Forecasting with Model Uncertainty: Representations and Risk Reduction*

Keisuke Hirano† Jonathan H. Wright‡

First version: October 14, 2013
This version: April 14, 2015

Abstract

We consider forecasting with uncertainty about the choice of predictor variables. The researcher wants to select a model, estimate the parameters, and use the parameter estimates for forecasting. We investigate the distributional properties of a number of different schemes for model choice and parameter estimation: in-sample model selection using the Akaike information criterion, out-of-sample model selection, and splitting the data into subsamples for model selection and parameter estimation. Using a weak-predictor local asymptotic scheme, we provide a representation result that facilitates comparison of the distributional properties of the procedures and their associated forecast risks. We develop a simulation procedure that improves the accuracy of the out-of-sample and split-sample methods uniformly over the local parameter space. We also examine how bootstrap aggregation (bagging) affects the local asymptotic risk of the estimators and their associated forecasts. Numerically, we find that for many values of the local parameter, the out-of-sample and split-sample schemes perform poorly if implemented in the conventional way. But they perform well, if implemented in conjunction with our risk-reduction method or bagging.

*We are grateful to Don Andrews, Marine Carrasco, Russell Davidson, Gary Chamberlain, Sylvia Gonçalves, Bruce Hansen, Serena Ng, Peter Phillips, and Jack Porter 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 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. 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 pseudo-likelihood setting with a fixed number k of potential parameters to be estimated, each of which has a coefficient that is local to zero. Selecting a forecasting model amounts to selecting a set of zero restrictions; in a regression setting, for example, this would indicate which predictors are excluded from the regression. Thus there are up to 2^k possible models among which we can 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 parts, 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 obtain asymptotic characterizations of these forecasting procedures under the local parameter sequence. A key step is to obtain an asymptotic representation of the partial sum process for the score function 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 represen-

tation 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. The influence of the ancillary noise term in the limiting asymptotic representation can be eliminated by a conditioning argument. We can implement this noise reduction via a simulation-and-averaging scheme; doing this is shown to uniformly improve the out-of-sample and split-sample methods asymptotically for a wide variety of loss functions.

This method is related to bootstrap aggregating (bagging) (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. Bagging has a smoothing effect that alters the risk properties of estimators, but the averaging over bootstrap draws 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[1].

We then numerically compare the various procedures, both in terms of their local asymptotic risk, and their finite-sample performance. In their standard forms there is no unambiguous rank ordering among the in-sample, out-of-sample and split-sample 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 simulation or bagging step changes the rank ordering substantially. Our simulation scheme has no effect on in-sample forecasts, but reduces the local asymptotic mean square prediction error of the out-of-sample and split-sample forecasts uniformly in the localization parameter, and the reductions are generally numerically large. Bagging can modestly reduce the local asymptotic mean square prediction error of the in-sample forecasts over some parts of the param-

[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 our simulation scheme or bootstrap aggregation step destroys this feature.

eter space, but it makes a more dramatic difference to the out-of-sample and split-sample forecasts. In our numerical work, we find no case in which bagging fails to reduce the local asymptotic mean square prediction error of out-of-sample and split-sample forecasts, although this is not guaranteed by our theoretical results.

For many values of the localization parameter, the incorporation of either our simulation scheme or bagging 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 either our simulation scheme or bagging provides the most accurate forecasts from among any of the methods considered here.

In the next section, we set up the model and introduce the various procedures we will evaluate. In Section 3, we derive asymptotic characterizations via our representation theorem for the partial sum process. Section 4 contains some extensions. Section 5 explores the asymptotic risk properties of the procedures numerically. Section 6 examines some finite-sample simulation evidence, and Section 7 concludes.

2 Pseudo-Likelihood Framework

We observe (y_t, x_t) for $t = 1, \ldots, T$ and wish to forecast y_{T+1} given knowledge of x_{T+1}. Let the pseudo log (conditional) likelihood be

$$\ell(\beta) = \sum_{t=1}^{T} \ell_t(\beta) = \sum_{t=1}^{T} f(y_t | x_t, \beta),$$

where f is a conditional density function and the parameter β is $k \times 1$. This framework could apply to cross-sectional regression of an outcome variable y_t on a $k \times 1$ vector of predictors x_t, h-step ahead forecasting regressions (where x_t are suitably lagged predictor variables), vector autoregression (where x_t contains lagged values of the vector y_t), and nonlinear regression models. There may be unmodeled dependence, subject to the large-sample distributional assumptions imposed below.

Model selection amounts to setting some elements of β to zero, and estimating the others. Thus there are up to 2^k possible models. Let $m \subset \{1, \ldots, k\}$ denote a model, with the interpretation that the elements of m indicate which coefficients of β are allowed to be nonzero. The set of possible models $\mathscr{M}$ is a subset of the power set of $\{1, \ldots, k\}$.

We consider a variety of strategies for model selection and parameter estimation. Each strategy will end up giving us an estimator for β, some elements of which will be zero. We denote this overall estimator as $\tilde{\beta}$. The strategies that we consider are:

1. **MLE** Set $\tilde{\beta}$ to the unrestricted (pseudo) maximum likelihood estimator, $\hat{\beta}$, that maximizes $\ell(\beta)$.

2. **JS** The positive-part James-Stein estimator uses the unrestricted estimate $\hat{\beta}$ and an estimate $\hat{V}$ of its asymptotic variance-covariance matrix. The JS estimator for $k > 2$ is

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

3. **Small Model** Set $\tilde{\beta} = 0$.

4. **AIC (In-sample)** For each model $m \in \mathscr{M}$, let $\hat{\beta}(m)$ denote the restricted pseudo-ML estimator that maximizes $\ell(\beta)$ subject to the zero restrictions implied by m. (Thus, $\hat{\beta}(m)$ is a $k \times 1$ vector with zeros in the locations implied by m.) Let $n(m)$ be the number of free parameters in model m. For each $m \in \mathscr{M}$ we calculate the AIC objective function

$$AIC(m) = 2\ell(\hat{\beta}(m)) - 2n(m),$$

   and choose the model m^* that maximizes $AIC(m)$. Then set $\tilde{\beta} = \hat{\beta}(m^*)$.

5. **Out of Sample** For each model m, we estimate the model recursively starting a fraction $\pi \in (0,1)$ of the way through the sample, and calculate its one-period-ahead predictive density to obtain a pseudo out-of-sample estimate of predictive performance. Let $\hat{\beta}_{1,t-1}(m)$ denote the pseudo maximum likelihood estimate for model m using observations 1 to $t-1$. For each m, we calculate

$$\sum_{t=[T\pi]+1}^{T} \ell_t(\hat{\beta}_{1,t-1}(m)).$$

   We then choose the model m that maximizes this predictive likelihood, and use the full sample for estimation of the model.[2]

[2] Several authors test for the statistical significance of differences in out-of-sample forecasting performance with one model, typically a simple benchmark, as the null (see, for example Diebold and Mariano (1995) and Hansen and Timmermann (2013)). Here we are instead thinking of selecting the model based on the point estimate of its pseudo out-of-sample predictive performance.

6. **Split-sample** For each model m, we calculate AIC using data up to a fraction π of the way through the sample:

$$AIC_{ss}(m) = 2\ell(\hat{\beta}_{1,[T\pi]}(m)) - 2n(m).$$

For $m^* = \arg\max AIC_{ss}(m)$, we use the second fraction of the sample to estimate the model parameters:

$$\tilde{\beta} = \hat{\beta}_{[T\pi]+1,T}(m^*).$$

Later in the paper, we also consider adding a bagging (bootstrap aggregation) step to the procedures described above.

Forecasts for y_{T+1} will typically depend on an estimate of β. Let $\tilde{\beta}$ be any of the estimators of β, including post-model selection estimators that set some elements of the coefficient vector to zero. We focus on obtaining limiting distributions for $\tilde{\beta}$ in a form that facilitates estimation and forecast risk comparisons.

2.1 Example: Prediction in a Regression Model

To illustrate our approach in a simple setting, we consider prediction using a standard regression model:

$$y_t = \beta' x_t + u_t, \tag{2.1}$$

where the u_t are i.i.d. with mean 0, finite variance σ^2, and $2+\delta$ finite moments for some $\delta > 0$. We assume x_t is a $k \times 1$ stationary vector that has been orthonormalized, so that $E[x_t x_t'] = I_k$. (The orthonormality of x_t is not essential for the analysis, but simplifies the notation.)

This model fits into the general pseudo-likelihood framework of Section 2, using the standard Gaussian likelihood. Then $\hat{\beta}$, the unrestricted pseudo-ML estimator of β, is the OLS estimator; and $\hat{\beta}(m)$, the restricted pseudo-ML estimator under model m, is the restricted OLS estimator using only the regressors indicated by m. Each model corresponds to some subset of the k regressors that are to be used for forecasting.

Consider forecasts of the form $\tilde{\beta}' x_{T+1}$, where $\tilde{\beta}$ can again be any of the estimators of β. The uncondi-

tional mean squared prediction error is

$$\begin{aligned} MSPE &= E\left[(y_{T+1} - \tilde{\beta}' x_{T+1})^2\right] = E\left[(u_{T+1} + (\beta - \tilde{\beta})' x_{T+1})^2\right] \\ &= \sigma^2 + E\left[(\beta - \tilde{\beta})'(\beta - \tilde{\beta})\right] + o\left(T^{-1}\right). \end{aligned} \tag{2.2}$$

The first term on the right hand side of (2.2) is the asymptotic forecast error neglecting parameter uncertainty, which is the same for all forecasts. The second term is $O(T^{-1})$ and differs across forecasting methods. We therefore normalize the mean square prediction error as:

$$NMSPE = T\left(MSPE - \sigma^2\right) = TE\left[(\beta - \tilde{\beta})'(\beta - \tilde{\beta})\right] + o(1).$$

We could also consider other measures of forecast performance, such as forecast accuracy conditional on x_{T+1}. We will develop approximations for the distribution of $\tilde{\beta}$ that can be used under a variety of loss functions.

To gain further intuition for our theoretical results in the next section, consider the special case where the u_t are i.i.d. $N(0,1)$ and the regressors are treated as fixed and satisfy $\frac{1}{T}\sum_{t=1}^{T} x_t x_t' = I_k$. Then the least squares estimator for the full set of parameters has an exact normal distribution:

$$\hat{\beta} = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t y_t \sim N\left(\beta, \sigma^2 I_k / n\right)$$

and $\hat{\beta}$ is a minimal sufficient statistic for β. If a procedure makes nontrivial use of information in the data other than that contained in $\hat{\beta}$, it is introducing an unnecessary source of randomness. In the next section we will obtain an asymptotic analog to this argument in the general pseudo-likelihood setting, and show how it applies to the various procedures we consider.

3 Local Asymptotics

In order to capture the role of parameter and model uncertainty in our analysis, the joint distribution of $\{(y_1, x_1), \ldots, (y_T, x_T)\}$ is assumed to be a triangular array with drifting parameters. Let $\{(y_1, x_1), \ldots, (y_T, x_T)\}$

have joint distribution P_T, and define the pseudo-true value of the parameter as

$$\beta_{0,T} = \arg\max_{\beta} \int \ell(\beta) dP_T.$$

We take the pseudo-true values (or functions of them) as our objects of interest. We will take limits as $T \to \infty$ under the assumption that

$$\beta_{0,T} = \frac{b}{\sqrt{T}}, \quad b \in \mathbb{R}^k.$$

This type of drifting sequence was also used by Claeskens and Hjort (2008) and Inoue and Kilian (2004) to study the large-sample properties of model selection procedures. It preserves the role of parameter uncertainty in the asymptotic approximations, unlike fixed-alternative asymptotics in which model selection can determine which coefficients are nonzero with probability approaching one. The analysis could be extended to allow some components of β to be localized away from zero, corresponding to situations where some components of β are known to be nonzero. We use $\to_d$ to denote weak convergence and $\to_p$ to denote convergence in probability under the sequence of measures $\{P_T\}_{T=1}^{\infty}$. Our results below depend crucially on the convergence properties of the partial sums of the pseudo-likelihood function. We make the following high level assumptions.

Assumption 3.1

$$T^{-1/2} \Sigma_{t=1}^{[Tr]} \frac{\partial \ell_t(\beta_{0,T})}{\partial \beta} \to_d B(r),$$

where $B(r)$ is a k-dimensional Brownian motion with covariance matrix Λ.

Assumption 3.2 *For all sequences β_T in a $T^{-1/2}$-neighborhood of zero,*

$$-T^{-1} \Sigma_{t=1}^{[Tr]} \frac{\partial^2 \ell_t(\beta_T)}{\partial \beta^2} \to_p r\Sigma.$$

These high-level assumptions would follow from conventional regularity conditions in correctly specified parametric models. In misspecified models, the assumptions require that the pseudo-true parameter sequence $\beta_{0,T}$ is related to the distribution of the data in a smooth way.

To gain intuition for the results to follow, consider the case where the parametric model with conditional likelihood $f(y_t|x_t, \beta)$ is correctly specified. Then, under standard regularity conditions, Assumptions 3.1

and 3.2 will hold with $\Lambda = \Sigma$. Furthermore, the model will be locally asymptotically normal (LAN), and possess a limit experiment representation (see for example van der Vaart 1998, Chs. 7-9). In particular, consider any estimator sequence $\tilde{\beta}$ with limiting distributions in the sense that

$$T^{1/2}\tilde{\beta} \to_d \mathscr{L}_b,$$

where the limit is taken under the drifting sequences of measures corresponding to $\beta_{0,T} = T^{-1/2}b$, and $\mathscr{L}_b$ is a law that may depend on b. Then the estimator $\tilde{\beta}$ has an asymptotic representation as a randomized estimator in a shifted normal model: if Y is a single draw from the $N(\Sigma b, \Sigma)$ distribution, and U is random variable independent of Y (with sufficiently rich support[3]), there exists an estimator $S(Y,U)$ with

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

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

We extend this type of asymptotic representation, in terms of an asymptotically sufficient component and an independent randomization, to the pseudo-likelihood setup. We do this by establishing a large-sample representation of the partial sum process for the score function that corresponds to the (Y,U) limit experiments in parametric LAN models.

From Assumptions 3.1 and 3.2, it follows that:

$$T^{-1/2}\Sigma_{t=1}^{[Tr]} \frac{\partial \ell_t(0)}{\partial \beta} \to_d B(r) + r\Sigma b =: Y(r)$$

Thus the partial sums of the score function evaluated at $\beta = 0$ converge to a Brownian motion with linear drift. By a standard argument, we can decompose this process into the sum of a normal random vector and a Brownian bridge:

[3]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.

Proposition 3.3 *Under Assumptions 3.1 and 3.2,*

$$T^{-1/2}\Sigma_{t=1}^{[Tr]}\frac{\partial \ell_t(0)}{\partial \beta} \to_d Y(r) = rY + U_B(r),$$

where $Y := Y(1) \sim N(\Sigma b, \Lambda)$*, and* $U_B(r)$ *is a* k*-dimensional Brownian bridge with covariance matrix* Λ*, where* U_B *is independent of* Y*.*

All proofs are given in Appendix A. This result decomposes the limit of the partial sums of the score function into two stochastic components, one of which depends on the local parameter b and one of which is ancillary.

Let $\Sigma(m)$ denote the $k \times k$ matrix that consists of the elements of Σ in the rows and columns indexed by m and zeros in all other locations, and let $H(m)$ denote the Moore-Penrose generalized inverse of $\Sigma(m)$. Then $T^{1/2}\hat{\beta} \to_d \Sigma^{-1}Y(1)$ and $T^{1/2}\hat{\beta}(m) \to_d H(m)Y(1,m)$, where $Y(r,m)$ denotes the $k \times 1$ vector with the elements of $Y(r)$ in the locations indexed by m and zeros elsewhere. This leads to the following asymptotic characterizations of the procedures:

Proposition 3.4 *Under Assumptions 3.1 and 3.2, we have the following limiting representations of the parameter estimation procedures:*

(i) Using unrestricted MLE:

$$T^{1/2}\hat{\beta} \to_d \Sigma^{-1}Y(1) \tag{3.1}$$

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

$$T^{1/2}\tilde{\beta} \to_d \Sigma^{-1}Y(1)\max(1 - \frac{k-2}{Y(1)'\Sigma^{-2}Y(1)}, 0) \tag{3.2}$$

(iii) Selecting the model using the AIC:

$$T^{1/2}\tilde{\beta} \to_d \sum_{m^*} H(m^*)Y(1,m^*)\mathbf{1}\{m^* = \arg\max_m [Y(1,m)'H(m)Y(1,m) - 2n(m)]\} \tag{3.3}$$

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

the sample:

$$T^{1/2}\tilde{\beta} \to_d \sum_{m^*} H(m^*)Y(1,m^*)\mathbf{1}\{m^* = \arg\max_m [2\int_\pi^1 \frac{Y(r,m)'}{r} H(m)dY(r) - \int_\pi^1 \frac{Y(r,m)'}{r} H(m)\frac{Y(r,m)}{r}dr]\} \tag{3.4}$$

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

$$T^{1/2}\tilde{\beta} \to_d \sum_{m^*} H(m^*)\frac{Y(1,m^*) - Y(\pi,m^*)}{1-\pi}\mathbf{1}\{m^* = \arg\max_m [\frac{1}{\pi}Y(\pi,m)'H(m)Y(\pi,m) - 2n(m)]\} \tag{3.5}$$

where $\sum_{m^*}$ denotes the summation over all the models in $\mathscr{M}$.

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 same asymptotic distribution as in equation (3.3). 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 pick the small model ($\beta = 0$) with probability converging to one. Part (iv) of the proposition can immediately be adapted to selecting the model minimizing out-of-sample error with a rolling estimation window, as long as the estimation window contains a fixed fraction of the sample size, but not if it instead contains a fixed number of observations as in Giacomini and White (2006).

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 (3.1) and a result very similar to equation (3.4).

3.1 Rao-Blackwellization

The estimators other than the out-of-sample and split-sample estimators can be viewed 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 (3.4) and (3.5) are functions of both Y and an independent Brownian bridge, $U_B(r)$. Thus the out-of-sample and

split-sample estimators are not shrinkage estimators asymptotically, and their representation in terms of $Y = Y(1)$ and $U = U_B$ suggests a way to improve them.

In the statistical experiment of observing the pair (Y, U), where $Y \sim N(\Sigma b, \Lambda)$ and U is ancillary, the variable Y is sufficient. Thus, for any estimator $S(Y, U)$, consider its conditional expectation

$$\tilde{S}(Y) := E[S(Y, U)|Y]. \tag{3.6}$$

By the Rao-Blackwell theorem, the risk of $\tilde{S}(Y)$ is less than or equal to that of $S(Y, U)$ for all b and for any convex loss function.

To implement the conditional estimators, we need consistent estimators $\hat{\Lambda} \to_p \Lambda$ and $\hat{\Sigma} \to_p \Sigma$. Dependence in the scores poses no problem, so long as $\hat{\Lambda}$ is a consistent estimate of the zero-frequency spectral density. Recall that $T^{1/2}\hat{\beta}(m) \to_d H(m)Y(1, m)$. Then take L independent artificially generated Brownian bridges $\{U_B^i(r)\}_{i=1}^L$ with covariance matrix $\hat{\Lambda}$. For each i, consider the estimators:

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

and

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

where $\hat{H}(m)$ is the Moore-Penrose inverse of $\hat{\Sigma}(m)$ and $U_B^i(r, m)$ is the vector with the elements of $U_B^i(r)$ in the locations indexed by m and zeros everywhere else. The next proposition gives their limiting distributions:

Proposition 3.5 *For each i:*

$$T^{1/2}\tilde{\beta}_{i,1} \to_d \sum_{m^*} H(m^*)\tilde{Y}_i(1,m^*)\mathbf{1}\{m^* = \arg\max_m [-\int_\pi^1 \frac{\tilde{Y}_i(r,m)'}{r} H(m) \frac{\tilde{Y}_i(r,m)}{r} dr + 2\int_\pi^1 \frac{\tilde{Y}_i(r,m)'}{r} H(m) d\tilde{Y}_i(r)]\}$$

and

$$T^{1/2}\tilde{\beta}_{i,2} \to_d \sum_{m^*} H(m^*) \frac{\tilde{Y}_i(1,m^*) - \tilde{Y}_i(\pi,m^*)}{1-\pi} \mathbf{1}\{m^* = \arg\max_m [\frac{1}{\pi}\tilde{Y}_i(\pi,m)' H(m) \tilde{Y}_i(\pi,m) - 2n(m)]\}$$

where $\tilde{Y}_i(r) = rY + U_B^i(r)$ and $\tilde{Y}_i(r,m)$ is a $k \times 1$ vector with the elements of $\tilde{Y}_i(r)$ in the locations indexed by m and zeros elsewhere. These are the same distributions as in equations (3.4) and (3.5).

These estimators can then be averaged over i. After this step of averaging over different realizations of the Brownian bridge, the asymptotic distributions depend on Y alone and are asymptotically the expectations of the out-of-sample and split-sample estimators conditional on Y. Note that this Rao-Blackwellization (henceforth, RB) does not apply to the in-sample estimator because there is no ancillary noise process to eliminate in this case.

In the special case of regression considered in Subsection 2.1, numerical calculations indicate that the limiting risk of the RB estimator is strictly lower than the original estimator for at least some values of b, implying that the out-of-sample and split-sample estimators are asymptotically inadmissible.

3.2 Linear Regression Model and Bagging

In the special case of the regression model with orthonormal regressors, considered in Subsection 2.1, we have $\Lambda = \Sigma = \sigma^{-2} I_k$. In this model, all of the estimators depend crucially on the partial sum process $T^{-1/2}\Sigma_{t=1}^{[Tr]} x_t y_t$ and it follows from Proposition 3.3 that:

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

and Proposition 3.4 will immediately apply.

In the linear regression model (subsection 2.1), we can also consider adding a *bagging* step to each of the procedures. Bagging, or bootstrap aggregation, was proposed by Breiman (1996) as a way to smooth predictive procedures. Bühlmann and Yu (2002) study the large-sample properties of bagging. The i^{th} 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 full model-selection and estimation procedure is then applied to the i^{th} bootstrap sample. This is repeated L times, and the L estimates are averaged to obtain the bagged estimate that can be used for forecasting. The following proposition provides a key result for obtaining the limiting distribution of a single bootstrap sample.

Proposition 3.6 *Let $\{x_t^*(i), y_t^*(i), t = 1, \ldots, T\}$ be the i^{th} bootstrap sample. In large samples*

$$T^{-1/2}\sigma^{-2}\Sigma_{t=1}^{[Tr]} x_t^*(i) y_t^*(i) \rightarrow_d r\,Y + V_i(r) =: Y_i^*(r),$$

where Y is as in Proposition 3.3 and $\{V_i(r)\}_{i=1}^{L}$ are $k \times 1$ Brownian motions with covariance matrix $\sigma^{-2} I$ 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 Proposition 3.3, except that the Brownian bridge $U_B(r)$ is replaced with a Brownian motion $V_i(r)$. Using Proposition 3.6, we can obtain asymptotic representations for a single bootstrap draw of the different procedures in analogy with (3.1)-(3.5):

Proposition 3.7 *In the i^{th} 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 unrestricted MLE:

$$T^{1/2}\tilde{\beta}_i \rightarrow_d \Sigma^{-1} Y_i^*(1) \tag{3.7}$$

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

$$T^{1/2}\tilde{\beta}_i \rightarrow_d \Sigma^{-1} Y_i^*(1) \max(1 - \frac{k-2}{Y_i^*(1)'\Sigma^{-2} Y_i^*(1)}, 0) \tag{3.8}$$

(iii) Selecting the model using the AIC:

$$T^{1/2}\tilde{\beta}_i \to_d \sum_{m^*} H(m^*) Y_i^*(1,m^*) \mathbf{1}\{m^* = \arg\max_m [Y_i^*(1,m)' H(m) Y_i^*(1,m) - 2n(m)]\} \tag{3.9}$$

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

$$\begin{aligned} T^{1/2}\tilde{\beta}_i \to_d \sum_{m^*} H(m^*) Y_i^*(1,m^*) \mathbf{1}\{m^* = \arg\max_m [2 \int_\pi^1 \frac{Y_i^*(r,m)'}{r} H(m) dY_i^*(r) \\ - \int_\pi^1 \frac{Y_i^*(r,m)'}{r} H(m) \frac{Y_i^*(r,m)}{r} dr]\} \end{aligned} \tag{3.10}$$

(v) Using the split-sample method:

$$\begin{aligned} T^{1/2}\tilde{\beta}_i \to_d \sum_{m^*} H(m^*) \frac{Y_i^*(1,m^*) - Y_i^*(\pi,m^*)}{1-\pi} \mathbf{1}\{m^* = \arg\max_m [\frac{1}{\pi} Y_i^*(\pi,m)' H(m) Y_i^*(\pi,m) \\ -2n(m)]\} \end{aligned} \tag{3.11}$$

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

The distribution of the parameter estimates from bagging are then given by averaging the expressions in equations (3.7)–(3.11) over L different draws of $V_i(r)$. In Appendix B, we also provide more concrete expressions for the in-sample and split-sample procedures, in their standard form, with RB, and with bagging, in the special case where $k = 1$.

For all of the bagged procedures characterized in Proposition 3.7, 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.

Bagging and our proposed RB procedure are closely related. RB uses simulation to integrate out the Brownian bridge $U_B(r)$. Bagging is asymptotically equivalent to replacing the Brownian bridge with a Brownian motion, and then integrating it out. Our RB approach can be used in any setting where we have consistent estimators of Λ and Σ, and does not require resampling the data. For this reason, it may be especially attractive when the data are dependent. Breiman (1996) gave a heuristic argument for why bagging weakly reduces mean square error, but in fact 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). On the other hand RB does indeed weakly reduce the local asymptotic risk for convex loss functions.

Because RB involves directly integrating out $U_B(r)$, it does not affect any procedure that does not depend on this ancillary noise (because $U_B(1) = 1$ full sample procedures won't depend on this noise). In particular, RB does not affect in-sample model selection with AIC. Bagging is different. Bagging replaces $U_B(r)$ with a Brownian motion and then integrates that out. But this affects the limiting distribution of all of the procedures that we consider, except for the unrestricted MLE. Bagging affects in-sample model selection with AIC, where it can be thought of as replacing hard thresholding with soft thresholding (see Appendix B for more discussion).

4 Extensions

In this section, we consider two extensions of the basic framework of our analysis, in the context of the linear regression model.

4.1 Unmodeled Structural Change

A variant of our basic regression model specifies that $y_t = \beta_t' x_t + u_t$ where $T^{1/2}\beta_{[Tr]} = W(r)$, where r may be either a stochastic or nonstochastic process. This allows various forms of structural breaks, and is similar to specifications used by Andrews (1993) and Elliott and Mueller (2014). For example, if $\beta_t = T^{-1/2}b + T^{-1/2}\tilde{b}1(t > [Ts])$, then $W(r) = b + \tilde{b}1(r > s)$. Or, if $\beta_t = T^{-1}\Sigma_{s=1}^{t}\eta_s$ with Gaussian shocks, then $W(r)$ is a Brownian motion. Proposition 4.1 gives the asymptotic distribution of the partial sum process

$T^{-1/2}\sigma^{-2}\Sigma_{t=1}^{[Tr]} x_t y_t$ in this variant of our basic model:

Proposition 4.1 *As $T \to \infty$, the partial sum process*

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

where $Z(r) \stackrel{d}{=} \Sigma \int_0^r W(s)ds + r\xi + U_B(r)$, $\xi \sim N(0,\Sigma)$ and $U_B(r)$ is an independent k-dimensional Brownian bridge with covariance matrix $\Sigma = \sigma^{-2} I$.

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 Propositions 3.4 and 3.7, with $Y(r)$ replaced by $Z(r)$ and $Y_i^*(r)$ replaced by $rZ(1) + \sigma V_i(r)$ everywhere. Alternatively, the researcher might be aware of the possibility of structural change, and might choose to select among models and estimate parameters using a rolling window. The estimators will then have limiting distributions that are simple extensions of those in Propositions 3.4 and 3.7. Other approaches for dealing with the possibility of parameter instability might be considered, but we leave this topic for future research.

4.2 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 $\hat{\beta}(1,m)$ denotes the parameter estimate from the model containing the variables indexed by m (with zeros in other locations), then we could estimate the parameter vector as $\Sigma_m w(m)\hat{\beta}(1,m)$, where Σ_m denotes the sum over all the models in $\mathscr{M}$ and the weights sum to 1. As examples of weighting schemes, we could set $w(m) = \frac{\exp(AIC(m)/2)}{\Sigma_{m^*}\exp(AIC(m^*)/2)}$ (Buckland, Burnham, and Augustin, 1997) or $w_i = \frac{\exp(-\hat{\sigma}^2(m))}{\Sigma_{m^*}\exp(-\hat{\sigma}^2(m^*))}$ where $AIC(m)$ and $\hat{\sigma}^2(m)$ denote the Akaike Information Criterion and out-of-sample mean of squared residuals in the model indexed by m. Alternatively, to do a combination version of the split-sample scheme, we could estimate the parameter vector as $\Sigma_m w(m)\hat{\beta}^*(\pi,m)$ where $w(m) = \frac{\exp(AIC(\pi,m)/2)}{\Sigma_m \exp(AIC(\pi,m)/2)}$ and $AIC(\pi,m)$ denotes the Akaike Information Criterion for the model indexed by m computed only over the first fraction π of the sample.

Proposition 4.2 *If the parameter vector is estimated by* $\Sigma_m w(m)\hat{\beta}(1,m)$ *then in large samples, the distributions of the alternative parameter estimates will be:*

$$\sigma^2 E\{\Sigma_m w(m) H(m) Y(1,m)\}$$

where

$$w(m) \propto \exp([Y(1,m)' H(m) Y(1,m) - 2n(m)]/2)$$

or

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

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

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

where

$$w(m) \propto \exp([\frac{1}{\pi} Y(\pi,m)' H(m) Y(\pi,m) - 2n(m)]/2)$$

The standard bagging step can be added to any of these methods for forecast combination and the resulting limiting distributions in the ith of L bootstrap samples are also given by Proposition 4.2, except with $Y(\cdot)$ and $Y(\cdot,m)$ replaced by $Y_i^*(\cdot)$ and $Y_i^*(\cdot,m)$ everywhere. Or RB can be added, and Proposition 4.2 would still apply, except with $Y(\cdot)$ and $Y(\cdot,m)$ replaced by $\tilde{Y}_i(\cdot)$ and $\tilde{Y}_i(\cdot,m)$.

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.

5 Numerical Work

In this section we numerically explore the root mean squared error

$$\sqrt{E\left[(T^{1/2}\tilde{\beta}-b)'(T^{1/2}\tilde{\beta}-b)\right]}, \tag{5.1}$$

the square of which is asymptotically equivalent to the NMSPE in the regression model example. Given the expressions in Propositions 3.4 and 3.7, we can simulate the asymptotic risk of different methods in their standard form, with RB, and with bagging[4] for different choices of the localization parameter b and the number of potential predictors k. None of the methods gives the lowest risk uniformly in b. Always using the big model is minmax, but due to the Stein phenomenon, it may be dominated by shrinkage estimators. In all cases, RB and bagging are implemented using 100 replications, the out-of-sample and split-sample methods both set $\pi = 0.5$, and we set $\Sigma = \Lambda = I_k$. The asymptotic risk is symmetric in b and is consequently shown only for non-negative b. The bagging results from Proposition 3.7 apply only in the special case of the linear regression model, but RB applies in the general pseudo-likelihood framework. Figure 1 plots the asymptotic risk of the standard in-sample, out-of-sample and split-sample methods, for the case $k = 1$ against b. Results with RB and bagging are also included.

Among the standard methods, 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 RB changes this ordering. RB reduces the risk of the out-of-sample and split-sample methods for all values of b, and makes them much more competitive. Bagging accomplishes much the same thing. The fact that bagging improves the out-of-sample and split-sample methods uniformly in b is just a numerical result, but it is also a theoretical result for RB. Neither bagging nor RB dominates the other in terms of risk. Bagging also helps with the in-sample method for some but not all values of b—this was also shown by Bühlmann and Yu (2002). Recall that RB does nothing to the in-sample method.

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 or RB step is included. Indeed, for all b, the best forecast is some method combined with bagging or RB.

[4]The results with bagging are based on Proposition 3.7, which applies only in the case of the linear regression model.

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 $b = (b_1, 0, \ldots, 0)'$. Figure 2 plots the risk for $k = 3$ against b_1 for in-sample, out-of-sample and split-sample methods in the standard form, with RB, and with bagging. The positive-part James-Stein estimator is also included. The split-sample method with either RB or bagging compares very favorably with the other alternatives.

We next consider the case with multiple potential predictors, but where the associated coefficients are all equal. We specify that $b = b_1 k^{-1/2} i$ where i denotes a $k \times 1$ vector of ones. Figure 3 plots the risk for $k = 3$ against b_1 for the different procedures. Again RB and bagging help the split-sample and out-of-sample methods a good deal.

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 risk for various pairs of possible forecasting methods. We find that in all cases, the, out-of-sample forecasts with RB or bagging dominate those without. For the bagging, this is a numerical result, but for RB it is a theoretical one, as discussed above. In this sense, one should never use the conventional out-of-sample forecasting methodology without some risk-reduction scheme to integrate out noise.

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

Figure 4 plots the risk for $k = 1$ against b for the in-sample, out-of-sample and split-sample forecast combination methods, in their standard form, with RB and with bagging. These are based on simulating the distributions in Proposition 4.2. 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 randomization step, using in-sample AIC weights does best for most values of b. Adding in RB/bagging allows better predictions to be made. RB/bagging reduces the risk of the combination forecasts with out-of-sample or split-sample weights uniformly in b. Once an RB/bagging step is added in, there is no clear winner among the in-sample, out-of-sample and split-sample forecast combination methods.

6 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 (2.1) 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 5 plots the simulated root normalized mean square prediction errors ($\sqrt{T*(MSPE-1)}$) against β for $k = 1$. Figure 6 and 7 repeat this for $k = 3$ where $\beta = (\beta_1, 0, 0)'$ and $\beta = \beta_1 k^{-1/2} i$, respectively, with the results plotted against β_1. In our simulations, we include bagging and RB of the out-of-sample and split-sample forecasts. 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 omitted from FIgures 5-7.

Figures 5-7 give very similar conclusions to the local asymptotic calculations reported in Figures 1-3. Without RB or bagging, the in-sample scheme generally gives the best forecasts, followed by out-of-sample, with the split-sample doing the worst. RB or bagging substantially improve the performance of the out-of-sample and split sample methods. In Figure 6, there are many values of β_1 for which the split-sample method with RB or bagging does best among all the model selection methods considered.

7 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, the out-of-sample and split-sample methods can be improved substantially by removing the impact of an ancillary noise term that appears in their limit representations, either through Rao-Blackwellization or bagging. For important ranges of the local parameters, these modified procedures are very competitive with in-sample methods. Our representation results highlight a noise-reduction aspect of bagging, and also leads to an alternative approach that dominates the out-of-sample and split-sample methods asymptotically and can be implemented without having to resample the data.

Appendix A: Proof of Propositions

Proof of Proposition 3.3: We have

$$T^{-1/2}\sum_{t=1}^{[Tr]}\frac{\partial \ell_t(0)}{\partial \beta} = T^{-1/2}\sum_{t=1}^{[Tr]}\frac{\partial \ell_t(\beta_0)}{\partial \beta} - T^{-1/2}\sum_{t=1}^{[Tr]}\frac{\partial^2 \ell_t(\beta_0)}{\partial \beta^2}\beta_0 + o_p(1)$$
$$\rightarrow_d r\Sigma b + B(r),$$

where $B(r)$ denotes a Brownian motion with covariance matrix Λ. Let $Y(r) = r\Sigma b + B(r)$, and let $Y = Y(1)$ which is $N(\Sigma b, \Lambda)$. Define $U_B = B(r) - rB(1)$. Then by a standard construction of the Brownian bridge process,

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

■

Proof of Proposition 3.4: Let $\hat{\beta}(m)$ denote the vector of coefficient estimates corresponding to the predictors indexed by m with zeros in all other locations. Equations (3.1) and (3.2) immediately follow because $T^{1/2}\hat{\beta} \rightarrow_d Y$.

The AIC objective function (to be maximized) is:

$$2\Sigma_{t=1}^{T} l_t(\hat{\beta}(m)) - 2n(m) = 2\Sigma_{t=1}^{T} l_t(0) + 2\hat{\beta}(m)'\frac{\partial l_t(y_t,0)}{\partial \beta} + \hat{\beta}(m)'\Sigma_{t=1}^{T}\frac{\partial^2 l_t(y_t,0)}{\partial \beta^2}\hat{\beta}(m) - 2n(m) + o_p(1)$$

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

$$Y(1,m)'H(m)Y(1,m) - 2n(m),$$

noting that $H(m)\Sigma H(m) = H(m)$.

The OOS objective function (to be maximized) is:

$$\Sigma_{t=[T\pi]+1}^{T} l_t(\hat{\beta}_{1,t-1}(m)) = \Sigma_{t=[T\pi]+1}^{T}[l_t(0) + \hat{\beta}_{1,t-1}(m)' \frac{\partial l_t(y_t,0)}{\partial \beta}$$
$$+\frac{1}{2}\hat{\beta}_{1,t-1}(m)' \frac{\partial^2 l_t(y_t,0)}{\partial \beta^2}\hat{\beta}_{1,t-1}(m)] + o_p(1)$$

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

$$2\int_{\pi}^{1} \frac{Y(r,m)'}{r} H(m) dY(r) - \int_{\pi}^{1} \frac{Y(r,m)'}{r} H(m) \frac{Y(r,m)}{r} dr$$

The AIC estimated over the first fraction π of the sample is:

$$2\Sigma_{t=1}^{T} l_t(\hat{\beta}_{1,[T\pi]}(m)) - 2n(m) = 2\Sigma_{t=1}^{T} l_t(0)$$
$$+2\hat{\beta}_{1,[T\pi]}(m)' \frac{\partial l_t(y_t,0)}{\partial \beta} + \hat{\beta}_{1,[T\pi]}(m)' \Sigma_{t=1}^{T} \frac{\partial^2 l_t(y_t,0)}{\partial \beta^2} \hat{\beta}_{1,[T\pi]}(m) - 2n(m) + o_p(1)$$

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

$$\frac{1}{\pi} Y(\pi,m)' H(m) Y(\pi,m) - 2n(m)$$

The limiting distributions in Proposition 3.4 all follow from these results and the facts that $T^{1/2}\hat{\beta} \rightarrow_d \Sigma^{-1} Y(1)$ and $T^{1/2}\hat{\beta}(m) \rightarrow_d H(m) Y(1,m)$. ■

Proof of Proposition 3.5: We know that $T^{1/2}\hat{\beta}(m) \rightarrow_d H(m)\tilde{Y}_i(1,m)$ and $\hat{H}(m) \rightarrow_p H(m)$. Hence $T^{1/2}\hat{\beta}(m) + \hat{H}(m)\frac{U_B^i(r,m)}{r} \rightarrow_d H(m)\frac{\tilde{Y}_i(r,m)}{r}$. The result follows immediately. ■

Proof of Proposition 3.6: Let $\{x_t^*(i), y_t^*(i)\}$ be the ith bootstrap sample and let $u_t^*(i) = y_t^*(i) - \beta' x_t^*(i)$, $t = 1, \ldots T$. From Theorem 2.2 of Park (2002), $T^{-1/2}\sigma^{-2}\sum_{t=1}^{[Tr]}(x_t^*(i)u_t^*(i) - T^{-1}\sum_{s=1}^{T} x_s u_s) \rightarrow_d V_i(r)$. Consequently $T^{-1/2}\sigma^{-2}\Sigma_{t=1}^{[Tr]} x_t^*(i) y_t^*(i) \rightarrow_d rY + V_i(r)$. ■

The proofs of Propositions 3.7 and 4.2 involve exactly the same calculations as in Proposition 3.4 and are hence omitted.

Proof of Proposition 4.1: We have

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

where $B(r)$ is a Brownian motion with covariance matrix Λ. ■

Appendix B: Shrinkage Representations in the case $k = 1$

In the case $k = 1$, and with $\Sigma = \Lambda$, some of the expressions in Propositions 3.4 and 3.7 can be simplified. For the AIC estimator in its standard form we have:

$$T^{1/2}\tilde{\beta} \to_d \Sigma^{-1} Y 1(|Y| > \sqrt{2\Sigma}). \tag{B1}$$

For the split-sample estimator, we have:

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

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

Rao-Blackwellization makes no difference to the AIC estimator, and equation (B1) continues to apply. For the split-sample estimator in the ith simulated sample, we have:

$$T^{1/2}\tilde{\beta}_i \to_d \Sigma^{-1}(Y - \sqrt{\frac{\pi\Sigma}{1-\pi}} z(i)) 1(|Y + \sqrt{\frac{(1-\pi)\Sigma}{\pi}} z(i)| > \sqrt{\frac{2\Sigma}{\pi}}) = (\Sigma^{-1} Y - \gamma z(i)) 1(|\gamma Y + z(i)| > \sqrt{\frac{2}{1-\pi}}),$$

where $z(i)$ is $N(0,1)$, and is independent of Y and $\gamma = \sqrt{\frac{\pi}{(1-\pi)\Sigma}}$. Thus for the overall split-sample estimator with RB, we have:

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

For the AIC estimator, with bagging, we have:

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

where $\kappa = \Sigma^{-1/2}$, shown in proposition 2.2 of Bühlmann and Yu (2002).[5] Comparing this to equation

[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.

(B1), in the context of the AIC estimator, bagging is effectively replacing a hard thresholding procedure with a soft thresholding counterpart.

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

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

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 $N(Y, \frac{\Sigma}{(1-\pi)})$, $z_2(i)|Y$ is $N(Y, \frac{\Sigma}{\pi})$ and the two are independent, conditional on Y. Thus for the overall split-sample with bagging estimator:

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

We have no such simplified expression for the out-of-sample estimators, but we still know from equations (3.4) and (3.10) that the limits of the out-of-sample estimators with RB and bagging are both functions of Y alone.

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.

ANDREWS, D. W. K. (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61(4), 821–856.

ASHLEY, R., C. W. GRANGER, AND R. SCHMALENSEE (1980): "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica*, 48, 1149–1167.

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.

CLAESKENS, G., AND N. L. HJORT (2008): *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

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

DIEBOLD, F. X., AND R. S. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

ELLIOTT, G., AND U. K. MUELLER (2014): "Pre and Post Break Parameter Inference," *Journal of Econometrics*, 180(2), 141–157.

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

GIACOMINI, R., AND H. WHITE (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

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

HANSEN, P. R., AND A. TIMMERMANN (2013): "Equivalence Between Out-of-Sample Forecast Comparisons and Wald Statistics," 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.

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.

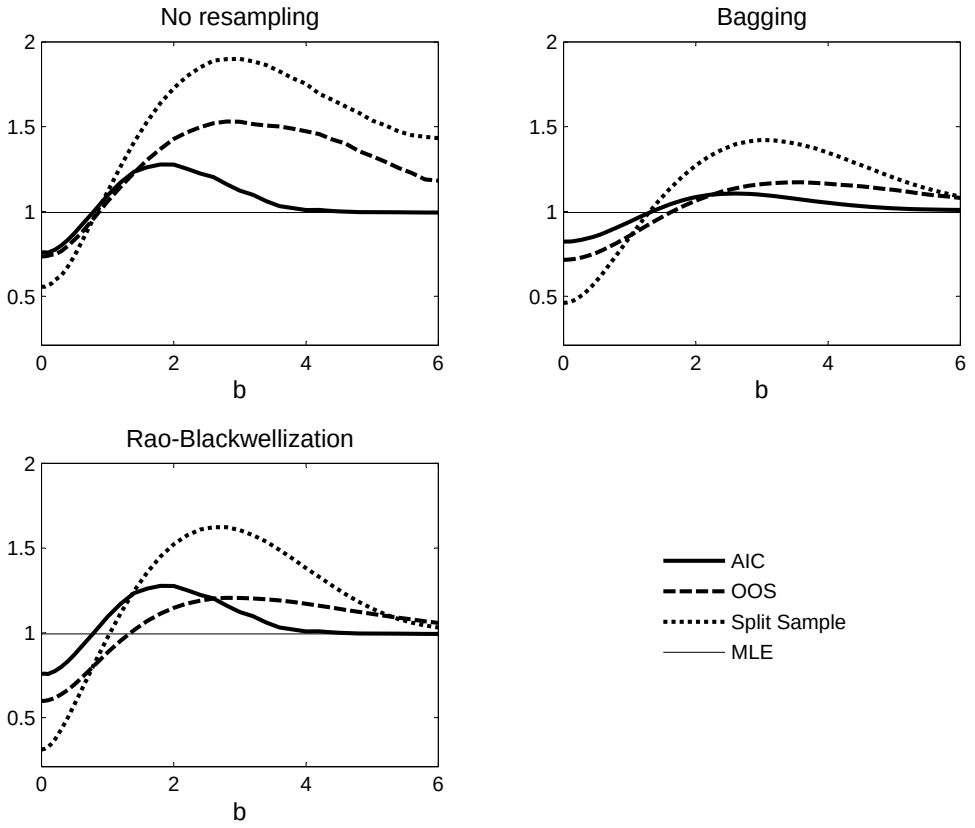
VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Table 1: Dominance Relations in Local Asymptotic Risk

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. OOSRB	–	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	–	–	OOSRB	OOSRB	OOSRB
MLE v. SSRB	–	–	–	SSRB	SSRB	SSRB	–	–	–	–	SSRB
AIC v. AICB	–	–	–	–	–	–	–	–	–	–	–
JS v OOSB	NA	NA	OOSB	–	–	–	NA	–	–	–	–
JS v SSB	NA	NA	SSB	SSB	SSB	SSB	NA	–	–	–	SSB
JS v OOSRB	NA	NA	OOSRB	OOSRB	–	–	NA	OOSRB	–	–	–
JS v SSRB	NA	NA	–	SSRB	SSRB	SSRB	NA	–	–	–	–
AIC v. OOSB	–	–	–	–	–	OOSB	–	–	–	–	–
AIC v. SSB	–	–	–	–	SSB	SSB	–	–	–	–	–
AIC v. OOSRB	–	–	OOSRB	OOSRB	OOSRB	OOSRB	–	–	–	–	OOSRB
AIC v. SSRB	–	–	–	–	SSRB	SSRB	–	–	–	–	–
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. OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB
OOS v. SSRB	–	SSRB	SSRB	SSRB	SSRB	SSRB	–	SSRB	SSRB	SSRB	SSRB
SS v. AICB	–	–	–	–	–	–	–	–	–	–	–
SS v. OOSB	–	–	–	–	–	–	–	–	–	–	–
SS v. SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB	SSB
SS v. OOSRB	–	–	–	–	–	–	–	–	–	–	–
SS v. SSRB	SSRB	SSRB	SSRB	SSRB	SSRB	SSRB	SSRB	SSRB	SSRB	SSRB	SSRB
AICB v. OOSB	–	–	OOSB	OOSB	OOSB	OOSB	–	–	–	OOSB	OOSB
AICB v. SSB	–	–	–	SSB	SSB	SSB	–	–	–	–	SSB
AICB v. OOSRB	–	OOSRB	OOSRB	OOSRB	OOSRB	OOSRB	–	–	OOSRB	OOSRB	OOSRB
AICB v. SSRB	–	–	–	–	SSRB	SSRB	–	–	–	–	–
OOSB v. SSB	–	–	–	–	SSB	SSB	–	–	–	–	–
OOSB v. OOSRB	–	–	OOSRB	OOSRB	OOSRB	OOSRB	–	–	–	OOSRB	OOSRB
OOSB v. SSRB	–	–	–	–	–	SSRB	–	–	–	–	–
SSB v. OOSRB	–	–	–	–	–	–	–	–	–	–	–
SSB v. SSRB	–	–	–	– –	–	–	–	–	–	–	–
OSSRB v. SSRB	–	–	–	–	–	–	–	–	–	–	–

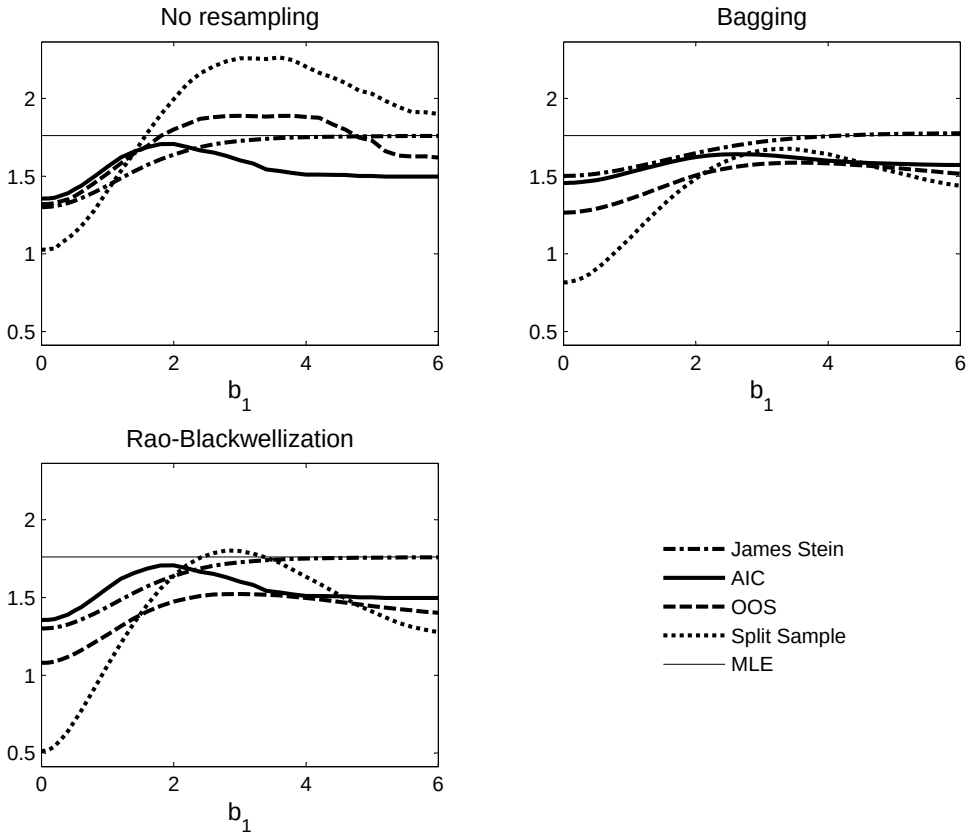
This table reports comparisons of local asymptotic risk between pairs of methods: maximum likelihood (MLE), positive-part James-Stein estimator (JS) applicable if $k > 2$, in-sample with AIC (AIC), the counterpart with bagging (AICB), out-of-sample (OOS), the counterpart with bagging/Rao-Blackwellization (OOSB/OOSRB), the split-sample method (SS) and the counterpart with bagging/Rao-Blackwellization (SSB/SSRB). 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 "–".

Figure 1: Local Asymptotic Risk ($k = 1$)



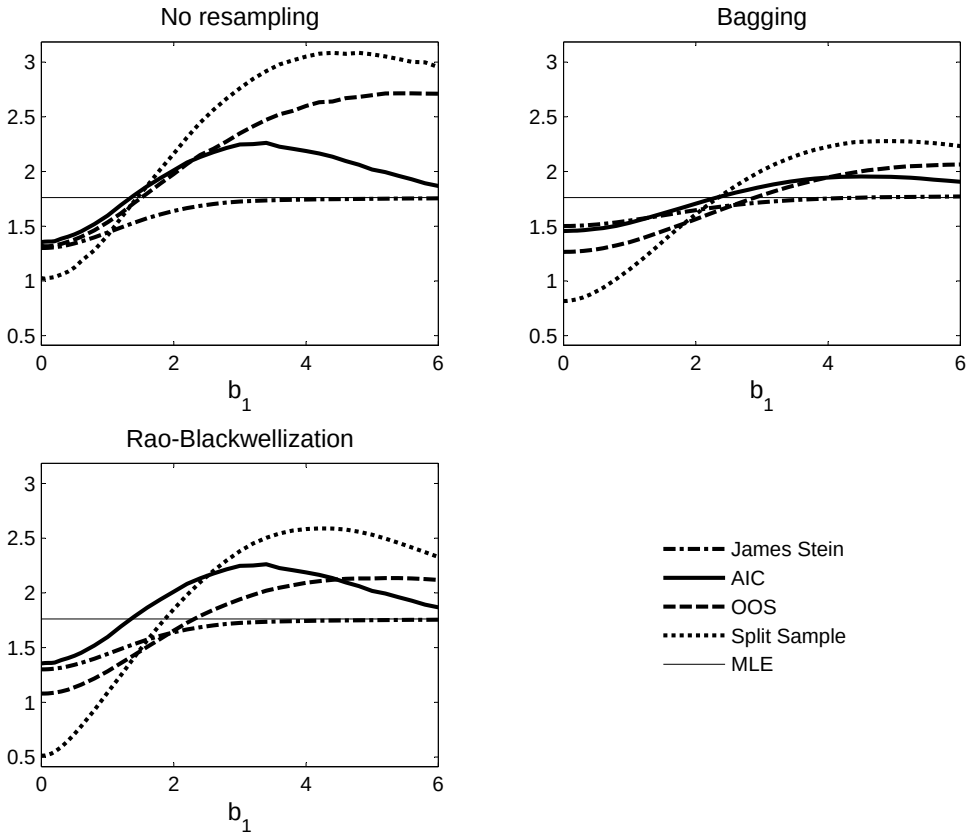
Notes: These are the simulated local asymptotic risk values, equation (5.1), for different procedures, plotted against b. Note that MLE is the same without any resampling, with bagging or with Rao-Blackwellization. AIC is the same without any resampling or with Rao-Blackwellization.

Figure 2: Local Asymptotic Risk ($k = 3$, Single Nonzero Coefficient)



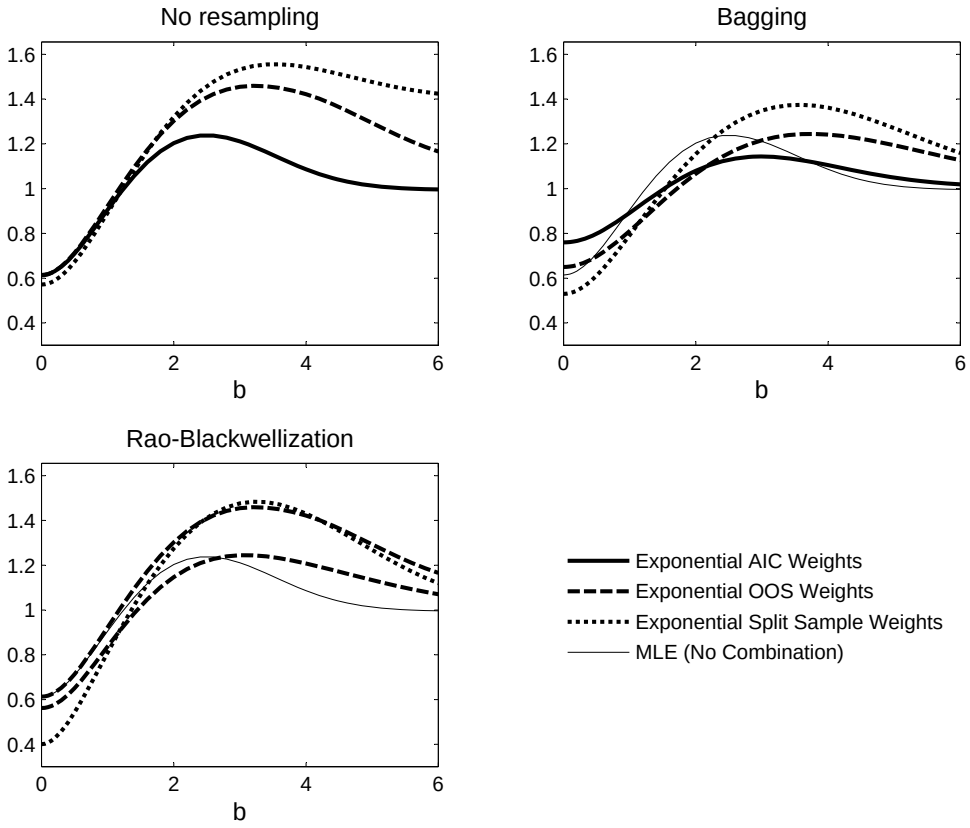
Notes: These are the simulated local asymptotic risk values, equation (5.1), for different procedures, plotted against b_1, where $b = (b_1, 0, \ldots, 0)'$. Note that MLE is the same without any resampling, with bagging or with Rao-Blackwellization. AIC is the same without any resampling or with Rao-Blackwellization.

Figure 3: Local Asymptotic Risk ($k = 3$, All Coefficients Equal)



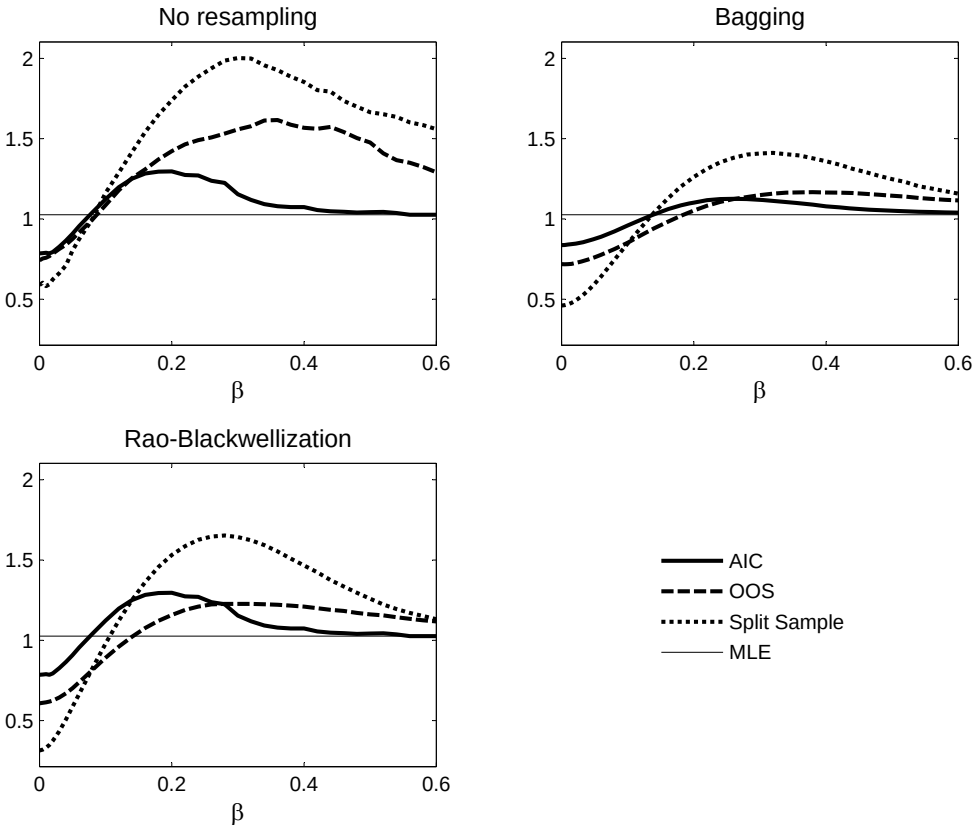
Notes: These are the simulated local asymptotic risk values, equation (5.1), for different procedures, plotted against b_1, where $b = b_1 k^{-1/2} i$. Note that MLE is the same without any resampling, with bagging or with Rao-Blackwellization. AIC is the same without any resampling or with Rao-Blackwellization.

Figure 4: Local Asymptotic Risk: Combination Forecasts ($k = 1$)



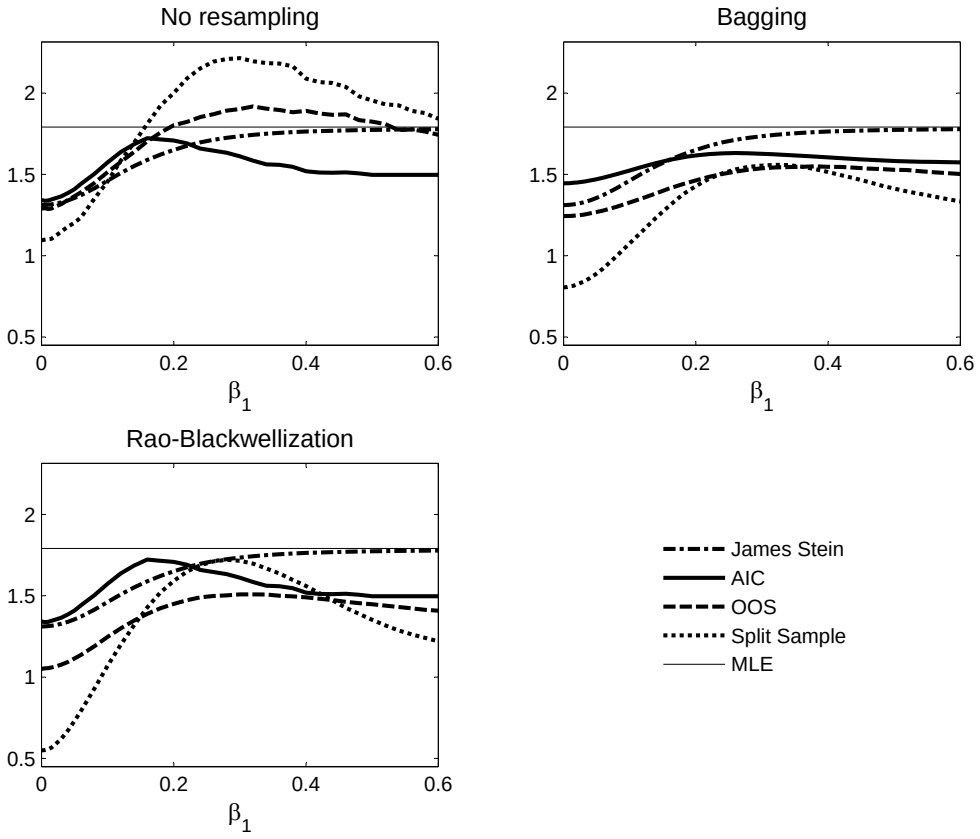
Notes: These are the simulated local asymptotic risk values, equation (5.1), for different procedures, plotted against b. Note that MLE is the same without any resampling, with bagging or with Rao-Blackwellization. Exponential AIC forecast combination is the same without any resampling or with Rao-Blackwellization.

Figure 5: Root Normalized Mean Square Prediction Errors ($k = 1$)



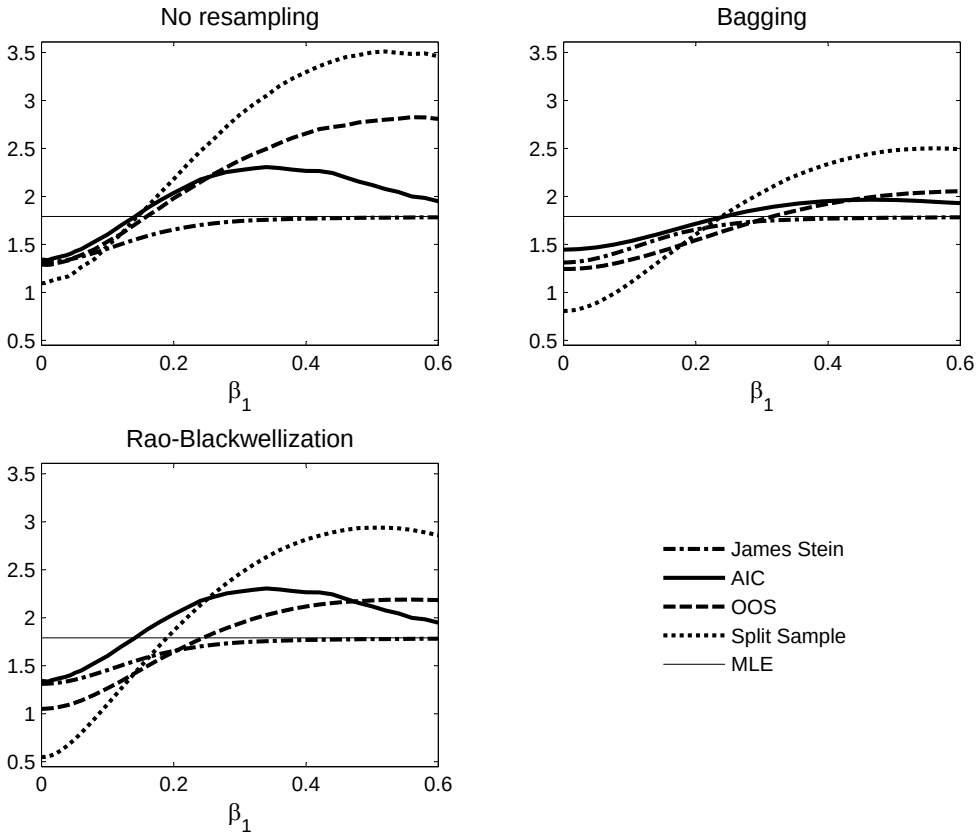
Notes: These are the simulated root normalized mean square prediction errors using different procedures, plotted against β. There is one possible predictor and the sample size is $T = 100$. Note that MLE is the same without any resampling, with bagging or with Rao-Blackwellization. AIC is the same without any resampling or with Rao-Blackwellization.

Figure 6: Root Normalized Mean Square Prediction Errors ($k = 3$, Single Nonzero Coefficient)



Notes: These are the simulated root normalized mean square prediction errors using different procedures, where $\beta = (\beta_1, 0, 0)'$, plotted against β_1. The sample size is $T = 100$. Note that MLE is the same without any resampling, with bagging or with Rao-Blackwellization. AIC is the same without any resampling or with Rao-Blackwellization.

Figure 7: Root Normalized Mean Square Prediction Errors ($k = 3$, All Coefficients Equal)



Notes: These are the simulated root normalized mean square prediction errors using different procedures, where $\beta = \beta_1 k^{-1/2} i$, plotted against β_1. The sample size is $T = 100$. Note that MLE is the same without any resampling, with bagging or with Rao-Blackwellization. AIC is the same without any resampling or with Rao-Blackwellization.