y}$, which will be justified in Proposition 3 below. It is simple to implement and computationally appealing. Nevertheless, it can be generalized in various ways. For example, an information based model averaging approach in line with Tu (2011) could be developed. □

Intuitively and computationally appealing, CFPC also enjoys some theoretical justification as presented as follows. Note that, in the case that $N \leq T$, a regression of

$y$ on $X$ would give

$$\hat{y} = x_1 b_1 + x_2 b_2 + \ldots + x_N b_N \equiv XB, \tag{4.14}$$

where $B = \text{diag}(b)$. Further from (4.11) we define a function $f(\cdot)$ such that

$$b = BL_1\Omega_1^{-1}L_1'BX'y = \text{diag}\,(b)\, L_1\Omega_1^{-1}L_1'\text{diag}\,(b)\, X'y \equiv f(b). \tag{4.15}$$

We now show that the true parameter $\beta$ is an asymptotic fixed point for $f(\cdot)$, by construction of $B = \text{diag}(b)$. We first state assumptions:

**Assumption 1:** (a) The process $\{X_t, y_t\}$ is jointly stationary and ergodic. (b) $E\left[X_t'(y_t - X_t\beta)\right] = 0$. (c) $\beta$ is an interior point of parameter space $\Theta$. (d) Assumptions A-F of Bai (2003) are satisfied for the factor structure (4.3) with $B = \text{diag}(\beta)$. (e) $||\Sigma_{XB}^{-1}\Sigma_E|| = O_p\,(N/T)$, where $\Sigma_\xi$ denotes the variance-covariance matrix of $\xi$, and $||\cdot||$ denotes a matrix norm. (f) $N^2/T \to 0$, as $N, T \to \infty$.

**Proposition 3:** *Under Assumption 1, the true parameter $\beta$ in (4.1) is an asymptotic fixed point for $f(\cdot)$ defined in (4.15), that is,*

$$\left(f(\beta) - \beta\right)_i = O_p\left(\max\left\{\frac{N}{\sqrt{T}}, \frac{N^2}{T}\right\}\right) = o_p(1) \quad \textit{for all } i,$$

*where $a_i$ denotes the $i$-th element of $a$.* □

The proof is in Appendix C.

**Remark 6** (fixed point): Proposition 3 justifies the construction of the supervision matrix $B = \text{diag}(b)$ in Proposition 2. When we start with $B = \text{diag}(b)$ such that $b$ is close to $\beta$, CFPC would give an estimate of $\beta$, $f(b)$, which is close enough to the true value $\beta$ in the sense of Proposition 3. □

## 4.5 Supervising Factor Models with Variable Selection

The previous section looks into supervised factor models from the perspective of supervising the formation of latent factors for a given set of original predictors. Before that step, we can consider selecting a subset of the predictor variables. Boivin and Ng (2006) raise the concern of the quality of data when researchers are ambitious to employ all data available from large panels. Through simulation and application examples, they show that factors extracted from a smaller set of variables are likely to perform no worse, and in many cases even better, in forecasting than those extracted from a lager set of series.

To forecast using a subset of variables when there are too much information has been a popular research topic and many methods have been developed to tackle the issue – see Miller (2002) and Hastie et al (2009). Variable selection in forecasting in the presence of many predictors is not as simple as in an AR model for which the lags have a natural order. Predictors are naturally not in order. Thus we can not determine which variables should be included and which are not unless we find ways to rank them. The principles used to rank the predictors can be in two categories: hard-thresholding and soft-thresholding.

### 4.5.1 Hard-Threshold Variable Selection

The method of hard-thresholding is to use a statistical test to determine if a particular predictor is significant in forecasting, without considering the effect of other predictors. Bair et al (2006) take this approach. (Although their model is termed as supervised principal component model, their supervision is in selection of predictors but not in computation of the principal components. Supervision there is only performed

via variable selection, but not directly through the factor computation process.) In this chapter, lags of $y_t$ are included as regressors with each individual $x_{it}$ to get the individual $t$-statistic as an indicator of the marginal predictive power of $x_{it}$, following Bai and Ng (2008). It involves the following steps: For each $i = 1, ..., N$, run the regression of $y_{t+h}$ on a constant, four lags of $\{y_{t-j}\}_{j=0}^{3}$ and $x_{it}$. Let $t_i$ denote the $t$-statistic associated with the $i$-th predictor $x_{it}$. Select those variables with $t_i$ larger than a threshold value at a given significance level and apply factor models to them. As we show in the empirical application of Section 4.6, the hard-threshold variable selection plays a critical role in forecasting in the sense that it can substantially reduce MSFE.

### 4.5.2 Soft-Threshold Variable Selection

Hard-threshold variable selection is highly likely to choose variables similar to each other (so called the "group effect"). In this sense, important information may be lost during the selection process. In contrast to the hard-thresholding which uses a single index to separate qualified predictors from others, soft-thresholding employs more flexible indices to select variables. There are several variable selection methods of this kind, see Tibshirani (1996), Efron et al (2004), and Zou and Hastie (2005) among many others.

In this chapter, we use the least angle regression or LARS of Efron et al (2004). LARS has gained its popularity in forecasting literature due to its comparative advantages. First, it gives relative ranking of predictors unlike hard-thresholding which gives the marginal predictive power of each predictor. Second, it avoids the group effect. Third, it is very fast and has the same order of computation complexity as OLS.

The LARS algorithm proceeds roughly as follows. Like classical forward selection we first find the predictor, say $x_{j_1}$ which is most correlated to the response $y$.

However, instead of taking the largest step in the direction of $x_{j_1}$ as in forward selection, we stop at the point where some other predictor, say $x_{j_2}$, has as much correlation with the current residual. Instead of continuing along $x_{j_1}$, LARS proceeds in a direction equiangular between the two predictors until a third variable $x_{j_3}$ makes its way into the "most correlated" set. LARS then proceeds equiangularly between $x_{j_1}$, $x_{j_2}$ and $x_{j_3}$, that is, along the "least angle direction," until a fourth variable enters, and so on. Readers interested in LARS are referred to Efron et al (2004) for detailed description of the algorithm and its satisfactory properties.

In the next section, we apply the LARS algorithm to first select 30 variables, as in Bai and Ng (2008), from the 131 predictors. Then we use the four factor methods, PCR, PLS, PCovR, CFPC, to the 30 variables in forecasting the monthly CPI inflation of U.S.

## 4.6 Empirical Applications

This section compares the methods described in the previous two sections. Variable of interest to forecast is the logarithm of PUNEW, i.e., CPI all items, using some or all of the 132 monthly time series predictors. Data used are available on Mark Watson's website: http//www.princeton.edu/mwatson. The data range from 1960:1 to 2003:12, with 528 monthly observations in total. These data are transformed by taking logs, first or second differences as suggested in Stock and Watson (2004). Following Stock and Watson (2002b), define

$$y_{t+h}^{h} = \frac{1200}{h} \cdot (y_{t+h} - y_t) - 1200 \cdot (y_t - y_{t-1}), \tag{4.16}$$

and

$$z_t = 1200 \cdot (y_t - y_{t-1}) - 1200 \cdot (y_{t-1} - y_{t-2}).$$

For $h = 1, 3, 6, 12, 18, 24, 30$ and 36, we form the factor-augmented forecast as, given information at time $t$,

$$\hat{y}_{t+h|t} = \hat{\alpha}_0 + \hat{\alpha}_1'(L)z_t + \hat{\beta}_1'(L)\hat{f}_t,$$

Here, $z_t$ is the set of lagged variables and $\hat{f}_t$ latent factors. The number of lags of $z_t$ and $\hat{f}_t$ are determined by the BIC with the maximum number of lags set to six when the sample size permits, and is reduced to four otherwise. Although we are forecasting the change in inflation, we will continue to refer to the forecasts as inflation forecasts.

As parameter instability is salient in economic time series, we employ two ways to tackle this difficulty in evaluating different forecasting schemes. First, note that for each time period $t$, the predictors are selected and the forecasting equation is re-estimated after new factors are estimated. We do not restrict the optimal predictors to be the same for every time period. Second, we consider 9 forecast subsamples: 1970:1-1979:12, 1980:1-1989:12, 1990:1-1999:12, 1970:1-1989:12, 1980:1-1999:12, 1970:1-1999:12, 1970:1-2003:12, 1980:1-2003:12, 1990:1-2003:12. For example, for subsample 1970:1-1979:12, the first $h$-step forecast of 1970:1 is based on estimation up to 1960:3-1970:1-$h$. The last forecast is for 1979:12, and it employs parameters estimated for the sample 1960:3-1979:12-$h$. That is, recursive scheme is used here, as in Bai and Ng (2008).

MSFE are used to examine the performance of different forecasting procedures. We denote RMSFE as the ratio of the MSFE for a given method relative to the MSFE of PCR model. Therefore, RMSFE less than one means that the specified method outperforms the PCR model in the forecasting practice considered.

*Tables 1-8 About Here*

Tables 1-8 report RMSFE for each of forecast horizons $h = 1, 3, 6, 12, 18, 24, 30, 36$. Column 1 lists the 9 out-of-sample forecasting subsamples. We report three panels of the RMSFE results depending on whether or how we conduct the variable selection prior to applying the factor models. The first panel of the results reported in Columns 2-5 is for factor models without variable selection, where we use the all 131 predictors to estimate the latent factors for PCR, CFPC, PLS, PCovR. Columns 6-9 and Columns 10-13 present RMSFE for the factor models after selecting the predictors. The second panel reported in Columns 6-9 uses the hard-threshold variable selection at 5% level with the critical value 1.65 for $t$ statistics. To keep in line with Bai and Ng (2008), the third panel of the results reported in Columns 10-13 uses the soft-thresholding variable selection via the LARS algorithm to select 30 variables. Note that PCR without variable selection is used as a benchmark (in each row) in computing the relative MSFEs and thus the values for PCR in Column 2 is 1.000 for all cases.

### 4.6.1 Supervision on Computation of Factors

One of the main objectives of this chapter is to examine the effect of supervision on the computation of latent factors. The main conclusion over this topic is summarized as below.

1. Although not reported in the table, the performance of AR(4) is generally as good as AR model with number of lags selected by BIC. However, the predictability of the univariate AR model decreases as forecasting horizon increases, reporting larger MSFE as horizon getting larger. This finding reasserts the need to use more information than just the target variable in economic time series forecasting.

2. CFPC is better than PCR, no matter variable selection is performed or not. Look-

ing at Column 3 for CFPC from Tables 1-8, for 62 out of 72 cases without variable selection, CFPC reports a RMSFE less than 1. In the case of hard threshold variable selection, 63 out of 72 cases favor CFPC (Column 7 for CFPC from Tables 1-8). Also, the LARS variable selection reports 64 out of 72 cases that are in favor of CFPC over PCR (Column 11 for CFPC from Tables 1-8).

3. PLS is not doing as well as one might expected. From Table 1, we can see that supervision on factor computation does not make PLS much better than PCR. And it is seen in Tables 1 and 2 that PLS could be very bad and unstable, reporting RMSFE larger than 2. However, as horizon increases, as in Table 6-8, PLS indeed improves over PCR a lot, reducing RMSFE even below 70%. Variable selection also improves the performance of PLS over PCR, as can be seen in the last two panels of Tables 1-8.

4. PCovR performs better than PCR most of the cases, with 64 out of 72 cases reporting RMSFE lower than 1 without variable selection. Its better predictability is also revealed after variable selection. For example, for $h = 36$, the subsample 90:1-99:12 reports RMSFE of PCovR as 0.187 while that of PCR is 0.834, with hard-thresholding variable selection.

By comparing the RMSFEs in each of the three panels from the tables, we conclude that, the supervision on the computation of factors does improve the predictability of the naive principal component. This improvement is quite substantial as noted above.

### 4.6.2 Supervision on Predictors

Next, let us take a look at the effects of variable selection on the predictability of factor models.

1. One notable observation from Tables 1-8 is that, variable selection does not make much difference for PCR, with RMSFE closely around 1 most of the cases. This finding is consistent with that reported in Stock and Watson (2002b).

2. Hard threshold variable selection can make CFPC even better. Most of the cases, hard threshold variable selection reports RMSFE smaller than that without variable selection. To the contrary, more than often, the soft-thresholding LARS variable selection worsens the predictive ability of CFPC.

3. PLS generally reports lower MSFE when variable selection is carried out in the first step. Hard threshold even makes PLS the best method for several cases. See Table 8 for the second and fourth subsamples for example.

4. For PCovR, the LARS variable selection makes it the best for several subsample when $h = 1$. For all other cases, hard threshold works better as a variable selection procedure to improve the performance of PCovR.

### 4.6.3 Effects of "Double" Supervision

It would be interesting to see that the above two parts on supervision leads to the essence of this chapter. The RMSFE reported for factor models after supervision on the computation of factors and also the selection of variable are generally lower than 1, as can be seen in the last two panels of Tables 1-8. Exception to this conclusion is for PLS with short forecasting horizons. In most of the cases, the reduction of MSFE relative to PCR is clearly noticeable. After variable selection, CFPC reports RMSFE as low as 40% in a lot of cases. PCovR can reduce RMSFE to be as low as 18.7%. The findings affirm the conjecture raised in Section 1 that the *double* supervision in selection of predictors and formation of latent factors should be carried out in forecasting practice.

### 4.6.4 Supervision and Forecasting Horizon

The effect of supervision over forecasting horizons $h$ is very clear, which can be seen by comparing the results across the eight tables. That can be visually presented by reporting the RMSFE numbers as a function of $h$. Figures 1-3 plot RMSFE over the eight values of forecast horizons $h = 1, 3, 6, 12, 18, 24, 30$ and 36. Figure 1 presents the RMSFEs without variable selections for each of nine out-of-sample forecasting subsamples. Figures 2-3 do the same with hard-thresholding variable selection and soft-thresholding variable selection, respectively. One salient feature of these figures is that the lines connecting the RMSFEs over $h$ are generally downward sloped for the three supervised factor models. That is, the superiority of supervised factor models is getting more and more significant as the forecasting horizon increases. On the other hand, the unsupervised factor model, PCR, has RMSFEs in Figures 2-3 moving up and down over the horizons with no slope pattern over forecasting horizon $h$.

*Figures 1-3 About Here*

Note that the forecast target variable $y_{t+h}^h$ defined in (4.16) is the average monthly changes over the $h$ months, and it may be easier to forecast when forecasting horizon $h$ is longer as it becomes smoother. The three supervised factor models are able to capture this feature in $y_{t+h}^h$ while PCR fails to do so. We also observe (although not reported for space) that neither AR(4) or AR models with number of lags selected by BIC capture this feature. This is seen from the RMSFE values for these univariate models, which are generally increasing over the forecasting horizons. Hence, it seems that richer information from multivariate environment benefits the factor models even

more especially for longer forecast horizons when they are supervised on the selection of the variables and on the computation of their latent factors.

### 4.6.5 Supervision and Number of Factors

Another important finding of this chapter (not reported) is that, for supervised factor models, the number of factors selected by BIC is less than that of PCR. This finding also favors the previous result that, with supervision, factor models tend to form better latent variables and thus need less indices to describe "the state of the economy", as termed in Heij, Groenen and van Dijk (2007). They report the result for PCovR and this chapter validates their conclusion for PLS and CFPC.

## 4.7 Conclusions

In exploiting high dimensional information from large number of predictors we wish to improve efficiency of a forecast and to enhance the robustness of a forecast. This chapter compares the forecasting performance of factor models in such data-rich environment. Our findings suggest that one can profit from supervising the computation of factors. Computation of latent factors may be doubly supervised via variable selection. Variable selection is generally useful for the supervised factor models. Interestingly, the effect of supervision gets even larger as forecast horizon increases and the supervision also helps a factor model achieving more parsimonious factor structure. Among the supervised factor models compared in this chapter, CFPC stands out for its superiority in predictive ability and its stability in performance. In general, the CFPC model generates most efficient and robust forecasts.

## Appendix A: NIPALS Algorithm for PCR

The intuition behind the working of the nonlinear iterative algorithm for PCR goes as follows. Formally,

$$\begin{aligned} E_1 &= X - f_1 p_1', \qquad E_2 = E_1 - f_2 p_2', \qquad \dots \\ E_h &= E_{h-1} - f_h p_h', \qquad \dots \qquad E_r = E_{r-1} - f_r p_r'. \end{aligned} \tag{4.17}$$

The NIPALS follows the steps for the computation of $f_h$:

1. take a vector $x_J$ from $X$ and call it $f_h$:
2. normalize $f_h$: $f_h' = f_h' / \parallel f_h' \parallel$
3. calculate $p_h'$:

$$p_h' = f_h' X \tag{4.18}$$

4. normalize $p_h' : p_h' = p_h' / \parallel p_h' \parallel$
5. calculate $f_h$:

$$f_h = X p_h \tag{4.19}$$

6. compare $f_h$ in step 2 with that obtained in step 5. If they are the same, stop. Otherwise go to step 2.

Note that the evolution of $p_h'$ and $f_h$ are described by (4.18) and (4.19). Substitute (4.19) into (4.18), we have

$$c p_h' = (X p_h)' X, \tag{4.20}$$

where $c$ is a constant that accounts for the normalization in step 4. This is equivalent to

$$0 = \left(X'X - cI_r\right) p_h. \tag{4.21}$$

This is exactly the eigenvalue/eigenvector equation for $X'X$ in PCR. Hence, the NIPALS algorithm gives the same principal components as derived by eigenvalue decomposition.

## Appendix B: NIPALS Algorithm for PLS

For the $X$ block: (1) take $u_{\text{start}} =$ some $y_J$ (instead of some $x_J$); (2) normalize $u$: $u = u/ \parallel u \parallel$; (3) $p' = u'X$; (4) normalize $p'$: $p' = p'/ \parallel p' \parallel$; (5) $f = Xp$.

For the $y$ block: (6) $q = f$ (instead of some $y_S$); (7) normalize $q$: $q = q/ \parallel q \parallel$; (8) $u' = y'q$; (9) normalize $u'$: $u' = u'/ \parallel u' \parallel$; (10) compare $f$ in step 5 with that in the preceding iteration step. If they are equal (up to a tolerance level) then stop; otherwise go to step 2.

By exchanging scores in step 1 and 6, the above algorithm supervises the computation of the $x$-score thus should improve the predictability of PLS over PCR. For the purpose of prediction, we can rewrite (4.8) as

$$\begin{aligned} E_h &= E_{h-1} - f_h p_h'; \qquad X = E_0, \\ G_h &= G_{h-1} - u_h q_h'; \qquad y = G_0, \end{aligned}$$

and a mixed relation is available as

$$G_h = G_{h-1} - b_h f_h q_h',$$

where $b_h = u_h' f_h / (f_h' f_h)$. Therefore,

$$\hat{y} = \sum \hat{u}_h q_h' = \sum b_h f_h q_h' = F B_0 Q', \tag{4.22}$$

where $B_0 = \text{diag}(b_1, \ldots, b_r)$.

Note that the $x$-score extracted in the $h^{\text{th}}$ iteration, $f_h$, is a linear combination of $E_{h-1}$, instead of as a direct function of original data matrix $X$. de Jong (1993) gives a direct relationship as $F = XR \equiv XW(P'W)^{-1}$, where $P = [p_1, \ldots, p_h]$ and $W = [E_0 u_1, \ldots, E_{h-1} u_h]$. Thus, (4.22) can be used for prediction as

$$\hat{y}_{\text{PLS}} = XRB_0Q'. \tag{4.23}$$

That is, we have $R = W\left(P'W\right)^{-1}$, $U = FB_0$ for the linear factor model framework, while $\beta$ in (4.1) is estimated by $b = RBQ'$.

## Appendix C: Proofs of Propositions

For matrix decomposition used later for proof, we adopt the following convention: for a $T \times N$ matrix $C$, it is decomposed in to two blocks $C_1$ and $C_2$, with $C_1$ containing its first $r$ columns $c_1, \ldots, c_r$ and $C_2$ containing the rest. That is, $C \equiv [C_1, C_2]$, where $C_1 = [c_1, \ldots, c_r]$ and $C_2 = [c_{r+1}, \ldots, c_N]$.

**Proof of Proposition 1:** The eigenvalue decomposition of $X'X$ is

$$X'X = R\Lambda R' = R_1\Lambda_1 R_1' + R_2\Lambda_2 R_2', \tag{4.24}$$

where $\Lambda = \text{diag}(\Lambda_1, \Lambda_2)$ is the eigenvalue matrix and $R = [R_1, R_2]$ is the eigenvector matrix corresponding to $\Lambda$. As $R$ is orthonormal with $R'R = I$,

$$R_1'X'XR_1 = \Lambda_1. \tag{4.25}$$

Stock and Watson (2002ab) has shown that the true factors can be consistently estimated by the first $r$ principal components of $X$. Therefore, we adopt that $\hat{F} = XR_1$. With $\hat{U} = \hat{F}$, the OLS estimator of the coefficient $Q$, $r \times 1$ vector, is given as

$$\begin{aligned} \hat{Q} &= \left(\hat{F}'\hat{F}\right)^{-1}\hat{F}'y \\ &= \left(R_1'X'XR_1\right)^{-1}R_1'X'y \\ &= \Lambda_1^{-1}R_1'X'y. \end{aligned} \tag{4.26}$$

Therefore, PCR forecast is formed as

$$\hat{y}_{\text{PCR}} = \hat{F}\hat{Q} = XR_1\Lambda_1^{-1}R_1'X'y.$$

□

**Proof of Proposition 2:** Consider a linear regression model for each $i = 1, 2, \ldots, N$,

$$y = x_i b_i + u_i, \tag{4.27}$$

where $b_i$ is estimated by

$$b_i = \left(x_i' x_i\right)^{-1} x_i' y. \tag{4.28}$$

Thus the prediction could be formed as

$$\hat{y}_i \equiv x_i b_i. \tag{4.29}$$

To write (4.29) in compact form,

$$\hat{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N] = [x_1 b_1, x_2 b_2, \ldots, x_N b_N] \equiv XB, \tag{4.30}$$

where $B = \text{diag}(b) = \text{diag}(b_1 \ldots b_N)$ is the diagonal matrix with $b_1, b_2, \ldots, b_N$ sitting on the diagonal. Parallel to (4.24), we also have its eigenvalue decomposition of $\hat{Y}'\hat{Y}$ as follows,

$$\hat{Y}'\hat{Y} = L\Omega L' = L_1'\Omega_1 L_1 + L_2\Omega_2 L_2'.$$

The principal component estimator of $F$ which is the first $r$ principal components of $\hat{Y}$, is therefore given as $\hat{F} = \hat{Y}L_1 = XBL_1$. Then consider the following regression,

$$y = \hat{F}Q + \varepsilon = \hat{Y}L_1 Q + \varepsilon. \tag{4.31}$$

The OLS estimation of the coefficient $Q$, $r \times 1$ vector, in (4.31) is given as

$$\hat{Q} = \left(\hat{F}'\hat{F}\right)^{-1} \hat{F}' y = \left(L_1' B X' X B L_1\right)^{-1} L_1' B X' y = \Omega_1^{-1} L_1' B X' y. \tag{4.32}$$

Therefore, a forecast can be formed as

$$\hat{y}_{\text{CFPC}} = \hat{Y}L_1\hat{Q} = XBL_1\Omega_1^{-1}L_1'BX'y. \tag{4.33}$$

□

**Proof of Proposition 3:** Rewrite (4.3) as

$$XB = C + E, \tag{4.34}$$

where $C = FP'$, is the common component of $XB$. Note that $C$ is estimated using principal component method as

$$\begin{aligned}\tilde{C} &= \hat{F}\hat{P}' \\ &= XBL_1(L_1'BX'XBL_1)^{-1}L_1'BX'XB \\ &= XBL_1\Omega_1^{-1}L_1'BX'XB.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\tilde{C}'\tilde{C} &= BX'XBL_1\Omega_1^{-1}L_1'BX'XBL_1\Omega_1^{-1}L_1'BX'XB \\ &= BX'XBL_1\Omega_1^{-1}L_1'BX'XB,\end{aligned}$$

which leads to

$$\begin{aligned}J &\equiv B\left(X'X/T\right)BL_1\Omega_1^{-1}L_1'BX'XB - (B\Sigma_X B - \Sigma_E) \qquad (4.35)\\ &= B\left(X'X/T\right)BL_1\Omega_1^{-1}L_1'BX'XB - (\Sigma_{XB} - \Sigma_E) \\ &= B\left(X'X/T\right)BL_1\Omega_1^{-1}L_1'BX'XB - \Sigma_C \\ &= \left(\tilde{C}'\tilde{C}/T\right) - \Sigma_C \\ &= \frac{1}{T}\left(\tilde{C}'\tilde{C} - \tilde{C}'C + \tilde{C}'C - C'C\right) + \left(\frac{1}{T}C'C - \Sigma_C\right) \\ &= \frac{1}{T}\tilde{C}'\left(\tilde{C} - C\right) + \frac{1}{T}\left(\tilde{C} - C\right)'C + \left(\frac{1}{T}C'C - \Sigma_C\right) \\ &\equiv \psi^1 + \psi^2 + \psi^3.\end{aligned}$$

Note that

$$\psi_{ij}^1 = \frac{1}{T}\sum_{t=1}^{T}\tilde{C}_{ti}\left(\tilde{C}_{tj} - C_{tj}\right) = O_p\left(\frac{1}{\sqrt{NT}}\right). \qquad (4.36)$$

which follows from Theorem 3 of Bai (2003) that $\tilde{C}_{it} - C_{it} = O_p\left(1/\sqrt{N}\right)$ under Assumption 1.d. Similarly, we have

$$\psi_{ij}^2 = O_p\left(\frac{1}{\sqrt{NT}}\right). \qquad (4.37)$$

By assumption 1.a, we have

$$\psi_{ij}^3 = O_p\left(\frac{1}{\sqrt{T}}\right). \tag{4.38}$$

(4.36), (4.37) and (4.38) lead to

$$J_{ij} = O_p\left(\frac{1}{\sqrt{NT}}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right) + O_p\left(\frac{1}{\sqrt{T}}\right) = O_p\left(\frac{1}{\sqrt{T}}\right).$$

Note that (4.35) is equivalent to

$$\begin{aligned} J &= B\Sigma_X[BL_1\Omega_1^{-1}L_1'BX'X - I_N - \\ &\quad (\Sigma_X + \epsilon_{NT}I_N)^{-1}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}]B + o_p(1) \\ &\equiv B\Sigma_X HB + o_p(1) \end{aligned} \tag{4.39}$$

where

$$H = \left(BL_1\Omega_1^{-1}L_1'BX'X - I_N - (\Sigma_X + \epsilon_{NT}I_N)^{-1}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}\right)$$

where $\epsilon_{NT}, \zeta_{NT}$ are sequences of small positive numbers such that $\epsilon_{NT}, \zeta_{NT} \to 0$ as $N, T \to \infty$. $\epsilon_{NT}$ and $\zeta_{NT}$ are introduced to guarantee the matrix inverse exists. To see

(4.39), note that the last term in the bracket of the right hand

$$\begin{aligned}
& B\Sigma_X(\Sigma_X+\epsilon_{NT}I_N)^{-1}(B+\zeta_{NT}I_N)^{-1}\Sigma_E(B+\zeta_{NT}I_N)^{-1}B \\
= \; & B(\Sigma_X+\epsilon_{NT}I_N)(\Sigma_X+\epsilon_{NT}I_N)^{-1}(B+\zeta_{NT}I_N)^{-1}\Sigma_E(B+\zeta_{NT}I_N)^{-1}B \\
& -B\epsilon_{NT}(\Sigma_X+\epsilon_{NT}I_N)^{-1}(B+\zeta_{NT}I_N)^{-1}\Sigma_E(B+\zeta_{NT}I_N)^{-1}B \\
= \; & B(B+\zeta_{NT}I_N)^{-1}\Sigma_E(B+\zeta_{NT}I_N)^{-1}B+o_p(1) \\
= \; & (B+\zeta_{NT}I_N)(B+\zeta_{NT}I_N)^{-1}\Sigma_E(B+\zeta_{NT}I_N)^{-1}B \\
& -\zeta_{NT}(B+\zeta_{NT}I_N)^{-1}\Sigma_E(B+\zeta_{NT}I_N)^{-1}B+o_p(1) \\
= \; & \Sigma_E(B+\zeta_{NT}I_N)^{-1}B+o_p(1) \\
= \; & \Sigma_E(B+\zeta_{NT}I_N)^{-1}(B+\zeta_{NT}I_N) \\
& -\Sigma_E(B+\zeta_{NT}I_N)^{-1}\zeta_{NT}I_N+o_p(1) \\
= \; & \Sigma_E+o_p(1) \qquad\qquad \text{as } \epsilon_{NT}=o_p(1) \text{ as } N,T\to\infty.
\end{aligned}$$

Note that (4.39) is true for all values of $\beta$. Therefore, it must be the case that

$$H_{ij}=O(J_{ij})=O_p\left(\frac{1}{\sqrt{T}}\right).$$

Define

$$K=BL_1\Omega_1^{-1}L_1'BX'X-I_N.$$

We have

$$\begin{aligned}
K &= H+(\Sigma_X+\epsilon_{NT}I_N)^{-1}(B+\zeta_{NT}I_N)^{-1}\Sigma_E(B+\zeta_{NT}I_N)^{-1} \\
&= H+\Sigma_{XB}^{-1}\Sigma_E+o_p(1),
\end{aligned}$$

i.e.,

$$\begin{aligned}
K_{ij} &= H_{ij}+O_p(N/T) \quad \text{(by Assumption 1.e)} \\
&= O_p\left(\max\left\{\frac{1}{\sqrt{T}},\frac{N}{T}\right\}\right).
\end{aligned} \tag{4.40}$$

By definition of $f(\beta)$,

$$\begin{aligned} f(\beta)-\beta &= \operatorname{diag}(\beta) L_1 \Omega_1^{-1} L_1' \operatorname{diag}(\beta) X'y - \beta \\ &= BL_1\Omega_1^{-1}L_1'BX'y - \beta \\ &= BL_1\Omega_1^{-1}L_1'BX'(X\beta+e) - \beta \\ &= \left(BL_1\Omega_1^{-1}L_1'BX'X - I_N\right)\beta + o_p(1) \qquad \text{(by Assumption 1.b)} \\ &= K\beta + o_p(1), \end{aligned}$$

and it follows from (4.40) that

$$(f(\beta)-\beta)_i = O_p\left(\max\left\{\frac{N}{\sqrt{T}}, \frac{N^2}{T}\right\}\right). \tag{4.41}$$

□

## References


Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica* 71(1), 135-171.

Bai, J. and Ng, S. (2008), "Forecasting Economic Time Series Using Targeted Predictors", *Journal of Econometrics* 146, 304-317.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006), "Prediction by Supervised Principal Components," *Journal of the American Statistical Association* 101:473, 119–137.

Boivin, J. and Ng, S. (2006), "Are More Data Always Better for Factor Analysis," *Journal of Econometrics* 132, 169–194.

Chan, Y.L., Stock, J.H., and Watson, M.W. (1999), "A Dynamic Factor Model Framework for Forecast Combination," *Spanish Economic Review* 1, 91-121.

de Jong, Sijmen (1992), "Principal Covariate Regression Part I. Theory," *Chemometrics and Intelligent Laboratory Systems*, 14, 155-164.

de Jong, Sijmen (1993), "SIMPLS: An Alternative Approach to Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, 18, 251-261.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics* 32:2, 407–499.

Fan, J. and R. Li. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties", *Journal of the American Statistical Association* 96, 1348-1360.

Garthwait, P.H. (1994), "An Interpretation of Partial Least Squares," *Journal of the American Statistical Association* Vol. 89, No. 425 pp.122-127.

Geladi, P. and Kowalski, B.R. (1986), "Partial Least-squares Regression: A Tutorial," *Analytica Chimica Acta* 185, 1-17.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning – data mining, inference, and prediction*, Second edition, Springer.

Heij, C., Groenen, P.J.F., and van Dijk, D. (2007), "Forecast Comparison of Principal Component Regression and Principal Covariate Regression," *Computational Statistics and Data Analysis* 51, 3612-3625.

Huang, J., Horowitz, J.L. and Ma, S. (2008), "Asymptotic Properties of Bridge Estimators in Sparse High-dimensional Regression Models", *Annals of Statistics* 36, 587-613.

Huang, H. and Lee, T.H. (2010), "To Combine Forecasts or To Combine Information?" *Econometric Reviews* 29, 534-570.

Mardia, K., Kent, J. and Bibby, J. (1980), *Multivariate Analysis*, Academic Press, London.

Miller, A. (2002), *Subset Selection in Regression*, Chapman & Hall/CRC.

Otto, M. and Wegscheider, W. (1985), "Spectrophotometric Multicomponent Applied to Trace Metal Determinations," *Analytical Chemistry* 57, 63-69.

Stock, J.H. and Watson, M.W. (2002a), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association* 97, 1167-1179.

Stock, J.H., and Watson, M.W. (2002b), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business Economical Statistics* 20,147-162.

Stock, J.H. and Watson, M.W. (2004), "Combination Forecasts of Output Growth in a Seven-Country Data Set", *Journal of Forecasting* 23, 405-430.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of Royal Statistical Society Series B* 58:1, 267–288.

Tu, Y. (2011), "Model Averaging PartiaL Effect (MAPLE) Estimation with Large Dimensional Data", UC Riverside.

Wold, H. (1966), "Estimation of Principal Components and Related Models by Iterative Least Squares," in *Multivariate Analysis,* Krishnaiaah, P.R. (ed.) (pp. 391-420). New York: Academic Press.

Wold, H. (1975), "Soft Modelling by Latent Variables: the non-linear iterative partial least squares approach," in *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett,* J. Gani (ed.). Academic Press, London.

Wold, S., Ruhe, A., Wold, H., and Dunn III, W.J. (1984), "The Collinearity Problem in Linear Regression, The Partial Least Squares Approach to Generalized Inverse," *SIAM Journal of Scientific and Statistical Computing* 5, 735-743

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties", *Journal of the American Statistical Association* 101(476), 1418-1429.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of Royal Statistical Society Series B* 67:2, 301–320.

Zou, H. and Zhang, H. H. (2009), "On the Adaptive Elastic-Net with a Diverging Number of Parameters", *The Annals of Statistics* 37(4):1773-1751

Table 4.1: RMSFE, h=1

| | NO VARIABLE SELECTION | | | | HARD THRESHOLD VARIABLE SELECTION | | | | LARS(30) VARIABLE SELECTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR |
| 70.1-79.12 | 1.000 | 1.138 | 2.057 | 1.024 | 1.085 | 1.018 | 1.167 | 1.093 | 1.005 | 1.078 | 1.060 | **0.985** |
| 80.1-89.12 | 1.000 | **0.863** | 1.109 | 0.931 | 1.046 | 0.922 | 1.060 | 0.969 | 1.001 | 0.927 | 0.911 | 0.926 |
| 90.1-99.12 | 1.000 | 0.994 | 1.291 | 1.003 | 1.054 | **0.949** | 1.079 | 0.985 | 1.008 | 0.986 | 1.040 | 0.993 |
| 70.1-89.12 | 1.000 | 0.988 | 1.538 | 0.973 | 1.064 | 0.966 | 1.108 | 1.025 | 1.003 | 0.996 | 0.978 | **0.953** |
| 80.1-99.12 | 1.000 | **0.898** | 1.158 | 0.951 | 1.048 | 0.929 | 1.065 | 0.973 | 1.003 | 0.943 | 0.946 | 0.944 |
| 70.1-99.12 | 1.000 | 0.989 | 1.497 | 0.978 | 1.062 | 0.963 | 1.103 | 1.018 | 1.004 | 0.994 | 0.989 | **0.960** |
| 70.1-03.12 | 1.000 | 1.007 | 1.497 | 0.979 | 1.066 | 0.985 | 1.136 | 1.018 | 1.003 | 0.998 | 0.992 | **0.964** |
| 80.1-03.12 | 1.000 | **0.943** | 1.224 | 0.957 | 1.056 | 0.969 | 1.121 | 0.981 | 1.002 | 0.959 | 0.959 | 0.954 |
| 90.1-03.12 | 1.000 | 1.057 | 1.389 | **0.993** | 1.071 | 1.035 | 1.209 | 0.999 | 1.004 | 1.004 | 1.028 | 0.994 |

Table 4.2: RMSFE, h=3

| | NO VARIABLE SELECTION | | | | HARD THRESHOLD VARIABLE SELECTION | | | | LARS(30) VARIABLE SELECTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR |
| 70.1-79.12 | 1.000 | 1.016 | 1.492 | **0.900** | 1.002 | 1.000 | 1.027 | 0.945 | 1.012 | 0.943 | 0.960 | 0.907 |
| 80.1-89.12 | 1.000 | 0.904 | 0.984 | 0.886 | 0.973 | **0.871** | 0.921 | 0.904 | **0.990** | 0.892 | 0.924 | 0.925 |
| 90.1-99.12 | 1.000 | 1.055 | 1.735 | 1.072 | 1.126 | 1.044 | 1.321 | 1.066 | 0.990 | 1.010 | 0.991 | 1.058 |
| 70.1-89.12 | 1.000 | 0.946 | 1.176 | **0.891** | 0.984 | 0.920 | 0.961 | 0.919 | 0.998 | 0.911 | 0.937 | 0.919 |
| 80.1-99.12 | 1.000 | 0.934 | 1.132 | 0.923 | 1.003 | **0.905** | 1.000 | 0.936 | 0.990 | 0.915 | 0.937 | 0.952 |
| 70.1-99.12 | 1.000 | 0.961 | 1.251 | **0.915** | 1.003 | 0.937 | 1.009 | 0.939 | 0.997 | 0.924 | 0.944 | 0.937 |
| 70.1-03.12 | 1.000 | 0.973 | 1.256 | **0.928** | 1.003 | 0.953 | 1.021 | 0.944 | 0.997 | 0.931 | 0.950 | 0.944 |
| 80.1-03.12 | 1.000 | 0.955 | 1.156 | 0.940 | 1.003 | 0.933 | 1.019 | 0.943 | 0.991 | **0.926** | 0.946 | 0.959 |
| 90.1-03.12 | 1.000 | 1.070 | 1.541 | 1.060 | 1.072 | 1.072 | 1.238 | 1.032 | **0.993** | 1.003 | 0.997 | 1.035 |

Table 4.3: RMSFE, h=6

| | NO VARIABLE SELECTION | | | | HARD THRESHOLD VARIABLE SELECTION | | | | LARS(30) VARIABLE SELECTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR |
| 70.1-79.12 | 1.000 | 0.902 | 1.267 | 0.908 | 0.922 | 0.824 | **0.806** | 0.847 | 0.983 | 0.842 | 0.824 | 0.829 |
| 80.1-89.12 | 1.000 | 0.895 | 1.119 | 0.937 | 0.971 | **0.842** | 0.929 | 0.886 | 1.001 | 1.005 | 1.056 | 0.953 |
| 90.1-99.12 | 1.000 | 0.964 | 1.656 | 0.996 | 1.068 | 0.975 | 1.374 | **0.955** | 0.985 | 0.971 | 1.027 | 1.006 |
| 70.1-89.12 | 1.000 | 0.898 | 1.180 | 0.925 | 0.951 | **0.835** | 0.878 | 0.870 | 0.994 | 0.938 | 0.960 | 0.902 |
| 80.1-99.12 | 1.000 | 0.910 | 1.229 | 0.949 | 0.991 | **0.870** | 1.021 | 0.900 | 0.998 | 0.998 | 1.050 | 0.964 |
| 70.1-99.12 | 1.000 | 0.907 | 1.243 | 0.935 | 0.966 | **0.853** | 0.944 | 0.881 | 0.993 | 0.942 | 0.969 | 0.916 |
| 70.1-03.12 | 1.000 | 0.927 | 1.261 | 0.959 | 0.972 | **0.874** | 0.977 | 0.893 | 0.993 | 0.954 | 0.984 | 0.932 |
| 80.1-03.12 | 1.000 | 0.939 | 1.258 | 0.984 | 0.998 | **0.899** | 1.062 | 0.915 | 0.998 | 1.010 | 1.064 | 0.983 |
| 90.1-03.12 | 1.000 | 1.045 | 1.600 | 1.098 | 1.063 | 1.036 | 1.387 | **0.986** | 0.989 | 1.023 | 1.085 | 1.055 |

Table 4.4: RMSFE, h=12

| | NO VARIABLE SELECTION | | | | HARD THRESHOLD VARIABLE SELECTION | | | | LARS(30) VARIABLE SELECTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR |
| 70.1-79.12 | 1.000 | 0.762 | 1.281 | 0.807 | 0.960 | 0.738 | 0.886 | 0.734 | 1.014 | **0.716** | 0.757 | 0.727 |
| 80.1-89.12 | 1.000 | 0.797 | 1.008 | 0.843 | 1.042 | **0.787** | 0.845 | 0.868 | 1.005 | 0.967 | 1.114 | 0.867 |
| 90.1-99.12 | 1.000 | **0.891** | 1.415 | 0.980 | 1.101 | 0.906 | 1.381 | 0.924 | 0.976 | 0.936 | 0.979 | 1.015 |
| 70.1-89.12 | 1.000 | 0.782 | 1.128 | 0.827 | 1.006 | **0.765** | 0.863 | 0.808 | 1.009 | 0.856 | 0.956 | 0.805 |
| 80.1-99.12 | 1.000 | 0.817 | 1.095 | 0.872 | 1.054 | **0.812** | 0.959 | 0.880 | 0.998 | 0.961 | 1.085 | 0.899 |
| 70.1-99.12 | 1.000 | 0.796 | 1.166 | 0.847 | 1.018 | **0.784** | 0.931 | 0.824 | 1.004 | 0.867 | 0.959 | 0.833 |
| 70.1-03.12 | 1.000 | 0.822 | 1.205 | 0.883 | 1.034 | **0.811** | 0.980 | 0.855 | 1.004 | 0.883 | 0.979 | 0.859 |
| 80.1-03.12 | 1.000 | 0.857 | 1.162 | 0.926 | 1.076 | **0.854** | 1.033 | 0.924 | 0.998 | 0.978 | 1.106 | 0.934 |
| 90.1-03.12 | 1.000 | 1.013 | 1.567 | 1.143 | 1.167 | 1.029 | 1.528 | 1.072 | **0.979** | 1.007 | 1.084 | 1.111 |

Table 4.5: RMSFE, h=18

| | NO VARIABLE SELECTION | | | | HARD THRESHOLD VARIABLE SELECTION | | | | LARS(30) VARIABLE SELECTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR |
| 70.1-79.12 | 1.000 | **0.596** | 1.083 | 0.673 | 0.895 | 0.629 | 0.793 | 0.656 | 1.021 | 0.630 | 0.633 | 0.688 |
| 80.1-89.12 | 1.000 | 0.707 | 0.878 | 0.701 | 1.026 | **0.686** | 0.765 | 0.701 | 1.028 | 0.760 | 0.895 | 0.710 |
| 90.1-99.12 | 1.000 | 1.084 | 1.527 | 1.012 | 1.313 | 1.055 | 1.509 | 0.991 | 0.974 | **0.907** | 1.007 | 1.047 |
| 70.1-89.12 | 1.000 | **0.657** | 0.971 | 0.688 | 0.966 | 0.660 | 0.778 | 0.681 | 1.025 | 0.701 | 0.776 | 0.700 |
| 80.1-99.12 | 1.000 | 0.767 | 0.981 | 0.750 | 1.071 | **0.744** | 0.883 | 0.747 | 1.019 | 0.783 | 0.913 | 0.763 |
| 70.1-99.12 | 1.000 | **0.696** | 1.023 | 0.718 | 0.998 | 0.697 | 0.846 | 0.710 | 1.020 | 0.720 | 0.797 | 0.732 |
| 70.1-03.12 | 1.000 | **0.722** | 1.056 | 0.750 | 1.023 | **0.722** | 0.886 | 0.735 | 1.018 | 0.746 | 0.820 | 0.765 |
| 80.1-03.12 | 1.000 | 0.804 | 1.039 | 0.800 | 1.106 | **0.783** | 0.947 | 0.786 | 1.016 | 0.821 | 0.942 | 0.816 |
| 90.1-03.12 | 1.000 | 1.145 | 1.605 | 1.147 | 1.389 | 1.126 | 1.588 | 1.086 | **0.973** | 1.038 | 1.108 | 1.187 |

Table 4.6: RMSFE, h=24

| | NO VARIABLE SELECTION | | | | HARD THRESHOLD VARIABLE SELECTION | | | | LARS(30) VARIABLE SELECTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR |
| 70.1-79.12 | 1.000 | **0.524** | 0.951 | 0.625 | 0.890 | 0.561 | 0.632 | 0.610 | 0.975 | 0.623 | 0.597 | 0.672 |
| 80.1-89.12 | 1.000 | 0.794 | 0.867 | 0.834 | 1.078 | 0.773 | **0.718** | 0.837 | 1.014 | 0.849 | 0.850 | 0.849 |
| 90.1-99.12 | 1.000 | 0.450 | 0.881 | 0.347 | 0.868 | 0.444 | 0.775 | 0.324 | 0.957 | 0.346 | 0.933 | **0.316** |
| 70.1-89.12 | 1.000 | **0.667** | 0.907 | 0.735 | 0.989 | 0.673 | 0.678 | 0.730 | 0.995 | 0.743 | 0.731 | 0.766 |
| 80.1-99.12 | 1.000 | 0.684 | 0.871 | 0.678 | 1.011 | **0.668** | 0.736 | 0.674 | 0.996 | 0.689 | 0.876 | 0.679 |
| 70.1-99.12 | 1.000 | **0.624** | 0.901 | 0.658 | 0.965 | 0.628 | 0.697 | 0.650 | 0.988 | 0.664 | 0.771 | 0.676 |
| 70.1-03.12 | 1.000 | **0.644** | 0.927 | 0.675 | 0.973 | 0.647 | 0.730 | 0.668 | 0.991 | 0.687 | 0.784 | 0.700 |
| 80.1-03.12 | 1.000 | 0.711 | 0.914 | 0.704 | 1.019 | **0.695** | 0.785 | 0.700 | 1.000 | 0.723 | 0.889 | 0.716 |
| 90.1-03.12 | 1.000 | 0.569 | 0.994 | 0.482 | 0.919 | 0.563 | 0.898 | **0.466** | 0.976 | 0.507 | 0.957 | 0.488 |

Table 4.7: RMSFE, h=24

| | NO VARIABLE SELECTION | | | | HARD THRESHOLD VARIABLE SELECTION | | | | LARS(30) VARIABLE SELECTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR |
| 70.1-79.12 | 1.000 | **0.513** | 0.934 | 0.672 | 0.911 | 0.539 | 0.619 | 0.579 | 0.986 | 0.612 | 0.592 | 0.625 |
| 80.1-89.12 | 1.000 | 0.761 | 0.809 | 0.844 | 1.029 | 0.760 | **0.704** | 0.832 | 1.002 | 0.842 | 0.862 | 0.822 |
| 90.1-99.12 | 1.000 | 0.343 | 0.765 | 0.285 | 0.831 | 0.343 | 0.654 | **0.240** | 1.020 | 0.308 | 0.707 | 0.290 |
| 70.1-89.12 | 1.000 | **0.658** | 0.861 | 0.773 | 0.980 | 0.668 | 0.669 | 0.727 | 0.995 | 0.746 | 0.750 | 0.740 |
| 80.1-99.12 | 1.000 | 0.620 | 0.794 | 0.656 | 0.962 | **0.619** | 0.687 | 0.632 | 1.008 | 0.662 | 0.810 | 0.643 |
| 70.1-99.12 | 1.000 | **0.586** | 0.839 | 0.661 | 0.946 | 0.593 | 0.666 | 0.615 | 1.001 | 0.646 | 0.740 | 0.637 |
| 70.1-03.12 | 1.000 | **0.600** | 0.864 | 0.673 | 0.953 | 0.607 | 0.684 | 0.629 | 1.001 | 0.668 | 0.753 | 0.663 |
| 80.1-03.12 | 1.000 | 0.639 | 0.833 | 0.673 | 0.971 | **0.637** | 0.713 | 0.651 | 1.008 | 0.693 | 0.824 | 0.680 |
| 90.1-03.12 | 1.000 | 0.435 | 0.874 | 0.389 | 0.875 | 0.434 | 0.727 | **0.351** | 1.018 | 0.445 | 0.761 | 0.444 |

Table 4.8: RMSFE, h=24

| | NO VARIABLE SELECTION | | | | HARD THRESHOLD VARIABLE SELECTION | | | | LARS(30) VARIABLE SELECTION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR | PCR | CFPC | PLS | PCovR |
| 70.1-79.12 | 1.000 | 0.546 | 0.847 | 0.715 | 0.938 | **0.501** | 0.505 | 0.616 | 1.039 | 0.668 | 0.594 | 0.667 |
| 80.1-89.12 | 1.000 | 0.713 | 0.781 | 0.834 | 1.049 | 0.692 | **0.634** | 0.830 | 0.995 | 0.782 | 0.845 | 0.819 |
| 90.1-99.12 | 1.000 | 0.280 | 0.612 | 0.238 | 0.834 | 0.279 | 0.523 | **0.187** | 1.021 | 0.261 | 0.506 | 0.287 |
| 70.1-89.12 | 1.000 | 0.643 | 0.808 | 0.784 | 1.003 | 0.613 | **0.580** | 0.741 | 1.013 | 0.735 | 0.740 | 0.756 |
| 80.1-99.12 | 1.000 | 0.547 | 0.716 | 0.605 | 0.967 | **0.534** | 0.591 | 0.583 | 1.005 | 0.582 | 0.715 | 0.614 |
| 70.1-99.12 | 1.000 | 0.546 | 0.756 | 0.639 | 0.958 | **0.524** | 0.565 | 0.593 | 1.015 | 0.608 | 0.678 | 0.631 |
| 70.1-03.12 | 1.000 | 0.561 | 0.768 | 0.650 | 0.962 | **0.539** | 0.581 | 0.605 | 1.015 | 0.628 | 0.690 | 0.645 |
| 80.1-03.12 | 1.000 | 0.567 | 0.735 | 0.622 | 0.972 | **0.555** | 0.613 | 0.600 | 1.005 | 0.610 | 0.731 | 0.636 |
| 90.1-03.12 | 1.000 | 0.361 | 0.670 | 0.324 | 0.863 | 0.360 | 0.583 | **0.275** | 1.019 | 0.368 | 0.570 | 0.377 |

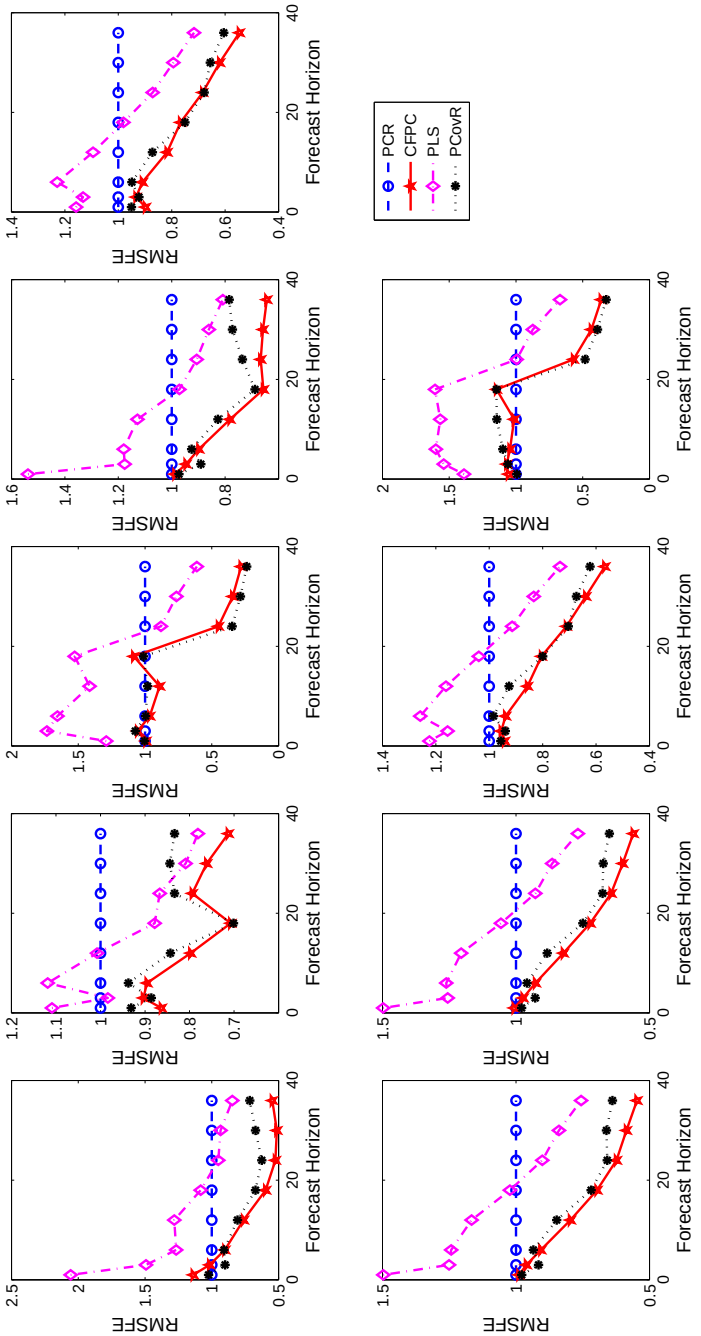



Figure 4.1: RMSFE without variable selection

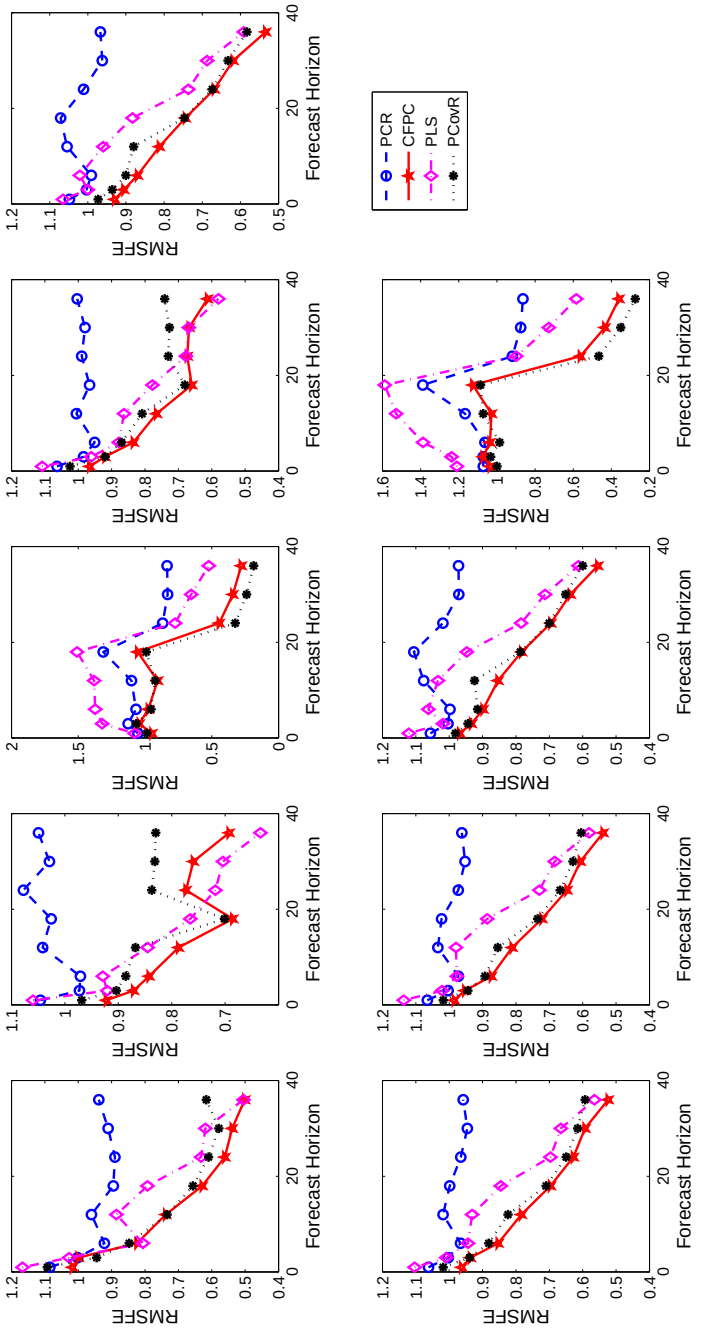



Figure 4.2: RMSFE with hardthreshold variable selection: threshold=1.65

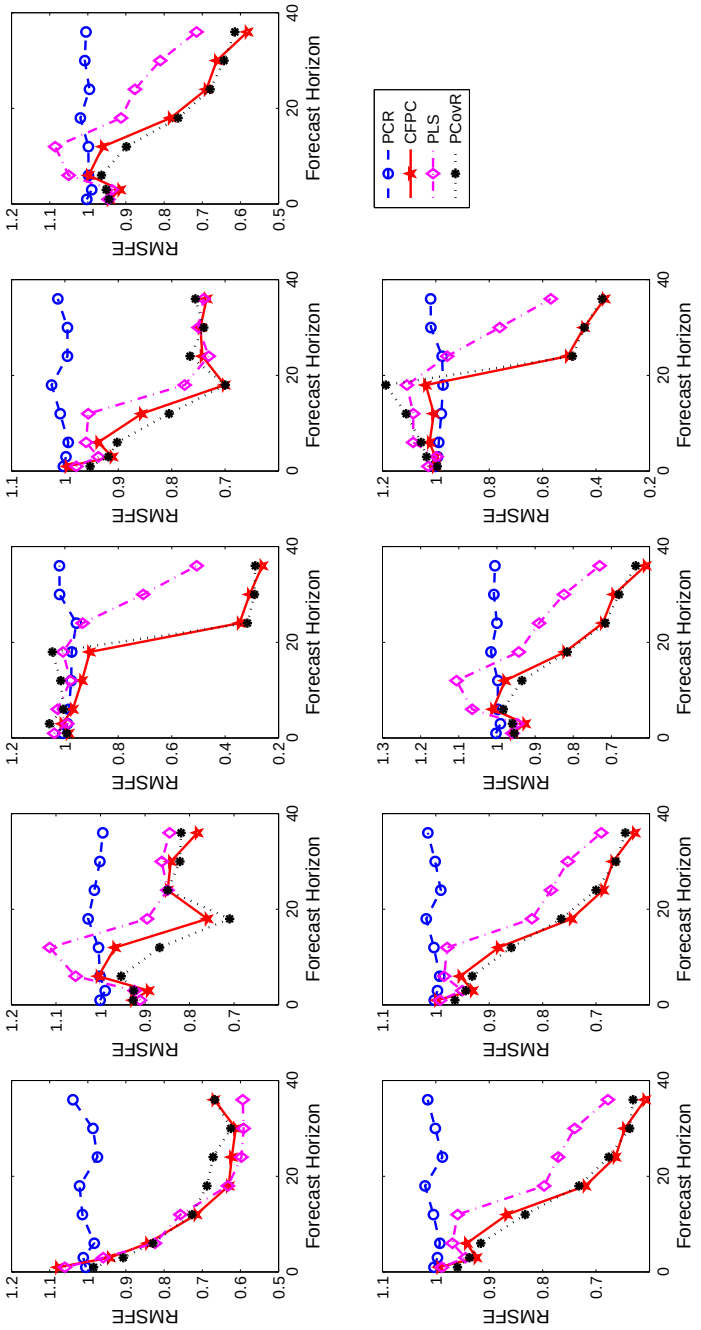



Figure 4.3: RMSFE with soft threshold variable selection: LARS(30)

# Part II

# Nonparametrics, Semiparametrics and Information Theoretic Econometrics

# Chapter 5

# Efficient Estimation of Nonparametric Simultaneous Equations Models

## 5.1 Introduction

Nonparametric structural models draw a lot of attention in recent years. However, simultaneous equations models considered so far impose different dependence structural relationship between the error terms and the instruments. One line of research, which can estimate the unknown structural up to a constant term, starts from Newey, Powell and Vella (1999). Recently, Su and Ullah (2008) proposed a three-step estimator that is more efficient than those of Pinkse (2000) and Newey and Powell (2003).

The chapters cited earlier all share the additive structure of the simultaneous equation models. Additive models are widely used in both theoretical economics and in econometric data analysis. See Linton (1997, 2000) and references there. Within the

framework of the single parameter linear exponential family, Linton (2000) exploits the additive structure of the nonparametric model and derive an estimator that can achieve oracle efficiency.

In this chapter, we exploit the procedure proposed in Su and Ullah (2008) one step further by exhausting the information contained in the additive structure of the simultaneous equation models. We follow a similar argument as in Linton (2000) to take the advantage of the additive structure. Thus we improve the estimator in Su and Ullah (2008) by first consistently estimating the nonparametric error term and then applying a local polynomial regression to consistently, and more importantly, efficiently estimate the nonparametric structure and its derivatives. The derived estimator achieves oracle efficiency as that in Linton (2000). Monte Carlo results show that our estimator is efficient compared to that in Su and Ullah (2008).

The organization of this chapter is as follows. Section 2 introduces our local polynomial estimator and proves its asymptotic properties. In section 3, we report Monte Carlo simulation results. Section 4 concludes.

## 5.2 Local Polynomial Estimator

We consider the regression model of Newey, Powell and Vella (1999) and Su and Ullah (2008):

$$\begin{cases} Y = g(X, Z_1) + \varepsilon, \qquad Z = (Z_1', Z_2')', \\ X = h(Z) + U,\ E(U|Z) = 0,\ E[\varepsilon|Z, U] = E[\varepsilon|U], \end{cases} \tag{5.1}$$

where $Y$ is an observable scalar random variable, g denotes the true, unknown structural function of interest, $X$ is $d_x \times 1$ vector of explanatory variables, $Z_1$ and $Z_2$ are $d_1 \times 1$ and $d_2 \times 1$ vectors of instrumental variables, $h \equiv (h_1, \ldots, h_{d_x})'$ is a $d_x \times 1$ vector of functions

of the instruments $Z$, and $U$ and $\varepsilon$ are disturbances. We are interested in estimating $g$ and its derivatives consistently.

Newey, Powell and Vella (1999) employed series approximations that exploit the additive structure of the model and propose a two-stage estimator of $g$, which is identified up to an additive constant if there is no functional relationship between $(X, Z_1)$ and $U$. They also derive consistency and asymptotic normality results for functional of their estimator. Su and Ullah (2008) develop a three-step kernel estimation procedure that can consistently estimate $g$ based on local polynomial regression and marginal integration techniques. They also establish the asymptotic distribution of their estimator under weak data dependence conditions. In addition, they provide simulation evidence which suggests the superior performance of their estimator compared to that proposed by Newey et al (1999).

Following Su and Ullah (2008), our estimation procedure is based on the following observation:

$$E[Y|X, Z, U] = g(X, Z_1) + E[\epsilon|U]. \tag{5.2}$$

Employing the law of iterated expectation gives,

$$m(X, Z_1, U) \equiv E[Y|X, Z_1, U] = g(X, Z_1) + E[\epsilon|U]. \tag{5.3}$$

Since $U$ is not observable, Su and Ullah (2008) used the estimated residual from the nonparametric regression of $X$ on $Z$ and estimated $g(x, z_1)$ up to a constant by first estimating $m(X, Z_1, U)$ and then integrating it over $U$.

Denote $m_u(U) = E[\epsilon|U]$ and note that the structure of (/refmxzu)implies that

$$\begin{aligned} g(X, Z_1) &= E[Y|X, Z_1, U] - E[\epsilon|U] \\ &= E[Y - E[\epsilon|U]|X, Z_1, U] \\ &= E[Y - m_u(U)|X, Z_1, U] \\ &= E[Y - m_u(U)|X, Z_1], \end{aligned}$$

if $U$ is observable and the functional form $m_u(\cdot)$ is known. Nevertheless, with their consistent estimators $\hat{U}$ and $\hat{m}_u(\cdot)$, we derive an estimator of $g(\cdot, \cdot)$ that can achieve the efficiency of the oracle estimator which requires the knowledge of both $U$ and $m_u(\cdot)$, following Linton (2000).

We state our estimation procedure as follows:

**1**. Proceed as in Su and Ullah (2008) procedure to get the estimators $\hat{h}(Z_t)$, $\hat{U}_t$, $\hat{m}(x, z_1, u)$ and $\hat{g}_Q(x, z_1)$.

**2**. Average $\hat{m}(x, z_1, u)$ over $(x, z_1)$ by a deterministic weight function $Q_1(x, z_1)$ to get an estimator of $m_u(u)$, $\hat{m}_u(u)$, with $\int_{\mathbb{R}^{d_x+d_1}} dQ_1(x, z_1) = 1$. We require that $Q_1$ has a bounded density on its support with respect to either Lebesgue measure or a counting measure in $\mathbb{R}^{d_x+d_1}$.

**3**. Obtain an estimator of $g(x, z_1)$ by a $p$-th order smoothing of $Y_t - \hat{m}_u(\hat{U}_t)$ on $X_t, Z_{1t}$ with kernel $K$ and bandwidth sequence $b = b(n)$. Denote the estimator as $\hat{g}^*(x, z_1)$.

Let $V \equiv (X, Z_1)'$ and $d \equiv d_x + d_1$. For the data set $\{X_t,\ Z_t\}_{t=1}^n$, the $p$-th order local polynomial regression of $Y_t - \hat{m}_u(\hat{U}_t)$ on $V_t$ can be obtained from the multivariate weighted least squared criterion:

$$nb^{-d} \sum_{t=1}^{n} K(\frac{\underline{V}_t - \underline{v}}{b}) \left[ Y_t - \hat{m}_u(\hat{U}_t) - \sum_{0 \leq |\underline{j}| \leq p} \theta_{\underline{j}}(\underline{v}) \, (\underline{V}_t - \underline{v})^{\underline{j}} \right]^2, \tag{5.4}$$

where $K$ is a nonnegative kernel function on $\mathbb{R}^d$ and $b = b(n)$ is a scalar bandwidth sequence. For other notations, we follow Masry (1996) and Su and Ullah (2008), $\underline{j} = (j_1, \ldots, j_d)'$, $\underline{j}! = \Pi_{i=1}^d j_i!$, $|\underline{j}| = \sum_{i=1}^d j_i$, $z^{\underline{j}} = \Pi_{i=1}^d z_i^{j_i}$, $\sum_{0 \leq |\underline{j}| \leq p} = \sum_{k=0}^p \sum_{j_1=0}^k \cdots \sum_{j_d=0}^k$, $b_{\underline{j}}(\underline{v}) = \frac{1}{\underline{j}!} D^{\underline{j}} g(\underline{y})|_{\underline{y}=\underline{v}}$, $D^{\underline{j}} g(\underline{y}) = \frac{\partial^{\underline{j}} g(\underline{y})}{\partial y_1^{j_1} \cdots \partial y_d^{j_d}}$. Minimizing (5.4) with respect to each $\theta_{\underline{j}}(\underline{v})$ gives an estimate $\hat{\theta}_{\underline{j}}(\underline{v})$. Note that $\underline{j}!\hat{\theta}_{\underline{j}}(\underline{v})$ estimates $D^{\underline{j}} g(\underline{v})$, that is, $\widehat{D^{\underline{j}} g(\underline{v})} \equiv \underline{j}!\hat{\theta}_{\underline{j}}(\underline{v})$. Therefore, $\hat{\theta}_0(\underline{v})$ is the estimator of $g(x, z_1)$ of interest. Arrange the distinct values of the d-tuple $b^{|\underline{j}|}\hat{\theta}_{\underline{j}}$ as a sequence in a lexicographical order in $\hat{\underline{\beta}}_{n,i}$, where $i = |\underline{j}|$. Then collect $\hat{\underline{\beta}}_{n,i}$, $0 \leq i \leq p$, as a column vector in the form $\hat{\underline{\beta}}_n = \left[\hat{\underline{\beta}}_{n,0}, \hat{\underline{\beta}}_{n,1}, \ldots, \hat{\underline{\beta}}_{n,p}\right]'$.Similarly, define $\underline{\beta}$ as the true value that corresponds to $\hat{\underline{\beta}}_n$ and denote $\sigma^2(\underline{v}) = var\left[Y_t - m_u(U_t)|\underline{V}_t = \underline{v}\right]$.

Before presenting our theorem, we introduce the following notations. Following Masry (1996), let $N_l = \begin{bmatrix} l+d+1 \\ d-1 \end{bmatrix}$ be the number of distinct $d$-tuples $\underline{j}$ with $|\underline{j}| = l$. Arrange these $N_l$ $d$-tuples as a sequence in a lexicographical order (with highest priority to last position so that $(0, \ldots, 0, i)$ is the first element in the sequence and $(i, 0, \ldots, 0)$ is the last element) and let $\phi_i^{-1}$ denote this one-to-one map. Denote $N = \sum_{l=0}^p N_l(d)$. For each $\underline{j}$ with $0 \leq \underline{j} \leq 2p$, let $\mu_{\underline{j}}(K_i) = \int_{R^{d^i}} \underline{w}^{\underline{j}} K(\underline{w}) d\underline{w}$. For each $\underline{j}$ with $0 \leq |\underline{j}| \leq p$, let $\gamma_{\underline{j}}(K_i) = \int_{R^d} \underline{u}^{\underline{j}} K^2(\underline{u}) d\underline{u}$. Define the $N \times N$ dimensional matrices $M$ and $\Gamma$, and the $N \times N_{p+1}$ matrix $B$ by

$$M = \begin{pmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p} \\ M_{1,0} & M_{1,1}^i & \cdots & M_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ M_{p,0} & M_{p,1} & \cdots & M_{p_i,p} \end{pmatrix}, \quad \Gamma = \begin{pmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \cdots & \Gamma_{0,p} \\ \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{p,0} & \Gamma_{p,1} & \cdots & \Gamma_{p,p} \end{pmatrix}, \quad B = \begin{pmatrix} M_{0,p+1} \\ M_{1,p+1} \\ \vdots \\ M_{p,p+1} \end{pmatrix},$$

where $M_{l,m}$ and $\Gamma_{l,m}$ are $N_l \times N_m$ dimensional matrices whole $(q, r)$ elements are, respectively, $\mu_{\phi_l(q)+\phi_m(r)}$ and $\gamma_{\phi_l(q)+\phi_m(r)}$. Note that the matrices $M$ and $\Gamma$ are essentially

multivariate moments of the kernels and higher order products of the kernels. In addition, $\underline{m}_{p+1}(\underline{v})$ collects $\frac{1}{\underline{k}!}\left(D^{\underline{k}}g\right)(\underline{v})$ in a lexicographical order.

We state the following asymptotic normality result for $\hat{\underline{\beta}}_n$.

**Theorem** Under Assumptions of Su and Ullah (2008) and $b = O(n^{-1/(d+2p+2)})$, we have

$$\left(nb^d\right)^{1/2}\left(\hat{\underline{\beta}}_n - \underline{\beta} - b^{p+1}M^{-1}B\underline{m}_{p+1}(\underline{v})\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(\underline{v})M^{-1}\Gamma M^{-1}/f(\underline{v})\right)$$

at continuity points $\underline{v}$ of $\left\{\sigma^2, f\right\}$ whenever $f(\underline{v}) > 0$, where $f(\underline{v})$ is the density function of $\underline{v} = (x, z_1)$.

**Proof:** See appendix.

**Remark**: Note that the term $\sigma^2(\underline{v}) = var\left[Y_t - m_u(U_t)|\underline{V}_t = \underline{v}\right]$ in the asymptotic variance depends on the knowledge of the unobserved error term $m_u(U_t)$. And note that the variance of an estimator that minimizes (5.4) with knowledge of $U_t$ has the same variance as our proposed estimator $\hat{\underline{\beta}}_n$. Thus, our estimator is oracle efficient in the sense of Linton (2000).

## 5.3 Monte Carlo Simulation

In this section, we perform Monte Carlo simulation to examine the properties of the estimator we proposed. We assume $E(\varepsilon) = 0$ and compare it with the estimators in Su and Ullah (2008), with data generating processes (DGPs) similar to theirs:

$$DGP1: \begin{cases} Y_t = 2\Phi(X_t) + \varepsilon_t, \\ X_t = Z_t - 0.2Z_t^2 + U_t. \end{cases} \qquad DGP2: \begin{cases} Y_t = \log(X_t) + \varepsilon_t, \\ X_t = 10 + \exp(0.1Z_t) + U_t. \end{cases}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal random variable. The error terms $\varepsilon_t$ and $U_t$, and the instrument $Z_t$ are generated according to

$$\varepsilon_t = \theta w_t + 0.3v_{Y_t},\ U_t = 0.5w_t + 0.2v_{X_t},\ Z_t = 1 + 0.5Z_{t-1} + 0.5v_{Z_t}, \tag{5.5}$$

Table 5.1: Relative Root Mean Squared Errors

| DGP | Mean | Median | Mean | Median | Mean | Median |
|---|---|---|---|---|---|---|
| N=100 | $\theta = 0.2$ | | $\theta = 0.5$ | | $\theta = 0.8$ | |
| 1 | 0.3844 | 0.5150 | 0.2306 | 0.4330 | 0.4170 | 0.4064 |
| 2 | 0.2596 | 0.3908 | 0.3705 | 0.5133 | 0.4589 | 0.5756 |
| N=400 | $\theta = 0.2$ | | $\theta = 0.5$ | | $\theta = 0.8$ | |
| 1 | 0.1361 | 0.3327 | 0.2152 | 0.3310 | 0.2398 | 0.3181 |
| 2 | 0.2848 | 0.4433 | 0.3253 | 0.5011 | 0.3595 | 0.5672 |

in which $v_{Y_t}, v_{X_t}, v_{Z_t}, w_t$ are *i.i.d.* sum of 48 independent random variables each uniformly distributed on $[-0.25, 0.25]$. Note that $v_{Y_t}, v_{X_t}, v_{Z_t}, w_t$ have bounded support $[-12, 12]$ and central limit theorem implies that these variables are nearly normally distributed. As seen in (5.5), correlation between $\varepsilon_t$ and $X_t$ is characterized by the parameter $\theta$, and we consider the following specification values: $\theta = 0.2, 0.5, 0.8$.The correlation between $\varepsilon_t$ and $X_t$ increases as $\theta$ increases and the problem of simultaneity is further magnified.

For each DGP and estimator, we consider two sample size: $n = 100$ and 400, with 200 repetitions for each n. We compute the mean of the root mean squared errors (RMSEs) of our estimator of $g(x)$ by averaging across the realized values of X and the 200 repetitions. These mean of RMSEs relative to those of Su and Ullah (2008) are reported in Table 1. Also, we report the median of the RMSEs of the two estimators obtained by averaging across the realized values of X only. It is clear from the results that the new estimation procedure gives more efficient estimator, the relative mean of RMSEs being all smaller than 1.

## 5.4 Conclusion

We propose a new estimator based on local polynomial regression and marginal integration techniques in this chapter. It is oracle efficient and it exhausts the information contained in the additive structure of the model. Our simulation results show that it is more efficient than the estimator in Su and Ullah (2008) in the sense that the MSE is much smaller.

# Appendix

**Proof of Theorem.** Denote $s_{n,\underline{j}} = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\underline{V}_t-\underline{v}}{b}\right)^{|\underline{j}|} K_b\left(\underline{V}_t-\underline{v}\right)$.Arrange the possible values of $s_{n,\underline{j}+\underline{k}}$ by a matrix $S_{n,|\underline{j}|,|\underline{k}|}$ in a lexicographical order with the $(l,m)$ element of $S_{n,|\underline{j}|,|\underline{k}|}$ given by

$$\left(S_{n,|\underline{j}|,|\underline{k}|}\right)_{l,m} = s_{n,\phi_{\underline{j}}(l)+\phi_{\underline{k}}(m)}.$$

The matrix $\left(S_{n,|\underline{j}|,|\underline{k}|}\right)$ is of dimension $N_{|\underline{j}|}\times N_{|\underline{k}|}$. Now define the $N\times N$ matrix $S_n$ by

$$S_n = \begin{pmatrix} S_{n,0,0} & S_{n,0,1} & \cdots & S_{n,0,p} \\ S_{n,1,0} & S_{n,1,1} & \cdots & S_{n,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,p,0} & S_{n,p,1} & \cdots & S_{n,p,p} \end{pmatrix}.$$

From the F.O.C. of the minimization criterion (5.4), we can derive

$$\underline{\hat{\beta}}_{-n} - \underline{\beta}_{-n} = S_n^{-1}\hat{\tau}_n^* + b^{p+1}S_n^{-1}B_n\underline{m}_{p+1}\left(\underline{v}\right) + o_p\left(b^{p+1}\right),$$

where $\hat{\tau}_n^* = \tau_n^* + \bar{J}_1 + \bar{J}_2$ is a compact form of

$$
\begin{aligned}
\hat{t}^{*}_{n,\underline{j}} &= \frac{1}{n}\sum_{t=1}^{n}\left[Y_t - \hat{m}_u(\hat{U}_t) - g(\underline{V}_t)\right]\left(\frac{\underline{V}_t - \underline{v}}{b_3}\right)^{\underline{j}} K_{3b_3}(\underline{V}_t - \underline{v}) \\
&= \frac{1}{n}\sum_{t=1}^{n}\left[Y_t - m_u(U_t) - g(\underline{V}_t)\right]\left(\frac{\underline{V}_t - \underline{v}}{b_3}\right)^{\underline{j}} K_{3b_3}(\underline{V}_t - \underline{v}) \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\left[m_u(U_t) - m_u(\hat{U}_t)\right]\left(\frac{\underline{V}_t - \underline{v}}{b_3}\right)^{\underline{j}} K_{3b_3}(\underline{V}_t - \underline{v}) \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\left[m_u(\hat{U}_t) - \hat{m}_u(\hat{U}_t)\right]\left(\frac{\underline{V}_t - \underline{v}}{b_3}\right)^{\underline{j}} K_{3b_3}(\underline{V}_t - \underline{v}) \\
&\equiv t^{*}_{n,\underline{j}} + J_{1,\underline{j}} + J_{2,\underline{j}}
\end{aligned}
$$

It follows from Masry (1996) that as $n \to \infty$, $S_n \overset{M.S.}{\to} Mf(\underline{v})$, $B_n \overset{M.S.}{\to} Bf(\underline{v})$ and

$$
\left(nb^d\right)\tau^{*}_n \overset{L}{\to} \mathcal{N}\left(0, \sigma^2(\underline{v}) f(\underline{v}) \Gamma\right).
$$

Thus, the asymptotic normality of $\hat{\underline{\beta}}_n$ depends properties of $J_{1,\underline{j}}$ and $J_{2,\underline{j}}$. First, it is easy to show that $\left(nb^d\right)^{1/2} J_{1,\underline{j}} = o_p(1)$, using Taylor series expansion similar to Su and Ullah (2008). Second, $\left(nb^d\right)^{1/2} J_{2,\underline{j}} = o_p(1)$. To see this, note that it is straightforward to show that,

$$
\begin{aligned}
m_u\left(\hat{U}_t\right) - \hat{m}_u(\hat{U}_t) &= \frac{1}{n}\sum_{s=1}^{n}\left[\hat{g}(X_s, Z_{1s}) - g(X_s, Z_{1s})\right] + \\
&\qquad \frac{1}{n}\sum_{s=1}^{n}\left[m\left(X_s, Z_{1s}, \hat{U}_t\right) - \hat{m}\left(X_s, Z_{st}, \hat{U}_t\right)\right].
\end{aligned}
$$

It follows from Su and Ullah (2008) that

$$
\hat{g}(X_s, Z_{1s}) - g(X_s, Z_{1s}) = o_p(1).
$$

and from Masry (1996) that

$$
\left[m\left(X_s, Z_{1s}, \hat{U}_t\right) - \hat{m}\left(X_s, Z_{st}, \hat{U}_t\right)\right] = o_p(1).
$$

Combining these results, we have $\left(nb^d\right)^{1/2} J_{2,\underline{j}} = o_p(1)$ following a similar argument as

in Su and Ullah (2008). Therefore, $\left(nb^d\right)\hat{\tau}_n^* \xrightarrow{L} \mathcal{N}\left(0, \sigma^2\left(\underline{v}\right) f\left(\underline{v}\right)\Gamma\right)$, which completes the proof of the theorem.□

## References


Linton, O., 1997. Efficient estimation of additive nonparametric regression models. *Biometrika* 84, 469–473.

Linton, O., 2000. Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory* 16, 502–523.

Masry, E., 1996. Multivariate regression estimation: local polynomial fitting for time series. *Stochastic Processes and their Applications* 65, 81–101.

Newey, W.K., Powell, J.L., 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71, 1565–1578.

Newey, W.K., Powell, J.L., Vella, F., 1999. Nonparametric estimation of triangular simultaneous equation models. *Econometrica* 67, 565–603.

Pinkse, J., 2000. Nonparametric two-step regression estimation when regressors and errors are dependent. *Canadian Journal of Statistics* 28, 289–300.

Su, L., Ullah, A., 2008. Local Polynomial Estimation of Nonparametric Simultaneous Equation Models. *Journal of Econometrics* 144, 193-218

# Chapter 6

# Model Averaging PartiaL Effect (MAPLE) Estimation with Large Dimensional Data

## 6.1 Introduction

We live in a world full of valuable information recorded by thousands of economic and financial variables. Economic researchers, policy makers and financial analysts are faced with these overwhelming economic signals. In theoretical macroeconomics, agents are forced to process all available quantities when they form expectations for future. In program evaluation, experts incorporate individual features such as gender, education, marriage status, family size, health status, etc. to analyze the treatment effect. In labor economics, newly available sources of data are called forth to advance theory and inform policy. In finance, equity premium is studied with thousands of financial variables, indices and macro policy variables.

This chapter, to our knowledge, serves as the first work to study the marginal effect of one variable on another in the large dimensional data setting, with the use of model averaging. This problem is typical in any field of economics, since artistic economic theory would suggest plenty of variables that would be potentially related to the variable of interest (Sala-i-Martin et al 2006). When it comes to estimation of such partial effect, the omission of other variables from the model would lead to biased estimate, fallible inference and result in misleading policy recommendation. In the following subsections, we first make clear the problem of estimation in the presence of large dimensional data, then review the related literature and finally spell out the contributions of the chapter.

### 6.1.1 Large Dimensional Data v.s. Small Models

With the advancement of computer technology, economic and financial data are more easily collected, shared and utilized in studies. Resources for Economists on the Internet[1] provides a wide range of economic topics with links to many different data sources. National Bureau of Economic Research[2] provides links to various data sources including macro data, industrial data, hospital data, demographic and vital statistics, patent and scientific papers data, and so forth. Penn World Table[3] provides purchasing power parity and national income accounts converted to international prices for 188 countries for years 1950-2004. In finance and business, Datastream by Thompson Financial[4] and Wharton research Data Services[5] provide researchers worldwide with

[1] http://rfe.org/

[2] http://www.nber.org/data/

[3] http://pwt.econ.upenn.edu/

[4] http://www.thomsonone.com/

[5] http://wrds-web.wharton.upenn.edu/wrds/

instant access to financial and marketing series. Yahoo[6] and the Federal Reserve Bank of St Louis maintain free data access to a wide variety of financial time series.

Economic models are introduced and estimated to analyze the linkage among economic variables and characterize the relationship of interest. With the principle of parsimony, researchers usually start with small models that focus on salient features of economic phenomena. For example, Keynes (1936) hypothesized that the major influence on individual consumption is personal income; Phillips (1958) and numerous work afterwards described an inverse relationship between money wage changes and unemployment in British economy; Mincer (1976) studied the direction of labor mobility resulting from minimum-wage imposition; Ashenfelter (1978) attribute current earning to past earnings and job training.

While these models are argued to explain economic phenomena, economists usually implicitly or explicitly require the environment under investigation hold *ceteris paribus*. This superiority is appreciated together with Occam's Razor in all scientific exploration. However, such a parsimony principle is better interpreted as a heuristic rather than an irrefutable principle of logic (Gernert, 2007). It has been maintained in economic modeling for mainly two reasons: First, analyzing the full model with all available economic variables would result in difficulties in parameter identification, estimation and model evaluation, driving us astray from the economic analysis originally designated.[7] The second reason that leads to simple models is that economics is more complex than it appears. Modeling methods available in mainstream science aim to

[6] `http://finance.yahoo.com/`

[7] With advancement in fuzzy analysis, set identification and inference has achieved significant progress. In economic applications, see Manski (1995, 2003, 2007), Imbens and Manski (2004), Santos (2011), Romano and Shaikh (2008, 2010), to name a few. Inference with large dimensional data is still left open.

separate important linkage from abounded noisy signals. This intrinsic feature limits inference in the presence of large data sets.

### 6.1.2 Related Literature

Dimensionality reduction techniques have been proposed and frequently used in forecasting literature when large dimensional data are present.The first line of research assumes that the data is generated by some underlying factors of smaller dimension and approaches the estimation of the common factors in a way fitting the problem at hand.[8] For recent work in this direction, see Bai and Ng (2010) and references therein. Another direction to achieve dimensionality reduction is variable selection. Selection is conducted by minimizing some objective loss functions, such as Akaike information criterion (AIC) or Bayesian information criterion (BIC). Early examples are forward variable selection, backward selection and stepwise selection etc. (Miller 2002). More recently the literature is overwhelmed by more sophisticated methods.[9] See Fan and Lv (2010) for a review.

Though popular in forecasting literature, dimensionality reduction methods have their own limitations when applying to partial effect estimation. In factor analysis, partial effect parameters are not estimated and factor loadings are hard to interpret. On the other side, variable selection is mostly concerned only with the explanation of the dependent variable by choosing a subset of regressors, but not with the estimation of the

[8]Popular examples are Principal Component Analysis (PCA) invented by Pearson (1901), factor analysis pioneered by Spearman (1904), Partial Least Square (PLS) developed by Wold (1966), Principal Covariate Regression (PCovR) proposed by De Jong and Kiers (1992), Supervised Factor Model (SFM) introduced by Tu and Lee (2011), and so forth.

[9]Examples include LASSO (Tibishirani 1996), SCAD (Fan and Li 2001), Elastic Net (Zou and Hastie 2005), group LASSO (Zou 2006), bridge estimator (Huang, Horowitz and Ma 2008) and so on.

partial effect. The key variable whose effect is of interest may be excluded from a variable selection procedure. Even though oracle properties of variable selection procedures (e.g. Huang, Horowitz and Ma 2008) have been established, these oracle properties do not provide a satisfactory answer in finite sample. First, when the variable of interest is not selected, oracle selection procedure such as bridge estimator would estimate the partial effect as zero. In this case, there is no way to do further inference such as constructing confidence interval or testing for the partial effect. That is, variable selection procedures would be over confident that the partial effect is zero when it is actually not. Second, even when the variable of interest is kept after the model selection procedure, the asymptotic distribution of the partial effect estimator depends on the true value of the partial effect and thus hard to provide valid inference in finite sample. See Leeb and Pötscher (2005, 2006, 2008abc, 2009), Pötscher (2009) and Pötscher and Schneider (2009, 2010) for problems that involve inferences with model selection procedures. Theoretical investigation of partial effect estimation with dimensionality reduction methods demands more effort before they become the working force.

Statisticians have long noticed that "all models are wrong but some are useful" (Box, 1979). This famous quote vividly describes a dilemma with which theoretical researchers are forced to face: models are misspecified. Taken as granted, we're in a position to estimate parameters of interest in misspecified models. For example, program evaluation researchers are evaluating the effects of the treatment with their misspecified model. The partial effect thus computed potentially suffers from model misspecification bias. Macro policy makers are predicting the effects of a counterfactual policy on the performance of economy, using a misspecified model. The prediction is as accurate as the model itself. Luckily for researchers that are concerned with partial effect parameters, potentially of low dimension, they are free of this misspecification problem, as to be

pointed out by this chapter. We specify a condition under which researchers who are interested in learning the partial effect parameters can well proceed with a misspecified model. Nevertheless, the parameter of interest should be correctly identified within the model. This is a big step, following White and Lu (2010), towards the estimation of economic sensible parameters rather than some statistical projection coefficients. It is important to point out that the identified partial effect parameters have the causal effect interpretation but the regression coefficients do not (e.g., White and Chalak 2006, White and Lu 2010). In a word, classical modeling and estimation approaches are contaminated with bias and new estimation techniques are called upon to derive more efficient estimators. This chapter suggests the use of model averaging to achieve this aim.

Model averaging, advocated by Bates and Granger (1969), works as an alternative to the factor approach or variable selection in the forecasting paradigm. Simple model averaging gains a lot of popularity in financial market forecasts, for example, Rapach et al (2010). Recently, Hansen (2007, 2008, 2009, 2010) proposed model averaging with Mallow's criterion to select the combining weights, while Hansen and Racine (2011) proposed Jackknife model averaging. Model averaging is shown to be promising in forecasting exercises due to at least three facts: First, averaging reduces variances while incurring small bias. Whenever the bias is relatively small compared to the variance reduction, model averaging performs better than individual models in Mean Squared Error (MSE) sense. Secondly, individual models are likely to be misspecified and exclude information that is incorporated in averaging models. This loss of information potentially degrades the power of a single model. Thirdly, model uncertainty is somehow reduced in averaging model attributing to the observation that it incorporates individual models

as special cases by properly assigning the weights, spanning a larger model space and reducing the chance of misspecification.

However, the power of model averaging for parameter estimation has not been fully explored. Hansen (2009) applied model averaging for parameter estimation in a structural break setting. The idea of averaging estimator dates back to Breiman (1996), where a **b**ootstrap method is implemented together with model **agg**regat**ing** (bagging, hereafter).[10] There is a large literature on Bayesian Model Averaging (BMA).[11] BMA takes a different perspective that the parameter of interest is random rather than have a true value. A prior on the parameter is required and a computing algorithm (e.g. MCMC) is needed to derive the BMA estimator. The dependence of the results on the prior and the algorithm adopted usually weaken the conclusions therefore arrived. More than often, convergence of the computing algorithms available (e.g. Metropolis-Hastings, Gibbs sampler, or $MC^3$, etc.) is hard to check in practice. See Hoeting et al (1999) for more details about these challenges faced by Bayesian researchers.

### 6.1.3 Contributions

This chapter contributes to the literature in the following regards: First, we lay out the conditions that help to identify the partial effect parameter of interest in a large dimensional model. We show that Conditional Mean Independence (CMI) is sufficient for this purpose. This is a weaker condition than conditional independence

[10] Breiman (1996) shows that bagging estimator has a smaller MSE in the i.i.d. case for the the purpose of prediction. Bulman and Yu (2002) establish the theoretical properties of bagging estimators, followed by Lee, Tu and Ullah (2011ab) and Tu (2011a) that adopt bagging for constrained parameter estimation in nonparametric setting.

[11] In economics, recent work on BMA includes Sala-i-martin et al (2004), Eicher et al (2009) and so on.

used in White and Lu (2010). When CMI does not hold, we state a weaker condition, Weak Conditional Mean Independence, that identifies the partial effect parameter when the number of observation is large. CMI conditions can be either implied by conditional independence (White and Chalak 2010, Su and White 2011) or easily checked using the nonparametric tests proposed by Li and Wang (1998) or Hsiao, Li and Racine (2007). An information-based approach that is easy to implement is also suitable to test CMI.[12] We emphasize that such estimated coefficients would have economic interpretation like causal effects only under identification. However, this identification issue is often ignored by empirical researchers, especially those who experiment with including and excluding explanatory variables till they get coefficient estimates agree with initial intuition.

Second, we consider the situation in which the parameter of interest can be identified in more than one model. This is often the case when we have large dimensional data. We propose two **m**odel **a**veraging **p**artia**l** **e**ffect (MAPLE) estimators in this setting. One of the estimators is generalized-method-of-moment based MAPLE (gMAPLE) and the other is entropy-based MAPLE (eMAPLE). The estimators are constructed from model averaging point of view, utilizing more than one model (potentially misspecified) to quantify such partial effect. They utilize more information than partial effect estimator derived from each individual model. Averaging in this way helps to wipe out the large bias lying in individual estimator and reduces variances, especially in small sample.

The gMAPLE estimator is constructed through combining all the moment conditions specified by individual models. A GMM-like objective function is used to derive the gMAPLE estimator. This estimator is different from the classical GMM estimator proposed by Hansen (1982) in the sense that each model has its own unique parameters

[12]See Tu (2011b) for more details.

other than the common partial effect parameter. This estimator looks similar to but differs from the GMM estimator of Seemingly Unrelated Regression models because of the common partial effect parameter in each model. gMAPLE estimator is the first attempt, as far as we know in the literature, to use moment conditions of more than one model to conduction inference on parameters of interest, while treating other parameters as pseudo ones.

The eMAPLE estimator is motivated from the Maximum Entropy point of view, i.e., to maximize the uncertainty of the model and data that is consistent with the moment conditions that identify the partial effect parameter. The main intuition is that the same set of data would occur with different probability if they are generated from different models. We introduce the concept of entropy of a model class in line with the classical notion of entropy of a random variable. We similarly define the conditional and joint entropy between a model class and random variables generated from that model class. Our eMAPLE estimator is constructed such that the conditional entropy of the model class given the observations is maximized. That is, the uncertainty of the model class is maximized given that the data is observed. Model averaging with entropy-based weights opens a new area of Maximum Entropic Econometrics (MEE). Other than estimating the probability of each observation in classical MEE, model averaging raises the question of probability of each individual model, instead of assigning equal probabilities. eMAPLE estimation is a novel statistical inference approach in that it introduces model uncertainty and model averaging into the entropy paradigm for parameter estimation. The inference based on the objective function (joint entropy) to construct confidence intervals or testing restrictions for the parameters of interest is easy to carry out and often better resembles the asymptotic results in finite sample than competing methods.

The third contribution of the chapter is the theoretical study of the two MAPLE estimators. We set up conditions under which our MAPLE estimators are consistent and asymptotically normal. The conditions for gMAPLE estimator are similar to those in the GMM literature. The conditions for eMAPLE estimator resemble those used in the Generalized Empirical Likelihood (GEL) literature.[13] Testing of nonlinear restrictions on the parameters is also considered. We show that the Wald, Rao's Score and Likelihood-ratio type tests based on our MAPLE estimators are asymptotically chi-squared distributed.

The fourth contribution is the thorough simulation study conducted to compare various partial effect estimators, including MMA, JMA, FOGLeSs etc.. Our gMAPLE and eMAPLE estimator are shown to have appealing finite sample properties in various Data Generating Processes, including factor model, large dimensional models, models with large number of irrelevant regressors and models with heterogeneous errors etc.. Evaluation measures including Mean Squared Errors, Mean Absolute Errors, Bias, Variance, Inter Quantile Range are used to compare the competing estimators. Our MAPLE estimators clearly stand out, especially in small samples, and even achieve the oracle efficiency lower bound in MSE in some designs (true design is small dimensional without heterogeneity, but with a large dimensional covariates). We also conduct simulations to examine the performance of the MAPLE based test statistics. Generally, these tests enjoy sizes closer to theoretical ones than other testing procedures including, e.g., FOGLeSs based tests.

Finally, we illustrate the use of MAPLE estimator in an economic application to evaluate the effect of inherited control on firm performance. We find that our MAPLE estimates confirm earlier findings by Pérez-González (2006) and White and Lu (2010)

[13] See Kitamura (2006) for a review.

that there is a negative effect, i.e., firms with family related CEOs tend to underperform those with family unrelated CEOs. However, confidence intervals constructed based on MAPLE estimators are much narrower than those based on FOGLeSs estimator, which indicates the superority the proposed approach.

Structure of the rest of this chapter is planned as follows: Section 2 presents the model and discusses the identification issues. Section 3 proposes the gMAPLE estimator, introduces the concept of entropy of models in the presence of model uncertainty and proposes the eMAPLE estimator. Section 4 presents the theoretical properties of the proposed MAPLE estimators. Section 5 studies the finite sample properties, via simulation experiments, of our estimator together with other competitors. Section 6 provides an illustration of our estimation approach with the dataset of Pérez-González(2006) in the study of the impact of inherited control on firm performance. Section 7 concludes and comments on future studies. All the technical proofs are collected in the Appendix.

## 6.2 The Model and Identification Condition

In this section, we introduce the model with large dimensional data and illustrate with six economic examples. We discuss the identification of the partial effect parameter of interest and present the key condition, Conditional Mean Independence (CMI), that serves for identification purpose. Other approaches for identification are also discussed in a concise way. In the end, We point out other related issues.

### 6.2.1 The Model

We present the model after introducing notations. Let $\mathbf{y}$ denote the $n \times 1$ dependent variable, $\mathbf{x}$ the $n \times 1$ exogenous independent variable whose partial effect on

$\mathbf{y}$ is of major interest, and $\mathbf{z}$ the large dimensional independent variables.

**Assumption A.1 (linearity)**:

$$y_i = \mathbf{x}_i^\tau \beta + \mathbf{z}_i^\tau \gamma + \varepsilon_i \ , (i = 1, 2, \cdots, n) \tag{6.1}$$

where $\beta$ is the partial effect vector of interest, $\gamma$ is a large-dimensional coefficient vector and $\varepsilon_i$ is the disturbance term.

**Assumption A.2 ($\alpha$-mixing stationarity)**: The large dimensional vector stochastic process $\{d_i\}_{i=1}^n \equiv \{y_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i\}_{i=1}^n$ is a stationary $\alpha$-mixing process with mixing coefficients $\alpha(j)$ satisfying $\sum_{j=1}^\infty j^2 \alpha^{\epsilon/(\epsilon+2)}(j) < \infty$ for $0 < \epsilon \leq 1$, where $\mathbf{w_i}$ is some instrumental vector.

**Assumption A.3 (moment restriction)**: All the instruments are orthogonal to the contemporaneous error term: $E(\mathbf{w}_{ik} \cdot \varepsilon_i) = \mathbf{0}$, for all $i$ and $k$ $(= 1, 2, \cdots, \dim(\mathbf{w}_i))$.

**Assumption RC (rank condition)**: $E[\mathbf{w}_i (\ \mathbf{x}_i^\tau,\ \mathbf{z}_i^\tau)]$ is finite with full column rank.

We comment on the strength of Assumption A and RC before presenting some economic examples.

**Remark A**: Assumption A.1 assumes that the relationship between $y$ and the covariates is linear. As we see later on, we require that $y$ be linear in unknown parameters. This assumption is not restrictive and can be extended in various ways. However, we will maintain this assumption only to clarify the presentation of our identification and estimation approach. In addition, the model as specified does not contain an intercept term. This is not restrictive either since a demean of the data would remove the intercept. We emphasize that the model is structural in the sense that the parameters, e.g., $\beta$ carry causal effect interpretation. More than often, a low dimensional parameter such as $\beta$ has economic policy implication but not others contained in $\gamma$. Our inference

is mainly concerned with $\beta$. Assumption A.2 is classical since that certain type of nonstationary process can be made stationary via transformations such as differencing or detrending. Dependence across observations is allowed by the $\alpha$-mixing condition. Assumption A.3 would meet since we include all possible explanatory variables in the regression. Anything else that is not explained in the dependent variable should be due to the pure random error term.

**Remark RC**: The rank condition is needed to identify all the unknown parameters, but often fails when $z_i$ is of large dimension. This is especially the case for economic models, contrasted to statistical models, since all economic variables are closely intertwined. In the case that a few economic variables are linearly dependent RC fails to hold. However, as argued earlier, economists more than often are concerned with only the partial effect parameter, $\beta$, but not the other coefficient vector $\gamma$. This observation is momentum since its implication is that we only need focus on identification and inference on $\beta$. This alleviates the need for Assumption RC and allows us to proceed with weaker condition such as Conditional Mean Independence. We will introduce CMI for identification after presenting some economic examples that highlight the importance of partial effect estimation in large dimensional data.

### 6.2.2 Examples

We briefly discuss some examples from macroeconomics, program evaluation and labor economics.

**Example 1** *(**Phillips Curve**) The famous historical inverse relationship between the rate of unemployment and the rate of inflation in the economy, usually termed as Phillips Curve (Fisher 1926; Phillips 1958), has been the focus of macro economy since its birth.*

*Yet this is short run phenomena. A cursory analysis of U.S. inflation and unemployment data 1953-92 reveals that there is no single curve that fits the data. However, this argument ignore the fact that the macro economy has been evolving over time and factors such as technological developments, institutional factors including macro policy are also affecting the curve. These factors might prove to be important, but do not change the relationship between unemployment rate and inflation rate. Therefore, the estimation of the Phillips Curve should incorporate other macro variables.*

**Example 2** *(**Consumption Hypothesis**) Keynes (1936) developed his theory of consumption and detailed the relationship between consumption and income in his famous book* "The General Theory of Employment, Interest and Money" *(Keynes, 1936). A function that relates consumption and income is usually estimated and Keynes' consumption theory is tested. The marginal propensity to consume (MPC), i.e., the rate at which consumption changes as income is changing, is the slope of the consumption function. According to Keynes, MPC should be in between 0 and 1. However, a consumption function that only has income as an explanatory cause suffers from potential misspecification bias. It bases consumption only on current income, but neglects other factors that also have important effects. One such factor is future income, which leads to Friedman's (1957) Permanent Income Hypothesis.*

**Example 3** *(**Treatment Effect**) Ashenfelter (1978) studied effect of training programs on earnings where individual characteristics such as gender, race, past earnings together with training variable.*

**Example 4** *(**Wage Equation**) Kruger (1993) examined the role of computers on the wage structure. A long list of variables such as gender, education, race, age, occupation, union status, hours, marriage status, experience and region are considered as important*

*factors when studying the effect of computers on wage.*

**Example 5** *(**Inherited Control**) Pérez-González (2006) used a large data set from 355 management transitions of publicly traded U.S. corporations to examine whether firms with family related incoming chief executive officers (CEOs) underperform in terms of operating profitability relatives to firms with unrelated incoming CEOs. 34 covariates are used including firm size, firm's past performance, board's R&D expenditure, departing CEO's separation conditions and incoming CEO's ownership, incoming CEO's characteristics, together with 17 year dummies. We will provide more analysis with this example in the empirical exercise in Section 6.*

**Example 6** *(**Economic Growth**) Sala-i-Martin et al (2004) studied the determinants of economic growth with 67 variables that correlate with economic growth with only 80 observations. This job would be in vain since we have a large number of unknowns compared to the number of observations. However, growth economists are interested to know whether a particular variable, e.g., human capital, is a determinant of economic growth, in the presence of large number of other covariates.*

### 6.2.3 Identification and Conditional Mean Independence

In this subsection, we look into the identification issue of the partial effect parameter $\beta$. We distinguish the identification problem for two cases: (i) when Assumption RC holds; and (ii) when Assumption RC fails. It is to be shown that Assumption RC, together with Assumption A.1, A.2 and A.3, are sufficient for $\beta$ to be identified. When Assumption RC fails, a further condition called conditional mean independence (CMI) is introduced to identify $\beta$. Tests to verify CMI and lower level conditions that imply CMI are reviewed.

Note that under Assumption RC, $E\left[\mathbf{w}_i\left(\ \mathbf{x}_i^\tau,\ \mathbf{z}_i^\tau\right)\right]$ is of full column rank. The moment restriction Assumption A.3 implies that,

$$E\left[\mathbf{w}_i\left(y_i - \mathbf{x}_i^\tau\beta - \mathbf{z}_i^\tau\gamma\right)\right] = \mathbf{0},$$

which is equivalent to

$$E\left[\mathbf{w}_i\left(\ \mathbf{x}_i^\tau,\ \mathbf{z}_i^\tau\right)\right]\begin{bmatrix}\beta\\ \\ \gamma\end{bmatrix} = E\left(\mathbf{w}_i y_i\right). \tag{6.2}$$

There is a unique solution to the above equation. This completes the identification of $\beta$.

#### 6.2.3.1 Conditional Mean Independence

If, on the other hand, Assumption RC fails, then $E\left[\mathbf{w}_i\left(\ \mathbf{x}_i^\tau,\ \mathbf{z}_i^\tau\right)\right]$ is singular. This leads to multiple solutions of $\beta$ in equation (6.2). Consequently, $\beta$ is underidentified. We find the following condition is needed for $\beta$ to be identified.

**Assumption CMI (conditional mean independence)**:

$$E\left(\mathbf{z}_i^2|\mathbf{x}_i, \mathbf{z}_i^1\right) = E\left(\mathbf{z}_i^2|\mathbf{z}_i^1\right) \tag{6.3}$$

where $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$ forms a partition of $\mathbf{z}_i$, i.e., $\mathbf{z}_i^\tau = \left[\mathbf{z}_i^{1\tau}, \mathbf{z}_i^{2\tau}\right]$ ,for $i = 1, 2, \cdots, n$.

CMI condition is quite commonly adopted in the literature of parameter identification. A similar form of CMI is used in Stock and Watson (2010, pp.232) to distinguish the role of variables of interest and control variables. Under CMI, the coefficient of the variable of interest is argued to have an interpretation of causal effect. In the case that $\mathbf{z}_i^2$ is univariate, tests of Li and Wang (1998) and Hsiao, Li and Racine (2007) can be easily adjusted to verify CMI condition. When $\mathbf{z}_i^2$ is multivariate, element-wise tests would apply.

Conditions stronger than CMI are, for example, conditional exogeneity and conditional independence. They have been imposed by Hahn (1998, 2004), White and Lu (2010) and White, Chalak and Lu (2010), to name a few, as major tools to study identification, treatment effect and Granger-Causality. Su and White (2007ab) suggest tests of conditional independence that are based on Hellinger metrics and empirical likelihood. White and Chalak (2010) provided tests for conditional exogeneity. See White and Chalak (2010), Su and White (2008) and references therein for details.

**Lemma 7** *$\beta$ is identified under Assumption 1, 2, 3, and CMI.*

The proof of Lemma 1 is given in Appendix A. More than often in economic modeling, we will assume the existence of a partition of $\mathbf{z}$ such that Assumption CMI is satisfied. As a result of Lemma 1, the partial effect parameter $\beta$ is identified.

When the set of $\mathbf{z}$ contains a large dimensional data, it is possible that more than one decomposition can be found such that (6.3) is satisfied. This is the case if $\mathbf{z}$ are linearly dependent. In this circumstance, we have competing models that all can identify $\beta$ according to Lemma 1. However, each model will produce a different estimate of $\beta$, for a given sample of observations. In practice, it is hard to tell which estimate is closer to the true value. An average estimate that aggregate these estimated values can be constructed, with weights inversely proportional to each individual variance. See White and Lu (2010) for example. However, the construction of this estimate requires the knowledge of variance of individual estimators. Estimates of the variance can be used in practice. Nevertheless, this estimation procedure is deemed to be inefficient since the estimation of $\beta$ takes into account different model specifications one-at-a-time and that the variance estimates are usually not accurate in finite sample. In the next

section, we propose a model averaging estimator that could potentially circumvent such difficulties and result in a more efficient estimator.

Next we investigate the more interesting situation when there is no such partition of $\mathbf{z}$ such that CMI holds. The direct consequence is that $\beta$ is not identifiable. We consider two cases (i) weak identification and (ii) no identification.

#### 6.2.3.2 Weak Conditional Mean Independence

**Assumption WCMI (weak conditional mean independence)**:

$$E\left(\mathbf{z}_i^2|\mathbf{x}_i, \mathbf{z}_i^1\right) = E\left(\mathbf{z}_i^2|\mathbf{z}_i^1\right) + \eta\mathbf{x}_i \tag{6.4}$$

where $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$ forms a partition of $\mathbf{z}_i$, i.e., $\mathbf{z}_i^\tau = \left[\mathbf{z}_i^{1\tau}, \mathbf{z}_i^{2\tau}\right]$ ,for $i = 1, 2, \cdots, n$, and $\eta$ is a matrix of the same dimension as $\mathbf{z}_i^2$, with Euclidean norm

$$||\eta|| = o\left(n^{-1/2}\right)$$

Under Assumption WCMI, $\beta$ is weakly identified. As sample size increase, the dependence of $\mathbf{z}_i^2$ on $\mathbf{x}_i$ becomes weaker and weaker. In the limit, condition WCMI becomes condition CMI. Therefore, $\beta$ is identified in the limit. This type of condition has been used in Belloni, et al (2011) to approximate the factor estimation.

**Lemma 8** *$\beta$ is weakly identified (identified in the limit) under Assumption 1, 2, 3, and WCMI.*

When WCMI condition fails, $\beta$ cannot be identified in any approximating models. This is a more interesting case, since the true model cannot be approximated arbitrarily well as we intend to. Our proposed estimator based on model averaging, tends to perform well for this difficult case, as shown in our simulation results in Section 5.

Before we proceed, a few things should be noted in sequence. First, when $\beta$ is not identified, estimators for $\beta$ using methods such as OLS are not targeting the correct partial effect. Inevitably, estimators are biased and their properties are hard to evaluate. In this circumstance, hardly any effort can be made towards the estimation of partial effect. Second, partial identification approaches advocated by Manski (2003) could be employed when WCMI fails, which is beyond the focus of this chapter. Third, the estimation of $\beta$ can be put into a general framework in which conditional moment restrictions summarize the model information. These restrictions take the form

$$E\left[g\left(y_i, \mathbf{x}_i, \mathbf{z}_i; \beta, \gamma\right)\right] = 0, \tag{6.5}$$

where $g\left(y_i, \mathbf{x}_i, \mathbf{z}_i; \beta, \gamma\right)$ has a known functional form, $\beta$ is the partial effect parameter of interest and $\gamma$ is a vector of pseudo parameters. Note that first, $g(\cdot)$ may be derived from a nonlinear model, thus it is not restricted to the model specified in (6.1). $\gamma$ can also be an infinite dimensional parameter such as a nonparametric function. Identifications of this type have been studied by Chen, et al (2011). A separate paper is written to study estimation in this semiparametric framework and leaves us to focus on the case when $\gamma$ is the coefficient of $\mathbf{Z}$.

## 6.3 Model Averaging PartiaL Effect Estimation

### 6.3.1 Model Uncertainty and Moment Uncertainty

We motivate the estimation of partial effect from the point of view of model uncertainty. Model (6.1) can be viewed as aggregated models from $\mathcal{M}$ with certain probabilities. For example, in state $s$, the dependent variable is generated through the following equation,

$$y_i = \mathbf{x}_i^\tau \beta + \mathbf{z}_{i,s}^\tau \gamma_s + \varepsilon_{i,s} \ (i = 1, 2, \cdots, n) \tag{6.6}$$

where $\mathbf{z}_{i,s}$ is a subset of $\mathbf{z}_i$ and $\gamma_s$ denotes corresponding coefficient vector. Denote the above model as $M_s$ and denote a collection of such models as $\mathcal{M}$. We emphasize that in (6.6), $\beta$ is identified via the CMI condition.

Ideally, if the observed data can be classified according to the state from which they are generated, we can estimate the coefficients $\beta$ and $\gamma$ within each state via LS whenever it applies. A second averaging procedure may be implemented after $\hat{\beta}_s$ is computed in state $s$ $(= 1, 2, \cdots, S)$ to derive a more efficient estimator $\hat{\beta}$ using an auxiliary regression. See White and Lu (2010) for such a construction via a pseudo regression of $\hat{\beta}_s$ on $\beta$. Nevertheless, classification of data into states is neither practical nor necessary. First, classification requires further information and renders the estimation even more complex. Inference after data classification or model selection raises challenging issues such as those in data snooping (White, 2000). See Berk et al (2009) and Berk et al (2011) for recent studies on this issue. Second, entropy-based inference is already suitable for this type of so-called ill-posed "inverse" problems. Partial effect estimation of $\beta$ amounts to estimating the model probability distribution $\mathbf{p}$ and model coefficients $\beta$ and $\gamma$. We present procedures that circumvent the classification difficulty as notified and achieves the estimation objective.

To put the analysis in a general framework, we present the estimation of $\beta$ from the model information characterized by moment constraints in the form of (6.5). To facilitate the presentation, we simplify our notations. Note that first, model (6.6) in state $s$ can be summarized by corresponding moment condition

$$E\left[g_s\left(d;\theta_0\right)\right] = 0, \tag{6.7}$$

where expectation is taken over random vector $d = (y, \mathbf{x}, \mathbf{z})$, with $g_s\left(\cdot;\cdot\right)$ denoting the moment restriction in $M_s$, and $\theta_0 = (\beta, \gamma_1, \ldots \gamma_S)$ collecting all the unknown parameters

in $S$ models. We emphasize that $\beta$ is the partial effect parameter of interest that is identified in each model, but not the projection coefficient vectors $\gamma_s, s = 1, \ldots S$.

### 6.3.2 gMAPLE

Facing parameter estimation problems identified by moment conditions via (6.7), it is natural to adopt the Generalized Method of Moment (GMM) approach proposed by Hansen (1982). We present the GMM estimator in the current setting. Denote $\bar{g}_s(d,\theta) = \frac{1}{n}\sum_{i=1}^{n} g_s(d_i,\theta)$ and $\bar{g}(d,\theta) = [\bar{g}_1^\tau(d,\theta), \ldots, \bar{g}_S^\tau(d,\theta)]^\tau$. The one-step GMM estimator with a weighting matrix $W$ is defined as

$$\hat{\theta}_{gMAPLE} = \arg\min_{\theta} \bar{g}^\tau(d,\theta) W \bar{g}(d,\theta). \tag{6.8}$$

The solution to this convex minimization problem can be easily found through numerical methods.

We need some notation to proceed. Define $\nabla_\theta g_s(d,\theta) = \partial g_s^\tau(d,\theta)/\partial\theta$, where $\partial g_s^\tau(d,\theta)/\partial\theta$ is the transpose of $\partial g_s(d,\theta)/\partial\theta$. Denote $G(s,\theta) = E[\nabla_\theta g_s(d,\theta)]$ and $V(s,\theta) = E[g_s(d,\theta) g_s^\tau(d,\theta)]$. Define $G(\theta) = (G^\tau(1,\theta), \ldots, G^\tau(S,\theta),)^\tau$, $V(\theta) = \text{diag}\,(V(1,\theta), \ldots, V(S,\theta))$ and use short notation $G = G(\theta_0)$, $V = V(\theta_0)$, $\Omega = E[g(d,\theta_0) g^\tau(d,\theta_0)]$. Following the GMM literature (see, e.g., Newey and McFadden, 1994), it is easy to establish the following theorem, under suitable set of additional assumptions on the moment conditions (6.7).

**Theorem 9** *The GMM estimator defined in (6.8) has the following properties:*

**(a)** $\hat{\theta}_{gMAPLE} \xrightarrow{p} \theta_0$.

**(b)** $\sqrt{n}\left(\hat{\theta}_{gMAPLE} - \theta_0\right) \xrightarrow{d} N\left(0, (G^\tau W G)^{-1} G^\tau W \Omega W G (G^\tau W G)^{-1}\right)$

An efficient two-step GMM estimator can be derived based on a first step estimator $\hat{\theta}_{gMAPLE1}$ that solves (6.8) by setting $W = I$, the identity matrix. The optimal weight matrix can be shown to be $W_{opt} = \Omega^{-1}$ that can be consistently estimated by

$$\hat{W}_{opt} = \left[ \frac{1}{n} \sum_{i=1}^{n} g\left(d_i, \hat{\theta}_{gMAPLE1}\right) g^{\tau}\left(d_i, \hat{\theta}_{gMAPLE1}\right) \right]^{-1}. \tag{6.9}$$

**Theorem 10** *The GMM estimator defined in (6.8) with $W = \hat{W}_{opt}$ have the following properties:*

**(a)** $\hat{\theta}_{gMAPLE} \xrightarrow{p} \theta_0$.

**(b)** $\sqrt{n}\left(\hat{\theta}_{gMAPLE} - \theta_0\right) \xrightarrow{d} N\left(0, I^{-1}\left(\theta_0\right)\right)$, with $I\left(\theta_0\right) = G^{\tau}\Omega^{-1}G$.

Note that (6.9) is a very large dimensional matrix in the current setting. Earlier results (e.g., Altonji and Segal 1994) show that GMM estimator with estimated optimal weighting matrix does not perform well in finite sample. The two-step optimal GMM can be even beaten by the one-step GMM that uses the naive identity weighting matrix. In practice, iterative GMM estimator and continuously updating GMM estimators can be used, see Hansen et al (1996).

### 6.3.3 eMAPLE

This section introduce entropy-based model averaging. We start by defining the entropy for model uncertainty of a given class of models. We then extend this concept and account for model uncertainty in the presence of random variables that are generated from the models. Similar concepts, such as entropy, joint entropy and conditional entropy, exist in the entropy literature, for example, as in Cover and Thomas (2006) or Golan et al (1996). However, to our knowledge, it is the first time to define these concepts for random models.

#### 6.3.3.1 Entropy

Imagine a world that is comprised of a finite number of states $s = 1, 2, \cdots, S$. In each state, the data generating process is described by a mechanism, called *model*. We denote $\mathcal{M}$ as a collection of such models, i.e., $\mathcal{M} = \{M_s : s = 1, 2, \cdots, S\}$, where $M_s$ describes the world in state $s$. Each state of the world, $s$, is associated with a probability $q_s$. We denote the probability space by the simplex $\triangle^S = \left\{ \mathbf{q} \in \mathbf{R}^S : q_s \geq 0, \sum_{s=1}^{S} q_s = 1 \right\}$.

**Definition 11** *Consider a class of models $\mathcal{M} = \{M_s : s = 1, 2, \cdots, S\}$, from which data are generated with probability distribution $\mathbf{q}(\mathcal{M}) = (q_1, q_2, \cdots, q_S)$. The **entropy** that characterizes the information uncertainty associated with $\mathcal{M}$ is defined as*

$$H(\mathbf{q}) = -\sum_{s=1}^{S} q_s \log q_s,$$

*where the convention $0 \cdot \log 0 = 0$ is taken.*

Here $\mathbf{q}(\mathcal{M}) = (q_1, q_2, \cdots, q_S)$ is the the probability mass function of models $M_1$, $M_2$,$\cdots$, $M_S$ that are contained in $\mathcal{M}$. It is abbreviated as $\mathbf{q}$ whenever no confusion occurs. As defined, $H(\mathbf{q})$ is a measure of the amount of uncertainty in the probability mass $\mathbf{q}(\mathcal{M})$ that describes the states of the world. It reaches a maximum when $q_s = 1/S$, for all $s = 1, 2, \cdots, S$, i.e., when the probability is uniform. This definition is consistent with entropy of a discrete random variable. See, for example, Cover and Thomas (2006) or Golan et al (1996) for more details.

Next, we extend the measure of uncertainty when there is an additional set of observations from the potential class of models $\mathcal{M}$. The following definition parallels that of joint entropy of two random variables. Let a random vector $D$ be defined on $\mathcal{D}$.

**Definition 12** *The **joint entropy** $H(\mathcal{M}, D)$ of the model class $\mathcal{M}$ and the random vector $D$ with a joint distribution $p(M, D)$ is defined as*

$$\begin{aligned} H(\mathcal{M}, D) &= -\sum_{M \in \mathcal{M}} \sum_{d \in \mathcal{D}} p(M, d) \log p(M, d) \\ &= -E \log p(\mathcal{M}, D). \end{aligned} \tag{6.10}$$

We further define the conditional entropy of a model class given a random vector as the expected value of the entropies of the conditional distributions averaged over the conditioning random vector.

**Definition 13** *The **conditional entropy** $H(\mathcal{M}|D)$ of the model class $\mathcal{M}$ given the random vector $D$ with a joint distribution $p(M, d)$ is defined as*

$$\begin{aligned} H(\mathcal{M}|D) &= \sum_{d \in \mathcal{D}} p(d) H(\mathcal{M}|D = d) \\ &= -\sum_{d \in \mathcal{D}} p(d) \sum_{M \in \mathcal{M}} p(M|d) \log p(M|D = d) \\ &= -\sum_{M \in \mathcal{M}} \sum_{d \in \mathcal{D}} p(M, d) \log p(M|d) \\ &= -E \log p(\mathcal{M}|D). \end{aligned} \tag{6.11}$$

Similarly, we can define the conditional entropy $H(D|\mathcal{M})$ of the random vector $D$ given the model class $\mathcal{M}$.

**Definition 14** *The **conditional entropy** $H(D|\mathcal{M})$ of the random vector $D$ given the model class $\mathcal{M}$ with a joint distribution $p(M, d)$ is defined as*

$$\begin{aligned} H(D|\mathcal{M}) &= \sum_{M \in \mathcal{M}} p(M) H(D|\mathcal{M} = M) \\ &= -\sum_{M \in \mathcal{M}} p(M) \sum_{M \in \mathcal{M}} p(d|M) \log p(d|\mathcal{M} = M) \\ &= -\sum_{d \in D} \sum_{M \in \mathcal{M}} p(M, d) \log p(d|M) \\ &= -E \log p(D|\mathcal{M}). \end{aligned} \tag{6.12}$$

The following theorem shows that the difference between the joint entropy defined in (6.10) and conditional entropy defined in (6.11) is the entropy of the conditioning random vector $D$. Similar result holds if we switch the model class and the random vector.

**Theorem 15** *(Chain rule)*

$$\begin{aligned} H(\mathcal{M}, D) &= H(\mathcal{M}|D) + H(D) \\ &= H(D|\mathcal{M}) + H(\mathcal{M}). \end{aligned}$$

Proof: The proof follows closely from that of Theorem 2.2.1 in Cover and Thomas (2006, p.17).

#### 6.3.3.2 eMAPLE estimator

Instead of approximating the expectation in (6.7) with a simple sample average as in GMM, we adopt

$$\sum_{i=1}^{n} p_{is} g_s(d_i, \theta_0) = 0$$

where $p_{is}$ is defined as probability of observing $d_i$ given that the model is $M_s$. That is, $p_{is} = p(d_i|M_s)$, $s = 1, \ldots, S$. Requirement of probability states that

$$p_{is} \geq 0 \text{ and } \sum_{i=1}^{n} p_{is} = 1, i = 1, \ldots, n, s = 1, \ldots, S.$$

For each parameter vector $\theta \in \mathbf{\Theta}$, define the set of probability measures:

$$\mathcal{P}(\theta) \equiv \left\{ \mathbf{p} = (p_1^\tau, \ldots, p_S^\tau)^\tau : \sum_{i=1}^{n} p_{is} g_s(d_i; \theta) = 0, \sum_{i=1}^{n} p_{is} = 1, p_s^\tau = (p_{is}) \geq 0, s = 1, \ldots, S. \right\}. \tag{6.13}$$

and

$$\mathcal{Q}(\theta) \equiv \left\{ \mathbf{q} = (q_1, \ldots, q_S)^\tau : \sum_{s=1}^{S} q_s = 1 \right\},$$

where $q_s = p(M_s)$, $s = 1, \ldots S$. To estimate the probabilities, it is natural to consider the following maximization problem,

$$\max_{[p^\tau, q^\tau]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} H(\mathcal{M}|D), \tag{6.14}$$

for each $\theta \in \mathbf{\Theta}$. That is, we simultaneously select the conditional probabilities, $p_{is}$, the probability of observing $d_i$ given model $M_s$, and the marginal probability of model $M_s$, $q_s$, to maximize the missing information between the class of model $\mathcal{M}$ and the observed data. This is the essential philosophy of maximum entropy econometrics.

To analyze directly the objective function in (6.14), ones needs to know the conditional entropy of a model class for a given data set. This is a conceptual challenge, since we need begin with the probability distribution of the model class for a given data set. This is exactly the difficulty researchers facing when model uncertainty presents, for example, in the Bayesian model averaging methods. However, we will circumvent this difficulty in entropy-based approach. We make use of the following theorem to rewrite the objective function.

**Theorem 16** *The solution in (6.14) solves*

$$\max_{[p^\tau, q^\tau]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} -\sum_{s=1}^{S}\sum_{i=1}^{n} q_s p_{is} \log p_{is} - \sum_{s=1}^{S} q_s \log q_s.$$

**Proof.** Note first that the joint entropy

$$\begin{aligned} H(\mathcal{M}, D) &= H(D|\mathcal{M}) + H(\mathcal{M}) \\ &= -\sum_{s=1}^{S}\sum_{i=1}^{n} q_s p_{is} \log p_{is} - \sum_{s=1}^{S} q_s \log q_s \end{aligned}$$

Therefore, we have

$$\begin{aligned}
[\hat{p}^\tau, \hat{q}^\tau]^\tau &= \arg \max_{[p^\tau, q^\tau]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} H(\mathcal{M}|D) \\
&= \arg \max_{[p^\tau, q^\tau]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} H(\mathcal{M}|D) + H(D) \\
&= \arg \max_{[p^\tau, q^\tau]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} H(D, \mathcal{M}) \\
&= \arg \max_{[p^\tau, q^\tau]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} H(D|\mathcal{M}) + H(\mathcal{M}) \\
&= \arg \max_{[p^\tau, q^\tau]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} -\sum_{s=1}^{S} \sum_{i=1}^{n} q_s p_{is} \log p_{is} - \sum_{s=1}^{S} q_s \log q_s.
\end{aligned}$$

This completes the proof. ■

Lagrange multipliers can be used to solve (6.14). The Lagrangian is

$$\begin{aligned}
\mathcal{L} &= -\sum_{s=1}^{S} \sum_{i=1}^{n} q_s p_{is} \log p_{is} - \sum_{s=1}^{S} q_s \log q_s - \mu \left[ \sum_{i=1}^{n} p_{is} - 1 \right] \\
&\quad - \sum_{s=1}^{S} \eta_s^\tau \sum_{i=1}^{n} g_s(d_i; \theta) p_{is} - \xi \left[ \sum_{s=1}^{S} q_s - 1 \right],
\end{aligned} \tag{6.15}$$

where $\mu$, $\eta_s^\tau$ and $\xi$ are Lagrange multipliers.

In the appendix, we show that the solution to (6.14) are

$$\hat{q}_s = \frac{1}{\sum_{s=1}^{S} \exp\left( -\sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is} \right)} \exp\left( -\sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is} \right), \tag{6.16}$$

and

$$\hat{p}_{is} = \frac{1}{\Upsilon_s(\lambda_s, \theta)} \exp\left[ -\lambda_s^\tau g_s(d_i; \theta) \right] \tag{6.17}$$

where

$$\Upsilon_s(\lambda_s, \theta) = \sum_{i=1}^{n} \exp\left[ -\lambda_s^\tau g_s(d_i; \theta) \right], \tag{6.18}$$

with $\lambda_s^\tau = \eta_s^\tau / \hat{q}_s$, $\lambda = (\lambda_1^\tau, \ldots, \lambda_S^\tau)^\tau$. In addition, each $\theta \in \mathbf{\Theta}$, $\hat{\lambda}_s^\tau$ solves

$$\sum_{i=1}^{n} g_s(d_i; \theta) \exp\left[ -\hat{\lambda}_s^\tau g_s(d_i; \theta) \right] = 0, \tag{6.19}$$

for all $s = 1, \ldots, S$.

We define the profile joint entropy (JE) at $\theta$ as

$$\begin{aligned} \mathrm{JE}(\theta) &= -\sum_{s=1}^{S}\sum_{i=1}^{n}\hat{q}_s\hat{p}_{is}\log\hat{p}_{is} - \sum_{s=1}^{S}\hat{q}_s\log\hat{q}_s \\ &= \log\Upsilon(\lambda,\theta), \end{aligned} \tag{6.20}$$

where the last equality is show in the Appendix, with

$$\Upsilon(\lambda,\theta) = \sum_{s=1}^{S}\Upsilon_s(\lambda_s,\theta) = \sum_{s=1}^{S}\sum_{i=1}^{n}\exp\left[-\lambda_s^{\tau}g_s(d_i;\theta)\right],$$

and $\lambda = (\lambda_1^{\tau},\ldots,\lambda_S^{\tau})^{\tau}$.

Our **e**ntropy-based **m**odel **a**veraging **p**artia**l** **e**ffect (eMAPLE) estimator of $\theta$ is thus defined as

$$\begin{aligned} \hat{\theta}_{eMAPLE} &= \arg\max_{\theta\in\boldsymbol{\Theta}}\mathrm{JE}(\theta) \\ &= \arg\max_{\theta\in\boldsymbol{\Theta}}\frac{1}{nS}\exp\{JE(\theta)\} \\ &= \arg\max_{\theta\in\boldsymbol{\Theta}}\frac{1}{nS}\Upsilon(\lambda,\theta) \\ &= \arg\max_{\theta\in\boldsymbol{\Theta}}\mathrm{JE}_n(\theta), \end{aligned} \tag{6.21}$$

where

$$\mathrm{JE}_n(\theta) = \frac{1}{nS}\Upsilon(\lambda,\theta) = \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n}\exp\left[-\lambda_s^{\tau}g_s(d_i;\theta)\right]$$

To implement our estimator, it's easily seen that the $\lambda_s^{\tau}$ solving (6.19) can be alternatively found as

$$\lambda_s^{\tau} = \arg\max_{\varsigma\in R^{\dim(g_s(x,\theta))}}\Upsilon_s(\varsigma,\theta),$$

Note that this is a well-defined finite dimensional unconstrained convex maximization problem that has a unique solution. Algorithms such as Newton-Raphson method can be easily applied. Once $\lambda_s^{\tau}$ is solved as a function of $\theta$, it can be substituted to (6.17) and (6.16), and consequently, (6.21) can be solved easily through numerical methods.

### 6.3.4 Alternative methods

An alternative is to optimally combine the estimators of parameters common to all models via an artificial regression (White and Lu, 2010). Denote the GLS estimator of $\beta_0$ (parameter of interest) from model $s$ as $\tilde{\beta}_s$ (we suppress the dependence on sample size $n$) and that of $\theta_0$ as $\tilde{\theta}_s$. Denote $\tilde{\beta}_n = \left[\tilde{\beta}_1^\tau, \ldots, \tilde{\beta}_S^\tau\right]^\tau$, $\tilde{\theta}_n = \left[\tilde{\theta}_1^\tau, \ldots, \tilde{\theta}_S^\tau\right]^\tau$ and $\Lambda$ a selection matrix such that $\tilde{\beta}_n = \Lambda\tilde{\theta}_n$. A combined estimator of $\beta_0$ can be formulated through the following regression,

$$\sqrt{n}\tilde{\beta}_n = \sqrt{n}\mathcal{I}\beta_0 + e, \tag{6.22}$$

where the $S\dim(\beta_0) \times S\dim(\beta_0)$ matrix of artificial regressor $\mathcal{I} \equiv \iota \otimes I_{\dim(\beta_0)}$, with $\iota$ being the $S\times 1$ vector of ones and $I_{\dim(\beta_0)}$ being the identity matrix of the same dimension as $\beta_0$, $e \sim N(0, \Sigma^*)$ is the artificial regression error with $\Sigma^* = \left((\Lambda G)^\tau \Omega^{-1}\Lambda G\right)^{-1}$. The Feasible Optimally combined GLS (FOGLeSs) estimator of White and Lu (2010) is defined as the FGLS estimator of (6.22):

$$\tilde{\beta}_n^* = \left(\mathcal{I}^\tau\hat{\Sigma}^{*-1}\mathcal{I}\right)^{-1}\mathcal{I}^\tau\hat{\Sigma}^{*-1}\tilde{\beta}_n, \tag{6.23}$$

where $\hat{\Sigma}^*$ is a consistent estimator of $\Sigma^*$ and satisfies

$$\sqrt{n}\left(\tilde{\beta}_n^* - \beta_0\right) \xrightarrow{d} N\left(0, \left(\mathcal{I}^\tau\Sigma^{*-1}\mathcal{I}\right)^{-1}\right).$$

Note that $\left(\mathcal{I}^\tau\Sigma^{*-1}\mathcal{I}\right)^{-1} = \left(\mathcal{I}^\tau\left((\Lambda G)^\tau\Omega^{-1}\Lambda G\right)\mathcal{I}\right)^{-1} \geq \Lambda\left(G^\tau\Omega^{-1}G\right)^{-1}\Lambda^\tau$. FOGLeSs estimator is not as efficient as optimal GMM estimator. We emphasize that, to implement FOGLeSs, $\hat{\Sigma}^*$ is needed to compute $\tilde{\beta}_n^*$ in (6.23).

However, it is important to explore our proposed entropy-based estimation approach for the following reasons. First, the above two-step estimation procedures require the first step consistent estimator of $\theta_0$, which will be used for the estimation

of the optimal weighting matrix. Inevitably, this would introduce finite sample bias for the second stage estimation. Second, the weighing matrix (either is of large dimension whose accuracy is more than often a concern when available data sample size is small.

## 6.4 Theoretical Properties

In this section, we present the theoretical properties of the eMAPLE estimator.

### 6.4.1 Consistency

We introduce some notations before stating the needed assumptions. Define $\nabla_\theta g_s(d,\theta) = \partial g_s^\tau(d,\theta)/\partial\theta$, where $\partial g_s^\tau(d,\theta)/\partial\theta$ is the transpose of $\partial g_s(d,\theta)/\partial\theta$. Denote $G(s,\theta) = E[\nabla_\theta g_s(d,\theta)]$ and $V(s,\theta) = E[g_s(d,\theta)\, g_s^\tau(d,\theta)]$.

**Assumption B.1** For each $\theta \neq \theta_0$ there exists a sub class $\mathcal{M}_\theta \subseteq \mathcal{M}$ such that $\Pr(M_s \in \mathcal{M}_\theta) > 0$, and $E\{g_s(d,\theta)\} \neq 0$ for each $M_s \in \mathcal{M}_\theta$.

**Assumption B.2** $E\{\sup_{\theta\in\mathbf{\Theta}} ||g_s(d,\theta)||^m\} < \infty$ for some $m \geq 8$, for all $s = 1,\ldots,S$.

**Assumption B.3** For all $s = 1,\ldots,S$,

**(i)** $\theta_m \to \theta \in \mathbf{\Theta} \Longrightarrow g_s(d,\theta_m) \to g_s(d,\theta)$ for almost every $d$;

**(ii)** $E[\sup_{\theta\in\mathbf{\Theta}} \| \partial g_s(d,\theta)/\partial\theta' \|] < \infty$, for all $s = 1,\ldots,S$.

**(iii)** $\sup_{\theta\in\mathcal{B}} \left|\partial g_s^{(i)}(d,\theta)/\partial\theta^{(j)}\right| \leq r(d), \sup_{\theta\in\mathcal{B}} \left|\partial^2 g_s^{(i)}(d,\theta)/\partial\theta^{(j)}\partial\theta^{(k)}\right| \leq t(d),\ w.p.1$ for some real valued functions $r(d)$ and $t(d)$ such that $Ed^\upsilon < \infty$ for some $\upsilon \geq 4$ and $Et(d) < \infty$.

**Assumption B.4** There is a closed ball around $\theta_0$, $\mathcal{B}$, such that for all $s = 1,\ldots,S$

**(i)** $G(s,\cdot)$ and $V(s,\cdot)$ are continuous w.p.1. on $\mathcal{B}$.

**(ii)** $\inf_{(\varsigma,s,\theta)} \varsigma' V(s,\theta)\varsigma > 0$ and $\sup_{(\varsigma,d,\theta)} \varsigma' V(s,\theta)\varsigma < \infty$ with $\theta \in \mathcal{B}$.

**Assumption B.5** $\lambda_s \in \left\{\gamma : \|\gamma\| \leq an^{-1/m}\right\}$ for some $a > 0$ and $m$ as in Assumption B.2.

**Remark:** Similar assumptions to the above ones are adopted in the EL literature (e.g., Kitamura, Tripathi and Ahn (2004)). Assumptions B.1 states that $\theta_0$ is identified jointly in all $S$ models. Assumption B.2 is needed to prove a Lemma C.1 in line with Lemma 3 of Owen (1990) or Lemma D.2 of Kitamura, Tripathi and Ahn (2004). Assumption B.3 impose regularity conditions on the moment function. Assumption B.4 imposes conditions on the first derivative of the moment condition and the variance-covariance matrix. Assumption B.5 is a technical assumption that leads to the asymptotic normality of e-MAPLE estimator.

**Theorem 17** *(consistency) Under Assumption A.1-3, B.1-4, e-MAPLE estimator is consistent, i.e.,* $\hat{\theta}_{eMAPLE} \to^p \theta_0$.

**Proof**: See the Appendix.

**Remark**: Theorem 4.1 shows that e-MAPLE estimator is consistent. The consistency comes as a result of the moment conditions that identify $\theta_0$ (that includes $\beta$). We emphasize that $\gamma_s$ in $\theta_0$ is pseudo projection coefficient vector in model $s$ and does not carry any economic interpretation.

### 6.4.2 Asymptotic Normality

**Theorem 18** *(asymptotic normality) Under additional Assumption B.5,*

$$\sqrt{n}\left[\hat{\theta}_{eMAPLE}-\theta_0\right] \xrightarrow{d} N\left(0, J^{-1}\left(\theta_0\right) I\left(\theta_0\right) J^{-1}\left(\theta_0\right)\right),$$

*where* $J\left(\theta_0\right)=G^{\tau}V^{-1}G$, $I\left(\theta_0\right)=G^{\tau}V^{-1}\Omega V^{-1}G$.

**Proof**: See the Appendix.

**Remark**: The theorem shows that, when there is no correlation among models in different states, i.e., $\Omega = V$, our e-MAPLE estimator $\hat{\theta}_{eMAPLE}$ is asymptotically efficient and it achieves the asymptotic variance lower bound, $G^{\tau}\Omega^{-1}G$, which is the expectation of the inverse of Fisher information matrix averaged across states of the world. Note that this lower bound agrees the variance of optimally weighted GMM estimator, where the optimal weight is used for each individual model. When $\Omega \neq V$, i.e., there is correlation among models in different states, the e-MAPLE estimator agrees with the GMM estimator that adopts weighting matrix $W = V$. This is suboptimal in the sense that it efficiently uses information of in each model only. As pointed out earlier, e-MAPLE estimator avoids the estimation of the large dimensional variance-covariance matrix, which makes it appealing in finite sample.

### 6.4.3 Hypothesis Testing

To construct tests of the possible nonlinear restrictions as follows:

$$H_0 : R\left(\theta_0\right) = r, \tag{6.24}$$

where $r$ is a $k \leq m$ dimensional vector of constants and $R\left(\cdot\right)$ is a known parametric function. Impose this restriction in the optimization of (6.21). Denote the constrained

solution by $\hat{\theta}^c$ and the Jacobian matrix of $R$ evaluated at $\theta_0$ as $A$, which is assumed to be of full row rank. We have the following theorem for the Wald, Rao's Score, and Likelihood Ratio-like test statistics.

**Theorem 19** *Test statistics of the restrictions (6.24),*

$$\begin{aligned} Wald_n &= n\left[R\left(\hat{\theta}\right)-r\right]'\left[A\hat{I}^{-1}\left(\theta_0\right)A\right]^{-1}\left[R\left(\hat{\theta}\right)-r\right], \\ LM_n &= ng\left(d,\hat{\theta}^c\right)\hat{V}^{-1}\hat{G}\hat{I}^{-1}\left(\theta_0\right)\hat{G}'\hat{V}^{-1}g\left(d,\hat{\theta}^c\right), \\ LR_n &= 2\left[JE_n\left(\hat{\theta}\right)-JE_n\left(\hat{\theta}^c\right)\right] \end{aligned}$$

*are asymptotically* $\chi^2_k$*, where* $\hat{V},\hat{G},\hat{I}^{-1}\left(\theta_0\right)$*,* $\hat{A}$*, are consistent estimates of* $V,G,I\left(\theta_0\right)$ *and* $A$*.*

**Proof**: The results follows from Theorem 3 and Amemiya (1985).

**Remark**: Tests based on g-MAPLE estimators can be similarly constructed, without any difficulty. However, in practice, we recommend the $LR_n$ test based on e-MAPLE estimator due to its easy implementation and nice finite sample properties as to be shown in the next section.

## 6.5 Finite Sample Investigations

In this section, we conduct simulation studies to examine the finite sample properties of e-MAPLE estimator, with a comparison to other estimators available in the literature. We include the ordinary least square estimator, Generalized Least Square (GLS) estimator with perfect knowledge of heterogeneity function, Feasible GLS with knowledge of heterogeneity functional form,1-step GMM (GMM1) estimator, 2-step optimal GMM (GMM2) estimator, the FOGLeSs estimator of White and Lu (2010), LASSO

estimator of Tibshirani (1996), the factor based estimator of Galbraith etc (2010) and the Mallows model averaging (MMA) estimator of Hansen (2007).

We perform sequentially five experiments for the investigation. We briefly describe our experiments before presenting the details. The first experiment is to study the performance in a factor model setting. The second experiment is look into the classical regression model with a large number of covariates. The third experiment is to amplify the role of efficient estimation in the presence of heterogeneous errors. The next experiment is to investigate the effects of the irrelevant covariates with homogeneous disturbance, which is replaced by heterogeneity in the last experiment. For all experiments considered, we consider sample size $n = 50, 100, 150, 200, 250$ and replicate the process 1,000 times. The covariates are kept fixed for each replication. The estimator of the parameter of interest, $\beta$, is the partial effect of $x$ on $y$. We report different criteria to evaluate estimators under investigation, including the Mean Squared Error (MSE), the Mean Absolute Error (MAE), Squared Bias (Bias$^2$), Variance (Var) and Inter Quantile Range (IQR) over 1,000 replications.

### 6.5.1 Experiment 1: factor model

We first consider a factor model, in which all the observed covariates are generated from some underlying factors $f_i$, according to the following DGP.

$$\text{DGP1}: \begin{cases} y_i = x_i^\tau \beta + z_i' \gamma + e_{1i}, \\ Z_i = f_i^\tau \xi_z + e_{2i}, \\ x_i = f_i^\tau \xi_x + e_{3i}, \end{cases}$$

$i = 1, ..., n$. We consider $n_f = \dim(f_i) = 3$ and generate $f_i$ from a multivariate normal distribution with random mean vector and random covariance matrix. $p = \dim(z_i) =$

$0.8n-2$, $\xi_z$ is generated in a similar way as $f_i$ but is normalized to a unit vector. $\xi_x$ is generated from uniform $U[0,3]$. We set $\beta=[2,3]'$, and $\gamma$ is generated from $U[0,0.3]$. $e_{ji}, j=1,2,3$, is independent standard normal error.

### 6.5.2 Experiment 2: regression model, case 1

We next consider the classical regression model that has a large dimensional observed covariates.

$$\text{DGP2}: \quad y_i = x_i^\tau \beta + z_i'\gamma + e_i,$$

where $x_i$ and $z_i$ are generated from independent standard normal distribution. $x_i$ and $z_i$ of the same dimension as those in DGP 1, and so are the values of $\beta$ and $\gamma$. $e_i$ is the independent standard normal error.

### 6.5.3 Experiment 3: regression model, case 2

While heterogeneity is more often the case than exception in economic data, we incorporate such a feature into DGP3.

$$\text{DGP3}: \quad y_i = x_i^\tau \beta + z_i'\gamma + v_i^m \cdot e_i,$$

We generate $x_i$, $z_i$, $\beta$, $\gamma$ and $e_i$ in the same way as in DGP2. We consider heterogeneity function $v_i^m$ for three different forms

$$\begin{aligned} v_i^1 &= \sqrt{x_{1i}^2 + x_{1i}^2} \\ v_i^2 &= \sqrt{x_{1i}^2 + 2x_{1i}^2} \\ v_i^3 &= \exp\left(-x_{1i}^2\right) \end{aligned}$$

### 6.5.4 Experiment 4: regression model, case 3

Note that in earlier experiments, the large dimensional covariate vector $(x_i, z_i)$ are causal in generating the dependent variable $y$. We next consider the case in which a large dimensional irrelevant covariates are available.

$$\text{DGP4}: \quad y_i = x_i^\tau \beta + z_{1i}' \gamma + e_i,$$

where $Z_{1i}$ is a subset of $Z_i$. $x_i$, $Z_i$, $\beta$, $\gamma$ and $e_i$ are generated in the same way as in DGP2.

### 6.5.5 Experiment 5: regression model, case 4

We further consider effects of heterogeneity in error term.

$$\text{DGP5}: \quad y_i = x_i^\tau \beta + z_{1i}' \gamma + u_i^m \cdot e_i,$$

where we consider three different form of heterogeneity,

$$\begin{aligned} u_i^1 &= \sqrt{x_{1i}^2 + 2x_{1i}^2}, \\ u_i^2 &= \sqrt{x_{1i}^2 + 5x_{1i}^2}, \\ u_i^3 &= \sqrt{x_{1i}^2 + 5x_{1i}^2 + 2\cos(x_{1i}x_{2i})}. \end{aligned}$$

Table 4-21 present the simulation results for experiment 1-5. To save space, we only report the squared bias and MSE for the estimators considered, with sample size 50 and 200. Other simulation results resemble and are available from the author upon requests. The findings are summarized as follows.

1. In experiment 1, factor estimator and MMA estimator suffer from huge bias. This leads to its bad performance in MSE. FOGLeSs, GMM and eMAPLE estimator

incur a small bias but enjoy a big reduction in variance, as shown as their MSE are much smaller than GLS estimator. Our proposed estimators are generally better than FOGLeSs estimator. Lasso is very attractive in small sample, due to the correlation in the factors and the regressors. However, it is much worse than MAPLE estimators when $n = 200$.

2. In experiment 2, MMA, JMA and LASSO estimators perform pretty well and even beat the oracle GLS estimator. The advantage of MAPLE estimator becomes clear when sample size is $n = 200$. MAPLE outperforms FOGLeSs estimator in all cases.

3. In experiment 3, when heterogeneity presents, the performance of the estimators considered is quite mixed. A smaller MSE of one parameter estimator is usually glued with a larger MSE of the other parameter estimator. However, in the third heterogeneity case, MAPLE estimators outperform others in small sample.

4. In experiment 4, we see the clear dominance of MAPLE estimators over other competing ones. Especially, eMAPLE estimator is performing as if it is the oracle GLS estimator in terms of MSE. All competitors perform quite close to GLS. LASSO becomes the worst among all methods.

5. In experiment 5, the dominance of MAPLE estimator remains when heterogeneity exists. Although they are not as good as the oracle GLS, but they are quite competing with the FGLS. LASSO remains the worst among all competitor, but perform slightly better than the naive OLS estimator.

### 6.5.6 Experiment 6: rejection probability

We consider to evaluate the size of the tests based on different estimators. We include GLS, FOGLeSs, GMM and our eMAPLE estimators. For tests based on

eMAPLE estimator, we only include the LR type test that is appealing in its computation. Other tests are based on the usual t-test statistic. We consider the following data generating process.

$$\text{DGP6-1} : y_i = 2x_i + z_{1i}'\gamma + e_i,$$

where $z_{1i}$ is generated in the same way as that in experiment 4. We report the results based on 1000 replications for sample size n=50 and 200.

Table 6.1: Rejection Probability: Homogeneous Case

| | $n = 50$ | | | $n = 200$ | | |
|---|---|---|---|---|---|---|
| $\alpha$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| OLS | 0.259 | 0.389 | 0.468 | 0.246 | 0.374 | 0.456 |
| GLS | 0.023 | 0.076 | 0.133 | 0.012 | 0.061 | 0.115 |
| FOGLeSs | 0.081 | 0.179 | 0.252 | 0.015 | 0.077 | 0.148 |
| gMAPLE1 | 0.034 | 0.079 | 0.152 | 0.013 | 0.055 | 0.116 |
| gMAPLE2 | 0.049 | 0.100 | 0.176 | 0.014 | 0.061 | 0.119 |
| eMAPLE | 0.027 | 0.075 | 0.132 | 0.013 | 0.053 | 0.115 |

We consider to evaluate the size of the tests. Table 1 presents the rejection probability when there is no heterogeneity. Test based on eMAPLE estimator tends to outperform all other competitors, including that based on oracle GLS estimator. Its rejection probabilities are very close to their nominal levels for both sample sizes. FOGLeSs estimator suffers from big size distortion.

To incorporate heterogeneity, we consider the following design,

$$\text{DGP6-2}: \quad y_i = x_i^\tau \beta + z_{1i}' \gamma + u_i \cdot e_i,$$

with

$$u_i = \log\left(3x_i^2\right).$$

Table 6.2: Rejection Probability: Heterogeneous Case

| | $n = 50$ | | | $n = 200$ | | |
|---|---|---|---|---|---|---|
| $\alpha$ | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| OLS | 0.240 | 0.366 | 0.455 | 0.249 | 0.386 | 0.463 |
| GLS | 0.017 | 0.068 | 0.126 | 0.013 | 0.040 | 0.086 |
| FOGLeSs | 0.124 | 0.235 | 0.306 | 0.030 | 0.090 | 0.153 |
| gMAPLE1 | 0.034 | 0.103 | 0.166 | 0.013 | 0.057 | 0.111 |
| gMAPLE2 | 0.044 | 0.125 | 0.194 | 0.015 | 0.052 | 0.114 |
| eMAPLE | 0.021 | 0.094 | 0.159 | 0.012 | 0.053 | 0.114 |

The performance of eMAPLE estimator remains satisfactory in the presence of heterogeneity. Although there is a distortion when sample size is 50, it beats GLS estimator when sample size becomes 200. FOGLeSs perform slightly better with heterogeneity, but still have serious size distortion.

## 6.6 Empirical Illustration

We illustrate the use of MAPLE estimator in the study of the impact of inherited control on firm performance. We adopt the data set that is originally analyzed

by Pérez-González(2006) and subsequently examined by White and Lu (2010). Pérez-Gonzálezuses data from 335 management transitions of publicly traded U.S. corporations to examine whether firms with family related incoming chief executive officers (CEOs) underperform in terms of operating profitability relative to firms with unrelated incoming CEOs. In this application, $x$ equals to 1 if the incoming CEO is related to the departing CEO, to the founder, or to a large shareholder by blood or marriage and otherwise it equals to 0. Operating return on assets (OROA) is used as a measure of firm performance. $y$ is the difference in OROA calculated as the three-year average after succession minus the three-year average before succession. We direct detailed data description to White and Lu (2010).

Following White and Lu (2010), we classify the covariates into firm size, firm's past performance, board characteristics, firm's R&D expenditure, departing CEO's separation conditions and incoming CEO's ownership, and incoming CEO's characteristics. We follow White and Lu (2010) to consider 5 model specifications that correspond to 5 states of the world. We report the estimated weights and e-MAPLE estimate in TABLE 7, together with associated the t-statistic. In TABLE 3, we include the estimator of White and Lu (2010) for comparison.

Table 6.3: Empirical results: Inherited control

| | FOGLeSs | gMPL 1 | gMPL2 | eMPL |
|---|---|---|---|---|
| Estimate | -0.0246 | -0.0283 | -0.0283 | -0.0221 |
| 95% C.I. | (-0.04409,-0.00510) | (-0.04805,-0.00862) | (-0.04606,-0.01057) | (-0.03300,-0.01200) |
| 95% C.I. length | 0.03899 | 0.03943 | 0.03550 | 0.02100 |
| eMAPLE model probability | | | | |
| 0.1996 | 0.2001 | 0.2001 | 0.2003 | 0.1999 |

We find that all the estimates are negative and all 95% confidence intervals are to the left of zero. The implication is that the effect of inherited control on firm performance is significant. This agree with the findings of White and Lu (2010) and Pérez-González(2006). A second finding from Table 3 is that confidence interval based on our eMAPLE estimator is much narrower than those based on FOGLeSs and gMAPLE estimators. Combined with findings in our simulation results, the eMAPLE estimator provides more accurate inference analysis.

## 6.7 Conclusion and Future Work

This chapter studies the estimation of marginal effect of one economic variable on another, in the presence of large amount of other economic variables. The chapter first points out that only small dimensional partial effect parameters have economic policy implication and therefore are economically sensible. Then we set up conditions to identify partial effect parameter of interest in high dimensional structural model. Based on identification of the parameter of interest, we consider the case that the partial effect parameter may be identified in more than one model. I propose two new model averaging estimator to estimate the partial effect estimator based on a GMM-like objective function and an entropy objective function. The two estimators are termed as gMAPLE and eMAPLE estimators. Asymptotic properties of MAPLE estimators are established under a suitable set of conditions. Simulation results show that the MAPLE estimator outperform other competitors in finite sample. An application of the MAPLE estimator to study the effect of inherited control on firm's performance is carried out to illustrate its use. We found that a negative effect does exist which is consistent with earlier find-

ings in the literature. The gain in using MAPLE estimator compared to FOGLeSs is revealed through the shorter confidence interval length.

This chapter opens directions for future studies in model averaging in numerous ways. It emphasizes the estimation of parameter of interest in large dimensional model via identification conditions and model averaging techniques. An information based test of the key identification condition, conditional mean independence, is under investigation by the author. A second direction is to apply MAPLE to study the determinants of economic growth following the work of Sala-i-Martin et al (2004). MAPLE can also be extended to the nonparametric and semiparametric models. the only challenge is the identification condition. As an alternative to entropy based approach, empirical likelihood (Owen 1988, 1990, 1991) based approach can be used for MAPLE as well. Moreover, information based variable selection and estimation is another direction to extend the current chapter.

# Appendix

## Proof of Lemmas

**Proof of Lemma 7.** Partition the coefficient of $\mathbf{z}_i$ as $\gamma_i^\tau = [\gamma_{1i}^\tau, \gamma_{2i}^\tau]$ corresponding to the partition of $\mathbf{z}_i^\tau = \left[\mathbf{z}_i^{1\tau}, \mathbf{z}_i^{2\tau}\right]$ .Under Assumption CMI,

$$\begin{aligned} E\left[y_i|\mathbf{x}_i, \mathbf{z}_{1i}\right] &= \alpha + \mathbf{x}_i\beta + \mathbf{z}_{1i}^\tau\gamma_{1i} + E\left[\mathbf{z}_{2i}^\tau\gamma_{2i}|\mathbf{x}_i, \mathbf{z}_{1i}\right] + E\left[\varepsilon_i|\mathbf{x}_i, \mathbf{z}_{1i}\right] \\ &= \alpha + \mathbf{x}_i\beta + \mathbf{z}_{1i}^\tau\gamma_{1i} + E\left[\mathbf{z}_{2i}^\tau|\mathbf{z}_{1i}\right]\gamma_{2i} \end{aligned} \tag{6.25}$$

where $E\left[\varepsilon_i|\mathbf{x}_i, \mathbf{z}_{1i}\right] = 0$ due to strict exogeneity of the regressors. Note that (6.25) implies that $\beta$ is identified in the regression of $y_i$ on $\mathbf{x}_i$ and $\mathbf{z}_{1i}$, with conditions as specified in Robinson (1988).

**Proof of Lemma 8.** Similar to the proof of Lemma 1, under Assumption WCMI, we can derive that

$$E\left[y_i|\mathbf{x}_i, \mathbf{z}_{1i}\right] = \alpha + \mathbf{x}_i\left(\beta + \eta\right) + \mathbf{z}_{1i}^\tau\gamma_{1i} + E\left[\mathbf{z}_{2i}^\tau|\mathbf{z}_{1i}\right]\gamma_{2i}.$$

Thus $(\beta + \eta)$ would be identified in the regression of $y_i$ on $\mathbf{x}_i$ and $\mathbf{z}_{1i}$, with conditions as specified in Robinson (1988). Since $||\eta|| = o\left(n^{-1/2}\right)$, with sample size gets large, Robinson's (1988) estimator of $(\beta + \eta)$ will converge to $\beta$.

## Derivation of Some Equations

This Appendix provides derivation of equation (6.16), (6.17), (6.19), (6.20).

The FOCs of the Lagrangian in (6.15) are:

$$\frac{\partial L}{\partial p_{is}} = -\hat{q}_s \log \hat{p}_{is} - \hat{q}_s - \hat{\mu}_s - \hat{\eta}_s^\tau g_s\left(d_i, \theta\right) = 0, \tag{6.26}$$

$$\frac{\partial L}{\partial q_s} = -\sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is} - 1 - \log \hat{q}_s - \hat{\xi} = 0, \tag{6.27}$$

$$\frac{\partial L}{\partial \mu_s} = \sum_{i=1}^{n} \hat{p}_{is} - 1 = 0, \tag{6.28}$$

$$\frac{\partial L}{\partial \eta_s} = \sum_{i=1}^{n} \hat{p}_{is} g_s\left(d_i; \hat{\theta}\right) = 0, \tag{6.29}$$

$$\frac{\partial L}{\partial \xi} = \sum_{s=1}^{S} \hat{q}_s - 1 = 0, \tag{6.30}$$

$$\frac{\partial L}{\partial \theta} = \sum_{s=1}^{S} \hat{\eta}_s^\tau \sum_{i=1}^{n} \hat{p}_{is} \nabla_\theta g_s\left(d_i, \hat{\theta}\right) = 0. \tag{6.31}$$

**(i)** Derivation of (6.16). From (6.27), we get

$$\hat{q}_s = \exp\left(-\sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is} - 1 - \hat{\xi}\right).$$

Combined this equation with (6.30), it gives (6.16).

**(ii)** Derivation of (6.17). Using (6.26), It's straightforward to show that

$$\hat{p}_{is} = \exp\left(-\frac{\hat{q}_s - \hat{\mu}_s - \hat{\eta}_s^\tau g_s\left(d_i, \theta\right)}{\hat{q}_s}\right),$$

With normalization in (6.28), we have

$$\begin{aligned}
\hat{p}_{is} &= \frac{\exp\left(\frac{-\hat{q}_s - \hat{\mu}_s - \hat{\eta}_s^\tau g_s(d_i,\theta)}{\hat{q}_s}\right)}{\sum_{i=1}^{n} \exp\left(\frac{-\hat{q}_s - \hat{\mu}_s - \hat{\eta}_s^\tau g_s(d_i,\theta)}{\hat{q}_s}\right)} \\
&= \frac{\exp\left(\frac{-\hat{\eta}_s^\tau g_s(d_i,\theta)}{\hat{q}_s}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{-\hat{\eta}_s^\tau g_s(d_i,\theta)}{\hat{q}_s}\right)} \\
&= \frac{1}{\Upsilon_s\left(\lambda_s,\theta\right)} \exp\left[-\lambda_s^\tau g_s\left(d_i;\theta\right)\right],
\end{aligned}$$

with $\lambda_s^\tau = \hat{\eta}_s^\tau / \hat{q}_s$, and $\Upsilon_s\left(\lambda_s,\theta\right) = \sum_{i=1}^{n} \exp\left[-\lambda_s^\tau g_s\left(d_i;\theta\right)\right]$. This proves (6.17).

**(iii)** Derivation of (6.19). Plugging (6.17) into (6.29) results

$$\sum_{i=1}^{n} \frac{g_s\left(d_i;\theta\right)}{\Upsilon_s\left(\lambda_s,\theta\right)} \exp\left[-\lambda_s^\tau g_s\left(d_i;\theta\right)\right] = 0.$$

Since $\Upsilon_s(\lambda_s,\theta) > 0$, this leads to (6.19).

**(iv)** Derivation of (6.20). We show this results in two steps. (a) Note that

$$\begin{aligned}
-\sum_{s=1}^{S}\hat{q}_s\log\hat{q}_s &= -\sum_{s=1}^{S}\hat{q}_s\log\left[\frac{\exp\left(-\sum\limits_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}\right)}{\sum\limits_{s=1}^{S}\exp\left(-\sum\limits_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}\right)}\right] \\
&= \sum_{i=1}^{n}\hat{q}_s\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is} \\
&\quad +\sum_{s=1}^{S}\hat{q}_s\log\sum_{s=1}^{S}\exp\left(-\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}\right) \\
&= \sum_{i=1}^{n}\hat{q}_s\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}+\log\sum_{s=1}^{S}\exp\left(-\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}\right).
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{JE}(\theta) &= -\sum_{s=1}^{S}\sum_{i=1}^{n}\hat{q}_s\hat{p}_{is}\log\hat{p}_{is}-\sum_{s=1}^{S}\hat{q}_s\log\hat{q}_s \\
&= -\sum_{s=1}^{S}\sum_{i=1}^{n}\hat{q}_s\hat{p}_{is}\log\hat{p}_{is}+\sum_{i=1}^{n}\hat{q}_s\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}+\log\sum_{s=1}^{S}\exp\left(-\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}\right) \\
&= \log\sum_{s=1}^{S}\exp\left(-\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}\right).
\end{aligned}$$

(b) Next,

$$\begin{aligned}
&-\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is} \\
&= -\sum_{i=1}^{n}\hat{p}_{is}\log\frac{\exp\left[-\lambda_s^{\tau}g_s(d_i;\theta)\right]}{\Upsilon_s(\lambda_s,\theta)} \\
&= \lambda_s^{\tau}\sum_{i=1}^{n}\hat{p}_{is}g_s(d_i;\theta)+\sum_{i=1}^{n}\hat{p}_{is}\log\Upsilon_s(\lambda_s,\theta) \\
&= \log\Upsilon_s(\lambda_s,\theta),
\end{aligned}$$

where we have used (6.29).

Putting (a) and (b) together leads to

$$\mathrm{JE}(\theta)=\log\sum_{s=1}^{S}\Upsilon_s(\lambda_s,\theta)=\log\Upsilon(\lambda,\theta),$$

which proves (6.20).

## Proof of Auxiliary Lemmas

**Lemma C.1** Under Assumption B.1-5, $\sup_{\theta\in\mathbf{\Theta},s=1,...,S,d_i}\left|\lambda_s^\tau g_s\left(d_i;\theta\right)\right|=o_p\left(1\right)$.

**Proof.** It follows from Lemma 3 of Owen (1990) or Lemma D.2 of Kitamura, Tripathi and Ahn (2004).

**Lemma C.2** Under Assumption B.1-5, $\sup_{\theta\in\mathbf{\Theta}}\left|\left|\lambda_s^\tau\left(\theta\right)-V^{-1}\left(s,\theta\right)Eg_s\left(d,\theta\right)\right|\right|=o_p\left(\left\|\lambda_s\right\|\right)$.

**Proof.** By (6.19), $\lambda_s^\tau$ solves

$$\sum_{i=1}^{n}g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^\tau g_s\left(d_i;\theta\right)\right]=0.$$

By Taylor's Theorem, there exists $\bar{\lambda}_s$ lying between 0 and $\lambda_s$ such that

$$0=\sum_{i=1}^{n}g_s\left(d_i;\theta\right)\left\{1-\lambda_s^\tau g_s\left(d_i;\theta\right)+\frac{\left(\bar{\lambda}_s^\tau g_s\left(d_i;\theta\right)\right)^2}{2}\right\}.$$

Rearranging terms leads to

$$\begin{aligned}\lambda_s &= \left[\frac{1}{n}\sum_{i=1}^{n}g_s\left(d_i;\theta\right)g_s^\tau\left(d_i;\theta\right)\right]^{-1}\sum_{i=1}^{n}g_s\left(d_i;\theta\right)/n\\ &\quad+\left[\frac{1}{n}\sum_{i=1}^{n}g_s\left(d_i;\theta\right)g_s^\tau\left(d_i;\theta\right)\right]\sum_{i=1}^{n}g_s\left(d_i;\theta\right)\frac{\left(\bar{\lambda}_s^\tau g_s\left(d_i;\theta\right)\right)^2}{2n}\\ &\equiv l_1+l_2,\end{aligned}$$

where

$$\begin{aligned}l_1 &= V^{-1}\left(s,\theta\right)Eg_s\left(d,\theta\right)+\left[\hat{V}^{-1}\left(s,\theta\right)-V^{-1}\left(s,\theta\right)\right]Eg_s\left(d,\theta\right)\\ &\quad+\hat{V}^{-1}\left(s,\theta\right)\left[\sum_{i=1}^{n}g_s\left(d_i;\theta\right)/n-Eg_s\left(d,\theta\right)\right]\\ &= V^{-1}\left(s,\theta\right)Eg_s\left(d,\theta\right)+o_p\left(1\right),\end{aligned}$$

by Assumption B.2 and B.4, and

$$\begin{aligned}
\|l_2\| &= \left\| \left[ \frac{1}{n} \sum_{i=1}^{n} g_s(d_i;\theta) g_s^{\tau}(d_i;\theta) \right]^{-1} \right\| \left\| \sum_{i=1}^{n} g_s(d_i;\theta) \frac{\left(\bar{\lambda}_s^{\tau} g_s(d_i;\theta)\right)^2}{2n} \right\| \\
&\leq \left\| \hat{V}^{-1}(s,\theta) \right\| \left\| \sum_{i=1}^{n} g_s^2(d_i;\theta) \right\|^{1/2} \left\| \sum_{i=1}^{n} \left[ \bar{\lambda}_s^{\tau} g_s(d_i;\theta) \right]^4 \right\|^{1/2} / n \\
&\leq \left\| \hat{V}^{-1}(s,\theta) \right\| \left\| \sum_{i=1}^{n} g_s^2(d_i;\theta) \right\|^{1/2} \sup \left\| \left[ \bar{\lambda}_s^{\tau} g_s(d_i;\theta) \right]^4 \right\| \\
&\leq \left\| \hat{V}^{-1}(s,\theta) \right\| \left\| \sum_{i=1}^{n} g_s^2(d_i;\theta) \right\|^{1/2} \sup \left\| \bar{\lambda}_s^{\tau} \right\|^4 \left\| g_s(d_i;\theta) \right\|^4 \\
&\leq \left\| \hat{V}^{-1}(s,\theta) \right\| \left\| \sum_{i=1}^{n} g_s^2(d_i;\theta) \right\|^{1/2} \left( \sup \|\lambda_s\| \left\| g_s(d_i;\theta) \right\| \right)^4 \\
&= o(1),
\end{aligned}$$

by Cauchy-Schwartz's inequality, Assumption B.2 and Lemma C.1.

**Lemma C.3** Under Assumption B.1-5, $\sup_{\theta \in \mathbf{\Theta}} ||\nabla_{\theta} \lambda_s^{\tau}(\theta) - V^{-1}(s,\theta) D(s,\theta)|| = o_p(1)$.

**Proof.** By (6.19), $\lambda_s^{\tau}$ solves

$$\sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\lambda_s^{\tau} g_s(d_i;\theta)\right] = 0.$$

Differentiating both sides with respect to $\theta$ gives

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \nabla_{\theta} g_s(d_i;\theta) \exp\left[-\lambda_s^{\tau} g_s(d_i;\theta)\right] \\
&\quad - \sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\lambda_s^{\tau} g_s(d_i;\theta)\right] \nabla_{\theta} \lambda_s^{\tau}(\theta) g_s(d_i;\theta) \\
&\quad - \sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\lambda_s^{\tau} g_s(d_i;\theta)\right] \lambda_s^{\tau}(\theta) \nabla_{\theta} g_s(d_i;\theta) \\
&\equiv l_1 - l_2 \nabla_{\theta} \lambda_s^{\tau}(\theta) + l_3.
\end{aligned}$$

The proof is completed after showing that (i) $\sup_{\theta \in \mathbf{\Theta}} ||l_1/n - D(s,\theta)|| = o_p(1)$; (ii) $\sup_{\theta \in \mathbf{\Theta}} ||l_2/n - V(s,\theta)|| = o_p(1)$;(iii) $\sup_{\theta \in \mathbf{\Theta}} ||l_3/n|| = o_p(1)$ and an application of triangular inequality.

We show (i) first. Note that

$$\begin{aligned} l_1/n &= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g_s\left(d_i;\theta\right) \exp\left[-\lambda_s^\tau g_s\left(d_i;\theta\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g_s\left(d_i;\theta\right) + o_p\left(1\right) \\ &= D\left(s,\theta\right) + o_p\left(1\right), \end{aligned}$$

by Lemma C.1 and a Law of Large Numbers.

We then show (ii). It is easily seen that

$$\begin{aligned} l_2/n &= \frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right) \exp\left[-\lambda_s^\tau g_s\left(d_i;\theta\right)\right] g_s\left(d_i;\theta\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right) g_s^\tau\left(d_i;\theta\right) + o_p\left(1\right) \\ &= V\left(s,\theta\right) + o_p\left(1\right), \end{aligned}$$

by Lemma C.1 and a Law of Large Numbers.

Finally, we show (iii).

$$\begin{aligned} \left\|l_3/n\right\| &= \left\|\frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right) \exp\left[-\lambda_s^\tau g_s\left(d_i;\theta\right)\right] \lambda_s^\tau\left(\theta\right) \nabla_\theta g_s\left(d_i;\theta\right)\right\| \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right) \lambda_s^\tau\left(\theta\right) \nabla_\theta g_s\left(d_i;\theta\right)\right\| + o_p\left(1\right) \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right)\right\| \sup_{\theta\in\mathbf{\Theta}} \left\|\lambda_s^\tau\left(\theta\right)\right\| \left\|\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g_s\left(d_i;\theta\right)\right\| + o_p\left(1\right) \\ &= o_p\left(1\right) \end{aligned}$$

by Lemma C.1, Assumption B.3, B.4 and a Law of large numbers.

## Proof of Main Theorems

**Proof of Theorem 4.1.**

Define $\mathrm{JE}_0\left(\theta\right) = -\frac{1}{S}\sum_{s=1}^{S} Eg_s^\tau\left(d_i;\theta\right) V\left(s,\theta\right) Eg_s\left(d_i;\theta\right) \equiv -\frac{1}{S}\sum_{s=1}^{S} h_s\left(s,\theta\right)$.

By Theorem 4.1.1 of Amemiya (1985), to prove $\hat{\theta} \to_p \theta_0$, we need only show that (i)

$\mathrm{JE}_0(\theta)$ is uniquely maximized at $\theta = \theta_0$ and (ii) $\sup_{\theta\in\mathbf{\Theta}} |\mathrm{JE}_n(\theta) - \mathrm{JE}_0(\theta)| \to_p 0$.

We first prove (i). By Assumption B.1 and B.4, $h_s(s,\theta) > 0$ for any $\theta \in \mathbf{\Theta}\backslash\{\theta_0\}$. However, $h_s(s,\theta_0) = Eg_s^\tau(d_i;\theta_0) V(s,\theta_0) Eg_s(d_i;\theta_0) = 0$ by (6.7). Thus $\mathrm{JE}_0(\theta) \geq 0$ with the unique minimizer $\theta = \theta_0$.

Next we show (ii). Applying Lemma C.2, write

$$\sum_{i=1}^n \lambda_s^\tau(\theta) g_s(d_i;\theta) = \sum_{i=1}^n Eg_s^\tau(d_i;\theta) V^{-1}(s,\theta) g_s(d_i;\theta) + o_p(1).$$

This leads to

$$\begin{aligned}
& |\mathrm{JE}_n(\theta) - \mathrm{JE}_0(\theta)| \\
= & \frac{1}{S}\left|\sum_{s=1}^S Eg_s^\tau(d_i;\theta) V(s,\theta) Eg_s(d_i;\theta) - \sum_{s=1}^S Eg_s^\tau(d_i;\theta) V^{-1}(s,\theta) \frac{1}{n}\sum_{i=1}^n g_s(d_i;\theta) + o_p(1)\right| \\
= & \frac{1}{S}\left|\sum_{s=1}^S Eg_s^\tau(d_i;\theta) V(s,\theta)\left[Eg_s(d_i;\theta) - \frac{1}{n}\sum_{i=1}^n g_s(d_i;\theta)\right] + o_p(1)\right| \\
\leq & \frac{1}{S}\sum_{s=1}^S |Eg_s^\tau(d_i;\theta) V(s,\theta) o_p(1) + o_p(1)| = o_p(1),
\end{aligned}$$

where we have used Assumption B.2, B.4 and a Law of Large numbers.

**Proof of Theorem 4.2.** Note that FOC of (6.21) is

$$\nabla_\theta \mathrm{JE}_n\left(\hat{\theta}\right) = 0.$$

By Taylor's Theorem, there exisit $\bar{\theta}$ lying between $\hat{\theta}$ and $\theta_0$, s.t.,

$$0 = \nabla_\theta \mathrm{JE}_n\left(\hat{\theta}\right) = \nabla_\theta \mathrm{JE}_n(\theta_0) + \nabla_{\theta\theta}\mathrm{JE}_n\left(\bar{\theta}\right)\left(\hat{\theta} - \theta_0\right).$$

This leads to

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\left[\nabla_{\theta\theta}\mathrm{JE}_n\left(\bar{\theta}\right)\right]^{-1}\left[\sqrt{n}\nabla_\theta \mathrm{JE}_n(\theta_0)\right].$$

We complete the proof by showing that (i) $\sqrt{n}\nabla_\theta \mathrm{JE}_n(\theta_0) \to N\left(0, \frac{1}{S^2} I(\theta_0)\right)$ and (ii) $-\nabla_{\theta\theta}\mathrm{JE}_n\left(\bar{\theta}\right) \to_p \frac{1}{S} J(\theta_0)$ and an application of Slutsky's Theorem.

(1) We first prove $\sqrt{n}\nabla_\theta \mathrm{JE}_n(\theta_0) \to N\left(0, \frac{1}{S^2} I^{-1}(\theta_0)\right)$. Note first that by (6.19), $\lambda_s^\tau$ solves

$$\sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] = 0.$$

Thus, we have

$$\begin{aligned}
\nabla_\theta \mathrm{JE}_n(\theta_0) &= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n} \nabla_\theta \lambda_s^\tau(\theta)\, g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&\quad + \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n} \lambda_s^\tau(\theta)\, \nabla_\theta g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n} \lambda_s^\tau(\theta)\, \nabla_\theta g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n} \left(\frac{1}{n}\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1} \nabla_\theta g_s(d_i;\theta) \\
&\quad \times \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] + o_p(1) \\
&\equiv \hat{U} + o_p(1)
\end{aligned}$$

We need to show that $n^{1/2}\hat{U} \to N\left(0, \frac{1}{S^2} I(\theta_0)\right)$.

Since $\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] = 1 - \lambda_s^\tau g_s(d_i;\theta) + o_p(1)$ by Assumption B.5. We have

$$\begin{aligned}
n^{1/2}\hat{U} &= n^{-1/2}\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n} \left(\frac{1}{n}\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1} \nabla_\theta g_s(d_i;\theta) \\
&\quad - n^{-1/2}\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n} \left(\frac{1}{n}\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1} \nabla_\theta g_s(d_i;\theta)\, \lambda_s^\tau g_s(d_i;\theta) \\
&= n^{-1/2}\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n} \left(\frac{1}{n}\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1} \nabla_\theta g_s(d_i;\theta) + o_p(1) \\
&= n^{-1/2}\hat{U}_1 + o_p(1).
\end{aligned}$$

Furthermore,

$$\begin{aligned} n^{-1/2}\hat{U}_1 &= \frac{1}{\sqrt{n}}\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n}\left(\sum_{i=1}^{n} g_s^{\tau}\left(d_i;\theta\right)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right) g_s^{\tau}\left(d_i;\theta\right)\right]^{-1}\nabla_{\theta} g_s\left(d_i;\theta\right)/n \\ &= \frac{1}{\sqrt{n}}\frac{1}{S}\sum_{i=1}^{n}\left\{\sum_{s=1}^{S} g_s^{\tau}\left(d_i;\theta\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right) g_s^{\tau}\left(d_i;\theta\right)\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta} g_s\left(d_i;\theta\right)\right)\right\} \\ &= \frac{1}{\sqrt{n}}\frac{1}{S}\sum_{i=1}^{n}\left\{\sum_{s=1}^{S} g_s^{\tau}\left(d_i;\theta\right) V^{-1}\left(s,\theta\right) G\left(s,\theta\right)\right\} + o_p\left(1\right) \\ &\equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_i \end{aligned}$$

where

$$\zeta_i = \frac{1}{S}\sum_{s=1}^{S} g_s^{\tau}\left(d_i;\theta\right) V^{-1}\left(s,\theta\right) G\left(s,\theta\right).$$

It is easily seen that $\zeta_i$ is an m.d.s. with variance

$$\begin{aligned} E\left[\zeta_i\zeta_i^{\tau}\right] &= E\left\{\sum_{s=1}^{S} g_s^{\tau}\left(d_i;\theta\right) V^{-1}\left(s,\theta\right) G\left(s,\theta\right)\right\}^2. \\ &= \frac{1}{S^2}\sum_{s,t=1}^{S} G^{\tau}\left(s,\theta\right) V^{-1}\left(s,\theta\right) E\left[g_s\left(d_i;\theta\right) g_t^{\tau}\left(d_i;\theta\right)\right] V^{-1}\left(t,\theta\right) G\left(t,\theta\right) \\ &= \frac{1}{S^2} G^{\tau} V^{-1}\Omega V^{-1} G \qquad \left(= \frac{1}{S^2} G^{\tau} V^{-1} G\right) \\ &\equiv \frac{1}{S^2} I\left(\theta_0\right). \end{aligned}$$

By a CLT for vector ergodic stationary m.d.s. (see, for example, Billingsley, 1961), we have

$$n^{1/2}\hat{U}_1 \to^d N\left(0, \frac{1}{S^2} I\left(\theta_0\right)\right).$$

(2) We then show that $-\nabla_{\theta\theta}\mathrm{JE}_n\left(\bar{\theta}\right) \to_p J\left(\theta_0\right)$. First,

$$\begin{aligned} -nS\nabla_{\theta\theta}\mathrm{JE}_n\left(\theta_0\right) &= \sum_{s=1}^{S}\sum_{i=1}^{n}\nabla_{\theta}\lambda_s^{\tau}\left(\theta\right) g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^{\tau} g_s\left(d_i;\theta\right)\right]\nabla_{\theta}\left[\lambda_s^{\tau}\left(\theta\right) g_s\left(d_i;\theta\right)\right] \\ &\quad + \sum_{s=1}^{S}\sum_{i=1}^{n}\nabla_{\theta}\lambda_s^{\tau}\left(\theta\right)\nabla_{\theta} g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^{\tau} g_s\left(d_i;\theta\right)\right] \\ &\quad + \sum_{s=1}^{S}\sum_{i=1}^{n}\lambda_s^{\tau}\left(\theta\right)\nabla_{\theta\theta} g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^{\tau} g_s\left(d_i;\theta\right)\right] \\ &\equiv u_1 + u_2 + u_3. \end{aligned}$$

We show that (i) $\|u_1/n\| = o_p(1)$, (ii) $\|u_2/(nS) - J(\theta_0)\| = o_p(1)$, and (iii) $\|u_3/n\| = o_p(1)$.

We first show (i) $\|u_1/n\| = o_p(1)$.

$$\begin{aligned}
\|u_1/n\| &= \left\| \frac{1}{nS} \sum_{s=1}^{S} \sum_{i=1}^{n} \nabla_\theta \lambda_s^\tau(\theta) g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \nabla_\theta \left[\lambda_s^\tau(\theta) g_s(d_i;\theta)\right] \right\| \\
&\leq \frac{1}{S} \sum_{s=1}^{S} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \lambda_s^\tau(\theta) g_s(d_i;\theta) \nabla_\theta \left[\lambda_s^\tau(\theta) g_s(d_i;\theta)\right] \right\| + o_p(1) \\
&\leq \frac{1}{S} \sum_{s=1}^{S} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \lambda_s^\tau(\theta) \right\| \left\| \frac{1}{n} \sum_{i=1}^{n} g_s(d_i;\theta) \right\| \\
&\quad \times \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\nabla_\theta \lambda_s^\tau(\theta)\right] g_s(d_i;\theta) \right\| + \left\| \frac{1}{n} \sum_{i=1}^{n} \lambda_s^\tau(\theta) \nabla_\theta g_s(d_i;\theta) \right\| \right\} + o_p(1) \\
&\leq o_p(1),
\end{aligned}$$

by Assumption B.3, B.4 and Lemma C.1, C.2.

We next show (ii) $\|u_2/(nS) - J(\theta_0)\| = o_p(1)$. Note that by Lemma C.3, we have

$$\begin{aligned}
u_2/(nS) &= \frac{1}{nS} \sum_{s=1}^{S} \sum_{i=1}^{n} \nabla_\theta \lambda_s^\tau(\theta) \nabla_\theta g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&= \frac{1}{nS} \sum_{s=1}^{S} \sum_{i=1}^{n} \nabla_\theta \lambda_s^\tau(\theta) \nabla_\theta g_s(d_i;\theta) + o_p(1) \\
&= \frac{1}{nS} \sum_{s=1}^{S} \sum_{i=1}^{n} G^\tau(s,\theta) V^{-1}(s,\theta) \nabla_\theta g_s(d_i;\theta) + o_p(1) \\
&= \frac{1}{S} \sum_{s=1}^{S} G^\tau(s,\theta) V^{-1}(s,\theta) \left( \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta g_s(d_i;\theta) \right) + o_p(1) \\
&= \frac{1}{S} \sum_{s=1}^{S} G^\tau(s,\theta) V^{-1}(s,\theta) G(s,\theta) + o_p(1), \\
&= \frac{1}{S} G^\tau V^{-1} G + o_p(1) = \frac{1}{S} J(\theta_0) + o_p(1).
\end{aligned}$$

by Assumption B.4 and a Law of Large Numbers.

Finally we show (iii) $\|u_3/n\| = o_p(1)$.

$$\begin{aligned}
\|u_3/n\| &= \left\| \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n} \lambda_s^{\tau}(\theta) \nabla_{\theta\theta} g_s(d_i;\theta) \exp\left[-\lambda_s^{\tau} g_s(d_i;\theta)\right] \right\| \\
&\leq \sum_{s=1}^{S} \left\| \frac{1}{n} \sum_{i=1}^{n} \lambda_s^{\tau}(\theta) \nabla_{\theta\theta} g_s(d_i;\theta) \right\| + o_p(1) \\
&\leq \sum_{s=1}^{S} \left\| \frac{1}{n} \sum_{i=1}^{n} \lambda_s^{\tau}(\theta) \right\| \left\| \frac{1}{n} \nabla_{\theta\theta} g_s(d_i;\theta) \right\| + o_p(1) \\
&\leq o_p(1),
\end{aligned}$$

by Assumption B.4 and Lemma C.2.

## References


Altonji, J. G., and Segal, L. M. (1996), "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, Vol. 14, 353-366.

Amemiya, T. (1985). *Advanced Econometrics*, Harvard University Press.

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings", *The Review of Economics and Statistics*, Vol. 60, No. 1, 47-57

Bai, J. and S. Ng (2010), "Instrumental Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26:6, 1577-1606.

Bates, J. and C.W.J., Granger (1969), "The Combination of Forecasts," *Operations Research Quarterly* 20 (4): 451–468.

Belloni, A., V. Chernozhukov, and C. Hansen (2011), "LASSO Methods for Gaussian Instrumental Variables Models", Working paper, MIT, Econ. Dept.

Berk, R., L., Brown, A., Buja, K., Zhang and L. Zhao (2011). "Valid Post-Selection Inference," Statistics Department, Wharton School, University of Pennsylvania, Discussion Paper.

Berk, R., L., Brown and L., Zhao (2009). "Statistical inference after model selection," *Journal of Quantitative Criminology*, 26, 217-236.

Billingsley, P. (1961), "The Lindeberg-Levy Theorem for Martingales," in *Proceedings of the American Mathematical Society*, Vol. 12, 788-792.

Box, G.E.P. (1979), "Robustness in the strategy of scientific model building," in *Robustness in Statistics*, R.L. Launer and G.N. Wilkinson, Editors. Academic Press:

New York.

Breiman, L. (1996): "Bagging Predictors," *Machine Learning*, 36, 105–139.

Bühlmann, P., and B. Yu (2002): "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961.

Chen, X., E., Tamer, and A., Torgovitsky (2011). "Sensitivity Analysis in a Semiparametric Likelihood Model: A Partial Identification Approach," Department of Economics, Yale University, Working paper.

Clyde, M. and E.I., George (2004), "Model Uncertainty," Statistical Science, Vol. 19, No. 1, 81-94.

Cover, T.M. and J.A., Thomas (2006), *Elements of Information Theory*, John Wiley & Sons, Inc.

De Jong, S., and H.A.L. Kiers (1992), "Principal covariate regression," *Chemometrics and Intelligent Laboratory Systems* 14, pp. 155-164.

Diebold, F.X., Rudebusch, G.D. and Aruoba, B. (2006), "The Macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach," *Journal of Econometrics*, 131, 309-338.

Eicher, T.S., A., Lenkoski and A.E., Raftery (2009), "Bayesian Model Averaging and Endogeneity Under Model Uncertainty: An Application to Development Determinants," Working Paper no. 94 Center for Statistics and the Social Sciences, University of Washington.

Fan, J. and R. Li (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of American Statistical Association.* 96 1348-1360.

Fisher, I. (1973), "I Discovered the Phillips Curve: A Statistical Relation between Unemployment and Price Changes," *The Journal of Political Economy*, Vol. 81, 496-502. Reprint of 1926 article by Irving Fisher.

Friedman, M. (1957), *A Theory of the Consumption Function*, Princeton University Press.

Galbraith, J.W. and V. Zinde-Walsh (2010), "Reduced-Dimension Control Regression," Department of Economics, McGill University, Working Paper.

Gernert, D. (2007), "Ockham's Razor and Its Improper Use," *Journal of Scientific Exploration*, Vol. 21, No. 1, pp. 135-40.

Golan, A., G., Judge, and D., Miller (1996), *Maximum Entropy Econometrics: Robust Eestimation with Limited Data*, John Wiley & Sons

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, Vol. 66, 315-331.

Hahn, J. (2004), "Functional Restriction and Efficiency in Causal Inference," *The Review of Economics and Statistics*, 86, 73-76.

Hansen, B.E. (2007), "Least Squares Model Averaging," *Econometrica*, Vol. 75, 1175-1189.

Hansen, B.E. (2008), "Least-square Forecast Averaging," *Journal of Econometrics*, Vol. 146, 342-350.

Hansen, B.E. (2009), "Averaging Estimators for Regression with a Possible Structure Break," *Econometric Theory* 35, 1498-1514.

Hansen, B.E. (2010), "Averaging Estimators for Autoregressions with a Near Unit Root," *Journal of Econometrics*, Vol. 158, 142-155.

Hansen, B.E., and J. Racine (2011), "Jackknife Model Averaging," *Journal of Econometrics*, forthcoming.

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

Hansen, L., J. Heaton and A. Yaron (1996), "Finite Sample Properites of Alternative GMM Estimators," *Journal of Business and Economic Statistics*, Vol. 14, 262-280.

Hoeting, J.A., D, Madigan, A.E., Raftery and C.T., Volinsky (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, Vol. 14, No. 4, 382-417.

Hsiao, C., Q., Li and J., Racine (2007), "A consistent model speci cation test with mixed discrete and continuous data," *Journal of Econometrics* 140, 802-826.

Huang, J., Horowitz, J. and Ma, S. (2008). "Asymptotic properties of bridge estimators in sparse high-dimensional regression models." *The Annals of Statistics* 36, 587-613.

Imbens, G. and C. Maski (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, Vol. 72, 1845-1857.

2004, pp. 1845-1857.

Imbens, G., R.H., Spady, and P. Johnson (1998), "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333-357.

Keynes, J. M. (1936), *The General Theory of Unemployment, Interest and Inflation*

Kitamura, Y., (2006), "Empirical Likelihood Methods in Econometrics: Theory and Practice," *Advances in Economics and Econometrics: Theory and Applications*, Nineth World Congress, Edited by R. Blundell, W., Newey, and T., Persson.

Kitamura, Y. and M., Stutzer (1997), "An Information-theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861-874.

Kitamura, Y., G., Tripathi, and H., Ahn (2004), "Empirical Likelihood-based Inference in Conditional Moment Restriction Models," *Econometrica*, Vol. 72, 1667-1714.

Kruger, A.B., (1993), "How Computers Have Changed the Wage Structure: Evidence from Microdata," 1984-1989, *The Quarterly Journal of Economics*, Vol. 108, No. 1, 33-60.

Lee, T.H., Y., Tu and A., Ullah (2011a),"Nonparametric and Semiparametric Regressions Subject to Monotonicity Constraints: Estimation and Forecasting," Working paper, UC, Riverside, Econ. Dept.

Lee, T.H., Y., Tu and A., Ullah (2011b),"Forecasting Equity Premium: Global Historical Mean Versus Local Historical Mean and Constraints," Working paper, UC, Riverside, Econ. Dept.

Leeb, H. and B.M., Pötscher (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory* 21, 29-59.

Leeb, H. and B.M., Pötscher (2006), "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators ?", *Annals of Statistics* 34, 2554-2591

Leeb, H. and B.M., Pötscher (2008a), "Recent Developments in Model Selection and Related Areas," *Econometric Theory*, 24, 319-322.

Leeb, H. and B.M., Pötscher (2008b), "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators ?" *Econometric Theory*, 24, 338-376.

Leeb, H. and B.M., Pötscher (2008c), "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," *Journal of Econometrics* 142, 201-211.

Leeb, H. and B.M., Pötscher (2009), "On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding," *Journal of Multivariate Analysis* 100, 2065-2082.

Li, Q., Wang, S. (1998), "A simple consistent bootstrap test for a parametric regression function," *Journal of Econometrics*, 87, 145-165.

Manski, C.F. (1995), *Identification Problems in the Social Sciences*, Harvard University Press

Manski, C.F. (2003), *Partial Identification of Probability Distributions*, Springer-Verlag

Manski, C.F. (2007), *Identification for Prediction and Decision*, Harvard University Press

Miler, Alan (2002), *Subset Selection in Regression*, Chapman & Hall/CRC.

Mincer, J. (1976). "Unemployment Effects of Minimum Wages," *Journal of Political Economy* 84, 87-104.

Newey, W. K., and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics* Vol. 4, Edited by Daniel McFadden and Robert Engle. Amsterdam, The Netherlands: Elsevier, North-Holland, 1999, chapter 36, pp. 2113-2245.

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika* 75, 237-49.

Owen, A. (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics* 18, 90-120.

Owen, A. (1991), "Empirical Likelihood for Linear Models," *The Annals of Statistics* 19, 1725-1747.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine* 2 (6): 559-572.

Pérez-González, F. (2006), "Inherited Control and Firm Perfance," American Economic Review 96, 1559-1588.

Phillips, A. W. (1958), "The Relationship between Unemployment and the Rate of Change of Money Wages in the United Kingdom 1861-1957," *Economica* 25 (100): 283-99.

Pötscher, B.M. (2009), "Confidence Sets Based on Sparse Estimators Are Necessarily Large," *Sankhya* 71-A, 1-18.

Pötscher, B.M. and U., Schneider (2009), "On the Distribution of the Adaptive LASSO Estimator," *Journal of Statistical Planning and Inference* 139, 2775-2790.

Pötscher, B.M. and U., Schneider (2010), "Confidence Sets Based on Penalized Maximum Likelihood Estimators," *Electronic Journal of Statistics* 10, 334-360.

Qin J. and J. Lawless (1994), "Empirical Likelihood and General estimating Equations," *The Annals of Statistics* 22, 300-325.

Racine, J. and B. Hansen (2011), "Jackknife Model Averaging,", *Journal of Econometrics*, *forthcoming*

Rapach, D., J., Strauss and G. Zhou (2010), "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *Review of Financial Studies* 23, 821-862.

Robinson, P.M. (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, Vol. 56, 931-954.

Romano, J.P. and A. Shaikh (2008), "Inference for Identi able Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, Vol. 138, 2786-2807.

Romano, J.P. and A. Shaikh (2010), "Inference for the Identi ed Set in Partially Identified Econometric Models," *Econometrica*, Vol. 78, 169-211.

Sala-i-Martin, X., G., Doppelhofer and R., Miller (2004), "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94(4): 813–835.

Santos A. (2011), "Inference in Nonparametric Instrumental Variables with Partial Identification," *Econometrica*, forthcoming.

Spearman, C. (1904), "General intelligence, objectively determined and measured," *American Journal of Psychology*, 15, 201-293.

Stock, J.D., and M. Watson (2010), *Introduction to Econometrics*, 3ed. Addison Wesley.

Su, L. and H. White (2007a), "A Consistent Characteristic Function-Based Test for Conditional Independence," *Journal of Econometrics*, 141, 807-834.

Su, L. and H. White (2007b), "Testing Conditional Independence via Empirical Likelihood," UCSD Dept. of Economics Discussion Paper.

Su, L. and H. White (2008), "A Nonparametric Hellinger Metric Test for Conditional Independence," *Econometric Theory*, 24, 829-864.

Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso," *Journal of Royal Statistical Society* Ser. B 58 267–288.

Tu, Y. (2011a), "Nonparametric Estimation with Economic Constraints via Trimming," Working paper

Tu, Y. (2011b), "Information-based Tests of Conditional Mean Independence," Working paper

Tu, Y. and T.H. Lee (2011), "Forecasting Using Supervised Factor Model," Working paper

White H. (2000), "A Reality Check for Data Snooping," *Econometrica*, Vol. 68, 1097-1126.

White, H. and K. Chalak (2006), "A Unified Framework for Defining and Identifying Causal Effects," Working paper, UCSD, Economics Dept.

White, H., K. Chalak, and X. Lu (2011), "Linking Granger Causality and the Pearl Causal Model with Settable Systems," *Journal of Machine Learning Research Workshop and Conference Proceedings*, 12, 1-29.

White, H. and X. Lu (2010), "Robustness Checks and Robustness Tests in Applied Economics," UCSD Dept. of Economics Discussion Paper.

Wold, Herman (1966). "Estimation of principal components and related models by iterative least squares," In P.R. Krishnaiaah (Ed.). *Multivariate Analysis.* (pp. 391-420) New York: Academic Press.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101, 1418-1429.

Zou, H. and T. Hastie (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Ser. B.* 67 301-320.

Table 6.4: Squared Bias (×100): DGP 1

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.01 | 0.01 | 0.01 | 580.45 | 2650.07 | 5672.03 | 7.25 | 2.69 | 10.66 | 10.66 | 11.01 |
| | $\theta_2$ | 0.00 | 0.00 | 0.00 | 1744.79 | 1353.36 | 2344.46 | 0.00 | 0.39 | 3.41 | 3.41 | 2.93 |
| n= 200 | $\theta_1$ | 0.00 | 0.00 | 0.00 | 5445.91 | 6859.52 | 7348.23 | 6.11 | 3.95 | 0.04 | 0.04 | 0.00 |
| | $\theta_2$ | 0.00 | 0.00 | 0.00 | 2910.09 | 7132.12 | 7690.01 | 6.04 | 10.69 | 2.22 | 2.22 | 1.69 |

Table 6.5: Mean Squared Error (×100): DGP 1

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 4.12 | 4.12 | 4.12 | 581.03 | 2653.12 | 5672.88 | 7.57 | 3.07 | 10.70 | 10.70 | 11.06 |
| | $\theta_2$ | 0.91 | 0.91 | 0.91 | 1744.84 | 1354.10 | 2344.68 | 0.09 | 0.51 | 3.42 | 3.42 | 2.94 |
| n= 200 | $\theta_1$ | 0.56 | 0.56 | 0.56 | 5445.99 | 6859.71 | 7348.34 | 6.17 | 4.03 | 0.04 | 0.04 | 0.01 |
| | $\theta_2$ | 0.81 | 0.81 | 0.81 | 2910.28 | 7132.44 | 7690.17 | 6.10 | 10.77 | 2.23 | 2.23 | 1.70 |

Table 6.6: Squared Bias (×100): DGP 2

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.00 | 0.00 | 0.00 | 2.68 | 1.83 | 1.99 | 0.11 | 6.57 | 5.31 | 5.88 | 5.24 |
| | $\theta_2$ | 0.02 | 0.02 | 0.02 | 5.30 | 1.24 | 2.44 | 0.00 | 3.80 | 3.18 | 4.34 | 3.24 |
| n= 200 | $\theta_1$ | 0.00 | 0.00 | 0.00 | 0.09 | 0.04 | 0.13 | 0.04 | 0.00 | 0.17 | 0.22 | 0.17 |
| | $\theta_2$ | 0.01 | 0.01 | 0.01 | 2.98 | 0.76 | 0.78 | 0.02 | 0.13 | 0.19 | 0.07 | 0.18 |

Table 6.7: Mean Squared Error (×100): DGP 2

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 7.64 | 7.64 | 7.64 | 5.11 | 5.83 | 5.25 | 6.75 | 10.01 | 7.75 | 8.57 | 7.68 |
| | $\theta_2$ | 7.05 | 7.05 | 7.05 | 7.24 | 4.71 | 4.95 | 4.06 | 6.94 | 5.14 | 6.44 | 5.20 |
| n= 200 | $\theta_1$ | 2.84 | 2.84 | 2.84 | 0.56 | 2.05 | 1.40 | 1.61 | 0.73 | 0.63 | 0.69 | 0.63 |
| | $\theta_2$ | 3.90 | 3.90 | 3.90 | 3.52 | 3.44 | 2.40 | 1.93 | 0.87 | 0.72 | 0.61 | 0.71 |

Table 6.8: Squared Bias (×100): DGP 3-1

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.01 | 0.02 | 0.03 | 10.83 | 5.26 | 8.19 | 0.79 | 9.20 | 11.18 | 12.13 | 11.22 |
| | $\theta_2$ | 0.01 | 0.00 | 0.00 | 0.83 | 5.46 | 9.23 | 1.30 | 16.25 | 17.73 | 18.64 | 17.76 |
| n= 200 | $\theta_1$ | 0.00 | 0.01 | 0.01 | 0.60 | 0.00 | 0.04 | 0.07 | 0.17 | 0.58 | 0.70 | 0.59 |
| | $\theta_2$ | 0.00 | 0.00 | 0.00 | 1.84 | 1.17 | 1.97 | 0.34 | 0.86 | 0.34 | 0.36 | 0.34 |

Table 6.9: Mean Squared Error (×100): DGP 3-1

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 20.01 | 15.26 | 15.49 | 18.33 | 15.17 | 16.66 | 20.09 | 17.81 | 18.38 | 19.36 | 18.43 |
| | $\theta_2$ | 24.85 | 22.58 | 23.26 | 9.18 | 16.99 | 18.32 | 18.16 | 25.21 | 24.49 | 25.22 | 24.53 |
| n= 200 | $\theta_1$ | 4.43 | 3.79 | 3.80 | 2.45 | 2.76 | 2.20 | 3.10 | 2.22 | 2.43 | 2.59 | 2.43 |
| | $\theta_2$ | 5.48 | 4.99 | 5.04 | 3.63 | 4.78 | 4.71 | 4.27 | 3.00 | 2.29 | 2.40 | 2.28 |

Table 6.10: Squared Bias (×100): DGP 3-2

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.05 | 0.00 | 0.01 | 11.84 | 11.74 | 15.00 | 0.81 | 0.75 | 0.15 | 0.09 | 0.14 |
| | $\theta_2$ | 0.06 | 0.02 | 0.00 | 0.02 | 0.01 | 0.01 | 0.00 | 1.66 | 2.99 | 3.97 | 3.06 |
| n= 200 | $\theta_1$ | 0.00 | 0.00 | 0.00 | 17.70 | 2.58 | 5.97 | 0.99 | 13.76 | 15.75 | 15.57 | 15.72 |
| | $\theta_2$ | 0.00 | 0.00 | 0.00 | 1.49 | 0.24 | 0.39 | 0.07 | 1.81 | 1.95 | 2.17 | 1.95 |

Table 6.11: Mean Squared Error (×100): DGP 3-2

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 29.75 | 20.48 | 21.57 | 19.99 | 22.91 | 24.33 | 29.58 | 12.74 | 8.88 | 9.12 | 8.86 |
| | $\theta_2$ | 19.72 | 17.79 | 18.88 | 13.68 | 13.66 | 13.50 | 18.22 | 17.42 | 15.91 | 16.92 | 15.98 |
| n= 200 | $\theta_1$ | 8.22 | 6.85 | 6.91 | 20.19 | 6.60 | 9.03 | 5.94 | 16.60 | 18.22 | 18.06 | 18.20 |
| | $\theta_2$ | 7.70 | 6.31 | 6.37 | 4.76 | 5.26 | 4.56 | 5.83 | 5.66 | 5.24 | 5.49 | 5.24 |

Table 6.12: Squared Bias (×100): DGP 3-3

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.00 | 0.00 | 400.00 | 0.00 | 3.15 | 1.33 | 0.16 | 0.18 | 0.06 | 0.19 | 0.09 |
| | $\theta_2$ | 0.00 | 0.01 | 900.00 | 13.22 | 13.36 | 12.29 | 0.08 | 0.08 | 0.10 | 0.01 | 0.09 |
| n= 200 | $\theta_1$ | 0.00 | 0.00 | 400.00 | 1.97 | 0.07 | 0.36 | 0.34 | 4.06 | 3.16 | 3.45 | 3.16 |
| | $\theta_2$ | 0.00 | 0.00 | 900.00 | 0.09 | 0.02 | 0.00 | 0.20 | 0.54 | 0.25 | 0.28 | 0.25 |

Table 6.13: Mean Squared Error (×100): DGP 3-3

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 3.32 | 2.47 | 400.00 | 0.10 | 4.62 | 1.84 | 1.79 | 0.56 | 0.17 | 0.35 | 0.20 |
| | $\theta_2$ | 1.88 | 1.26 | 900.00 | 13.59 | 13.95 | 12.61 | 1.18 | 0.86 | 0.47 | 0.50 | 0.48 |
| n= 200 | $\theta_1$ | 0.38 | 0.26 | 400.00 | 1.99 | 0.40 | 0.61 | 0.57 | 4.19 | 3.18 | 3.47 | 3.18 |
| | $\theta_2$ | 0.71 | 0.56 | 900.00 | 0.30 | 0.62 | 0.48 | 0.62 | 0.82 | 0.45 | 0.49 | 0.45 |

Table 6.14: Squared Bias (×100): DGP 4

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.01 | 0.00 | 0.00 | 0.05 | 0.01 | 0.01 | 0.16 | 0.25 | 0.14 | 0.11 | 0.14 |
| | $\theta_2$ | 0.00 | 0.00 | 0.00 | 0.04 | 0.08 | 0.08 | 0.26 | 0.27 | 0.16 | 0.12 | 0.16 |
| n= 200 | $\theta_1$ | 0.00 | 0.00 | 0.00 | 0.02 | 0.07 | 0.08 | 0.07 | 0.02 | 0.00 | 0.01 | 0.00 |
| | $\theta_2$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.06 | 0.01 | 0.01 | 0.01 |

Table 6.15: Mean Squared Error (×100): DGP 4

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 6.79 | 1.88 | 1.88 | 1.82 | 2.27 | 1.97 | 4.88 | 2.43 | 1.89 | 1.94 | 1.88 |
| | $\theta_2$ | 8.88 | 2.36 | 2.36 | 2.33 | 3.19 | 2.73 | 6.28 | 3.03 | 2.37 | 2.46 | 2.37 |
| n= 200 | $\theta_1$ | 2.30 | 0.53 | 0.53 | 0.56 | 0.62 | 0.61 | 1.15 | 0.60 | 0.53 | 0.54 | 0.53 |
| | $\theta_2$ | 2.21 | 0.60 | 0.60 | 0.58 | 0.64 | 0.61 | 1.11 | 0.71 | 0.59 | 0.60 | 0.59 |

Table 6.16: Squared Bias (×100): DGP 5-1

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.01 | 0.00 | 0.01 | 9.18 | 4.12 | 5.22 | 0.25 | 2.41 | 0.81 | 0.38 | 0.72 |
| | $\theta_2$ | 0.02 | 0.00 | 0.02 | 2.67 | 0.70 | 1.23 | 0.15 | 1.66 | 0.86 | 0.22 | 0.74 |
| n= 200 | $\theta_1$ | 0.00 | 0.00 | 0.00 | 0.04 | 0.08 | 0.12 | 0.09 | 0.04 | 0.01 | 0.02 | 0.01 |
| | $\theta_2$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.05 | 0.06 | 0.01 | 0.00 | 0.01 |

Table 6.17: Mean Squared Error (×100): DGP 5-1

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 42.30 | 4.91 | 6.55 | 17.77 | 16.16 | 15.94 | 31.91 | 12.26 | 9.72 | 9.24 | 9.57 |
| | $\theta_2$ | 23.13 | 6.41 | 8.38 | 16.16 | 13.98 | 14.53 | 21.18 | 15.10 | 13.43 | 12.59 | 13.23 |
| n= 200 | $\theta_1$ | 8.22 | 1.24 | 1.28 | 2.53 | 2.66 | 2.67 | 4.48 | 2.73 | 2.48 | 2.49 | 2.48 |
| | $\theta_2$ | 7.70 | 1.64 | 1.65 | 3.27 | 3.48 | 3.47 | 5.54 | 3.41 | 3.30 | 3.30 | 3.30 |

Table 6.18: Squared Bias (×100): DGP 5-2

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.01 | 0.01 | 0.02 | 0.06 | 0.16 | 0.54 | 0.12 | 0.01 | 0.01 | 0.01 | 0.01 |
| | $\theta_2$ | 0.00 | 0.01 | 0.04 | 6.99 | 11.86 | 20.58 | 0.18 | 0.09 | 0.20 | 0.25 | 0.19 |
| n= 200 | $\theta_1$ | 0.03 | 0.00 | 0.00 | 0.16 | 0.10 | 0.02 | 0.01 | 0.02 | 0.04 | 0.02 | 0.04 |
| | $\theta_2$ | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.08 | 0.11 | 0.00 | 0.01 | 0.00 | 0.01 |

Table 6.19: Mean Squared Error (×100): DGP 5-2

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 106.66 | 9.44 | 21.20 | 26.45 | 29.91 | 27.52 | 78.98 | 24.91 | 25.06 | 20.46 | 24.38 |
| | $\theta_2$ | 140.79 | 14.21 | 21.09 | 67.04 | 68.65 | 88.93 | 143.22 | 40.32 | 50.58 | 43.07 | 50.53 |
| n= 200 | $\theta_1$ | 9.97 | 1.54 | 1.80 | 3.85 | 3.88 | 3.71 | 7.40 | 3.88 | 3.76 | 3.59 | 3.75 |
| | $\theta_2$ | 16.38 | 3.73 | 4.07 | 7.56 | 7.66 | 7.85 | 11.80 | 8.46 | 7.71 | 7.41 | 7.71 |

Table 6.20: Squared Bias (×100): DGP 5-3

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 0.00 | 0.04 | 0.06 | 0.26 | 0.16 | 0.04 | 0.22 | 0.27 | 0.05 | 0.00 | 0.05 |
| | $\theta_2$ | 0.26 | 0.01 | 0.00 | 0.14 | 1.29 | 3.39 | 0.67 | 0.05 | 0.04 | 0.00 | 0.04 |
| n= 200 | $\theta_1$ | 0.02 | 0.01 | 0.01 | 0.09 | 0.06 | 0.01 | 0.01 | 0.03 | 0.04 | 0.09 | 0.04 |
| | $\theta_2$ | 0.06 | 0.00 | 0.00 | 0.02 | 0.04 | 0.14 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 |

Table 6.21: Mean Squared Error (×100): DGP 5-3

| | $\theta_0$ | OLS | GLS | FGLS | factor | MMA | JMA | LASSO | FOGLS | gMPL1 | gMPL2 | eMPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n= 50 | $\theta_1$ | 71.76 | 19.48 | 25.39 | 22.83 | 27.98 | 25.91 | 66.38 | 29.01 | 24.19 | 25.02 | 24.21 |
| | $\theta_2$ | 108.90 | 22.82 | 27.67 | 27.68 | 35.08 | 35.65 | 84.66 | 33.65 | 28.85 | 30.36 | 28.88 |
| n= 200 | $\theta_1$ | 17.61 | 3.02 | 3.28 | 4.76 | 5.03 | 4.90 | 10.57 | 5.06 | 4.77 | 4.82 | 4.78 |
| | $\theta_2$ | 29.56 | 5.57 | 6.01 | 8.49 | 9.05 | 9.12 | 19.83 | 8.53 | 8.41 | 8.43 | 8.41 |

# Chapter 7

# Testing Additive Separability of Error Term in Nonparametric Structural Models

## 7.1 Introduction

Economic models that incorporate stochastic features usually proceed by specifying the relationship between an observed dependent variable (or variable of interest), a set of observed independent variables (or explanatory variables), and some unobservable random term represented by error (or shock). This chapter examines how to deal with this unobserved error in the econometric modeling process and whether it enters the econometric model as a separable additive component or as a nonseparable element.

Although economic theory dictates that some economic variables are important for the causal effects of others, rarely does it state exactly how the interaction takes place or how unobserved factors affect the variable of interest. For example, labor

economists tend to adopt years of education, years on a current job, experience in labor force, age, etc., as main sources to explain the variation in workers' earnings. Early literature adopts a simple linear regression model, or a log-linear equation, which incorporates the unobserved effects as an additive component. This simple model is later generalized to nonlinear regression models and more recently nonparametric models. Again all these models, parametric or nonparametric, share the same feature that the unobserved random term enters the model in an additive manner. While this is a conventional assumption undertaken in regression literature, it could be highly unlikely that many economic structures do fall into this group. In other words, the error term could be nonseparable from the main economic structure.

When the error term is nonadditive, the conventional identification and estimation approaches for additive nonparametric models are not applicable anymore. Therefore, new approaches to identification and estimation are called upon for nonseparable nonparametric models. A great deal of efforts has been made towards this direction in the past two decades or so. Roehrig (1988) provides a general condition for the identification of nonparametric equation systems. Brown and Matzkin (1998) present an extremum estimator for the nonparametric simultaneous equations studied in Roehrig (1988), with a generalized identification condition formulated. Matzkin (1999) proposes a maximum rank correlation estimator for a nonparametric model in which the dependent variable is monotone in an unobservable random error term. The investment function considered in Olley and Pakes (1996) is also nonseparable but strictly monotone in the unobservable productivity variable, which is then used to get rid of the unobservable by expressing it as a function of the observed variables. Heckman and Vytlacil (1999, 2001) consider models in which potential outcomes are nonseparable in unobservable terms. Blundell and Powell (2003) estimate the average structural function

in nonparametric nonseparable models in the presence of endogeneity. Matzkin (2003) presents estimators for nonparametric nonadditive models and shows their asymptotic characteristics under a set of assumptions that may be implied by economic theory. Altonji and Matzkin (2005) adopt a conditional independence assumption to estimate the average derivative of a nonparametric function and the distribution of the unobservable random term, when the unobservable is nonadditive and the regressors are endogenous. Briesch, Chintagunta and Matzkin (2007) provide a method to estimate discrete choice models with unobserved heterogeneity that enters the subutility function nonadditively. Heckman, Matzkin and Nesheim (2010) establish nonparametric identification of structural functions and distributions in general nonparametric nonadditive hedonic models by relaxing the assumptions of additive marginal utility and additive marginal product function adopted in Ekeland, Heckman and Nesheim (2004). Altonji, Ichimura and Otsu (2011) present a simple method to estimate the marginal effects of observable variables on a limited dependent variable, when the dependent variable is a nonseparable function of observables and unobservables.

Albeit the literature is flooded with approaches that are capable of tackling both separable and nonseparable nonparametric models, there is no valid method available to distinguish which model is more appropriate for the problem confronted by the researchers. We believe that there are at least four reasons that amplify the urgent need for some convincing testing procedures to detect the way through which the unobservable random term enters the economic structure. They are: (1) The economic meaning of an unobservable random term varies from case to case; (2) The identification and statistical properties of the estimated underlying economic structure depend on whether additive separability holds; (3) The identification and estimation of other economic structures also relies on the separability properties; and (4) There is a lack

of consistent testing procedures to detect additive separability of unobservables in the literature. These are described below.

**Economic meaning of an unobservable.** An additive unobservable takes on the traditional explanation as measurement error of the variable of interest, or a level shift of the dependent variable due to some random shocks to the economy, or some minor factors other than the included regressors that may affect the dependent variable. A nonadditive unobservable random term, on the other hand, may adopt explanations such as a *heterogeneity* parameter in a utility function, the productivity shock or utility value for some unobserved attributes, etc. See, for example, Heckman (1974), Heckman and Willis (1977), McFadden (1974), and Lancaster (1979), among others. A clarification of the additivity property of the unknown economic structure helps to identify the economic meaning of an unobservable, which facilitates further evaluation of sources of heterogeneity, improvement of productivity for firms and better economic policy proposals.

**Identification and statistical properties of the estimators.** Classical additive nonparametric models can be identified under standard conditional moment restrictions, and estimated, for example, by conventional nonparametric kernel or sieve methods. The consistency and asymptotic normality of these nonparametric estimators have been well understood. In contrast, methods to identify and estimate nonadditive nonparametric functions are relatively new in the literature and have not yet been fully explored. Matzkin (2003) presents an estimator of the nonseparable nonparametric random function, and shows that it is consistent and asymptotically normal under certain identification conditions. She argued that her identification conditions are not very strong since they may be implied by some economic theory and are rather straightforward to derive if certain parametric functional forms are tolerated. Yet, one concern

regarding these identification conditions is that the underlying economic theory itself is subject to valid tests, not to mention its implications or the parametric functional forms that are implicitly needed to facilitate the formulation of identification conditions. Therefore, there is a potentially high cost of applying these conditions for identification purposes. Although direct comparisons of the asymptotic properties of estimators in additive and nonadditive models are not available, simulations in Matzkin (2003) show that estimators under the correct additive restriction have much smaller variance and mean squared errors than those without imposing additive restriction. For this reason, it seems prudent to develop a test for additive separability in nonparametric models before embarking on estimation and inference.

**Estimation of other economic structure.** Quite often it is also of interest to estimate other sensible economic structure. Examples are available in the policy evaluation literature. Heckman and Vytlacil (2005) construct models with heterogeneity in response to treatment among otherwise observationally identical people. The nonparametric selection model proposed with testable restrictions can be used to "unify the treatment literature, identify different treatment effects, link the literature on treatment effects to the literature in structural econometrics, and interpret the implicit economic assumptions underlying instrumental variables and matching models." However, they point out that the results of the chapter and even the entire recent literature on instrumental variable estimators with heterogeneous responses "relies critically on the assumption that the treatment choice equation has a representation in the additive separable form." Although the marginal treatment effect (MTE) can be extended in the nonseparable case and it is policy invariant, the linear instrumental variable (LIV) does not identify MTE. Heckman and Vytlacil (2001) show that, even after some transformation, the defined MTE is still not identified through LIV, and MTE defined in this

way precludes getting treatment parameters via integration. Furthermore, Heckman and Vytlacil (2005) also notice that nonseparability will lead to failure of the index sufficiency. In other words, additive structure simplifies the estimation of other economic structure. Yet, there is no convincing testing procedure to provide such evidence that the economic structure under investigation is indeed additive.

**Lack of specification tests for separability.** Since Hausman (1978) a large literature on testing for the correct specification of functional forms has developed; see Bierens (1982, 1990), Ruud (1984), Newey (1985), Tauchen (1985), White (1987), Robinson (1989), Wooldridge (1992), Yatchew (1992), Härdle and Mammen (1993), Hong and White (1995), Zheng (1996), Andrews (1997), Bierens and Ploberger (1997), Li and Wang (1998), Stinchcombe and White (1998), and Hsiao, Li and Racine (2007), among others. Some of these tests are only consistent against some specific alternatives while others are consistent against all global alternatives. Although much progress has been made towards econometric model specification, almost all the literature has been confined to functional forms that only accommodate additive random errors. A few exceptions include Hoderlein, Su and White (2011, **HSW** hereafter) and Lu and White (2011, **LW** hereafter). The former paper proposes a nonparametric test for monotonicity in unobservables in nonparametric nonseparable structural models whereas the latter considers a nonparametric test for additive separability in structural models based on a test for conditional independence. As **LW** argue, many important identification results in the econometrics literature depend on the separability of structural equations, and when correctly imposed, separability helps achieve estimation efficiency in various scenarios. Thus it is desirable to consider tests for separability.

In this chapter we propose a consistent testing procedure that is able to differentiate an additively separable model from a nonadditive one. Like **LW,** we consider

testing the null hypothesis of additive separability in a nonparametric structural model (see eq. (7.1) below) under a conditional exogeneity condition (see Assumption I.3 below). Unlike **LW,** we follow **HSW** and also assume a monotonicity condition to identify the structural equation without imposing additive separability because our testing strategy requires the identification and estimation of the nonparametric structural function under both the null and the alternative. Note that the monotonicity condition is naturally guaranteed under the null but it may not be ensured under the alternative. **LW** do not need to impose such a condition under the alternative because they transform their test of additive separability to a test of conditional independence, which is implied by but in general does not imply the null. So they avoid the identification and estimation of the nonparametric structural model under the alternative. The cost is that their test is not consistent against all global alternatives because of the gap between the implied hypothesis and the original null hypothesis.[1] In contrast, our test is based on the estimate of the partial derivative of the structural function with respect to the unobservable which is identically one under the null and not otherwise. We shall study the asymptotic distributions of our test under the null hypothesis and a sequence of Pitman local alternatives and establish the consistency of our test.

The rest of the chapter is structured as follows. Section 7.2 states our testing problem and presents the test statistic. Section 7.3 provides asymptotic properties of our proposed test. We perform a small set of Monte Carlo experiments in Section 8.6 to investigate the finite sample size and power behavior of our test. In Section 8.8, we conclude and remark on future research. All proofs are relegated to the appendix.

[1]Interestingly, **LW** show that by imposing monotonicity in unobservables for the nonparametric structural function, they can establish the equivalence between the conditional independence and additive separability hypotheses. In this case, their test is also consistent.

## 7.2 Testing Additive Separability

The model of interest can be formulated as

$$Y = m(X, \varepsilon) \tag{7.1}$$

where $Y$ and $X$ are observables, $\varepsilon$ is an unobserved random shock, $m(\cdot, \cdot)$ is an unknown but smooth function defined on $\mathcal{X} \times \mathcal{E}$, where $\mathcal{X} \subset \mathbb{R}^{d_X}$ and $\mathcal{E} \subset \mathbb{R}$. $m(\cdot, \cdot)$ is termed as "nonadditive random function" by Matzkin (2003). We are interested in testing whether the random error $\varepsilon$ enters the model as an additive term.

### 7.2.1 Identification

The model specified in (7.1) is generally not identified without further restriction. For testing purpose, we only consider the situation in which $m(\cdot, \cdot)$ is identified. Matzkin (2003, 2007) studies the identification issue extensively. **HSW** revisit the identification issue and give a set of identification conditions that are analogous to Specification I in Matzkin (2003) but much easier to use. The identification conditions in **HSW** require the existence of a control variable $Z$ such that $X$ is independent of $\varepsilon$ given $Z$, or in short, $X \perp \varepsilon \mid Z$. We shall use $\mathcal{Z}$ to denote the support of $Z$, and $G(\cdot|x, z)$ to denote the conditional cumulative distribution function (CDF) of $Y$ given $(X, Z) = (x, z)$.

Following **HSW**, we make the following identification assumptions.

**Assumption I.1** For all $x \in \mathcal{X}$, $m(x, \cdot)$ is strictly increasing.

**Assumption I.2** There exists $\bar{x} \in \mathcal{X}$ such that $m(\bar{x}, e) = e$ for all $e \in \mathcal{E}$.

**Assumption I.3** $X \perp \varepsilon \mid Z$, where $Z$ is not measurable$-\sigma(X)$.

**Assumption I.4** For each $(x, z) \in \mathcal{X} \times \mathcal{Z}$, $G(\cdot|x, z)$ is invertible.

**Remark 1.** I.1-I.4 parallels Assumptions A.2, A.3, B.1, B.2, respectively, in **HSW**. I.1 and I.3 are also analogous to Assumptions I.2 and I.3 in Matzkin (2003) and I.2 corresponds to their Specification I discussed in their Section 3.1 where an assumption similar to I.4 is also implicitly made.

**Remark 2.** As **HSW** remark, given I.1 and the structural functional relationship in (7.1), for any $\bar{x} \in \mathcal{X}$ there exists a function, say $\bar{m}$, for which I.1 and I.2 hold. This implies that under I.1, any point in $\mathcal{X}$ can play the role of $\bar{x}$ in I.2. Given this $\bar{x}$, we can replace $m$ with $\bar{m}$, such that $\bar{m}(x, \cdot)$ is strictly increasing for all $x \in \mathcal{X}$, and $\bar{m}(\bar{x}, \varepsilon) = \varepsilon$ a.s. With this normalization in mind, we can drop the reference to $\bar{m}$ and simply work with $m$, as what is stated in I.2. In what follows, we simply choose a particular value $\bar{x}$, such as the vector of medians of $X$, and adopt the normalization rule $m(\bar{x}, e) = e$.

The following lemma summarizes some of the identification results in **HSW**.

**Lemma 20** *Suppose (7.1) and Assumptions I.1-I.4 hold. Then*

$$m(x, e) = G^{-1}(G(e|\bar{x}, z) \mid x, z) \qquad \forall (e, x, z) \in \mathcal{E} \times \mathcal{X} \times \mathcal{Z}, \text{ and}$$

$$\varepsilon = G^{-1}(G(Y|X, z) \mid \bar{x}, z) \qquad \forall z \in \mathcal{Z}.$$

The above identification result lays down the foundation for our test of additive separability. It says that under I.1-I.4, the structural response function $m(\cdot, \cdot)$ and the unobserved error term $\varepsilon$ can be identified. In addition, the first result in the above lemma implies that

$$D_e m(x, e) \equiv \frac{\partial m(x, e)}{\partial e} = \frac{g(e \mid \bar{x}, z)}{g(m(x, e) \mid x, z)} \tag{7.2}$$

where $D_e m(\cdot, \cdot)$ denotes the partial derivative of $m(\cdot, \cdot)$ with respect to its second argument, and $g(\cdot|, x, z)$ is the conditional probability density function (PDF) of $Y$ given

$(X, Z) = (x, z)$. Note that the partial derivative $D_e m(x, e)$ is also identified provided $g$ is well defined. Note also that $z$ appears only on the right hand side of (7.2).

### 7.2.2 Hypotheses

Given the model specified in (7.1) we are interested in testing whether $m(\cdot, \cdot)$ is additively separable, that is, whether there exist some measurable functions $m_1(\cdot)$ and $m_2(\cdot)$ such that $m(X, \varepsilon) = m_1(X) + m_2(\varepsilon)$ almost surely (a.s.). Therefore the null hypotheses is

$$\mathbb{H}_0 : m(X, \varepsilon) = m_1(X) + m_2(\varepsilon) \text{ a.s.} \tag{7.3}$$

for some measurable functions $m_1(\cdot)$ and $m_2(\cdot)$, and the alternative hypothesis is the negation of $\mathbb{H}_0$ :

$$\mathbb{H}_1 : P[m(X, \varepsilon) = m_1(X) + m_2(\varepsilon)] < 1 \tag{7.4}$$

for all measurable functions $m_1(\cdot)$ defined on $\mathcal{X}$ and $m_2(\cdot)$ on $\mathcal{E}$.

The simulation experiment in Matzkin (2003) shows that the nonparametric estimate of an additive model without imposing the additive restriction is significantly worse than that with the additive restriction correctly imposed. This highlights the importance of testing the additivity structure of the unknown relationship between the observables and unobservables.

Under I.1, $m_2(\cdot)$ is strictly increasing in (7.3). Given I.2 and $\mathbb{H}_0$ in (7.3), we have

$$m(\bar{x}, \varepsilon) = m_1(\bar{x}) + m_2(\varepsilon) = \varepsilon \text{ a.s.},$$

implying that $m_2(\varepsilon) - \varepsilon$ is a constant with probability one. Therefore we observe that under $\mathbb{H}_0$ and I.1-I.2,

$$D_e m(X, \varepsilon) \equiv \frac{\partial m(X, \varepsilon)}{\partial e} = 1 \text{ a.s.} \tag{7.5}$$

This observation is very important because it motivates us to propose a test based on the derivative of $m(\cdot,\cdot)$ with respect to its second argument. In particular, we will consider a test for $\mathbb{H}_0$ based on the following weighted $L_2$-distance measure between $D_e m(x,e)$ and 1:

$$J = \int \left[D_e m(x,e) - 1\right]^2 a_0(x,e)\, dP(x,e) \tag{7.6}$$

where $P(\cdot)$ is the joint CDF of $X$ and $\varepsilon$ and $a_0(\cdot,\cdot)$ is a nonnegative weight function defined on $\mathcal{X}_0 \times \mathcal{E}_0$, where $\mathcal{X}_0$ and $\mathcal{E}_0$ are a compact subset of $\mathcal{X}$ and $\mathcal{E}$, respectively. [2]

### 7.2.3 Estimation and test statistic

Let $\{(Y_i, X_i, Z_i),\ i = 1, ..., n\}$ denote a random sample for $(Y, X, Z)$ that has support $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$. Let $U_i \equiv (X_i', Z_i')'$. Let $u \equiv (x', z')'$ be a $d \times 1$ vector, $d \equiv d_X + d_Z$, where $x$ is $d_X \times 1$ and $z$ is $d_Z \times 1$. Let $W_i \equiv (Y_i, U_i')'$ and $w \equiv (y, u')'$.

To propose a feasible test statistic, we need to estimate $G(\cdot|u)$, $G^{-1}(\cdot|u)$, $g(\cdot|u)$, and $\varepsilon_i$. Throughout, we rely on local constant estimates.[3] First, we estimate $G(y|u)$ by

$$\hat{G}_b(y|u) \equiv \frac{1}{n}\sum_{i=1}^{n} K_b(U_i - u)\, 1\{Y_i \le y\} / \hat{g}_b(u)$$

where $\hat{g}_b(u) \equiv \frac{1}{n}\sum_{i=1}^{n} K_b(U_i - u)$, $K_b(\cdot) \equiv K(\cdot/b)/b$, $K(\cdot)$ a kernel function defined on $\mathbb{R}^d$, $b \equiv b(n)$ is a bandwidth parameter, and $1\{\cdot\}$ is the usual indicator function. Then we estimate $G^{-1}(\cdot|u)$ by inverting $\hat{G}_b(\cdot|u)$ to obtain

$$\hat{G}_b^{-1}(\cdot\,|u) = \inf\left\{y \in \mathbb{R} : \hat{G}_b(y|u) \ge \cdot\right\},$$

[2] Here and below we restrict $(x, z, e)$ to $\mathcal{X}_0 \times \mathcal{Z}_0 \times \mathcal{E}_0$ because we need to estimate $G(e|x,z)$ and its inverse $G^{-1}(\cdot \mid x, z)$ which can not be estimated sufficiently well if $G(e|x,z)$ is close to either 0 or 1, say, when $(x, z, e)$ lies at the boundary of its support $\mathcal{X} \times \mathcal{Z} \times \mathcal{E}$.

[3] Alternatively one can follow **HSW** and apply the local polynomial method to obtain all necessary estimates. But we find that the local constant method is less computational expensive than the latter.

which is well defined if $K$ is always nonnegative such that $\hat{G}_b(y|u)$ is always between zero and one and monotone in $y$. Nevertheless, to reduce the bias of these kernel estimates, we permit the use of a higher order kernel for $K$ when $d$ is large (e.g., $d \geq 4$). In this case, we may only consider the estimates $\hat{G}_b$ and $\hat{G}_b^{-1}$ on a subset of the observations for which $\hat{G}_b$ lies on a compact subset of $(0,1)$ for large $n$, which is also required in our asymptotic analysis.

Similarly, we estimate the conditional PDF $g(y|u)$ of $Y_i$ given $U_i = u$ by

$$\hat{g}_c(y|u) = \sum_{i=1}^{n} L_c(W_i - w) / \sum_{i=1}^{n} L_c(U_i - u)$$

where $L_c(\cdot) \equiv L(\cdot/c)/c$, $L(\cdot)$ a kernel function defined on $\mathbb{R}^d$ or $\mathbb{R}^{d+1}$, and $c \equiv c(n)$ is a bandwidth parameter.[4]

With $\hat{G}_b$, and $\hat{G}_b^{-1}$ on hand, Lemma 20 motivates us to estimate $m(x,e) = G^{-1}(G(e|\bar{x},z) \mid x,z)$ by

$$\hat{m}_b(x,e) = \int \hat{G}_b^{-1}(\hat{G}_b(e|\bar{x},z)|x,z)\, dH(z) \tag{7.7}$$

and $\varepsilon_i = G^{-1}(G(Y_i|X_i,z) \mid \bar{x},z)$ by

$$\hat{\varepsilon}_i = \int \hat{G}_b^{-1}(\hat{G}_b(Y_i|X_i,z) \mid \bar{x},z) dH(z), \tag{7.8}$$

where $H(\cdot)$ is a CDF that has a PDF $h(\cdot)$ with compact support $\mathcal{Z}_0 \subset \mathcal{Z}$. Note that here we suppress the dependence of $\hat{m}_b$ and $\hat{\varepsilon}_i$ on $H$ and that of $\hat{\varepsilon}_i$ on $b$. Like **HSW**, the use of $H$ helps to eliminate the variability of estimators of $m(x,e)$ and $\varepsilon_i$ based on an arbitrary choice of $z$. In view of the fact that the left hand side of (7.2) does not depend

[4]We abuse the notation a little bit. The multivariate kernel function $L$ can be defined either on $\mathbb{R}^d$ for $U_i$ or $\mathbb{R}^{d+1}$ for $W_i$, which is self evident from its argument.

on $z$, we propose to estimate $D_e m(x,e)$ by[5]

$$\widehat{D_e m}(x,e) = \int \frac{\hat{g}_c(e|\bar{x},z)}{\hat{g}_c(\hat{m}_b(x,e) \mid x,z)} dH(z). \tag{7.9}$$

Based on (7.6), we can consider either

$$\begin{aligned} \tilde{J}_n &= n^{-1}\sum_{i=1}^{n}\left[\widehat{D_e m}(X_i,\hat{\varepsilon}_i)-1\right]^2 a_0\left(X_i,\hat{\varepsilon}_i\right) \\ &= n^{-1}\sum_{i=1}^{n}\left[\int \frac{\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z)}{\hat{g}_c(\hat{m}_b(X_i,\hat{\varepsilon}_i)|X_i,z)}\, dH(z)-1\right]^2 a_0\left(X_i,\hat{\varepsilon}_i\right), \end{aligned} \tag{7.10}$$

or

$$\breve{J}_n = n^{-1}\sum_{i=1}^{n}\left[\int \frac{\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}\, dH(z)-1\right]^2 a_0\left(X_i,\hat{\varepsilon}_i\right). \tag{7.11}$$

Both $\tilde{J}_n$ and $\breve{J}_n$ can be regarded as a sample analogue of $J$ defined in (7.6). The only difference between $\tilde{J}_n$ and $\breve{J}_n$ is that the former relies on the estimate $\hat{m}_b(X_i,\hat{\varepsilon}_i)$ of $m\left(X_i,\varepsilon_i\right)$ whereas the latter utilizes the fact that $m\left(X_i,\varepsilon_i\right)=Y_i$ regardless of whether $\mathbb{H}_0$ holds or not. It turns out the analysis of $\breve{J}_n$ is significantly less complicated than that of $\tilde{J}_n$.

To simplify the analysis further, in view of $\varepsilon_i = m^{-1}\left(X_i,Y_i\right)$ where $m^{-1}\left(x,\cdot\right)$ denotes the inverse of $m\left(x,\cdot\right)\ \forall x\in\mathcal{X}$, we define $a\left(X_i,Y_i\right)=a_0\left(X_i,m^{-1}\left(X_i,Y_i\right)\right)$ and consider the following simpler test statistic

$$\hat{J}_n = n^{-1}\sum_{i=1}^{n}\left[\int \frac{\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}\, dH(z)-1\right]^2 a\left(X_i,Y_i\right). \tag{7.12}$$

Apparently, the support of $a_0$ and that of $a$ are closely related to each other, and the nonnegativity of $a$ is inherited from that of $a_0$. We will make assumptions on the support of $a$ directly so that $\hat{J}_n$ is well defined. Let $\mathcal{Z}_0$ denote the compact support

[5]When $G$ and $G^{-1}$ are estimated by the local polynomial regressions, the asymptotic distributions of $\hat{m}_b(x,e)$, $\hat{\varepsilon}_i$, and $\widehat{D_e m}(x,e)$ are quite complicated and studied in **HSW**.

of $h(\cdot) \equiv \partial H(\cdot)/\partial z$ that is a proper subset of $\mathcal{Z}$. Let $\mathcal{X}_0 \times \mathcal{Y}_0$ denote the compact support of $a(\cdot,\cdot)$ where $\mathcal{X}_0$ and $\mathcal{Y}_0$ are a proper subset $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\mathcal{E}_0$ denote the support of $\varepsilon_i = \int G^{-1}(G(Y_i|X_i,z) \mid \bar{x},z)dH(z)$ when $(X_i,Y_i)$ is constrained to lie on $\mathcal{X}_0 \times \mathcal{Y}_0$. $\hat{G}_b(y|x,z)$ will be bounded away below from 0 and above from 1 for all $(y,x,z) \in \mathcal{Y}_0 \times \mathcal{X}_0 \times \mathcal{Z}_0$ for sufficiently large sample size $n$ by the consistency of $\hat{G}_b$. This will ensure $\hat{\varepsilon}_i = \int \hat{G}_b^{-1}(\hat{G}_b(Y_i|X_i,z) \mid \bar{x},z)dH(z)$ to be well defined for observations with nonzero value of $a(X_i,Y_i)$.

We study the asymptotic distribution of $\hat{J}_n$ in the next section.

## 7.3 Asymptotic Distribution

In this section we first present assumptions that are used in deriving the asymptotic distribution of our test statistic $\hat{J}_n$. Then we study its asymptotic distribution under the null hypothesis and a sequence of Pitman local alternatives. We also prove the consistency of the test and propose a bootstrap method to obtain the bootstrap $p$-value.

### 7.3.1 Assumptions

Let $\boldsymbol{j} \equiv (j_1,...,j_d)$ be a $d$-vector of non-negative integers and $|\boldsymbol{j}| \equiv \sum_{i=1}^{d} j_i$. To study asymptotic distribution of our test statistic, we use the following assumptions.

**Assumption A.1** Let $W_i \equiv (Y_i, X_i', Z_i')'$, $i = 1,2,...,n$, be IID random variables with $W_i$ distributed identically to $(Y, X', Z')$.

Let $g(w)$, $g(u)$, and $g(y|u)$ denote the PDF of $W_i$, that of $U_i$, and the conditional PDF of $Y_i$ given $U_i = u$, respectively. Let $\mathcal{U} \equiv \mathcal{X} \times \mathcal{Z}$ and $\mathcal{U}_0 \equiv \mathcal{X}_0 \times \mathcal{Z}_0$. Let $\mathcal{Y}_0 \equiv [\underline{y}, \bar{y}]$ denote a proper subset of $\mathcal{Y}$.

**Assumption A.2** $(i)$ $g(u)$ is continuous in $u \in \mathcal{U}$, and $g(y|u)$ is continuously differentiable in $y \in \mathcal{Y}$ for all $u \in \mathcal{U}$.

$(ii)$ There exist $C_1, C_2 \in (0, \infty)$ such that $C_1 \leq \inf_{u \in \mathcal{U}_0} g(u) \leq \sup_{u \in \mathcal{U}_0} g(u) \leq C_2$ and $C_1 \leq \inf_{(y,u) \in \mathcal{Y}_0 \times \mathcal{U}_0} g(y|u) \leq \sup_{(y,u) \in \mathcal{Y}_0 \times \mathcal{U}_0} g(y|u) \leq C_2$.

**Assumption A.3** $(i)$ There exist $\underline{\tau}, \bar{\tau} \in (0,1)$ such that $\underline{\tau} \leq \inf_{u \in \mathcal{U}_0} G\left(\underline{y}|u\right) \leq \sup_{u \in \mathcal{U}_0} G(\bar{y}|u) \leq \bar{\tau}$ and $\underline{\tau} \leq \inf_{z \in \mathcal{Z}_0} G\left(\underline{y}|\bar{x}, z\right) \leq \sup_{z \in \mathcal{Z}_0} G(\bar{y}|\bar{x}, z) \leq \bar{\tau}$.

$(ii)$ $G(\cdot|u)$ admits the PDF $g(\cdot|u)$ and is equicontinuous: $\forall \epsilon > 0$, $\exists \delta > 0$ : $|y - \tilde{y}| < \delta \Rightarrow \sup_{u \in \mathcal{U}_0} |G(y|u) - G(\tilde{y}|u)| < \epsilon$. For each $y \in \mathcal{Y}_0$, $G(y \mid \cdot)$ has all partial derivatives up to order $r_1$ where $r_1 \geq 2$ is an even integer.

$(iii)$ Let $D^{\mathbf{j}} G(y|u) \equiv \partial^{|\mathbf{j}|} G(y|u) / \partial^{j_1} u_1 ... \partial^{j_d} u_d$ where $u = (u_1, ..., u_d)'$. For each $y \in \mathcal{Y}_0$, $D^{\boldsymbol{j}} G(y \mid \cdot)$ with $|\boldsymbol{j}| = r_1$ is uniformly bounded and Lipschitz continuous on $\mathcal{U}_0$ : for all $u$, $\tilde{u} \in \mathcal{U}_0$, $|D^{\boldsymbol{j}} G(y \mid u) - D^{\boldsymbol{j}} G(y \mid \tilde{u})| \leq C_3 ||u - \tilde{u}||$ for some $C_3 \in (0, \infty)$ where $\|\cdot\|$ is the Euclidean norm.

$(iv)$ For each $u \in \mathcal{U}_0$ and for all $y$, $\tilde{y} \in \mathcal{Y}_0$, $|D^{\boldsymbol{j}} G(y \mid u) - D^{\boldsymbol{j}} G(\tilde{y} \mid u)| \leq C_4 |y - \tilde{y}|$ for some $C_4 \in (0, \infty)$ where $|\boldsymbol{j}| = r_1$.

**Assumption A.4** The joint PDF $g(w)$ of $W_i$ has all $r_2$th partial derivatives that are uniformly continuous on $\mathcal{Y}_0 \times \mathcal{U}_0$ where $r_2 \geq 2$ is an even integer.

**Assumption A.5** $(i)$ The distribution function $H(z)$ admits a PDF $h(z)$ that is continuous on $\mathcal{Z}_0$.

$(ii)$ The weight function $a(\cdot, \cdot)$ is a nonnegative function that is uniformly bounded on its compact support $\mathcal{X}_0 \times \mathcal{Y}_0$.

**Assumption A.6** $(i)$ For some even integer $r_1 \geq 2$, the kernel $K$ is a product kernel of the bounded symmetric kernel $k : \mathbb{R} \to \mathbb{R}$ satisfying $\int_{\mathbb{R}} v^i k(v) dv = \delta_{i0}$ $(i = 0, 1, ..., r_1 - 1)$, $\int_{\mathbb{R}} v^{r_1} k(v) dv < \infty$, and $k(v) = O\left((1 + |v|^{r_1 + 1 + \epsilon})^{-1}\right)$ for some $\epsilon > 0$, where $\delta_{ij}$

is Kronecker's delta.

(*ii*) For some even integer $r_2 \geq 2$, the kernel $L$ is a product kernel of the bounded symmetric kernel $l : \mathbb{R} \to \mathbb{R}$ satisfying $\int_{\mathbb{R}} v^i l(v)dv = \delta_{i0}$ $(i = 0, 1, ..., r_2 - 1)$, $\int_{\mathbb{R}} v^{r_2} l(v)dv < \infty$, and $l(v) = O\left((1 + |v|^{r_2+1+\epsilon})^{-1}\right)$ for some $\epsilon > 0$.

**Assumption A.7** As $n \to \infty$, $b \to 0$, $c \to 0$, and the following conditions are satisfied:

(*i*) $nc^{d_X+1}/\log n \to \infty$, $nc^{2r_2+d_X+\frac{1}{2}} \to 0$, $nb^{2r_1}c^{d_X+\frac{1}{2}} \to 0$,

(*ii*) $nb^{d_X}c/(\log n)^2 \to \infty$, $nb^{2(d_X+d_Z)}c^{-d_X-\frac{1}{2}}/(\log n)^2 \to \infty$,

(*iii*) $c^{\frac{1}{2}}(c/b)^{d_X}\left[1 + c^{2r_2}\log n\right] \to 0$, and $b^{2r_1}c^{-\frac{1}{2}}\log n \to 0$.

A.1-A.3 parallel Assumptions C.1-C.3 in **HSW**. As in **HSW**, the IID requirement in A.1 is standard in cross-section studies but can be relaxed to allow for weakly dependent time series observations. A.2-A.4 and A.6 are standard for nonparametric local constant estimation of conditional CDF and PDF when a higher order kernel may be called upon. Note that we permit the use of higher order kernel for either $K$ and $L$ but neither is necessary if $d = d_X + d_Z$ is small, see the discussions below. A.5 specifies the weak conditions on the probability weight $H$ and the weight function $a(\cdot, \cdot)$. In the simulations we simply choose $H$ to be a scaled beta distribution that has a compact support $\mathcal{Z}_0$ and specify $a$ as an indicator function with compact support $\mathcal{X}_0 \times \mathcal{Y}_0$. A.7 appropriately restricts the choices of bandwidth sequences and the orders of kernel functions.

Note that if we choose $b = c \propto n^{-1/\alpha}$ for some $\alpha > 0$, then A.7(*iii*) is automatically satisfied and A.7(i)-(ii) would require $nb^{d_X+1}/(\log n)^2 \to \infty$, $nb^{2r_2+d_X+\frac{1}{2}} \to 0$, $nb^{2r_1+d_X+\frac{1}{2}} \to 0$, and $nb^{d_X+2d_Z-\frac{1}{2}}/(\log n)^2 \to \infty$. The last set of conditions are met provided

$$d_X + 2d_Z - \frac{1}{2} < \alpha < d_X + \frac{1}{2} + 2\min(r_1, r_2). \tag{7.13}$$

Apparently, (7.13) requires $\min(r_1, r_2) > d_Z - \frac{1}{2}$. In the case where $d_Z = 1$ or 2, we can choose $r_1 = r_2 = 2$ and $\alpha \in (d_X + 2d_Z - \frac{1}{2},\, d_X + \frac{9}{2})$ such that (7.13) is satisfied. In this case, there is no need to use higher order kernels for either $K$ or $L$.

More generally, we can consider choosing $b \propto n^{-1/\alpha}$ and $c \propto n^{-\kappa/\alpha}$. Then A.7 would require

$$\max\left\{(d_X+1)\kappa,\ d_X+\kappa,\ 2(d_X+d_Z)-(d_X+\frac{1}{2})\kappa\right\} < \alpha < \min\left\{d_X+\frac{1}{2}+2r_2,\ (d_X+\frac{1}{2})\kappa+2r_1\right\}$$

where $d_X/\left(d_X+\frac{1}{2}\right) < \kappa < 4r_1$. Due to the "curse of dimensionality" in nonparametric estimation, we expect that typical values of $d_X$ and $d_Z$ are 1, 2, or 3 such that $d_X + d_Z \leq 4$ for realistic applications, in which case we can verify that the above conditions can be satisfied for a variety of combinations for $\alpha, \kappa$, $r_1$ and $r_2$. In particular, to ensure the conditional CDF estimate $\hat{G}_b(y|x,z)$ to lie between zero and 1 and to be monotone in $y$, it is always possible to restrict our attention to the use of a second order kernel for $K$ (i.e., $r_1 = 2$) for properly chosen $\alpha$, $\kappa$ and $r_2$. In particular, if $d_Z \leq 2$, we recommend using the same second order kernel for $K$ and $L$ (implying that $r_1 = r_2 = 2$) and setting $b = c \propto n^{-1/\alpha}$. So one only needs to choose a sequence of bandwidth.

### 7.3.2 Asymptotic null distribution

In this section, we study the asymptotic behavior of the test statistic in (7.12). To state the next result, we write $\tilde{w} \equiv (\tilde{y}, \tilde{x}', \tilde{z}')'$ and introduce the following notation:

$$\zeta_0(W_i, W_j) \equiv \int g(Y_i|X_i,z)^{-1}\left[g(\bar{x},z)^{-1}\bar{L}_{cj,(\varepsilon_i,\bar{x},z)} - g(X_i,z)^{-1}\bar{L}_{cj,(Y_i,X_i,z)}\right]dH(z) \tag{7.14}$$

and

$$\varphi(w,\tilde{w}) \equiv E\left[\zeta_0(W_i,w)\,\zeta_0(W_i,\tilde{w})\,a(X_i,Y_i)\right], \tag{7.15}$$

where $\bar{L}_{ci,w} = L_{ci,w} - E(L_{ci,w})$, and $L_{ci,w} = L_c(W_i - w)$.[6] We define the asymptotic bias and variance respectively by

$$\mathbb{B}_n \equiv n^{-1} c^{d_X + \frac{1}{2}} \sum_{i=1}^{n} \varphi(W_i, W_i) \text{ and } \sigma_n^2 = 2c^{2d_X+1} E[\varphi(W_1, W_2)^2].$$

The following theorem establishes the asymptotic null distribution of the $\hat{J}_n$ test statistic.

**Theorem 21** *Suppose Assumptions I.1-I.4 and A.1-A.7 hold. Then under* $\mathbb{H}_0$, *we have* $nc^{d_X+\frac{1}{2}}\hat{J}_n - \mathbb{B}_n \xrightarrow{d} N\left(0, \sigma_0^2\right)$, *where* $\sigma_0^2 \equiv \lim_{n\to\infty} \sigma_n^2$.

The proof of the above theorem is extremely involved. After a long and arduous effort, we can demonstrate that the key building block in obtaining the asymptotic bias and variance of the test statistic $\hat{J}_n$ is $\zeta_0(W_i, W_j)$. The first term, $g(Y_i|X_i, z)^{-1} g(\bar{x}, z)^{-1} \bar{L}_{cj,(\varepsilon_i,\bar{x},z)}$, in the definition of $\zeta_0$ reflects the influence of the numerator estimator $\hat{g}_c(\varepsilon_i|\bar{x}, z)$ in the definition of $\hat{J}_n$ in (7.12), whereas the second term $g(Y_i|X_i, z)^{-1} g(X_i, z)^{-1} \bar{L}_{cj,(Y_i,X_i,z)}$ embodies the effect of the denominator estimator $\hat{g}_c(Y_i|X_i, z)$. Like the test statistic in **HSW**, these two terms contribute to the asymptotic bias of $\hat{J}_n$ symmetrically but to the asymptotic variance asymmetrically due to different roles played by $\bar{x}$ (the normalization point) and $X_i$ (data). A careful analysis of $\mathbb{B}_n$ indicates that both terms contribute to the asymptotic bias of $\hat{J}_n$ to the order of $O\left(c^{-1/2}\right)$. On the other hand, a detailed study of $\sigma_n^2$ shows that they contribute asymmetrically to the asymptotic variance: the asymptotic variance of $\hat{J}_n$ is mainly determined by the numerator estimator, whereas the role played by the denominator estimator is asymptotically negligible. See **HSW** for further discussion of similar phenomena in a different context. They also explain

[6] Even though $X_i, Z_i, Y_i$, and $\varepsilon_i$ all enter the definition of $\zeta_0$, we can still use $W_i = (Y_i, X_i', Z_i')'$ to summarize these variables because $\varepsilon_i = m^{-1}(X_i, Y_i)$ is measurable under Assumption I.1 and the continuity of $m(\cdot, \cdot)$.

why we need $c^{d_X+\frac{1}{2}}$ instead of the usual term $c^{(d_X+1)/2}$ as the normalization constant in the front of $\hat{J}_n$, which unavoidably reduces the size of the class of local alternatives that this test has power to detect.

To implement, we need consistent estimates of the asymptotic bias and variance. Let

$$\hat{\zeta}_0(W_i, W_k) \equiv \int \hat{g}_c(Y_i|X_i, z)^{-1} \left[\hat{g}_c(\bar{x}, z)^{-1} \hat{L}_{cj,(\hat{\varepsilon}_i,\bar{x},z)} - \hat{g}_c(X_i, z)^{-1} \hat{L}_{cj,(Y_i,X_i,z)}\right] dH(z)$$

where $\hat{L}_{cj,w} = L_{cj,w} - \frac{1}{n}\sum_{k=1}^{n} L_{ck,w}$, and $\hat{g}_c(x, z)$ is a kernel estimator of the PDF $g(x, z)$ by using kernel $L$ and bandwidth $c$. We propose estimating the asymptotic bias $\mathbb{B}_n$ by

$$\hat{\mathbb{B}}_n = n^{-2} c^{d_X+\frac{1}{2}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left[\hat{\zeta}_0(W_i, W_j)\right]^2 a(X_i, Y_i)$$

and the asymptotic variance $\sigma_n^2$ by

$$\hat{\sigma}_n^2 = \frac{2c^{2d_X+1}}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \left[\frac{1}{n}\sum_{l=1}^{n} \hat{\zeta}_0(W_l, W_i)\hat{\zeta}_0(W_l, W_j) a(X_l, Y_l)\right]^2.$$

It is tedious but straightforward to show $\hat{\mathbb{B}}_n - \mathbb{B}_n = o_P(1)$ and $\hat{\sigma}_n^2 - \sigma_n^2 = o_P(1)$. Then the following feasible test statistic

$$T_n \equiv \left(nc^{d_X+\frac{1}{2}}\hat{J}_n - \hat{\mathbb{B}}_n\right) / \sqrt{\hat{\sigma}_n^2} \tag{7.16}$$

is asymptotically distributed as $N(0, 1)$ and we reject the null for large value of $T_n$.

### 7.3.3 Local power property and consistency

To study the local power of the $T_n$ test, consider the sequence of Pitman local alternatives:

$$\mathbb{H}_1(\gamma_n) : D_e m(x, e) = 1 + \gamma_n \delta_n(x, e), \tag{7.17}$$

where $\gamma_n \to 0$ as $n \to \infty$, and $\delta_n$ is a non-constant measurable function with $\mu_0 \equiv \lim_{n\to\infty} E[\delta_n(X_1, \varepsilon_1)^2 a(X_1, Y_1)] < \infty$.

**Theorem 22** *Suppose Assumptions I.1-I.4 and A.1-A.7 hold. Then under* $\mathbb{H}_1(\gamma_n)$ *with* $\gamma_n = n^{-1/2}c^{-d_X/2-1/4}$, $T_n \xrightarrow{d} N(\mu_0/\sigma_0, 1)$. *That is, the asymptotic local power function of* $T_n$ *is given by* $P(T_n > z|\mathbb{H}_1(\gamma_n)) = 1-\Phi(z-\mu_0/\sigma_0)$, *where* $\Phi$ *is the standard normal CDF.*

Theorem 22 implies that the $T_n$ test has non-trivial power against Pitman local alternatives that converge to zero at rate $n^{-1/2}c^{-d_X/2-1/4}$, provided $0 < \mu_0 < \infty$. As remarked above, this rate is different from the usual nonparametric rate $n^{-1/2}c^{-(d_X+1)/4}$ or $n^{-1/2}c^{-(d_X+d_Z+1)/4}$ when $(d_X+1)$ or $(d_X+d_Z+1)$ dimensional nonparametric objects need to be estimated.

The following theorem shows that the test is consistent.

**Theorem 23** *Suppose Assumptions I.1-I.4 and A.1-A.7 hold. Suppose that* $\mu_A \equiv E\{[D_e m(X_i, \varepsilon_i) - 1]^2 a(X_i, Y_i)\} > 0$. *Then* $P(T_n > \lambda_n) \to 1$ *as* $n \to \infty$ *for any non-stochastic sequence* $\lambda_n = o(nc^{d_X+1/2})$.

### 7.3.4 A bootstrap version of the test

It is well known that nonparametric tests based on their asymptotic normal null distributions may perform poorly in finite samples. As an alternative, we can rely on bootstrap to obtain a bootstrap $p$-value.

To obtain the bootstrap replicates of $W_i = (Y_i, X_i', Z_i')'$, we need to impose various restrictions. First, we need to impose the identification conditions given in Assumption I.1 and I.3. Fortunately, these can be well handled by following the local smooth bootstrap procedure of **HSW** (see also Su and White (2008)). Second, we need to impose the null of additive separability. In view of the discussion in Section 2.2, under

$\mathbb{H}_0$ and Assumption I.2, we have

$$m(x,e) = \bar{m}_1(x) + e$$

for some measurable function $\bar{m}_1$ whose exact structure depend on the choice of the normalization point $\bar{x}$. This motivates us to estimate $\bar{m}_1(x)$ by

$$\hat{m}_{1,b}(x) = \int \hat{m}_b(x,e) dQ(e)$$

where $Q(\cdot)$ is a proper CDF on $\mathbb{R}$. Then $\hat{m}_{1,b}(x)$ is consistent for $\bar{m}_1(x) + \int e dQ(e)$ provided $\hat{m}_b(x,e)$ is consistent for $m(x,e)$. The last claim can be established as in **HSW** and the term $\int e dQ(e)$ is constant, which does not affect the asymptotic distribution of our bootstrap test statistic if we generate the bootstrap data $Y_i^*$ through this relationship. See Step 3 below.

Let $\mathcal{W}_n \equiv \{W_i\}_{i=1}^n$. Following Su and White (2008) and **HSW**, we draw bootstrap resamples $\{X_i^*, Y_i^*, Z_i^*\}_{i=1}^n$ based on the following smoothed local bootstrap procedure:

1. For $i = 1, ..., n$, obtain a preliminary estimate of $\varepsilon_i$ as $\hat{\varepsilon}_i = \int \hat{G}_b^{-1}(\hat{G}_b(Y_i|X_i,z) \mid \bar{x}, z) dH(z)$.

2. Draw a bootstrap sample $\{Z_i^*\}_{i=1}^n$ from the smoothed kernel density $\tilde{f}_Z(z) = n^{-1}\sum_{i=1}^n \mathbf{\Phi}_{\alpha_z}(Z_i - z)$, where $\mathbf{\Phi}_\alpha(z) = \alpha^{-d_Z}\mathbf{\Phi}(z/\alpha)$ where $\mathbf{\Phi}(\cdot)$ is a product kernel formed from the standard normal PDF $\phi(\cdot)$, and $\alpha_z > 0$ is a bandwidth parameter.

3. For $i = 1, ..., n$, given $Z_i^*$, draw $X_i^*$ and $\varepsilon_i^*$ independently from the smoothed conditional density $\tilde{f}_{X|Z}(x|Z_i^*) = \sum_{j=1}^n \mathbf{\Phi}_{a_x}(X_j - x)\mathbf{\Phi}_{\alpha_z}(Z_j - Z_i^*) / \sum_{l=1}^n \mathbf{\Phi}_{\alpha_z}(Z_l - Z_i^*)$

and $\tilde{f}_{\varepsilon|Z}(e|Z_i^*) = \sum_{j=1}^n \mathbf{\Phi}_{\alpha_e}(\hat{\varepsilon}_j - e)\Phi_{\alpha_z}(Z_j - Z_i^*) / \sum_{l=1}^n \mathbf{\Phi}_{\alpha_z}(Z_l - Z_i^*)$, respectively, where $\alpha_z$, $\alpha_x$, and $\alpha_e$ are given bandwidths.[7]

4. For $i = 1, ..., n$, generate the bootstrap analogue of $Y_i$ as $Y_i^* = \hat{m}_{1,b}(X_i^*) + \varepsilon_i^*$.

5. Compute a bootstrap statistic $T_n^*$ in the same way as $T_n$ with $\{(Y_i^*, X_i^*, Z_i^*)\}_{i=1}^n$ replacing $\mathcal{W}_n$.

6. Repeat Steps 2-5 $B$ times to obtain bootstrap test statistics $\left\{T_{nj}^*\right\}_{j=1}^B$. Calculate the bootstrap $p$-values $p^* \equiv B^{-1}\sum_{j=1}^B 1\left(T_{nj}^* \geq T_n\right)$ and reject the null hypothesis if $p^*$ is smaller than the prescribed nominal level of significance.

## 7.4 Monte Carlo Simulations

In this section, we conduct a small set of Monte Carlo simulations to examine the finite sample performance of our test. We first consider the following two data generating processes (DGPs) for the level study:

DGP 1: $Y_i = X_i + \varepsilon_i$,

DGP 2: $Y_i = \Phi(X_i) - \frac{1}{2} + \varepsilon_i$,

where $i = 1, \ldots, n$, $\Phi(\cdot)$ is the standard normal CDF, $X_i = 0.25 + Z_i - 0.25Z_i^2 + \upsilon_{1i}$, $\varepsilon_i = 0.5Z_i + \upsilon_{2i}$ and $Z_i, \upsilon_{1i}$ and $\upsilon_{2i}$ are IID $N(0,1)$ and mutually independent. Clearly, the error terms in DGPs 1-2 are additively separable and we use the above two DGPs to evaluate the finite sample level behavior of our test. Note that

$$m(x,e) = \begin{cases} x + e & \text{in DGP 1,} \\ \Phi(x) - \frac{1}{2} + e & \text{in DGP 2.} \end{cases}$$

[7] We abuse the notation $\mathbf{\Phi}$ a little bit here: $\mathbf{\Phi}_\alpha(z) = \alpha^{-d_Z}\mathbf{\Phi}(z/\alpha)$ and $\mathbf{\Phi}_\alpha(x) = \alpha^{-d_X}\mathbf{\Phi}(z/\alpha)$. So the argument of $\mathbf{\Phi}$ can be of dimension $d_X$ or $d_Z$. The bandwidths here are all set according to the Silverman's rule of thumb in our simulations below.

In both designs, $m(x,\cdot)$ is strictly monotone for each $x$ and $m(\bar{x},e)=e$ for $\bar{x}=0$. The other two identification conditions used throughout the chapter are easily verified.

To study the finite sample power behavior of our test, we consider the following four DGPs:

DGP 3: $Y_i=(0.5+0.1X_i^2)\varepsilon_i$,

DGP 4: $Y_i=\Phi\left(\left(X_i+1\right)\varepsilon_i/4\right)\left(X_i+1\right)$,

DGP 5: $Y_i=X_i+\varepsilon_i-\frac{\delta X_i^2}{0.1+\exp(\varepsilon_i)}$,

DGP 6: $Y_i=\Phi\left(X_i\right)-\frac{1}{2}+\varepsilon_i-\frac{\delta(\sin X_i)^2}{0.1+\varepsilon_i^3}$,

DGP 7: $Y_i=X_i+\varepsilon_i+\frac{\delta X_i}{0.1+\exp(\varepsilon_i)}$,

DGP 8: $Y_i=\Phi\left(X_i\right)-\frac{1}{2}+\varepsilon_i+\frac{\delta\sin X_i}{0.1+\varepsilon_i^2}$,

where $i=1,\ldots,n$, $X_i$, $\varepsilon_i$ and the instrument $Z_i$ are generated as in DGPs 1-2, and $\delta$ is a parameter that adapts the corresponding DGP for different simulation purposes.

DGPs 3 and 4 are used by **HSW** to test for the monotonicity in the unobservable ($\varepsilon_i$ here). It is easy to verify that the identification conditions specified in Assumptions I.1-I.4 are all satisfied for DGPs 3-4. But these two DGPs do not satisfy the additive separability condition.

When $\delta=0$, DGPs 5 and 7 (resp. DGPs 6 and 8) reduce to DGP 1 (resp. DGP 2). For other values of $\delta$, the structural function $m(x,e)$ implied by DGPs 5-8 is not additively separable in error terms. In addition, DGPs 5 and 6 satisfy all the identification conditions specified in Assumptions I.1-I.4; DGPs 7 and 8 violate Assumption I.1 and but satisfies the other identification conditions (e.g., $m(0,e)=e$ regardless of the value of $\delta$ in DGPs 7-8). In this case, we can investigate the robustness of the finite sample power behavior of our test against non-monotonicity under the alternative.

To construct our standardized test statistic $T_n$ in (7.16), we need to compute sequentially $\hat{J}_n$, $\hat{\mathbb{B}}_n$ and $\hat{\sigma}_n$. We first obtain local constant estimates $\hat{G}_b(y|u)$, $\hat{G}_b^{-1}(\tau|u)$, $\hat{g}_c(y|u)$ and $\hat{\varepsilon}_i = \int \hat{G}_b^{-1}(\hat{G}_b(Y_i|X_i,z) \mid \bar{x},z)dH(z)$ by using standard normal kernel function and Silverman's rule of thumb for bandwidth choice, i.e., $b = c = \left(1.06S_X n^{-1/5}, 1.06S_Z n^{-1/5}\right)$ with $S_X$ and $S_Z$ being the sample standard deviation of $\{X_i\}$ and $\{Z_i\}$, respectively. We choose $H(z)$ to be a scaled beta(3,3) distribution on $[\zeta_\kappa, \zeta_{1-\kappa}]$, where $\zeta_\kappa$ denotes the $\kappa$-th sample quantile of $\{Z_i\}$ and $\kappa = 0.05$. $N = 30$ evenly-spaced points are chosen for numerical integration. We set $a(X_i,Y_i) = 1\{\zeta_{\lambda,X} \leq X_i \leq \zeta_{1-\lambda,X}\} \times 1\{\zeta_{\lambda,Y} \leq Y_i \leq \zeta_{1-\lambda,Y}\}$, where, e.g., $\zeta_{\lambda,X}$ is the $\lambda$-th sample quantile of $\{X_i\}$ and $\lambda = 0.0125$. For the computation of $\hat{\mathbb{B}}_n$ and $\hat{\sigma}_n$, we need to further compute $\hat{g}_c(x,z)$ with a standard normal kernel function and bandwidth $c$ chosen as before. The same trimming function $a(X_i,Y_i)$ and weight function $H(z)$ are utilized everywhere.

To obtain the bootstrap $p$-values, we follow the procedure stated in Section 3.4 to compute the rejecting probabilities. We consider two sample sizes ($n = 100$ and 200) with 250 replications. Due to the high computational burden, we only use $B = 100$ bootstrap resamples in each replication. Before conducting the bootstrap with $B = 100$, we study the sensitivity of the test to the bandwidth $b$ as suggested by Giacomini, Politis and White (2007), using the warp-speed bootstrap procedure based on a single bootstrap resample. We find that the our test is not very sensitive to the choice of $b = \left(c_1 S_X n^{-1/5}, c_1 S_Z n^{-1/5}\right)$ as long as $c_1$ is between 1 and 2. We report the results for $c_1 = 1.06$. In addition, we consider $\delta = 1$ in DGPs 5-8.

Table 1 reports the empirical level of our bootstrapped test for DGPs 1-2 where the nominal levels are 1%, 5% and 10%. We see that the level of our test is

Table 7.1: Empirical level for DGPs 1-2

| DGP | $n$ | 1% | 5% | 10% |
|---|---|---|---|---|
| 1 | 100 | 0.008 | 0.044 | 0.116 |
| | 200 | 0.012 | 0.048 | 0.112 |
| 2 | 100 | 0.004 | 0.040 | 0.088 |
| | 200 | 0.008 | 0.052 | 0.108 |

Table 7.2: Empirical power for DGPs 3-8

| DGP | $n$ | 1% | 5% | 10% |
|---|---|---|---|---|
| 3 | 100 | 0.908 | 0.992 | 0.996 |
| | 200 | 0.980 | 0.996 | 1 |
| 4 | 100 | 0.996 | 1 | 1 |
| | 200 | 1 | 1 | 1 |
| 5 | 100 | 0.896 | 0.912 | 0.984 |
| | 200 | 0.924 | 0.952 | 0.992 |
| 6 | 100 | 0.872 | 0.904 | 0.952 |
| | 200 | 0.916 | 0.936 | 0.988 |
| 7 | 100 | 0.440 | 0.468 | 0.492 |
| | 200 | 0.484 | 0.524 | 0.572 |
| 8 | 100 | 0.476 | 0.544 | 0.648 |
| | 200 | 0.504 | 0.584 | 0.692 |

fairly well behaved and it gets closer to the nominal level as the sample size increases. Table 2 presents the empirical power of our bootstrapped test at various nominal levels. Surprisingly our test has fantastic power to reject additive separability for DGPs 3-4. The power is also reasonably good and increases as the sample size doubles in DGPs 5-8. Comparing the results for DGPs 7-8 with DGPs 5-6, we observe that the power performance of our test is adversely affected by the violation of the monotonicity assumption.

## 7.5 Concluding Remarks

The prevalent additivity error structure has been an important assumption in many economic and econometric models. This chapter develops a simple consistent test to detect whether this critical assumption holds in the presence of economic data. The

test is motivated from the simple observation that the partial derivative of the unknown structural function with respect to the unobserved error term is one under the null hypothesis of additive separability and certain identification conditions. We derive the asymptotic distributions of our test statistic under the null and a sequence of Pitman local alternatives and prove its consistency. We also propose a bootstrap version of the test. Monte Carlo simulations are conducted to examine the finite sample performance of the bootstrapped test. The test enjoys proper size and reasonable power in finite samples.

There are some interesting topics for further research. First, under the same set of identification conditions considered in this chapter, one can develop other tests for additive separability. For example, one may consider a test based on the observation that the cross derivatives with respect to the regressor and the error term is zero under additivity. But this would need consistent estimate of cross derivatives and thus is expected to be less powerful. For another example, we can consider the estimation of the structural function under both the null and the alternative, and base a test on the weighted $L_2$ distance between these estimates. To this goal, one needs to develop an estimate of the structural function under the additive separability condition. Under Assumption I.3 and the null: $m(X,\varepsilon) = \bar{m}_1(X) + \varepsilon$, $E(Y|X,Z) = \bar{m}_1(X) + E(\varepsilon|Z)$. This motivates us to obtain a consistent estimate $\tilde{m}_1(x)$ of $\bar{m}_1(x)$ by using either the marginal integration or backfitting technique. Then we can compare this estimate with $\hat{m}_{1,b}(x)$ used in Section 3.4. The theoretical study of this test is left for future research.

Second, one may consider relaxing some of the identification conditions used to identify and estimate the nonparametric structural function under the alternative. For example, one may follow **LW** and relax the monotonicity assumption. The problem is that without monotonicity, one cannot identify $m(x,e)$ or its partial derivative with

respect to $e$ under the alternative without further assumptions. It is interesting to know whether it is possible to develop a consistent test in this case. Alternatively, one may consider relaxing the conditional exogeneity condition: $X \perp \varepsilon \mid Z$. Again, without this assumption, one cannot identify $m(x, e)$ or its partial derivative as in this chapter. Some other assumptions have to be in place.

# Appendix

## Proof of Some Technical Lemmas

In this appendix, we prove some technical lemmas that are used in the establishment of the main results in Section 3.

Recall that $\mathcal{U}_0 \equiv \mathcal{X}_0 \times \mathcal{Z}_0$, $U_i \equiv (X_i', Z_i')'$, $u \equiv (x', z')'$, $W_i \equiv (Y_i, U_i')'$ and $w \equiv (y, u')'$. Let $1_i \equiv 1\{X_i \in \mathcal{X}_0, Y_i \in \mathcal{Y}_0\}$. Define

$$\begin{aligned} V_{n,b}(y;u) &\equiv \frac{1}{n}\sum_{i=1}^{n} K_b(U_i - u)\left[1\{Y_i \le y\} - G(y|U_i)\right] = \frac{1}{n}\sum_{i=1}^{n} K_{bi,u}\bar{\mathbf{1}}_i(y), \\ \mathbf{V}_{n,c}(y;u) &\equiv \frac{1}{n}\sum_{i=1}^{n}\{L_c(W_i - w) - E\left[L_c(W_i - w)\right]\} = \frac{1}{n}\sum_{i=1}^{n}\bar{L}_{ci,w}, \end{aligned}$$

where $K_{bi,u} \equiv K_b(U_i - u)$, $L_{ci,w} = L_c(W_i - w)$, $\bar{L}_{ci,w} = L_{ci,w} - E(L_{ci,w})$, and $\bar{\mathbf{1}}_i(y) = 1\{Y_i \le y\} - G(y|U_i)$. Let

$$\nu_{1b} \equiv n^{-1/2}b^{-d_X/2}\sqrt{\log n},\ \nu_{2b} \equiv n^{-1/2}b^{-(d_X+d_Z)/2}\sqrt{\log n},\ \text{and}\ \nu_{3b} \equiv n^{-1/2}b^{-(d_X+d_Z+1)/2}\sqrt{\log n}.$$

$\nu_{1c}$, $\nu_{2c}$, and $\nu_{3c}$ are similarly defined.

**Lemma 24** *Suppose that Assumptions A.1-A.3, A.6(i) and A.7 hold. Let $\mathcal{T}_0 = [\underline{\tau}, \bar{\tau}]$ denote a compact subset of $(0,1)$. Then*

*(a) $\hat{G}_b(y|u) - G(y|u) = g(u)^{-1} V_{n,b}(y;u) + O_P(\nu_{2b}^2 + b^{r_1})$ uniformly in $(y,u) \in \mathbb{R} \times \mathcal{U}_0$,*

*(b) $\hat{G}_b^{-1}(\tau|u) - G^{-1}(\tau|u) = O_P(\nu_{2b} + b^{r_1})$ uniformly in $(\tau,u) \in \mathcal{T}_0 \times \mathcal{U}_0$,*

*(c) $\hat{G}_b^{-1}(\tau|u) - G^{-1}(\tau|u) = -\frac{V_{n,b}(G^{-1}(\tau|u);u)\{1+o(1)\}}{g(G^{-1}(\tau|u)|u)g(u)} + O_P(\nu_{2b}^2 + b^{r_1})$ uniformly in $(\tau,u) \in \mathcal{T}_0 \times \mathcal{U}_0$.*

**Proof.** For (a), we make the following bias-variance decomposition:

$$\hat{G}_b(y|u) - G(y|u) = \hat{g}_b(u)^{-1}\frac{1}{n}\sum_{i=1}^{n} K_b(U_i - u)\left[G(y|U_i) - G(y|u)\right] + \hat{g}_b(u)^{-1} V_{n,b}(y;u)$$

By Assumptions A.1-A.3 and A.6(i) and the standard arguments in kernel estimation (e.g., Masry (1996a, 1996b), Hansen (2008)), $\sup_{u\in\mathcal{U}_0}|\hat{g}_b(u) - g(u)| = O_P(\nu_{2b} + b^{r_1})$,

$\sup_{u\in\mathcal{U}_0}|\frac{1}{n}\sum_{i=1}^n K_b(U_i-u)[G(y|U_i)-G(y|u)]| = O_P(b^{r_1})$, and $\sup_{u\in\mathcal{U}_0}|V_{n,b}(y;u)| = O_P(\nu_{2b})$. It follows that uniformly in $u\in\mathcal{U}_0$,

$$\hat{G}_b(y|u) - G(y|u) = g(u)^{-1} V_{n,b}(y;u) + O_P\left(b^{r_1}+\nu_{2b}^2\right).$$

By the same argument as used in the proof of Theorem 4.1 of Boente and Fraiman (1991), we can show that the last result also holds uniformly in $y\in\mathbb{R}$ under Assumption A.3.

For (b), noting that $\hat{G}_b\left(\hat{G}_b^{-1}(\tau|u)|u\right) = \tau = G\left(G^{-1}(\tau|u)|u\right)$, we have

$$\left|G\left(\hat{G}_b^{-1}(\tau|u)|u\right) - G\left(G^{-1}(\tau|u)|u\right)\right| = \left|G\left(\hat{G}_b^{-1}(\tau|u)|u\right) - \hat{G}_b\left(\hat{G}_b^{-1}(\tau|u)|u\right)\right| \le \sup_{y\in\mathbb{R}}\left|G(y|u) - \hat{G}_b(y|u)\right|$$

So the pointwise consistency of $\hat{G}_b^{-1}(\tau|u)$ follows from that of $\hat{G}_b(y|u)$ and the continuity of $G(\cdot|u)$. By Assumption A.3(ii) and the first order Taylor expansion,

$$G\left(\hat{G}_b^{-1}(\tau|u)|u\right) - G\left(G^{-1}(\tau|u)|u\right) = \left[\hat{G}_b^{-1}(\tau|u) - G^{-1}(\tau|u)\right] g\left(\tilde{G}^{-1}(\tau|u)|u\right)$$

where $\tilde{G}^{-1}(\tau|u)$ lies between $\hat{G}_b^{-1}(\tau|u)$ and $G^{-1}(\tau|u)$. Therefore by (a) and Assumption A.2(ii)

$$\begin{aligned}\sup_{(\tau,u)\in\mathcal{T}_0\times\mathcal{U}_0}\left|\hat{G}_b^{-1}(\tau|u) - G^{-1}(\tau|u)\right| &\le \frac{\sup_{(\tau,u)\in\mathcal{T}_0\times\mathcal{U}_0}\left|G\left(\hat{G}_b^{-1}(\tau|u)|u\right) - G\left(G^{-1}(\tau|u)|u\right)\right|}{\inf_{(\tau,u)\in\mathcal{T}_0\times\mathcal{U}_0} g\left(\tilde{G}^{-1}(\tau|u)|u\right)} \\ &\le \frac{\sup_{u\in\mathcal{U}_0}\sup_{y\in\mathbb{R}}\left|G(y|u) - \hat{G}_b(y|u)\right|}{\inf_{(\tau,u)\in\mathcal{T}_0\times\mathcal{U}_0} g\left(\tilde{G}^{-1}(\tau|u)|u\right)} = O_P(\nu_{2b}+b^{r_1}).\end{aligned}$$

To obtain the uniform Bahadur representation for $\hat{G}_b^{-1}(\tau|u)$, we apply the Hadamard differentiability of the (conditional) quantile operator (see e.g., Doss and Gill (1992, Theorem 1)) to obtain

$$\hat{G}_b^{-1}(\tau|u) - G^{-1}(\tau|u) = \frac{\hat{G}_b(G^{-1}(\tau|u)|u) - \tau}{g\left(G^{-1}(\tau|u)|u\right)}\{1+o(1)\}.$$

This together with (a) implies that $\hat{G}_b^{-1}(\tau|u) - G^{-1}(\tau|u) = -\frac{V_{n,b}(G^{-1}(\tau|u);u)\{1+o(1)\}}{g(G^{-1}(\tau|u)|u)g(u)} + O_P(\nu_{2b}^2+b^{r_1})$. ■

If $G(y|x,z) \in \mathcal{T}_0 = [\underline{\tau}, \bar{\tau}] \subset (0,1)$ for $(y,x,z) \in \mathcal{Y}_0 \times \mathcal{X}_0 \times \mathcal{Z}_0$, by Lemma 24(a) $\hat{G}_b(y|x,z) \in \mathcal{T}_0^\epsilon$ with probability approaching 1 (wpa.1) as $n \to \infty$, where $\mathcal{T}_0^\epsilon \equiv [\underline{\tau} - \epsilon, \bar{\tau} + \epsilon] \subset (0,1)$ for some $\epsilon > 0$. Note that the result in Lemma 24(c) also holds uniformly in $(\tau, u) \in \mathcal{T}_0^\epsilon \times \mathcal{U}_0$ wpa.1.

**Lemma 25** *Suppose that Assumptions A.1-A.4, A.6 and A.7 hold. Then*
*(a)* $\sup_{\tilde{y},y \in \mathcal{Y}_0, |\tilde{y}-y| \leq M(v_{2b}+b^{r_1})} \sup_{u \in \mathcal{U}_0} \sqrt{nc^{d_X+1/2}} \|V_{n,b}(\tilde{y};u) - V_{n,b}(y;u)\| = o_P(1)$*;*
*(b)* $\sup_{\tilde{y},y \in \mathcal{Y}_0, |\tilde{y}-y| \leq M(v_{2b}+b^{r_1})} \sup_{u \in \mathcal{U}_0} \sqrt{nc^{d_X+1/2}} \|\mathbf{V}_{n,c}(\tilde{y};u) - \mathbf{V}_{n,c}(y;u)\| = o_P(1)$.

**Proof.** The proof is analogous to that of Lemma A.3 in **HSW** and thus omitted. ■

**Lemma 26** *Suppose that Assumptions A.1-A.4, A.6 and A.7 hold. Then for any* $\delta_n = O(\nu_{2b} + b^{r_1})$*, we have*
*(a)* $\hat{G}_b(a+\delta_n|u) - \hat{G}_b(a|u) = g(a|u)\delta_n + o_P\left(n^{-1/2}c^{-d_X/2-1/4}\right)$ *uniformly in* $u \in \mathcal{U}_0$,
*(b)* $\hat{G}_b^{-1}(a+\delta_n|u) - \hat{G}_b^{-1}(a|u) = g(G^{-1}(a|u)|u)^{-1}\delta_n + o_P\left(n^{-1/2}c^{-d_X/2-1/4}\right)$ *uniformly in* $u \in \mathcal{U}_0$,

**Proof.** By Lemma 24, $\hat{G}_b(a+\delta_n|u) - \hat{G}_b(a|u) = [G(a+\delta_n|u) - G(a|u)] + g(u)^{-1}[V_{n,b}(a+\delta_n;u) - V_{n,b}(a;u)] + O_P(\nu_b^2 + b^{r_1})$. By Assumption A.4 and Taylor expansions, the first term on the right hand side of the last expression is $g(a|u)\delta_n + O(\delta_n^2)$. By Lemma 25, $V_{n,b}(a+\delta_n;u) - V_{n,b}(a;u) = o_P(n^{-1/2}c^{-d_X/4-1/4})$ uniformly in $u \in \mathcal{U}_0$. Thus (a) follows by Assumption A.7. The proof of (b) is analogous and thus omitted.

■

**Lemma 27** *Suppose Assumptions A.1-A.4, A.6 and A.7 hold. Then uniformly in* $(y,u) \in \mathcal{Y}_0 \times \mathcal{U}_0$,
*(a)* $\hat{g}_c(y|u) - g(y|u) = g(u)^{-1}\mathbf{V}_{n,c}(y;u) + O_P(c^{r_2} + \nu_{3c}^2)$,
*(b)* $\mathbf{V}_{n,c}(y;u) = O_P(\nu_{3c})$,
*(c)* $\hat{g}_c(y+\delta_n|u) - \hat{g}_c(y|u) = D_y g(y|u)\delta_n + o_P\left(n^{-1/2}c^{-d_X/2-1/4}\right)$ *for any* $\delta_n = O(\nu_{2b} + b^{r_1})$,
*where* $D_y g(y|u) \equiv \partial(g(y|u))/\partial y$.

**Proof.** Recall $W_i \equiv (Y_i, U_i')'$ and $w \equiv (y, u')'$. We make the following bias-variance decomposition:

$$\hat{g}_c(y|u) - g(y|u) = \hat{g}_c(u)^{-1} \frac{1}{n} \sum_{i=1}^{n} \{g(y,u) - E[L_c(W_i - w)]\} + \hat{g}_c(u)^{-1} \mathbf{V}_{n,c}(y;u)$$

By Assumptions A.1, A.4 and A.6(ii) and the standard arguments in kernel estimation, $\sup_{u \in \mathcal{U}_0} |\hat{g}_c(u) - g(u)| = O_P(\nu_{2c} + c^{r_2})$, $\sup_{w \in \mathcal{W}_0} |\frac{1}{n}\sum_{i=1}^{n} E[L_c(W_i - w)] - g(y,u)| = O_P(c^{r_2})$, and $\sup_{w \in \mathcal{W}_0} |\mathbf{V}_{n,c}(y;u)| = O_P(\nu_{3c})$. Thus (a) and (b) follow. Furthermore, $\hat{g}_c(y + \delta_n|u) - \hat{g}_c(y|u) = [g(y + \delta_n|u) - g(y|u)] + g(u)^{-1}[\mathbf{V}_{n,c}(y + \delta_n; u) - \mathbf{V}_{n,c}(y;u)] + O_P(c^{r_2} + \nu_{3c}^2)$. Then (c) follows from Taylor expansions and Lemma 25. ■

**Lemma 28** *Suppose that Assumptions A.1-A.4, A.6 and A.7 hold. Then uniformly in* $i$,

*(a)* $(\hat{\varepsilon}_i - \varepsilon_i) 1_i = s_{\varepsilon n,i} 1_i \{1 + o(1)\} + o_P(n^{-1/2} c^{-d_X/2 - 1/4})$,

*(b)* $(\hat{\varepsilon}_i - \varepsilon_i) 1_i = O_P(v_{1b} + b^{r_1})$,

*where* $s_{\varepsilon n,i} = \int \frac{1}{g(G^{-1}(\tau_{iz}|\bar{x},z)|\bar{x},z)} [-\frac{V_{n,b}\left(G^{-1}(\tau_{iz}|\bar{x},z);\bar{x},z\right)}{g(\bar{x},z)} + \frac{V_{n,b}(Y_i;X_i,z)}{g(X_i,z)}] dH(z)$ *and* $\tau_{iz} \equiv G(Y_i|X_i,z)$.

**Proof.** Let $\hat{\tau}_{iz} \equiv \hat{G}_b(Y_i|X_i,z)$. Then $(\hat{\varepsilon}_i - \varepsilon_i) 1_i = \varepsilon_{1i} + \varepsilon_{2i}$, where

$$\begin{aligned}
\varepsilon_{1i} &\equiv \left[\int \hat{G}_b^{-1}(\hat{\tau}_{iz}|\bar{x},z) dH(z) - \int G^{-1}(\hat{\tau}_{iz}|\bar{x},z) dH(z)\right] 1_i, \text{ and} \\
\varepsilon_{2i} &\equiv \left[\int G^{-1}(\hat{\tau}_{iz}|\bar{x},z) dH(z) - \int G^{-1}(\tau_{iz}|\bar{x},z) dH(z)\right] 1_i.
\end{aligned}$$

By Lemmas 24 and 25,

$$\begin{aligned}
\varepsilon_{1i} &= -\int \frac{V_{n,b}\left(G^{-1}(\hat{\tau}_{iz}|\bar{x},z);\bar{x},z\right)\{1 + o(1)\}}{g\left(G^{-1}(\hat{\tau}_{iz}|\bar{x},z)|\bar{x},z\right) g(\bar{x},z)} dH(z) 1_i + O_P(\nu_{2b}^2 + b^{r_1}) \\
&= -\int \frac{V_{n,b}\left(G^{-1}(\tau_{iz}|\bar{x},z);\bar{x},z\right)\{1 + o(1)\}}{g\left(G^{-1}(\tau_{iz}|\bar{x},z)|\bar{x},z\right) g(\bar{x},z)} dH(z) 1_i + o_P(n^{-1/2} c^{-d_X/2 - 1/4}),
\end{aligned}$$

and

$$\begin{aligned}
\varepsilon_{2i} &= \int g(G^{-1}(\tau_{iz}|\bar{x},z)|\bar{x},z)^{-1} (\hat{\tau}_{iz} - \tau_{iz}) \, dH(z) 1_i + O_p\left(b^{2r_1} + v_{2b}^4\right) \\
&= \int g(G^{-1}(\tau_{iz}|\bar{x},z)|\bar{x},z)^{-1} g(X_i,z)^{-1} V_{n,b}(Y_i;X_i,z) \, dH(z) 1_i + o_P(n^{-1/2} c^{-d_X/2 - 1/4}).
\end{aligned}$$

Combining these results yields (a). (b) follows from (a) and the standard arguments as used in showing $\sup_{u\in\mathcal{U}_0}|V_{n,b}(y;u)| = O_P(\nu_{2b})$. ■

**Lemma 29** *Suppose that Assumptions A.1-A.4, A.6 and A.7 hold. Then*

*(a)* $\alpha_1(x,e) \equiv \int\left[\frac{\hat{g}_c(e|\bar{x},z)}{\hat{g}_c(y|x,z)} - \frac{g(e|\bar{x},z)}{g(y|x,z)}\right]dH(z) = s_{1n}(x,e) + O_P(c^{r_2} + n^{-1}c^{-(d_X+1)}\log n)$ *uniformly in* $(e,x)\in\mathcal{E}_0\times\mathcal{X}_0$,

*(b)* $\alpha_{2i} \equiv \int\frac{\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z)-\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}dH(z)1_i = s_{2i}1_i + o_P(n^{-1/2}c^{-d_X/2-1/4})$ *uniformly in* $i$,

*where* $y = m(x,e)$, $s_{1n}(x,e) = \int\frac{1}{g(y|x,z)}\left[\frac{\mathbf{V}_{n,c}(e;\bar{x},z)}{g(\bar{x},z)} - D_e m(x,e)\frac{\mathbf{V}_{n,c}(y;x,z)}{g(x,z)}\right]dH(z)$, *and* $s_{2n,i} = \int\frac{D_y g(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)}\,dH(z)s_{\varepsilon n,i}$.

**Proof.** First, observe that $\alpha_1(x,e) = \alpha_{11}(x,e) + \alpha_{12}(x,e)$, where $\alpha_{11}(x,e) = \int\hat{g}_c(y|x,z)^{-1}[\hat{g}_c(e|\bar{x},z) - g(e|\bar{x},z)]dH(z)$, and $\alpha_{12}(x,e) = \int g(e|\bar{x},z)[\hat{g}_c(y|x,z)^{-1} - g(y|x,z)^{-1}]dH(z)$. By Lemma 27(i), we can show that

$$\begin{aligned}\int[\hat{g}_c(y|x,z) - g(y|x,z)]^2\,dH(z) &= \int g(x,z)^{-2}\mathbf{V}_{n,c}(y;x,z)^2\,dH(z) + O_P(c^{2r_2} + \nu_{3c}^4)\\ &= O_P(v_{1c}^2c^{-1} + c^{2r_2}) \text{ uniformly in } (y,x)\in\mathcal{Y}_0\times\mathcal{X}_0 \quad (7.18)\end{aligned}$$

By Lemma 27, (7.18) and the Cauchy-Schwarz inequality, we have that uniformly in $(e,x)\in\mathcal{E}_0\times\mathcal{X}_0$

$$\begin{aligned}\alpha_{11}(x,e) &= \int g(y|x,z)^{-1}[\hat{g}_c(e|\bar{x},z) - g(e|\bar{x},z)]\,dH(z)\\ &\quad - \int\hat{g}_c(y|x,z)^{-1}g(y|x,z)[\hat{g}_c(y|x,z) - g(y|x,z)][\hat{g}_c(e|\bar{x},z) - g(e|\bar{x},z)]\,dH(z)\\ &= \int g(y|x,z)^{-1}[\hat{g}_c(e|\bar{x},z) - g(e|\bar{x},z)]\,dH(z) + O_P(n^{-1}c^{-(d_X+1)}\log n + c^{2r_2})\\ &= \int g(y|x,z)^{-1}g(\bar{x},z)^{-1}\mathbf{V}_{n,c}(e;\bar{x},z)\,dH(z) + O_P(c^{r_2} + \nu_{3c}^2),\end{aligned}$$

and

$$\begin{aligned}\alpha_{12}(x,e) &= -\int g(e|\bar{x},z)\hat{g}_c(y|x,z)^{-1}g(y|x,z)^{-1}[\hat{g}_c(y|x,z) - g(y|x,z)]dH(z)\\ &= -\int g(e|\bar{x},z)g(y|x,z)^{-2}[\hat{g}_c(y|x,z) - g(y|x,z)]dH(z) + O_P(n^{-1}c^{-(d_X+1)}\log n + c^{2r_2})\\ &= -\int g(e|\bar{x},z)g(y|x,z)^{-2}g(x,z)^{-1}\mathbf{V}_{n,c}(y;x,z)\,dH(z) + O_P(n^{-1}c^{-(d_X+1)}\log n + c^{r_2}).\end{aligned}$$

Then by (7.2) we have that uniformly in $(e,x) \in \mathcal{E}_0 \times \mathcal{X}_0$

$$\begin{aligned}\alpha_1(x,e) &= \int \frac{1}{g(y|x,z)}\left[\frac{\mathbf{V}_{n,c}(e;\bar{x},z)}{g(\bar{x},z)} - D_e m(x,e)\frac{\mathbf{V}_{n,c}(y;x,z)}{g(x,z)}\right] dH(z) \\ &\quad + O_P(c^{r_2} + n^{-1}c^{-(d_X+1)}\log n) \\ &= s_{1n}(x,e) + O_P(c^{r_2} + n^{-1}c^{-(d_X+1)}\log n).\end{aligned}$$

For (b), note that $\alpha_{2i} = \alpha_{21i} - \alpha_{22i}$, where $\alpha_{21i} = \int \frac{\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z) - \hat{g}_c(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)}\, dH(z)\, 1_i$, and $\alpha_{22i} =$

$\int \frac{[\hat{g}_c(Y_i|X_i,z) - g(Y_i|X_i,z)][\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z) - \hat{g}_c(\varepsilon_i|\bar{x},z)]}{\hat{g}_c(Y_i|X_i,z)g(Y_i|X_i,z)} dH(z)\, 1_i$. By Lemmas 27 and 28(i),

$$\alpha_{21i} = \int \frac{D_y g(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)}\, dH(z)\,(\hat{\varepsilon}_i - \varepsilon_i)\, 1_i + o_P(n^{-1/2}c^{-d_X/2-1/4}) = s_{2n,i}1_i + o_P(n^{-1/2}c^{-d_X/2-1/4}),$$

where $o_P(n^{-1/2}c^{-d_X/2-1/4})$ holds uniformly in $i$. By Assumption A.2, Lemmas 27 and 28(ii), and (7.18), we have that uniformly in $i$

$$\begin{aligned}\alpha_{22i} &= \int \frac{[\hat{g}_c(Y_i|X_i,z) - g(Y_i|X_i,z)]\,[\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z) - \hat{g}_c(\varepsilon_i|\bar{x},z)]}{g(Y_i|X_i,z)g(Y_i|X_i,z)} dH(z)\, 1_i \{1 + o(1)\} \\ &= \int |\hat{g}_c(Y_i|X_i,z) - g(Y_i|X_i,z)|\, dH(z) O_P(v_{1b} + b^{r_1}) \\ &= \left[\int \{[\hat{g}_c(Y_i|X_i,z) - g(Y_i|X_i,z)]\}^2\, dH(z)\right]^{1/2} 1_i O_P(v_{1b} + b^{r_1}) \\ &= O_P(n^{-1/2}c^{-(d_X+1)/2}\sqrt{\log n} + c^{r_2}) O_P(v_{1b} + b^{r_1}) = o_P(n^{-1/2}c^{-d_X/2-1/4}).\end{aligned}$$

It follows that $\alpha_{2i} = s_{2n,i}1_i + o_P(n^{-1/2}c^{-d_X/2-1/4})$ uniformly in $i$. ■

## Proof of the Main Results

### Proofs of Theorem 21 and 22

We only prove Theorem 22 as the proof of Theorem 21 is a special case. To conserve space, let $a_i \equiv a(X_i, Y_i)$. We first make the following decomposition:

$$\begin{aligned} nc^{d_X+\frac{1}{2}}\hat{J}_n &= c^{d_X+\frac{1}{2}}\sum_{i=1}^{n}\left\{\int \frac{\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z)-\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}\,dH(z)+\int\left[\frac{\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}-1\right]dH(z)\right\}^2 a_i \\ &= c^{d_X+\frac{1}{2}}\sum_{i=1}^{n}\left\{\int\left[\frac{\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}-1\right]dH(z)\right\}^2 a_i \\ &\quad +c^{d_X+\frac{1}{2}}\sum_{i=1}^{n}\left[\int\frac{\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z)-\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}\,dH(z)\right]^2 a_i \\ &\quad +2c^{d_X+\frac{1}{2}}\sum_{i=1}^{n}\int\frac{\hat{g}_c(\hat{\varepsilon}_i|\bar{x},z)-\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}\,dH(z)\int\left[\frac{\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}-1\right]dH(z)a_i \\ &\equiv \hat{J}_{n1}+\hat{J}_{n2}+2\hat{J}_{n3},\ \text{say.} \end{aligned}$$

Propositions 30, 31 and 32 study $\hat{J}_{n1}, \hat{J}_{n2}$ and $\hat{J}_{n3}$, respectively. Combining the results in these propositions yields $nc^{d_X+\frac{1}{2}}\hat{J}_n = J_n + \mu_0 + o_P(1)$, where $J_n = c^{d_X+\frac{1}{2}}\sum_{i=1}^{n} s_{n,i}^2 a_i$,

$$s_{n,i} = \int g_{1iz}^{-1}\left[\frac{\mathbf{V}_{n,c}(\varepsilon_i;\bar{x},z)}{g(\bar{x},z)}-\frac{\mathbf{V}_{n,c}(Y_i;X_i,z)}{g(X_i,z)}\right]dH(z) = n^{-1}\sum_{j=1}^{n}\zeta_0(W_i,W_j), \quad (7.19)$$

$g_{1iz} = g(Y_i|X_i,z)$, and $\zeta_0(W_i,W_j) = \int g_{1iz}^{-1}\left[g(\bar{x},z)^{-1}\bar{L}_{cj,(\varepsilon_i,\bar{x},z)} - g(X_i,z)^{-1}\bar{L}_{cj,(Y_i,X_i,z)}\right]dH(z)$ is as defined in (7.14). The rest of the proof follows that of **HSW** closely.

First, using $\zeta_0$, we can write $J_n$ as a third order $V$-statistic:

$$J_n = c^{d_X+\frac{1}{2}}\sum_{i=1}^{n}\left[n^{-1}\sum_{j=1}^{n}\zeta_0(W_i,W_j)\right]^2 a_i = n^{-2}c^{d_X+\frac{1}{2}}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{i_3=1}^{n}\zeta(W_{i_1},W_{i_2},W_{i_3}),$$

where $\zeta(W_{i_1},W_{i_2},W_{i_3}) \equiv \zeta_0(W_{i_1},W_{i_2})\zeta_0(W_{i_1},W_{i_3})a_{i_1}$. To study the asymptotic distribution of $J_n$, we need to use the $U$-statistic theory (e.g., Lee (1990)). Let $\varphi(w_{i_1},w_{i_2}) \equiv E[\zeta(W_1,w_{i_1},w_{i_2})]$, and $\bar{\zeta}(w_{i_1},w_{i_2},w_{i_3}) \equiv \zeta(w_{i_1},w_{i_2},w_{i_3}) - \varphi(w_{i_2},w_{i_3})$. Then we can decompose $J_n$ as follows

$$\begin{aligned} J_n &= n^{-1}c^{d_X+\frac{1}{2}}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\varphi(W_{i_1},W_{i_2}) + n^{-2}c^{d_X+\frac{1}{2}}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{i_3=1}^{n}\bar{\zeta}(W_{i_1},W_{i_2},W_{i_3}) \\ &\equiv J_{1n}+J_{2n},\ \text{say.} \end{aligned}$$

Consider $J_{2n}$ first. Write $E(J_{n2}^2) = n^{-4}c^{2d_X+1}\sum_{i_1,\ldots,i_6=1}^n E\left[\bar{\zeta}\left(W_{i_1},W_{i_2},W_{i_3}\right)\bar{\zeta}\left(W_{i_4},W_{i_5},W_{i_6}\right)\right.$ Observing that $E\left[\bar{\zeta}\left(W_{i_1},w_{i_2},w_{i_3}\right)\right] = E\left[\bar{\zeta}\left(w_{i_1},W_{i_2},w_{i_3}\right)\right] = E[\bar{\zeta}(w_{i_1},w_{i_2},W_{i_3})] = 0$, $E[\bar{\zeta}(W_{i_1},W_{i_2},W_{i_3})\bar{\zeta}(W_{i_4},W_{i_5},W_{i_6})] = 0$ if there are more than three distinct elements in $\{i_1,\ldots,i_6\}$. In view of this, we can show that

$$E(J_{2n}^2) = O(n^{-1}c^{-d_X-1} + n^{-2}c^{-2d_X-1} + n^{-3}c^{-2d_X-2}) = o(1).$$

Then $J_{2n} = o_P(1)$ by the Chebyshev inequality.

For $J_{1n}$, let $\varphi(W_i,W_j) = \int \zeta_0(\tilde{w},W_i)\,\zeta_0(\tilde{w},W_j)\,a\left(\tilde{x},m^{-1}(\tilde{x},\tilde{y})\right)dG(\tilde{w})$, where $G(\cdot)$ is the CDF of $W_i$.Then $J_{1n} = \mathbb{B}_n + \mathbb{V}_n$, where $\mathbb{B}_n = n^{-1}c^{d_X}\sum_{i=1}^n \varphi(W_i,W_i)$ and $\mathbb{V}_n = 2n^{-1}c^{d_X+\frac{1}{2}}\sum_{1\le i<j\le n}\varphi(W_i,W_j)$ contribute to the asymptotic bias and variance of our test statistic, respectively. Observing that $\mathbb{V}_n$ is a second-order degenerate $U$-statistic, we can easily verify that all the conditions of Theorem 1 of Hall (1984) are satisfied and a central limit theorem applies to it: $\mathbb{V}_n \xrightarrow{d} N\left(0,\sigma^2\right)$, where $\sigma^2 = \lim_{n\to\infty}\sigma_n^2$ and $\sigma_n^2 = 2c^{2d_X+1}E\left[\varphi(W_1,W_2)\right]^2$.[8] ■

**Proposition 30** *$\hat{J}_{n1} = c^{d_X+\frac{1}{2}}\sum_{i=1}^n s_{n,i}^2 a_i + \mu_0 + o_P(1)$ under $\mathbb{H}_1(\gamma_n)$.*

**Proof.** To begin with, we decompose $\hat{J}_{n1}$ as follows:

$$\begin{aligned}
\hat{J}_{n1} &= c^{d_X+\frac{1}{2}}\sum_{i=1}^n\left\{\int\left[\frac{\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)} - \frac{g(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)}\right]dH(z)\right\}^2 a_i \\
&\quad + c^{d_X+\frac{1}{2}}\sum_{i=1}^n\left\{\int\left[\frac{g(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)} - 1\right]dH(z)\right\}^2 a_i \\
&\quad + 2c^{d_X+\frac{1}{2}}\sum_{i=1}^n\int\left[\frac{\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)} - \frac{g(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)}\right]dH(z)\int\left[\frac{g(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)} - 1\right]dH(z)a_i \\
&= J_{n11} + J_{n12} + 2J_{n13}.
\end{aligned}$$

Using Lemma 29 and the fact that $D_e m(x,e) = 1 + \gamma_n\delta_n(x,e)$ under $\mathbb{H}_1(\gamma_n)$, we can

[8] Write $\zeta_0(W_i,W_j) = \int g_{1iz}^{-1}g(\bar{x},z)^{-1}\bar{L}_{cj,(\varepsilon_i,\bar{x},z)}dH(z) - \int g_{1iz}^{-1}g(X_i,z)^{-1}\bar{L}_{cj,(Y_i,X_i,z)}dH(z) \equiv \zeta_{1ij} - \zeta_{2ij}$, say. A careful calculation suggests that both $\zeta_{1ij}$ and $\zeta_{2ij}$ contributes to the asymptotic bias of $J_{1na}$ but only $\zeta_{1ij}$ contributes to the asymptotic variance of $J_{1na}$.

show that

$$\begin{aligned} J_{n11} &= c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} s_{1n,i}^2 a_i + n c^{d_X+\frac{1}{2}} O_P((c^{r_2} + n^{-1} c^{-(d_X+1)} \log n)^2) \\ &= c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} s_{n,i}^2 a_i + o_P(1) \end{aligned}$$

where $s_{1n,i} = s_{1n}(X_i, \varepsilon_i)$ and $s_{n,i}$ is defined in (7.19). By (7.2), (7.17), and the weak law of large numbers (WLLN), we have

$$\begin{aligned} J_{n12} &= c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} [D_e m(X_i, \varepsilon_i) - 1]^2 a_i = c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \gamma_n^2 \delta_n(X_i, \varepsilon_i)^2 a_i + o_P(1) \\ &= n^{-1} \sum_{i=1}^{n} \delta_n(X_i, \varepsilon_i)^2 a_i + o_P(1) \xrightarrow{P} \lim_{n\to\infty} E\left[\delta_n(X_i, \varepsilon_i)^2 a(X_i, Y_i)\right] \equiv \mu_0. \end{aligned}$$

For $J_{n13}$, by Lemma 29 and (7.17), we have

$$\begin{aligned} J_{n13} &= c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \int \left[\frac{\hat{g}_c(\varepsilon_i|\bar{x}, z)}{\hat{g}_c(Y_i|X_i, z)} - \frac{g(\varepsilon_i|\bar{x}, z)}{g(Y_i|X_i, z)}\right] dH(z) \gamma_n \delta_n(X_i, \varepsilon_i) a_i \\ &= \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i s_{1n,i} \delta_n(X_i, \varepsilon_i) + n \gamma_n c^{d_X+\frac{1}{2}} O_P(c^{r_2} + n^{-1} c^{-(d_X+1)} \log n) \\ &= \bar{J}_{n13} + o_P(1), \end{aligned}$$

where $\bar{J}_{n13} \equiv \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i s_{n,i} \delta_n(X_i, \varepsilon_i)$. Note that $\bar{J}_{n13} = \bar{J}_{n131} + \bar{J}_{n132}$, where $\bar{J}_{n13s} = n^{-1} \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \zeta_{sij} \delta_n(X_i, \varepsilon_i)$ for $s = 1, 2$, where $\zeta_{1ij}$ and $\zeta_{2ij}$ are defined in footnote 7.5. We further write $\bar{J}_{n131} = n^{-1} \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i \zeta_{1ii} \delta(X_i, \varepsilon_i)$ $+ n^{-1} \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j\neq i}^{n} a_i \zeta_{1ij} \delta_n(X_i, \varepsilon_i)$. It is easy to show that the first term is $O_P(\gamma_n c^{d_X+\frac{1}{2}})$ and the second term is $O_P(c^{1/4})$ by moment calculations. It follows that $\bar{J}_{n131} = o_P(1)$. Similarly, $\bar{J}_{n132} = n^{-1} \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i \zeta_{2ii} \delta_n(X_i, \varepsilon_i) + n^{-1} \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j\neq i}^{n} a_i$ $\zeta_{2ij} \delta_n(X_i, \varepsilon_i) = O_P(\gamma_n c^{-\frac{1}{2}}) + O_P(n^{-\frac{1}{2}} c^{-\frac{1}{4}} + c^{\frac{1}{2} d_X + \frac{1}{4}}) = o_P(1)$. It follows that $J_{n13} = o_P(1)$.

Combining the above results yields the desired result: $\hat{J}_{n1} = c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} s_{n,i}^2 a_i +$ $\mu_0 + o_P(1)$. ■

**Proposition 31** $\hat{J}_{n2} = o_P(1)$ *under* $\mathbb{H}_1(\gamma_n)$.

**Proof.** By Lemma 29(ii) and the Cauchy-Schwarz inequality, we have

$$\hat{J}_{n2} \leq 2c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i s_{2n,i}^2 + 2nc^{d_X+\frac{1}{2}} o_P((n^{-1/2}c^{-d_X/2-1/4})^2) = 2J_{n2} + o_P(1),$$

where $J_{n2} \equiv c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i \beta_i^2 s_{\varepsilon n,i}^2$ and $\beta_i \equiv \int \frac{D_y g(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)} dH(z)$. Let $g_{2iz} \equiv g(G^{-1}(\tau_{iz}|\bar{x},z)|\bar{x},z)$ where recall $\tau_{iz} = G(Y_i|X_i,z)$. Then

$$\begin{aligned} J_{n2} &= c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i \beta_i^2 \left[ \frac{1}{n} \sum_{j=1}^{n} (-\eta_{1ij} + \eta_{2ij}) \right]^2 \\ &= n^{-2} c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_i \beta_i^2 (\eta_{1ij}\eta_{1ik} + \eta_{2ij}\eta_{2ik} - 2\eta_{1ij}\eta_{2ik}) \\ &= J_{n21} + J_{n22} + J_{n23}, \text{ say,} \end{aligned}$$

where $\eta_{1ij} \equiv \int g_{2iz}^{-1} g(\bar{x},z)^{-1} K_{bj,(\bar{x},z)} \bar{\mathbf{1}}_j \left(G^{-1}(\tau_{iz}|\bar{x},z)\right) dH(z)$, $\eta_{2ij} \equiv \int g_{2iz}^{-1} g(X_i,z)^{-1} K_{bj,(X_i,z)} \bar{\mathbf{1}}_j(Y_i)$ $dH(z)$, and e.g., $J_{n211} \equiv n^{-2} c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_i \beta_i^2 \eta_{1ij}\eta_{1ik}$. For $J_{n21}$, we decompose it as follows:

$$\begin{aligned} J_{n21} &= n^{-2} c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} \sum_{k=1,k\neq i,j}^{n} a_i \beta_i^2 \eta_{1ij}\eta_{1ik} + n^{-2} c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} a_i \beta_i^2 \eta_{1ij}^2 \\ &\quad + 2n^{-2} c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} a_i \beta_i^2 \eta_{1ii}\eta_{1ij} + n^{-2} c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i \beta_i^2 \eta_{1ii}^2 \\ &\equiv J_{n211} + J_{n212} + J_{n213} + J_{n214}, \text{ say.} \end{aligned}$$

In view of $E(J_{n211a}) = 0$, $E\left(J_{n211}^2\right) = O\left(c^{2d_X+1} b^{-2d_X}\right)$, $E|J_{n212}| = E(J_{n212}) = O(c^{d_X+\frac{1}{2}} b^{-d_X})$, $E|J_{n213}| = O(c^{d_X+\frac{1}{2}} b^{-d_X})$, and $E|J_{n214}| = O(n^{-1} c^{d_X+\frac{1}{2}} b^{-2d_X})$, we have $J_{n21} = O_P(c^{d_X+\frac{1}{2}} b^{-d_X} + n^{-1} c^{d_X+\frac{1}{2}} b^{-2d_X}) = o_P(1)$ by the Chebyshev and Markov inequalities. By the same token, we can show that $J_{n22} = o_P(1)$. Then $J_{n23} = o_P(1)$ by the Cauchy-Schwarz inequality. Consequently, we have shown that $J_{n2} = o_P(1)$. ■

**Proposition 32** *$\hat{J}_{n3} = o_P(1)$ under $\mathbb{H}_1(\gamma_n)$.*

**Proof.** Following the proof of Propositions 30 and 31, we can show that

$$\begin{aligned}
\hat{J}_{n3} &= c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} s_{2n,i} \left[ s_{1n,i} + \gamma_n \delta(X_i, \varepsilon_i) \right] a_i + o_P(1) \\
&= c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} s_{2n,i} s_{1n,i} a_i + \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} s_{2n,i} \delta(X_i, \varepsilon_i) a_i + o_P(1) \\
&\equiv J_{n31} + J_{n32} + o_P(1), \text{ say.}
\end{aligned}$$

We prove the lemma by demonstrating that $J_{n31} = o_P(1)$ and $J_{n32} = o_P(1)$. Recall $\zeta_{1ij} \equiv \int g_{1iz}^{-1} g(\bar{x}, z)^{-1} \bar{L}_{cj,(\varepsilon_i,\bar{x},z)} dH(z)$ and $\zeta_{2ij} \equiv \int g_{1iz}^{-1} g(X_i, z)^{-1} \bar{L}_{cj,(Y_i,X_i,z)} dH(z)$. Let $\bar{\zeta}_{2ij} = D_e m(X_i, \varepsilon_i) \zeta_{2ij}$. Then

$$\begin{aligned}
J_{n31} &= c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i \frac{1}{n} \sum_{j=1}^{n} \beta_i \left( -\eta_{1ij} + \eta_{2ij} \right) \frac{1}{n} \sum_{k=1}^{n} \left( \zeta_{1ik} - \bar{\zeta}_{2ik} \right) \\
&= n^{-2} n^{-2} c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_i \beta_i \left( -\eta_{1ij} \zeta_{1ik} + \eta_{1ij} \bar{\zeta}_{2ik} + \eta_{2ij} \zeta_{1ik} - \eta_{2ij} \bar{\zeta}_{2ik} \right) \\
&\equiv -J_{n311} + J_{n312} + J_{n313} - J_{n314}, \text{ say.}
\end{aligned}$$

As in the analysis of $J_{n211}$, we can readily demonstrate that $J_{n31s} = o_P(1)$ by straightforward moment calculations and the Chebyshev/Markov inequalities. Thus $J_{n31} = o_P(1)$. Note that $J_{n32} = J_{n321} + J_{n322}$, where $J_{n32s} = n^{-1} \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \beta_i \eta_{sij} \delta_n(X_i, \varepsilon_i)$ for $s = 1, 2$. We further write $J_{n321} = n^{-1} \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} a_i \beta_i \eta_{1ii} \delta_n(X_i, \varepsilon_i) +$ $n^{-1} \gamma_n c^{d_X+\frac{1}{2}} \sum_{i=1}^{n} \sum_{j \neq i}^{n} a_i \beta_i \eta_{1ij} \delta_n(X_i, \varepsilon_i)$. It is easy to show that the first term is $O_P(\gamma_n c^{d_X+\frac{1}{2}})$ and the second term is $O_P((c/b)^{d_X/2} c^{1/4})$ by moment calculations. It follows that $J_{n321} = o_P(1)$. Similarly, $J_{n321} = o_P(1)$. Thus we have shown that $J_{n32} = o_P(1)$.

■

**Proof of Theorem 23**

The proof is simpler than that of Theorem 22. Under $\mathbb{H}_1$, we can readily apply Lemmas 29, 28, and the WLLN to obtain

$$\begin{aligned}
\hat{J}_n &= n^{-1}\sum_{i=1}^{n}\left\{\int\left[\frac{\hat{g}_c(\varepsilon_i|\bar{x},z)}{\hat{g}_c(Y_i|X_i,z)}-1\right]dH(z)\right\}^2 a_i + o_P(1) \\
&= n^{-1}\sum_{i=1}^{n}\left\{\int\left[\frac{g(\varepsilon_i|\bar{x},z)}{g(Y_i|X_i,z)}-1\right]dH(z)\right\}^2 a_i + o_P(1) \\
&= n^{-1}\sum_{i=1}^{n}\left[D_e m(X_i,\varepsilon_i)-1\right]^2 a_i + o_P(1) \xrightarrow{P} E\left\{\left[D_e m(X_i,\varepsilon_i)-1\right]^2 a_i\right\}.
\end{aligned}$$

The result follows by noticing that $\hat{\sigma}_n^2 = O_P(1)$ and $\hat{B}_n = o_P\left(nc^{d_X+1/2}\right)$ under $\mathbb{H}_1$. ■

**References**


Altonji, J. G., H. Ichimura and T. Otsu (2011). Estimating Derivatives in Nonseparable Models with Limited Dependent Variables. *Econometrica*, forthcoming.

Altonji, J. G., and R. L. Matzkin (2005). Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors. *Econometrica* 73:1053-1102.

Andrews, D. W. K. (1997). A Conditional Kolmogorov Test. *Econometrica* 65: 1097-1128.

Blundell, R., and J. L. Powell (2003). Endogeneity in Nonparametric and Semiparametric Regression Models. In: M. Dewatripont, L. P. Hansen, and S. J. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress*, Vol. II, pp.312-357. Cambridge University Press, Cambridge.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics* 20:105-134.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica* 58: 1443-1458.

Bierens, H. J. and W. Ploberger (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica* 65: 1129-1151.

Boente, G. and R. Fraiman (1991). Strong Uniform Convergence Rates for Some Robust Equivariant Nonparametric Regression Estimates for Mixing Processes. *International Statistical Review* 59: 355-372.

Briesch, R., P. Chintagunta, and R. L. Matzkin (2007). Nonparametric Discrete Choice Models with Unobserved Heterogeneity. Mimeo, Northwestern University.

Brown, D. J., and R. L. Matzkin (1998). Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand. Mimeo, Northwestern University.

Doss, H. and R. D. Gill (1992). An Elementary Approach to Weak Convergence for Quantile Processes, with Applications to Censored Survival Data. *Journal of the American Statistical Association* 87: 869-877.

Ekeland, I., J. J. Heckman and L. Nesheim (2004). Identification and Estimation of Hedonic Models. *Journal of Political Economy* 112(S1), S60-109. Paper in Honor of Sherwin Rosen: A Supplement to Volume 112.

Giacomini, R., D. N. Politis and H. White (2007). A Warp-Speed Method for Conducting Monte Carlo Experiments Involving Bootstrap Estimators. *Discussion Paper*, Dept. of Economics, UCSD.

Hall, P. (1984). Central Limit Theorem for Integrated Square Error Properties of

Multivariate Nonparametric Density Estimators. *Journal of Multivariate Analysis* 14: 1-16.

Hansen, B. E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory* 24: 726-748.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21: 1926-1947.

Hausman, J. A. (1978). Specification Testing in Econometrics. *Econometrica* 46: 1251-1271.

Heckman, J. J. (1974). Effects of Child-Care Programs on Women's Work Effort. *Journal of Political Economy* 82: S136-S163.

Heckman, J., R. Matzkin and L. Nesheim (2010). Nonparametric Estimation of Nonadditive Hedonic Models. *Econometrica* 78: 1569-1591.

Heckman, J. J. and E. J. Vytlacil (1999). Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Science* 96: 4730-4734.

Heckman, J. J. and E. J. Vytlacil (2001). Local Instrumental Variables. In: C. Hsiao, K. Morimune, and J. Powell (eds.) *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, pp. 1-146. Cambridge University Press, Cambridge.

Heckman, J. J. and E. Vytlacil (2005). Structural Equations, Treatment Effects, and Economic Policy evaluation. *Econometrica* 73, 669-738.

Heckman, J. J. and R. Willis (1977). A Beta-Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women. *Journal of Political Economy*

85: 27-58.

Hoderlein, S., L. Su and H. White (2011). Specification Testing for Nonparametric Structural Models with Monotonicity in Unobservables. *Discussion Paper*, Dept. of Economics, UCSD.

Hong, Y. and H. White (1995). Consistent Specification Testing via Nonparametric Series Regression,*Econometrica*, 63, 1133-1159.

Hsiao, C., Q. Li and J. Racine (2007). A Consistent Model Specification Test with Mixed Discrete and Continuous data. *Journal of Econometrics* 140: 802–826.

Lancaster, T. (1979). Econometric Methods for the Analysis of Unemployment. *Econometrica* 47: 939-956.

Lee, A. J. (1990). *U-statistics: Theory and Practice.* Taylor & Francis Group, New York.

Li, Q., Wang, S. (1998). A simple Consistent Bootstrap Test for a Parametric Regression Function. *Journal of Econometrics* 87: 145–165.

Lu, X. and H. White (2011). Testing for Separability in Structural Equations. *Discussion Paper*, Dept. of Economics, UCSD.

Masry, E. (1996a). Multivariate Regression Estimation: Local Polynomial Fitting for Time Series. *Stochastic Processes and their Applications* 65: 81–101.

Masry, E. (1996b). Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency rates. *Journal of Time Series Analysis* 17: 571–599.

Matzkin R. L. (1999). Nonparametric Estimation of Nonadditive Random Functions. Mimeo, North- western University

Matzkin, R. L. (2003). Nonparametric Estimation of Nonadditive Random Functions. *Econometrica* 71: 1339-1375.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (eds.), *Frontiers in Econometrics*, . pp.105-142, Academic Press, New York.

Newey, W. K. (1985). Generalized Method of Moments Specification Testing. *Journal of Econometrics* 29: 229-256.

Olley, G. S. and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica* 64: 1263-1297.

Robinson, P. (1989). Hypothesis Testing in Semiparametric and Nonparametric Models for Econometric Time Series. *Review of Economic Studies* 56: 511-534.

Roehrig, C. S. (1988). Conditions for Identification in Nonparametric and Parametric Models. *Econometrica* 56: 433-447.

Ruud, P. A. (1984). Tests of Specification in Econometrics. *Econometric Reviews* 3: 211-242.

Stinchcombe, M. and H. White (1998). Consistent Specification Testing with Nuisance Parameters Present Only Under Alternative. *Econometric Theory* 14: 295-324.

Su, L. and H. White (2008). A Nonparametric Hellinger Metric Test for Conditional Independence. *Econometric Theory* 24: 829-864.

Tauchen, G. (1985). Diagnostic Testing and Evaluation of Maximum Likelihood Models. *Journal of Econometrics* 30: 415443.

White, H. (1987). Specification testing in dynamic models. In T. F. Bewley (eds.), *Advances in econometrics,* Fifth world congress, Vol. I. Cambridge University Press, Cambridge.

Wooldridge, J. M. (1992). A Test for Functional Form Against Nonparametric Alternatives. *Econometric Theory* 8: 452-475.

Yatchew, A. J. (1992). Nonparametric Regression Tests Based on Least Squares. *Econometric Theory* 8: 435-451.

Zheng, X. (1996). Consistent Test of Functional Form via Nonparametric Estimation Technique. *Journal of Econometrics* 75: 263-289.

# Chapter 8

# Inference in Semiparametric Partial Threshold Models

## 8.1 Introduction

Since the seminal work of Robinson (1988), partially linear model, one of the simplest semiparametric models, starts to gain its popularity in economic applications. It has been used to model hedonic price fluctuations (Anglin and Gencay, 1996), to estimate Engel curves (Blundell, Duncan and Pendakur, 1998), to estimate the relationship between weather and electricity sales (Engle, Granger, Rice and Weiss, 1986), to predict the mean effect of a change in the distribution of some policy-related variables (Stock, 1989), to study the production frontier of the U.S. banking industry (Adams, Berger and Sickles, 1999), and estimate price and income elasticities in the presence of endogeneity (Yatchew and No, 2001). Besides progress in applied work, theoretical effort has also been made. The simple *i.i.d.* model was extended by Li (1996), Fan and Li (1999) into time series setting, and by Li and Stengos (1996) into panel data framework. Hardle, Liang and Gao (2000) have a thorough treatment for partially linear models.

Yet, the stability of the parameter of the linear component does not receive enough attention. One direct approach to evaluate the stability is to estimate the par-

tially linear model for appropriately selected different subsamples and then compare the estimators. However, this approach is not statistically well developed. Another approach to tackle this problem is to consider a nonlinear generalization of the semiparametric partially linear model by incorporating a threshold component which is usually called sample splitting. Caner (2002), Caner and Hansen (2004), and Hansen (1996, 1999, 2000, 2004) investigate the theoretical properties of the threshold models. Gonzalo and Pitarakis (2002) consider issues of estimation, model selection in threshold models and estimating the number of thresholds. Hansen (2000) develops the theory of estimation of threshold model with exogenous regressors, derives the asymptotic properties of the threshold parameter estimator, constructs the confidence interval and deals with testing issues. The asymptotic distribution of derived estimator is nonstandard yet free of nuisance parameters under the specified assumptions, parallel to the results in change point theory by Picard (1985) and Bai (1997).

The primary purpose of this chapter is to develop the estimation of the semiparametric partially threshold models and derive the testing statistics. The contribution of this chapter is seven folds. First, we propose an estimation procedure to consistently estimate the threshold parameters, the slope parameters and the nonparametric component. In the presence of the nonparametric component compared to threshold model considered in Hansen (2000), our model needs special treatment of the nonparametric component in order to make the estimation feasible. We adopt Robinson's (1988) approach to eliminate the nonparametric component and estimate the conditional expectations showing up in the reduced form using kernel based approach.

Second, we prove the asymptotic properties of the threshold parameter estimator, which has a convergence rate relying on a parameter regarding the threshold effect. Its asymptotic distribution is nonstandard, yet has a similar form to that of the thresh-

old model. Third, a likelihood ratio statistic that can be used to test the threshold effect is developed and it is shown to have an asymptotic chi-squared distribution. Confidence interval estimators are also constructed based on the likelihood ratio statistic.

Fourth, we show that the slope parameters are asymptotically normally distributed and testing issues are also considered. Fifth, we prove that the nonparametric component estimator is asymptotically normally distributed and achieves oracle efficiency as if the threshold parameter is known.

Sixth, we examine the finite sample properties of our estimator and compare them with those of partially linear models and nonparametric models. The simulation results shows the necessity of adopting the semiparametric threshold model when there is such effect, in case of large samples. Seventh, we apply our model to study consumer demand as did in Blundell et al (1998). Also, we test the existence of the threshold effect and we find a threshold effect in domestic fuel Engel curve.

The rest of the chapter is organized as follows. In section 8.2, we lay out our model, present notations and comment the relationship of our models with other popular ones in the literature. Section 8.3 describes the proposed estimation procedure. Section 8.4 presents asymptotic distribution theory for the proposed estimators. Section 8.5 discusses the testing of a threshold effect. Monte Carlo experiments are performed in section 8.6 to examine the finite sample properties of our estimators and in the following section, we conduct an application to study consumer demand. Section 8.8 concludes and comments on future research. All technical proofs are collected in Appendix.

## 8.2 Model

The model of interest is

$$\begin{aligned} y_i &= \theta_1' X_i + g(Z_i) + e_i, \text{ if } q_i \leq \gamma \\ y_i &= \theta_2' X_i + g(Z_i) + e_i, \text{ if } q_i > \gamma \end{aligned}$$

or in compact form

$$y_i = \theta_1' X_i 1(q_i \leq \gamma) + \theta_2' X_i 1(q_i > \gamma) + g(Z_i) + e_i, \tag{8.1}$$

with $\{X_i, Z_i, q_i, y_i, \}_{i=1}^n$ being the observed sample, $e_i$ as the disturbance term and others as model parameters, $1(\cdot)$ being the indicator function. $X_i$ is $m \times 1$ vector of exogenous regressors and $Z_i$ is $p \times 1$ vector of exogenous regressors. $q_i$ is a real-valued and continuous threshold variable that is independent of $Z_i$, and $y_i$ is the real-valued dependent variable. $\gamma \in [\, \underline{\gamma}, \bar{\gamma}\, ]$, is the threshold parameter assumed to be unknown. $\theta_1$ and $\theta_2$ are slope parameters which are of dimension $m \times 1$. $g(\cdot)$ is a real-valued unknown function defined on $\mathbb{R}^p$. Further, we assume that $E(e_i \mid X_i, Z_i) = 0$. Note that since the model has both a threshold component and a nonparametric component, we call it "semiparametric partial threshold model," parallel to the semiparametric partial linear model.

**Remark 1:** Note that $q_i$ may be part of $X_i$ but is assumed to be independent of $Z_i$. When $q_i$ is a time index, the model is a semiparametric generalization of the models considered in change point theory, see Picard (1985) and Bai (1997). When $q_i$ is a discrete variable such as gender, there is no need to estimate the threshold parameter. Therefore, it is interesting to assume that $q_i$ is a continuous variable. The case in which $q_i$ and $Z_i$ are dependent is a more complicated situation that is beyond this paper.□

**Remark 2:** The threshold parameter $\gamma$ is assumed to be unknown. Otherwise, with the knowledge of $\gamma$, the model is simply the partially linear model considered in Robinson

(1988). The estimation can be done by splitting the sample based on $q_i$ first and then applying Robinson's approach. However, the split of the sample will lead to the loss of efficiency in the estimation of the nonparametric component. Even if $\gamma$ is known, our proposed estimation approach will trivially apply. □

**Remark 3 (identification):** Note that $X_i$ cannot contain a constant and that $X_i$ and $Z_i$ cannot have common component for the identification of $\theta_1$ and $\theta_2$.

First, suppose that $X_i$ contains a constant, that is, $\theta_1$ and $\theta_2$ include intercepts, $\alpha_1$ and $\alpha_2$, respectively. Then $\alpha_1$ and $\alpha_2$ will not be identified separately from the unknown function $g(\cdot)$. To see this, for any non-zero constant $c$, note that

$$\begin{aligned} & \alpha_1 1(q_i \leq \gamma) + \alpha_2 1(q_i > \gamma) + g(Z_i) \\ = \; & (\alpha_1 + c) 1(q_i \leq \gamma) + (\alpha_2 + c) 1(q_i > \gamma) + [g(Z_i) - c] \\ \equiv \; & \tilde{\alpha}_1 1(q_i \leq \gamma) + \tilde{\alpha}_2 1(q_i > \gamma) + \tilde{g}(Z_i). \end{aligned}$$

That is, the sum of the new intercepts and the new $g(\cdot)$ are observationally equivalent to the sum of the old ones. Therefore, the unknown function $g(\cdot)$ creates the problem of unidentification of the intercepts.

Second issue related to identification of the slope parameter arises from the estimation procedure that we propose to approach the model. As will be seen in the next section, the identification of $\theta_1$ and $\theta_2$ will require that $\Phi \equiv E\left\{[X - E(X \mid Z)][X - E(X \mid Z)]'\right\}$ be positive definite. Hence $X$ cannot contain a constant and components of $X$ cannot be a deterministic function of $Z$. Otherwise, $\Phi$ becomes singular since $X - E(X \mid Z) = 0$. See Robinson (1988) and Matzkin (2007) for more discussion on identification issue. □

## 8.3 Estimation

### 8.3.1 Estimation of model parameters

#### 8.3.1.1 Infeasible estimation procedure

In this subsection, we introduce an infeasible procedure to estimate the model parameters of (8.1) based on the observation that $E(e_i \mid Z_i) = E(E(e_i \mid X_i, Z_i) \mid Z_i) = 0$, by law of iterated expectation. Therefore, taking expectation of (8.1) conditional on $Z_i$ gives,

$$E(y_i \mid Z_i) = \theta_1' E(X_i \mid Z_i) 1(q_i \leq \gamma) + \theta_2' E(X_i \mid Z_i) 1(q_i > \gamma) + g(Z_i). \tag{8.2}$$

Subtracting (8.2) from (8.1) leads to

$$y_i - E(y_i \mid Z_i) = \theta_1' [X_i - E(X_i \mid Z_i)] 1(q_i \leq \gamma) + \theta_2' [X_i - E(X_i \mid Z_i)] 1(q_i > \gamma) + e_i.$$

Let $\tilde{y}_i = y_i - E(y_i \mid Z_i)$, $\tilde{X}_i = X_i - E(X_i \mid Z_i)$, $\tilde{X}_i(\gamma) = \tilde{X}_i 1(q_i \leq \gamma)$, $\delta_n = \theta_2 - \theta_1$, $\theta = \theta_2$, then we have

$$\begin{aligned} \tilde{y}_i &= \theta_1' \tilde{X}_i 1(q_i \leq \gamma) + \theta_2' \tilde{X}_i 1(q_i > \gamma) + e_i \\ &= \theta' \tilde{X}_i + \delta_n \tilde{X}_i(\gamma) + e_i. \end{aligned} \tag{8.3}$$

We can write the model (8.3) in compact form, by defining $\tilde{y} = [\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n]'$, $\tilde{X} = \left[\tilde{X}_1', \tilde{X}_2', ..., \tilde{X}_n'\right]'$, $\tilde{X}_\gamma = \tilde{X}(\gamma) = \left[\tilde{X}_1(\gamma)', \tilde{X}_2(\gamma)', ..., \tilde{X}_n(\gamma)'\right]'$, and $e = [e_1, e_2, ..., e_n]'$, as

$$\tilde{y} = \tilde{X}\theta + \tilde{X}_\gamma \delta_n + e. \tag{8.4}$$

Therefore, the parameters of the model are $(\theta, \delta_n, \gamma)$, which can be estimated by least squares (LS), jointly minimizing the sum of squared residuals (SSE) as

$$\tilde{S}_n(\theta, \delta_n, \gamma) = \left(\tilde{y} - \tilde{X}\theta + \tilde{X}_\gamma \delta_n\right)' \left(\tilde{y} - \tilde{X}\theta + \tilde{X}_\gamma \delta_n\right). \tag{8.5}$$

Note that for given $\gamma$, (8.4) is linear in $(\theta', \delta_n')$. Therefore, the estimation of the slope parameters can be easily done via conditional OLS by concentrating out $\gamma$. That is, the OLS estimators of the slope parameters, $\beta \equiv (\theta', \delta_n')'$, are derived by regressing $\tilde{y}$ on $X^* \equiv \left[\tilde{X}, \tilde{X}_\gamma\right]$ as, for given $\gamma$,

$$\begin{pmatrix} \tilde{\theta}(\gamma) \\ \tilde{\delta}_n(\gamma) \end{pmatrix} = \begin{bmatrix} \tilde{X}'\tilde{X} & \tilde{X}'\tilde{X}_\gamma \\ \tilde{X}_\gamma'\tilde{X} & \tilde{X}_\gamma'\tilde{X}_\gamma \end{bmatrix}^{-1} \begin{pmatrix} \tilde{X}' \\ \tilde{X}_\gamma{}' \end{pmatrix} \tilde{y},$$

i.e.,

$$\tilde{\beta}(\gamma) = \begin{pmatrix} \tilde{\theta}(\gamma) \\ \\ \tilde{\delta}_n(\gamma) \end{pmatrix} = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}\tilde{y} \tag{8.6}$$

Conditional on $\gamma$, the concentrated SSE is

$$\tilde{S}_n(\gamma) = \tilde{S}_n(\theta(\gamma), \delta_n(\gamma), \gamma) = \left(\tilde{y} - \tilde{X}\theta(\gamma) + \tilde{X}_\gamma\delta_n(\gamma)\right)' \left(\tilde{y} - \tilde{X}\theta(\gamma) + \tilde{X}_\gamma\delta_n(\gamma)\right).$$

Hence, $\gamma$ can be estimated by

$$\tilde{\gamma} = \arg\min_{\gamma \in \Gamma_n} \tilde{S}_n(\gamma) \tag{8.7}$$

with $\Gamma_n = \Gamma \cap \{q_1, q_2, ..., q_n\}$. Note that the minimization is well-defined since $S_n(\gamma)$ is to be evaluated for at most $n$ different values of $\gamma$.

With estimated $\tilde{\gamma}$ given by (8.7), the OLS estimators of the slopes can be formulated by substituting $\tilde{\gamma}$ for $\gamma$ in (8.6). Therefore, the estimator of $\beta$ is

$$\tilde{\beta} = \tilde{\beta}(\tilde{\gamma}). \tag{8.8}$$

Note that idea of the approach taken here to estimate the slope parameters goes as follows. First, we eliminate the nonparametric component to simplify the model to be of the form of Hansen (2000). Then we adopt Hansen's LS method estimator to estimate the threshold and slopes. However, in the process of getting rid of $g(\cdot)$, we introduce the unknown conditional expectation terms $E(X_i \mid Z_i)$ and $E(y_i \mid Z_i)$ into

the transformed model (8.3), which makes the estimation procedure infeasible. The estimation of these conditional expectations are discussed in the following subsection.

#### 8.3.1.2 Feasible estimation procedure

This subsection presents the feasible estimation procedure for the model parameters. Since $\tilde{X}_i = X_i - E(X_i \mid Z_i)$ and $\tilde{y}_i = y_i - E(y_i \mid Z_i)$ are unknown, we can replace them by consistent nonparametric kernel counterparts, $X_i - \hat{X}_i$ and $y_i - \hat{y}_i$, with

$$\begin{aligned} \hat{X}_i &\equiv \hat{E}(X_i \mid Z_i) = n^{-1} \sum_{j=1}^{n} X_j K_h(Z_i, Z_j) / \hat{f}(Z_i), \\ \hat{y}_i &\equiv \hat{E}(y_i \mid Z_i) = n^{-1} \sum_{j=1}^{n} y_j K_h(Z_i, Z_j) / \hat{f}(Z_i), \end{aligned}$$

and

$$\hat{f}(Z_i) = n^{-1} \sum_{j=1}^{n} K_h(Z_i, Z_j)$$

where

$$K_h(Z_i, Z_j) = \Pi_{s=1}^{p} h_s^{-1} k\left(\frac{Z_{is} - Z_{js}}{h_s}\right)$$

with $k(\cdot)$ being the univariate kernel function and $h = (h_1, h_2, ..., h_p)$ the bandwidth.

Note that the random denominator $\hat{f}(Z_i)$ can cause technical difficulties in the derivation of the asymptotic distribution of the feasible estimator of $\beta$. Robinson (1988) apply a function to "trim out" the observations with small denominator values, $\hat{f}(Z_i)$. Concern with the trim out approach is that not all information is to be used since it leaves out some sample observations. Another approach to tackle the small value problem of $\hat{f}(Z_i)$ is to weight the regression equation (8.3) by the density itself. Following Li (1996), multiply (8.3) by $\hat{f}_i = \hat{f}(Z_i)$,

$$\begin{aligned} \tilde{y}_i \hat{f}_i &= \theta_1' \tilde{X}_i \hat{f}_i 1(q_i \leq \gamma) + \theta_2' \tilde{X}_i \hat{f}_i 1(q_i > \gamma) + e_i \hat{f}_i \qquad (8.9) \\ &= \theta' \tilde{X}_i \hat{f}_i + \delta_n \tilde{X}_i(\gamma) \hat{f}_i + e_i \hat{f}_i. \end{aligned}$$

Denote $\bar{y}_i = \tilde{y}_i \hat{f}_i, \bar{X}_i = \tilde{X}_i \hat{f}_i, \bar{X}_i(\gamma) = \tilde{X}_i(\gamma) \hat{f}_i$, and denote $\bar{y} = [\bar{y}_1, \bar{y}_2, ..., \bar{y}_n], \bar{X} = \left[\bar{X}_1, \bar{X}_2, ..., \bar{X}_n\right]$, $\bar{X}_\gamma = \left[\bar{X}_1(\gamma), \bar{X}_2(\gamma), ..., \bar{X}_n(\gamma)\right]$ and $\bar{X}^* = \left[\bar{X}, \bar{X}_\gamma\right]$. The LS estimator of $\beta \equiv (\theta', \delta_n')'$ is given by, for given $\gamma$,

$$\hat{\beta}(\gamma) = \left(\bar{X}^{*\prime}\bar{X}^*\right)^{-1} \bar{X}^{*\prime}\bar{y}. \tag{8.10}$$

And we estimate $\gamma$ as by minimizing the SSE

$$\begin{aligned}
\bar{S}_n(\gamma) &= \bar{S}_n(\theta(\gamma), \delta_n(\gamma), \gamma) \\
&= \left(\bar{y} - \bar{X}\theta(\gamma) + \bar{X}_\gamma \delta_n(\gamma)\right)' \left(\bar{y} - \bar{X}\theta(\gamma) + \bar{X}_\gamma \delta_n(\gamma)\right) \\
&= \left(\bar{y} - \bar{X}^* \hat{\beta}(\gamma)\right)' \left(\bar{y} - \bar{X}^* \hat{\beta}(\gamma)\right),
\end{aligned}$$

that is,

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma_n} \bar{S}_n(\gamma). \tag{8.11}$$

Substituting $\hat{\gamma}$ into the slope estimator $\hat{\beta}(\gamma)$ in (8.10) gives the feasible estimator of $\beta$ as

$$\hat{\beta} = \hat{\beta}(\hat{\gamma}) \tag{8.12}$$

To solve the minimization problem, see Hansen (1999) for details of computation issues.

### 8.3.2 Estimation of the nonparametric component

To estimate the nonparametric function, note that (8.2) implies that

$$\begin{aligned}
g(Z_i) &= E\left(y_i - \theta_1' X_i 1(q_i \le \gamma) - \theta_2' X_i 1(q_i > \gamma) \mid Z_i\right) \\
&= E\left(y_i - \theta' X_i - \delta_n' X_i(\gamma) \mid Z_i\right).
\end{aligned}$$

Therefore, with a consistent estimator of $\beta$, $\hat{\beta} = \left(\hat{\theta}', \hat{\delta}_n'\right)'$, $g(z)$ is consistently estimated as

$$\hat{g}(z) = \frac{\sum_{j=1}^n \left(y_j - \hat{\theta}' X_j - \hat{\delta}_n' X_j(\hat{\gamma})\right) K_h(z, Z_j)}{\sum_{j=1}^n K_h(z, Z_j)}, \tag{8.13}$$

the asymptotic distribution of which is shown, under conditions in Theorem 4 below, to be same as the following infeasible estimator that makes use of the true value of the threshold and slope parameters,

$$\tilde{g}(z) = \frac{\sum_{j=1}^{n} (y_j - \theta' X_j - \delta_n' X_j(\gamma)) K_h(z, Z_j)}{\sum_{j=1}^{n} K_h(z, Z_j)}. \tag{8.14}$$

## 8.4 Distribution theory

### 8.4.1 Assumption

We state our assumptions after the definition of a class of functions and some moments notations to be used later.

**Definition 1:** For $\alpha > 0$, and integer $v \geq 2$, define $\mathcal{G}_v^\alpha$ as the class of smooth functions such that if $g \in \mathcal{G}_v^\alpha$, then $g$ is $v$ times differentiable; $g$ and its partial derivatives all satisfy the Lipschitz-type conditions such as $\|g(z_1) - g(z_2)\| \leq H_g(z) \|z_1 - z_2\|$, for a continuous function $H_g(z)$ with finite $\alpha$th moment. □

Let $\varphi(\cdot)$ and $f(\cdot)$ be the density function of $q_i$ and $Z_i$, respectively. Denote $\varphi_i = \varphi(q_i)$ , $\varphi = \varphi(\gamma_0)$ and $f_i = f(z_i)$.

**Definition 2:** Define the moment functionals (following Hansen, 2000)

$$\begin{aligned}
M_1(\gamma) &= E\left[\tilde{X}_i \tilde{X}_i' 1(q_i \leq \gamma)\right], \ \bar{M}_1(\gamma) = E\left[\tilde{X}_i \tilde{X}_i' f_i^2 1(q_i \leq \gamma)\right] \\
M_2(\gamma) &= E\left[\tilde{X}_i \tilde{X}_i' 1(q_i > \gamma)\right], \ \bar{M}_2(\gamma) = E\left[\tilde{X}_i \tilde{X}_i' f_i^2 1(q_i > \gamma)\right] \\
D(\gamma) &= E\left[\tilde{X}_i \tilde{X}_i' \mid q_i = \gamma\right], \bar{D}(\gamma) = E\left[\tilde{X}_i \tilde{X}_i' f_i^2 \mid q_i = \gamma\right] \\
V(\gamma) &= E\left[\tilde{X}_i \tilde{X}_i' e_i^2 \mid q_i = \gamma\right], \bar{V}(\gamma) = E\left[\tilde{X}_i \tilde{X}_i' e_i^2 f_i^4 \mid q_i = \gamma\right] \\
\Omega_1(\gamma) &= E\left[\tilde{X}_i \tilde{X}_i' e_i^2 \mid q_i \leq \gamma\right], \bar{\Omega}_1(\gamma) = E\left[\tilde{X}_i \tilde{X}_i' e_i^2 f_i^4 \mid q_i \leq \gamma\right] \\
\Omega_2(\gamma) &= E\left[\tilde{X}_i \tilde{X}_i' e_i^2 \mid q_i > \gamma\right], \bar{\Omega}_2(\gamma) = E\left[\tilde{X}_i \tilde{X}_i' e_i^2 f_i^4 \mid q_i > \gamma\right]
\end{aligned}$$

and denote $M = E\left[\tilde{X}_i \tilde{X}_i'\right]$, $\bar{M} = E\left[\tilde{X}_i \tilde{X}_i' f_i^2\right]$, $D = D(\gamma_0)$, $V = V(\gamma_0)$, $\bar{D} = \bar{D}(\gamma_0)$,

$\bar{V} = \bar{V}(\gamma_0)$, $\Omega_1 = \Omega_1(\gamma_0)$, $\Omega_2 = \Omega_2(\gamma_0)$, $\bar{\Omega}_1 = \bar{\Omega}_1(\gamma_0)$, $\bar{\Omega}_2 = \bar{\Omega}_2(\gamma_0)$ where $\gamma_0$ denote the true value of $\gamma$. □

**Assumption 1.**

1.1 $\{X_i, Z_i, q_i, y_i,\}_{i=1}^n$ is strictly stationary, ergodic and $\rho$-mixing, with $\rho(\tau) = O\left(\tau^{1+\epsilon}\right)$ for some small $\epsilon > 0$.

1.2 PDF of $Z_i$, $f(\cdot) \in \mathcal{G}_{v+1}^{\infty}$, has three-times bounded continuous derivatives, $g(\cdot) \in \mathcal{G}_v^4$, and $E(X \mid z) \in \mathcal{G}_v^4$ for integer $v \geq 2$.

1.3 $E(e_i \mid I_{i-1}) = 0$, where $I_{i-1}$ denotes the information set available up to $i$.

1.4 $E \mid \tilde{X}_i \mid^4 < \infty$ and $E \mid \tilde{X}_i e_i \mid^4 < \infty$.

1.5 For all $\gamma \in \Gamma$, $E\left[\mid \tilde{X}_i \mid^4 e_i^4 \mid q_i = \gamma\right] \leq C$, and $E\left[\mid \tilde{X}_i \mid^4 \mid q_i = \gamma\right] \leq C$ for some $C < \infty$, and $\varphi(\gamma) \leq \bar{\varphi} < \infty$.

1.6 $\varphi(\cdot)$, $D(\cdot)$, and $V(\cdot)$ are continuous at $\gamma_0$.

1.7 $\delta_n = cn^{-\lambda}$, with $c \neq 0$ and $0 < \lambda < \frac{1}{2}$.

1.8 $c'Dc > 0$, $c'Vc > 0$, and $\varphi > 0$.

1.9 $M > M(\gamma) > 0$ for all $\gamma \in \Gamma$.

**Remark 4.** The above assumptions are quite similar to those in Hansen (2000, 2004), which are needed to prove the asymptotic properties of the (infeasible) estimators of $\beta$ and $\gamma$.

Assumption 1.1 is trivially satisfied for *i.i.d.* observations and accommodates weakly dependent time series process. Stationarity excludes trends and unit root process. Assumption 1.2 places restrictions on smoothness and moment of, the density of $Z_i$, the unknown nonparametric component $g(\cdot)$, and the conditional expectation of $X$ given $Z = z$. Assumption 1.3 states that $e_i$ is a martigale difference sequence with respect to available information set. Assumption 1.4 and 1.5 impose the boundedness of moments. Assumption 1.6 requires the continuity of the density function of the

threshold variable at $\gamma_0$. Assumption 1.7 dictates that the difference between regression slopes is getting smaller as sample size gets larger, which indicates that the asymptotic approximation holds when $\delta_n$ is small. The magnitude of $\delta_n$ is modeled through the pre-specified parameter $\lambda$, the smaller a value of which makes the model less restrictive. The introduction of $\delta_n$ is for the purpose of achieving an asymptotic distribution free of nuisance parameters, as did in Hansen (1996, 2000), and Chan (1993) in the change point theory. Assumption 1.8 and 1.9 are full rank conditions for the defined moments, to exclude degenerate asymptotic distributions and multicollinearity. □

**Assumption 2.**

2.1 $K(\cdot)$ is a product kernel, the univariate kernel $k(\cdot)$ is a bounded $v$th order kernel, and $k(s) = O\left(1/\left[1+\mid s \mid\right]^{v+1}\right)$.

2.2 $E(e_i^2|z) = \sigma^2(z)$ belongs to $\mathcal{G}_1^2$.

2.3 As $n \to \infty$, $n(h_1...h_p)^2 \to \infty$, and $n\sum_{s=1}^{p} h_s^{4v} \to 0$.

**Remark 5:** This set of assumptions, together with Assumption 1, are for the purpose of deriving the asymptotic property of the feasible estimators and the nonparametric component. Assumption 2.1 states that the multivariate kernel function is the product of univariate kernels that satisfying certain regularity conditions. It controls the order of the bias term in the density estimation. Assumption 2.2 and 2.3 play similar roles as those in Fan and Li (1999) to prove the $\sqrt{n}$-consistency of the slope estimator of partial linear models. □

### 8.4.2 Asymptotic distribution theory

This subsection states the asymptotic properties of the estimators described in the previous section.

#### 8.4.2.1 Threshold estimate

We need the following definition to present our asymptotic distribution.

**Definition 3:** A two-sided Brownian motion $B(r) : \mathbb{R} \to [0, \infty)$ is defined as

$$B(r) = \begin{cases} B_1(-r), & \text{if } r < 0, \\ 0, & \text{if } r = 0, \\ B_2(r), & \text{if } r < 0, \end{cases}$$

with $B_1(r)$ and $B_2(r)$ being independent standard Brownian motions on $[0, \infty)$. □

**Proposition 1:** Suppose that Assumption 1 holds, we have

$$n^{1-2\lambda}(\tilde{\gamma} - \gamma_0) \to_d \eta\xi,$$

where

$$\eta = \frac{c'Vc}{(c'Dc)^2 \varphi}$$

and

$$\xi = \arg\max_{r \in \mathbb{R}} \left[ -\frac{1}{2} \mid r \mid + B(r) \right].$$

□

The likelihood ratio statistic defined as

$$LR_1(\gamma) = n \frac{\tilde{S}_n(\gamma) - \tilde{S}_n(\tilde{\gamma})}{\tilde{S}_n(\tilde{\gamma})},$$

can be used to test the null hypothesis $H_0 : \gamma = \gamma_0$. We have the following theorem that establishes the asymptotic distribution of the statistic.

**Proposition 2:** Suppose that Assumption 1 holds and that $e_i$ is *i.i.d.* $N\left(0, \sigma^2\right)$, we have

$$LR_1(\gamma_0) \to_d \tau\varsigma,$$

where

$$\tau = \frac{c'Vc}{\sigma^2 (c'Dc)},$$

and

$$\varsigma = \max_{r \in \mathbb{R}} \left[ - \mid r \mid + 2B(r) \right],$$

with distribution function $P(\varsigma \leq x) = \left(1 - e^{-x/2}\right)^2$. $\square$

**Remark 6:** The above propositions establish the asymptotic distribution of the threshold estimator and that of the likelihood ratio test for hypothesis on $\gamma$, which is nonstandard yet free of nuisance parameter under conditional homoskedasticity $E(e_i \mid q_i) = \sigma^2$, in the same line as Hansen (2000). The closed form distribution function $P(\cdot)$ can easily generate the $p$-values for observed test statistics. See TABLE I of Hansen (2000) for critical values of the statistic. $\square$

The proof of Proposition 1 and Proposition 2 goes through by verifying that Assumption 1 in Hansen (2000) holds. Therefore, the two propositions follow and they will be used to show the properties for our feasible estimators in the following theorems.

**Theorem 1:** Suppose that Assumption 1 and 2 hold, we have

$$n^{1-2\lambda} (\hat{\gamma} - \gamma_0) \rightarrow_d \bar{\eta} \xi,$$

with

$$\bar{\eta} = \frac{c' \bar{V} c}{\left(c' \bar{D} c\right)^2 \varphi}$$

and $\xi$ being defined as in Proposition 1. $\square$

**Theorem 2:** Suppose that Assumption 1 and 2 hold, and that $e_i$ is $i.i.d.$ $N\left(0, \sigma^2\right)$, we have

$$LR_2(\gamma_0) \rightarrow_d \tau \varsigma,$$

with

$$\bar{\tau} = \frac{c' \bar{V} c}{\sigma^2 \left(c' \bar{D} c\right)},$$

, $\varsigma$ being defined in Proposition 2 and

$$LR_2(\gamma) = n\frac{\bar{S}_n(\gamma) - \bar{S}_n(\hat{\gamma})}{\bar{S}_n(\hat{\gamma})}.$$

□

**Remark 7:** The above two theorems establish the asymptotic distributions of the feasible threshold estimator and the likelihood ratio statistic. The asymptotic distributions are similar to those of Theorem 1 and 2, respectively. Thus, our feasible estimator $\hat{\gamma}$, of $\gamma_0$, achieves the oracle efficiency that can only be achieved with known conditional expectations $E(X_i \mid Z_i)$ and $E(y_i \mid Z_i)$. So does the likelihood ratio test statistic. □

**Remark 8:** The nuisance parameter $\bar{\tau}$ in the asymptotic distribution in Theorem 2 can be consistently estimated through either a polynomial regression or a kernel regression. Denote $l_{1i} = (\delta_n' x_i)^2 \left(e_i^2 f_i^2/\sigma(x,z)\right)$ and $l_{2i} = (\delta_n' x_i)^2 f_i^2$. We have

$$\bar{\tau} = \frac{E(l_{1i}|q_i = \gamma_0)}{E(l_{2i}|q_i = \gamma_0)}. \tag{8.15}$$

Replace with the unobserved variables by their sample counterparts, we have $\hat{l}_{1i} = \left(\hat{\delta}_n' x_i\right)^2 \left(\hat{e}_i^2 \hat{f}_i^2/\hat{\sigma}(x,z)\right)$ and $l_{2i} = \left(\hat{\delta}_n' x_i\right)^2 \hat{f}_i^2$. Therefore, (8.15) can be estimated by a quadratic regression as

$$\bar{\tau} = \frac{\hat{\varkappa}_{10} + \hat{\varkappa}_{11}\hat{\gamma} + \hat{\varkappa}_{12}\hat{\gamma}^2}{\hat{\varkappa}_{20} + \hat{\varkappa}_{21}\hat{\gamma} + \hat{\varkappa}_{22}\hat{\gamma}^2}$$

where $\hat{l}_{ji} = \hat{\varkappa}_{10} + \hat{\varkappa}_{11} q_i + \hat{\varkappa}_{12} q_i^2 + \hat{\epsilon}_{ji}$ is the OLS fit, for $j = 1$ and 2, or by Nadaraya-Watson kernel estimator as

$$\bar{\tau} = \frac{\sum_{i=1}^n K_h(\hat{\gamma} - q_i)\hat{l}_{1i}}{\sum_{i=1}^n K_h(\hat{\gamma} - q_i)\hat{l}_{2i}},$$

where $K_h(s) = \frac{1}{h}K\left(\frac{u}{h}\right)$ with $h$ being the bandwidth and $K(s)$ being kernel function such as normal kernel or Epanechnikov. For details about kernel estimation, see Pagan and Ullah (1999) or Li and Racine (2007), for example.

#### 8.4.2.2 Slope parameters

This subsection presents the asymptotic distribution of the estimators of the slope parameters.

**Proposition 3:** Suppose that Assumption 1 and 2 hold,

$$n^{1/2}\left(\tilde{\beta}-\beta\right) \rightarrow_d \tilde{W} \sim N\left(0, V_1\right)$$

with $V_1$ being the standard asymptotic covariance matrix if $\gamma = \gamma_0$ were fixed.

Or

$$\begin{aligned} n^{1/2}\left(\tilde{\theta}_1-\theta_1\right) &\rightarrow {}_dN\left(0,\tilde{V}_1\right) \\ n^{1/2}\left(\tilde{\theta}_2-\theta_2\right) &\rightarrow {}_dN\left(0,\tilde{V}_2\right) \end{aligned}$$

where

$$\begin{aligned} \tilde{V}_1 &= \left(M_1'\Omega_1 M_1\right)^{-1}, \\ \tilde{V}_2 &= \left(M_2'\Omega_2 M_2\right)^{-1}. \end{aligned}$$

□

Proposition 3 follows from Lemma A.12 of Hansen (2000), the proof of which is therefore omitted. With this result, we are ready to establish the following theorem:

**Theorem 3:** Suppose that Assumption 1 and 2 hold,

$$n^{1/2}\left(\hat{\beta}-\beta\right) \rightarrow_d W \sim N\left(0, V_2\right)$$

with $V_2$ being the standard asymptotic covariance matrix if $\gamma = \gamma_0$ were fixed.

Or

$$\begin{aligned} n^{1/2}\left(\hat{\theta}_1-\theta_1\right) &\rightarrow {}_dN\left(0,\bar{V}_1\right) \\ n^{1/2}\left(\hat{\theta}_2-\theta_2\right) &\rightarrow {}_dN\left(0,\bar{V}_2\right) \end{aligned}$$

where

$$\begin{aligned}\bar{V}_1 &= \left(\bar{M}_1'\bar{\Omega}_1\bar{M}_1\right)^{-1}, \\ \bar{V}_2 &= \left(\bar{M}_2'\bar{\Omega}_2\bar{M}_2\right)^{-1}.\end{aligned}$$

□

**Remark 9:** The above theorem establishes the asymptotic distributions of the feasible slope estimator. The asymptotic normality and the covariance matrix indicate that we can approximate the distribution of $\theta_1$ and $\theta_2$ by traditional normal distribution, treating $\gamma$ as known. Therefore, the first stage estimation of $\gamma$ does not contribute to the loss of efficiency of the slope estimator. With the asymptotic distribution, we are ready to construct confidence intervals for both slope parameters, which is omitted here.

□

#### 8.4.2.3 Nonparametric component

This subsection presents the asymptotic distribution of the estimator of the nonparametric component. First, we state the result for the infeasible estimator given in (8.14) as in the following proposition.

**Proposition 4:** Under Assumption 1 and 2,

$$\left(nh_1...h_p\right)^{1/2}\left(\tilde{g}\left(z\right)-g\left(z\right)-\sum_{s=1}^{p}h_s^2B_s\left(z\right)\right)\rightarrow_d N\left(0,V\left(z\right)\right)$$

with

$$\begin{aligned}B_s\left(z\right) &= \frac{1}{2}k_2\left[f\left(z\right)g_{ss}\left(z\right)+2f_s\left(z\right)g_s\left(z\right)\right], \\ V\left(z\right) &= k^p\sigma^2\left(z\right)/f\left(z\right).\end{aligned}$$

□

Proof of proposition 4 follows from Theorem 18.4 of Li and Racine (2007). With Proposition 4, we can derive the following theorem for the feasible nonparametric component estimator.

**Theorem 4:** Under Assumption 1, 2 and the conditions, $\hat{\beta} - \beta = O\left(n^{-1/2}\right)$ and $\hat{\gamma} - \gamma_0 = O\left(n^{2\lambda-1}\right) = o(1)$,

$$(nh_1...h_p)^{1/2}\left(\hat{g}(z) - g(z) - \sum_{s=1}^{p} h_s^2 B_s(z)\right) \to_d N(0, V(z))$$

with

$$\begin{aligned} B_s(z) &= \frac{1}{2}k_2\left[f(z)\, g_{ss}(z) + 2f_s(z)\, g_s(z)\right], \\ V(z) &= k^p \sigma^2(z)/f(z). \end{aligned}$$

□

**Remark 10:** The asymptotic property of the nonparametric component estimator coincides that of partial linear model considered in Fan and Li (1996). The bias and variance term keep the same order therefore the introduction of the threshold parameter and its estimation do not lead any loss of efficiency for the estimation of the nonparametric component.

## 8.5 Testing for a threshold

It is very important to test whether the threshold effect conjectured is statistically significant or not. We consider the test for hypothesis of no threshold,

$$H_0 : \theta_1 = \theta_2.$$

The null is equivalent to $H_0' : \delta_n = 0$. Under the null, the threshold $\gamma$ is not identified, which leads to the collapse of the standard distribution of classical test statistics. We consider the likelihood ratio statistic for such a test. This typically called 'Davies'

problem (Davies, 1977, 1987) has been investigated by Andrew and Ploberger (1994) and Hansen (1996). We follow the suggestion of Hansen (1996) to adopt a bootstrap procedure to simulate the asymptotic distribution of the likelihood ratio test. See Hansen (1996, 1999) for more details.

## 8.6 Monte Carlo

In this section, we conduct Monte Carlo simulations to examine the finite sample performance of our proposed model, with comparison to nonparametric models and semiparametric partial linear models of Robinson (1988). We compute the coverage probability for nominal 90% confidence intervals of the threshold estimator. We compute the out-of-sample Mean Squared Errors of these models to examine their predictability.

### 8.6.1 DGP

We consider the following Data Generating Process (DGP), for $i = 1, ..., n$,

$$\begin{aligned} y_i &= \theta x_i + \delta_n x_i 1_{\{q_i \leq \gamma\}} + \exp\{-z_i\} + \varepsilon_i \\ x_i &\sim N(2,1), \\ z_i &\sim U[0,1], \varepsilon_i \sim N(0,1). \end{aligned}$$

with $\theta = 1, \gamma = 2$, $\delta_n = \{0.25, 0.50, 1.00, 2.00\}$ and $n = \{50, 100, 200, 400\}$. $q_i$ is assumed to be $N(2,1)$, independent of $x_i$. For nonparametric model, it is estimated via local constant kernel estimator,

$$\hat{m}(x,z) = \frac{\sum_{i=1}^{n} k_h(x_i, z_i; x, z)\, y_i}{\sum_{i=1}^{n} k_h(x_i, z_i; x, z)},$$

where $k_h(x_i, z_i; x, z) = k\left(\frac{x_i - x}{h_x}\right) \times k\left(\frac{z_i - z}{h_z}\right)$, and $k(\cdot)$ is chosen as second order Epanechnikov kernel, $h_x$ and $h_z$ are chosen to minimize the integrated mean squared error through cross-validation. For partial linear model, we first estimate $\beta$ as proposed in

Robinson (1988) and then estimate the nonparametric function using local constant kernel estimator.

### 8.6.2 Monte Carlo results

We report results of two kinds. First, we compute the coverage probability for the threshold confidence interval, i.e., the fraction of simulations in which the estimated confidence intervals covers the true value of the threshold. The coverage probability at 90% nominal level is reported in TABLE 1.

Second, we compare the out-of-sample performance of the proposed model with nonparametric model and partial linear model. We evaluate the estimated model at 400 grid data points formed by $x$ and $z$, in which $x$ takes values from 1 to 3 with equal increments, and $z$ takes values from 0.2 to 0.8 with equal increments. We compute the relative Mean Squared Errors using following formula,

$$RMSE_{base} = \left( \frac{MSE_{pt}}{MSE_{base}} - 1 \right) \times 100,$$

where "pt" refers to partial threshold model, and "base" refers to either nonparametric model or partially linear model. The computed RMSE is the percentage improvement achieved by partial threshold model in terms of Mean Squared Error. Negative values of RMSE indicates that partial threshold model outperforms the base model with the absolute value measuring the percentage improvement. Relative MSE for our proposed model based on partially linear model is shown in TABLE 2 and that based on nonparametric model is shown in TABLE 3.

Several findings are in sequence. It is conclusive from TABLE 2 and TABLE 3 that the semiparametric partial threshold model with the proposed estimation approach beats both nonparametric model and partially linear model, when $\delta_n$ is large and/or when $n$ is large. This confirms assumption 1.7 that the difference in the slope param-

eter should shrink as sample size diverges. First, for small $\delta_n$ and small $n$, partially linear model outperforms the proposed model. This reveals that uncertainty induced by the first stage estimation of the threshold parameter deteriorates the model when the threshold effect is very small. As the accuracy of the estimate of the threshold value critically depends on the sample (size), it is no wonder that ignoring the threshold effect in small sample improves over the partial threshold model, in out-of-sample performance. Second, for small $n$ and moderate values of $\delta_n$, nonparametric exhibits better out-of-sample predictive power. This again shows the big effect of first stage estimation of threshold parameter on the subsequent nonparametric estimation. When sample size gets larger or the threshold effect gets larger, specification of the correct model by bringing in threshold effect and additive structure becomes valuable.

TABLE 1 shows that the coverage probabilities are generally above but close to the nominal 90% level. This implies that the confidence intervals are generally conservative. The proposed estimation procedure is able to estimate the threshold parameter value but the accuracy is subject to large uncertainty. This, together with the above finding, shows that the imperfect performance of the proposed model is largely due to the first stage estimation of the threshold parameter, which is very sensitive to sample size.

[Insert TABLE 1, 2, 3 here]

## 8.7 Application

In this section, we apply our model to study consumer demand where semiparametric methods find a lot of applications. We extend the partially linear model of Blundell et al (1998) by considering the threshold effects in the linear component. As

argued in Blundell et al (1998), it is important to recover an accurate specification of Engel curve relationship and they have shown the importance of the restriction placed on the shape of Engel curve. Nevertheless, we conjecture that there might be threshold effect in the linear component of the model which might jeopardize the shape of the estimated Engel curve.

We consider the following specification,

$$w_{ij} = \theta_1' X_i + g(Z_i) + e_{ij}, \tag{8.16}$$

where $w_{ij}$ is the budget share of the $j$th good for individual $i$, $X_i$ is the log of total expenditure for individual $i$. We adopt the data set investigated in Blundell et al (1998), drawn from 1980-1982 British Family Expenditure Surveys[1]. We consider Engel curve relationship for six broad categories: Food, Fuel, Clothing, Alcohol, Transport and Other good. From the available data, we take Age as the variable $Z_i$ that enters the Engel curve nonparametrically. We conjecture that the total income would have an influence on the Engel coefficient, therefore it is taken as threshold variable in our model. See Table I of Blundell et al (1998) for data description statistics of the 1519 observations.

We modify the Lagrangian Multiplier (LM) test developed in Hansen (1996), to test the existence of threshold effect. Since $\gamma$ is not identified under the null hypothesis of no threshold effect, we perform the bootstrap analog to compute the $p$-values, with estimated residual $\hat{e}_{ij}$. We generating the bootstrap dependent variable from the distribution $N\left(0, \hat{e}_{ij}^2\right)$ and keeping the regressors on the right-hand side of (8.16) fixed.

Figure 1 depicts, for Food Engel curve, the normalized likelihood ratio statistic sequence $LR_n^*(\gamma)$ as a function of the threshold, which shows that there is no evidence for the existence of threshold. The generated $p$-value is 0.537, which favors the null

[1]The data set is available at http://qed.econ.queensu.ca/jae/1998-v13.5/blundell-duncan-pendakur/

of no threshold. Figure 2 presents the case for Fuel Engel curve where a threshold exist. The bootstrapped $p$-value is 0 and the estimated threshold value is $\hat{\gamma} = 350.0$ as shown in Figure 3. However the estimated confidence interval for the estimate is $[20.0, 1110.0]$, which is rather broad, suggesting considerable uncertainty about the value of the threshold. The estimation result is presented in TABLE 4. Figures 4-7 display the normalized likelihood ratio statistic sequence against the threshold for Engel curve in other good classes, where there is no threshold. To save space, we do not report estimation results for these cases.

[Insert TABLE 4 here]

We conclude that there is a threshold in Engel curve for domestic fuel consumption and there is significant uncertainty regarding its value. Although the Engel coefficients are not significantly different from each other, this difference might lead to wrong policy recommendation. Therefore, it is necessary for us to build up a general model as in (8.16), although we did not find any threshold effect in Engel curves in other classes of consumption, which favors the popular semiparametric partial linear model.

[Insert Figure 1-7 here]

## 8.8 Concluding remarks

This chapter considers the estimation of semiparametric partial threshold model which is a combination of the popular model considered in Robinson (1988) and that in Hansen (2000). We develop a two-step approach to estimate the threshold parameter and the slope parameter, as well as the nonparametric component. We prove the consistency of these estimators and derive their asymptotic distributions. The estimators preserve the same asymptotic distributions as their counterparts in Robinson (1988)

and Hansen (2000). We examine finite sample properties of our distribution theory with Monte Carlo simulations, followed by an application of the proposed model to study consumer demand. Our results show the necessity of taking into account of both the threshold effect and nonlinear relationship signaling through the nonparametric component.

There are several directions towards future research regarding the proposed semiparametric partial threshold models. First, the explanatory variables considered in this chapter are assumed to be exogenous, while this is rarely the case in practice. Therefore, instrumental variable approach to the proposed model calls for attention and it is under investigation by the author. Second, extension to panel data setting is also of great interest, see Hansen (1999) and Fan and Stengos (1996) for valuable reference. Third, the model can be modified by applying a smooth weighting function instead of an indicator function as having been done in Seo and Linton (2007). Effort towards this direction has been made in the estimation of the threshold model, while it is interesting to see similar methodology progress in our proposed models.

Semiparametric partial threshold models do not receive much attention yet, partly due to its complex model structure. However, when we realized the instability of the estimate of the slope parameter of the partial linear model, the literature will be directed towards research on this edge. Hopefully, this work will attract more research into this area.

## Acknowledgements

We would like to express our deep gratitude to Aman Ullah and Goloria Gonzalez-Rivera for their helpful comments. And we thank Bruce E. Hansen for his generosity of offering his matlab code available online[2].

[2]Matlab code for application results of Hansen (2000) is available at Hansen's Webpage:

## Appendix A: Proof of Theorem

In Appendix A, we first prove some lemmas that will be used in the proof of the theorems, in which we will apply some useful lemmas presented in Appendix B.

**Lemma A.1:** $\hat{\gamma} \to_p \gamma_0$.

**Proof:** Parallel to the proof of lemma 5 of Hansen (2000), we can apply Lemma B.1, Lemma B.12, Assumption 1.7 to show that $\hat{\gamma}$ minimizes $\bar{S}_n(\gamma)$, which is uniquely minimized at $\gamma_0$. Therefore, applying Theorem 2.1 of Newey and Mcfadden (1994), we have $\hat{\gamma} \to_p \gamma_0$.□

**Lemma A.2:** $n^{\lambda}\left(\hat{\theta} - \theta\right) = o_p(1)$ and $n^{\lambda}\left(\hat{\delta} - \delta_n\right) = o_p(1)$.

**Proof:** Parallel to the proof of lemma 6 of Hansen (2000), we can apply Lemma A.1, Lemma B.1, Lemma B.12, Assumption 1.9 to prove the results.□

**Lemma A.3:** $n^{1-2\lambda}(\hat{\gamma} - \gamma_0) = O_p(1)$.

**Proof:** Parallel to the proof of lemma 9 of Hansen (2000), applying Lemma A.1, Lemma B.1, Lemma B.12, Assumption 1.9, we can prove the results.□

**Lemma A.4:** $\sqrt{n}\left(\underline{\hat{\theta}} - \underline{\hat{\theta}}(\gamma_0)\right) \to_p 0$, and $\sqrt{n}\left(\underline{\hat{\theta}}(\gamma_0) - \underline{\hat{\theta}}\right) \to_d W \sim N(0, V_\theta)$, with

$$W = \begin{pmatrix} \bar{M} & \bar{M}(\gamma_0) \\ \bar{M}(\gamma_0) & \bar{M}(\gamma_0) \end{pmatrix}^{-1} \begin{pmatrix} J \\ J(\gamma_0) \end{pmatrix}$$

**Proof:** The proof follows through the steps in proving lemma A.12 of Hansen (2000), with the adaption in the definition of $J_n = n^{-1/2}\bar{X}'\bar{e}$ and $J_n(\gamma)$, applying Lemma B.5 and Lemma B.8, Lemma B.9 and Lemma B.12.□

Define $Q_n(v) = S_n\left(\hat{\theta}, \hat{\delta}, \gamma_0\right) - S_n\left(\hat{\theta}, \hat{\delta}, \gamma_0 + v/a_n\right)$

---

http://www.ssc.wisc.edu/~bhansen/progs/ecnmt_00.html

**Lemma A.5:** On any compact set $\Psi$, $Q_n(v) \Rightarrow Q(v) = -c'\bar{D}c\varphi \mid v \mid +2\sqrt{c'\bar{V}c\varphi}B(v)$

**Proof:** Similar to the proof of Lemma A.13, we can show that

$$Q_n(v) = -G_n^*(v) + 2c'R_n(v) + L_n(v),$$

where $G_n^*(\cdot)$ and $R_n(\cdot)$ are defined as in Lemma B.9 and Lemma B.10, and $L_n(v) \Rightarrow 0$ by Lemma B.9, Lemma B.10, Lemma B.12 and Lemma A.4. Applying Lemma B.9 and Lemma B.10 again gives that $Q_n(v) \Rightarrow -c'\bar{D}c\varphi \mid v \mid +2c'\bar{B}(v) = -c'\bar{D}c\varphi \mid v \mid +2\sqrt{c'\bar{V}c\varphi}B(v)$. □

**Proof of Theorem 1:** Note that the proof of Lemma A.3 implies that $n^{1-2\lambda}(\hat{\gamma} - \gamma_0) = \arg\max_v Q_n(v) = O_p(1)$. Since $Q(v)$ is continuous, has a unique maximum, and $\lim_{|v|\to\infty} Q(v) = -\infty$ almost surely. Theorem 2.7 of Kim and Pollard (1990) and Lemma A.5 imply that

$$\begin{aligned}
n^{1-2\lambda}(\hat{\gamma} - \gamma_0) &\to_d \arg\max_v Q(v) \\
&= \arg\max_v \left[-c'\bar{D}c\varphi \mid v \mid +2\sqrt{c'\bar{V}c\varphi}B(v)\right] \\
&= \frac{c'\bar{V}c}{\left(c'\bar{D}c\right)^2\varphi} \arg\max_r \left[-\frac{c'\bar{V}c\varphi}{c'\bar{D}c\varphi} \mid r \mid +2\sqrt{c'\bar{V}c\varphi}B\left(\frac{c'\bar{V}c\varphi}{\left(c'\bar{D}c\right)^2\varphi}r\right)\right] \\
&= \frac{c'\bar{V}c}{\left(c'\bar{D}c\right)^2\varphi} \arg\max_r \left[-\frac{c'\bar{V}c\varphi}{c'\bar{D}c\varphi} \mid r \mid +2\frac{c'\bar{V}c\varphi}{c'\bar{D}c\varphi}B(r)\right] \\
&= \frac{c'\bar{V}c}{\left(c'\bar{D}c\right)^2\varphi} \arg\max_r \left[-\frac{1}{2} \mid r \mid +B(r)\right] = \bar{\eta}\xi,
\end{aligned}$$

where in line 3, we have made change-of-variable $v = \left(\frac{c'\bar{V}c\varphi}{\left(c'\bar{D}c\right)^2\varphi}r\right)$. □

**Proof of Theorem 2:** Note that

$$\begin{aligned}\hat{\sigma}^2(x,z)\,LR_2(\gamma_0) - Q_n(\hat{v}) &= \left[\bar{S}_n\left(\underline{\hat{\theta}}(\gamma_0),\gamma_0\right) - \bar{S}_n\left(\underline{\hat{\theta}},\hat{\gamma}\right)\right] - \left[S_n\left(\underline{\hat{\theta}},\gamma_0\right) - \bar{S}_n\left(\underline{\hat{\theta}},\hat{\gamma}\right)\right] \\ &= \bar{S}_n\left(\underline{\hat{\theta}}(\gamma_0),\gamma_0\right) - S_n\left(\underline{\hat{\theta}},\gamma_0\right) \\ &= \left(\underline{\hat{\theta}}(\gamma_0) - \underline{\hat{\theta}}\right)' \bar{X}_\gamma^{*\prime}\bar{X}_\gamma^*\left(\underline{\hat{\theta}}(\gamma_0) - \underline{\hat{\theta}}\right) \to 0\end{aligned}$$

by Assumption 1.9, Lemma B.1 and Lemma A.4.

Therefore, Lemma A.5 and continuous mapping imply that

$$\begin{aligned}LR_2(\gamma_0) &= \frac{Q_n(\hat{v})}{\hat{\sigma}^2(x,z)} + o_p(1) \\ &\to_d \frac{\sup Q(v)}{\sigma^2(x,z)} \\ &= \frac{1}{\sigma^2(x,z)}\sup_{r\in\mathbb{R}}\left[-c'\bar{D}c\varphi \mid v \mid +2\sqrt{c'\bar{V}c\varphi}B(v)\right] \\ &= \frac{c'\bar{V}c}{\sigma^2\left(c'\bar{D}c\right)}\sup_{r\in\mathbb{R}}\left[-\frac{c'\bar{V}c\varphi}{c'\bar{D}c\varphi} \mid r \mid +2\sqrt{c'\bar{V}c\varphi}B\left(\frac{c'\bar{V}c\varphi}{\left(c'\bar{D}c\right)^2\varphi}r\right)\right] \\ &= \frac{c'\bar{V}c}{\sigma^2\left(c'\bar{D}c\right)}\sup_{r\in\mathbb{R}}\left[- \mid r \mid +2B(r)\right] \\ &= \bar{\tau}\varsigma.\end{aligned}$$

The distribution function $P(\varsigma \le x) = \left(1 - e^{-x/2}\right)^2$ follows from Proposition 2. □

**Proof of Theorem 3:** The proof follows from Lemma A.4. □

**Proof of Theorem 4:** With Proposition 4, we only need show that

$$\hat{g}(z) - \tilde{g}(z) = O_p\left(n^{-1/2}\right) = o_p\left((nh_1...h_p)^{-1/2} + \sum_{s=1}^{p} h_s^2\right).$$

Note that from (8.13) and (8.14), we have

$$\begin{aligned}
\hat{g}(z)-\tilde{g}(z) &= \frac{\sum_{j=1}^{n}\left(y_j-\hat{\theta}'X_j-\hat{\delta}_n'X_j(\hat{\gamma})\right)K_h(z,Z_j)}{\sum_{j=1}^{n}K_h(z,Z_j)} \\
&\quad -\frac{\sum_{j=1}^{n}\left(y_j-\theta'X_j-\delta_n'X_j(\gamma)\right)K_h(z,Z_j)}{\sum_{j=1}^{n}K_h(z,Z_j)} \\
&= \frac{\sum_{j=1}^{n}\left[\left(\theta-\hat{\theta}\right)'X_j+\left(\delta_n'X_j(\gamma)-\hat{\delta}_n'X_j(\hat{\gamma})\right)\right]K_h(z,Z_j)}{\sum_{j=1}^{n}K_h(z,Z_j)} \\
&= \frac{\sum_{j=1}^{n}\left(\theta-\hat{\theta}\right)'X_jK_h(z,Z_j)}{\sum_{j=1}^{n}K_h(z,Z_j)}+\frac{\sum_{j=1}^{n}\left(\delta_n'X_j(\gamma)-\hat{\delta}_n'X_j(\hat{\gamma})\right)K_h(z,Z_j)}{\sum_{j=1}^{n}K_h(z,Z_j)} \\
&\equiv \hat{m}_1(z)+\hat{m}_2(z),
\end{aligned}$$

which indicates that $\hat{m}_1(z)$ and $\hat{m}_2(z)$ are the local constant kernel estimators of $m_1(z)$ and $m_2(z)$, where

$$\left(\theta-\hat{\theta}\right)'X_j=m_1(z_j)+u_j$$

and

$$\delta_n'X_j(\gamma)-\hat{\delta}_n'X_j(\hat{\gamma})=m_2(z_j)+v_j.$$

Therefore, we have, from Li and Racine (2007),

$$\begin{aligned}
\hat{m}_1(z) &= m_1(z)+O_p\left((nh_1...h_p)^{-1/2}+\sum_{s=1}^{p}h_s^2\right) \\
&= E\left[\left(\theta-\hat{\theta}\right)'X_j|z\right]+O_p\left((nh_1...h_p)^{-1/2}+\sum_{s=1}^{p}h_s^2\right) \\
&= O\left(n^{-1/2}\right)+O_p\left((nh_1...h_p)^{-1/2}+\sum_{s=1}^{p}h_s^2\right) \\
&= O_p\left(n^{-1/2}\right)
\end{aligned}$$

and

$$\begin{aligned}
\hat{m}_2(z) &= m_2(z)+O_p\left((nh_1...h_p)^{-1/2}+\sum_{s=1}^{p}h_s^2\right) \\
&= E\left[\delta_n'X_j(\gamma)-\hat{\delta}_n'X_j(\hat{\gamma})|z\right]+O_p\left((nh_1...h_p)^{-1/2}+\sum_{s=1}^{p}h_s^2\right) \\
&= O\left(n^{-\min\{\lambda,1-2\lambda\}}\right)+O_p\left((nh_1...h_p)^{-1/2}+\sum_{s=1}^{p}h_s^2\right)
\end{aligned}$$

where we have used the result of the following,

$$\begin{aligned}
\delta_n' X_j(\gamma) - \hat{\delta}_n' X_j(\hat{\gamma}) &= \delta_n' X_j(\gamma) - \delta_n' X_j(\hat{\gamma}) + \delta_n' X_j(\hat{\gamma}) - \hat{\delta}_n' X_j(\hat{\gamma}) \\
&= \delta_n' \left(X_j(\gamma) - X_j(\hat{\gamma})\right) + \left(\delta_n' - \hat{\delta}_n'\right) X_j(\hat{\gamma}) \\
&= O\left(n^{-\lambda}\right) + O_p\left(n^{2\lambda-1}\right) \\
&= O_p\left(n^{-\min\{\lambda, 1-2\lambda\}}\right)
\end{aligned}$$

Thus we have proved

$$\begin{aligned}
\hat{g}(z) - \tilde{g}(z) &= \hat{m}_1(z) + \hat{m}_2(z) \\
&= O_p\left(n^{-1/2}\right) + O\left(n^{-\min\{\lambda, 1-2\lambda\}}\right) + O_p\left((nh_1...h_p)^{-1/2} + \sum_{s=1}^{p} h_s^2\right) \\
&= O_p\left(n^{-1/2}\right),
\end{aligned}$$

which completes the proof. □

## Appendix B: Some Useful Lemma

**Lemma B.1:** If $\{w_i\}$ is strictly stationary and ergodic, $E \mid \phi(w_i) \mid < \infty$, and $w_i$ has a continuous distribution, then

$$\sup \mid \frac{1}{n} \sum_{i=1}^{n} \phi(w_i) 1(w_i \leq \gamma) - E\left[\phi(w_i) 1(w_i \leq \gamma)\right] \mid \rightarrow_{a.s.} 0$$

**Proof:** This is lemma 1 in Hansen (1996).□

Define $h_i(\gamma_1, \gamma_2) = \mid x_i e_i f_i^2 \mid\mid 1(q_i \leq \gamma_1) - 1(q_i \leq \gamma_2) \mid$ and $k_i(\gamma_1, \gamma_2) = \mid x_i f_i \mid\mid 1(q_i \leq \gamma_1) - 1(q_i \leq \gamma_2) \mid$.

**Lemma B.2:** There is a $C_1 < \infty$ such that for $\underline{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \bar{\gamma}$, and $r \leq 4$,

$$\begin{aligned}
E h_i^r(\gamma_1, \gamma_2) &\leq C_1 \mid \gamma_2 - \gamma_1 \mid, \\
E k_i^r(\gamma_1, \gamma_2) &\leq C_1 \mid \gamma_2 - \gamma_1 \mid.
\end{aligned}$$

**Proof:** The proof follows through the steps in proving lemma A.1 of Hansen (2000), by applying the assumption that $f_i$ is bounded.□

**Lemma B.3:** There is a $C_2 < \infty$ such that for $\underline{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \bar{\gamma}$,

$$E \mid \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ h_i^2 (\gamma_1, \gamma_2) - E h_i^2 (\gamma_1, \gamma_2) \right] \mid^2 \leq C_2 \mid \gamma_2 - \gamma_1 \mid,$$
$$E \mid \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ k_i^2 (\gamma_1, \gamma_2) - E h_i^2 (\gamma_1, \gamma_2) \right] \mid^2 \leq C_2 \mid \gamma_2 - \gamma_1 \mid .$$

**Proof:** The proof follows through the steps in proving lemma A.2 of Hansen (2000), by applying the assumption that $f_i$ is bounded.□

Define

$$J_n (\gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i e_i f_i^2 1 (q_i \leq \gamma) .$$

**Lemma B.4:** There are $K_1$, $K_2 < \infty$ such that for all $\gamma_1$, $\varepsilon > 0$, $\vartheta > 0$, and $\varrho \geq \frac{1}{n}$, if $\sqrt{n} \geq K_2 / \vartheta$, then

$$P \left( \sup_{\gamma_1 \leq \gamma \leq \gamma_1 + \varrho} \mid J_n (\gamma) - J_n (\gamma_1) \mid > \vartheta \right) \leq \frac{K_1}{\vartheta^4}.$$

**Proof:** The proof follows through the steps in proving lemma A.3 of Hansen (2000), by applying the assumption that $f_i$ is bounded.□

**Lemma B.5:** $J_n (\gamma) \Rightarrow J (\gamma)$, a mean-zero Gaussian process with almost surely continuous sample paths, where "$\Rightarrow$" denotes weak convergence with respect to the uniform metric.

**Proof:** This is lemma A.4 of Hansen (2000), with the adaption in the definition of $J_n (\gamma)$.□

**Lemma B.6:** Let $g \in \mathcal{G}_\alpha^v$ and $f \in \mathcal{G}_\alpha^v$, where $v \geq 2$ is an integer. Suppose that $k$ is the $v$-th order kernel, then

$$\text{(i)} \quad \mid EK \left( \frac{Z - z}{h} \right) - h^p f (z) \mid \leq h^{p+v} G_f (z) , \text{ uniformly in } z;$$
$$\text{(ii)} \quad \mid E \left[ g (Z) - g (z) \right] K \left( \frac{Z - z}{h} \right) \mid \leq h^{p+v} G_g (z) , \text{ uniformly in } z,$$

where $G_f (\cdot)$ and $G_g (\cdot)$ have $\alpha$-th finite moments.

**Proof:** This is lemma B.4 of Fan and Li (1999).□

Define $G_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(c'x_i f_i^2\right)^2 \mid 1(q_i \le \gamma) - 1(q_i \le \gamma_0) \mid$ and $K_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} k_i(\gamma_0, \gamma)$. Let $a_n = n^{1-2\lambda}$.

**Lemma B.7:** There exists constants $B > 0$, $0 < d < \infty$, and $0 < k < \infty$, such that for all $\vartheta > 0$, and $\varepsilon > 0$, there exists a $\bar{v} < \infty$ satisfying, for all $n$,

$$P\left(\inf_{\frac{\bar{v}}{a_n} \le |\gamma - \gamma_0| \le B} \frac{G_n(\gamma)}{\mid \gamma - \gamma_0 \mid} < (1-\vartheta)\, d\right) \le \varepsilon$$

and

$$P\left(\sup_{\frac{\bar{v}}{a_n} \le |\gamma - \gamma_0| \le B} \frac{K_n(\gamma)}{\mid \gamma - \gamma_0 \mid} < (1+\vartheta)\, d\right) \le \varepsilon.$$

**Proof:** The proof follows through the steps in proving lemma A.7 of Hansen (2000), with the adaption in the definition of $G_n(\gamma)$ and $K_n(\gamma)$.□

**Lemma B.8:** For $\vartheta > 0$ and $\varepsilon > 0$, there exists some $\bar{v} < \infty$ such that for any constants $B < \infty$,

$$P\left(\sup_{\frac{\bar{v}}{a_n} \le |\gamma - \gamma_0| \le B} \frac{\mid J_n(\gamma) - J_n(\gamma_0) \mid}{\mid \gamma - \gamma_0 \mid} > \vartheta\right) \le \varepsilon.$$

**Proof:** The proof follows through the steps in proving lemma A.8 of Hansen (2000), with the adaption in the definition of $J_n(\gamma)$.□

Define $G_n^*(v) = a_n G_n(\gamma_0 + v/a_n)$ and $K_n^*(v) = a_n K_n(\gamma_0 + v/a_n)$

**Lemma B.9:** Uniformly on compact sets $\Psi$,

$$G_n^*(v) \to_p c'\bar{D}c\varphi$$

and

$$K_n^*(v) \to_p \mid \bar{D}\varphi v \mid$$

**Proof:** The proof follows through the steps in proving lemma A.10 of Hansen (2000), with the adaption in the definition of $G_n^*(v)$ and $K_n^*(v)$, applying Lemma B.3 and (the proof of) Lemma B.7.□

Define $R_n(v) = \sqrt{a_n}\left(J_n(\gamma_0 + v/a_n) - J_n(\gamma_0)\right)$

**Lemma B.10:** On any compact sets $\Psi$,

$$R_n(v) \Rightarrow \bar{B}(v)$$

where $\bar{B}(v)$ is a vector Brownian motion with covariance matrix $E\left(\bar{B}(1)\bar{B}(1)'\right) = \bar{V}\varphi$

**Proof:** The proof follows through the steps in proving lemma A.11 of Hansen (2000), with the adaption in the definition of $R_n(v)$, applying Lemma B.4, Lemma B.5, Theorem 24.3 of Davidson (CLT for Martingale Difference Sequence), Markov's inequality and Theorem 16.1 of Billingsley (1968).□

**Lemma B.11:** Assuming that $\{Z_i\}$ satisfies Assumption 1.1, the density function $f(\cdot)$ satisfies Assumption 1.2, the univariate kernel function, $k(\cdot)$ satisfies Assumption 2.1 and bandwidth $h_s$ satisfies Assumption 2.2, we have

$$E\left\{\left[\hat{f}(z) - f(z)\right]^2\right\} = O\left(\left(\sum_{s=1}^{p} h_s^2\right)^2 + (nh_1...h_p)^{-1}\right)$$

**Proof:** This is Theorem 18.1 of Li and Racine (2007).□

**Lemma B.12:** $\hat{f}(z) - f(z) = o_p(1)$

**Proof:** The proof follows directly from applying Theorem A.7 of Li and Racine (2007) and Lemma B.11.□

## References


Adams, R.M., Berger, A.N. and Sickles, R.C. (1999), "Semiparametric approaches to stochastic panel frontiers with applications in the banking industry," *Journal of Business and Economics Statistics* 17, 349-358

Andrews, D.W.K. and Ploberger, V. (1994), "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica* 62, 1383–1414

Anglin, P. and Gencay, R. (1996), "Semiparametric estimation of a hedonic price function," *Journal of Applied Econometrics* 11, 633-648

Bai, J. (1997), "Estimation of a change point in multiple regression models," *Review of Economics and Statistics* 79, 551–563

Barnett, S.A. and Plutarchos, S. (1998), "Nonlinear response of firm investment to q: Testing a model of convex and non-convex adjustment costs," *Journal of Monetary Economics* 42, 261–188

Billinsley, P. (1968), *Convergence of Probability Measures* Wiley Press, New York.

Blundell, R., Duncan, A., and Pendakur, K. (1998), "Semiparametric estimation and consumer demand," *Journal of Applied econometrics* 13, 435-461

Caner, M. (2002), "A note on LAD estimation of a threshold model," *Econometric Theory* 18, 800–814

Caner, M. and Hansen B.E. (2004), "Instrumental variable estimation of a threshold model," Econometric Theory 20, 813-843

Chan, K.S. (1993), "Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model," *Annals of Statistics* 21, 520–533

Davidson, J. (1994), *Stochastic Limit Theory: An Introduction for Econometricians* Oxford University Press, Oxford.

Davies, R.B. (1977), "Hypothesis testing when a nuisance parameter is present only under the alternative," *Biometrika* 64, 247–254

Davies, R.B. (1987), "Hypothesis testing when a nuisance parameter is present only under the alternative," *Biometrika* 74, 33–43

Engle, R.F., Granger, C.W.J., Rice, J. and Weiss, A. (1986), "Semiparametric estimates of the relation between weather and electricity demand," *Journal of the American Statistical Association* 76, 817-823

Erickson, T. and Whited, T.M. (2000), "Measurement error and the relationship between investment and q," *Journal of Political Economy* 108, 1027–1057

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, Chapman & Hall, London.

Fan, Y. and Li, Q. (1999), "Root-n-consistent estimation of partially linear time series models," *Journal of Nonparametric Statistics* 11, 251-269

Farrell, L., Morgenroth, E. and Walker, I. (1999), "A time series analysis of U.K. lottery sales: long and short run price elasticities," *Oxford Bulletin of Economics and Statistics 61*, 513–626

Fazzari, S.M., Hubbard, R.G. and B.C. Petersen (1988), "Financing constraints and corporate investment," *Brookings Papers on Economic Activity* 141–195

Gonzalo, J. and Pitarakis, J.Y. (2002), "Estimation and Model Selection Based Inference in Single and Multiple Threshold Models," *Journal of Econometrics* 110,

319-352.

Hall, P. and Heyde, C.C. (1980), *Martingale Limit Theory and Its Application.* Academic Press, New York.

Hansen, B.E. (1996), "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica* 64, 413– 430

Hansen, B.E. (1999), "Threshold effects in non-dynamic panels: Estimation, testing, and inference," *Journal of Econometrics* 93, 345–368

Hansen, B.E. (2000), "Sample splitting and threshold estimation," *Econometrica* 68, 575– 603

Hardle, W., Liang, H. and Gao, J., (2000), *Partial linear models*, Heidelberg: Physica-Verlag.

Kim, J. and Pollard, D. (1990), "Cube root asymptotics," *Annals of Statistics* 18, 191–219

Li, Q. (1996), "On the root-n-consistent semiparametric estimation of partially linear models," Economic Letters 51, 277-285

Li, Q. and Racine, J.S. (2007), *Nonparametrics Econometrics: Theory and Practice* Princeton University Press, Princeton and Oxford

Li, Q. and Stengos, T. (1996), "Semiparametric estimation of partially linear panel data models," *Journal of Econometrics* 71, 389-397

Matzkin, R.L. (2007), "Nonparametric identification," in *Handbook of Econometrics* 6B, Edited by: James J. Heckman and Edward E. Leamer, 5307-5368

Pagan, A. and Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge University Press, Cambridge.

Picard, D. (1985), "Testing and estimating change-points in time series," *Advances in Applied Probability* 17, 841–867

Robinson, P.M. (1983), "Nonparametric estimators for time series," *Journal of Time Series Analysis* 4, 185-297

Robinson, P.M. (1983), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931-954

Seo, M.H. and Linton, O. (2007), "A smoothed least squares estimator for threshold regression models," *Journal of Econometrics* 141, 704-735

Stock, J.H. (1989), "Nonparametric policy analysis," *Journal of the American Statistical Association* 84, 567-575

Yatchew, A.J. and No, J.A. (2001), "Household gasoline demand in Canada," *Econometrica* 69, 1697-1709

Table 8.1: Coverage Probability for nominal 90% confidence interval for threshold parameter

| $\delta_n$ | $n$=50 | $n$=100 | $n$=200 | $n$=400 |
|---|---|---|---|---|
| 0.25 | 0.93 | 0.96 | 0.92 | 0.94 |
| 0.50 | 0.93 | 0.97 | 0.92 | 0.96 |
| 1.00 | 0.90 | 0.92 | 0.96 | 0.95 |
| 2.00 | 0.93 | 0.94 | 0.96 | 0.98 |

Table 8.2: RMSE: Partial Linear Model

| $\delta_n$ | $n$=50 | $n$=100 | $n$=200 | $n$=400 |
|---|---|---|---|---|
| 0.25 | 176.4 | 91.7 | 31.5 | -18.7 |
| 0.50 | 63.7 | -5.9 | -43.5 | -69.2 |
| 1.00 | -19.2 | -48.9 | -77.8 | -89.3 |
| 2.00 | -44.7 | -67.1 | -85.0 | -95.0 |

Table 8.3: RMSE: Nonparametric Model

| $\delta_n$ | $n$=50 | $n$=100 | $n$=200 | $n$=400 |
|---|---|---|---|---|
| 0.25 | 76.1 | 11.9 | 2.5 | -16.7 |
| 0.50 | 68.5 | 23.1 | -13.9 | -46.9 |
| 1.00 | 24.3 | -23.4 | -67.2 | -83.7 |
| 2.00 | -17.7 | -52.7 | -79.6 | -93.3 |

Table 8.4: Threshold Estimation: Fuel Engel Curve

| | Regime 1: Tot Inc≤350.0 | Regime 2: Tot Inc>350.0 |
|---|---|---|
| Obs | 1508 | 11 |
| DoF | 1507 | 10 |
| SSE | 0.005239 | 0.000072 |
| Res Var | 0.000003 | 0.000007 |
| R-squared | 0.148232 | 0.023394 |

Table 8.5: Slope Parameter Estimates for Regime 1

| Variable | Estimate | St Error |
|---|---|---|
| Log Totexp | -0.050514 | 0.003983 |

Table 8.6: Slope Parameter Estimates for Regime 2

| Variable | Estimate | St Error |
|---|---|---|
| Log Totexp | -0.005386 | 0.027016 |

Figure 8.1: F test for threshold: Food Engel Curve

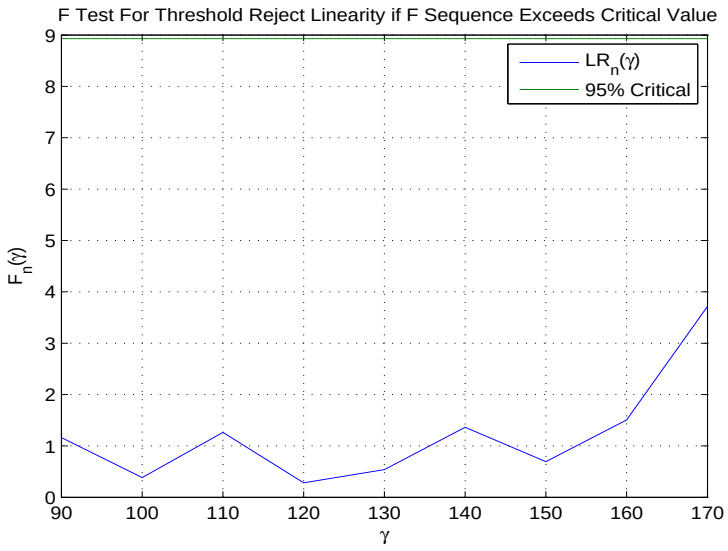


Figure 8.2: F test for threshold: Fuel Engel Curve

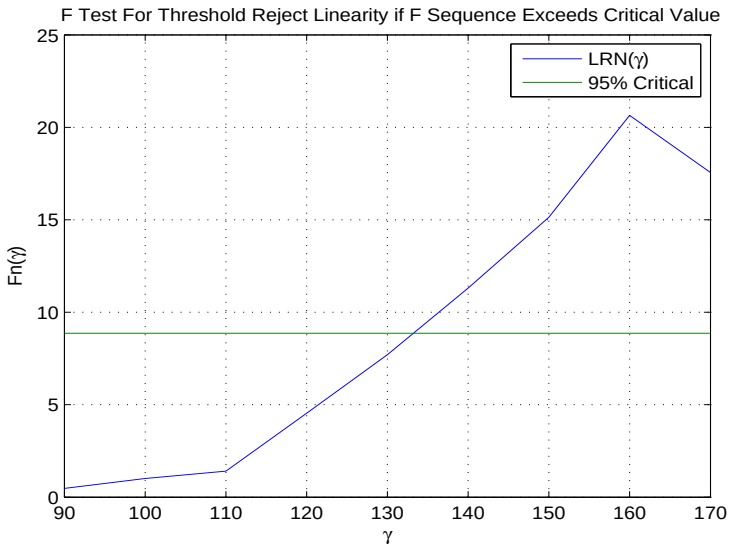



Figure 8.3: Threshold estimate: Fuel Engel Curve

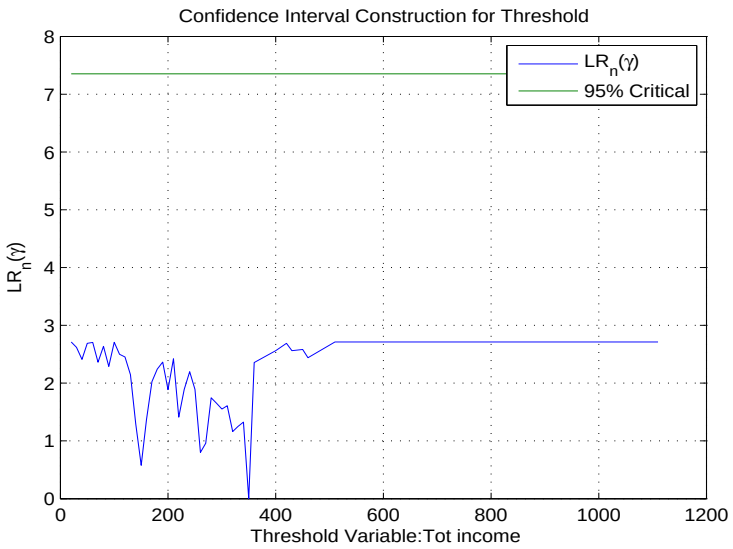


Figure 8.4: F test for threshold: Clothing Engel Curve

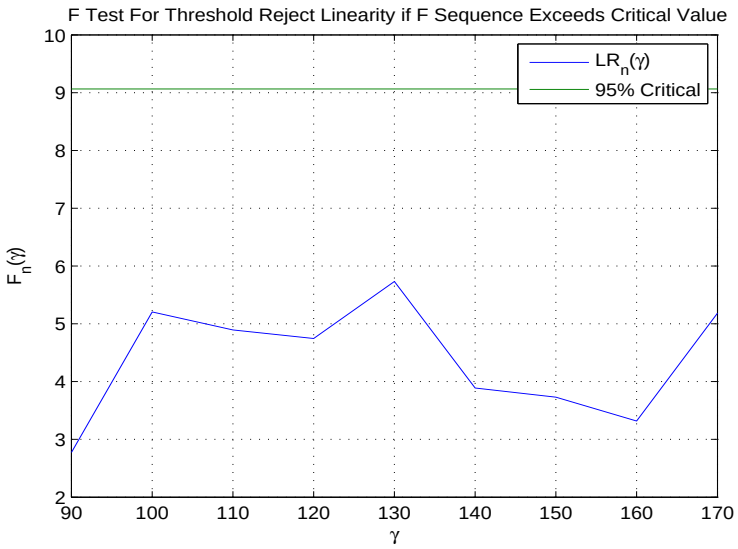



Figure 8.5: F test for threshold: Alcohol Engel Curve

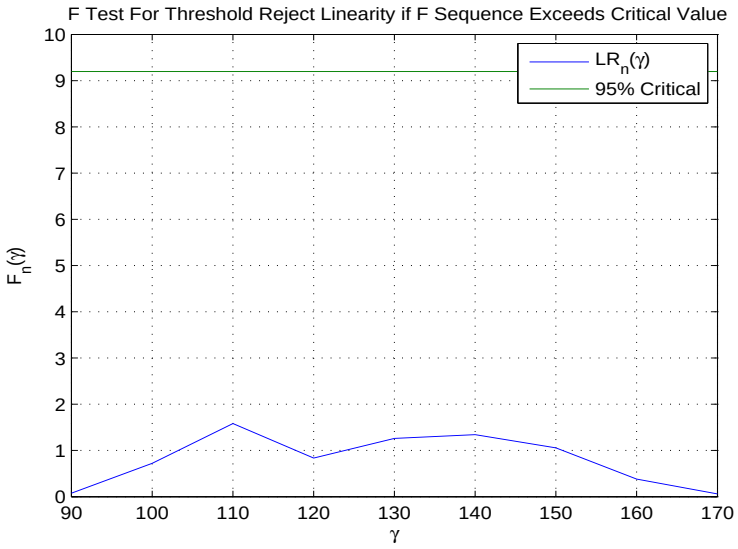


Figure 8.6: F test for threshold: Transport Engel Curve

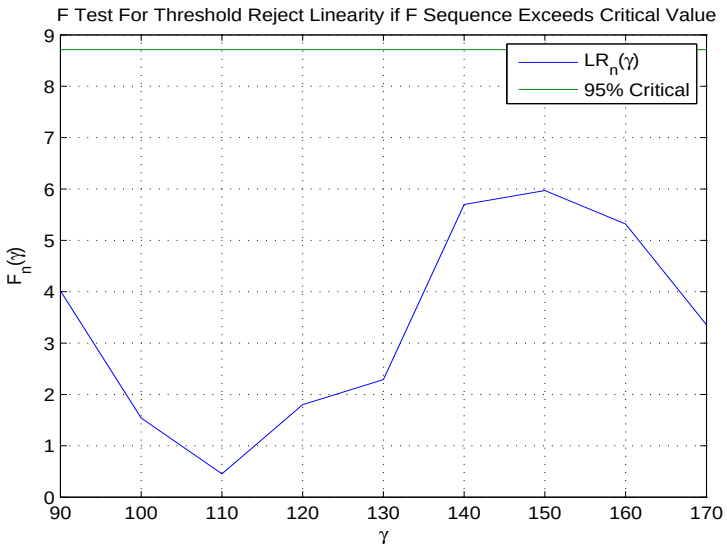
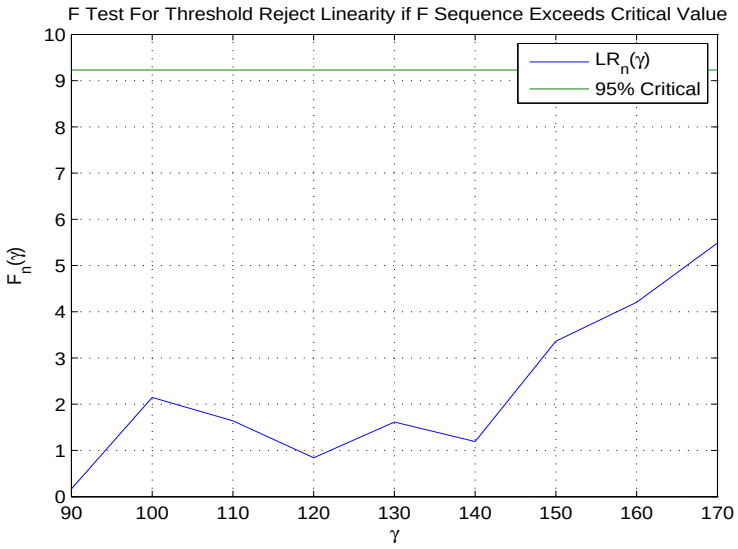



Figure 8.7: F test for threshold: Other Good Engel Curve



# Chapter 9

# Conclusions and Future Work

In this thesis, we investigated how information should be analyzed in economic studies. In particular, we studied how to incorporate economic constraints (monotonicity, positivity) into structure estimation and forecasting. And we studied the properties of supervised factor models that help to separate noises from useful information content contained in large dimensional data. We proposed estimators that efficiently use the additive structure in the simultaneous equation models and tests for the validity of such kind additive separable error structure. We extended the semiparametric partial linear model into semiparametric partial threshold model.

In addition, we made use of the entropy literature and extended the related concepts to account for model uncertainty. We proposed the entropy-based model averaging estimators to evaluate the partial effect parameters that is of economic importance, in large dimensional data set.

There are various directions that the above work could be extended and further developed.

First, nonparametric estimation under restrictions with the use of bagging can be extended for other type of constraints, including concavity/convexity, homogeneity, unimodality, symmetry, etc.

Second, bagging as a way to smooth indicator functions can be applied to other shrinkage type of estimators, including Stein-rule estimator, LASSO type estimator, ridge estimator, etc. The theoretical properties are more challenging to investigate. The extension of bagging in time series and panel data framework is also of theoretical and practical importance.

Third, entropy as a measure of information content can be more widely used in economics research. The proposed entropy-based model averaging can be extended to nonparametric models and semiparametric models. Also, it can used for shrinkage-type estimation for partial effects of interest.

Fourth, entropy-based variable selection and estimation is not developed and deserves our attention. A unified framework that can simultaneously perform variable selection and parameter estimation is attractive to investigate.