How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation*

Atsushi Inoue† Lutz Kilian‡
North Carolina State University University of Michigan

November 2, 2004

Abstract

This article explores the usefulness of bagging methods in forecasting economic time series. We focus on the widely studied question of whether the inclusion of indicators of real economic activity lowers the prediction mean-squared error of forecast models of U.S. consumer price inflation. We compare the accuracy of simulated out-of-sample forecasts of inflation based on the bagging method to that of alternative forecast methods, including factor model forecasts, shrinkage estimator forecasts, combination forecasts and Bayesian model averaging. We find that bagging in this application performs almost as well as or better than the best alternatives. Our analysis demonstrates that significant improvements in forecasting accuracy can be obtained over existing methods, and we illustrate how researchers can determine whether such gains are likely in a given application.

JEL: C22, C52, C53

KEYWORDS: Bootstrap aggregation; Bayesian model averaging; Forecast combination methods; Factor model forecasts; Shrinkage estimation; Forecast 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, Clara Vega, Mark Watson, Jonathan Wright and Arnold Zellner. We thank seminar participants at Caltech, the Center for Financial Studies, the ECB, Johns Hopkins, Michigan State, Maryland, Purdue, and Tokyo. We have also benefited from comments received at the 2003 Triangle Econometrics Conference, the 2004 Financial Econometrics Conference in Waterloo, the 2004 Forecasting Conference at Duke, the 2004 North American Summer Econometric Society Meeting at Brown, the 2004 NBER-NSF Time Series Conference at SMU, and the 2004 Midwest Econometrics Group Meeting at Northwestern.

†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 few (if any) of these predictors individually are likely to have high predictive power. This problem is particularly relevant in economic forecasting, because economic theory rarely puts tight restrictions on the set of potential predictors. In addition, often alternative proxies of the same variable are available to the economic forecaster. A case in point are forecasts of consumer price inflation, which may involve a large number of alternative measures of real economic activity such as the unemployment rate, industrial production growth, housing starts, capacity utilization rates in manufacturing, or the number of help wanted postings, to name a few.

It is well known that forecasts generated using only one of these proxies tend to be unreliable and unstable (see, e.g., Cecchetti, Chu and Steindel 2000, Stock and Watson 2003). On the other hand, including all proxies (even if feasible) is thought to lead to overfitting and poor out-of-sample forecast accuracy. This fact suggests that we use formal statistical methods for selecting the best subset of these predictors. Standard methods of comparing all possible combinations of predictors by means of an information criterion function, however, become computationally infeasible when the number of potential predictors is moderately large.[1]

One strategy in this situation is to combine forecasts from many models with alternative subsets of predictors. For example, one could use the mean, median or trimmed mean of these forecasts as the final forecast or one could use regression-based weights for forecast combination (see Bates and Granger 1969, Stock and Watson 2003). There is no reason, however, for simple averages to be optimal, and the latter approach of regression-based weights tends to perform poorly in practice, unless some form of shrinkage estimation is used (see, e.g., Stock and Watson 1999). More sophisticated methods of forecast model averaging weight individual forecasts by the posterior probabilities of each forecast model (see, e.g., Min and Zellner 1993, Avramov 2002, Cremers 2002, Wright 2003a and Koop and Potter 2003 for applications in econometrics). This Bayesian model averaging (BMA) approach has been used successfully in forecasting inflation by Wright (2003b). An alternative strategy involves shrinkage estimation of the unrestricted model that includes all potentially relevant predictors. Such methods are routinely used for example in the literature on Bayesian vector autoregressive models (see Litterman 1986). A third strategy is to reduce the dimensionality of the regressor set by extracting the principal components from the set of potential 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 not more than moderately large relative to the sample size, a fourth strategy is to rely on a testing procedure for deciding which predictors to include in the

[1] See Inoue and Kilian (2003) for a discussion of this and related approaches to 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).

forecast model and which to drop. For example, we may fit a model including all potentially relevant predictors, conduct a two-sided 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. The predictive accuracy of simple pre-test strategies, however, may be greatly enhanced by application of the bagging technique, leading to a fifth strategy that will be the focus of this paper.

Bagging is a statistical method designed to reduce the out-of-sample prediction mean-squared error of forecast models selected by unstable decision rules such as pre-tests. 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 pre-test rule to each of the resamples, and averaging the forecasts from the models selected by the pre-test on 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. This heuristic argument suggests that bagging will reduce the prediction mean squared error of the regression model after variable selection. 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 lower the asymptotic prediction mean-squared error (PMSE) of regressions with a single regressor 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 asymptotic forecast mean squared errors in many cases.

In this article, we explore the usefulness of bagging methods in forecasting economic time series. In section 2, we briefly review the theory behind bagging, and - drawing on the analysis of the single-regressor model in Bühlmann and Yu (2002) - provide some intuition for how and when bagging works. We then show how the bagging proposal may be adapted to applications involving dynamic multiple regression models with possibly serially correlated and heteroskedastic errors.

In section 3, we study the finite-sample properties of bagging using a number of stylized data generating processes designed to capture some of the typical features of economic forecasting problems. The simulation results suggest that bagging, while no panacea, is a promising alternative to existing forecasting methods when the number of predictors is moderately large relative to the sample size. For example, we show that the bagging forecast can be substantially more accurate than the factor model forecast when the number of predictors is smaller than the sample size, even when the factor model is the true model.

While these simulation results are encouraging, they are not dispositive. We therefore recommend that, in practice, researchers choose between the alternative forecasting methods based

on the ranking of their recursive PMSE in simulated out-of-sample forecasts. In section 4, 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, 2003), Marcellino et al. (2003), Bernanke and Boivin (2003), Forni et al. (2003) and Wright (2003b).

We show that bagging is a very accurate forecasting procedure in this empirical application. It outperforms the benchmark model involving only lags of inflation, the unrestricted model and the factor models with rank 1, 2, 3, or 4 and different lag structures. Given that bagging may be viewed as a shrinkage estimator, we also compare its performance to Bayesian shrinkage estimators. We find that bagging forecasts in some cases are almost as accurate as the forecast from the best Bayesian shrinkage estimator and in others more accurate. Bagging also is more accurate than forecast combination methods such as equal-weighted forecasts of models including one indicator of real economic activity at a time or the type of BMA studied by Wright (2003b). Finally, we show that bagging forecasts - depending on the horizon - are almost as accurate as or somewhat more accurate than BMA forecasts generated using the method of Raftery, Madigan and Hoeting (1997) that is based on randomly selected subsets of the predictors. We conclude in section 5.

2 How Does Bagging Work?

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.

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. Note that - unlike Bühlmann and Yu (2002) - we re-estimate the model after variable selection. For $x_t \in \Re^M$, we define the predictor from the unrestricted model (UR), the predictor from the fully restricted model (FR), and the pre-test (PT) predictor conditional on x_{T-h+1} by

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

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 UR model forecast is based on the fitted values of a regression including all M potential predictors. The FR model forecast emerges when all predictors are dropped, as in the well-known no-change forecast model of asset returns. The latter forecast sometimes is also referred to as a random walk forecast in the literature, as y_{t+h} in economic applications often refers to a percentage growth rate.

The pre-test strategy that we analyze is particularly simple. 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 $h = 1$ or 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).

2.1 Algorithm for Bagging Dynamic Regression Models

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* $(T-h) \times (M+1)$:

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

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 *(see, e.g., Hall and Horowitz 1996, Gonçalves and White 2004).*

(*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* $\widehat{H}^{*-1}\widehat{S}^*\widehat{H}^{*-1}$ *where*

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

$\varepsilon^*_{t+h} = y^*_{t+h} - \widehat{\beta}^{*\prime}x^*_t$*, 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:*

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

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 $m = h$ (see, e.g., Gonçalves and Kilian 2004). Otherwise $m > h$. 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 pre-testing strategy discussed here, and there is no reason to believe that our t-test strategy is optimal. For example, we could use F-tests for variables that are highly correlated or that measure related concepts (say, asset prices or monetary aggregates). Nevertheless, the simple t-test strategy studied in this paper appears to work well in many cases.

Bagging could also be applied to other methods of forecasting. For bagging to offer the potential for asymptotic improvements, however, it is necessary that the predictor to be bagged involve some hard threshold (such as the decision of whether to include or exclude a given predictor). In the absence of such a threshold, the bagged predictor would be expected to perform similarly to the original predictor. It is also important that the decision rule select

different models with positive probability. For example, the Schwarz Information Criterion is not a promising candidate for bagging (even when M is small) because it will select one forecast model with probability one asymptotically, so there is no scope for asymptotic improvements in the prediction mean-squared error.

The performance of bagging will in general depend on the significance level chosen for pre-testing. 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 taken up in section 4.

2.2 Asymptotic Properties of Bagging: A Simple Example

The fundamental problem in choosing a forecast model is that of resolving the bias-variance trade-off that arises when the regressors have only weak predictive power. 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) who consider bagging a linear model with one local-to-zero regressor when the data are i.i.d.. Here we consider the special case of an intercept for expository purposes. Suppose that $\beta = \delta T^{-1/2}$, $x_t = 1\ \forall t$, ε_t is distributed $iid(0,1)$, and $h = 1$.

The forecasts from the unrestricted model, the fully restricted model and the pre-test model can be written as

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

It may be shown that:

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

where $z \sim N(0,1)$ and $\xi \sim N(\delta, 1)$.Ultimately, we are interested in the asymptotic PMSE of each method, which here takes the form $E[(y_{T+1} - \hat{y}(x_T))^2] = 1 +$ forecast bias^2 + forecast variance. The first term of the PMSE expression is beyond the forecaster's control, but by choosing

between different forecast methods, the forecaster may be able to reduce the sum of the second term and the third term, which we will refer to as the mean-squared error (MSE) of the predictor. In our example, the asymptotic MSE expressions are:

$$
\begin{aligned}
T\ E[(E(y_{T+1}) - \hat{y}^{UR})^2] &= 1 + o(1), && (2)\\
T\ E[(E(y_{T+1}) - \hat{y}^{FR})^2] &= \delta^2, && (3)\\
T\ E[(E(y_{T+1}) - \hat{y}^{PT})^2] &= E[(\xi - \delta)I(|\xi| > 1.96) + \delta I(|\xi| \leq 1.96)]^2 + o(1), && (4)\\
T\ E[(E(y_{T+1}) - \hat{y}^{BA})^2] &= E[(\delta - \xi + \xi\Phi(1.96 - \xi) - \sqrt{\Sigma}\phi(1.96 - \xi) \\
&\quad -\xi\Phi(-1.96 - \xi) + \sqrt{\Sigma}\phi(-1.96 - \xi)]^2 + o(1). && (5)
\end{aligned}
$$

The aim of the forecaster is to choose the predictor that minimizes the MSE of the forecast, which in general is a function of the drift term δ.

Figure 1 shows that for $\delta > 1$ the UR predictor has lower MSE than the FR predictor, for $\delta = 1$ both models are tied and for $\delta < 1$ the FR model is asymptotically more accurate. Moreover, although the PT predictor protects the user from choosing the UR forecast when δ is close to zero and the FR forecast when δ is large, the PT forecast for any given choice of δ is always dominated by either the UR or the FR model.[3]

How does the MSE of the BA predictor compare to that of the PT predictor? Note that the asymptotic MSE expression for the BA predictor does not depend on the indicator function, reflecting the smoothing implied by bootstrap aggregation. Although this smoothing should typically help to reduce the forecast variance relative to the PT predictor, it is not obvious a priori whether bagging the pre-test predictor will also improve the MSE. Figure 2 investigates this question. The upper panel shows the squared bias of the two predictors. Although bagging does reduce the bias somewhat for most values of δ, the gains are small. The second panel, in contrast, shows dramatic reductions in variance relative to the pre-test estimator for most δ, which, as shown in the third panel, result in substantial improvements in the overall accuracy measured by the MSE. 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. Nevertheless, as Figure 1 shows, the BA predictor in turn is dominated by the FR predictor for low values of δ and by the UR predictor for large values of δ. Only for a small range of δ values near one is bagging the asymptotically best strategy.

This simple example based on Bühlmann and Yu (2002) conveys two valuable insights that hold more generally: First, bagging under certain conditions *can* yield asymptotic improvements in the PMSE. This fact suggests that it deserves further study. Second, the extent of these asymptotic improvements depends very much on unobservable features of the data. Under some conditions bagging may actually result in a lower asymptotic PMSE. This seems especially likely when the signal-to-noise ratio in the data is very weak, as in forecasting asset returns for example. This is not a limitation of the bagging method alone, of course, but simply a reflection of the bias-variance trade-off in forecasting. The same type of problem would arise with any other forecasting method in the literature.

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

These insights generalize directly to the multiple regression case, as the following examples show. Suppose that we have multiple predictors with different drift values, 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 can 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. Once again bagging yield a reduction in the asymptotic MSE. Specifically, bagging yields an MSE of only 2.53 compared with 3, 3.97 and 5 for the other three forecast methods. It is interesting to note that the bagging forecast MSE may be lower than the MSE of either the unrestricted or the fully restricted 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, i.e., in parameter regions where bagging would not work well in the single-regressor context. On the other hand, it is possible to construct examples, in which bagging is asymptotically dominated by other predictors. For example, when $\delta_1 = 0$, $\delta_2 = 0.1$, and $\delta_3 = 0.2$, both the PT predictor and the FR predictor will have lower asymptotic MSE than the BA forecast.

3 Finite-Sample Properties of Bagging in Multiple Regression

There are two main limitations of the asymptotic analysis in the preceding section. First, in more general settings, it is difficult to work out analytical solutions for the asymptotic PMSE of the bagging method in multiple regression, and indeed not particularly informative since we do not know the properties of the data generating process and cannot consistently estimate the relevant parameter δ (or is multiple regression analogue). Second, nothing ensures that the finite-sample properties of bagging are similar to its asymptotic properties. We therefore turn to a simulation study to investigate the potential of bagging to improve the finite-sample PMSE relative to the unrestricted forecast model, various restricted forecast models and the pre-test forecast model.

The design of the simulation study will aim to reproduce some of the typical features of economic forecasting problems, notably the relatively weak predictive power of most potential predictors and the potentially strong degree of comovement among predictors. The latter is modeled by imposing a common factor structure involving a small number of common components, ranging from 1 through 4. A question of practical interest is whether the bagging strategy is competitive with alternative approaches to forecasting from large data sets with those specific features. Hence, we will broaden the scope of alternative forecast methods under consideration to include factor model forecasts.

Our aim here is not to argue that one approach in general will be superior to alternative approaches. Clearly, the design of our simulation study is too limited to make such a case, and we would not expect such clear-cut results in any case given the earlier theoretical analysis. 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 in finite samples, and we want to show that the bagging approach is not dominated by existing forecasting techniques in situations when M is large enough to prevent

the use of information criteria, but still much smaller than T.

In the simulation study we deliberately abstract from the existence of lagged dependent variables. In other words, we think of a benchmark model that is just white noise. We also model the potential predictors as uncorrelated over time. These simplifications help reduce the computational cost of the simulation study and should not make a difference asymptotically. Judging by the close coherence of the simulation results and the empirical results in section 4, this simplification seems acceptable.

3.1 Simulation Design: Multiple Regression Models

Rather than being universal, the PMSE 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.

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

$$y_{t+1} \;=\; \beta' x_t + \varepsilon_{t+1} \tag{6}$$

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. In the multiple regression case, δ will be an M-dimensional vector, and the population R^2 of the forecast model for finite T will be directly linked to δ by $R^2 = \delta'\delta/(T + \delta'\delta)$. Since δ cannot be consistently estimated, this suggests using a grid of values $R^2 \epsilon \{0.25, 0.5, 0.75\}$. Below we will show simulation results for $R^2 = 0.25$ and $R^2 = 0.5$ only, since the results for $R^2 = 0.75$ are similar to those for $R^2 = 0.5$.

3.1.1 Design of Slope Parameters in Forecast Model

The first design issue is the choice of the slope parameter vector β. A plausible scenario in economic forecasting is that most (if not all) regression slopes are close to zero, but some predictors are relatively more important than others. We attempt to capture this feature 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. 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. Thus, only the relative magnitude of the elements of β matters, not their absolute magnitude.

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.

$$x_t \sim NID(0_{30\times 1}, I_{30})$$

While instructive, this first case, in which all predictors are orthonormal in population, 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 $r \in \{1, 2, 3, 4\}$.

Case 2.

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

where $F_t \sim N(0, I_r)$, $\eta_t \sim N(0_{30\times 1}, I_{30})$, *and* Λ is an $M \times r$ 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. Note that all measures of forecast accuracy are true out-of-sample measures. 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 fully restricted model, a model including only the intercept, 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 $r \in \{1, 2, 3, 4\}$. 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 $\phi(L) = 0$ 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 3, 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 intercept only and fully restricted 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.

3.2.2 Case 2: Common Factors among Predictors

As noted earlier, case 1 is a useful benchmark, but 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 the case when the true model is a factor model of rank 1. In that case, for $R^2 = 0.25$, the unrestricted, fully restricted and intercept only 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 the seven 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 Discussion

The evidence in Tables 2 and 3 suggests that bagging forecasts often are more accurate than forecasts from factor models, even when the predictor data were generated from a factor model. This may seem odd at first in that usually imposing correct structure should improve forecast accuracy. It is important to understand why that argument fails here.

First, note that factor model forecasts are based on regressions of the variable to be predicted on estimated factors, as opposed to the true factors. As shown by Bai and Ng (2003), ignoring the estimation error will in general distort forecasts. To overcome this generated regressor problem the number of predictors must be large relative to the number of time series observations. 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). In the applications we considered M was large, but distinctly smaller than T. Hence, it is not surprising that the factor model forecast did not perform well.

Our evidence shows that bagging provides a useful alternative to factor model forecasts in precisely those situations, in which the standard justification for forecasting from estimated dynamic factor models is questionable. Conversely, we note that bagging methods tend to become infeasible when the number of predictors is large relative to the sample size. This happens because near-singularity problems arise in computing the least-squares estimator of the unrestricted model, on which the bootstrap approximation is based.[4] Factor models do not suffer from this limitation. Thus, bagging forecasts and factor model forecasts have been designed with different situations in mind and are best viewed as complements.

Second, for given M and T, the ranking of bagging forecasts and forecasts from dynamic factor models will depend crucially on the strength of the common factor component relative to

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

the idiosyncratic noise component in the set of predictors. When the factor structure is weak, as it appears to be in the empirical example upon which we based our choice of R^2_{pseudo} for the simulation study, bagging may be much more accurate in small samples than imposing the true 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.

These two points may be illustrated further by two examples from the simulation study. Figures 3 and 4 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 3 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 and coheres well with what we would expect based on theory. The ranking itself also depends on the population R^2 of the unrestricted forecast model. 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.

Figure 4 illustrates the importance of the strength of the factor component in the predictor data. All results are based on $M = 30$ and $T = 100$. Figure 4 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.

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

While the simulation results in section 3 are encouraging, we have no way of knowing a priori whether the data generating process in a given empirical application will favor bagging or some other forecasting method because the relative accuracy of the forecasting methods 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 alternative forecasting methods 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 this section for a typical forecast problem in economics.

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 (2003b), among others. The choice of the benchmark model is conventional (see, e.g., Stock and Watson 2003, Forni et

al. 2003) as is the focus on the PMSE. 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 upon 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.

4.1 Comparison with Dynamic Factor Model Forecasts

The alternative forecasting strategies under consideration in the first round of comparisons include the benchmark model involving only an intercept and 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 block bootstrap with $m = 12$. Finally, we also fit factor models with rank $r \in \{1, 2, 3, 4\}$ 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. When such problems arise, the corresponding entry in the table has been left blank. We also show results based on the SIC with an upper bound of 2 lags. For larger upper bounds, again 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:

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

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 average of the squared forecast errors obtained by recursively re-estimating the model at each point in time t 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 4a 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 4a 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 PMSE 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 4b 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.

4.2 Comparison with Bayesian Shrinkage Estimators

The bagging method also has similarities with shrinkage estimators such as the Stein-type estimator or the Bayesian shrinkage estimator used by Litterman (1986) in a different context. Thus, it is natural to compare the accuracy of bagging to that of the shrinkage estimator. A Bayesian approach is convenient in this context because it allows us to treat the parameters of the benchmark model differently from the parameters of the real economic indicators. Note that the use of prior distributions in this context does not reflect subjectively held beliefs, but simply is a device for controlling the degree of shrinkage. To facilitate the exposition and to preserve consistency with related studies, in the remainder of the paper we will include at most

one lag of each indicator of real economic activity. The Bayesian shrinkage estimator is applied to the model:

$$\pi^h_{t+h|t} = \widehat{\alpha} + \sum_{k=1}^{p} \widehat{\phi}_k \pi_{t-k} + \sum_{j=1}^{M} \widehat{\beta}_j x_{j,t}$$

We postulate a diffuse Gaussian prior for $(\alpha, \phi_1, ..., \phi_p)$. The prior mean is based on the fitted values of a regression of inflation on lagged inflation and the intercept over the pre-sample period, as proposed by Wright (2003b). In our case, the pre-sample period includes 1947.1-1971.3. The prior variance is infinity. We use a different prior mean for each combination of h and p used in the benchmark model. For the remaining parameters we postulate a Gaussian prior with mean zero and standard deviation $\lambda \epsilon \{0.01, 0.05, 0.1,\ 0.2, 0.3, 0.4, 0.5, 1, 2, 5, 100\}$ for the standardized data. For $\lambda = \infty$, the shrinkage estimator reduces to the least-squares estimator of the unrestricted model. All prior covariances are set to zero. For further details on the implementation of this estimator see Lütkepohl (1993, ch. 5.4).

Table 5a shows selected results of the grid search over λ. We find that for $h = 1$ a moderate degree of shrinkage helps reduce the PMSE. The optimal degree of shrinkage is near $\lambda = 0.5$; as λ declines further, the PMSE ratio quickly starts deteriorating. The best shrinkage estimator is slightly more accurate than the bagging estimator at the one-month horizon with a ratio of 81 percent compared with 83 percent for bagging. In contrast, at the one-year horizon, the unrestricted model with a ratio of 70 percent is more accurate than any shrinkage estimator, and bagging is substantially more accurate than either with a ratio of 58 percent. We conclude that bagging in this application performs almost as well or better than Bayesian shrinkage estimators, depending on the horizon.

It is important to note that the performance of the shrinkage estimator in Table 5a reflects a grid search over the parameter space of the prior, whereas for the bagging estimator we relied on an ad hoc choice of the nominal size of $\alpha = 0.05$. There is no reason for that choice to be optimal in this application, and one might expect that in particular for $h = 1$ the performance of bagging could be improved by a judicious choice of α. Table 5b shows the results for a grid search over $\alpha \epsilon \{0.01, 0.05, 0.10, 0.20, 0.50\}$. [RESULTS TO BE ADDED]

4.3 Comparison with Bayesian Model Averaging: One Extra Predictor at a Time

Recently, there has been mounting evidence that forecast combination methods are a promising approach to improving forecast accuracy. For example, Stock and Watson (2003) have shown that simple methods of forecast combination such as using the median forecast from a large set of models may effectively reduce the instability of inflation forecasts and lower their prediction mean-squared errors. In its simplest form, forecast combination methods assign equal weight to all possible combinations of the benchmark model and one extra predictor at a time. More recently, Wright (2003b) has shown that the accuracy of forecast combination methods may be improved upon further by weighting the individual forecast models based on the posterior probabilities associated with each forecast model. in this subsection, we will expand the list of competitors of bagging to include Wright's BMA method. A key difference between our papers

is that Wright imposes one lag of inflation only in the benchmark model, whereas we allow for potentially more than one lag of inflation. Otherwise our approaches are identical.

As before, for the benchmark model we follow Wright (2003b) in postulating a diffuse Gaussian prior with the prior mean based on the fitted values of a regression of inflation on lagged inflation and the intercept over the pre-sample period. For the remaining parameters we postulate a Gaussian prior with mean zero and a prior standard deviation of $\phi \epsilon \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2, 5, 100\}$ for the standardized data. Again the prior treats the predictors as independent. The prior probability for each forecast model is $1/M$, as in the equal-weighted forecast combination. For $\phi = 0$, the BMA method of forecast combination reduces to the equal-weighted method. Table 5c presents selected results of the grid search over ϕ.

We find that, as in Wright (2003b), the BMA method is clearly superior to the equal-weighted forecast combination method. Table 5c also shows the PMSE ratio of the median forecast. This alternative combination forecast was inferior to both the BMA forecast and the equal-weighted forecast. The best results for the BMA method at the one-month horizon are achieved with $\phi = 0.1$. At the one-year horizon an even tighter prior of $\phi = 0.05$ works best. These results are of course problem-specific. For example, for Wright's (2003b) quarterly data set much larger prior standard deviations appear to work best.

With a ratio of 90 percent the BMA method in our application is more accurate than the factor model forecast at the one-month horizon, but somewhat less accurate than the bagging forecast. At the one-year horizon, on the other hand, the best BMA forecast with a ratio of 84 percent is inferior to the factor model forecast and much less accurate than the bagging forecast. We conclude that in this application bagging clearly outperforms the BMA method.

4.4 Comparison with Bayesian Model Averaging: Randomly Chosen Subsets of Extra Predictors

Papers on forecast combination methods for inflation typically restrict the forecast models under consideration to include only one indicator of real economic activity at a time. There is no reason for this approach to be optimal, whether we use equal weights or posterior probability weights. In fact, a complete Bayesian solution to this problem that provides optimal predictive ability would involve averaging over all possible forecast model combinations (see Madigan and Raftery 1994). The problem is that such a systematic comparison of all possible subsets of such indicators would be computationally prohibitive in realistic situations. In our example, there are $2^{26} = 67,108,864$ possible combinations of predictors to be considered. In response to this problem, Raftery, Madigan and Hoeting (1997) proposed an alternative method of BMA for linear regression models based on a randomly selected subsets of predictors that approximates the Bayesian solution to searching over all models.[5] The random selection is based on a Markov Chain Monte Carlo (MCMC) algorithm that moves through the forecast model space. Unlike Wright's method, this algorithm involves simulation of the posterior distribution and is quite computationally intensive. Our results are based on 5000 draws from the posterior distribution at each point in time.

[5]See Sala-i-Martin, Doppelhofer and Miller (2004) for a similar approach to BMA in a different context. Also see George and McCulloch (1993) for an alternative stochastic search variable selection algorithm.

MATLAB code for the Raftery et al. algorithm is publicly available at http://www.spatial-econometrics.com. We modified the Raftery et al. approach to ensure that the benchmark model including only lags of inflation and the intercept is retained in each random selection. For the models of the benchmark model we use a diffuse Gaussian prior identical to the priors used for the Wright (2003b) method. For the remaining parameters of the forecast prior the algorithm involves a Gaussian prior with mean zero and hyperparameters $\nu = 2.58$, $\lambda = 0.28$, and $\phi \epsilon \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2, 5, 100\}$, where ϕ measures the prior standard deviation of the standardized predictor data (see Raftery et al. for further details). We report a subset of the empirical results in Table 5d. We also experimented with $\phi = 2.85$, the value recommended by Raftery et al. for a generic linear model, but the results were clearly worse than for our preferred value of ϕ below.

We find that a value of about $\phi = 0.01$ works best for $h = 1$ and $\phi = 0$ for $h = 12$. This version of BMA produces clearly more accurate results than the restricted version involving only one extra predictor at a time. Compared to Table 5c, at the one-month horizon the PMSE ratio falls from 90 percent to 80 percent and at the one-year horizon from 84 percent to 62 percent. Thus, for $h = 1$, this BMA method is somewhat more accurate than bagging using the default value of $\alpha = 0.05$; for $h = 12$, however, bagging promises somewhat higher accuracy with a ratio of 58 percent.

5 Conclusion

Recently, there has been increased interest in forecast methods that allow the user to extract the relevant information from a large set of potentially relevant predictors. One such method is bootstrap aggregation of forecasts (or *bagging* for short). Bagging is intended to reduce the out-of-sample prediction mean-squared error of forecast models selected by unstable decision rules such as pre-tests. This article explored the usefulness of bagging methods in forecasting economic time series. We first described how to implement the bagging idea in the context of dynamic multiple regression models with possibly serially correlated and heteroskedastic errors. Using asymptotic theory and simulation evidence we then showed that bagging, while no panacea, is a promising alternative to existing forecasting methods in many cases.

In particular, we showed that bagging forecasts may outperform factor model forecasts, even when the predictor data are generated by a factor model. This will tend to happen when the number of potential predictors is smaller than the sample size. The explanation is that factor model forecasts are based on regressions of the variable to be predicted on estimated factors, as opposed to the true factors. As shown by Bai and Ng (2003), the estimation error will in general distort forecasts. To overcome this generated regressor problem the number of predictors must be large relative to the number of time series observations. In the applications we considered the number of predictors was large, but distinctly smaller than the sample size. Hence, it is not surprising that the factor model forecast did not perform well. Bagging provides a useful alternative to factor model forecasts in precisely those situations, in which the standard justification for forecasting from estimated dynamic factor models is questionable. Conversely, we note that bagging methods tend to become infeasible when the number of predictors is large relative to the sample size. This happens because near-singularity problems arise in computing

the least-squares estimator of the unrestricted model, on which the bootstrap approximation is based. Factor models do not suffer from this limitation. Thus, bagging forecasts and factor model forecasts have been designed with different situations in mind and are best viewed as complements.

We also showed that, for a given number of predictors, the relative ranking of bagging forecasts and forecasts from dynamic factor models will in general depend 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.

Whether bagging is likely to improve out-of-sample forecast accuracy in a given application may be assessed based on a simulated out-of-sample forecast exercise. For illustrative purposes, we considered the widely studied question of whether the inclusion of indicators of real economic activity lowers the prediction mean-squared error of forecast models of U.S. CPI inflation. Over a twenty-year period, we compared the accuracy of simulated out-of-sample forecasts based on the bagging method to that of alternative forecast methods for U.S. inflation, including forecasts from a benchmark model that includes only lags of inflation, forecasts from the unrestricted model that includes all potentially relevant predictors, forecasts from models with a subset of these predictors selected by pre-tests, forecasts from estimated factor models, forecasts from models estimated by shrinkage estimators, standard combination forecasts and finally forecasts obtained by state-of-the-art methods of Bayesian model averaging.

We found that bagging greatly reduces the prediction mean squared error of forecasts of U.S. CPI inflation at horizons of one month and one year relative to the unrestricted, fully restricted and pre-test model forecasts. Consistent with our simulation evidence, bagging forecasts in this application also were more accurate than forecasts from estimated factor models. In addition, in this application, bagging performed better than the method of Bayesian model averaging recently proposed by Wright (2003b), and - depending on the horizon - almost as well as or somewhat better than forecasts based on Bayesian shrinkage estimators or on the method of Bayesian model averaging proposed by Raftery et al. (1997). Our analysis demonstrated that significant improvements in forecasting accuracy can be obtained over existing methods, and we illustrated how researchers can determine whether such gains are likely in a given application.

We note, however, that the analysis of bagging presented in this paper assumes a covariance stationary environment and abstracts 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. An interesting avenue for future research would be the development of bagging methods that allow for smooth structural change.

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. Avramov, D. (2002), "Stock Return Predictability and Model Uncertainty," *Journal of Financial Economics*, 64, 423-458.
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. Bates, J.M., and C.W.J. Granger (1969), "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451-468.
4. Bernanke, B.S., and J. Boivin (2003), "Monetary Policy in a Data-Rich Environment," *Journal of Monetary Economics*, 50, 525-546.
5. Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 36, 105-139.
6. Bühlmann, P. and B. Yu (2002), "Analyzing Bagging," *Annals of Statistics*, 30, 927-961.
7. Cecchetti, S., R. Chu, and C. Steindel (2000), "The Unreliability of Inflation Indicators," *Federal Reserve Bank of New York Current Issues in Economics and Finance*, 6, 1-6.
8. Cremers, K.J.M. (2002), "Stock Return Predictability: A Bayesian Model Selection Perspective," *Review of Financial Studies*, 15, 1223-1249.
9. 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.
10. 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.
11. 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*, 50, 1243-1255.
12. George, E.I., and R.E.McCulloch (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881-890.
13. Gonçalves, S. and L. Kilian (2004), "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form," *Journal of Econometrics*, 123, 89-120.
14. Gonçalves, S. and H. White (2004), "Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models," *Journal of Econometrics*, 119, 199-220.
15. Hall, P. and J.L. Horowitz (1996), "Bootstrap critical values for tests based on generalized method of moments estimators," *Econometrica,* 64, 891–916.
16. Inoue, A., and L. Kilian (2003), "On the Selection of Forecast Models," Working Paper No. 214, European Central Bank.
17. Inoue, A. and M. Shintani (2003), "Bootstrapping GMM Estimators for Time Series," forthcoming: *Journal of Econometrics*.
18. Koop, G., and S. Potter (2003), "Forecasting in Large Macroeconomic Panels Using Bayesian Model Averaging," *Federal Reserve Bank of New York Staff Report*, 163.
19. Litterman, R.B. (1986), "Forecasting with Bayesian Vector Autoregressions - Five Years of Experience," *Journal of Business and Economic Statistics*, 4, 25-38.

20. Lütkepohl, H. (1993), *Introduction to Multiple Time Series Analysis*, Springer-Verlag: Berlin.

21. Madigan, D., and A.E. Raftery (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535-1546.

22. Marcellino, M., J.H. Stock and M.W. Watson (2003), "Macroeconomic Forecasting in the Euro Area: Country-Specific versus Area-Wide Information," *European Economic Review*, 47, 1-18.

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

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

25. Raftery, A.E., D. Madigan, and J.A. Hoeting (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179-191.

26. Sala-i-Martin, G. Doppelhoffer, and R.I. Miller (2004), "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94, 813-835.

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

28. 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.

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

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

31. 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.

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

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

34. Wright, J.H. (2003a), "Bayesian Model Averaging and Exchange Rate Forecasts," *International Finance Discussion Papers*, No. 779, Board of Governors of the Federal Reserve System.

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

Table 1. Asymptotic Forecast MSE of Regression with Three Orthonormal Predictors:

Forecast Methods	$MSE = \sum_{i=1}^{M} MSE(i)$
UR	3.000
FR	5.000
PT	3.964
BA	2.530

Source: Based on $\delta_1 = 0$, $\delta_2 = 1$, $\delta_3 = 2$ 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
Intercept only	0.915	0.917	0.900	0.897	0.890	0.910	0.908
Fully restricted	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	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
Intercept only	1.051	1.050	1.034	1.048	1.041	1.021	1.051
Fully restricted	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	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
Intercept only	1.054	1.046	1.062	1.059	1.044	1.068	1.062
Fully restricted	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	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
Intercept only	1.090	1.113	1.075	1.089	1.100	1.083	1.082
Fully restricted	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	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
Intercept only	1.125	1.135	1.096	1.077	1.075	1.101	1.110
Fully restricted	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	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
Intercept only	1.389	1.392	1.350	1.345	1.338	1.349	1.373
Fully restricted	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	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
Intercept only	1.122	1.120	1.081	1.114	1.100	1.056	1.121
Fully restricted	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	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
Intercept only	1.1438	1.1264	1.1631	1.1533	1.1219	1.1760	1.1617
Fully restricted	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	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
Intercept only	1.2461	1.3029	1.2143	1.2459	1.2700	1.2279	1.2274
Fully restricted	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	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
Intercept only	1.360	1.384	1.288	1.242	1.238	1.304	1.322
Fully restricted	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	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 4a. Out-of-Sample Forecast Accuracy:
U.S. Inflation Forecasts: 1 Month Ahead
Evaluation Period: 1983.8-2003.7

	Models with Indicators of Economic Activity						
	PMSE Relative to 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	-	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 PMSE 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 4b. Out-of-Sample Forecast Accuracy:
U.S. Inflation Forecasts: 12 Months Ahead
Evaluation Period: 1983.8-2003.7

	Models with Indicators of Economic Activity						
	PMSE 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 PMSE 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 $m = 12$.

Table 5. Out-of-Sample Forecast Accuracy:
U.S. Inflation Forecasts: 1 Month and 12 Months Ahead
Evaluation Period: 1983.8-2003.7

(a)	**Shrinkage Estimator of Unrestricted Model**						
	PMSE Relative to Benchmark						
	Bayesian shrinkage estimator					**UR**	**BA**
	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 100$	$\lambda = \infty$	$\alpha = 0.05$
$h = 1$	**0.809**	0.826	0.843	0.865	0.885	0.885	0.833
$h = 12$	0.710	0.703	0.696	0.695	0.695	0.695	**0.582**

(b)	**Bagging Estimator with Size α**				
	PMSE Relative to Benchmark				
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.50$
$h = 1$		0.833			
$h = 12$		0.582			

(c)	**Bayesian Model Averaging: One Extra Predictor at a Time**							
	PMSE Relative to Benchmark							
	Median	**Equal-weighted**	**BMA**					**BA**
		$\phi = 0$	$\phi = 0.01$	$\phi = 0.05$	$\phi = 0.1$	$\phi = 0.2$	$\phi = 0.3$	$\alpha = 0.05$
$h = 1$	0.993	0.974	0.970	0.910	0.904	0.915	0.919	**0.833**
$h = 12$	0.947	0.871	0.852	0.843	0.885	0.948	0.958	**0.582**

(d)	**Bayesian Model Averaging: Random Sets of Extra Predictors**								
	PMSE Relative to Benchmark								
	Equal-weighted	**BMA**							**BA**
	$\phi = 0$	$\phi = 0.01$	$\phi = 0.05$	$\phi = 0.1$	$\phi = 0.2$	$\phi = 0.5$	$\phi = 1$	$\phi = 2$	$\alpha = 0.05$
$h = 1$	0.820	**0.804**	0.817	0.819	0.827	0.828	0.832	0.839	0.833
$h = 12$	0.622	0.643	0.676	0.681	0.666	0.656	0.645	0.646	**0.582**

SOURCE: See Tables 4a and 4b.

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

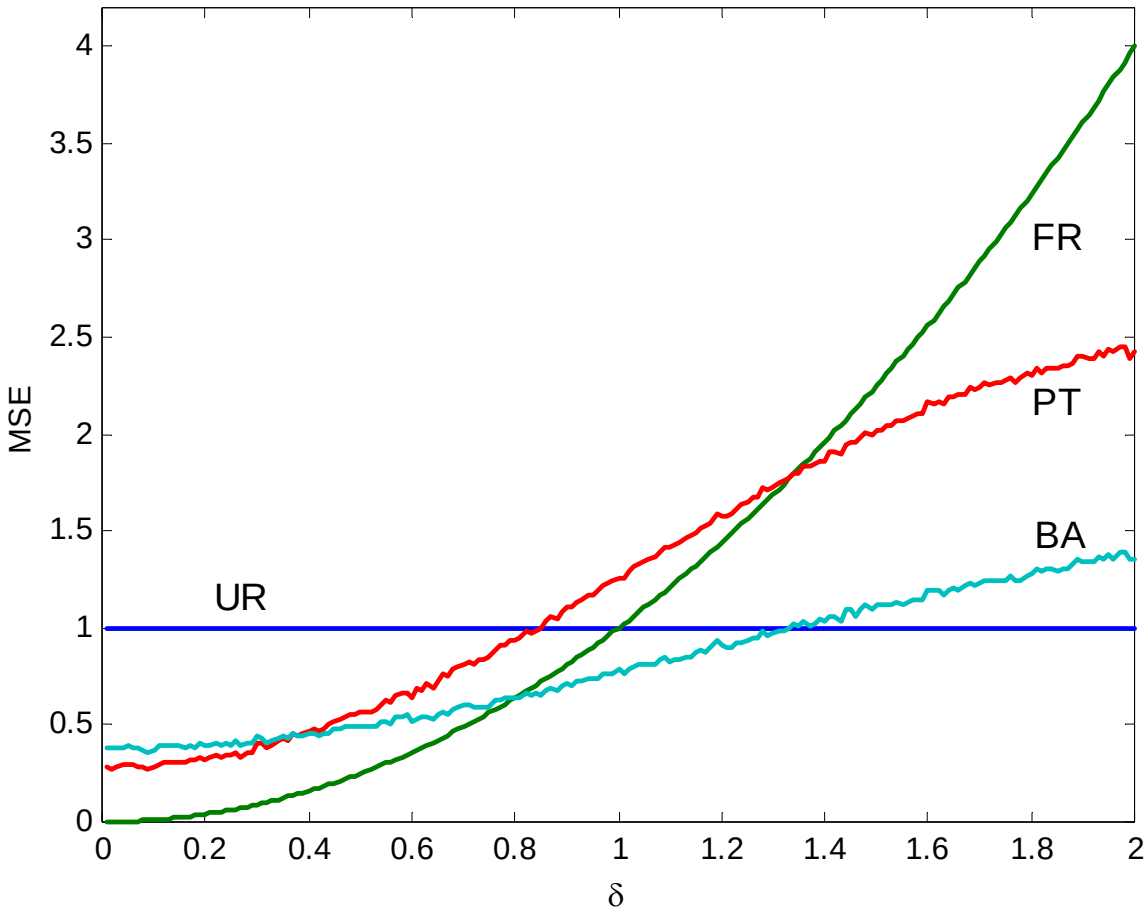

Figure 2: Squared Asymptotic Bias, Asymptotic Variance and Asymptotic MSE of PT and BA Predictors

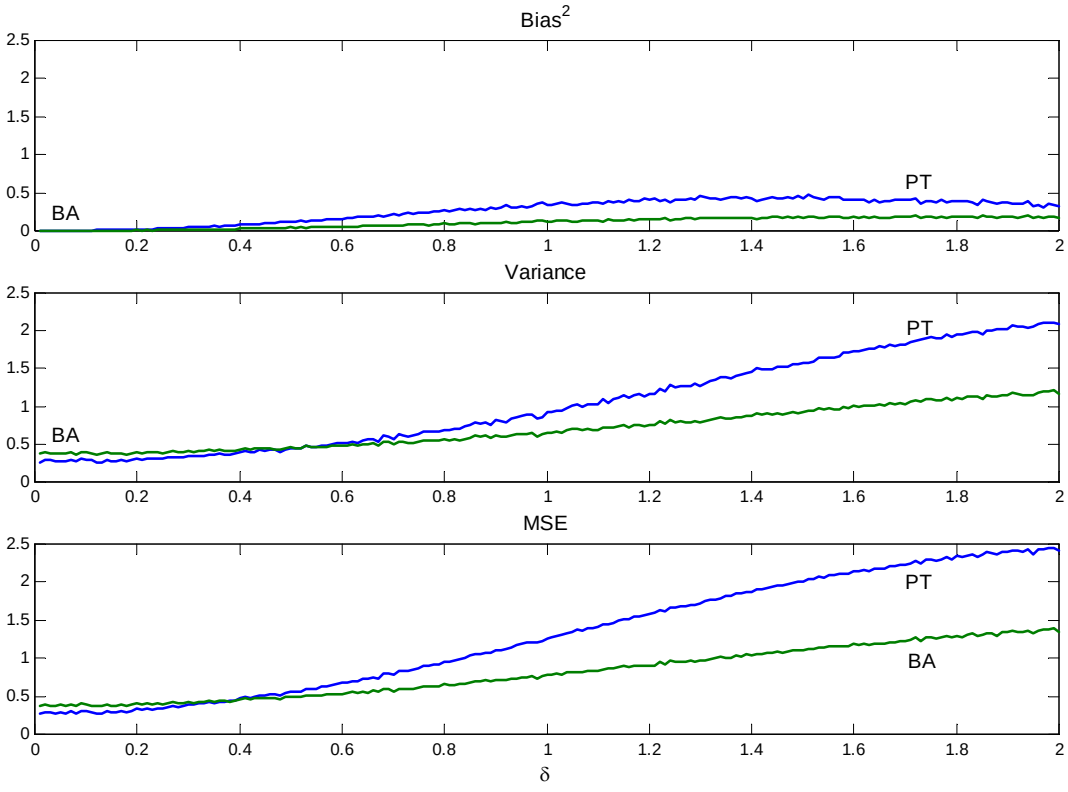

Figure 3: Gains from Bagging as a Function of M

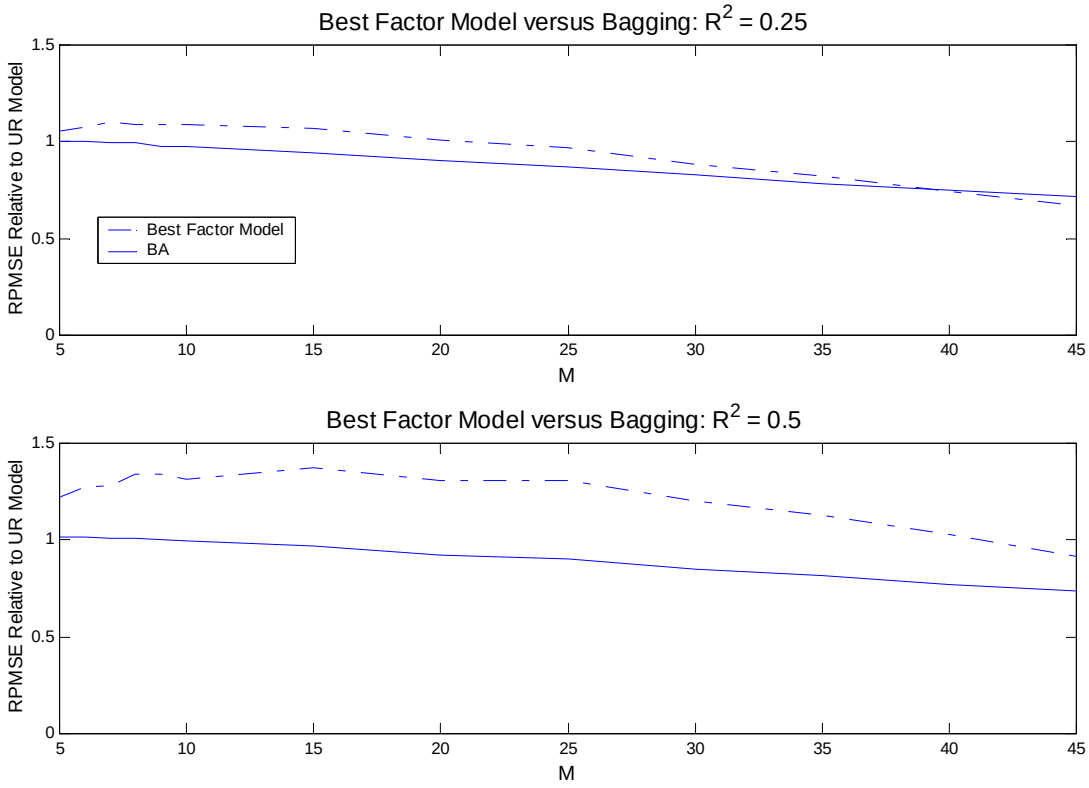

Figure 4: Gains from Bagging as a Function of the Strength of the Common Component