Bagging Time Series Models[*]

Atsushi Inoue[†] Lutz Kilian[‡]

North Carolina State University University of Michigan
CEPR

January 12, 2004

Abstract

A common problem in out-of-sample prediction is that there are potentially many relevant predictors that individually have only weak explanatory power. We propose bootstrap aggregation of pre-test predictors (or bagging for short) as a means of constructing forecasts from multiple regression models with local-to-zero regression parameters and errors subject to possible serial correlation or conditional heteroskedasticity. Bagging is designed for situations in which the number of predictors (M) is moderately large relative to the sample size (T). We show how to implement bagging in the dynamic multiple regression model and provide asymptotic justification for the bagging predictor. A simulation study shows that bagging tends to produce large reductions in the out-of-sample prediction mean squared error and provides a useful alternative to forecasting from factor models when M is large, but much smaller than T. We also find that bagging indicators of real economic activity greatly redcues the prediction mean squared error of forecasts of U.S. CPI inflation at horizons of one month and one year.

JEL: C22, C52, C53
KEYWORDS: Bootstrap aggregation; Forecasting; Model selection; Pre-testing.

[*] The first draft of this paper was written in January 2002. We thank Bob Stine for stimulating our interest in bagging. We acknowledge helpful discussions with Todd Clark, Silvia Gonçalves, Peter R. Hansen, Mike McCracken, Serena Ng, Barbara Rossi and Clara Vega. We thank seminar participants at Michigan State University, the University of Maryland, the University of Tokyo, and the 2003 Triangle Econometrics Conference.

[†] Department of Agricultural and Resource Economics, Box 8109, North Carolina State University, Raleigh, NC 27695-8109. E-mail: atsushi_inoue@ncsu.edu.

[‡] Department of Economics, University of Michigan, 611 Tappan Street, Ann Arbor, MI 48109-1220. E-mail: lkilian@umich.edu.

1 Introduction

A common problem in out-of-sample prediction is that the researcher suspects that many predictors are potentially relevant, but none of these predictors individually is likely to have high predictive power. If the number of predictors is at least moderately large, the usual approach of comparing all possible combinations of predictors by means of an information criterion function is computationally infeasible.[1] One strategy in this situation is to combine forecasts from many models with alternative subsets of predictors. For example, one could use the median or the trimmed mean of these forecasts as the final forecast (see Stock and Watson 2003) or one could use regression-based weights for forecast combination. The latter tend to perform poorly in practice, unless some form of shrinkage estimation is used (see, e.g., Wright 2003). Alternatively, we might extract the principal components in the set of predictors. If the data are generated by an approximate dynamic factor model, then factors estimated by principal components can be used for efficient forecasting under quite general conditions. (see, e.g., Stock and Watson 2002a, 2000b; Bai and Ng 2003).[2] If the number of predictors is moderately large relative to the sample size, a third strategy is to rely on a testing procedure for deciding which predictors to include in the forecast model and which to drop. For example, we may fit a model including all potentially relevant predictors, conduct a t-test for each predictor and discard all insignificant predictors prior to forecasting. Such pre-tests lead to inherently unstable decision rules in that small alterations in the data set may cause a predictor to be added or to be dropped. This instability tends to inflate the variance of the forecasts and may undermine the accuracy of pre-test forecasts in applied work. In this paper we will show that the predictive accuracy of simple pre-test strategies may be greatly enhanced by application of the bagging technique.

Bagging is a statistical method designed to improve the forecast accuracy of models selected by unstable decision rules. The term *bagging* is short for *bootstrap aggregation* (see Breiman 1996). In essence, bagging involves fitting the unrestricted model including all potential predictors to the original sample, generating a large number of bootstrap resamples from this approximation of the data, applying the decision rule to each of the resamples, and averaging the forecasts from the models selected by the decision rule for each bootstrap sample.

By averaging across resamples, bagging effectively removes the instability of the decision rule. Hence, one would expect the variance of the bagged prediction model to be smaller than that of the model that would be selected based on the original data. Especially when the decision rule is unstable, this variance reduction may be substantial. In contrast, the forecast bias of the prediction model is likely to be of similar magnitude, with or without bagging. Thus, one would expect bagging to reduce the prediction mean squared error of the regression model after variable selection.

This heuristic argument led Breiman (1996) to expect that bagging may in general improve the forecast accuracy of learning rules, of which regression forecasts after variable selection are

[1] See Inoue and Kilian (2003) for a discussion of this and related approaches of ranking competing forecast models. The difficulty in using information criteria when the number of potential predictors, M, is large is that the criterion must be evaluated for 2^M combinations of predictors. For $M > 20$ this task tends to become computationally prohibitive.

[2] A closely related approach to extracting common components has been developed by Forni et al. (2000, 2001) and applied in Forni et al. (2003).

just one example. Indeed, there is substantial evidence of such reductions in practice. There are some counterexamples, however, in which this intuition fails and bagging does not improve forecast accuracy. This fact has prompted increased interest in the theoretical properties of bagging. Bühlmann and Yu (2002) recently have investigated the ability of bagging to improve the forecast accuracy of regressions on an intercept when the data are i.i.d. They show that bagging does not always improve on pre-testing, but nevertheless has the potential of achieving dramatic reductions in forecast mean squared errors for a wide range of processes.

This result may seem to suggest that bagging would be useful for prediction, but this is not necessarily the case because bagging in turn may be dominated by forecasts based on the unrestricted model that includes all potential predictors or by the zero mean forecast that emerges when all predictors are dropped, as in the well-known no-change forecast model of asset returns. The former model generates unbiased, but high variance forecasts; the latter model generates biased, but zero variance forecasts. We show that in the simple setting considered by Bühlmann and Yu (2002) bagging will rarely be the best forecast method because it is almost always dominated either by the unrestricted forecast model or by the zero mean forecast model, depending on how close the slope parameters of the unrestricted model are to zero in population.

We note that it is important for bagging to work well relative to the unrestricted and the zero mean model that there be many predictors in the unrestricted model that are heterogeneous in a sense made more precise in the paper. We observe that this condition is likely to be met in economic forecasting from large data sets. Motivated by this observation, we extend the theoretical analysis of Bühlmann and Yu (2002) to multiple dynamic regression models. Our analysis allows for cross-sectional as well as serial correlation in the regression error, which arises naturally in constructing multi-period forecasts. We also allow for conditional heteroskedasticity in the regression error.

We study the asymptotic properties of bagging forecasts in this setting under the assumption that the regression parameters of the unrestricted forecast model are local-to-zero. For a stylized model we characterize analytically the effect of bagging on the asymptotic prediction mean-squared error (PMSE). We derive and compare the asymptotic PMSE of the unrestricted model, the fully restricted model, the pre-test model and the bagged pre-test model. We are able to show formally that no model will be the best forecast model for all possible data designs, but that the range for which bagging is more accurate than the alternative methods is far greater than in the single-regressor model. Simulation evidence suggests that for empirically relevant settings bagging tends to yield substantial improvements in forecast accuracy relative to the pre-test model, the unrestricted model and the zero mean model, when the number of predictors is moderately large relative to the sample size.

A practically interesting question is whether the bagging strategy is competitive with alternative approaches such as forecast combinations or forecasts from dynamic factor models. In this paper, we focus on the latter alternative. Using a number of stylized data generating processes, we provide simulation evidence that neither bagging forecasts nor factor model forecasts are most accurate in all settings, but that bagging forecasts tend to perform well relative to factor model forecasts when the number of predictors is large, but clearly smaller than the sample size. In many cases, the bagging forecast is more accurate than the factor model forecast, even when the factor model is the true model. This simulation evidence illustrates the potential of bagging to achieve substantial reductions in prediction mean squared errors.

Further research is needed to see whether these preliminary results hold more generally. While the simulation results are encouraging, we cannot be sure that our simulation design captures all important features of the data encountered in applied work. The development of realistic designs for simulation studies of this kind is still in its infancy. Given the difficulty of generalizing the results of our simulation study, we recommend that, in practice, researchers choose between the bagging strategy and the dynamic factor model strategy based on the ranking of their recursive PMSE in simulated out-of-sample forecasts.

We illustrate this approach for a typical forecasting problem in economics. Specifically, we investigate whether one-month and twelve-month ahead CPI inflation forecasts for the United States may be improved upon by adding indicators of real economic activity to models involving only lagged inflation rates. This empirical example is in the spirit of recent work by Stock and Watson (1999), Marcellino et al. (2003), Bernanke and Boivin (2003), Forni et al. (2003) and Wright (2003). We show that bagging the pre-test is by far the most accurate forecasting procedure in these empirical examples. It outperforms the benchmark autoregressive model, the unrestricted model and factor models with rank 1, 2, 3, or 4 and different lag structures. It also is more accurate than simple forecast combination methods.

The remainder of the paper is organized as follows. In section 2 we formally define bagging and present the theoretical arguments in favor of bagging. We also characterize the conditions under which we would expect bagging to work. Section 3 contains a stylized simulation study that pins the bagging strategy against some natural competitors. In section 4 we present the empirical application. We conclude in section 5. The raw data for the empirical study are described in the Data Appendix.

2 The Theory of Bootstrap Aggregation

Consider the forecasting model:

$$y_{t+h} \; = \; \beta' x_t + \varepsilon_{t+h}, \qquad h = 1, 2, 3, ... \tag{1}$$

where ε_{t+h} denotes the h-step ahead linear forecast error, β is an M-dimensional column vector of parameters and x_t is a column vector of M predictors at time period t. We presume that y_t and x_t are stationary processes or have been suitably transformed to achieve stationarity. We further assume that the predictors are weak in the sense that the coefficients of x_t are local to zero, i.e.,

$$\beta \; = \; \delta T^{-\frac{1}{2}} \tag{2}$$

where δ is an M-dimensional column vector and T is the sample size available to the forecaster at date T. This assumption facilitates the exposition. Note that our asymptotic results on bagging will still go through if some elements of β are not local to zero.

Let $\hat{\beta}$ denote the ordinary least-squares (OLS) estimator of β in (1) and let t_j denote the t-statistic for the null that β_j is zero in the unrestricted model, where β_j is the jth element of β. Further, let $\widehat{\gamma}$ denote the OLS estimator of the forecast model after variable selection. For

$x_t \in \Re^M$, we define the predictor from the unrestricted model (UR), the zero-mean predictor, and the pre-test (PT) predictor conditional on x_{T-h+1} by

$$\begin{aligned}
\hat{y}^{UR}(x_{T-h+1}) &= \hat{\beta}' x_{T-h+1}, \\
\hat{y}^{NC}(x_{T-h+1}) &= 0, \\
\hat{y}^{PT}(x_{T-h+1}) &= 0, \text{ if } |t_j| < 1.96\ \forall j \text{ and } \hat{y}^{PT}(x_{T-h+1}) = \widehat{\gamma}' S_T x_{T-h+1} \text{ otherwise,}
\end{aligned}$$

where S_T is the stochastic selection matrix obtained from

$$\begin{bmatrix} I(|t_1| > 1.96) & 0 & \cdots & 0 \\ 0 & I(|t_2| > 1.96) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I(|t_M| > 1.96) \end{bmatrix}$$

by deleting rows of zeros.

The unrestricted model forecast is based on the fitted values of a regression including all M potential predictors. The zero mean forecast emerges when all predictors are dropped, as in the well-known no-change forecast model of asset returns. We will refer to the zero-mean predictor as the no-change (NC) predictor throughout this paper, as y_{t+h} in economic applications typically refers to percentage growth rates.

The pre-test strategy that we analyze is particularly simple, but asymptotically justified. We first fit the unrestricted model that includes all potential predictors. We then conduct two-sided t-tests on each slope parameter at the 5% level using critical values based on the conventional asymptotic approximation. We discard the insignificant predictors and re-estimate the final model, before generating the forecast. In constructing the t-statistic we use appropriate standard errors that allow for serial correlation and/or conditional heteroskedasticity. Specifically, when the forecast model is correctly specified, the pre-test strategy may be implemented based on White (1980) robust standard errors for $h = 1$ and West (1997) robust standard errors for $h > 1$. If the forecast model is misspecified, the pre-test strategy must be based on nonparametric robust standard errors such as the HAC estimator proposed by Newey and West (1987). Note that this pre-test predictor is admissible in the sense of Inoue and Kilian (2003). Asymptotically, it will select a model with minimum PMSE among all possible combinations of predictors. The same is true for the unrestricted model and the no-change model under our assumptions.

The bootstrap aggregated or bagging predictor is obtained by averaging the pre-test predictor across bootstrap replications.

Definition 1. [Bagging] The bagging predictor is defined as follows:

(i) Arrange the set of tuples $\{(y_{t+h}, x_t')\}$, $t = 1, ..., T-h$. in the form of a matrix of dimension $(T-h) \times (M+1)$.

$$\begin{matrix} y_{1+h} & x_1' \\ \vdots & \vdots \\ y_T & x_{T-h}' \end{matrix}$$

When $x_t\varepsilon_{t+h}$ *is serially correlated, construct bootstrap samples* $(y^*_{1+h}, x'^*_1)\}, ..., \{(y^*_T, x'^*_{T-h})\}$ *by drawing with replacement blocks of* m rows of this matrix*, where the block size* m is chosen to capture the dependence in the error term. *This procedure is known as the blocks-of-blocks bootstrap (see, e.g., Hall and Horowitz 1996, Gonçalves and White 2003). When* $x_t\varepsilon_{t+h}$ *is serially uncorrelated, construct bootstrap samples* $(y^*_{1+h}, x'^*_1)\}, ..., \{(y^*_T, x'^*_{T-h})\}$*by sampling with replacement from the rows of the same matrix. This amounts to setting* $m = 1$ *in implementing the blocks-of-blocks bootstrap method. This procedure is also known as the pairwise bootstrap.*

(ii) *For each bootstrap sample, compute the bootstrap pre-test predictor conditional on* x_{T-h+1}

$$\hat{y}^{*PT}(x_{T-h+1}) = 0, \text{ if } |t^*_j| < 1.96\ \forall j \text{ and } \hat{y}^{*PT}(x_{T-h+1}) = \widehat{\gamma}^{*\prime} S^*_T x_{T-h+1} \text{ otherwise,}$$

where $\widehat{\gamma}^*$*and* S^*_T *are the bootstrap analogues of* $\widehat{\gamma}$*and* S_T*, respectively. In constructing* $|t^*_j|$ *we compute the variance of* $\sqrt{T}\widehat{\beta}^*$*as* $H^{*-1}\widehat{S}^*H^{*-1}$ *where*

$$\begin{aligned}\widehat{S}^* &= \frac{1}{bm}\sum_{k=1}^{b}\sum_{i=1}^{m}\sum_{j=1}^{m}(x^*_{(k-1)m+i}\varepsilon^*_{(k-1)m+i+h})'(x^*_{(k-1)m+j}\varepsilon^*_{(k-1)m+j+h}),\\ \widehat{H}^* &= \frac{1}{bm}\sum_{k=1}^{b}\sum_{i=1}^{m}(x^{*\prime}_{(k-1)m+i}x^*_{(k-1)m+i}),\end{aligned}$$

$\varepsilon^*_{t+h} = y^*_{t+h} - \widehat{\beta}^* x^*_t$, *and* b *is the integer part of* T/m *(see, e.g., Inoue and Shintani 2003).*

(iii) *The bagged predictor is the expectation of the bootstrap pre-test predictor across bootstrap samples, conditional on* x_{T-h+1}*:*

$$\hat{y}^{BA}(x_{T-h+1}) = E^*[\widehat{\gamma}^{*\prime} S^*_T x_{T-h+1}],$$

where E^* *denotes the expectation with respect to the bootstrap probability measure. The bootstrap expectation in* (iii) *may be evaluated by simulation:*

$$\hat{y}^{BA}(x_{T-h+1}) = \frac{1}{B}\sum_{i=1}^{B}\widehat{\gamma}^{*i\prime} S^{*i}_T x_{T-h+1},$$

where $B = \infty$ *in theory. In practice,* $B = 100$ *tends to provide a reasonable approximation.*

An important design parameter in applying bagging is the block size m. If the forecast model at horizon h is correctly specified in that $E(\varepsilon_{t+h}|\Omega_t) = 0$, where Ω_t denotes the date t information set, then $m = h - 1$. Otherwise $m > h - 1$. In the latter case, data-dependent rules such as calibration may be used to determine m (see, e.g., Politis, Romano and Wolf 1999).

Bagging can in principle be applied to any pre-testing strategy, not just to the specific pretesting strategy discussed here. It may seem that bagging could also be applied to other methods of forecasting. This is not necessarily the case. For bagging to be theoretically justified, it is essential that the predictor to be bagged involve some hard threshold (such as the decision of whether to include or exclude a given predictor). It would not be possible, for example, to justify bagging a given factor model, as such models are inherently smooth. Similarly, the Schwarz Information Criterion does not lend itself to bagging because it will select one forecast model with probability one asymptotically.

The performance of bagging will in general depend on the significance level chosen for pretesting. Throughout this paper we have set the nominal significance level to 5 percent. As we will show, this choice tends to work well. In practice, one could further refine the performance of bagging by comparing the accuracy of the bagging forecast method for alternative nominal significance levels in simulated out-of-sample forecasts. This question is beyond the scope of this paper.

2.1 Asymptotic Properties of Bagging

The key assumption for establishing the validity of bagging is the asymptotic normality of the OLS estimator which holds under the following conditions:

Assumption 1.

(a) $T^{-1/2}\sum_{t=1}^{T} x_t\varepsilon_t \xrightarrow{d} N(0,\Omega)$ where Ω is positive definite.

(b) $(1/T)\sum_{t=1}^{T} x_t x_t' \xrightarrow{p} E(x_t x_t')$ where $E(x_t x_t')$ is positive definite.

(c) There is a consistent estimator of Ω, $\hat{\Omega}$, i.e., $\hat{\Omega}-\Omega = o_p(1)$.

(d) $T^{-1/2}\sum_{t=1}^{T}[x_t^*\varepsilon_t^* - E^*(x_t^*\varepsilon_t^*)] \xrightarrow{d} N(0,\Omega)$ in probability conditional on the data.

(e) $(1/T)\sum_{t=1}^{T} x_t^* x_t^{*'} - E^*(x_t^* x_t^{*\prime}) = o_p(1)$ conditional on the data.

(f) There is an $M\times M$ matrix $\hat{\Omega}^*$ such that $\hat{\Omega}^* - \hat{\Omega} = o_p(1)$ conditional on the data.

(g) $E^*\|\hat{\gamma}^{*\prime}S_T^*\|^{1+\varsigma} = O_p(1)$, conditional on the data for some $\varsigma > 0$.

More primitive assumptions that imply Assumptions 1(a)-(c) can be found in Gallant and White (1988), for example. When the data are dependent, Hall and Horowitz (1996), Andrews (2002), Gonçalves and Kilian (2003), Gonçalves and White (2003) and Inoue and Shintani (2003) provide more primitive assumptions that imply Assumptions 1(d)-(f). Assumption 1(g) is a uniform integrability condition (see Billingsley 1995, p. 338).

We treat M as fixed with respect to T, but all our theoretical results will go through with minor modifications when M is allowed to increase with T at a rate not exceeding $\sqrt{T}$. The usual motivation for allowing M to increase is that the fixed M asymptotics may provide a poor small sample approximation when M is large relative to T. In the current context, however, the fixed M asymptotics work quite well, given that M is distinctly smaller than T.

Bühlmann and Yu (2002) consider bagging a linear model with one local-to-zero regressor in the form of an intercept when the data are i.i.d.. The following theorem generalizes Proposition 2.2 of Bühlmann and Yu (2002) to dynamic multiple regression models with possible serial correlation and conditional heteroskedasticity in the error term. Note that - unlike Bühlmann and Yu (2002) - we re-estimate the model after variable selection.

Theorem 1. [Asymptotic Properties of Forecasts] Suppose that Assumptions 1 (a)–(g) hold. Then

$$\begin{aligned}
T^{1/2}\hat{y}^{UR}(x) &\xrightarrow{d} \xi' x, && (3)\\
\hat{y}^{NC}(x) &= 0, && (4)\\
T^{1/2}\hat{y}^{PT}(x) &\xrightarrow{d} \xi' S'[SE(x_t x_t')S']^{-1} Sx I(|\xi_j| > \sqrt{\Sigma_{jj}}c \text{ for some } j), && (5)\\
T^{1/2}\hat{y}^{BA}(x) &\xrightarrow{d} E\{\eta' S^{*'}[S^* E(x_t x_t')S^{*'}]^{-1} S^* x I(|\eta_j| > \sqrt{\Sigma_{jj}}c \text{ for some } j)|\xi\} && (6)
\end{aligned}$$

where $\xi \sim N(\delta, \Sigma)$, $\Sigma = [E(x_t x_t')]^{-1}\Omega[E(x_t x_t')]^{-1}$, S is the stochastic selection matrix obtained from

$$\begin{bmatrix} I(|\xi_1| > c\sqrt{\Sigma_{11}}) & 0 & \cdots & 0 \\ 0 & I(|\xi_2| > c\sqrt{\Sigma_{22}}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I(|\xi_M| > c\sqrt{\Sigma_{MM}}) \end{bmatrix}$$

by deleting rows of zeros, Σ_{jj} is the (j,j)-th entry of Σ, S^* is defined as S with ξ replaced by η, and $\eta|\xi \sim N(\xi, \Sigma)$.

Proof of Theorem 1.

The result for the no-change predictor holds trivially. Since $T^{1/2}\hat{\beta} \xrightarrow{d} N(\delta, \Sigma) \equiv \xi$ by Assumptions 1(a) and (b), the result for the unrestricted predictor follows immediately. The result for the pretest predictor follows from Assumptions 1(a)-(c) and the continuous mapping theorem because the set of discontinuity points is of measure zero. The result for the bagging predictor follows from Assumptions 1(d)-(f) given that Assumption 1(g) implies the uniform integrability in probability of $\widehat{\gamma}^{*\prime} S_T^*$.

From the viewpoint of economic forecasting it is instructive to study the properties of the prediction mean squared error (PMSE) of these forecasting models. Note that in general

$$\begin{aligned}
E[(y_{T+1} - \hat{y}(x_{T-h+1}))^2|x_{T-h+1}] &= E[(y_{T+1} - E(y_{T+1}|x_{T-h+1}))^2|x_{T-h+1}] \\
&\quad +[E(y_{T+1}|x_{T-h+1}) - \hat{y}(x_{T-h+1})]^2 \\
&= \sigma^2 + [E(y_{T+1}|x_{T-h+1}) - E(\hat{y}(x_{T-h+1}))]^2 \\
&\quad +Var(\hat{y}(x_{T-h+1})|x_{T-h+1}), \qquad (7)
\end{aligned}$$

where $\sigma^2 = E\left[y_{T+1} - E(y_{T+1}|x_{T-h+1})\right]^2$. The last expression shows that the PMSE can be decomposed into three terms: the population PMSE of the forecast model, the squared bias of the forecasts and the variance of the forecasts. By choosing between different forecast models, the forecaster may be able to reduce the second term or the third term, but the first term is beyond the forecaster's control. We will refer to the sum of the last two terms as the mean-squared error (MSE) of the predictor. The aim is to choose a predictor that minimizes this MSE. The MSE expressions implied by Theorem 1 are too complicated to be compared analytically in general. In the next subsection we will therefore simplify the model to allow us to build intuition for the likely performance of bagging predictors compared to other predictors..

2.2 Why Bagging Tends to Work in Multiple Regression

The fundamental problem in choosing a forecast model is that of resolving the bias-variance trade-off that arises when predictors are weak in the sense described above. Clearly, there is a gain from choosing a more parsimonious model than the true model when the bias from underfitting is small relative to the reduction in estimation uncertainty. On the other hand, when the predictor is sufficiently strong, the bias from underfitting will outweigh the variance reduction. We illustrate this phenomenon with a stylized example involving only a single regressor. The example is based on Bühlmann and Yu (2002). Suppose that $\beta = \delta T^{-1/2}$, $x_t = 1 \; \forall t$, ε_t is distributed $iid(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 1$. For expository purposes, we will focus on $h = 1$. We also will assume that $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \Sigma + o(1)$ in computing the PMSEs. Then the forecasts from the unrestricted model, the zero mean (or no change) model and the pre-test model can be written as

$$\begin{aligned} \hat{y}^{UR} &= \hat{\beta}, \\ \hat{y}^{NC} &= 0, \\ \hat{y}^{PT} &= \widehat{\beta} I(|T^{1/2}\hat{\beta}| > 1.96), \\ \hat{y}^{BA} &= \frac{1}{B}\sum_{i=1}^{B} \widehat{\beta}^{*i} I(|T^{1/2}\hat{\beta}^{*i}| > 1.96). \end{aligned}$$

By an application of Theorem 1

$$\begin{aligned} T^{1/2}\hat{y}^{UR} &\overset{d}{\rightarrow} (\delta + z), \\ T^{1/2}\hat{y}^{NC} &= 0, \\ T^{1/2}\hat{y}^{PT} &\overset{d}{\rightarrow} (\delta + z) I(|\delta + z| > 1.96), \\ T^{1/2}\hat{y}^{BA} &\overset{d}{\rightarrow} (\xi - \xi\Phi(1.96 - \tilde{\xi}) + \sqrt{\Sigma}\phi(1.96 - \tilde{\xi}) + \xi\Phi(-1.96 - \tilde{\xi}) \\ &\quad\; -\sqrt{\Sigma}\phi(-1.96 - \tilde{\xi})), \end{aligned}$$

where $z \sim N(0, 1)$ and $\widetilde{\xi} = \xi/\sqrt{\Sigma}$. Note that the asymptotic PMSE in this model is of the form $E[(y_{T+1} - \hat{y}(x_T))^2 | x_T] = \sigma_\varepsilon^2 + \text{bias}^2 + \text{variance}$. We will focus on the last two terms.

2.2.1 Comparing the MSE of the Pre-test Predictor, the No-change Predictor and the Unrestricted Predictor

We are interested in evaluating how the bias and variance of the UR, NC and PT predictors affects their MSE. The magnitude of the MSE of each predictor is a function of the Pitman drift,

δ. Given the bias-variance trade-off described earlier, the MSE ranking of the unrestricted and the no-change forecast model will not be the same for all δ. The drift term δ in turn is directly linked to the population R^2 of the forecast model for finite T by $R^2 = \delta'\delta/(T+\delta'\delta)$.

For example, the MSE of the no-change forecast is a monotonically increasing function of δ. In the limit, when the $\delta = 0$, the no-change forecast has zero bias *and* zero variance and hence zero MSE. Although its variance is always zero, its bias equals δ and hence increases linearly with δ. The unrestricted forecast, in contrast, has the same MSE of 1 for all δ, because it will be unbiased with constant variance of 1. Thus, for $\delta > 1$ the unrestricted predictor has lower MSE than the no-change predictor, for $\delta = 1$ both models are tied and for $\delta < 1$ the no-change model is asymptotically more accurate (see Figure 1).

This result suggests that perhaps by selecting a subset of the predictors in the unrestricted model, we may be able to improve on the forecast accuracy of the unrestricted and the no-change model. Figure 1 shows that this strategy does not work in general because the pre-test forecast is always dominated by the unrestricted or the no-change model. Intuitively, this happens because the third term in the asymptotic PMSE expression is inflated by the excess variability induced by testing.[3]

2.2.2 Comparing the MSE of the Bagging Predictor and the Pre-test Predictor

One way of improving on the pre-test predictor is to bootstrap-aggregate the pre-test estimator. It may be shown that in the Bühlmann and Yu (2002) example that the PMSE of the pre-test predictor and of the bagging predictor, respectively, are

$$\begin{aligned}
T\ E[(E(y_{T+1}) - \hat{y}^{PT})^2] &= E[(\xi-\delta)I(\widetilde{|\xi|} > 1.96) + \delta I(\widetilde{|\xi|} \le 1.96))]^2 \\
&\quad + o(1), && (8) \\
T\ E[(E(y_{T+1}) - \hat{y}^{BA})^2] &= E[(\delta - \xi + \xi\Phi(1.96-\tilde{\xi}) - \sqrt{\Sigma}\phi(1.96-\tilde{\xi}) \\
&\quad - \xi\Phi(-1.96-\tilde{\xi}) + \sqrt{\Sigma}\phi(-1.96-\tilde{\xi}))]^2 \\
&\quad + o(1). && (9)
\end{aligned}$$

Note that the asymptotic PMSE expression for the bagging predictor does not depend on the indicator function, reflecting the smoothing implied by bootstrap aggregation. Although this smoothing should typically help to reduce the variance of the predictor relative to the pre-test predictor, it is not obvious a priori whether bagging the pre-test predictor will also improve the PMSE. Intuitively, we would expect bagging to reduce the PMSE by reducing the third term of the PMSE expression, while leaving the second term largely unchanged.

Figure 2 compares the performance of the pre-test predictor and the bagging predictor for this model. The upper panel of Figure 2 shows the squared bias of the predictors. Although bagging actually reduces the bias somewhat for most values of δ, the gains are small. Basically, bagging does not affect bias very much in one or the other direction. The second panel, in

[3] For a related discussion of the MSE of inequality constrained estimators see Thomson and Schmidt (1982).

contrast, shows dramatic reductions in variance relative to the pre-test estimator for most δ, which, as shown in the third panel, results in substantial improvements in the overall accuracy measured by the PMSE, consistent with Breiman's conjecture. Figure 2 illustrates the potential of the bagging principle to improve forecast accuracy relative to the pre-test. Although this improvement does not occur for all values of δ, it does for a wide range of δ.

2.2.3 Comparing the MSE of the Bagging Predictor, the No-change Predictor and the Unrestricted Predictor

Given that bagging has been shown to improve the accuracy of pre-tests in Figure 2, one might hope that this improvement would perhaps be large enough to beat the unrestricted and no-change forecast models, but the analysis of our stylized example suggests otherwise. Although bagging indeed asymptotically improves on pre-testing for all but the smallest values of δ, as Figure 1 shows, bagging in turn is dominated by the no-change forecast for low values of δ and by the unrestricted model for large values of δ. Only for a very small range of δ values near one is bagging the asymptotically best strategy. This conclusion is disturbing in that we clearly cannot count on being in the small region for which bagging works, when doing applied work. This conclusion holds for regression with one predictor as well as for regression with many predictors, when all predictors are orthonormal with the same value of δ.

2.2.4 On the Role of Heterogeneity in the Predictors

There are reasons, however, to be more optimistic about the likely performance of bagging in practice. Suppose that instead of having multiple predictors with the same δ we have multiple predictors with different values of δ, say, three orthonormal predictors with $\delta_1 = 0$, $\delta_2 = 1$, and $\delta_3 = 2$. If we evaluate those predictors at $x_{iT} = 1$, $i = 1, 2, 3$, then we may read off their respective asymptotic MSEs from Figure 1. For each forecast model, the combined forecast MSE may be computed as the sum of the forecast MSEs for each δ_i, provided that ε_t is i.i.d. The results of this exercise are shown in Table 1. Note that the combined bagging forecast MSE may be lower than the combined MSE of either the unrestricted or the no-change forecast, even when not all δ_i are near 1. In fact, in our example two out of three δ_i are far from 1. Nevertheless, the combined bagging MSE is only 2.53 compared to 3, 3.97 and 5 for the other predictors. This example illustrates that - at least under i.i.d. innovations - we would expect bagging to work for a much wider range of δ, when the orthonormal predictors are heterogeneous in the sense of having different slope parameters.

We conjecture that heterogeneity in the predictors is likely to improve the accuracy of bagging forecasts, even when the predictors are not orthonormal and the innovations are not i.i.d., making the computation of the combined MSE nontrivial. In this more general case, we may define heterogeneity as follows:

Definition 2.[Heterogeneity] Let $y_{t+h} = \beta' x_t + \varepsilon_{t+h} = \beta'\Sigma^{1/2}\Sigma^{-1/2}x_t + \varepsilon_{t+h} = \widetilde{\beta}'_T\widetilde{x}_t + \varepsilon_{t+h}$, *where* $\widetilde{\beta}'_T = \beta'\Sigma^{1/2}$ *and* $\Sigma = [E(x_t x_t')]^{-1}\sum_{j=0}^{\infty} E(x_t\varepsilon_{t+h}\varepsilon_{t+h-j}x_{t-j})[E(x_t x_t')]^{-1}$. *Then a vector of predictors is said to be heterogeneous if*

$$\widetilde{\beta}_T' \neq \kappa \ell_M$$

holds for any $\kappa \epsilon \mathbb{R}$ / $\{0\}$ where ℓ_M is an M-dimensional vector of ones.

The existence of such heterogeneity in empirical applications seems likely, if for no other reason than that there are differences in the quality of predictor data and in their a priori relevance (see, e.g., Boivin and Ng 2003). The conjecture that heterogeneity in this sense tends to improve the performance of bagging predictors is supported by simulation evidence, as we will show shortly. Nevertheless, it is important to be clear that heterogeneity in this sense is neither necessary nor sufficient for bagging forecasts to be more accurate than forecasts from the restricted, the unrestricted and the pre-test model.

It is not necessary because, as we have shown, it is entirely possible that bagging forecasts are the most accurate forecasts even in the homogeneous case. This situation arises, for example, when $\delta_i = 1.1 \; \forall i$ in the orthonormal model underlying Table 1. Conversely, the existence of heterogeneity does not guarantee that the bagging forecast has the lowest MSE. Rather than being universal, the accuracy gains from bagging - and indeed the ranking of all methods - will depend on design features such as the population R^2 of the forecast model (as a scalar summary measure of the vector δ) and the number of predictors, M. Although we do not present detailed results for the performance of bagging as a function of M, we note that sizable gains in accuracy may arise in practice for as few as five predictors. The extent of these gains depends crucially on the population R^2 (or, equivalently, the vector δ).

This point again may be illustrated in the context of the simple orthonormal model of Table 1. For example, when $\delta_1 = 0$, $\delta_2 = 0.1$, and $\delta_3 = 0.2$, both the pre-test forecast and the no change forecast will be more accurate than the bagging forecast. When $\delta_1 = 1.9, \delta_2 = 2$,and $\delta_3 = 2.1$, in contrast, the bagging forecast has lower MSE than the no-change and pre-test forecasts, but higher MSE than the unrestricted forecast. In neither case, the bagging forecast has the lowest MSE because the δ_i are clustered entirely to the left or entirely to the right of the range where bagging yields improvements (see Figure 1). It is only when the δ includes values to both sides of that range that heterogeneity may help boost the performance of bagging. For example, if we had two orthonormal predictors with $\delta_1 = 0$, and $\delta_2 = 2$,the MSE of the bagging forecast would be 1.76 compared with an MSE of 2.71 for the pre-test forecast, 3 for the forecast from the unrestricted model and 4 for the no-change forecast.[4]

Although there is no compelling reason that heterogeneity alone will benefit bagging, we will show that it typically does. The following two figures illustrate both the role of the population R^2 in bagging multiple regression models and the extent to which heterogeneity may extend the range for which bagging works well. All results were computed based on $M = 30$ orthonormal predictors for a sample size of $T = 100$.

Figure 3 is the direct multiple-regression analogue of the analysis of the single-regressor model in Figure 2. We postulate M orthonormal predictors with the same slope parameter.

[4] It would be of obvious interest for empirical work to design a test of whether bagging can be expected to improve forecast accuracy, but, since the vector δ cannot be estimated consistently, it is not possible to design such a test. This means that, although we can describe the conditions under which bagging can be expected to improve forecast accuracy, these conditions cannot be verified in practice.

Thus, the predictors are homogeneous in the sense of Definition 2, and there is a one-for-one mapping between $\delta_i = \delta$, $i = 1, ..., M$, and the population R^2. As expected, the results in Figure 3 are qualitatively the same as in Figure 2. For small values of R^2 the bagging forecast is dominated by the no-change forecast and for large values of R^2 the unrestricted model dominates the bagging model. Only for $0.2 < R^2 < 0.5$ the bagging predictor is the best choice (see vertical bars). Thus, the range of R^2 values for which bagging can be expected to work is rather small.

In contrast, Figure 4 shows the same experiment with different slope parameters for each orthonormal predictor. Thus, the predictors are heterogeneous in the sense of Definition 2.[5] Figure 4 shows that under heterogeneity the bagging predictor dominates the other predictors for $0.17 < R^2 < 0.93$ (see vertical bars). The range for which bagging works best more than doubles, improving our confidence that bagging would perform well in practice. This simulation example illustrates our point about the beneficial effects of heterogeneity on the likely performance of bagging in multiple regression problems. Note that we do not claim that the particular ranges found in this example are representative for applied work. Rather our point is that heterogeneity will tend to increase the scope for improvements in forecast accuracy from bagging.

3 Simulation Evidence

We now turn to a simulation study to investigate the potential of bagging to improve forecast accuracy relative to the unrestricted forecast model, various restricted forecast models and the pre-test forecast model. An additional question of practical interest is whether the bagging strategy is competitive with alternative approaches to forecasting from large data sets such as forecast combinations or forecasts from dynamic factor models. We focus on the latter alternative, because of earlier evidence that dynamic factor models tend to be more accurate than mean or median-based combination forecasts (see Stock and Watson 1999). We postpone for future research a comparison with Bayesian methods of regression-based forecast combination (see, e.g., Wright 2003).

Our aim here is not to argue that bagging will in general be superior to those alternative approaches. Clearly, the design of our study is too limited to make such a case. Rather we want to illustrate the importance of various design features that will affect the relative performance of bagging and other forecasting techniques. We also want to illustrate the potential gains from bagging and we want to show that the bagging approach is not dominated by existing forecasting techniques in situations when M is large, but still much smaller than T.

3.1 Simulation Design

All simulations in this paper are based on $M = 30$ and $T = 100$. This setting is intended to capture the assumption that the number of predictors is large, but distinctly smaller than the sample size. We postulate that

$$y_{t+1} = \beta' x_t + \varepsilon_{t+1} \tag{10}$$

[5] Specifically, we use design 3 described in section 3.1. Similar results are obtained for other heterogeneous designs. Figures 3 and 4 are based on averages across 5000 trials with B=100.

where $\varepsilon_t \sim NID(0,1)$. The innovation variance of ε_t can be set to one without loss of generality, since we will scale the variance of $\beta' x_t$ to maintain a given R^2 of the forecast model. We will consider $R^2 = 0.25$ and $R^2 = 0.5$ in the simulation study.

3.1.1 Design of Slope Parameters in Forecast Model

The first design issue is the choice of the slope parameter vector β. As discussed in the previous section, the relative performance of bagging is expected to improve when the predictors are heterogeneous in the sense of Definition 2. We attempt to capture this heterogeneity in our simulation design by exploring a number of different profiles for the slope parameters. As a benchmark we include a vector of constants in design 1. When the regressors are orthonormal, this design imposes homogeneity in the sense of Definition 2. The other six designs imply heterogeneity. Designs 4 and 5 are step functions. The remaining designs incorporate smooth decays, some slow and others (like the exponential design 6) very rapid decays, resulting in a few regressors with relatively high predictive power and many regressors with negligible predictive power.

Design 1. $\beta = c_1[1,1,...,1]'$.
Design 2. $\beta = c_2[30,29,28...,1]'$.
Design 3. $\beta = c_3[1,1/2,1/3...,1/30]'$.
Design 4. $\beta = c_4[1_{1\times 15},0_{1\times 15}]$.
Design 5. $\beta = c_5[1_{1\times 8},0_{1\times 22}]$.
Design 6. $\beta = c_6[e^{-1},e^{-2},...,e^{-30}]'$.
Design 7. $\beta = c_7[\sqrt{30},\sqrt{29},...,1]'$.

The scaling constants c_i, $i = 1,..,7$, are chosen, given the variance of x_t, such that the population R^2 of the forecasting model is the same across all profiles.

3.1.2 Design of Regressor Matrix

The second design issue is the data generating process for the vector of predictors, x_t. For expository purposes we begin by postulating that the predictors are uncorrelated Gaussian white noise.

Case 1.

$$x_t \sim NID(0, I_{30})$$

While instructive, this first case, in which all predictors are orthonormal, is implausible in that most economic data show a fair degree co-movements. It is this co-movement that motivated the development of factor models. The remaining data generating processes for x_t are therefore based on factor models with ranks of $r \epsilon \{1,2,3,4\}$.

Case 2.

$$x_t = \Lambda F_t + \eta_t$$

where $F_t \sim N(0, I_r)$, $\eta_t \sim N(0_{30\times1}, I_{30})$, *and* Λ is an Mxr matrix of parameters.

Since we do not know the value of Λ, we replace the elements of the parameter matrix Λ by independent random draws from the standard normal distribution. For each draw of Λ we compute the root PMSE for each model based on 5,000 Monte Carlo trials. Since the results may differ across draws, in the simulation study we report average root PMSE ratios based on 30 draws of Λ.

3.1.3 Controlling the Strength of the Factor Component in the Predictors

A third design feature is the relative importance of the idiosyncratic component η_t relative to the factor component ΛF_t in the DGP for x_t. We measure the explanatory power of the factor component by the pseudo-R^2 measure

$$R^2_{pseudo} = \frac{tr(\Lambda\Lambda')}{tr(\Lambda\Lambda' + \Sigma_{\eta_t})},$$

where Σ_{η_t} denotes the covariance matrix of η_t and tr denotes the trace operator. We chose this

ratio to match the pseudo-R^2 measure found in our empirical application in section 4. In the limit, for $R^2_{pseudo} = 0$, the factor model reduces to the orthonormal model of case 1. The data suggest values of approximately $R^2_{pseudo} = 0.2$ for $r = 1$, $R^2_{pseudo} = 0.3$ for $r = 2$, $R^2_{pseudo} = 0.4$ for $r = 3$, and $R^2_{pseudo} = 0.5$ for $r = 4$.

3.2 Simulation Results

The simulation results are presented in Tables 2 and 3. For each panel of the table, we normalize the results relative to the root PMSE (RPMSE) of the true model. We show results for the unrestricted model, the zero-mean (or no change) model, the estimated mean (or constant-change) model, the pre-test model discussed earlier, and the bagging model. Bagging results are based on $B = 100$ throughout. We also investigate factor models with rank $r \epsilon \{1, 2, 3, 4\}$. The factor model forecasts are based on the regression

$$y_{t+1} = \alpha + \phi(L)y_t + \theta(L)\ \widehat{F}_t + \varepsilon_{t+1}$$

with $\beta(L) = 0$ and $\theta(L) = \theta$ imposed in the simulation study.

3.2.1 Case 1: Orthonormal Predictors

Table 2 shows the results for a population R^2 of 0.25 and Table 3 the corresponding results when the slope coefficients have been scaled to imply $R^2 = 0.5$. The first panel of each table presents results for orthonormal white noise predictors. This case is interesting primarily because it is closest to the simplified assumptions used in section (2.2) when we discussed the intuition for bagging. For $R^2 = 0.25$ bagging improves on the true model for all designs with gains in the range of 10 to 20 percentage points of the RPMSE. Bagging also improves on the pre-test forecast with two exceptions. For design 5, bagging and pre-test forecasts are tied; for the exponential design 6, pre-test forecasts are more accurate than bagging forecasts. Both are much more accurate than the alternatives. As expected, the bagging forecast is more accurate than the factor model forecast for all designs. This is not surprising since there is no factor structure in population. Nevertheless, even factor models routinely outperform the true model, reflecting a favorable bias-variance trade-off. The additional gains from bagging range from 2 to 10 percentage points of the RPMSE of the true model.

For $R^2 = 0.5$, in contrast, imposing incorrect factor structure harms the factor models, indicating that the bias induced by imposing a factor structure outweighs the reduction in variance. The constant-change and no-change models perform poorly for the same reason. Bagging forecasts are once again more accurate than the true model with one important exception. For design 1, the true unrestricted model is even more accurate than the bagging model. This result underscores our theoretical point about the importance of heterogeneity in the predictors. Design 1 mimics the situation in which all predictors are perfectly homogeneous. In contrast, all other design incorporate varying degrees of heterogeneity.

3.2.2 Case 2: Common Factors among Predictors

The simulation results for the first case confirm the view that bagging will tend to work well when the predictors are heterogeneous. Case 1, however, is unrealistic in that it treats the predictors as uncorrelated. In economic applications most predictors show co-movement of varying degrees. This co-movement may be approximated by factor models. We therefore will focus on factor model data generating processes for the predictors in the remaining panels of Tables 2 and 3. We begin with case when the true model is a factor model of rank 1. In that case, for $R^2 = 0.25$, the unrestricted, no-change and constant-change forecasts perform poorly relative to the true model. Factor model forecasts perform well, regardless of the rank imposed. Pre-test forecasts perform erratically. In contrast, bagging forecasts do well across the board. They are more accurate than forecasts from any factor model considered, regardless of the design, with percentage gains close to 10 percentage points in some cases. Turning to the results for $R^2 = 0.5$ in Table 3, we see that the relative advantages of bagging forecasts increase further with percentage gains of more than 30 percentage points relative to the true factor model in some cases. Interestingly, even the unrestricted model outperforms the factor model in this case, although not by as much as the bagging forecast. These results reflect the relatively low R^2_{pseudo} for rank 1 models, which makes it hard to extract reliably the true factor structure in small samples. The bagging method does not impose any structure and hence is more robust.

The third panel in Tables 2 and 3 shows qualitatively similar results for rank 2 data generating processes. As the rank increases further, the results for $R^2 = 0.25$ become more mixed. In many cases, the factor model is somewhat more accurate than bagging, but only if the researcher imposes a rank close enough to the true rank. Typically, underestimation of the rank results in increases in the RPMSE. Although not the best forecast model in all cases, bagging remains quite competitive in most cases. It always outperforms some factor models and all other forecast models under consideration. In two of six heterogeneous designs it even outperforms all factor models. Moreover, for $R^2 = 0.5$ the bagging forecast is more accurate than any of the factor models, in some cases by more than 20 percentage points.

3.2.3 Sensitivity Analysis

All results in Tables 1 and 2 are based on the same choices for R^2, R^2_{pseudo} and M. We conclude this subsection with some sensitivity analysis that illustrates the role of these design parameters. Figures 5 and 6 are based on a representative draw from the rank 1 factor model data generating process for design 3. These simulation examples are not intended as concrete advice for practitioners, but are designed to illustrate the trade-offs that govern the ranking of bagging forecasts and factor model forecasts in practice. Figure 5 shows that the gains in accuracy from bagging forecasts decline - relative to using the best factor model forecast among models with $r \in \{1, 2, 3, 4\}$ - as M increases. This result simply reflects the fact that - all else equal - a larger M allows the more precise estimation of the factor component.

The ranking itself also depends on the population R^2 of the unrestricted forecast model. Low R^2 favors factor models because in that case the gains from imposing the correct factor structure are largest. In the example, for $R^2 = 0.25$ the best factor model outperforms bagging for M in excess of about 38; for $R^2 = 0.5$ bagging forecasts remain the more accurate forecasts even for $M = 45$, but here as well the gains from bagging decline with M.

Similarly important is the question of how strong the factor component in the predictor data is. Figure 6 shows that the gains in accuracy from bagging decline relative to the best factor model, as R^2_{pseudo} increases, as one would expect when the factor model is the true model. Again the range of R^2_{pseudo}, for which bagging is more accurate than factor models increases with R^2.

While the simulation results are encouraging, we have no way of knowing a priori whether the data generating process in a given empirical application will favor bagging or factor model forecasts because that ranking will depend on unknown features of the data generating process. Given the difficulty of generalizing the results of our simulation study, we recommend that, in practice, researchers choose between the bagging strategy and the dynamic factor model strategy based on the ranking of their recursive PMSE in simulated out-of-sample forecasts. The model with the lower recursive PMSE up to date $T - h$ will be chosen for forecasting y_{T+1}. We will illustrate this approach in the next section for a typical forecast problem in economics.

4 Application: Do Indicators of Real Economic Activity Improve the Accuracy of U.S. Inflation Forecasts?

We investigate whether one-month and twelve-months ahead U.S. CPI inflation forecasts may be improved upon by adding indicators of real economic activity to models involving only lagged inflation rates. This empirical example is in the spirit of recent work by Stock and Watson (1999), Bernanke and Boivin (2003), Forni et al. (2003), and Wright (2003). The choice of the benchmark model is conventional (see, e.g., Stock and Watson 2003, Forni et al. 2003). The lag order of the benchmark model is determined by the AIC subject to an upper bound of 12 lags. The optimal model is determined recursively in real time, so the lag order may change as we move through the sample.

Since there is no universally agreed on measure of real economic activity we consider 26 potential predictors that can be reasonably expected to be correlated with real economic activity. A complete variable list is provided in the Data Appendix. We obtain monthly data for the United States from the Federal Reserve Bank of St. Louis data base (FRED). We convert all data with the exception of the interest rates into annualized percentage growth rates. Interest rates are expressed in percent. Data are used in seasonally adjusted form where appropriate. All predictor data are standardized (i.e., demeaned and scaled to have unit variance and zero mean), as is customary in the factor model literature. We do not attempt to identify and remove outliers.

The alternative forecasting strategies under consideration include the benchmark model involving only lags of monthly inflation and seven models that include in addition at least some indicators of economic activity. The unrestricted (UR) model includes one or more lags of all 26 indicators of economic activity as separate regressors in addition to lagged inflation. The pre-test (PT) model uses only a subset of these additional predictors. The subset is selected using 2-sided t-tests for each predictor at the 5% significance level. Forecasts are generated from the subset model. The bagging (BA) forecast is the average of these pre-test predictors across 100 bootstrap replications. For the one-month ahead forecast model there is no evidence of serial correlation in the unrestricted model, so we use White (1980) robust standard errors for the pre-tests and the pairwise bootstrap. For the twelve-month ahead-forecast we use West (1997) standard errors with a truncation lag of 11 and the blocks-of-blocks bootstrap with $m = 12$. Finally, we also fit factor models with rank $r \; \epsilon \; \{1, 2, 3, 4\}$ to the 26 potential predictors and generate forecasts by adding one or more lagged values of this factor to the benchmark model (DFM).

We compute results for the UR, PT, and BA methods for up to three lags of the block of indicator variables in the unrestricted model. Note that adding more lags tends to result in near-singularity problems, when the estimation window is short. Even for three lags of the 26 indicator variables, there are near-singularity problems at the beginning of the recursive sample. We also show results based on the SIC with an upper bound of 2 lags. For larger upper bounds, near-singularity problems tend to arise at the beginning of the sample. In contrast, dynamic factor models are more parsimonious and hence allow for richer dynamics. We show results for models including up to five additional lags of the estimated factor. We also allow the lag order q to be selected by the SIC. The SIC generally produced more accurate forecasts than the AIC. The results are robust to the upper bound on the lag order.

To summarize, the forecast methods under consideration are:

$$
\begin{aligned}
Benchmark &: \quad \pi^h_{t+h|t} = \widehat{\alpha} + \sum\nolimits_{k=1}^{p} \widehat{\phi}_k \pi_{t-k} \\
UR &: \quad \pi^h_{t+h|t} = \widehat{\alpha} + \sum\nolimits_{k=1}^{p} \widehat{\phi}_k \pi_{t-k} + \sum\nolimits_{l=1}^{q} \sum\nolimits_{j=1}^{M} \widehat{\beta}_{jl} x_{jt-l+1} \\
PT &: \quad \pi^h_{t+h|t} = \widehat{\alpha} + \sum\nolimits_{k=1}^{p} \widehat{\phi}_k \pi_{t-k} + \sum\nolimits_{l=1}^{q} \sum\nolimits_{j=1}^{M} \widehat{\gamma}_{jl} I(|t_{jl}| > 1.96) x_{jt-l+1} \\
BA &: \quad \pi^h_{t+h|t} = \frac{1}{100} \sum_{i=1}^{100} \left(\widehat{\alpha}^* + \sum\nolimits_{i=1}^{p} \widehat{\phi}^*_i \pi_{t-i} + \sum\nolimits_{l=1}^{q} \sum\nolimits_{j=1}^{M} \widehat{\gamma}^*_{jl} I(|t^*_{jl}| > 1.96) x_{jt-l+1} \right) \\
DFM &: \quad \pi^h_{t+h|t} = \widehat{\alpha} + \sum\nolimits_{k=1}^{p} \widehat{\phi}_k \pi_{t-k} + \sum\nolimits_{l=1}^{q} \widehat{\theta}_l \ \widehat{F}_{t-l+1}
\end{aligned}
$$

where π^h_{t+h} denotes the rate of inflation over the period t to $t+h$.

The accuracy of each method is measured by the square root of the average of the squared forecast errors ($RPMSE$) obtained by recursively re-estimating the model at each point in time and forecasting π^h_{t+h}. Note that we also re-estimate the lag orders at each point in time, unless noted otherwise. The evaluation period consists of 240 observations covering the most recent twenty years in the sample. Table 4 shows the results for one-month ahead forecasts of U.S. CPI inflation ($h = 1$). The best results for each method are shown in bold face. Table 4 shows that bagging the pre-test is by far the most accurate forecasting procedure. The bagging forecast outperforms the benchmark autoregressive model, the unrestricted model and factor models with rank 1, 2, 3, or 4. The gains in forecast accuracy are 16 percentage points of the RPMSE of the AR benchmark. Dynamic factor models, in contrast, outperform the benchmark at best by 3 percentage points. These results are robust to extending or shortening the evaluation period of 240 observations.

One would expect that imposing the factor structure becomes more useful at longer forecast horizons. Table 5 shows the corresponding results for a horizon of twelve months ($h = 12$). In that case, the benchmark model no longer is an autoregression. Here as well the bagging forecast is by far the most accurate forecast. The accuracy gains are even larger with 44 percentage points relative to the benchmark model. Dynamic factor models also perform well, as expected, but the best factor model is still less accurate than the bagging model by 11 percentage points. This result is perhaps surprising in that the dynamics allowed for in bagging are much more restrictive than for factor models. Using the SIC for selecting the lag order q at each point in time does not necessarily improve the accuracy of the forecast relative to fixed lag structures.

We also computed the forecast accuracy of the median forecast from all possible models including just one indicator of economic activity at a time in addition to lagged inflation rates. We found that this combination forecast, was typically inferior to the best factor model forecast and always inferior to the bagging forecast. This finding is consistent with the results of Stock and Watson (1999), who investigated a number of different methods of combining inflation forecasts and found them to be inferior to dynamic factor model forecasts.

5 Conclusion

Bagging is designed for situations, in which there is a moderately large number of predictors relative to the sample size. The bagging method discussed here is likely to be useful in a variety of contexts. We may be interested in forecasting macroeconomic aggregates from sectoral or regional components or from latent variables that are imperfectly measured by observables. Apart from the inflation examples used in this paper, the same tools may be applied, for example, to predict stock returns, exchange returns, growth rates of output, sales, productivity or consumption, leading economic indicators or unemployment rates.

Like other methods of combining information from many predictors - such as dynamic factor models or forecast combination methods - bagging does not necessarily improve forecast accuracy in all cases. Nevertheless, bagging has the potential to yield substantial improvements in accuracy in practice. We showed that it tends to perform better than the unrestricted forecast model as well as the fully restricted model under mild conditions. Bagging also tends to perform better than the pre-test forecast model. Finally, it often performs better than factor models, even when the true model has a factor structure. The simulation and empirical evidence, however, should not be interpreted to mean that bagging should routinely replace dynamic factor models. Rather the bagging method should be viewed as complementary to the use of dynamic factor models in forecasting for two reasons.

First, dynamic factor models are designed for data sets with large M relative to T and may not work well, unless M is large enough.[6] In contrast, bagging is designed for forecasting situations in which M is large, but clearly smaller than T. In this sense, one might say that bagging is designed to work in precisely those situations, in which the standard justification for forecasting from estimated dynamic factor models is questionable. Conversely, bagging cannot be expected to work well in situations for which dynamic factor models have been designed. Bagging becomes infeasible when the number of predictors, M, is too large relative to T, because near-singularity problems arise in computing the least-squares estimator of the unrestricted model, on which the bootstrap approximation is based.[7]

Second, as we have shown, the ranking of bagging forecasts and forecasts from dynamic factor models will in general depend on features of the unknown data generating process. For example, the relative performance will depend crucially on the strength of the common factor component relative to the idiosyncratic noise component in the set of predictors. When the factor structure is weak, as it appears to be in our empirical example, bagging may be much more accurate in small samples than imposing a factor structure, even when the data are truly generated by the factor model. In contrast, when the data are well approximated by a common factor model, bagging clearly cannot be expected to outperform the dynamic factor model forecast. Given the difficulties of arriving at general results for the ranking of bagging relative to factor model forecasts, we provided some guidance as to how to choose between competing forecast approaches in practice based on simulated out-of-sample forecasts.

[6] Specifically, standard asymptotic theory for forecasts from dynamic factor models postulates that $T/M \to 0$. This is often heuristically interpreted as requiring that M be as large as T or larger for fixed T (see, e.g., Bai and Ng 2003).

[7] In our simulation study, these near-singularity problems for the OLS estimator arose for $M > 50$ when $T = 100$.

An important caveat is that the theory of bagging presented in this paper assumes a covariance stationary environment. This means that all variables must be transformed to stationarity prior to the analysis. It also means that we abstract from the possibility of structural change. The same is true of the standard theory of forecast combination, which relies on information pooling in a stationary environment. In contrast, the theory for factor models does allow for some forms of smooth structural change, although not when $M < T$. An interesting avenue for future research would be the development of bagging methods that allow for smooth structural change. A second direction for future research would be a comparison of bagging forecast methods, which have been derived from a frequentist perspective, with Bayesian methods of forecast model combination of the type recently considered by Wright (2003).

Data Appendix

All data are for the United States. The sample period for the raw data is 1971.4-2003.7. This choice is dictated by data constraints. The variable codes are from FRED:

$INDPRO$	industrial production
$HOUST$	housing starts
$HSN1F$	house sales
$NAPM$	purchasing managers index
$HELPWANT$	help wanted index
TCU	capacity utilization
$UNRATE$	unemployment rate
$PAYEMS$	nonfarm payroll employment
$CIVPART$	civilian participation rate
$AWHI$	average weekly hours
$MORTG$	mortgage rate
$MPRIME$	prime rate
$CD1M$	1-month CD rate
$FEDFUNDS$	Federal funds rate
$M1SL$	M1
$M2SL$	M2
$M3SL$	M3
$BUSLOANS$	business loans
$CONSUMER$	consumer loans
$REALN$	real estate loans
$EXGEUS$	DM/USD rate (extrapolated using the Euro/USD rate)
$EXJPUS$	Yen/USD rate
$EXCAUS$	Canadian Dollar/USD rate
$EXUSUK$	USD/British Pound rate
$OILPRICE$	WTI crude oil spot price
$TRSP500$	SP500 stock returns

References

1. Andrews, D.W.K., (2002), "Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators," *Econometrica*, 70, 119–162.

2. Bai, J., and S. Ng (2003), "Confidence Intervals for Diffusion Index Forecasts with a Large Number of Predictors" mimeo, Department of Economics, University of Michigan.

3. Bernanke, B.S., and J. Boivin (2003), "Monetary Policy in a Data-Rich Environment," *Journal of Monetary Economics*, 50, 525-546.

4. Billingsley, P. (1995), *Probability and Measure*, 3rd ed. John Wiley and Sons: New York.

5. Boivin, J., and S. Ng (2003), "Are More Data Always Better for Factor Analysis?" mimeo, Department of Economics, University of Michigan.

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

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

8. Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), "The Generalized Factor Model: Identification and Estimation," *Review of Economics and Statistics*, 82, 540-554.

9. Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2001), "The Generalized Factor Model: One-Sided Estimation and Forecasting," mimeo, ECARES, Free University of Brussels.

10. Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2003), "Do Financial Variables Help Forecasting Inflation and Real Activity in the Euro Area," *Journal of Monetary Economics*, 44, 1243-1255.

11. Gallant, A.R. and H. White (1988), *A Unified Theory of Estimation and Inference in Nonlinear Models*, Basil Blackwell: Oxford.

12. Gonçalves, S. and L. Kilian (2003), "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form," forthcoming: *Journal of Econometrics*.

13. Gonçalves, S. and H. White (2003), "Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models," forthcoming: *Journal of Econometrics*.

14. Hall, P. and J.L. Horowitz (1996), "Bootstrap critical values for tests based on generalized method of moments estimators," *Econometrica,* 64, 891–916.

15. Inoue, A., and L. Kilian (2003), "On the Selection of Forecast Models," Working Paper No. 214, European Central Bank.

16. Inoue, A. and M. Shintani (2003), "Bootstrapping GMM Estimators for Time Series," forthcoming: *Journal of Econometrics*.

17. Marcellino, M., J.H. Stock and M.W. Watson (2003), "Macroeconomic Forecasting in the Euro Area: Country-Specific versus Area-Wide Information," forthcoming: *European Economic Review*.

18. Newey, W., and K. West (1987), "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703-708.

19. Politis, D.N., J.P. Romano and M. Wolf (1999), *Subsampling*, Springer-Verlag: New York.

20. Stock, J.H., and M.W. Watson (1999), "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293-335.

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

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

23. Stock, J.H., and M.W. Watson (2003), "Forecasting Output and Inflation: The Role of Asset Prices ," *Journal of Economic Literature*, 41, 788-829.

24. Thomson, M., and P. Schmidt (1982), "A Note on the Comparison of the Mean Square Error of Inequality Constrained Least-Squares and Other Related Estimators ," *Review of Economics and Statistics*, 64, 174-176.

25. West, K. (1997), "Another Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Journal of Econometrics*, 76, 171-191

26. White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test of Heterogeneity," *Econometrica*, 48, 817-838.

27. Wright, J.H. (2003), "Forecasting U.S. Inflation by Bayesian Model Averaging," *International Finance Discussion Papers*, No. 780, Board of Governors of the Federal Reserve System.

Table 1. MSE of Three Heterogeneous Predictors:

Models	$MSE = \sum_{i=1}^{M} MSE(i)$
Unrestricted	3.000
No Change	5.000
Pre-test	3.964
Bagging pre-test	2.530

Source: Asymptotic analysis based on $\delta_1 = 0$, $\delta_2 = 1$, $\delta_3 = 2$ in stylized example with orthonormal predictors evaluated at $x_{iT} = 1$, $i = 1, 2, 3$.

Table 2. Out-of-Sample Forecast Accuracy for R^2= 0.25: RPMSE Normalized Relative to True Model

	Rank 0 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Constant change	0.915	0.917	0.900	0.897	0.890	0.910	0.908
No Change	0.907	0.911	0.894	0.887	0.881	0.902	0.899
Pre-test	0.980	0.963	0.814	0.931	0.887	0.757	0.956
Bagging pre-test	0.877	0.872	0.814	0.859	0.845	0.793	0.869
DFM rank 1	0.901	0.903	0.894	0.888	0.878	0.899	0.896
DFM rank 2	0.896	0.895	0.891	0.884	0.878	0.902	0.891
DFM rank 3	0.893	0.893	0.886	0.881	0.872	0.896	0.889
DFM rank 4	0.890	0.889	0.883	0.878	0.870	0.892	0.887

	Rank 1 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	1.142	1.142	1.143	1.145	1.146	1.142	1.143
Constant change	1.051	1.050	1.034	1.048	1.041	1.021	1.051
No Change	1.041	1.039	1.024	1.038	1.032	1.011	1.041
Pre-test	1.100	1.081	0.947	1.075	1.034	0.884	1.093
Bagging pre-test	0.900	0.984	0.941	0.984	0.972	0.916	0.988
DFM rank 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DFM rank 2	0.997	0.997	0.996	0.997	0.997	0.996	0.997
DFM rank 3	0.995	0.995	0.993	0.996	0.995	0.992	0.995
DFM rank 4	0.994	0.994	0.991	0.995	0.993	0.990	0.994

	Rank 2 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	1.159	1.156	1.159	1.164	1.163	1.168	1.161
Constant change	1.054	1.046	1.062	1.059	1.044	1.068	1.062
No Change	1.044	1.036	1.052	1.048	1.034	1.059	1.052
Pre-test	1.099	1.081	0.962	1.084	1.043	0.907	1.100
Bagging pre-test	0.997	0.990	0.953	0.994	0.981	0.936	0.997
DFM rank 1	1.022	1.019	1.021	1.023	1.018	1.024	1.023
DFM rank 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DFM rank 3	0.999	0.999	0.999	1.000	0.999	1.000	0.999
DFM rank 4	0.997	0.999	0.998	0.998	0.997	0.999	0.997

	Rank 3 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	1.195	1.199	1.191	1.195	1.193	1.188	1.193
Constant change	1.090	1.113	1.075	1.089	1.100	1.083	1.082
No Change	1.079	1.102	1.064	1.078	1.089	1.072	1.071
Pre-test	1.131	1.127	0.988	1.107	1.078	0.922	1.120
Bagging pre-test	1.019	1.019	0.973	1.011	1.001	0.948	1.015
DFM rank 1	1.050	1.060	1.047	1.059	1.059	1.044	1.048
DFM rank 2	1.025	1.031	1.014	1.025	1.015	1.013	1.022
DFM rank 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DFM rank 4	1.001	1.002	1.000	1.000	1.001	1.000	1.001

	Rank 4 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	1.216	1.219	1.211	1.209	1.206	1.213	1.213
Constant change	1.125	1.135	1.096	1.077	1.075	1.101	1.110
No Change	1.114	1.123	1.084	1.066	1.065	1.090	1.099
Pre-test	1.141	1.137	1.009	1.097	1.073	0.949	1.125
Bagging pre-test	1.027	1.027	0.986	1.013	1.003	0.967	1.022
DFM rank 1	1.073	1.075	1.055	1.047	1.048	1.052	1.065
DFM rank 2	1.037	1.036	1.025	1.027	1.021	1.020	1.033
DFM rank 3	1.002	1.011	1.010	1.013	1.010	1.007	1.011
DFM rank 4	1.000	1.000	1.000	1.000	1.000	1.000	1.000

SOURCE: Based on 5000 trials.

Table 3. Out-of-Sample Forecast Accuracy for R^2= 0.5:
RPMSE Normalized Relative to True Model

	Rank 0 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Constant change	1.389	1.392	1.350	1.345	1.338	1.349	1.373
No Change	1.376	1.382	1.345	1.327	1.323	1.343	1.358
Pre-test	1.340	1.176	0.881	1.086	0.889	0.760	1.265
Bagging pre-test	1.012	0.972	0.851	0.944	0.878	0.797	0.990
DFM rank 1	1.347	1.349	1.316	1.312	1.295	1.316	1.333
DFM rank 2	1.319	1.318	1.287	1.287	1.279	1.283	1.308
DFM rank 3	1.297	1.296	1.261	1.264	1.248	1.253	1.286
DFM rank 4	1.272	1.267	1.237	1.240	1.226	1.230	1.264

	Rank 1 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	0.808	0.808	0.807	0.812	0.813	0.807	0.809
Constant change	1.122	1.120	1.081	1.114	1.100	1.056	1.121
No Change	1.111	1.109	1.070	1.103	1.0900	1.045	1.111
Pre-test	1.036	0.942	0.719	0.903	0.756	0.632	0.992
Bagging pre-test	0.795	0.772	0.688	0.764	0.722	0.650	0.786
DFM rank 1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DFM rank 2	0.981	0.982	0.980	0.982	0.981	0.979	0.981
DFM rank 3	0.965	0.965	0.961	0.967	0.965	0.960	0.965
DFM rank 4	0.950	0.951	0.943	0.953	0.949	0.941	0.950

	Rank 2 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	0.8507	0.8482	0.8521	0.8561	0.8532	0.8605	0.8534
Constant change	1.1438	1.1264	1.1631	1.1533	1.1219	1.1760	1.1617
No Change	1.1326	1.1154	1.1519	1.1440	1.1105	1.1651	1.1504
Pre-test	1.0649	0.9719	0.7585	0.9509	0.8069	0.6761	1.0358
Bagging pre-test	0.8268	0.8030	0.7238	0.8007	0.7574	0.6925	0.8204
DFM rank 1	1.0599	1.0540	1.0583	1.0633	1.0513	1.0633	1.0639
DFM rank 2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
DFM rank 3	0.9839	0.9853	0.9846	0.9857	0.9839	0.9860	0.9836
DFM rank 4	0.9693	0.9720	0.9705	0.9706	0.9682	0.9713	0.9684

	Rank 3 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	0.9163	0.9225	0.9096	0.9158	0.9148	0.9062	0.9133
Constant change	1.2461	1.3029	1.2143	1.2459	1.2700	1.2279	1.2274
No Change	1.2333	1.2898	1.2017	1.2334	1.2573	1.2151	1.2149
Pre-test	1.1312	1.0691	0.8084	1.0190	0.8807	0.7117	1.0933
Bagging pre-test	0.8796	0.8666	0.7681	0.8487	0.8098	0.7261	0.8672
DFM rank 1	1.1418	1.1647	1.1363	1.1646	1.1642	1.1272	1.1414
DFM rank 2	1.0704	1.0857	1.0470	1.0718	1.0477	1.0426	1.0632
DFM rank 3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
DFM rank 4	0.9878	0.9889	0.9853	0.9862	0.9879	0.9856	0.9871

	Rank 4 Data Generating Process						
Forecast Model	**Design 1**	**Design 2**	**Design 3**	**Design 4**	**Design 5**	**Design 6**	**Design 7**
Unrestricted	0.973	0.978	0.964	0.961	0.957	0.968	0.967
Constant change	1.360	1.384	1.288	1.242	1.238	1.304	1.322
No Change	1.346	1.370	1.274	1.230	1.225	1.290	1.309
Pre-test	1.190	1.129	0.856	1.054	0.933	0.763	1.144
Bagging pre-test	0.919	0.907	0.811	0.880	0.843	0.774	0.905
DFM rank 1	1.220	1.226	1.177	1.157	1.158	1.174	1.200
DFM rank 2	1.121	1.119	1.090	1.097	1.081	1.086	1.111
DFM rank 3	1.046	1.043	1.039	1.047	1.041	1.037	1.044
DFM rank 4	1.000	1.000	1.000	1.000	1.000	1.000	1.000

SOURCE: Based on 5000 trials.

Table 4. Out-of-Sample Forecast Accuracy:
U.S. Inflation Forecasts: 1 Month Ahead
Evaluation Period: 1983.8-2003.7

	Models with Indicators of Economic Activity						
	RPMSE Relative to AR Benchmark at h=1						
Lags of				**DFM**			
Indicators	**UR**	**PT**	**BA**	**rank 1**	**rank 2**	**rank 3**	**rank 4**
1	**0.885**	**0.899**	**0.833**	0.985	0.991	**1.036**	**0.978**
2	1.168	0.925	0.862	**0.969**	**0.983**	1.049	1.021
3	1.668	1.017	14.302	0.984	1.000	1.055	1.049
4	-	-	-	0.990	1.013	1.094	1.089
5	-	-	-	0.993	1.019	1.123	1.142
6	-	-	-	0.998	1.012	1.168	1.185
SIC	0.885	0.899	0.836	0.984	1.014	1.135	1.066

SOURCE: The sample period of the raw data is 1971.4-2003.7. The RPMSE is based on the average of the squared recursive forecast errors. All pre-tests are based on White (1980) robust standard errors. The bagging results are based on the pairwise bootstrap.

Table 5. Out-of-Sample Forecast Accuracy:
U.S. Inflation Forecasts: 12 Months Ahead
Evaluation Period: 1983.8-2003.7

	Models with Indicators of Economic Activity						
	RPMSE Relative to Benchmark at h=12						
Lags of				**DFM**			
Indicators	**UR**	**PT**	**BA**	**rank 1**	**rank 2**	**rank 3**	**rank 4**
1	**0.695**	1.190	0.582	0.720	0.739	0.785	0.731
2	0.838	**1.046**	**0.564**	0.674	0.691	0.746	**0.704**
3	1.207	1.061	0.672	**0.668**	**0.685**	0.755	0.743
4	-	-	-	0.673	0.687	0.774	0.790
5	-	-	-	0.686	0.703	0.784	0.829
6	-	-	-	0.708	0.732	0.803	0.884
SIC	0.695	1.190	0.582	0.776	0.700	**0.738**	0.830

SOURCE: The sample period of the raw data is 1971.4-2003.7. The RPMSE is based on the average of the squared recursive forecasts errors. All pre-tests are based on West (1997) robust standard errors. The bagging results are based on blocks of length $m = 11$.

Figure 1: MSE of Alternative Predictors in Single-Regressor Model

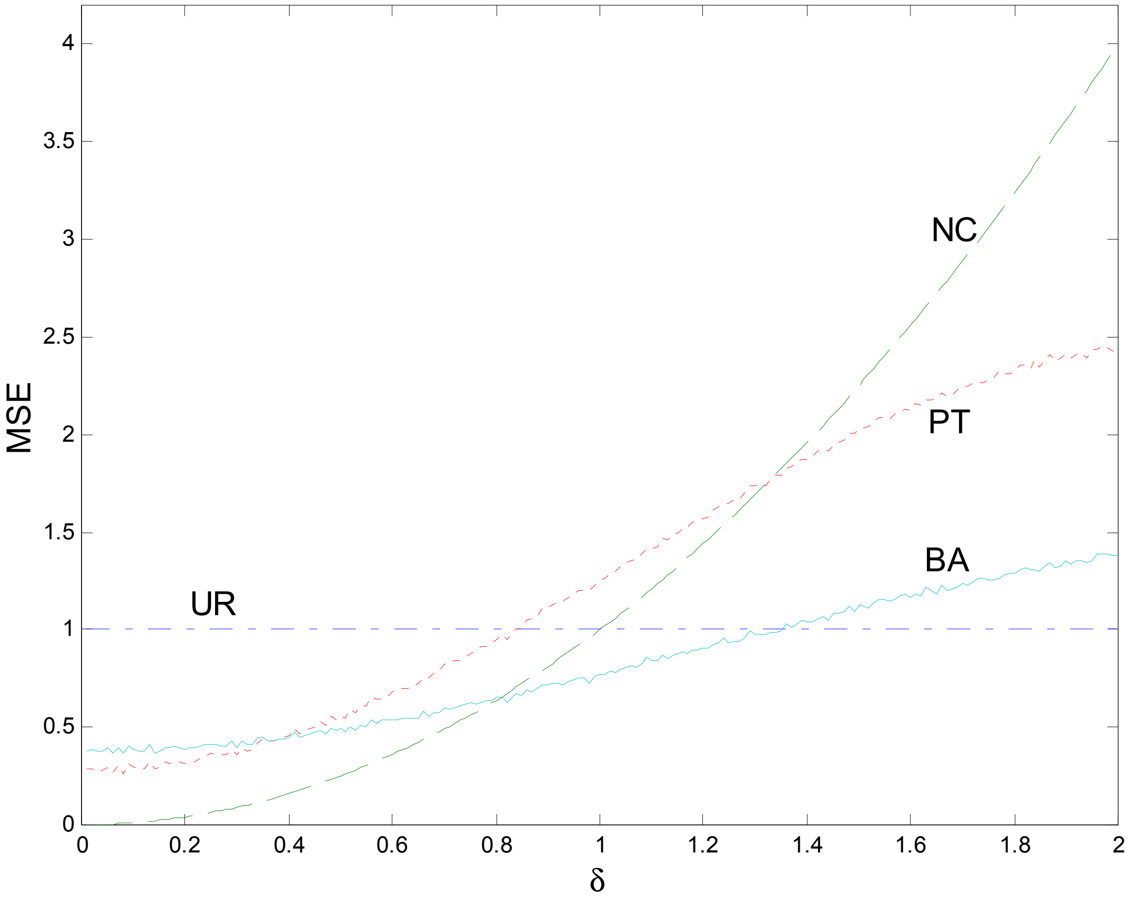


Figure 2: Bias^2, Variance and MSE of Pre-test and Bagging Predictors

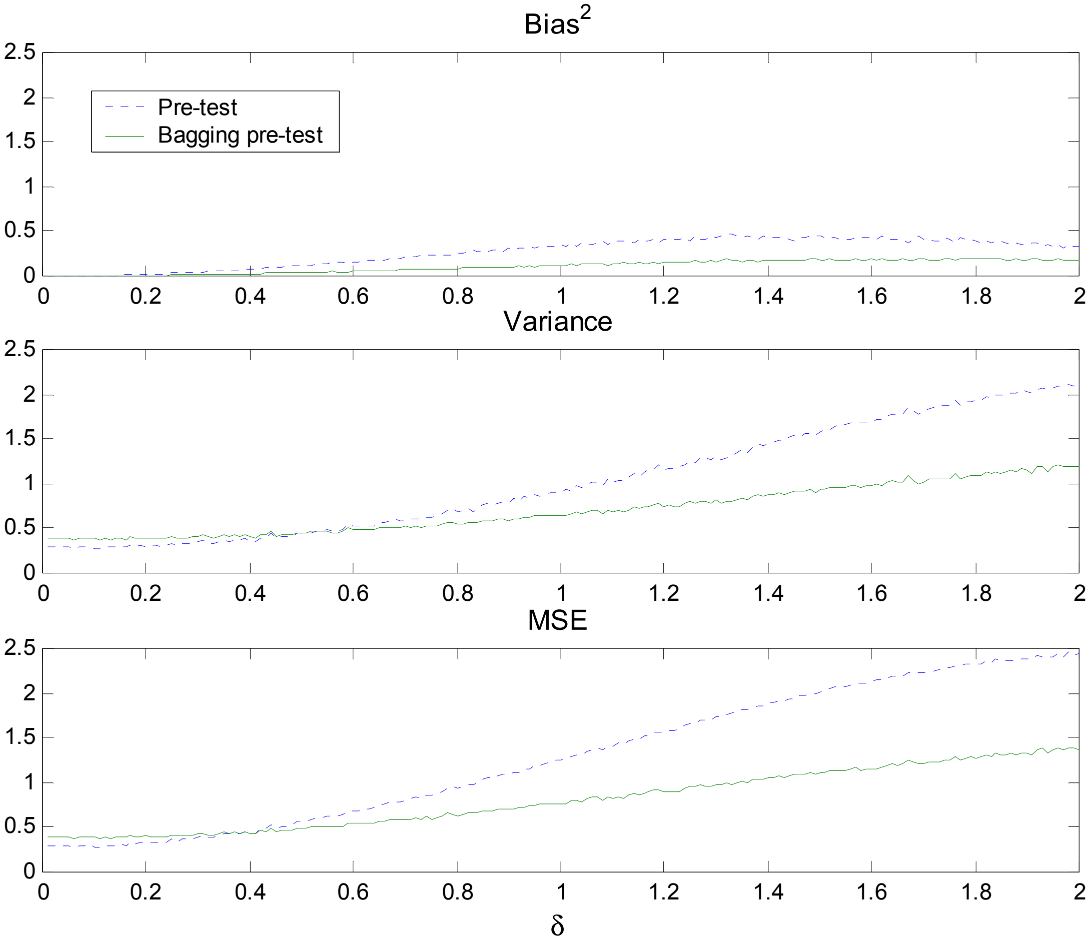


Figure 3: MSE of Alternative Predictors in Multiple-Regressor Model

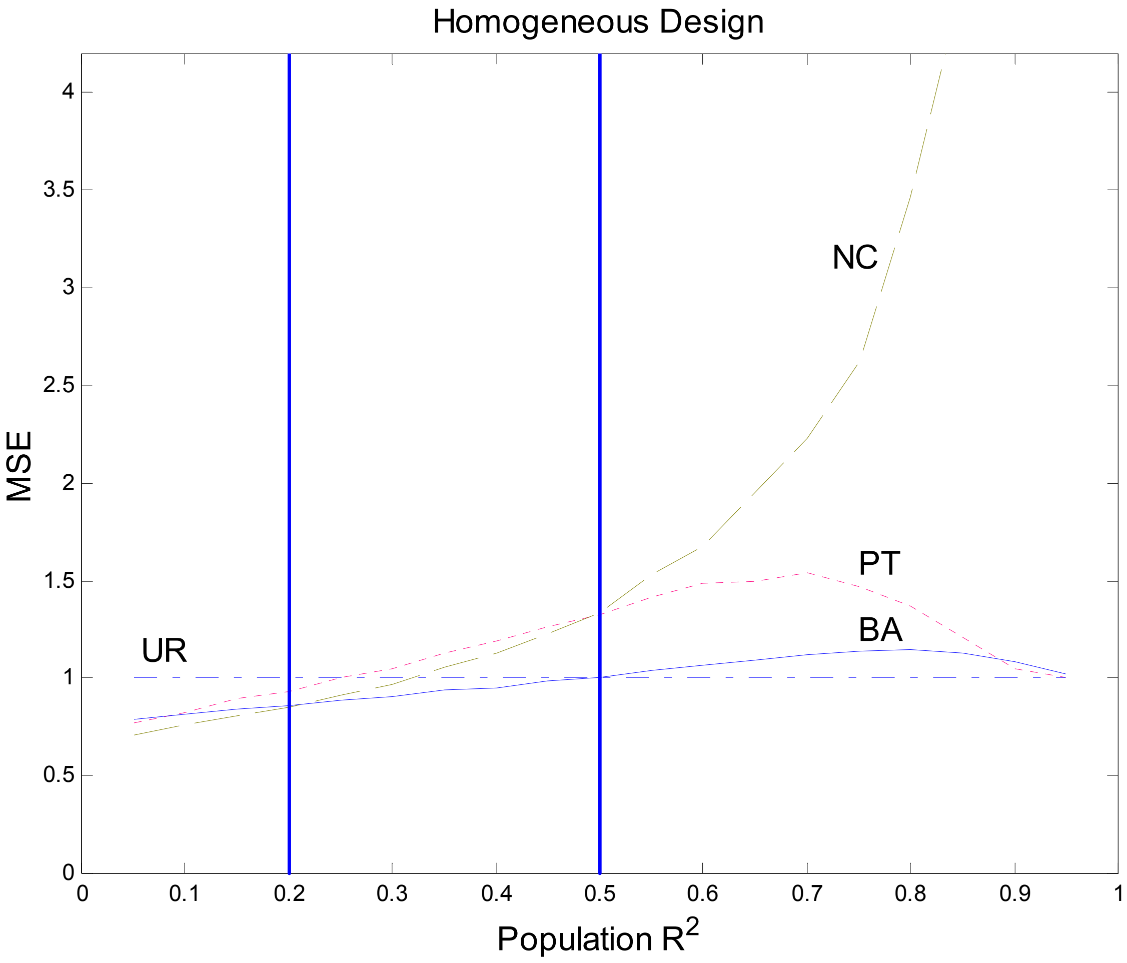


Figure 4: MSE of Alternative Predictors in Multiple-Regressor Model

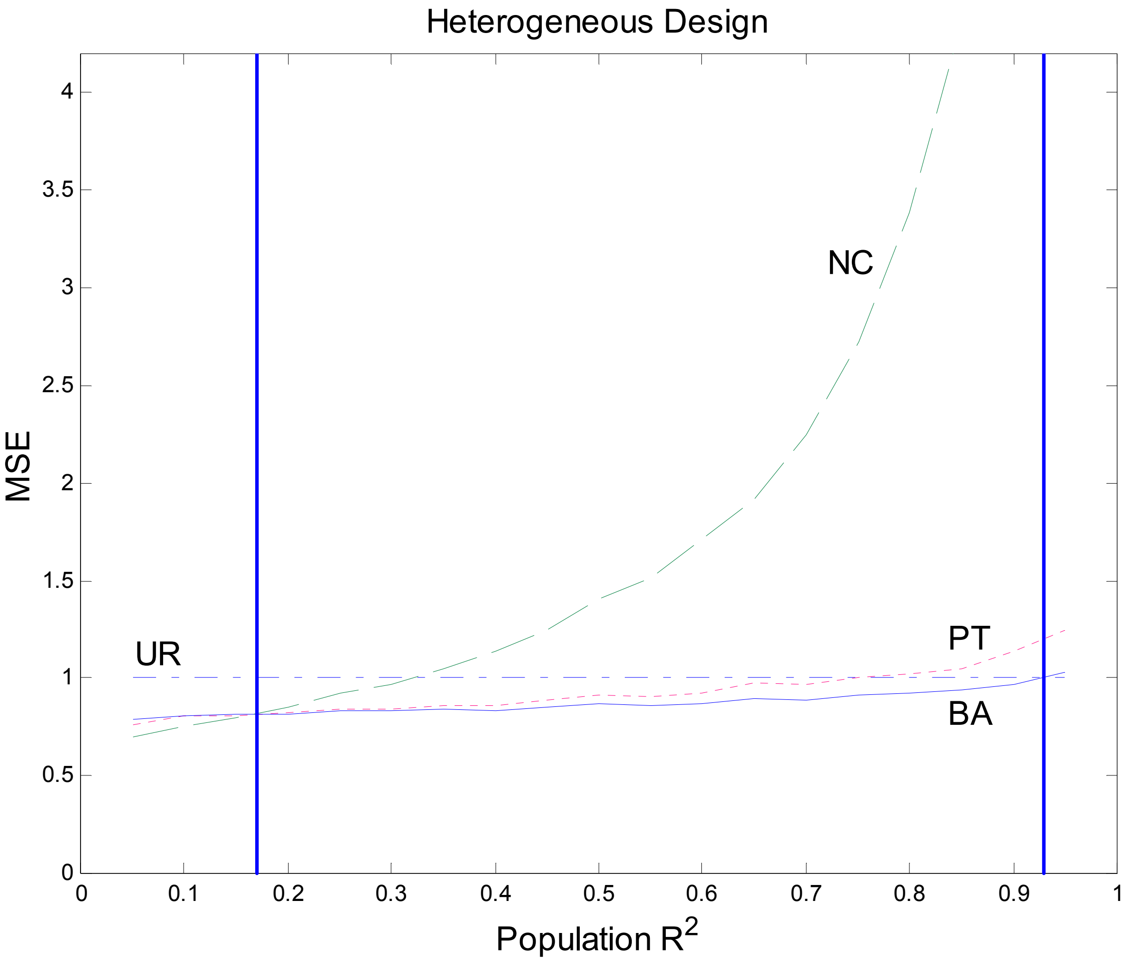


Figure 5: Gains from Bagging as a Function of M

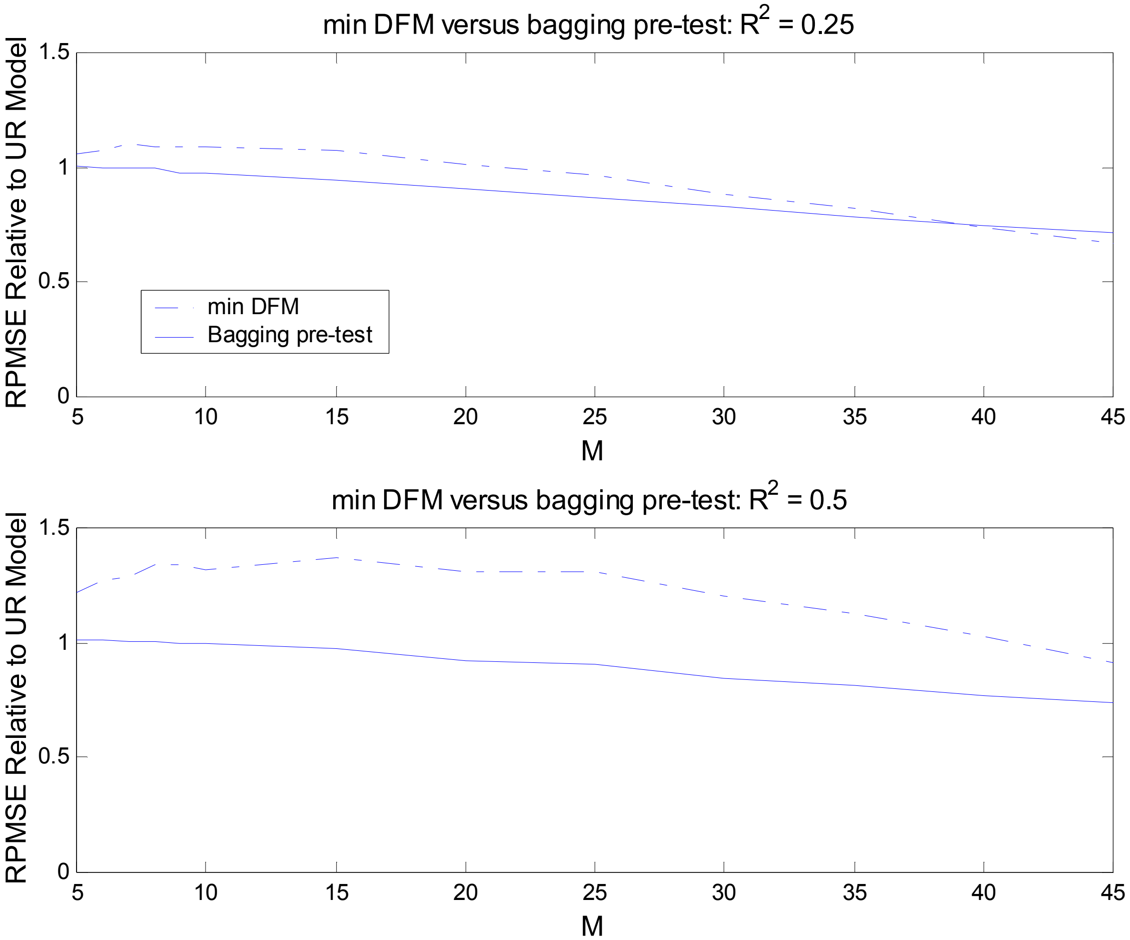
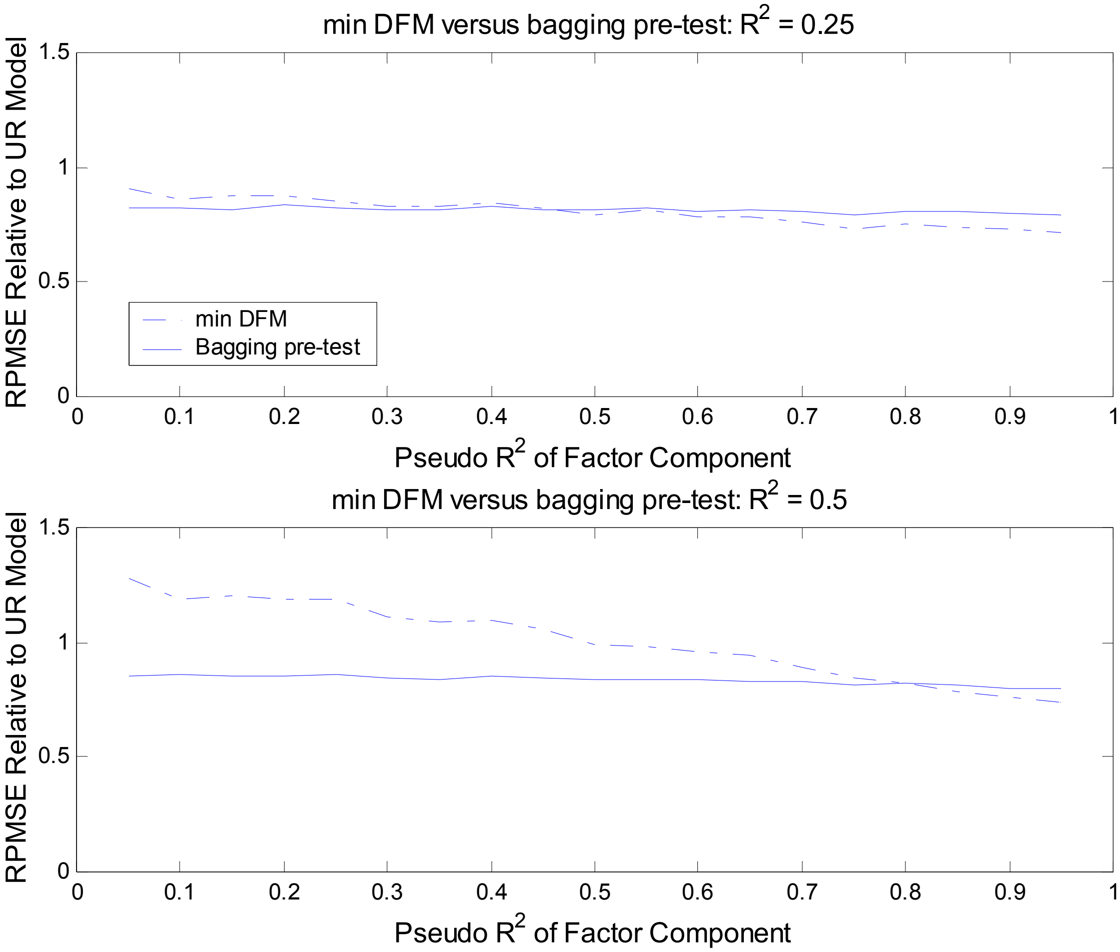


Figure 6: Gains from Bagging as a Function of the Strength of the Common Component


min DFM versus bagging pre-test: R^2 = 0.25
RPMSE Relative to UR Model
min DFM
Bagging pre-test
Pseudo R^2 of Factor Component
min DFM versus bagging pre-test: R^2 = 0.5
RPMSE Relative to UR Model
Pseudo R^2 of Factor Component