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

Tae-Hwy Lee†
Department of Economics
University of California, Riverside

Yundong Tu‡
Guanghua School of Management and
Center for Statistical Science
Peking University

Aman Ullah§
Department of Economics
University of California, Riverside

First Version: September 2011
This Version: September 2012

Abstract

This paper 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 for illustration. We introduce a new forecasting evaluation criterion based on the second order stochastic dominance in the size of forecast errors and compare models over different sizes of forecast errors. Imposing monotonicity constraint can mitigate the chance of making large size forecast errors.

Key Words: Local monotonicity, Bagging, Asymptotic mean squared errors, Second order stochastic dominance, Equity premium prediction.

JEL Classification: C14; C50; C53; G17.

---

*The authors would like to thank three anonymous referees and seminar participants at California Econometrics Conference at Stanford University, Midwest Econometrics Group at Washington University St. Louis, Conference in Honor of Halbert White at UC San Diego, Asian Meeting of the Econometric Society at Korea University, Singapore Management University, Office of the Comptroller of the Currency of US Department of Treasury, UCR, USC, Chinese Academy of Sciences, Shanghai University of Finance and Economics, and University of Tasmania, for valuable suggestions.

†Department of Economics, University of California, Riverside, CA 92521. E-mail: taelee@ucr.edu.

‡Corresponding author. Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, China 100871. Phone: +86 10 62760219. E-mail: yundong.tu@gsm.pku.edu.cn.

§Department of Economics, University of California, Riverside, CA 92521. E-mail: aman.ullah@ucr.edu.

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 paper 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 paper is to consider nonlinear models subject to local monotonicity constraint and their bagging versions.

Nonparametric estimation with constraints has long history that dates back to the work of Brunk (1955). Classical references on estimation under restriction include Barlow et al (1972), Ramsay (1988), Mammen (1991), Matzkin (1994) and Chen (2007), to name a few. 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 Du et al (2013) 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 a 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 a simulation study to demonstrate how the asymptotic results work in finite samples. 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 paper 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 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 paper would follow when x_t is a vector, except that such extensions would raise issues such as the curse of dimensionality or what notion of monotonicity to impose that deserve further effort to explore. In forecasting, the s-step ahead forecast of y_{n+s} at time $t = n$ given that $x_n = x$ is defined as

$$m_{n,s}(x) = E(y_{n+s}|x_n = x). \tag{1}$$

Sometimes a priori constraint 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 $m_{n,s}(x)$ is positive, monotonic, homogeneous, homothetic, and etc. To this end, estimators or forecasts may be subject to constraints. In this paper, 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.

2.1 Parametric Estimation with Constraints

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

$$m_{n,s}(x) = \alpha + \beta x \tag{2}$$

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

$$\tilde{\alpha} = \bar{y}_n - \tilde{\beta}\bar{x}_n \tag{3}$$

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 β through thresholding using an indicator function as done by Campbell and Thompson (2008),

$$\begin{aligned} \bar{\beta} &= 1_{\{\tilde{\beta}>0\}} \cdot \tilde{\beta}, \\ \bar{\alpha} &= 1_{\{\tilde{\beta}>0\}} \cdot \tilde{\alpha} + 1_{\{\tilde{\beta}\leq 0\}} \cdot \bar{y}_n. \end{aligned} \tag{4}$$

Note that the relationship between $\bar{\alpha}$ and $\bar{\beta}$ remains as in (3)

$$\bar{\alpha} = \bar{y}_n - \bar{\beta}\bar{x}_n, \tag{5}$$

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

$$\hat{\beta} = \frac{1}{J}\sum_{j=1}^{J} \bar{\beta}^{*(j)} \equiv E^* \bar{\beta}^*, \tag{6}$$

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

$$\hat{\alpha} \equiv \bar{y}_n - \hat{\beta}\bar{x}_n = E^* \bar{\alpha}^*. \tag{7}$$

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 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 Du et al (2013) and Henderson and Parmeter (2009). Alternatively, we propose here to use bagging to impose constraints in nonparametric kernel regression models.

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

$$\hat{m}_{n,s}(x) = \frac{1}{n-s}\sum_{t=1}^{n-s} A_t(x) y_{t+s}, \tag{8}$$

where $A_t(x)$ is the weighting function. For example, $A_t(x) = k\left(\frac{x_t - x}{h}\right) / \sum_{t=1}^{n-s} k\left(\frac{x_t - x}{h}\right)$ for the Nadaraya-Watson estimator. Hall and Huang (2001) suggested an estimator

$$\hat{m}_{n,s}(x|\mathbf{p}) = \sum_{t=1}^{n-s} p_t A_t(x) y_{t+s}, \tag{9}$$

where $\mathbf{p} = (p_1, \ldots, p_{n-s})'$. Note that (8) is a special case of (9) with the uniform weights $\mathbf{p}_0 = (\frac{1}{n-s}, \ldots, \frac{1}{n-s})'$. $\mathbf{p}$ is to be estimated by $\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} D(\mathbf{p})$ subject to the constraints and additional conditions such as $\sum_{t=1}^{n-s} p_t = 1$ and $\mathbf{p} \geq \mathbf{0}$, 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. In the case of monotonicity, the constraint is $\partial \hat{m}_{n,s}(x|\mathbf{p})/\partial x > 0$.

2.2.2 Nonparametric Estimation with Constraints: Bagging

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

$$\begin{aligned} y_t &= m(x_t) + u_t = m(x) + (x_t - x)m^{(1)}(x) + v_t \\ &= \alpha(x) + x_t\beta(x) + v_t = X_t\delta(x) + v_t \end{aligned} \tag{10}$$

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

$$\sum_{t=1}^{n} v_t^2 K_h(x_t, x) = \sum_{t=1}^{n} (y_t - X_t\delta(x))^2 K_h(x_t, x), \tag{11}$$

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

$$\tilde{\delta}(x) = (\mathbf{X}'\mathbf{K}(x)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(x)\mathbf{y}, \tag{12}$$

where $\mathbf{X}$ is an $n \times (k+1)$ matrix with the tth 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 $\bar{\beta}(x)$

$$\bar{\beta}(x) = 1_{\left\{\tilde{\beta}(x) > 0\right\}} \cdot \tilde{\beta}(x), \tag{13}$$

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

$$\hat{\beta}(x) = \frac{1}{J}\sum_{j=1}^{J} \bar{\beta}(x)^{*(j)} = E^* \bar{\beta}(x)^*. \tag{14}$$

Observing (3) and (5), the unconstrained LLLS estimator is

$$\tilde{\alpha}(x) = \bar{y}(x) - \tilde{\beta}(x)\bar{x}(x), \tag{15}$$

where

$$\begin{aligned} \bar{y}(x) &= \frac{\sum_{t=1}^{n} K_h(x_t, x) y_t}{\sum_{t=1}^{n} K_h(x_t, x)} = (\mathbf{i}'\mathbf{K}(x)\mathbf{i})^{-1}\mathbf{i}'\mathbf{K}(x)\mathbf{y}, \\ \bar{x}(x) &= \frac{\sum_{t=1}^{n} K_h(x_t, x) x_t}{\sum_{t=1}^{n} K_h(x_t, x)} = (\mathbf{i}'\mathbf{K}(x)\mathbf{i})^{-1}\mathbf{i}'\mathbf{K}(x)\mathbf{x}, \end{aligned} \tag{16}$$

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

$$\bar{\alpha}(x) = \bar{y}(x) - \bar{\beta}(x)\bar{x}(x), \tag{17}$$

$$\hat{\alpha}(x) = \bar{y}(x) - \hat{\beta}(x)\bar{x}(x), \tag{18}$$

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 (10) we have the unconstrained NP forecast,

$$\begin{aligned} \tilde{m}(x) &= \tilde{\alpha}(x) + x\tilde{\beta}(x) = \bar{y}(x) - \tilde{\beta}(x)\bar{x}(x) + x\tilde{\beta}(x) \\ &= \bar{y}(x) - \tilde{\beta}(x)\left[\bar{x}(x) - x\right]. \end{aligned} \tag{19}$$

Similarly, we get the constrained NP forecast

$$\bar{m}(x) = \bar{y}(x) - \bar{\beta}(x)\left[\bar{x}(x) - x\right], \tag{20}$$

and the bagged constrained NP forecast

$$\hat{m}(x) = \bar{y}(x) - \hat{\beta}(x)\left[\bar{x}(x) - x\right]. \tag{21}$$

2.3 Semiparametric Estimation with Constraints

Let us consider the model

$$\begin{aligned} y &= \alpha + \beta x + u \\ &= \alpha + \beta x + E(u|x) + [u - E(u|x)] \\ &= \alpha + \beta x + E(u|x) + v \\ &= m(x) + v \end{aligned} \tag{22}$$

where $m(x) = \alpha + \beta x + E(u|x)$, $E(u|x) \neq 0$,and $v = u - E(u|x)$ such that $E(v|x) = 0$. In model (22) the linear component in many cases plays the guiding role, like the benchmark linear model in the forecasting of equity premium (Section 7), while the nonparametric component of x, $E(u|x)$, behaves like a type of unknown departure or correction for the misspecified linear model. Since such departure is unknown, it is not unreasonable to treat $E(u|x)$ as a nonparametrically unknown function, and the model $m(x)$ in (22) as semiparametric. In recent literature, Glad (1998) and Martins-Filho et al (2008) have discussed the issue of reducing estimation biases through using a potentially misspecified parametric form in the first step rather than simply nonparametrically estimating the conditional mean function $m(x) = E(y|x)$. The function of interest, $m(x)$, is then estimated by a two step procedure. This two step estimator of $m(x)$ is consistent and asymptotically normal, see Martins-Filho et al (2008). In the first step α and β are obtained by the OLS estimation, and the second step involves an LLLS estimation of $E(u|x)$ by using 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 the NP regression. Thus the LLLS estimator can be represented by $\tilde{\xi}(x) - \tilde{\eta}(x)(\bar{x}(x) - x)$. This two-step algorithm leads to an unconstrained SP estimator of $m(\cdot)$ as

$$\begin{aligned} \tilde{m}_{sp}(x) &= \tilde{\alpha} + \tilde{\beta}x + \tilde{\xi}(x) - \tilde{\eta}(x)(\bar{x}(x) - x) \\ &= \tilde{\alpha} + \tilde{\xi}(x) - \tilde{\eta}(x)\bar{x}(x) + \left\{\tilde{\beta} + \tilde{\eta}(x)\right\} x, \end{aligned} \tag{23}$$

the slope of which is estimated by

$$\tilde{\beta}(x) \equiv \tilde{\beta} + \tilde{\eta}(x). \tag{24}$$

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

$$\bar{\beta}(x) = 1_{\left\{\tilde{\beta}(x) > 0\right\}} \cdot \tilde{\beta}(x) \tag{25}$$

When $\tilde{\beta}(x) \leq 0$, the slope of the regression function is zero, i.e., $\bar{\beta}(x) = 0$. 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

$$\bar{m}_{sp}(x) = \tilde{m}_{sp}(x) \cdot 1_{\left\{\tilde{\beta}(x) > 0\right\}} + \tilde{m}_{lc}(x) \cdot 1_{\left\{\tilde{\beta}(x) \leq 0\right\}}, \tag{26}$$

where $\tilde{m}_{lc}(x) = \bar{y}(x)$ is the local constant kernel estimator of $m(x)$ as in (16).

With having $\bar{m}_{sp}(x)$ obtained, similarly to (6), we get the bagging constrained SP forecast from

$$\hat{m}_{sp}(x) = \frac{1}{J}\sum_{j=1}^{J} \bar{m}_{sp}^{*(j)}(x) = E^{*}\bar{m}_{sp}^{*}(x), \tag{27}$$

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

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.

3.1 Constrained Parametric Estimator

Theorem 1. Let the unconstrained estimator $\tilde{\beta}$ of β have a cumulative distribution function (CDF) 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 β_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 β_1 to β_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.

Lovell and Prescott (1970) and Judge and Yancey (1986) consider the case in which $\tilde{\beta}$ has a normal distribution. Judge and Yancey (1986, 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 δ, 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 β with

$$\begin{aligned} &\gamma(n)\,\sigma_{\beta}^{-1}\left(\tilde{\beta}-\beta\right) \xrightarrow{d} Z \\ &\gamma(n)\,\sigma_{m}^{-1}\left(\tilde{m}(x)-m(x)\right) \xrightarrow{d} Z \end{aligned} \tag{28}$$

and Z is a random variable with CDF $F(\cdot)$, where σ_β is the asymptotic standard deviation of $\tilde{\beta}$ and σ_m is that of $\tilde{m}(x)$, and $\lim_{n\to\infty}\gamma(n)=\infty$. Then for $\bar{\beta}=\max\{\tilde{\beta},\beta_1\}$ and some given constant β_1, we have,

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

If we further assume that

$$\gamma(n)\,\sigma_{\beta}^{-1}\left(\beta-\beta_1\right)=b, \tag{29}$$

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

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

Remark 2(a). Theorem 2 states 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 β is on the boundary, the limiting CDF compresses all the mass of negative values at 0. Part 4 establishes the local asymptotic distribution of $\bar{\beta}$ that depends on the drift parameter b with asymptotic bias and variance given in part 5 and 6. 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 4 becomes that in part 1. Similarly, part 2 is reproduced with part 4 when $b = 0$. Part 3 presents the limiting distribution of $\bar{m}(x)$, the constrained estimator of $m(x) = E(y|x) = \alpha + \beta x$. The local asymptotic result for $\bar{m}(x)$ (and other estimators of $m(x)$ in the following sections) with local drift parameter b is complicated to establish and requires further conditions. We did not explore this in this paper.

Remark 2(b). Theorem 2 only requires $\tilde{\beta}$ satisfy some limiting theorem with asymptotic standard deviation σ_β. This is a very weak 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 β is a Pitman type drift to the specified bound β_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. □

3.2 Bagged Constrained Parametric Estimator

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

$$\begin{aligned} \gamma(n)\sigma_\beta^{-1}\left(\tilde{\beta} - \beta\right) &\xrightarrow{d} Z, \\ \gamma(n)\sigma_\beta^{-1}\left(\tilde{\beta}^* - \tilde{\beta}\right) &\xrightarrow{d} Z, \end{aligned} \tag{30}$$

with Z being a standard normal random variable, where σ_β 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 β_1 satisfies

$$\gamma(n)\sigma_\beta^{-1}(\beta - \beta_1) = b, \tag{31}$$

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

4. If $\gamma(n)\sigma_m^{-1}(\tilde{m}(x)-m(x)) \xrightarrow{d} Z$ and $\beta > \beta_1$, then $\gamma(n)\sigma_m^{-1}(\bar{m}(x)-m(x)) \xrightarrow{d} Z$.

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 β_1, the lower bound of β, 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). (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). It is this bootstrap failure for $\bar{\beta}$ that leads to Theorem 3. 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. □

Figure 1 About Here

4 Sampling Properties of Nonparametric Estimators

We consider sampling properties of NP estimators under constraint and their bagging versions.

4.1 Constrained Nonparametric Estimator

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

$$\begin{aligned} &\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}(x)-\beta(x)\right) \xrightarrow{d} Z, \\ &\gamma_2(n,h)\,\sigma_m^{-1}(x)\left(\tilde{m}(x)-m(x)-B_m(x)\right) \xrightarrow{d} Z, \end{aligned} \tag{32}$$

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 $\tilde{m}(x)$, $B_m(x) = \frac{1}{2}h^2 m^{(2)}(x)\int v^2 k(v)\,dv + o_p(h^2)$ 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 $\beta(x) > \beta_1(x)$, $\gamma_2(n,h)\,\sigma_m^{-1}(x)\,[\bar{m}(x) - m(x) - B_m(x)] \xrightarrow{d} Z$.

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

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

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. However, when the constraint is invalid, $\bar{\beta}(x)$ will be inconsistent.
Remark 4(b). The constrained estimator of $m(x)$, $\bar{m}(x)$, has the same asymptotic property as the unconstrained nonparametric estimator, when the constraint is strict, as established in part 3 of Theorem 4. This first order equivalence agrees with that of the estimators of Mammen (1991). 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. However, when the constraint is invalid, the constraint estimator is generally inconsistent[1]. Thus, a test based on the difference between the constrained estimator and unconstrained estimator could be developed to check the validity of the constraint. Other distribution-free tests could also be applied for this purpose, see, e.g., Lee, Linton and Whang (2009), and Delgado and Escanciano (2012).

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,

$$\begin{aligned} &\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}(x) - \beta(x)\right) \xrightarrow{d} Z, \\ &\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}^*(x) - \tilde{\beta}(x)\right) \xrightarrow{d} Z, \end{aligned} \tag{33}$$

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 $\tilde{\beta}(x)$.

[1]One exception is when both the constraint is invalid and $\beta_1(x) = 0$. In this case, the constrained estimator is the Nadaraya-Watson estimator that remains consistent.

Define $\bar{\beta}(x) = \max\left\{\tilde{\beta}(x), \beta_1(x)\right\}$, with some known $\beta_1(x) < \beta(x)$ that satisfies

$$\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\,(\beta(x) - \beta_1(x)) = b(x), \tag{34}$$

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

$$\gamma_2(n,h)\,\sigma_m^{-1}(x)\,[\hat{m}(x) - m(x) - B_m(x)] \xrightarrow{d} Z. \tag{35}$$

Remark 5(a). Condition (33) requires that bootstrap works for the slope estimator $\tilde{\beta}(x)$. See Hall (1992) or Horowitz (2001) for validity of bootstrap for local nonparametric estimators. Note that this condition may rule out some range of bandwidths, which is an important issue that deserves separate studies. For this paper, we consider the use of cross-validation to select the optimal bandwidth for the unconstrained estimator and use that same bandwidth for the bagged estimator. The choice of the bandwidth for bagging estimator is left for future research.
Remark 5(b). 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.

□

5 Sampling Properties of Semiparametric Estimators

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

5.1 Constrained Semiparametric Estimator

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

$$\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}(x)-\beta(x)\right)\xrightarrow{d}Z, \tag{36}$$

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 $\beta(x)>\beta_1(x)$, the semiparametric estimator has

$$\gamma_2(n,h)\,\sigma_m^{-1}(x)\left[\bar{m}_{sp}(x)-m(x)-B_m(x)\right]\xrightarrow{d}Z, \tag{37}$$

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

$$B_m(x)=\frac{1}{2}h^2m^{(2)}(x)\int v^2k(v)\,dv+o_p\left(h^2\right), \tag{38}$$

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

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. $\lim_{n\to\infty}\gamma_1(n,h)\,\sigma_\beta^{-1}\left[\bar{\beta}(x)-\beta(x)\right]=Z_{b(x)}1_{\left[Z_{b(x)}>0\right]}-b(x)$.
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 result shows that the estimation of $m(x)$ via the SP method is a consistent estimator of the true function $m(x)$, the same property that is possessed by the NP estimator but not by the parametric estimator under misspecification. 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 (2008). Part 4 considers the local asymptotics for the constrained slope estimator, with asymptotic bias and variance given in Parts 5 and 6.

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,

$$\begin{aligned} \gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}(x)-\beta(x)\right) &\xrightarrow{d} Z, \\ \gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\tilde{\beta}^*(x)-\tilde{\beta}(x)\right) &\xrightarrow{d} Z, \end{aligned} \tag{39}$$

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

$$\gamma_1(n,h)\,\sigma_\beta^{-1}(x)\left(\beta(x)-\beta_1(x)\right)=b(x), \tag{40}$$

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

$$B_m(x)=\frac{1}{2}h^2m^{(2)}(x)\int v^2k(v)\,dv+o_p\left(h^2\right) \tag{41}$$

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

$$\gamma_2(n,h)\,\sigma_m^{-1}(x)\left[\hat{m}_{sp}(x)-m(x)-B_m(x)\right]\xrightarrow{d} Z. \tag{42}$$

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.

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

$$\text{DGP}: \; y_{t+1} = ax_t^3 + e_{t+1}, \tag{43}$$

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

$$\text{Model}: \; y_{t+s} = m(x_t) + u_{t+s}. \tag{44}$$

For a given evaluation predictor value x, we are interested in forming a forecast $\hat{y}_{n+s} = m_{n,s}(x|I_n)$, 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, $s = 1$ 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 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 $P = 1000$. The simulation takes 90 minutes to run on a quad-core laptop. The computer codes are available on the authors' websites.

In the empirical application of Section 7, $s = 1, 6, 12$, 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 s months ahead forecasts of the equity premium y_{n+s}, with n moving forward from 1960M1 to 2005M12, resulting in the total of $P = (552 - s)$ evaluation points over the 46 years. Once $\hat{y}_{n+s}$ is obtained, we define the forecast error $\hat{u}_{n+s} \equiv y_{n+s} - \hat{y}_{n+s}$. We use the $(552 - s)$ squared forecast errors $\left\{\hat{u}_{n+s}^2\right\}_{n=1960M1}^{2005M12-s}$ 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 $P = 552 - s$.

6.1 Forecasting Models

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

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

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

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

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)

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

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

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

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

$$\begin{aligned} m_{n,s}^{\text{SP}}(x|I_n) &= \tilde{m}_{sp}(x) \text{ as defined in (23)}, \\ m_{n,s}^{\text{SP-P}}(x|I_n) &= \bar{m}_{sp}(x) \text{ as defined in (26)}, \\ m_{n,s}^{\text{SP-P-B}}(x|I_n) &= \hat{m}_{sp}(x) \text{ as defined in (27)}. \end{aligned}$$

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

$$MSE_{\mathrm{M}} = \frac{1}{P}\sum_{\forall x}\hat{u}^2(x). \tag{45}$$

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

$$100R^2 = 100 \times \left(1 - \frac{MSE_{\mathrm{M}}}{MSE_{\mathrm{HM}}}\right). \tag{46}$$

This is the out-of-sample R^2 (multiplied by 100) as reported in Campbell and Thompson (2008). We also report the decomposition of MSE into squared bias and variance (averaged over 1000 evaluation points) for the conditional mean estimators.

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

$$SOSD^{\mathrm{M}}(r) = \int_0^r \left[F^{\mathrm{M}}(s) - F^{\mathrm{HM}}(s)\right] ds, \qquad r > 0, \tag{47}$$

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 (1000 equidistant evaluation points in the range of squared forecast errors) in Table 2 and Table 3, 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 $c(\cdot)$. For asymmetric convex loss functions, convex loss stochastic dominance criterion could be adopted. We do not explore this issue here but direct interested readers to Granger (1999 Chapter 3) for more details. 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 7.

6.3 Simulation Results

The simulation results are presented in Table 1 and Table 2. Table 1 presents the variance, squared bias and MSE of the estimators of the conditional mean $m(x)$. It is clear from Table 1 that bagging estimators generally have larger bias compared to the constrained estimators. For example, when $\sigma_x^2 = 2$, the squared bias of NP-P is 0.282 while that for NP-P-B is 0.431. However, the reduction in variance is substantial via bagging, as can be seen that the variance for NP-P is 8.471 and it is reduced to be 7.434 for NP-P-B. This leads to an improvement in MSE for the bagging constrained estimators. Thus, we see similar properties of the conditional mean estimator as those of the slope estimators, as depicted in Figure 1, although the constraint is imposed on the slope $\beta(x)$ but not on the conditional mean $m(x)$.

We summarize the findings in Table 2 as follows:

First, note the varying slope of the cubic curve in the DGP in (43). A larger value of σ_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 2 confirms this by showing that the gains from imposing the monotonicity constraints and from bagging is large in this case. $100R^2$ is 42.0, 52.8, 58.2 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 $\sigma_x^2 = 4$ (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 2 again confirms that by showing the small gains from imposing the monotonicity constraints and from bagging. For example, $100R^2$ is 91.8, 92.1, 92.5 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 σ_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 σ_x^2. When the range of the evaluation x points expands to the stronger nonlinear part of the cubic curve, there are larger gains by using nonlinear models (NP and SP) over the linear model (L). When $\sigma_x^2 = 5$ (large), $100R^2$ is about 63 for L, while it is much larger, nearly 96 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 σ_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 even better than NP-HH (unless σ_x^2 is too large).

Fifth, bagging enhances the performance of the constrained NP/SP models (unless σ_x^2 is too large). 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 σ_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 3) 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 σ_x^2 is not too small. This signals the serious nonlinearity in the DGP. NP and SP are quite competing, with NP possibly 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 2, 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
Table 2 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.

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-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 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 6.2. In Table 3 and Figure 2, we only present the results for $s = 1$ as the results for $s = 6, 12$ (available upon request) show the same patterns with respect to nonlinearity and monotonicity. For bagging estimators in time series setting, the block bootstrap method is used. We consider the block length to be 1, 4 and 12 but the main results do not change much. Therefore, we only report the result for block length equal to 4. See Härdle, Horowitz and Kreiss (2003) and references therein for details of block bootstrap method for time series.

7.1 Empirical Results

We summarize the findings from Table 3 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 (2008). 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.

Fourth, the average SOSD and maximum SOSD measures in Table 3 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 3 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 (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.

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.

Figure 2 About Here

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), Du et al (2013), 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 β_1, which implies that $F_{\bar{\beta}}(z) = 0$ if $z < \beta_1$. When $z = \beta_1$, we have $F_{\bar{\beta}}(z) = \Pr\left(\bar{\beta} < \beta_1\right) + \Pr\left(\bar{\beta} = \beta_1\right) = \Pr\left(\tilde{\beta} \leq \beta_1\right) = F_{\tilde{\beta}}(\beta_1) = F_{\tilde{\beta}}(z)$. When $z > \beta_1$, $F_{\bar{\beta}}(z) = \Pr\left(\bar{\beta} \leq z\right) = \Pr\left(\bar{\beta} < \beta_1\right) + \Pr\left(\bar{\beta} = \beta_1\right) + \Pr\left(\beta_1 < \bar{\beta} \leq z\right) = F_{\tilde{\beta}}(\beta_1) + \Pr\left(\beta_1 < \tilde{\beta} \leq z\right) = F_{\tilde{\beta}}(z)$.

(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 $z \in R$,

$$\begin{aligned}
& \Pr\left(\gamma(n)\sigma_\beta^{-1}\left(\bar{\beta}-\beta\right)<z\right)=\Pr\left(\gamma(n)\sigma_\beta^{-1}\left(\max\{\tilde{\beta},\beta_1\}-\beta\right)<z\right) \\
= & \Pr\left(\gamma(n)\sigma_\beta^{-1}\left(\max\{\tilde{\beta},\beta_1\}-\beta\right)<z|\tilde{\beta}<\beta_1\right)\times\Pr\left(\tilde{\beta}<\beta_1\right) \\
& +\Pr\left(\gamma(n)\sigma_\beta^{-1}\left(\max\{\tilde{\beta},\beta_1\}-\beta\right)<z|\tilde{\beta}\geq\beta_1\right)\times\Pr\left(\tilde{\beta}\geq\beta_1\right) \\
= & \Pr\left(\gamma(n)\sigma_\beta^{-1}\left(\beta_1-\beta\right)<z\right)\times\Pr\left(\tilde{\beta}<\beta_1\right)+ \\
& \Pr\left(\gamma(n)\sigma_\beta^{-1}\left(\tilde{\beta}-\beta\right)<z|\tilde{\beta}\geq\beta_1\right)\times\Pr\left(\tilde{\beta}\geq\beta_1\right)
\end{aligned}$$

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

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

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

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

(ii)

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

(iii)

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

and (iv)

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

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

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

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 $z = 0$,

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

When $z < 0$,

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

(3) is trivial to show thus omitted here.

To prove (4), note that

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

Therefore, we have (5)

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

by Lemma 1, and (6)

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

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 (6). □

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

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

Therefore,

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

We have (i)

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

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

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

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

(2) From (1), we get

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

where we used Lemma 2.

(3) We need to prove that

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

where we used Lemma 2.

The proof for part (4) is trivial. □

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 $\bar{m}(x) = \tilde{m}_{LC}(x) \cdot 1_{\left\{\tilde{\beta}(x) \leq \beta_1(x)\right\}} + \tilde{m}_{LL}(x) \cdot 1_{\left\{\tilde{\beta}(x) > \beta_1(x)\right\}}$.

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

where,

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

and

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

Note that

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

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 (1-3) parallel those in Theorem 3. We prove part (4). 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

$$\begin{aligned} \hat{m}(x) - \tilde{m}(x) &= \left[\tilde{\beta}(x) - \hat{\beta}(x)\right] \times [\bar{x}(x) - x] \\ &= \left\{\left[\tilde{\beta}(x) - \beta(x)\right] + \left[\beta(x) - \hat{\beta}(x)\right]\right\} \times [\bar{x}(x) - x] \\ &= O\left(\frac{1}{\gamma_1(n,h)}\right) \times O\left(\frac{1}{\gamma_2(n,h)}\right) = o\left(\frac{1}{\gamma_2(n,h)}\right). \end{aligned}$$

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

Proof of Theorem 6. The proofs for part (1) and (2), (4), (5) and (6) follows that in Theorem 2. We only need to 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* (2008) 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 (1-3) follows in steps similar to part (1-3) of Theorem 5. We prove part (4). 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

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

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

B 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**) $E\left[Z^21_{\{Z_b>0\}}\right]=-b\varphi(b)+\Phi(b)$. (**d**) $E\left[Z_b1_{\{Z_b>0\}}\right]=\varphi(b)+b\Phi(b)$. (**e**) $E\left[Z_b^21_{\{Z_b>0\}}\right]=\Phi(b)+b\varphi(b)+b^2\Phi(b)$.

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**) $EZ^21_{\{Z_b>0\}}=\int_{-b}^{\infty}z^2\varphi(z)\,dz=-\int_{-b}^{\infty}z\varphi'(z)\,dz=-z\varphi(z)\,|_{-b}^{\infty}+\int_{-b}^{\infty}\varphi(z)\,dz=-b\varphi(b)+\Phi(b)$. (**d**) $E\left[Z_b1_{\{Z_b>0\}}\right]=EZ1_{\{Z_b>0\}}+bE1_{\{Z_b>0\}}=\varphi(b)+b\Phi(b)$. (**e**) $E\left[Z_b^21_{\{Z_b>0\}}\right]=E\left[(Z+b)^2\,1_{\{Z_b>0\}}\right]=EZ^21_{\{Z_b>0\}}+b^2E1_{\{Z_b>0\}}+2bE\left[Z1_{\{Z_b>0\}}\right]=\Phi(b)-b\varphi(b)+b^2\Phi(b)+2b\varphi(b)=\Phi(b)+b\varphi(b)+b^2\Phi(b)$.

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.

Barlow, R.E., D.J. Bartholomew, J.M. Bremner, and H.D. Brunk, 1972, Statistical inference under order restrictions: The theory and application of isotonic regression. Wiley, New York.

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.

Chen, X., 2007, Large sample sieve estimation of semi-nonparmetric models, in: J.J. Heckman and E. Leamer, (Eds.), Handbook of Econometrics, Vol. 6B. North-Holland, Amsterdam, pp. 5549-5632.

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

Delgado, M.A. and J.C. Escanciano, 2012, Distribution-free tests of stochastic monotonicity. Journal of Econometrics 170(1), 68-75.

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

Du, P., C.F. Parmeter and J.S. Racine, 2013, Nonparametric kernel regression with multiple predictors and multiple shape constraints. Statistica Sinica 23, 1343-1372.

Fama, E.F. and K.R. French, 2002, The equity premium. 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., 1992, The bootstrap and edgeworth expansion. Springer-Verlag, New York.

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

Härdle, W., J. Horowitz, and J.P. Kreiss, 2003, Bootstrap methods for time series. International Statistical Review 71, 435-459.

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.

Horowitz, J.L, 2001, The bootstrap, in: J.J. Heckman and E. Leamer, (Eds.), Handbook of Econometrics, Vol. 5. North-Holland, Amsterdam, pp. 3159-3228.

Judge, G.G. and T.A. Yancey, 1986, Improved methods of inference in econometrics, in: Theil, H. and H. Glejser, (Eds.), Studies in Mathematical and Managerial Economics, Vol. 34. North-Holland, Amsterdam.

Lee, S., O. Linton and Y.-J. Whang, 2009, Testing for stochastic monotonicity. Econometrica 77, 585-602.

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.

Lovell, M.C. and E. Prescott, 1970, Multiple regression with inequality constraints: Pretesting bias, hypothesis testing and efficiency. Journal of the American Statistical Association 65, 913-925.

Mammen, E., 1991, Estimating a smooth monotone regression function. Annals of Statistics 19(2), 724–740.

Mammen, E., 1992, When does bootstrap work? Asymptotic results and simulations, Lecture Note in Statistics Vol. 77. Springer-Verlag, New York.

Martins-Filho, C., S. Mishra and A. Ullah, 2008, A class of improved parametrically guided nonparametric regression estimators. Econometric Reviews 27, 542-573.

Matzkin, R.L., 1994, Restrictions of economic theory in nonparametric methods, in: R.F. Engle and D.L. McFadden, (Eds.), Handbook of Econometrics, Vol. 4. North-Holland, Amsterdam, pp. 2523-2558.

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

Ramsay, J.O., 1988, Monotone regression splines in action. Statistical Science 3(4), 425-441.

Table 1: Simulation Results: Variance, Squared Bias and Mean Squared Error (×100)

	$\sigma_x^2 = 2$			$\sigma_x^2 = 3$			$\sigma_x^2 = 4$			$\sigma_x^2 = 5$		
	var	bias^2	mse	var	bias^2	mse	var	bias^2	mse	var	bias^2	mse
HM	0.467	17.960	18.428	0.507	54.510	55.017	0.510	122.963	123.472	0.397	232.513	232.910
L	1.996	6.561	8.558	1.660	20.222	21.882	1.688	43.049	44.737	1.554	85.397	86.951
L-P	1.954	6.540	8.494	1.660	20.222	21.882	1.688	43.049	44.737	1.554	85.397	86.951
L-P-B	1.843	6.458	8.301	1.697	20.166	21.863	1.765	43.196	44.961	1.572	86.140	87.712
NP	10.555	0.328	10.883	9.962	0.422	10.384	9.332	0.777	10.109	8.572	1.315	9.887
NP-P	8.471	0.282	8.752	9.243	0.364	9.607	8.981	0.733	9.713	8.594	1.302	9.896
NP-P-B	7.434	0.431	7.866	8.146	0.569	8.715	8.352	0.941	9.293	7.551	2.760	10.311
NP-HH	9.020	1.004	10.024	14.811	0.511	15.322	8.851	0.809	9.660	8.476	1.418	9.894
SP	10.339	0.336	10.674	9.790	0.417	10.207	9.340	0.777	10.117	8.580	1.300	9.880
SP-P	8.428	0.276	8.705	9.097	0.359	9.456	8.968	0.733	9.701	8.576	1.304	9.880
SP-P-B	7.310	0.388	7.698	7.956	0.539	8.495	8.355	0.922	9.277	7.560	2.676	10.237

Table 2: 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	53.561	32.489	49.891	60.226	38.739	56.152	63.768	41.888	59.878	62.667	41.640	59.321
L-P	53.907	32.809	50.265	60.226	38.739	56.152	63.768	41.888	59.878	62.667	41.640	59.321
L-P-B	54.952	33.545	51.178	60.261	38.758	56.195	63.586	41.741	59.701	62.341	41.395	59.005
NP	40.941	21.253	38.133	81.126	53.981	75.646	91.813	64.038	86.217	95.755	68.268	90.584
NP-P	52.504	30.085	48.905	82.538	55.203	76.945	92.133	64.325	86.511	95.751	68.267	90.583
NP-P-B	57.315	33.633	53.408	84.159	56.589	78.445	92.474	64.648	86.845	95.573	68.096	90.408
NP-HH	45.604	24.683	42.506	72.151	46.698	67.247	92.177	64.376	86.563	95.752	68.265	90.581
SP	42.076	22.102	39.198	81.447	54.269	75.953	91.806	64.029	86.208	95.758	68.273	90.589
SP-P	52.764	30.310	49.148	82.812	55.446	77.203	92.143	64.336	86.523	95.758	68.273	90.589
SP-P-B	58.227	34.323	54.260	84.559	56.942	78.822	92.487	64.657	86.855	95.605	68.126	90.439

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

Table 3: 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 1: Asymptotic variance, Asymptotic squared bias, and Asymptotic mean squared error of constrained estimator (CE) and bagging constrained estimator (BCE)

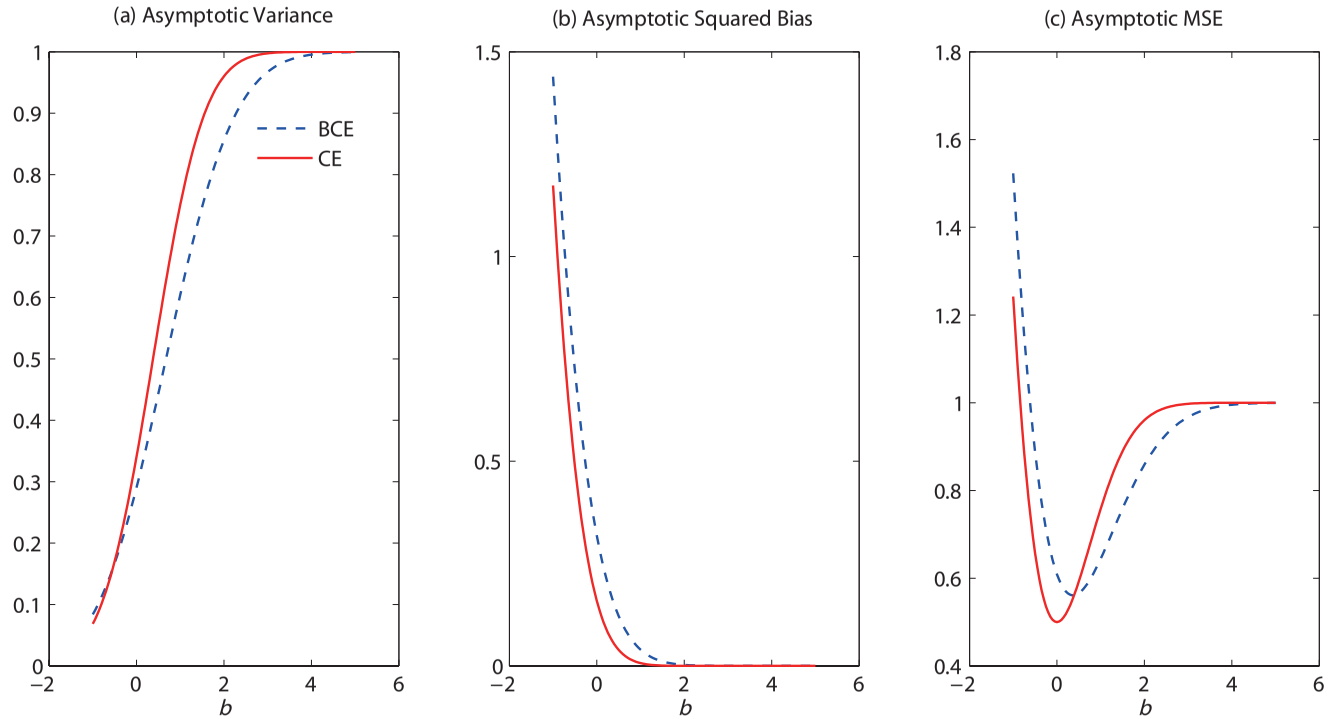

Figure 2: SOSD for lty