Forecasting with Model Uncertainty: Representations and Risk Reduction[*]

Keisuke Hirano[†] Jonathan H. Wright[‡]

First version: October 14, 2013
This version: January 1, 2014

Preliminary and incomplete; please do not circulate

Abstract

We consider forecasting with weak predictors. The researcher wants to select a model (a set of predictors), estimate the parameters, and use this for forecasting. We investigate the local asymptotic mean square prediction error (MSPE) of different forecasting schemes: including in-sample using the Akaike information criterion, out-of-sample forecasting, and splitting the data into subsamples for model selection and parameter estimation. We consider all these methods both with and without bootstrap aggregation (bagging). We develop an asymptotic representation result that facilitates comparison of the procedures. Numerically, we find that for many values of the local parameter, the out-of-sample and split-sample schemes with bagging perform well. We also show that an alternative form of bagging uniformly improves the accuracy of the out-of-sample and split-sample methods.

[*] We are grateful to Marine Carrasco, Russell Davidson, Gary Chamberlain, Sylvia Gonçalves and Serena Ng for very helpful discussions. The usual disclaimer applies.

[†] Department of Economics, University of Arizona, 1130 E. Helen St., Tucson AZ 85721. Email: hirano@u.arizona.edu

[‡] Department of Economics, Johns Hopkins University, 3400 North Charles St., Baltimore MD 21218. Email: wrightj@jhu.edu

1 Introduction

In this paper, we reconsider the problem of forecasting when there is uncertainty about which variables to include in the forecasting model. As is well known, a model that fits well in sample may not be good for forecasting—a model may fit well in-sample, only to turn out to be useless in prediction. Consequently, it is common practice to select the model based on pseudo-out-of-sample fit from a sequence of recursive or rolling predictions. Parameters are then estimated over the whole sample period. The idea of using an out-of-sample criterion was advocated by Ashley, Granger, and Schmalensee (1980) and Clark (2004), and is very intuitive: it is what a researcher could have done at the time. But alternatively, one might select the model based on in-sample fit, but adjust for overfitting by using an information criterion, such as the Akaike Information Criterion (AIC) (Akaike, 1974), as advocated by Inoue and Kilian (2006).

We consider a setting with a fixed number k of potential predictors, each of which has a coefficient that is local to zero. Selecting a forecasting model amounts to selecting a subset of the k potential predictors—there are thus $M = 2^k$ possible models among which we must choose. Having chosen the model, we then have to estimate the parameters and use these for forecasting. Although some model will be best in terms of predictive accuracy, the local-to-zero nesting means that we can never consistently select that model. We consider various methods of model selection and forecasting: including using in-sample fit with the AIC information criterion; selecting the model based on recursive pseudo-out-of-sample forecast accuracy and then using the whole dataset for parameter estimation; and splitting the sample into two, using one part for model selection and the other for parameter estimation. We call this last method the split-sample approach. Unlike the first two methods, it is not commonly used in practice. But it does ensure independence between parameter estimates and model selection, unlike methods based on in-sample fit (Leeb and Pötscher, 2005; Hansen, 2009), and also unlike the standard out-of-sample approach. We also consider the addition of a bootstrap aggregation (bagging) step (Breiman, 1996) in which the data are resampled, the forecasting method is applied to the resampled data, and the resulting forecasts are then averaged over all the bootstrap samples.

We obtain asymptotic characterizations of these forecasting procedures under the local parametrization. A key step is to obtain an asymptotic representation of a certain partial sum process as the sum of a term that is directly informative about the local parameters, and another term that is an independent Gaussian process. This allows us to provide a limit-experiment type representation of the procedures, from which we can calculate normalized local asymptotic mean square prediction errors up to $O(T^{-1})$ terms. We show that the recursive pseudo-out-of-sample and split-sample procedures are inefficient, in the sense that their limit distributions depend on the ancillary Gaussian noise process.

Our characterizations also suggest ways to improve upon these procedures. Bagging has a smoothing effect that alters the risk properties of estimators, but it can also reduce the influence of the extraneous noise term in the out-of-sample and split-sample methods. Earlier theoretical work on bagging, notably Bühlmann and Yu (2002), emphasized its smoothing effect but not the noise reduction effect. We also develop an alternative version of bagging that solely removes the impact of the ancillary noise term; doing this is shown to uniformly improve the out-of-sample and split-sample methods asymptotically.

We then numerically compare the various procedures, both in terms of their local asymptotic risk, and their finite-sample performance. Without bagging there is no unambiguous rank ordering among these three methods, but we find that for many values of the localization parameter, in-sample forecasting using the AIC gives the most accurate forecasts, out-of-sample prediction does worse, and the split-sample method does worst of all. This is intuitive because the out-of-sample and split-sample schemes are in some sense wasting data, and is essentially the argument of Inoue and Kilian (2004) and Inoue and Kilian (2006) for the use of in-sample rather than out-of-sample predictability tests. However, introducing the bagging step, or our alternative to bagging, changes the rank ordering substantially. It generally reduces the local asymptotic mean square prediction error of the in-sample forecasts, but makes a more dramatic difference to the out-of-sample and split-sample forecasts[1]. In our numerical work, we find no case in which standard bagging fails to reduce the local asymptotic mean square prediction error of out-of-sample and split-sample forecasts. Meanwhile, we show theoretically that the alternative form of bagging improves the accuracy of out-of-sample and split-sample forecasts uniformly in the localization parameter.

For many values of the localization parameter, the incorporation of the bagging step entirely reverses the relative ordering of the in-sample, out-of-sample, and split-sample prediction methods. When the true model includes only a single predictor and the number of candidate predictors is large, we find that the use of the split-sample approach with a bagging step provides the most accurate forecasts from among any of the methods considered here.

In the next section, we set up the local parametrization, introduce the various procedures we will evaluate, and derive asymptotic characterizations via our representation theorem for the partial sum process. Section 3 explores the asymptotic and finite-sample risk properties of the procedures through a series of numerical experiments. Section 4 extends the results to the multi-step forecasting case with serially correlated errors and to vector autoregressions. Section 5 concludes.

2 Local Asymptotics

The setup is a standard regression model in which

$$y_t = \beta' x_t + u_t \tag{1}$$

where u_t is iid with mean 0, finite variance σ^2 and $2+\delta$ finite moments for some $\delta > 0$, x_t is a $k \times 1$ stationary vector such that $E(x_t x_t^{'}) = \Sigma_{xx}$ where Σ_{xx} is finite and nonsingular and k is fixed. We observe data on $\{x_t, y_t\}$ for $t = 1, \ldots, T$, along with x_{T+1}. Our goal is to forecast y_{T+1}. We assume that the slope coefficient is local to zero as $\beta = \beta_T = T^{-1/2} b$,

[1] One other useful feature of the out-of-sample forecasting setup is that it can be constructed to use only real-time data which precisely mimics the data available to a researcher in the presence of data revisions. Unfortunately, adding a bootstrap aggregation step destroys this feature.

where $b \in \mathbb{R}^k$ is the local parameter. This local parametrization was also used by Inoue and Kilian (2004).

The researcher is unsure as to which elements of x_t to include in the model, and considers eleven possible strategies:

1. **Big model**. Let $\hat{\beta} = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t y_t$ be the unrestricted OLS estimator. Estimate the full model (with no restrictions on β) using OLS on the whole dataset and use this estimate for forecasting.

2. **The positive-part James-Stein estimator**. Estimate the large model on the whole dataset by OLS giving the estimator $\hat{\beta}$ and let $\hat{V}$ denote its estimated asymptotic variance-covariance matrix. Then estimate the parameter vector as

$$\hat{\beta} \max(1 - \frac{k-2}{T\hat{\beta}'\hat{V}^{-1}\hat{\beta}}, 0)$$

if $k > 2$, and use this for forecasting (James and Stein, 1961; Baranchik, 1964).

3. **Small model**. Impose $\beta = 0$ and use this for forecasting.

4. **In-sample**. Fit the model to the data and use AIC to decide which combination of the x_ts to use (including all and none). With $k = 1$, of course this amounts to just picking the big or the small model, but in general, there are 2^k possible models to choose among.

5. **Out-of-sample**. Estimate the model recursively starting a fraction π of the way through the sample. Pick between among the possible models depending on which predicts best out of sample. Then estimate the chosen model over the full sample and use for forecasting.[2]

6. **Split-sample**. Apply AIC to the first π of the sample and select the best model. Then estimate the chosen model over the remainder of the sample and use for forecasting. Note that there is no overlap between the model selection and estimation samples in this scheme.

7. **Big model with bagging.** This means resampling with replacement from the pairs $\{x_t, y_t\}$ with replacement, constructing the forecast using the resampled data as in (1), and

[2] As Clark (2004) notes, some researchers define "out-of-sample" forecasting to mean using the whole dataset for model selection and then estimating parameters only on a subset of the data. We are following the Ashley, Granger, and Schmalensee (1980) and Clark (2004) definition of out-of-sample forecasting.

then averaging the resulting forecasts over L different bootstrap samples.

8. **Positive-part James-Stein estimator with bagging.** Applying bagging to the positive-part James-Stein estimator.

9. **In-sample with bagging.** Apply bagging to the in-sample/AIC scheme.

10. **Out-of-sample with bagging**. Apply bagging to the out-of-sample scheme. Note that this means applying bagging to the entire methodology of selecting a model based on pseudo-out-of-sample forecast accuracy and then estimating the parameters of the selected model. It's not just an out-of-sample evaluation of bagging forecasts.

11. **Split-sample with bagging**. Apply bagging to the split-sample scheme.

Each forecast can be written in the form $\tilde{\beta}' x_{T+1}$ (typically, some elements of the coefficient vector are constrained to be zero). We consider the unconditional mean square prediction error:

$$MSPE = E(y_{t+1} - \tilde{\beta}' x_{T+1})^2 = E(u_{T+1} + (\beta - \tilde{\beta})' x_{T+1})^2$$

$$= \sigma^2 + E[(\tilde{\beta} - \beta)' \Sigma_{xx} (\tilde{\beta} - \beta)] + o(T^{-1}) \qquad (2)$$

The first term on the right hand side of (2) is the asymptotic forecast error neglecting parameter uncertainty, which is the same for all forecasts. The second term is $O(T^{-1})$ and this differs across forecasts. Because it is only this second term that we can do anything about, we define the normalized mean square prediction error as $NMSPE = T(MSPE - \sigma^2) = E[T^{1/2}(\tilde{\beta} - \beta)' \Sigma_{xx} T^{1/2} (\tilde{\beta} - \beta)]$, and consider this exclusively henceforth. This can be viewed as the regret risk corresponding to normalized squared error loss for $\tilde{\beta}$. Alternatively, we could consider forecast accuracy conditional on a particular value for x_{T+1}, which would lead to a different criterion, but we focus on the unconditional criterion for the remainder of the paper.

The model in (1), for any fixed σ and fixed distribution for u_t satisfying suitable regularity conditions, is locally asymptotically normal (LAN). As a result, it has a limit experiment representation (see for example van der Vaart, Chs. 7-9). In particular, consider any estimator sequence $\tilde{\beta}$ with limit distributions in the sense that

$$T^{1/2}\tilde{\beta} \rightarrow_{d,b} \mathcal{L}_b,$$

where $\rightarrow_{d,b}$ denotes convergence in distribution[3] under the local parameter b, and $\mathcal{L}_b$ is a law that may depend on b. Then, the estimator $\tilde{\beta}$ has an asymptotic representation as a randomized estimator in the shifted normal model: if Y is a single draw from the $N(b, \sigma^2\Sigma_{xx}^{-1})$ distribution, and U is random variable independent of Y (with sufficiently rich support[4]), there exists an estimator $S(Y, U)$ with

$$S(Y, U) \sim \mathcal{L}_b$$

for all b. That is, the sequence $T^{1/2}\tilde{\beta}$ is asymptotically equivalent to the randomized estimator S under all values of the local parameter.

While this general asymptotic representation is very powerful, it leaves open how to find the representation S corresponding to any particular estimator. However, we can specialize the result to obtain useful characterizations of the procedures we are considering. All of the estimators we consider depend crucially on the partial sum process

$$T^{-1/2}\sum_{t=1}^{[Tr]} x_t y_t = T^{-1/2}\sum_{t=1}^{[Tr]} x_t(x_t' b/\sqrt{T} + u_t).$$

This converges to a k-dimensional Brownian motion with drift, and by the properties of Brownian motion, we can decompose the limit into two terms:

Theorem 1: As $T \rightarrow \infty$, the partial sum process $T^{-1/2}\sum_{t=1}^{[Tr]} x_t y_t \rightarrow_d \Sigma_{xx} Y(r)$ where $Y(r) \overset{d}{=} rY + U_B(r)$, $Y \sim N(b, \sigma^2\Sigma_{xx}^{-1})$ and $U_B(r)$ is a k-dimensional Brownian bridge with covariance matrix $\sigma^2\Sigma_{xx}^{-1}$ that is independent of Y.

[3] In the sequel, we will use $\rightarrow_d$ to denote convergence in distribution under b when the dependence on b is clear from the context.

[4] Typically, a representation $S(Y, U)$ exists for U distributed uniform on $[0, 1]$, but for our results below, it is useful to allow U to have a more general form.

The proofs of the Theorems are given in the Appendix. With Theorem 1, we can obtain explicit asymptotic representations of estimator sequences in terms of the asymptotically sufficient component $Y = Y(1)$ and (possibly) an independent stochastic component $U = U_B$. For $g \subset \{1, \ldots, k\}$, let $\hat{\beta}(g)$ denote the vector of OLS coefficient estimates corresponding to the predictors indexed by g, with zeros in all other locations. Let $Y(r, g)$ denote the $k \times 1$ vector with the elements of $Y(r)$ in the locations indexed by g and zeros elsewhere. Define $n(g)$ as the number of elements in g. Let $\Sigma_{xx}(g)$ denote the kxk matrix that consists of the elements of Σ_{xx} in the rows and columns indexed by g and zeros in all other locations, and let $H(g) = \Sigma_{xx}(g)^{+}\Sigma_{xx}$, where $\Sigma_{xx}(g)^{+}$ denotes the Moore-Penrose inverse of $\Sigma_{xx}(g)$. It follows immediately from Theorem 1 that $T^{1/2}\hat{\beta} \rightarrow_d Y(1)$ and $T^{1/2}\hat{\beta}(g) \rightarrow_d H(g)Y(1, g)$. Theorem 2 provides the asymptotic distribution of the various estimates incorporating model selection without bagging.

Theorem 2: In large samples, the distribution of the parameter estimates (without bagging) is as follows:

(i) Using the big model:

$$T^{1/2}\tilde{\beta} \rightarrow_d Y(1) \tag{3}$$

(ii) Using the positive-part James-Stein estimator:

$$T^{1/2}\tilde{\beta} \rightarrow_d Y(1)\max(1 - \frac{k-2}{Y(1)'Y(1)}, 0) \tag{4}$$

(iii) Selecting the model using the AIC:

$$\begin{aligned} T^{1/2}\tilde{\beta} \rightarrow_d \sum_{g^*} H(g^*)Y(1, g^*)\mathbf{1}\{g^* = \arg\min_g [Y(1,g)'H(g)\Sigma_{xx}H(g)Y(1,g) \\ -2Y(1,g)'H(g)\Sigma_{xx}Y(1) + 2n(g)\sigma^2]\mathbf{\}} \end{aligned} \tag{5}$$

(iv) Selecting the model minimizing recursive out-of-sample error starting a fraction π of the way through the sample:

$$T^{1/2}\tilde{\beta} \rightarrow_d \sum_{g^*} H(g^*)Y(1,g^*)\mathbf{1}\{g^* = \arg\min_g [\int_\pi^1 \frac{Y(r,g)'}{r} H(g)\Sigma_{xx}H(g)\frac{Y(r,g)}{r}dr$$
$$-2\int_\pi^1 \frac{Y(r,g)'}{r}H(g)\Sigma_{xx}dY(r)]\} \tag{6}$$

(v) Using the split-sample method, using the first fraction π of the sample for model selection and the rest for parameter estimation:

$$T^{1/2}\tilde{\beta} \rightarrow_d \sum_{g^*} H(g^*)\frac{Y(1,g^*)-Y(\pi,g^*)}{1-\pi}\mathbf{1}\{g^* = \arg\min_g [\frac{1}{\pi}Y(\pi,g)'H(g)\Sigma_{xx}H(g)Y(\pi,g)$$
$$-\frac{2}{\pi}Y(\pi,g)'H(g)\Sigma_{xx}Y(\pi) + 2n(g)\sigma^2]\} \tag{7}$$

where $\sum_{g^*}$ denotes the summation over the $M = 2^k$ possible models. Using the small model, we just have $\tilde{\beta} = 0$.

The asymptotic NMSPEs are given by the expected sum of squared deviations of the asymptotic distributions in Theorem 2 from b.

Inoue and Kilian (2004) considered the local power of some in-sample and out-of-sample tests of the hypothesis that $\beta = 0$. They derived equation (3) and a result very similar to equation (6).

Of course, there are other criteria besides AIC that we could use for in-sample model selection. Some of these are asymptotically equivalent to AIC, such as Mallows' C_p criterion (Mallows, 1973) or leave-one-out cross-validation. Using any of these information criteria for in-sample model selection will give the asymptotic NMSPE in equation (5). Alternatively, one could use the Bayes information criterion (BIC). In the present setting, because the penalty term goes to zero at a rate slower than T^{-1}, the BIC will always pick the small model ($\beta = 0$).

The limiting distributions in equations (3)-(5) are all functions of Y alone (note that $Y(1) = Y$ because $U_B(1) = 0$). They are also easy to interpret via the limit of experiments framework: OLS corresponds to the estimator $S = Y$ in the limit experiment $Y \sim N(b, \sigma^2\Sigma_{xx}^{-1})$. The positive-part James-Stein estimator is asymptotically equivalent to a shrinkage estimator in the normal shift model, and the limiting distribution of the AIC procedure is equivalent to applying AIC model selection to the normal shift model. The estimators other than the out-of-sample and split-sample estimators can all be thought of as generalized shrinkage estimators (Stock and Watson, 2012) as their limiting distributions are of the form: $T^{1/2}\tilde{\beta} \to_d Yg(Y)$ for some nonlinear function $g(Y)$. In contrast, the limiting distributions in equations (6) and (7) are functions of both Y and an independent Brownian bridge, $U_B(r)$. Thus the out-of-sample and split sample schemes are based on Y but also on an additional random noise component. This means that they are not just shrinkage estimators, asymptotically. Moreover, based on this representation, it seems plausible that these procedures can be improved upon, a point to which we will return below.

Equations (5), (6) and (7) simplify considerably in the case of orthonormal predictors ($\Sigma_{xx} = I_k$). In this case, the estimators using the AIC, out-of-sample, and split sample strategies are

$$T^{1/2}\tilde{\beta} \to_d \sum_{g^*} Y(1,g^*)\mathbf{1}\{g^* = \arg\min_g [Y(1,\bar{g})'Y(1,\bar{g}) + 2n(g)\sigma^2]\}$$

$$T^{1/2}\tilde{\beta} \to_d \sum_{g^*} Y(1,g^*)\mathbf{1}\{g^* = \arg\min_g [\int_\pi^1 \frac{Y(r,g)'}{r}\frac{Y(r,g)}{r}dr - 2\int_\pi^1 \frac{Y(r,g)'}{r}dY(r)]\}$$

and

$$T^{1/2}\tilde{\beta} \to_d \sum_{g^*} \frac{Y(1,g^*) - Y(\pi,g^*)}{1-\pi}\mathbf{1}\{g^* = \arg\min_g [\frac{1}{\pi}Y(\pi,\bar{g})'Y(\pi,\bar{g}) + 2n(g)\sigma^2]\}$$

respectively, where $\bar{g} = \{1, 2, \ldots, k\} \setminus g$.

Next we consider the asymptotic distributions of the bagged versions of the forecast procedures. The ith bagging step resamples from the pairs $\{(x_t, y_t), t = 1, \ldots, T\}$ with replacement to form a pseudo-sample $\{x_t^*(i), y_t^*(i), t = 1, \ldots, T\}$. The forecast on the ith bootstrap sample can be written as $\tilde{\beta}_i' x_T$. This is repeated L times, and the L forecasts are averaged to obtain the bagged forecast. The next theorem provides a key result for obtaining the limiting distribution of a single bootstrap sample.

Theorem 3: Let $\{x_t^*(i), y_t^*(i), t = 1, \ldots, T\}$ be the ith bootstrap sample. In large samples $T^{-1/2}\Sigma_{t=1}^{[Tr]} x_t^*(i) y_t^*(i) \rightarrow_d \Sigma_{xx}(rY + V_i(r)) \equiv \Sigma_{xx} Y_i^*(r)$ where Y is as in Theorem 1 and $\{V_i(r)\}_{i=1}^L$ are $k \times 1$ Brownian motions with covariance matrix $\sigma^2 \Sigma_{xx}^{-1}$ that are independent of Y and of each other.

Thus the limiting distribution of a single bootstrap draw for the partial sums process mimics the result in Theorem 1, except that the Brownian Bridge U_B is replaced with a Brownian motion V_i. Using this result, we obtain asymptotic representations for a single bootstrap draw of the forecast procedures.

Theorem 4: In the ith bootstrap sample ($i = 1, \ldots, L$), in large samples, the distributions of the alternative parameter estimates including a bagging step are as follows:
(i) Using the big model:

$$T^{1/2}\tilde{\beta}_i \rightarrow_d Y_i^*(1) \tag{8}$$

(ii) Using the positive-part James-Stein estimator:

$$T^{1/2}\tilde{\beta}_i \rightarrow_d Y_i^*(1) \max(1 - \frac{k-2}{Y_i^*(1)' Y_i^*(1)}, 0) \tag{9}$$

(iii) Selecting the model using the AIC:

$$T^{1/2}\tilde{\beta}_i \rightarrow_d \sum_{g^*} H(g^*) Y_i^*(1, g^*) \mathbf{1}\{g^* = \arg\min_g [Y_i^*(1,g)' H(g) \Sigma_{xx} H(g) Y_i^*(1,g) \\ -2 Y_i^*(1,g)' H(g) \Sigma_{xx} Y_i^*(1) + 2n(g)\sigma^2]\} \tag{10}$$

(iv) Selecting the model minimizing out-of-sample error:

$$T^{1/2}\tilde{\beta}_i \to_d \sum_{g^*} H(g^*) Y_i^*(1,g^*) \mathbf{1}\{g^* = \arg\min_g [\int_\pi^1 \frac{Y_i^*(r,g)'}{r} H(g)\Sigma_{xx} H(g) \frac{Y_i^*(r,g)}{r} dr$$

$$-2\int_\pi^1 \frac{Y_i^*(r,g)'}{r} H(g)\Sigma_{xx} dY_i^*(r)]\} \tag{11}$$

(v) Using the split-sample method:

$$T^{1/2}\tilde{\beta}_i \to_d \sum_{g^*} H(g^*) \frac{Y_i^*(1,g^*) - Y_i^*(\pi,g^*)}{1-\pi} \mathbf{1}\{g^* = \arg\min_g [\frac{1}{\pi} Y_i^*(\pi,g)' H(g)\Sigma_{xx} H(g) Y_i^*(\pi,g)$$

$$-\frac{2}{\pi} Y_i^*(\pi,g)' H(g)\Sigma_{xx} Y_i^*(\pi) + 2n(g)\sigma^2]\} \tag{12}$$

where $\sum_{g^*}$ denotes the summation over the M possible models and $Y_i^*(r,g)$ is a $k \times 1$ vector with the elements of $Y_i^*(r)$ in the locations indexed by g and zeros elsewhere.

The distribution of the parameter estimates from bagging are then given by averaging the expressions in equations (8)–(12) over L different draws of $V_i(r)$. Finally, the asymptotic NMSPEs are given by the expected sum of squared elements of these asymptotic distributions.

For all of the bagged procedures characterized in Theorem 4, averaging over the L draws for $V_i(r)$ implies that their limiting distributions depend on Y alone. In the case of the big model (the full OLS estimator), integrating over $V_i(r)$ leads to the same limit as the original OLS estimator without bagging, and the inclusion of the bagging step is asymptotically irrelevant. However, for the other procedures, bagging changes their asymptotic distributions. In the case of the out-of-sample and split-sample procedures, bagging results in limiting distributions that do not depend on random elements other than Y. This suggests that bagging may be particularly effective in improving the risk properties of these procedures.

2.1 Shrinkage Representations in the case $k = 1$

In the case $k = 1$, some of the expressions in Theorems 2 and 4 can be simplified. Without bagging, for the AIC estimator we have:

$$T^{1/2}\tilde{\beta} \to_d Y1(|Y| > \sqrt{2}\tilde{\sigma})$$

where $\tilde{\sigma} = \sigma\Sigma_{xx}^{-1/2}$. For the split sample estimator, we have:

$$T^{1/2}\tilde{\beta} \to_d z_1 1(|z_2| > \tilde{\sigma}\sqrt{2/\pi})$$

where $z_1 = Y + \frac{U_B(1)-U_B(\pi)}{1-\pi}$ and $z_2 = Y + \frac{U_B(\pi)}{\pi}$. By direct calculations, z_1 is $N(b, \frac{\tilde{\sigma}^2}{1-\pi})$, z_2 is $N(b, \frac{\tilde{\sigma}^2}{\pi})$ and z_1 and z_2 are mutually independent.

As observed earlier, without bagging, the estimators other than the out-of-sample and split-sample estimators have shrinkage representations of the form: $T^{1/2}\tilde{\beta} \to_d Yg(Y)$ for some nonlinear function $g(Y)$. When we add in the bagging step, all of the estimators, including the out-of-sample and split-sample estimators, have shrinkage representations of this form. For the AIC estimator, with bagging, we have:

$$T^{1/2}\tilde{\beta} \to_d Y - Y\Phi(\frac{\sqrt{2}\sigma - Y}{\tilde{\sigma}}) + \tilde{\sigma}\phi(\frac{\sqrt{2}\sigma - Y}{\tilde{\sigma}}) + Y\Phi(\frac{-\sqrt{2}\sigma - Y}{\tilde{\sigma}}) - \tilde{\sigma}\phi(\frac{-\sqrt{2}\sigma - Y}{\tilde{\sigma}})$$

shown in proposition 2.2 of Bühlmann and Yu (2002) for the case $\Sigma_{xx} = 1$.[5]

Meanwhile, for the split-sample estimator in the ith bootstrap sample, we have:

$$T^{1/2}\tilde{\beta}_i \to_d z_1(i)1(|z_2(i)| > \sigma\sqrt{2/\pi})$$

where $z_1(i) = Y + \frac{V_i(1)-V_i(\pi)}{1-\pi}$ and $z_2(i) = Y + \frac{V_i(\pi)}{\pi}$. By direct calculations, $z_1(i)|Y$ is $N(Y, \frac{\tilde{\sigma}^2}{1-\pi})$,

[5] Indeed, given the orthonormal setting, even if $k > 1$, if we sort the coefficient estimates by their absolute magnitude and apply AIC sequentially to these models, dropping variables one at a time as long as called for by the information criterion, then the above two expressions will apply to each element of $\tilde{\beta} - \beta$ (Bühlmann and Yu, 2002; Stock and Watson, 2012). But the use of the AIC that we are considering in this paper is to select among all 2^k possible models and so no such simplification is available in this case.

$z_2(i)|Y$ is $N(Y, \frac{\tilde{\sigma}^2}{\pi})$ and the two are independent, conditional on Y. Thus for the overall split-sample with bagging estimator:

$$T^{1/2}\tilde{\beta} \rightarrow_d Y - Y\Phi(\frac{\sqrt{2}\sigma - \sqrt{\pi}Y}{\tilde{\sigma}}) + Y\Phi(\frac{-\sqrt{2}\sigma - \sqrt{\pi}Y}{\tilde{\sigma}})$$

We have no such simplified expression for the out-of-sample with bagging estimator, but we still know from equation (11) that the limit is a function of Y alone.

2.2 Alternative Bagging Scheme

The asymptotic distribution for the out-of-sample and split-sample estimators in Theorem 2 suggests a way to improve them. Both estimators have asymptotic representations that are nontrivial functions of both $Y = Y(1) \sim N(b, \sigma^2\Sigma_{xx}^{-1})$, and $U = U_B$ where U_B is a Brownian bridge independent of Y. In the statistical experiment corresponding to observing the pair (Y, U), the variable Y is sufficient. Thus, for an estimator $S(Y, U)$, consider the estimator

$$\tilde{S}(Y) = E[S(Y, U)|Y].$$

By the Rao-Blackwell theorem, the risk of $\tilde{S}(Y)$ is less than or equal to that of $S(Y, U)$, for any convex loss function. In particular, the conditional estimator will have weakly lower NMSPE than the original estimator, for all values of b. Our numerical calculations below indicate that the risk is strictly lower, for at least some values of b. Hence, the out-of-sample and split-sample procedures are asymptotically inadmissible.

To implement the conditional estimators, we can employ a variant of the bagging scheme. Recall that $T^{1/2}\hat{\beta}(g) \rightarrow_d H(g)Y(1, g)$. We could take independent artificially generated Brownian bridges $\{U_B^i(r)\}_{i=1}^L$ with covariance matrix $\hat{\sigma}\hat{\Sigma}_{xx}^{-1}$, where $\hat{\sigma}^2$ and $\hat{\Sigma}_{xx}$ are consistent estimators of σ^2 and Σ_{xx}, respectively. Then, for each i, consider the estimators:

$$\sum_{g^*}\hat{\beta}(g^*)\mathbf{1}\{g^* = \arg\min_g[\int_\pi^1[T^{1/2}\hat{\beta}(g)+\hat{H}(g)\frac{U_B^i(r,g)}{r}]'\hat{\Sigma}_{xx}[T^{1/2}\hat{\beta}(g)+\hat{H}(g)\frac{U_B^i(r,g)}{r}]dr$$
$$-2\int_\pi^1[T^{1/2}\hat{\beta}(g)+\hat{H}(g)\frac{U_B^i(r,g)}{r}]'\hat{\Sigma}_{xx}T^{1/2}\hat{\beta}dr$$
$$-2\int_\pi^1[T^{1/2}\hat{\beta}(g)+\hat{H}(g)\frac{U_B^i(r,g)}{r}]'\hat{\Sigma}_{xx}dU_B^i(r)\}$$

where $\hat{H}(g)=\hat{\Sigma}_{xx}(g)^+\hat{\Sigma}_{xx}$ and

$$\sum_{g^*}[\hat{\beta}(1,g^*)-T^{-1/2}\frac{U_b^i(\pi,g^*)}{1-\pi}]$$
$$\mathbf{1}\{g^* = \arg\min_g[\frac{1}{\pi}[T^{1/2}\pi\hat{\beta}(g)+\hat{H}(g)U_B^i(\pi,g)]'\hat{\Sigma}_{xx}[T^{1/2}\pi\hat{\beta}(g)+\hat{H}(g)U_B^i(\pi,g)]$$
$$-\frac{2}{\pi}[T^{1/2}\pi\hat{\beta}(g)+\hat{H}(g)U_B^i(\pi,g)]'\hat{H}(g)\hat{\Sigma}_{xx}[T^{1/2}\pi\hat{\beta}+U_B^i(\pi)]+2n(g)]\}$$

where $U_B^i(r,g)$ is the vector with the elements of $U_B^i(r)$ in the locations indexed by g and zeros everywhere else. Let $\tilde{Y}_i(r) = rY + U_B^i(r)$ and $\tilde{Y}_i(r,g)$ be a $k\times 1$ vector with the elements of $\tilde{Y}_i(r)$ in the locations indexed by g and zeros elsewhere. For each i, the above two estimators have the distributions:

$$T^{1/2}\tilde{\beta}_i \to_d \sum_{g^*}H(g^*)\tilde{Y}_i(1,g^*)\mathbf{1}\{g^* = \arg\min_g[\int_\pi^1\frac{\tilde{Y}_i(r,g)'}{r}H(g)\Sigma_{xx}H(g)\frac{\tilde{Y}_i(r,g)}{r}dr$$
$$-2\int_\pi^1\frac{\tilde{Y}_i(r,g)'}{r}H(g)\Sigma_{xx}d\tilde{Y}_i(r)]\} \qquad (13)$$

and

$$T^{1/2}\tilde{\beta}_i \to_d \sum_{g^*}H(g^*)\frac{\tilde{Y}_i(1,g^*)-\tilde{Y}_i(\pi,g^*)}{1-\pi}\mathbf{1}\{g^* = \arg\min_g[\frac{1}{\pi}\tilde{Y}_i(\pi,g)'H(g)\Sigma_{xx}H(g)\tilde{Y}_i(\pi,g)$$
$$-\frac{2}{\pi}\tilde{Y}_i(\pi,g)'H(g)\Sigma_{xx}\tilde{Y}_i(\pi)+2n(g)\sigma^2]\} \qquad (14)$$

respectively, which are the same distributions as in equations (6) and (7). Then average these estimators over i. After this step of averaging over different realizations of the Brownian bridge, the asymptotic distributions depend on Y alone. Thus, use of these estimators represents a form of bagging of the out-of-sample and split-sample estimators where the bootstrap aggregation is done by taking draws from the part of the asymptotic distribution that is independent of Y.

Note that this alternative form of bagging does not apply to the in-sample estimator because there is no ancillary noise process to eliminate in this case.

The approach of averaging over different realizations of a Brownian bridge does not have the form of a standard bootstrap. But the conditional estimators can also be implemented via a scheme where we resample the data. Let $\hat{\beta}$ be the OLS estimate, using the full sample and all predictors and let $e_t = y_t - x_t^{'}\hat{\beta}$. Suppose that there are L bootstrap draws and let $\{\theta_t(i)\}$ be ± 1, each with equal probability and independent across t and i, for $t = 1, ...T$ and $i = 1, ...L$. Let $z_t = x_t y_t = x_t x_t^{'}\hat{\beta} + x_t e_t$ and define the ith bootstrap draws of z_t as:

$$z_t^*(i) = x_t x_t^{'}\hat{\beta} + \theta_t(i) x_t e_t - T^{-1}\Sigma_{s=1}^{T}\theta_s(i) x_s e_s \tag{15}$$

All the estimators considered are functions of $\{z_t\}_{t=1}^{T}$ alone. Our propose resampling scheme replaces z_t by $z_t^*(i)$, computes the estimator, and then averages the resulting estimator over i. Note that this scheme does not resample $\{y_t\}$ or $\{x_t\}$ separately, but rather draws from the scores $\{z_t\}$ directly.

In each bootstrap sample, the partial sum process is:

$$\begin{aligned}
T^{-1/2}\Sigma_{t=1}^{[Tr]} z_t^*(i) &= T^{-1/2}\Sigma_{t=1}^{[Tr]} x_t x_t^{'}\hat{\beta} + T^{-1/2}\Sigma_{t=1}^{[Tr]}\theta_t x_t e_t - T^{-1/2} r \Sigma_{s=1}^{T}\theta_s x_s e_s \\
&= T^{-1/2}\Sigma_{t=1}^{[Tr]} x_t x_t^{'}\hat{\beta} + T^{-1/2}\Sigma_{t=1}^{[Tr]}\theta_t x_t \varepsilon_t - T^{-1}\Sigma_{t=1}^{[Tr]}\theta_t x_t^{'} x_t T^{1/2}(\hat{\beta} - \beta) \\
&- rT^{-1/2}\Sigma_{s=1}^{T}\theta_s x_s \varepsilon_s + rT^{-1}\Sigma_{s=1}^{T}\theta_s x_s^{'} T^{1/2}(\hat{\beta} - \beta) \\
&= T^{-1/2}\Sigma_{t=1}^{[Tr]} x_t x_t^{'}\hat{\beta} + T^{-1/2}\Sigma_{t=1}^{[Tr]}\theta_t x_t \varepsilon_t - rT^{-1/2}\Sigma_{s=1}^{T}\theta_s x_s \varepsilon_s + o_p(1) \\
&\rightarrow_d \Sigma_{xx}\tilde{Y}_i(r).
\end{aligned}$$

Thus, the out-of-sample and split sample estimators applied to the bootstrapped data will have the limiting distributions in equations (13) and (14). Averaging over i then gives the required conditional estimators.

Breiman (1996) gave a heuristic argument for why bagging weakly reduces mean square error, but in fact the standard form of bagging can increase mean square error. The calculations of Bühlmann and Yu (2002) showed this for the case of estimation with AIC model selection. See also Andreas and Stuetzle (2000) and Friedman and Hall (2007). However the alternate form of bagging, based on Rao-Blackwellization, does indeed weakly reduce risk, asymptotically.

2.3 Unmodelled Structural Change

A variant of our basic model specifies that $y_t = \beta_t' x_t + u_t$ where $\beta_0 = 0$, $\beta_t = T^{-1}\Sigma_{s=1}^{t}\eta_s$ and the η_ts are iid with mean zero, variance σ_η^2 and 2+δ finite moments and are independent of u_t. Theorem 5 gives the asymptotic distribution of the partial sum process $T^{-1/2}\Sigma_{t=1}^{[Tr]}x_t y_t$ in this case:

Theorem 5: As $T \to \infty$, the partial sum process

$$T^{-1/2}\sum_{t=1}^{[Tr]} x_t y_t \to_d \Sigma_{zz} Z(r)$$

where $Z(r) \overset{d}{=} \sigma_\eta \int_0^r W(s)ds + r\xi + U_B(r)$, $\xi \sim N(0, \sigma^2\Sigma_{xx}^{-1})$, $W(r)$ is a standard k-dimensional Brownian motion, $U_B(r)$ is a k-dimensional Brownian bridge with covariance matrix $\sigma^2\Sigma_{xx}^{-1}$, and ξ, $W(r)$ and $U_B(r)$ are all mutually independent.

Suppose that the researcher ignores the possibility of structural change, and simply uses the available estimators for forecasting. The limiting distributions of the estimators will be as in Theorems 2 and 4 with $Y(r)$ replaced by $Z(r)$ and $Y_i^*(r)$ replaced by $rZ(1) + \sigma V_i(r)$ everywhere.

2.4 Model Combination

It may also be appealing to combine forecasts made from multiple models, instead of selecting a single model (Bates and Granger (1969) and Timmermann (2006)). Recalling that $\hat{\beta}(1,g)$ denotes the parameter estimate from the model containing the variables indexed by g (with zeros in other locations), then we could estimate the parameter vector as $\Sigma_g w(g)\hat{\beta}(1,g)$, where Σ_g denotes the sum over all 2^k possible models and the weights sum to 1. As examples of weighting schemes, we could set $w(g) = \frac{\exp(-AIC(g)/2)}{\Sigma_{g*}\exp(-AIC(g^*)/2)}$ (Buckland, Burnham, and Augustin, 1997) or $w_i = \frac{\exp(-\hat{\sigma}^2(g))}{\Sigma_{g*}\exp(-\hat{\sigma}^2(g^*))}$ where $AIC(g)$ and $\hat{\sigma}^2(g)$ denote the Akaike Information Criterion and out-of-sample mean of squared residuals in the model indexed by g. Alternatively, to do a combination version of the split-sample scheme, we could estimate the parameter vector as $\Sigma_g w(g)\hat{\beta}^*(\pi,g)$ where $w(g) = \frac{\exp(-AIC(\pi,g)/2)}{\Sigma_g \exp(-AIC(\pi,g)/2)}$ and $AIC(\pi,g)$ denotes the Akaike Information Criterion for the model indexed by g computed only over the first fraction π of the sample.

Theorem 6: If the parameter vector is estimated by $\Sigma_g w(g)\hat{\beta}(1,g)$ then in large samples, the distributions of the alternative parameter estimates will be:

$$\sigma^2 E\{\Sigma_g Y(1,g)w(g))\}$$

where

$$w(g) \propto \exp(-[Y(1,g)'H(g)\Sigma_{xx}H(g)Y(1,g) - 2Y(1,g)'H(g)\Sigma_{xx}Y(1) + 2n(g)\sigma^2]/2)$$

or

$$w(g) \propto \exp(-[\int_\pi^1 \frac{Y(r,g)'}{r}H(g)\Sigma_{xx}H(g)\frac{Y(r,g)}{r}dr - 2\int_\pi^1 \frac{Y(r,g)'}{r}H(g)\Sigma_{xx}dY(r)])$$

for exponential AIC and mean square prediction error weights, respectively. Meanwhile, if the parameter vector is instead estimated by $\Sigma_g w(g)\hat{\beta}^*(\pi,g)$ with exponential AIC weights, then in large samples, the distributions of the alternative parameter estimates will be:

$$\sigma^2 E\{\Sigma_{g^*} \frac{Y(1,g) - Y(\pi,g)}{1-\pi} w(g)\}$$

where

$$w(g) \propto \exp(-[\frac{1}{\pi} Y(\pi,g)' H(g) \Sigma_{xx} H(g) Y(\pi,g) - \frac{2}{\pi} Y(\pi,g)' H(g) \Sigma_{xx} Y(\pi) + 2n(g)\sigma^2]/2)$$

The standard bagging step can be added to any of these methods for forecast combination and the resulting asymptotic NMSPEs in the ith of L bootstrap samples are also given by Theorem 6, except with $Y(.)$ and $Y(.,g)$ replaced by $Y_i^*(.)$ and $Y_i^*(.,g)$ everywhere. Or the alternate bagging step can be added, and Theorem 6 would still apply, except with $Y(.)$ and $Y(.,g)$ replaced by $\tilde{Y}_i(.)$ and $\tilde{Y}_i(.,g)$.

An alternative and more standard way to obtain combination weights for the out-of-sample forecasting scheme would be to weight the forecasts by the inverse mean square error (Bates and Granger (1969) and Timmermann (2006)). Under our local asymptotics, this will give each model equal weight in large samples.

2.5 Numerical Work

Given the expressions in Theorems 2 and 4, we can simulate the asymptotic NMSPE for different choices of the localization parameter b and the number of potential predictors k. None of the methods gives the lowest asymptotic NMSPE uniformly in b. Always using the big model is minmax, but is not necessarily the best choice with other loss functions. In all cases, bagging is implemented using 100 bootstrap replications, the out-of-sample and split-sample methods both set $\pi = 0.5$, and we set $\Sigma_{xx} = I_k$. The asymptotic NMSPEs are all symmetric in b and are consequently shown only for non-negative b.

Figure 1 plots the root of the asymptotic NMSPE of the in-sample, out-of-sample and split-sample methods, both with and without bagging for the case $k = 1$ against b. For the out-of-sample and split-sample procedures, results are also shown using the alternative form of bagging described in subsection 2.2: the asymptotic distributions consist of equations (13) and (14), averaged over i.

Among the methods *without* bagging, selecting the model in-sample by AIC does better than the out-of-sample scheme for most values of b, which in turn dominates the split-sample method. But bootstrap aggregation changes things. Bagging helps with the in-sample method for some but not all values of b—this was also found by Bühlmann and Yu (2002). It makes a more dramatic difference for the out-of-sample and split-sample methods. For these, either form of bagging reduces the asymptotic NMSPE for all values of b, and makes the out-of-sample and split-sample methods much more competitive. The fact that bagging improves these methods uniformly in b is just a numerical result for the standard form of bagging, but it is also a theoretical result for the alternative form of bagging. Neither the standard nor alternative form of bagging dominates the other in terms of local asymptotic NMSPE.

Among all the prediction methods represented in Figure 1, which one the researcher would ultimately would want to use depends on b, which is in turn not consistently estimable. But the split-sample and out-of-sample methods do best for many values of the localization parameter, as long as the bagging step is included. Indeed, for all b, the best forecast is some method using bagging.

We next consider the case where the number of potential predictors k is larger, but only one parameter actually takes on a nonzero value (of course the researcher does not know this). Without loss of generality, we let the nonzero element of b be the first element and so specify that $b = (b_1, 0, ...0)'$. Figure 2 plots the root asymptotic NMSPE for $k = 3$ against b_1 for in-sample, out-of-sample and split sample methods both without and with bagging. The positive-part James-Stein estimator is also included. The split-sample method with either the standard or alternative form of bagging compares very favorably with the other alternatives.

We finally consider the case where b has k elements and we do a grid search over $\bar{k}$ of these elements, setting the remaining elements to zero. As this is done by grid search, it is only feasible for $\bar{k} = 1, 2$. In Table 1, we list the cases in which one method dominates another one uniformly over the nonzero elements of b in terms of local asymptotic NMSPE for various pairs of possible forecasting methods. We find that in all cases, the, out-of-sample forecasts with bagging (either standard or alternate form) dominate those without. For the standard form of bagging, this is a numerical result, but for the alternate form it is a theoretical one, as discussed above. In this sense, one should never use the conventional out-of-sample forecasting methodology without bagging. Also, the split-sample forecasts with bagging (either standard or alternate form) dominate both the out-of-sample forecasts without bagging and the split sample forecasts without bagging.

In Table 1, if $k \geq 5$ and $\bar{k} = 1$ then the split-sample scheme with either form of bagging dominates in-sample forecasting (with or without bagging), the maximum-likelihood estimator and the James-Stein estimator. Thus it seems that the split-sample forecasting scheme with bagging does best if the model is sparse—there are multiple coefficients, most of which are equal to zero. The out-of-sample scheme with the alternate form of bagging dominates in-sample forecasting (with ot without bagging) and the maximum-likelihood estimator if $k \geq 4$ and $\bar{k} = 1$.

Figure 3 plots the root asymptotic NMSPE for $k = 1$ against b for the in-sample, out-of-sample and split sample forecast combination methods, without bagging, with the standard form of bagging, and with the alternate form of bagging. These are based on simulating the distributions in Theorem 6. The combination forecasts are generally better than forecasts based on selecting an individual model. Nonetheless, with combined forecasts as with individual forecasts, in the absence of a bagging step, using in-sample AIC weights does best for most values of b. Adding in a bagging step allows better predictions to be made. The alternate form of bagging reduces the asymptotic NMSPEs of the combination forecasts with out-of-sample or split-sample weights uniformly in b. Once a bagging step is added in, there is no clear winner among the in-sample, out-of-sample and split sample forecast combination methods.

3 Monte Carlo Simulations

The results in the previous section are based on a local asymptotic sequence. The motivation for this is to provide a good approximation to the finite sample properties of different forecasting methods while retaining some assurance that they are not an artifact of a specific simulation design. As some check that the local asymptotics are indeed relevant to small samples, we did a small simulation consisting of equation (1) with standard normal errors, independent standard normal regressors, a sample size $T = 100$, and different values of k. In each simulation we drew $T + 1$ observations on y_t and x_t, used the first T for model selection and parameter estimation according to one of the methods discussed above. Then given x_{T+1}, we worked out the prediction for y_{T+1}, and computed the mean square prediction error (MSPE).

Figure 4 plots the simulated root normalized mean square prediction errors ($\sqrt{T * (MSPE - 1)}$) against β for $k = 1$. Figure 5 repeats this for $k = 3$ where $\beta = (\beta_1, 0, 0)'$ against β_1. In our simulations, we include the alternative form of the out-of-sample and split sample bagging forecasts in equations (13) and (14). The Monte-Carlo simulation also included results from out-of-sample and split sample bagging forecasts using the resampling scheme in equation (15), but this turned out to give very similar results, and so these latter results are not reported. Our simulation also included results using leave-one-out cross-validation, but these were not surprisingly very close to the in-sample fit with the AIC, and so are again omitted.

Figures 4 and 5 give very similar conclusions to the local asymptotic calculations reported in Figures 1 and 2. Without bootstrap aggregation, the in-sample scheme generally gives the best forecasts, followed by out-of-sample, with the partitioned sample doing the worst. Bootstrap aggregation switches this around.

4 Incorporating Serial Correlation

4.1 Multi-step Forecasting

We can extend the results in section 2 to an h-period-ahead forecasting model with possibly serially correlated errors in which

$$y_{t+h} = \beta' x_t + u_t$$

where u_t is h-dependent with mean 0, finite variance σ^2 and 2+δ finite moments for some $\delta > 0$, x_t is a $k \times 1$ stationary vector such that $E(x_t x_t') = \Sigma_{xx}$, and h and k are fixed. Let $\frac{\omega^2}{2\pi}$ be the zero-frequency spectral density of u_t. We observe data on $\{x_t, y_{t+h}\}$ for $t = 1, ..T - h$. The slope coefficient is again local to zero as $\beta = T^{-1/2}b$. The goal is to predict y_{T+h} given x_T. In large samples, the distribution of the parameter estimates without bagging is exactly as in Theorem 2, except with $Y(r)$ replaced by $rY_\omega + \frac{\omega}{\sigma} U_B(r)$ where $Y_w \sim N(b, \omega^2 \Sigma_{xx}^{-1})$ and $U_B(r)$ is, as before, a Brownian bridge with covariance matrix $\sigma^2 \Sigma_{xx}^{-1}$ and is independent of Y_ω.

Consider bootstrap samples in which we resample from the pairs $\{x_t, y_{t+h}\}$ without replacement. Let $\{x_t^*(i), y_{t+h}^*(i)\}$ be the ith bootstrap sample, $t = 1, ...T - h$. Then,

$$T^{-1/2} \Sigma_{t=1}^{[Tr]-h} x_t^*(i) y_{t+h}^*(i) \to_d \Sigma_{xx}\{rY_\omega + V_i(r)\} \equiv \Sigma_{xx} Y_{i,\omega}^*(r)$$

where $V_i(r)$ is, as before, a Brownian motion with covariance matrix $\sigma^2 \Sigma_{xx}^{-1}$. Thus, in the multi-step case, Theorem 4 will also go through, except with $Y_i^*(r)$ replaced by $Y_{i,\omega}^*(r)$ everywhere.

4.2 Vector Autoregressions

The results in section 2 can also be extended to cover a vector autoregression (VAR). The VAR is of course a widely-used tool for forecasting with multivariate series that may have persistence. Consider a stationary VAR in which y_t is a px1 vector such that:

$$y_t = \mu + A_1 y_{t-1} ... + A_k y_{t-k} + u_t$$

where u_t is iid with mean zero and variance-covariance matrix Σ. The VAR can be written in the form:

$$y_t = Bx_t + u_t$$

where x_t is a $(pk+1)$x1 vector and B is a px$(pk+1)$ matrix of parameters. Suppose that we have a set of candidate models, each of which consists of zero restrictions on B. We can estimate the unrestricted model (by OLS or by the positive-part James-Stein estimator). Alternatively, we can select among the possible models by the AIC, OOS or SS methods, and then use the chosen model for estimation. Suppose that $B = CT^{-1/2}$.

All estimators will depend on $T^{-1}\Sigma_{t=1}^{[Tr]} x_t x_t'$ which converges in probability to $r\Omega_{xx}$ where $\Omega_{xx} = E(x_t x_t') = [0_{p\text{x}1} \; I_k \otimes \Sigma] + o_p(1)$ and on

$$T^{-1/2}\Sigma_{t=1}^{[Tr]} y_t x_t' \to_d [rC + B(r)]\Omega_{xx} \equiv Y(r)\Omega_{xx}$$

where $vec(B(r))$ is a $p(pk+1)$-dimensional Brownian motion with variance-covariance matrix $\Omega_{xx}^{-1} \otimes \Sigma$. We can rewrite $Y(r)$ as

$$rC + rB(1) + B(r) - rB(1) = rY + U_B(r)$$

where $vec(Y) \sim N(vec(C), \Omega_{xx}^{-1} \otimes \Sigma)$ and $vec(U_B(r))$ is an independent Brownian bridge with variance-covariance matrix $\Omega_{xx}^{-1} \otimes \Sigma$. Thus all of these estimators, except for OOS or SS, are asymptotically functions of Y alone. But OOS and SS are functions of both Y and $U_B(r)$.

Concretely, let $Y(r, g)$ denote the px$(pk+1)$ matrix with the elements of $Y(r)$ in the locations indexed by g and zeros elsewhere. Define $n(g)$ as the number of elements in g. Let $\Omega_{xx}(g)$ denote the matrix that consists of the elements of Ω_{xx} in the rows and columns indexed by g and zeros in all other locations, and let $H(g) = \Omega_{xx}(g)^{+}\Omega_{xx}$. If $\tilde{B}$ is the estimator of B after AIC model selection then:

$$T^{1/2}\tilde{B} \to_d \sum_{g^*} Y(1,g^*)H(g^*)\mathbf{1}\{g^* = \arg\min_g [Y(1,g)H(g)\Omega_{xx}H(g)Y(1,g)' \\ -2Y(1,g)H(g)\Omega_{xx}Y(1)' + 2n(g)\log|\Sigma|]\}. \tag{16}$$

If $\tilde{B}$ is the estimator of B minimizing recursive the determinant of the out-of-sample error variance covariance matrix starting a fraction π of the way through the sample, then:

$$T^{1/2}\tilde{B} \to_d \sum_{g^*} Y(1,g^*)H(g^*)\mathbf{1}\{g^* = \arg\min_g [\int_\pi^1 \frac{Y(r,g)}{r}H(g)\Omega_{xx}H(g)\frac{Y(r,g)'}{r}dr \\ -2\int_\pi^1 \frac{Y(r,g)}{r}H(g)\Omega_{xx}dY(r)']\}. \tag{17}$$

And if $\tilde{B}$ is the estimator of B using the split-sample method, then:

$$T^{1/2}\tilde{B} \to_d \sum_{g^*} \frac{Y(1,g^*) - Y(\pi,g^*)}{1-\pi}H(g^*)\mathbf{1}\{g^* = \arg\min_g [\frac{1}{\pi}Y(\pi,g)H(g)\Omega_{xx}H(g)Y(\pi,g)' \\ -\frac{2}{\pi}Y(\pi,g)H(g)\Omega_{xx}Y(\pi)' + 2n(g)\log|\Sigma|]\}. \tag{18}$$

Bagging provides another alternative. In this context, bagging consists of resampling $\{x_t, y_t\}$ with replacement, yielding a series $\{x_t^*(i), y_t^*(i)\}$ in the ith bootstrap replication. Each estimator can then be applied, with the resulting estimates averaged across bootstrap replications. In this case, we have:

$$T^{-1/2}\Sigma_{t=1}^{[Tr]}y_t^*(i)x_t^*(i)' \to_d [rY + V_i(r)]\Omega_{xx} \equiv Y_i^*(r)\Omega_{xx}$$

where $V_i(r)$ is a Brownian motion with variance-covariance matrix $\Omega_{xx}^{-1}\otimes\Sigma$ that is independent of Y, while $T^{-1/2}\Sigma_{t=1}^{[Tr]}x_t^*(i)x_t^*(i)' \to_p r\Omega_{xx}$. In the ith bootstrap sample ($i = 1, \ldots, L$), in large samples, the asymptotic distributions of the AIC, OOS and SS parameter estimates including a bagging step are given by equations (16), (17) and (18), except replacing $Y(r)$ by $Y_i^*(r)$ everywhere. So the estimates averaged across bootstrap draws will depend on Y alone.

5 Conclusion

When forecasting using k potential predictors, each of which has a coefficient that is local to zero, there are several competing methods, none of which is most accurate uniformly in the localization parameter. Optimizing the in-sample fit, as measured by the Akaike information criterion, generally does better than out-of-sample or split-sample methods. However, adding in a bootstrap aggregation step changes this. For important ranges of the localization parameter, the approach of splitting the sample into model selection and parameter estimation pieces, coupled with a bootstrap aggregation step, also performs well. Our representation results highlight a noise-reduction aspect of bagging, and also leads to alternative procedures that are similar to bagging and dominate the out-of-sample and split-sample methods.

Appendix: Proof of Theorems

Proof of Theorem 1: We have

$$\begin{aligned}
T^{-1/2}\sum_{t=1}^{[Tr]} x_t y_t &= T^{-1/2}\sum_{t=1}^{[Tr]} x_t x_t'\beta + T^{-1/2}\sum_{t=1}^{[Tr]} x_t u_t \\
&= T^{-1}\sum_{t=1}^{[Tr]} x_t x_t' b + T^{-1/2}\sum_{t=1}^{[Tr]} x_t u_t \\
&\to_d \Sigma_{xx}(rb + B(r)),
\end{aligned}$$

where $B(r)$ denotes a Brownian motion with covariance matrix $\sigma^2\Sigma_{xx}^{-1}$. Let $Y(r) = rb + B(r)$, and let $Y = Y(1)$ which is $N(b, \sigma^2\Sigma_{xx}^{-1})$. Define $U_B = B(r) - rB(1)$. Then

$$\begin{aligned}
Y(r) &= rb + B(r) \\
&= rb + rB(1) + B(r) - rB(1) = rY + U_B(r).
\end{aligned}$$

■

Proof of Theorem 2: Let $\hat\beta(g)$ denote the vector of OLS coefficient estimates corresponding to the predictors indexed by g with zeros in all other locations. Equations (3) and (4) immediately follow because $T^{1/2}\hat\beta \to_d Y$. The in-sample AIC is:

$$\begin{gathered}
\ln(T^{-1}\textstyle\sum_{t=1}^{T}(y_t - \hat\beta(g)'x_t)^2) + \frac{2n(g)}{T} \\
= \ln(T^{-1}\textstyle\sum_{t=1}^{T} y_t^2) + \ln(1 + \frac{1}{T^{-1}\sum_{t=1}^{T} y_t^2}[\hat\beta(g)'T^{-1}\sum_{t=1}^{T} x_t x_t'\hat\beta(g) - 2\hat\beta(g)'T^{-1}\sum_{t=1}^{T} x_t y_t]) + \frac{2n(g)}{T} \\
= \ln(T^{-1}\textstyle\sum_{t=1}^{T} y_t^2) + \frac{1}{T^{-1}\sum_{t=1}^{T} y_t^2}[\hat\beta(g)'T^{-1}\sum_{t=1}^{T} x_t x_t'\hat\beta(g) - 2\hat\beta(g)'T^{-1}\sum_{t=1}^{T} x_t y_t] \\
+ \frac{2n(g)}{T} + o_p(T^{-1}) = \ln(T^{-1}\textstyle\sum_{t=1}^{T} y_t^2) + \frac{1}{T\sigma^2}\{Y(1,g)'H(g)\Sigma_{xx}H(g)Y(1,g) - \\
2Y(1,g)'H(g)\Sigma_{xx}Y(1) + 2n(g)\sigma^2\} + o_p(T^{-1})
\end{gathered}$$

which is asymptotically the same, up to the same affine transformation across all models, as

$$Y(1,g)'H(g)\Sigma_{xx}H(g)Y(1,g) - 2Y(1,g)'H(g)\Sigma_{xx}Y(1) + 2n(g)\sigma^2$$

proving (5).

Let $\hat{\beta}(r,g)$ denote the vector of OLS coefficient estimates corresponding to the predictors indexed by g with zeros in all other locations, when the estimation is conducted only on the first fraction r of the sample. Then $T^{1/2}\hat{\beta}(r,g) \rightarrow_d \frac{1}{r}H(g)Y(r,g)$. The recursive pseudo-out-of-sample recursive mean square prediction error, starting a fraction π of the way through the sample is:

$$\begin{gathered}\tfrac{1}{T(1-\pi)}\textstyle\sum_{t=[T\pi]+1}^{T}(y_t-\hat{\beta}(t/T,g)'x_t)^2 = \\ \tfrac{1}{T(1-\pi)}\textstyle\sum_{t=[T\pi]+1}^{T} y_t^2 + \tfrac{1}{T(1-\pi)}\sum_{t=[T\pi]+1}^{T}\hat{\beta}(t/T,g)'x_t x_t'\hat{\beta}(t/T,g) - \tfrac{2}{T(1-\pi)}\sum_{t=[T\pi]+1}^{T}\hat{\beta}(t/T,g)'x_t y_t \\ = \tfrac{1}{T(1-\pi)}[\textstyle\int_\pi^1 \frac{Y(r,g)'}{r}H(g)\Sigma_{xx}H(g)\frac{Y(r,g)}{r}dr - 2\int_\pi^1 \frac{Y(r,g)'}{r}H(g)\Sigma_{xx}dY(r)] + o_p(T^{-1})\end{gathered}$$

neglecting a term that is constant across models, which proves (6).

Let $\hat{\beta}^*(\pi,g)$ denote the vector of OLS coefficient estimates corresponding to the predictors indexed by g with zeros in all other locations, when estimation is conducted only on the sample *excluding* the first fraction π. Then $T^{1/2}\hat{\beta}^*(\pi,g) \rightarrow_d H(g)\frac{Y(1,g)-Y(\pi,g)}{1-\pi}$. The AIC estimated over the first fraction π of the sample is

$$\begin{gathered}\ln(\tfrac{1}{\pi T}\textstyle\sum_{t=1}^{[\pi T]}(y_t-\hat{\beta}(\pi,g)'x_t)^2) + \frac{2n(g)}{\pi T} \\ = \ln(\tfrac{1}{\pi T}\textstyle\sum_{t=1}^{[\pi T]} y_t^2) + \ln(1+\frac{1}{\frac{1}{\pi T}\sum_{t=1}^{[\pi T]} y_t^2}[\frac{1}{\pi T}\hat{\beta}(\pi,g)'\sum_{t=1}^{[\pi T]} x_t x_t'\hat{\beta}(\pi,g) - 2\hat{\beta}(\pi,g)'\frac{1}{\pi T}\sum_{t=1}^{[\pi T]} x_t y_t]) \\ +\tfrac{2n(g)}{\pi T} \\ = \ln(\tfrac{1}{\pi T}\textstyle\sum_{t=1}^{[\pi T]} y_t^2) + \frac{1}{\frac{1}{\pi T}\sum_{t=1}^{[\pi T]} y_t^2}[\frac{1}{\pi T}\hat{\beta}(\pi,g)'\sum_{t=1}^{[\pi T]} x_t x_t'\hat{\beta}(\pi,g) - 2\hat{\beta}(\pi,g)'\frac{1}{\pi T}\sum_{t=1}^{[\pi T]} x_t y_t]) \\ +\tfrac{2n(g)}{\pi T} + o_p(T^{-1}) \\ = \ln(\tfrac{1}{\pi T}\textstyle\sum_{t=1}^{[\pi T]} y_t^2) + \frac{1}{\pi T\sigma^2}\{\pi\frac{Y(\pi,g)}{\pi}'H(g)\Sigma_{xx}H(g)\frac{Y(\pi,g)}{\pi} - 2\frac{Y(\pi,g)}{\pi}'H(g)\Sigma_{xx}Y(\pi) + 2n(g)\sigma^2\} + \\ o_p(T^{-1})\end{gathered}$$

which is asymptotically the same, up to the same affine transformation across all models, as

$$\tfrac{1}{\pi}Y(\pi,g)'H(g)\Sigma_{xx}H(g)Y(\pi,g) - \tfrac{2}{\pi}Y(\pi,g)'H(g)\Sigma_{xx}Y(\pi) + 2n(g)\sigma^2$$

proving (7). ■

Proof of Theorem 3: Let $\{x_t^*(i), y_t^*(i)\}$ be the ith bootstrap sample and let $u_t^*(i) = y_t^*(i) - \beta' x_t^*(i)$, $t = 1, ...T$. From Theorem 2.2 of Park (2002), $T^{-1/2}\sum_{t=1}^{[Tr]}(x_t^*(i)u_t^*(i) - T^{-1}\sum_{t=1}^{T} x_t u_t) \to_d \Sigma_{xx} V_i(r)$. Consequently $T^{-1/2}\Sigma_{t=1}^{[Tr]} x_t^*(i) y_t^*(i) \to_d \Sigma_{xx}(rY + V_i(r))$. ■

The proofs of Theorem 4 and 6 involve exactly the same calculations as in Theorem 2 and are hence omitted.

Proof of Theorem 5: We have

$$\begin{aligned}
T^{-1/2}\sum_{t=1}^{[Tr]} x_t y_t &= T^{-1/2}\sum_{t=1}^{[Tr]} x_t x_t' \beta_t + T^{-1/2}\sum_{t=1}^{[Tr]} x_t u_t \\
&= T^{-3/2}\sum_{t=1}^{[Tr]} x_t x_t' \Sigma_{s=1}^{t} \eta_s + T^{-1/2}\sum_{t=1}^{[Tr]} x_t u_t \\
&\to_d \Sigma_{xx}\{\sigma_\eta \int_0^r V(s)ds + \sigma B(r)\} = \Sigma_{xx}\{\sigma_\eta \int_0^r V(s)ds + r\xi + \sigma U_B(r)\}.
\end{aligned}$$

where $B(r)$ is a Brownian motion with covariance matrix Σ_{xx}^{-1}. ■

References

Akaike, H. (1974): "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Andreas, B., and W. Stuetzle (2000): "Bagging does not always Decrease Mean Squared Error," mimeo.

Ashley, R., C. W. Granger, and R. Schmalensee (1980): "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica*, 48, 1149–1167.

Baranchik, A. J. (1964): "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," Stanford University Department of Statistics Technical Report 51.

Bates, J. M., and C. W. Granger (1969): "The combination of forecasts," *Operations Research Quarterly*, 20, 451–468.

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

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.

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

Clark, T. E. (2004): "Can Out-of-Sample Forecast Comparisons help Prevent Overfitting?," *Journal of Forecasting*, 23, 115–139.

Friedman, J. H., and P. Hall (2007): "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference*, 137, 669–683.

Hansen, P. R. (2009): "In-Sample Fit and Out-of-Sample Fit: Their Joint Distribution and its Implications for Model Selection," mimeo.

Inoue, A., and L. Kilian (2004): "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?," *Econometric Reviews*, 23, 371–402.

——— (2006): "On the Selection of Forecasting Models," *Journal of Econometrics*, 130, 273–306.

James, W., and C. Stein (1961): "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379.

Leeb, H., and B. M. Pötscher (2005): "Model selection and inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

Mallows, C. L. (1973): "Some Comments on Cp," *Technometrics*, 15, 661–675.

Park, J. (2002): "An Invariance Principle for Sieve Bootstrap in Time Series," *Econometric Theory*, 18, 469–490.

Stock, J. H., and M. W. Watson (2012): "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business and Economic Statistics*, 30, 481–493.

Timmermann, A. (2006): "Forecast Combination," in *Handbook of Economic Forecasting*, ed. by C. W. Granger, G. Elliott, and A. Timmermann, Amsterdam. North Holland.

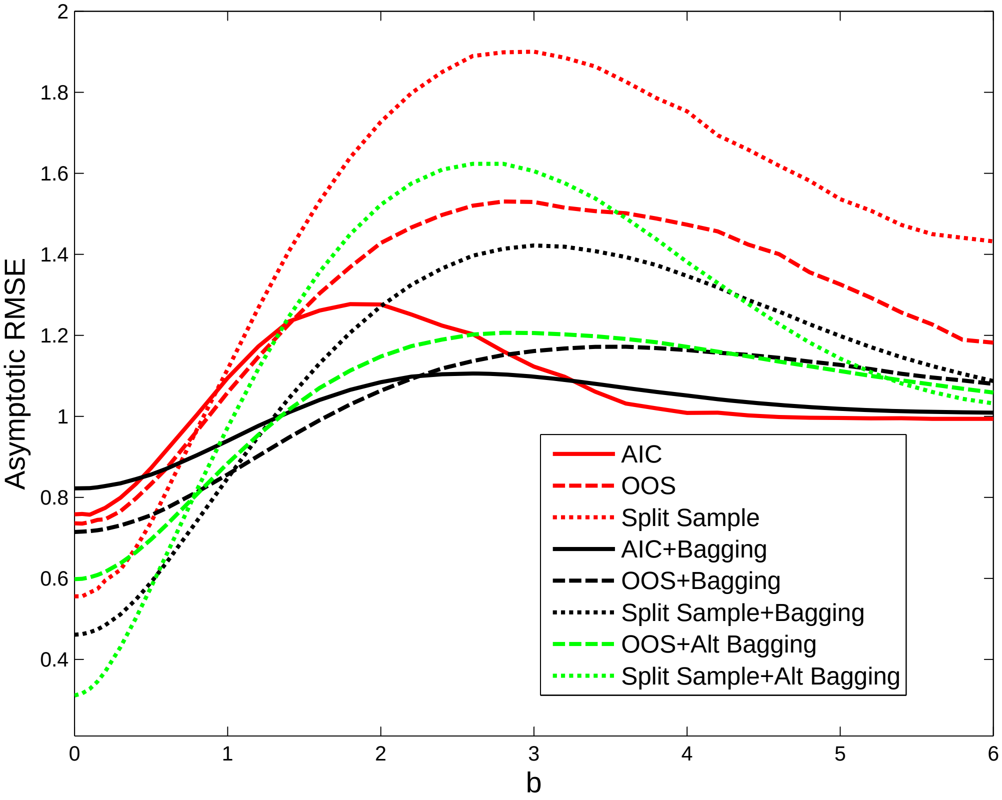
Van Der Vaart, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Table 1: Dominance Relations in Local Asymptotic NMSPE

k	1	2	3	4	5	6	2	3	4	5	6
$\bar{k}$	1	1	1	1	1	1	2	2	2	2	2
MLE v. AICB	–	AICB	AICB	AICB	AICB	AICB	–	–	AICB	AICB	AICB
MLE v. OOSB	–	OOSB	OOSB	OOSB	OOSB	OOSB	–	–	OOSB	OOSB	OOSB
MLE v. SSB	–	–	SSB	SSB	SSB	SSB	–	–	–	SSB	SSB
MLE v. OOSB2	–	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	–	–	OOSB2	OOSB2	OOSB2
MLE v. SSB2	–	–	–	SSB2	SSB2	SSB2	–	–	–	–	SSB2
AIC v. AICB	–	–	–	–	–	–	–	–	–	–	–
JS v OOSB	NA	NA	OOSB	–	–	–	NA	–	–	–	–
JS v SSB	NA	NA	SSB	SSB	SSB	SSB	NA	–	–	–	SSB
JS v OOSB2	NA	NA	OOSB2	OOSB2	–	–	NA	OOSB2	–	–	–
JS v SSB2	NA	NA	–	SSB2	SSB2	SSB2	NA	–	–	–	–
AIC v. OOSB	–	–	–	–	–	OOSB	–	–	–	–	–
AIC v. SSB	–	–	–	–	SSB	SSB	–	–	–	–	–
AIC v. OOSB2	–	–	OOSB2	OOSB2	OOSB2	OOSB2	–	–	–	–	OOSB2
AIC v. SSB2	–	–	–	–	SSB2	SSB2	–	–	–	–	–
OOS v. AICB	–	–	–	–	–	–	–	–	–	–	–
OOS v. OOSB	OOSB	OOSB	OOSB	OOSB	OOSB	OOSB	OOSB	OOSB	OOSB	OOSB	OOSB
OOS v. SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB
OOS v. OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2
OOS v. SSB2	–	SSB2	SSB2	SSB2	SSB2	SSB2	–	SSB2	SSB2	SSB2	SSB2
SS v. AICB	–	–	–	–	–	–	–	–	–	–	–
SS v. OOSB	–	–	–	–	–	–	–	–	–	–	–
SS v. SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB
SS v. OOSB2	–	–	–	–	–	–	–	–	–	–	–
SS v. SSB2	SSB2	SSB2	SSB2	SSB2	SSB2	SSB2	SSB2	SSB2	SSB2	SSB2	SSB2
AICB v. OOSB	–	–	OOSB	OOSB	OOSB	OOSB	–	–	–	OOSB	OOSB
AICB v. SSB	–	–	–	SSB	SSB	SSB	–	–	–	–	SSB
AICB v. OOSB2	–	OOSB2	OOSB2	OOSB2	OOSB2	OOSB2	–	–	OOSB2	OOSB2	OOSB2
AICB v. SSB2	–	–	–	–	SSB2	SSB2	–	–	–	–	–
OOSB v. SSB	–	–	–	–	SSB	SSB	–	–	–	–	–
OOSB v. OOSB2	–	–	OOSB2	OOSB2	OOSB2	OOSB2	–	–	–	OOSB2	OOSB2
OOSB v. SSB2	–	–	–	–	–	SSB2	–	–	–	–	–
SSB v. OOSB2	–	–	–	–	–	–	–	–	–	–	–
SSB v. SSB2	–	–	–	––	–	–	–	–	–	–	–
OSSB2 v. SSB2	–	–	–	–	–	–	–	–	–	–	–

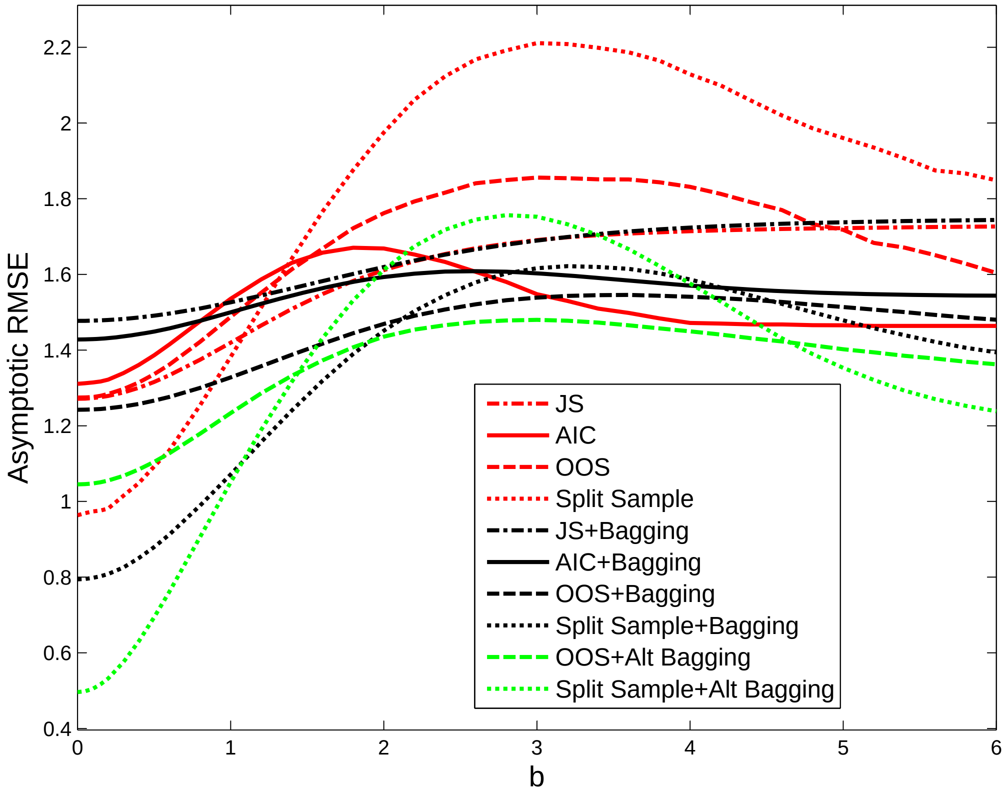
This table reports comparisons of local asymptotic RMSPE between pairs of methods: maximum likelihood (MLE), positive-part James-Stein estimator (JS) applicable if $k > 2$, in-sample with AIC (AIC), the counterpart with bagging (AICB), out-of-sample (OOS), the counterpart with standard/alternate bagging (OOSB/OOSB2), the split-sample method (SS) and the counterpart with standard/alternate bagging (SSB/SSB2). Results are shown for different numbers of predictors k. For each pairwise comparison, the table lists which method is uniformly dominant when only $\bar{k}$ of the predictors are in fact nonzero. If neither is dominant, then the entry in the table is "–".

Fig. 1: Local Asymptotic Root Normalized Mean Square Prediction Errors ($k = 1$)



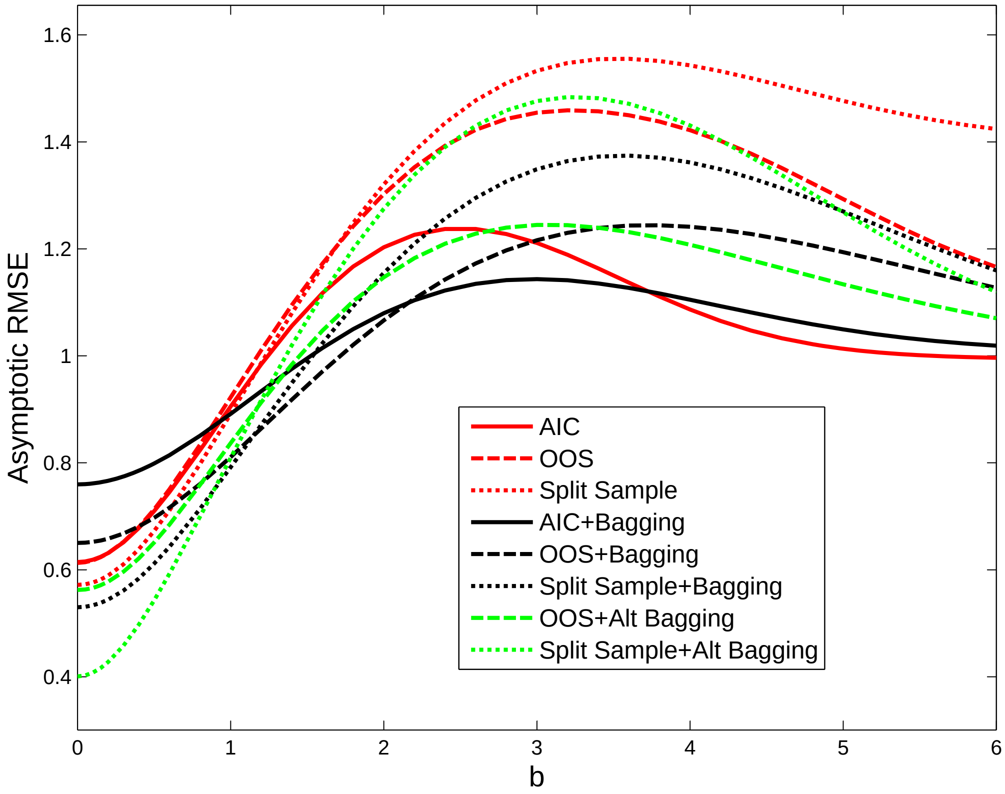
Notes: These are the simulated root normalized mean square prediction errors given by the square root of the expressions in equations (5)- (12), plotted against b.

Fig. 2: Local Asymptotic Root Normalized Mean Square Prediction Errors ($k = 3$)



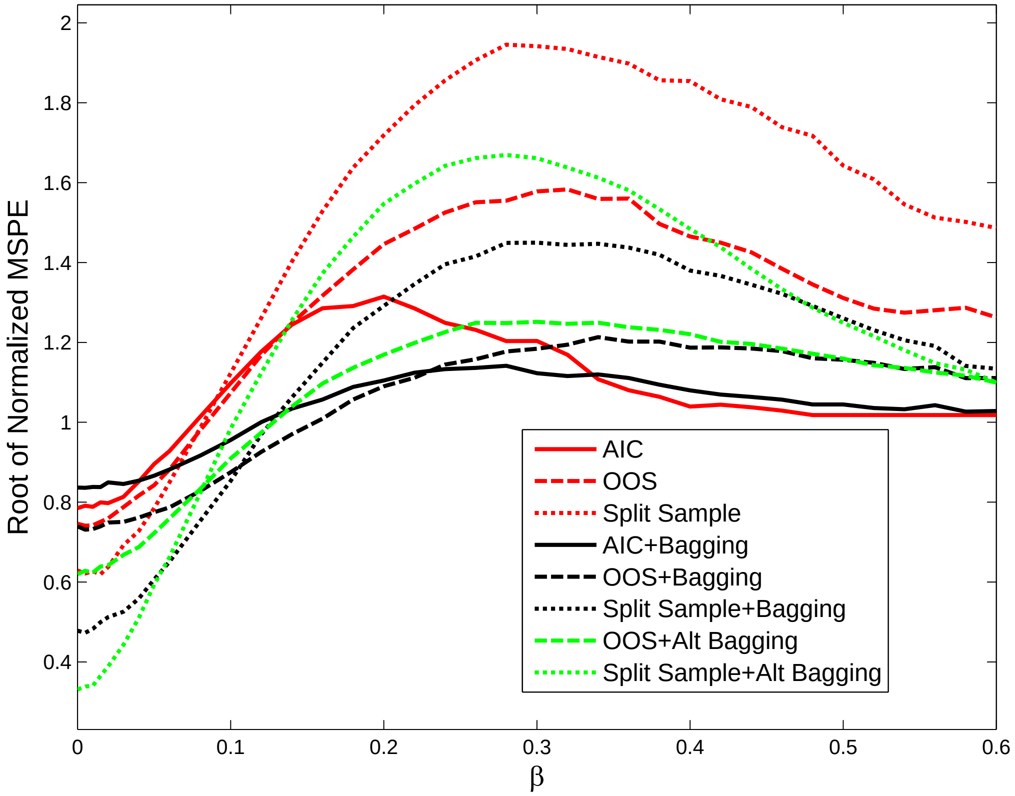
Notes: These are the simulated root normalized mean square prediction errors given by the square root of the expressions in equations (5)- (12), plotted against b.

Fig. 3: Local Asymptotic Root Normalized Mean Square Prediction Errors: Combination Forecasts ($k = 1$)



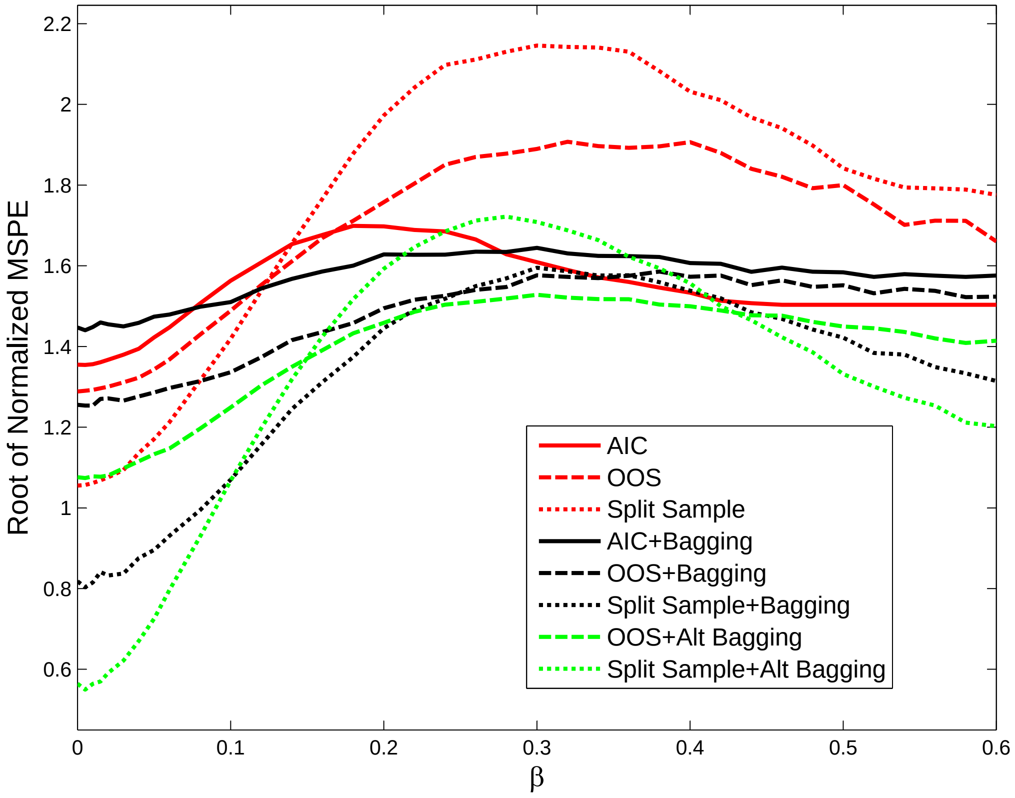
Notes: These are the simulated root normalized mean square prediction errors given by the square root of the expressions in Theorem 6, plotted against b. The forecast combination methods using exponential AIC, exponential out-of-sample, or exponential split-sample weighting schemes, as described in subsection 2.4.

Fig. 4: Root Normalized Mean Square Prediction Errors ($k = 1$)



Notes: These are the simulated root normalized mean square prediction errors from the Monte-Carlo simulation described in the text. The sample size is $T = 100$, there is one predictor, and the root of the normalized mean square prediction errors is plotted against β.

Fig. 5: Root Normalized Mean Square Prediction Errors ($k = 3$)



Notes: These are the simulated root normalized mean square prediction errors from the Monte-Carlo simulation described in the text. The sample size is $T = 100$, there are three possible predictors, and the root of the normalized mean square prediction errors is plotted against β_1 (the other coefficents are equal to zero.