How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation

Atsushi INOUE and Lutz KILIAN

This article focuses on the widely studied question of whether the inclusion of indicators of real economic activity lowers the prediction mean squared error of forecasting models of U.S. consumer price inflation. We propose three variants of the bagging algorithm specifically designed for this type of forecasting problem and evaluate their empirical performance. Although bagging predictors in our application are clearly more accurate than equally weighted forecasts, median forecasts, ARM forecasts, AFTER forecasts, or Bayesian forecast averages based on one extra predictor at a time, they are generally about as accurate as the Bayesian shrinkage predictor, the ridge regression predictor, the iterated LASSO predictor, or the Bayesian model average predictor based on random subsets of extra predictors. Our results show that bagging can achieve large reductions in prediction mean-squared errors even in such challenging applications as inflation forecasting; however, bagging is not the only method capable of achieving such gains.

KEY WORDS: Bayesian model averaging; Bootstrap aggregation; Factor model; Forecast combination; Forecasting model selection; Pre testing; Shrinkage estimation.

1. INTRODUCTION

A common situation in out-of-sample prediction is that many potentially useful predictors are available to the forecaster, but few (if any) of these predictors have high predictive power, and many of the potential predictors are correlated. This situation is particularly relevant for economic forecasting, because economic theory rarely puts tight restrictions on the set of potential predictors. In addition, usually several alternative proxies for the same economic variable are available to the forecaster. A case in point is the widely studied problem of forecasting consumer price inflation based on measures of real economic activity. There are many alternative measures of real economic activity, including the unemployment rate, industrial production growth, housing starts, capacity utilization rates in manufacturing, and the number of help wanted postings, all of which are believed to have some predictive power for consumer price inflation. It is well known that forecasts generated using only one of these indicators 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 for real activity is thought to lead to overfitting and poor out-of-sample forecast accuracy. Moreover, standard methods of ranking all possible combinations of predictors by means of an information criterion function and selecting the combination that minimizes the criterion, as discussed by Inoue and Kilian 2006, become computationally impractical when there are many potential predictors. This suggests that this forecasting problem is a natural candidate for the application of *bootstrap aggregation* or *bagging* methods, as discussed by Breiman (1996) and Bühlmann and Yu (2002).

Bagging is a statistical method designed to reduce the out-of-sample prediction mean squared error (PMSE) of forecasting models selected by unstable decision rules such as pretests. Bagging involves generating a large number of bootstrap resamples of the original forecasting problem, applying a pretest model selection rule to each of the resamples, and averaging the forecasts from the models selected by the pretest on each bootstrap sample. Although bagging has been found to work well in many statistical applications, to date little attention seems to have been devoted to the problem of bagging dynamic linear regression models with correlated regressors. This latter model is widely used in economic applications. In this article we propose three variants of the bagging algorithm specifically designed for this type of forecasting model: the BA method for models with possibly correlated regressors, the CBA method for models with orthogonalized regressors, and the BA^F method for factor models. We compare the forecast accuracy of these bagging methods relative to one another as well as relative to some of the leading alternatives to bagging discussed in the literature.

One of those leading alternatives 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 use regression-based weights for forecast combination (see Bates and Granger 1969; Stock and Watson 2003). There is no reason for simple averages to be optimal, however, and the latter approach of regression-based weights tends to perform poorly in practice (see, e.g., Stock and Watson 1999). More sophisticated methods of forecast averaging include adaptive regression by mixing, or ARM (see Yang 2001, 2003), and aggregation of forecasts through exponential reweighting, or AFTER (see Yang 2004). Alternatively, one could weight individual forecasts by the posterior probabilities of each forecasting model (see, e.g., Min and Zellner 1993; Avramov 2002; Cremers 2002; Wright 2003a; Koop and Potter 2004 for applications in econometrics). This Bayesian model averaging (BMA) approach has been used successfully in forecasting inflation by Wright (2003b). A second leading alternative 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). Shrinkage estimation also could be implemented based on the well-known ridge estimator. Yet another variation on this idea is the least-absolute shrinkage and selection operator (LASSO) of Tibshirani (1996), which combines features of shrinkage and model selection. A third leading alternative is to

Atsushi Inoue is Associate Professor, Department of Economics, University of British Columbia, Vancouver, BC V6T 1Z1, Canada and Associate Professor, Department of Agricultural and Resource Economics, North Carolina State University, Raleigh, NC 27695 (E-mail: *ainoue@interchange.ubc.ca*). Lutz Kilian is Associate Professor, Department of Economics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *lkilian@umich.edu*). This article has benefited from discussions with numerous people, most notably Tae-Hwy Lee and Mark Watson, the editor, and two anonymous referees.

Journal of the American Statistical Association
June 2008, Vol. 103, No. 482, Applications and Case Studies
DOI 10.1198/016214507000000473

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 factor model, then factors estimated by principal components analysis can be used for forecasting under quite general conditions (see, e.g., Stock and Watson 2002a,b). A closely related approach to extracting common components has been developed by Forni, Hallin, Lippi, and Reichlin (2000, 2005) and applied by Forni et al. (2003).

The remainder of the article is organized as follows. In Section 2 we show how the bagging proposal may be adapted to applications involving dynamic linear multiple regression with possibly serially correlated and heteroscedastic errors that are typical of the inflation forecasting problem. We discuss applications of bagging in the correlated regressor model as well as in factor models. In Section 3 we investigate whether adding indicators of real economic activity to models involving lagged inflation rates improves the accuracy of 1- and 12-month-ahead forecasts of U.S. consumer price inflation, as measured by reductions in the prediction mean-squared error. This empirical application is in the spirit of recent work by Stock and Watson (1999, 2003), Marcellino, Stock, and Watson (2003), Bernanke and Boivin (2003), Forni et al. (2003), and Wright (2003b), among others. We show that none of the forecasting methods considered uniformly dominates the other methods in this application. Nevertheless, the results are fairly clear-cut in that some forecasting methods perform well at both 1-month and 1-year horizons, whereas other methods do not.

We demonstrate that there is no clear ranking between the standard bagging method for unorthogonalized predictors (BA) and the orthogonalized bagging method (CBA). Both perform about equally well. The bagging predictor in always more accurate than the unrestricted model and typically at least as accurate as the pretest predictor. Although bagging predictors are clearly more accurate than equally weighted forecasts, median forecasts, ARM forecasts, AFTER forecasts, and Bayesian forecast averages based on one extra predictor at a time, they are generally about as accurate as the Bayesian shrinkage estimator, the ridge regression predictor, the iterated LASSO predictor, and the Bayesian model average predictor based on random subsets of extra predictors. At the 1-month horizon, all of these methods achieve gains relative to the inflation-only model of 16–18%. At the 1-year horizon, the gains increase to 35–40%. The ARM and AFTER predictors based on random subsets of extra predictors also perform well in some cases, but their performance is more uneven across horizons.

We find that the bagging predictors based on the regression model typically are more accurate than a factor model predictor of given rank r for the same data set. Although bagging the r largest principal components (BA^F) tends to improve the forecast accuracy of a factor model for conventional choices of r, often the gains are not as large as those for bagging methods designed for standard regression models.

In Section 4 we conclude that bagging can achieve large reductions in prediction mean-squared errors, even in such challenging applications as inflation forecasting. The gains in accuracy compare favorably with the benchmark model and with results reported in previous studies. But bagging is not the only method capable of achieving such gains. The high accuracy of the ridge regression predictor in particular suggests that similar accuracy is feasible at a much lower computational cost than required for bagging or for BMA based on random subsets of predictors. Recently proposed asymptotic approximations to bagging methods for orthogonalized predictors may eliminate the need for computer simulation in bagging, however. It will be of interest to see how these new approaches compare with the full-fledged bagging approach used in this article.

2. THREE PROPOSALS FOR BAGGING CORRELATED REGRESSORS IN DYNAMIC REGRESSOR MODELS

Most of the existing literature on bagging has focused on cross-sectional settings. Only recently has interest grown in applications of bagging to time series models. For example, Lee and Yang (2006) in related work studied the properties of bagging in binary prediction problems and quantile prediction of economic time series data. Whereas those authors focused on the ability of bagging to improve forecast accuracy under asymmetric loss functions, we study the ability of the bagging predictor to reduce the PMSE. More specifically, we are concerned with the usefulness of bagging methods in forecasting economic time series from dynamic linear multiple regression models with correlated regressors. Such forecasting models are routinely used by practitioners, but no attempt has been made to use bagging methods in this context. Our application is motivated by the common problem of assessing the incremental predictive power of a vector of possibly mutually correlated predictors, the prototypical example of which is the inflation forecasting problem that is examined in Section 3. We consider two alternative frameworks, the correlated regressor model and the factor model, and discuss how to implement bagging in each case.

2.1 Correlated Regressor Model

Consider the forecasting model

$$y_{t+h} = \alpha' w_t + \beta' x_t + \varepsilon_{t+h}, \qquad h = 1, 2, 3, \ldots, \tag{1}$$

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 possibly correlated predictors at time period t. We presume that y_t and x_t are joint covariance stationary processes or have been suitably transformed to achieve covariance stationarity. w_t is a vector of predetermined variables, such as deterministic regressors or lagged dependent variables with coefficient vector α. Model selection is applied only to the elements of x_t. For expository purposes, we suppress $\alpha' w_t$ in the description of the bagging algorithms.

2.1.1 Bagging Unorthogonalized Predictors. We begin by establishing some notation. Suppose that we are interested in predicting the scalar y_{T+h} based on x_T, where T denotes the most recent observation available to the forecaster and $x_t \in \Re^M$. Let $\widehat{\beta}$ 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 0 in the unrestricted model, where β_j is the jth element of β. Furthermore, let $\widehat{\gamma}$ denote the OLS estimator of the forecasting model after variable selection. Then the predictor from the unrestricted model (UR), the predictor from the fully restricted model (FR), and the pretest (PT) predictor conditional on x_T are

$$\hat{y}^{\text{UR}}_{T+h}(x_T) = \widehat{\beta}' x_T,$$

$$\hat{y}^{\text{FR}}_{T+h}(x_T) = 0,$$

$$\hat{y}^{\text{PT}}_{T+h}(x_T) = 0, \quad \text{if } |t_j| < c \ \forall j \qquad \text{and}$$

$$\hat{y}^{\text{PT}}_{T+h}(x_T) = \widehat{\gamma}' S_T x_T \quad \text{otherwise,}$$

where S_T is the stochastic selection matrix obtained from the $M \times M$ diagonal matrix with (i, i)th element $I(|t_i| > c)$ by deleting rows of 0's and c is the critical value of the pretest.

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 PT model forecast is obtained as follows. We first fit the unrestricted model that includes all potential predictors, then conduct two-sided t-tests on each slope parameter based on a prespecified critical value c. We discard the insignificant predictors and reestimate the final model before generating the PT forecast. In constructing the t-statistic we use appropriate standard errors that allow for serial correlation and/or conditional heteroscedasticity. Specifically, when the error term follows an MA$(h-1)$ process, the pretest strategy may be implemented based on White (1980) robust standard errors for $h = 1$ or West (1997) robust standard errors for $h > 1$. For more general error structures, nonparametric robust standard errors, such as the HAC estimator proposed by Newey and West (1987), would be appropriate.

The bootstrap aggregated or bagging predictor is obtained by averaging the pretest predictor across bootstrap replications. In principle, bagging can be applied to any pretesting strategy, not just to the specific pretesting strategy discussed here, and there is no reason to believe that our t-test strategy is the best choice. Nevertheless, the simple t-test strategy studied here appears to work well in many cases. Throughout the article we use * to denote the bootstrap draws of the data and the bootstrap equivalents of previously defined estimators. For example, if $\widehat{\beta}$ denotes the OLS estimator of β in model (1) conditional on the original data set $(y_{1+h}, x_1'), \ldots, (y_T, x_{T-h}')$, then $\widehat{\beta}^*$ denotes the corresponding OLS estimator based on the bootstrap data set $(y^*_{1+h}, x_1'^*), \ldots, (y^*_T, x'^*_{T-h})$.

Proposal 1 (BA method). The bagging predictor in the standard regression framework is defined as follows:

a. Arrange the set of tuples $\{(y_{t+h}, x_t')\}$, $t = 1, \ldots, T-h$, 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 $(y^*_{1+h}, x_1'^*), \ldots, (y^*_T, x'^*_{T-h})$ 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).

b. For each bootstrap sample, compute the bootstrap pretest predictor conditional on x_T,

$$\hat{y}^{*\text{PT}}_{T+h}(x_T) = 0, \quad \text{if } |t^*_j| < c \ \forall j \qquad \text{and}$$

$$\hat{y}^{*\text{PT}}_{T+h}(x_T) = \widehat{\gamma}^{*\prime} S^*_T x_T \quad \text{otherwise,}$$

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

$$\widehat{V}^* = \frac{1}{bm} \sum_{k=1}^{b} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(x^*_{(k-1)m+i} \varepsilon^*_{(k-1)m+i+h}\right) \times \left(x^*_{(k-1)m+j} \varepsilon^*_{(k-1)m+j+h}\right)',$$

$$\widehat{H}^* = \frac{1}{bm} \sum_{k=1}^{b} \sum_{i=1}^{m} \left(x^*_{(k-1)m+i} x^{*\prime}_{(k-1)m+i}\right),$$

$\varepsilon^*_{t+h} = y^*_{t+h} - \widehat{\beta}^{*\prime} x^*_t$, and b is the integer part of $(T-h)/m$ (see, e.g., Inoue and Shintani 2006).

c. The bagged predictor is the expectation of the bootstrap pretest predictor across bootstrap samples, conditional on x_T,

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

where E^* denotes the expectation with respect to the bootstrap probability measure. The bootstrap expectation in part c may be evaluated by simulation,

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

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 forecasting model at horizon h is specified correctly 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).

The performance of bagging will in general depend on the critical value chosen for pretesting, not unlike the way in which shrinkage estimators depend on the degree of shrinkage. In practice, we consider a grid of alternative values of c.

2.1.2 Bagging Orthogonalized Predictors. One seeming drawback of Proposal 1 is that when predictors are correlated, the effective size of the t-tests on individual predictors cannot be controlled. This fact suggests an alternative approach to bagging in which the predictors are orthogonalized before conducting the t-tests. When the regressor matrix is of full column rank, this may be accomplished by computing the orthogonalized predictor $\widetilde{x}_t = P'^{-1} x_t$, where P is the Cholesky decomposition of $E(x_t x_t')$, that is, the unique $M \times M$ upper-triangular matrix such that $P'P = E(x_t x_t')$. This allows us to express the forecasting model (1) equivalently as

$$y_{t+h} = \alpha' w_t + \beta' \widetilde{x}_t + \varepsilon_{t+h}, \qquad h = 1, 2, 3, \ldots, \qquad (2)$$

where $\alpha' w_t$ is suppressed from now on for notational convenience. The unrestricted and fully restricted predictors will be unaffected by this transformation. We now introduce the pretest estimator for this transformed model, designated CPT,

$$\hat{y}^{\text{CPT}}_{T+h}(\widetilde{x}_T) = 0, \quad \text{if } |t_j| < c \ \forall j \qquad \text{and}$$

$$\hat{y}^{\text{CPT}}_{T+h}(\widetilde{x}_T) = \widehat{\gamma}' S_T \widetilde{x}_T \quad \text{otherwise},$$

where the notation is analogous to the correlated regressor case. The corresponding bagging predictor is described next.

Proposal 2 (CBA method). The bagging predictor for the orthogonalized regressors may be obtained through a Cholesky decomposition as follows:

a. Arrange the set of tuples $\{(y_{t+h}, x_t')\}$, $t = 1, \dots, T-h$, 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 $(y^*_{1+h}, x_1'^*), \dots, (y^*_T, x'^*_{T-h})$ 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.

b. Compute the orthogonalized predictor $\widetilde{x}_t = P'^{-1} x_t$, where P is the Cholesky decomposition of $E(x_t x_t')$, that is, the $M \times M$ upper triangular matrix such that $P'P = E(x_t x_t')$. For each bootstrap sample, compute the bootstrap pretest predictor conditional on $\widetilde{x}_T$,

$$\hat{y}^{*\text{CPT}}_{T+h}(\widetilde{x}_T) = 0, \quad \text{if } |t^*_j| < c \ \forall j \qquad \text{and}$$

$$\hat{y}^{*\text{CPT}}_{T+h}(\widetilde{x}_T) = \widehat{\gamma}^{*\prime} S^*_T \widetilde{x}_T \quad \text{otherwise},$$

where $\widehat{\gamma}^*$ and S^*_T are the bootstrap analogs of $\widehat{\gamma}$ and S_T, applied to the orthogonalized predictor model. When constructing $|t^*_j|$, we compute the variance of $\sqrt{T}\widehat{\beta}^*$ as $\widehat{H}^{*-1}\widehat{V}^*\widehat{H}^{*-1}$, where

$$\widehat{V}^* = \frac{1}{bm} \sum_{k=1}^{b} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\widetilde{x}^*_{(k-1)m+i} \widetilde{\varepsilon}^*_{(k-1)m+i+h}\right) \times \left(\widetilde{x}^*_{(k-1)m+j} \widetilde{\varepsilon}^*_{(k-1)m+j+h}\right)',$$

$$\widehat{H}^* = \frac{1}{bm} \sum_{k=1}^{b} \sum_{i=1}^{m} \left(\widetilde{x}^*_{(k-1)m+i} \widetilde{x}^{*\prime}_{(k-1)m+i}\right),$$

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

c. The bagged predictor is the expectation of the bootstrap pretest predictor across bootstrap samples, conditional on $\widetilde{x}_T$,

$$\hat{y}^{\text{CBA}}_{T+h}(\widetilde{x}_T) = E^*[\widehat{\gamma}^{*\prime} S^*_T \widetilde{x}_T],$$

where E^* denotes the expectation with respect to the bootstrap probability measure. The bootstrap expectation in part a may be evaluated by simulation based on B bootstrap replications,

$$\hat{y}^{\text{CBA}}_{T+h}(\widetilde{x}_T) = \frac{1}{B} \sum_{i=1}^{B} \widehat{\gamma}^{*i\prime} S^{*i}_T \widetilde{x}_T .$$

Although the size of the pretest is easier to control in this framework, it is unclear a priori whether the CBA method will select superior forecasting models. In the context of bagging the purpose of the pretest is to select a forecasting model with lower PMSE, not to uncover the true relationships in the data. Notwithstanding the existence of size distortions, the BA method may lower the PMSE even in the presence of correlated regressors. An interesting question to be addressed in the empirical section is whether the performance of bagging may be improved by orthogonalizing the predictors.

2.2 Factor Models

Bagging methods for the correlated regressor model are not designed to handle situations where the regressor matrix is of reduced rank. A leading example of a reduced-rank structure is a factor model. In that case the forecasting model reduces to

$$y_{t+h} = \alpha' w_t + \beta' f_t + \varepsilon_{t+h}, \qquad h = 1, 2, 3, \dots, \tag{3}$$

where f_t denotes a vector of the r largest factors that may be extracted from the set of N potential predictors by principal components analysis (see, e.g., Stock and Watson 2002a,b). We denote that estimator by $\widehat{f}_t$. By construction, the estimated principal components or factors are orthogonal. If $N, T \to \infty$, then $\widehat{f}_t$ is consistent for f_t (see Stock and Watson 2002a, thm. 1, p. 1169). As before, w_t denotes a vector of predetermined regressors that we condition on in making forecasts. For notational convenience, we again suppress $\alpha' w_t$ in the description of the bagging algorithm.

2.2.1 Bagging Factor Predictors. Although principal components analysis generates N factors, in practice, researchers have typically focused on a small number of factors in generating forecasts from factor models, while ignoring information contained in the other factors. The obvious question is whether applying bagging to a suitably chosen subset of the principal components may improve forecast accuracy. We consider an r-dimensional subset of the N principal components that includes the r largest principal components where $r < T$. It is straightforward to adapt the bagging method to this situation. As before, we start by defining the pretest estimator. The unrestricted predictor in the factor model will be equivalent to the standard factor model forecast. We refer to this predictor as UR^F. We denote the pretest predictor in the factor model by PT^F. By analogy to the notation used for the correlated regressor model,

$$\hat{y}^{\text{PT}^F}_{T+h}(\widehat{f}_T) = 0, \quad \text{if } |t_j| < c \ \forall j \qquad \text{and}$$

$$\hat{y}^{\text{PT}^F}_{T+h}(\widehat{f}_T) = \widehat{\gamma}' S_T \widehat{f}_T \quad \text{otherwise}.$$

The corresponding bagging predictor is described in Proposal 3.

Proposal 3 (BA^F method). The bagging predictor in the factor model framework is defined as follows:

a. Use principal components analysis to extract the r largest common factors from the $T \times N$ matrix X of potential predictors. Denote the date t observation of these factor estimates by the $r \times 1$ vector $\widehat{f}_t$.

b. Arrange the set of tuples $\{(y_{t+h}, \widehat{f}_t')\}$, $t = 1, \dots, T-h$, in the form of a matrix of dimension $(T-h) \times (r+1)$,

$$\begin{matrix} y_{1+h} & \widehat{f}_1' \\ \vdots & \vdots \\ y_T & \widehat{f}_{T-h}' \end{matrix} .$$

Construct bootstrap samples $(y^*_{1+h}, \widehat{f}_1'^*), \dots, \{(y^*_T, \widehat{f}'^*_{T-h})$ 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, and subsequently orthogonalizing the bootstrap factor draws through principal components.

c. For each bootstrap sample, compute the bootstrap pretest predictor conditional on $\widehat{f}_T$,

$$\hat{y}^{*\mathrm{PT}^F}_{T+h}(\widehat{f}_T) = 0, \quad \text{if } |t^*_j| < c \ \forall j \qquad \text{and}$$

$$\hat{y}^{*\mathrm{PT}^F}_{T+h}(\widehat{f}_T) = \widehat{\gamma}^{*\prime} S^*_T \widehat{f}_T \quad \text{otherwise,}$$

where $\widehat{\gamma}^*$ and S^*_T are the bootstrap analogs of $\widehat{\gamma}$ and S_T, applied to the factor model. When constructing $|t^*_j|$, compute the variance of $\sqrt{T}\widehat{\beta}^*$ as $\widehat{H}^{*-1}\widehat{V}^*\widehat{H}^{*-1}$, where

$$\widehat{V}^* = \frac{1}{bm}\sum_{k=1}^{b}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\widehat{f}^*_{(k-1)m+i}\widehat{\varepsilon}^*_{(k-1)m+i+h}\right) \times \left(\widehat{f}^*_{(k-1)m+j}\widehat{\varepsilon}^*_{(k-1)m+j+h}\right)',$$

$$\widehat{H}^* = \frac{1}{bm}\sum_{k=1}^{b}\sum_{i=1}^{m}\left(\widehat{f}^*_{(k-1)m+i}\widehat{f}^{*\prime}_{(k-1)m+i}\right),$$

$\widehat{\varepsilon}^*_{t+h} = y^*_{t+h} - \widehat{\beta}^{*\prime}\widehat{f}^*_t$, and b is the integer part of $(T-h)/m$.

d. The bagged predictor is the expectation of the bootstrap pretest predictor across bootstrap samples, conditional on $\widehat{f}_T$,

$$\hat{y}^{\mathrm{BA}^F}_{T+h}(\widehat{f}_T) = E^*[\widehat{\gamma}^{*\prime} S^*_T \widehat{f}_T],$$

where E^* denotes the expectation with respect to the bootstrap probability measure. The bootstrap expectation in part d may be evaluated by simulation based on B bootstrap replications,

$$\hat{y}^{\mathrm{BA}^F}_{T+h}(\widehat{f}_T) = \frac{1}{B}\sum_{i=1}^{B}\widehat{\gamma}^{*i\prime} S^{*i}_T \widehat{f}_T.$$

3. DO INDICATORS OF REAL ECONOMIC ACTIVITY IMPROVE THE ACCURACY OF U.S. INFLATION FORECASTS?

In this section we return to the problem that motivated the development of the bagging methods in the previous section and investigate whether 1-month- and 12-months-ahead U.S. consumer price inflation forecasts may be improved by adding indicators of real economic activity to models involving only lagged inflation rates. This empirical application 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.

We follow the common practice of choosing between competing forecasting methods based on the ranking of their recursive PMSEs in simulated out-of-sample forecasts. The idea is to apply each forecasting method in real time, using only information available at each point in time, and compare the resulting forecasts with the actual realizations of inflation. As time progresses, the forecaster recursively updates the information set and generates a new forecast. Thus a given forecasting method may select different models at each point in time. The recursive PMSE of each forecasting method is obtained by averaging the sequence of squared forecast errors.

The choice of the inflation-only benchmark model is conventional (see, e.g., Stock and Watson 2003; Forni et al. 2003), as is the focus on the PMSE. The measure of inflation is based on the seasonally adjusted urban consumer price index. The lag order of the benchmark model is determined by the Akaike information criterion subject to an upper bound of 12 lags. Because the lag order of the benchmark model is selected recursively in real time, it may change as we move through the sample.

The benchmark model is compared with various alternative forecasting strategies that exploit in addition information about indicators of real economic activity. Because there is no universally agreed-on measure of real economic activity, we consider 30 potential extra predictors that a priori can be expected to be correlated with real economic activity and are available at monthly frequency. These predictors include production data, labor market data, monetary and financial data, and external data. A complete variable list and the data sources are provided at the end of the article. Note that measures of wage cost and productivity are not available at a monthly frequency for our sample period.

We use monthly data for 1971.4–2003.7. The starting point is dictated by data constraints. We set aside the last 20 years of data as our forecast evaluation period. 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 mean 0), as is customary in the factor model literature. We do not attempt to identify and remove outliers.

3.1 Unrestricted, Pretest, and Bagging Forecasts in the Correlated Regressor Model

The alternative forecasting strategies under consideration in the first round of comparisons include the benchmark model involving only an intercept, the current value of inflation and lags of monthly inflation, as well as five alternative methods that also include at least some indicators of economic activity. The unrestricted regression model (UR) includes current values of all 30 indicators of economic activity as separate regressors in addition to current and lagged inflation. The pretest predictors (PT and CPT) use only a subset of these additional predictors. The subsets for the pretest strategy are selected using two-sided t-tests for each predictor. We experimented with a range of critical values, $c \in \{.3853, .6745, 1.2816, 1.4395, 1.6449, 1.9600, 2.2414, 2.5758, 2.8070, 3.0233, 3.2905, 3.4808, 3.8906, 4.4172, 5.3267\}$. To conserve space, only results for the value of c that produced the lowest recursive PMSE are reported.

The bagging forecasts (BA and CBA) are computed as the average of the corresponding pretest forecasts across 100 bootstrap replications with $M = 30$. We consider the same range of critical values as in the case of the pretest predictors and report only the best result. For the 1-month-ahead forecasting model, there is no evidence of serial correlation in the unrestricted model, so we use White (1980) robust standard errors for the pretests and the pairwise bootstrap. For the 12-month ahead-forecast, we use West (1997) standard errors with a truncation lag of 11 and the block bootstrap with $m = 12$. To summarize, the following forecasting methods are under consideration:

$$\text{Benchmark:} \quad \pi^h_{t+h|t} = \widehat{\nu} + \sum_{k=1}^{p}\widehat{\phi}_k \pi_{t-k+1},$$

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

$$\text{PT:}\quad \pi^h_{t+h|t} = \widehat{\nu} + \sum_{k=1}^{p} \widehat{\phi}_k \pi_{t-k+1} + \sum_{j=1}^{M} \widehat{\gamma}_j I(|t_j| > c) x_{j,t},$$

$$\text{CPT:}\quad \pi^h_{t+h|t} = \widehat{\nu} + \sum_{k=1}^{p} \widehat{\phi}_k \pi_{t-k+1} + \sum_{j=1}^{M} \widehat{\gamma}_j I(|t_j| > c) \widetilde{x}_{j,t},$$

$$\text{BA:}\quad \pi^h_{t+h|t} = \frac{1}{100} \sum_{i=1}^{100} \left(\widehat{\nu}^{*i} + \sum_{k=1}^{p} \widehat{\phi}^{*i}_k \pi_{t-k+1} + \sum_{j=1}^{M} \widehat{\gamma}^{*i}_j I(|t^{*i}_j| > c) x_{j,t} \right),$$

and

$$\text{CBA:}\quad \pi^h_{t+h|t} = \frac{1}{100} \sum_{i=1}^{100} \left(\widehat{\nu}^{*i} + \sum_{k=1}^{p} \widehat{\phi}^{*i}_k \pi_{t-k+1} + \sum_{j=1}^{M} \widehat{\gamma}^{*i}_j I(|t^{*i}_j| > c) \widetilde{x}_{j,t} \right),$$

where π^h_{t+h} denotes the rate of inflation over the period t to $t+h$ and the superscript i denotes parameter estimates for the ith bootstrap replication.

The accuracy of each forecasting method is measured by the average of the squared forecast errors obtained by recursively reestimating the model at each point in time t and forecasting π^h_{t+h}. Throughout the article, all results are normalized relative to the recursive PMSE of the benchmark model, such that a ratio below 1 indicates that the method in question is more accurate than the benchmark model.

Table 1 summarizes the results for the unrestricted model, the two pretest methods and the two bagging methods. It shows results for both the 1-month-ahead forecasts of U.S. consumer price inflation ($h = 1$) and the corresponding results for 1-year-ahead forecasts ($h = 12$). Table 1 shows that all forecasting methods under consideration beat the lagged inflation-only benchmark model. Even the UR model constitutes a clear improvement over the benchmark model, with PMSE gains of 8% for $h = 1$ and 30% for $h = 12$. Pretesting further improves forecasting accuracy relative to the unrestricted model. The PT predictor increases the accuracy gains to 14% relative to the benchmark model at $h = 1$ and is about as accurate as the UR model at $h = 12$. The CPT predictor improves accuracy by only 10% at $h = 1$, but lowers the PMSE by 36% at $h = 12$.

Table 1. U.S. consumer price inflation forecasts, evaluation period: 1983.8–2003.7

	Correlated regressor model Recursive PMSE relative to benchmark model				
Horizon	UR	PT	BA	CPT	CBA
1 month	.923	.857	.818	.897	.842
12 months	.703	.697	.653	.636	.637

Table 1 suggests that the PT predictor can in turn be improved on by bagging. The PMSE gains from using the BA predictor are 18% at $h = 1$, and 35% at $h = 12$. The additional reduction in PMSE is about 4 percentage points at both horizons. The corresponding gains for the CBA predictor relative to the benchmark model are 16% at $h = 1$, and 36% at $h = 12$, making the CBA predictor about equally accurate as or more accurate than the CPT predictor. Table 1 also suggests that neither of the two bagging predictors dominates the other.

3.2 Unrestricted, Pretest, and Bagging Forecasts in Factor Models

The empirical results in Table 1 have been obtained under the premise that the regressor matrix is of full column rank. An alternative view is that indicators of real economic activity are well approximated by factor models. That interpretation suggests that we impose the factor structure on the forecasting model and examine the implied UR^F, PT^F, and BA^F predictors. Each of this predictors is based on the first r of the 30 principal components that can be constructed from the set of indicators of real economic activity. The principal components are constructed as done by Stock and Watson (2002a,b). To summarize, the methods under consideration are

$$\text{UR}^F:\quad \pi^h_{t+h|t} = \widehat{\nu} + \sum_{k=1}^{p} \widehat{\phi}_k \pi_{t-k+1} + \sum_{j=1}^{r} \widehat{\beta}_j \widehat{f}_{j,t},$$

$$\text{PT}^F:\quad \pi^h_{t+h|t} = \widehat{\nu} + \sum_{k=1}^{p} \widehat{\phi}_k \pi_{t-k+1} + \sum_{j=1}^{r} \widehat{\gamma}_j I(|t_j| > c) \widehat{f}_{j,t},$$

and

$$\text{BA}^F:\quad \pi^h_{t+h|t} = \frac{1}{100} \sum_{i=1}^{100} \left(\widehat{\nu}^{*i} + \sum_{k=1}^{p} \widehat{\phi}^{*i}_k \pi_{t-k+1} + \sum_{j=1}^{r} \widehat{\gamma}^{*i}_j I(|t^{*i}_j| > c) \widehat{f}_{j,t} \right).$$

The benchmark model is the same as in the preceding section. We do not report results for models with more lags of the estimated factors to conserve space. We found that the performance of these models rarely improves with more than one lag of the principal components.

Table 2 presents results for $r \in \{1, 2, 3, 4, 5, 6, 7, 8\}$, which covers most cases of practical interest. It is rare for users of factor models to use more than three principal components in generating forecasts. The UR^F results in Table 2 suggest that factor model forecasts are reasonably accurate in this application at the 1-year horizon with gains ranging from 27% to 36% relative to the benchmark model, but less accurate at the 1-month horizon with gains ranging from −2% to 8%, depending on the rank r.

Table 2. U.S. consumer price inflation forecasts, evaluation period: 1983.8–2003.7

	Factor model Recursive PMSE relative to benchmark model					
	1-month horizon			12-month horizon		
Rank	UR^F	PT^F	BA^F	UR^F	PT^F	BA^F
1	.971	.967	.975	.636	.668	.632
2	.988	.988	.970	.687	.668	.683
3	1.016	1.000	.975	.708	.634	.659
4	.953	.952	.925	.688	.645	.626
5	.961	.963	.929	.698	.686	.640
6	.961	.960	.934	.702	.697	.659
7	.960	.957	.929	.729	.710	.707
8	.919	.916	.891	.699	.677	.678
30	.923	.905	.823	.703	.708	.755

Nevertheless, Table 2 shows that for $h = 1$, the UR^F forecast can usually be improved on. With the exception of $r = 1$, the BA^F predictor is always more accurate than the UR^F and PT^F predictors at the 1-month horizon. For $r = 1$, the BA^F predictor is marginally less accurate than the UR^F predictor, whereas the pretest marginally improves on the standard factor model forecast. Overall, the gains from bagging for common choices of r are fairly systematic at $h = 1$, albeit rarely larger than 3 percentage points. For $h = 12$, the results are more mixed. Although the BA^F predictor improves on the UR^F predictor for all $r \leq 8$ and the gains can be as large as 6 percentage points, in some cases the PT^F predictor is even more accurate. Nevertheless, the results in Table 2 underscore the potential of bagging to improve the accuracy of factor model forecasts.

Whereas there is clear evidence of gains in forecast accuracy relative to the UR^F predictor at both horizons, Table 2 also shows that for commonly used values of r, the accuracy of the BA^F predictor, although roughly similar to that of the BA and CBA predictors at $h = 12$, is far inferior at $h = 1$. It may seem that this result is simply an artifact of the smaller information set used in constructing the BA^F predictor and could be overturned by allowing the information set to expand. But this is not the case. When all 30 principal components are included, as in the last row of Table 2, the UR^F model reduces to the UR model. In the latter case, the BA^F predictor achieves large in gains in accuracy at $h = 1$ with a PMSE reduction of 18% relative to the benchmark model. This gain in accuracy is similar in magnitude to that for the other bagging predictors in Table 1. But for $h = 12$, the BA^F predictor is not only less accurate than the UR^F predictor, but, more importantly, also it far less accurate than the BA and CBA predictors in Table 1. This evidence suggests that the BA^F predictor is not as robust as the BA and CBA predictors.

3.3 Forecasts Based on Shrinkage Estimation of the Correlated Regressor Model

Tables 1 and 2 provide compelling evidence that bagging predictors tend to improve forecast accuracy relative to unrestricted and pretest predictors. These are not the only relevant alternatives, however. Because the bagging method involves features reminiscent of shrinkage estimation, it is only natural

Table 3. U.S. consumer price inflation forecasts, evaluation period: 1983.8-92003.7

	Shrinkage methods Recursive PMSE relative to benchmark model		
Horizon	Bayesian shrinkage	Ridge regression	LASSO
1 month	.817	.807	.825
12 months	.632	.623	.596

to compare the accuracy of bagging predictors to that of predictors based on alternative shrinkage methods. Here we consider three versions of shrinkage estimators that have been used by practitioners: Bayesian shrinkage estimation with a Gaussian prior centered on zero, ridge regression, and the LASSO. The results for these three methods are summarized in Table 3.

3.3.1 Bayesian Shrinkage Estimation. Bayesian shrinkage estimation has a long tradition in econometric forecasting (see, e.g., Litterman 1986). 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. The Bayesian shrinkage estimator is applied to the model,

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

We postulate a diffuse Gaussian prior for $(\nu, \phi_1, \ldots, \phi_p)$. The prior mean is based on the fitted values of a regression of inflation on lagged inflation and the intercept over the presample period, as proposed by Wright (2003b). In our case, the presample 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 $\lambda \in \{.01, .05, .1, .2, .3, .4, .5, 1, 2, 5, 100\}$ for the standardized data. For $\lambda = \infty$, the shrinkage estimator reduces to the OLS estimator of the unrestricted model. All prior covariances are set to 0. Further details on the implementation of this estimator have been provided by Lütkepohl (1993, chap. 5.4). We only report results for the value of λ that generated the lowest recursive PMSE.

Table 3 shows that this Bayesian shrinkage predictor is about as accurate as the BA and CBA methods. At $h = 1$, it lowers the PMSE by 18% relative to the benchmark model; at $h = 12$, the gains are 37%.

3.3.2 Ridge Regression. The Bayesian shrinkage estimator may be interpreted as a classical ridge regression estimator. For comparison, we also include the standard ridge regression estimator, defined by

$$(\widehat{\nu}^R, \widehat{\phi}^R, \widehat{\beta}^R) = \arg\min \left\{ \sum_{t=1}^{T} \left(\pi^h_{t+h|t} - \nu - \sum_{k=1}^{p} \phi_k \pi_{t-k+1} - \sum_{j=1}^{M} \beta_j x_{j,t} \right)^2 + \lambda \sum_{j=1}^{M} \beta_j^2 \right\},$$

for $0 < \lambda < \infty$. This problem can be solved by standard numerical optimization methods. Because the estimator depends on λ, we considered a grid of values $\lambda \in \{.5, 1, 2, 3, 4, 5, 10, 20, 50, 100, 150, 200\}$. For $\lambda = 0$, the constraint is not binding, and the ridge regression estimator reduces to OLS. Given this estimate, the ridge regression forecasts are generated from

$$\pi^h_{t+h|t} = \widehat{\nu}^R + \sum_{k=1}^{p} \widehat{\phi}^R_k \pi_{t-k+1} + \sum_{j=1}^{M} \widehat{\beta}^R_j x_{j,t}.$$

Only the result with the lowest recursive PMSE is reported in Table 3. Table 3 shows that the ridge regression predictor is marginally more accurate than the bagging predictor and the Bayesian shrinkage predictor. It achieves PMSE reductions of 19% relative to the benchmark model at $h = 1$ and 38% at $h = 12$.

3.3.3 LASSO. One potential disadvantage of ridge regression is that parameter estimates may be shrunk toward 0, but never are 0 exactly. This feature is an artifact of the imposition of a quadratic loss function in ridge estimation and may undermine forecast accuracy. To allow some parameters to be exactly 0 requires the use of an absolute loss function, as in the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996). The LASSO is designed to shrink some regression coefficients toward 0, while setting others equal to 0. Thus it combines features of model selection with features of shrinkage estimation. The LASSO estimator is defined by

$$(\widehat{\nu}^L, \widehat{\phi}^L, \widehat{\beta}^L) = \arg\min \left\{ \sum_{t=1}^{T} \left(\pi^h_{t+h|t} - \nu - \sum_{k=1}^{p} \phi_k \pi_{t-k+1} - \sum_{j=1}^{M} \beta_j x_{j,t} \right)^2 + \lambda \sum_{j=1}^{M} |\beta_j| \right\}$$

or, equivalently,

$$(\widehat{\nu}^L, \widehat{\phi}^L, \widehat{\beta}^L) = \arg\min \left\{ \sum_{t=1}^{T} \left(\pi^h_{t+h|t} - \nu - \sum_{k=1}^{p} \phi_k \pi_{t-k+1} - \sum_{j=1}^{M} \beta_j x_{j,t} \right)^2 \right\}$$

$$\text{subject to} \quad \sum_{j=1}^{M} |\beta_j| \le \tau.$$

In practice, we solve the latter problem by an iterative algorithm. First, we compute the LASSO estimator of β conditional on the OLS estimates of ν and ϕ in the UR model using the algorithm for models without predetermined regressors described by Tibshirani (1996, p. 278). Then we solve for the estimates of ν and ϕ conditional on the first-round LASSO estimate for β, and recompute $\widehat{\beta}^L$ given the revised estimates of ν and ϕ. This process is iterated until convergence. Given the final estimate, the LASSO forecasts are generated from

$$\pi^h_{t+h|t} = \widehat{\nu}^L + \sum_{k=1}^{p} \widehat{\phi}^L_k \pi_{t-k+1} + \sum_{j=1}^{M} \widehat{\beta}^L_j x_{j,t}.$$

Because the estimator depends on τ, we considered a grid of values $\tau \in \{1, 2, 3, 4, 5, 10, 20, 50, 100\}$. For sufficiently large τ, the constraint is not binding, and the LASSO estimator reduces to OLS. Only the result with the lowest recursive PMSE is reported in Table 3.

Table 3 shows that the iterated LASSO predictor is marginally less accurate than the BA and CBA methods at $h = 1$, with PMSE gains of 18% relative to the benchmark model, and somewhat more accurate at $h = 12$, with gains of 40% relative to the benchmark model. Interestingly, the LASSO predictor does not dominate the computationally less demanding ridge regression predictor.

3.4 Forecast Combinations

Although one should not read too much into the (minor) differences in forecast accuracy between alternative shrinkage methods, the results in Table 3 underscore that some form of shrinkage, be it in the form of bagging or another method, is highly advantageous in this application. Yet another branch of the economic forecasting literature has explored the use of forecast combinations as an alternative to shrinkage estimation. In this section we consider several variations of this method, starting with the least sophisticated methods. The results are summarized in Tables 4 and 5.

3.4.1 Forecast Combination: One Extra Predictor at a Time. Equally Weighted Forecast Combinations. 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 PMSEs.

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. Each forecast receives weight $1/M$. Effectively this amounts to averaging across M forecasts. The idea of using averages of forecasts to improve forecast accuracy dates back to work of Bates and Granger (1969). Equally weighted forecast combinations, despite their simplicity, have been found to produce highly accurate forecasts in a wide range of applications. Table 4 shows that this forecasting approach in our application does not work well. Although equally weighted forecasts beat the benchmark model, the PMSE reductions are limited to 3% at $h = 1$ and to 15% at $h = 12$.

One potential problem with using the arithmetic mean of forecasts is its sensitivity to outliers. For that reason, some researchers prefer to use the median forecast or the trimmed mean

Table 4. U.S. consumer price inflation forecasts, evaluation period: 1983.8–2003.7

	Forecast combination methods, one extra predictor at a time Recursive PMSE relative to benchmark model				
Horizon	Equal-weighted	Median	ARM	AFTER	BMA Wright
1 month	.968	.990	.933	1.018	.904
12 months	.846	.941	.681	.724	.686

Table 5. U.S. consumer price inflation forecasts, evaluation period: 1983.8–2003.7

	Forecast combination methods, random subsets of predictors Recursive PMSE relative to benchmark model		
Horizon	ARM	AFTER	BMA Raftery et al.
1 month	.815	.882	.821
12 months	.690	.620	.618

of the M forecasts instead. Table 4 shows that this modification does little to improve forecast accuracy. In fact, the median forecast is distinctly less accurate than the equally weighted forecast, with gains relative to the benchmark model of only 1% at $h = 1$ and 6% at $h = 12$. Similar results are obtained for the trimmed mean.

ARM Forecasts. A possible explanation for the relatively poor performance of mean forecasts is the assumption of equal weights. In principle, one should be able to improve accuracy of forecast combination methods by letting the data choose the forecast weights. One such method is *adaptive regression by mixing* (ARM), as proposed by Yang (2001, 2003). Two steps are involved. In the first step, the first part of the sample available to the forecaster at a given point in time is used to estimate the forecasting models $i = 1, \dots, M$, where each model includes one of the M extra predictors of interest. In the second step, each of these fitted models is used to generate predictions of the realizations of inflation in the remainder of the sample. Appropriate weights for the M forecasting models are constructed based on their predictive success in the second part of the sample. The lower the PMSE of model i in the second part of the sample, the higher its weight in the forecasting model. (For a detailed description of this algorithm see Yang 2003, p. 786.) Rather than splitting the sample in half, as described by Yang (2003), we conducted a grid search allowing between 30% and 90% of the sample to be allocated for initial estimation in increments of 10%. The results reported herein are based on the most favorable results. Table 4 shows that ARM works moderately well in our context. The PMSE gains at $h = 1$ are 7%; at $h = 12$ they rise to 32%, making ARM superior to the median forecast and the equally weighted combination forecast.

AFTER Forecasts. Because the ARM method involves splitting the sample, and both the estimation and the evaluation half of the sample tend to be short in our application, unreliable estimates of the model parameters and noisy measures of predictive performance may limit the success of this method. This suggests that we consider alternative methods that retain the idea of weighting forecasting models by some measure of predictive performance without relying on sample splitting. One such method is aggregation of forecasts through exponential reweighting (AFTER), as proposed by Yang (2004, pp. 186–188). Another such method, Bayesian model averaging of the M candidate models as applied by Wright (2003a,b), is considered later.

The AFTER method works as follows. The $i = 1, \dots, M$ forecasting models are estimated on the full sample available to the forecaster. The first recursive forecast is obtained by assigning equal weights to all forecasting models. Subsequent recursive forecasts are based on recursively updated weights that reflect the predictive accuracy of each candidate model over the length of the recursive forecasting exercise up to that point in time. The weight function used in implementing this algorithm is given in eq. (4) of Yang (2004). The conditional variance estimates used in (4) are obtained from eq. (5) of Yang (2004). Table 4 shows that the AFTER algorithm does not work well at the 1-month horizon. In fact, it is the only competitor to perform worse than the inflation-only benchmark model. At the 12-month horizon, there is significant improvement, with gains of 28%, but even in that case AFTER remains inferior to ARM. This result illustrates that methods that do not involve sample splitting are not necessarily superior to methods that do.

Forecasts Based on Bayesian Model Averaging. In related work, Wright (2003b) has shown that the accuracy of equally weighted forecast combination methods may be improved on further by weighting the individual forecasting models based on the posterior probabilities associated with each forecasting model, which makes Wright's BMA method a natural competitor to consider. But whereas Wright imposes one lag of inflation in the benchmark model, we allow for potentially more than one lag of inflation in implementing his procedure.

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 presample period. For the remaining parameters, we postulate a Gaussian prior with mean 0 and a prior standard deviation of $\phi \in \{0, .01, .05, .1, .2, .3, .4, .5, 1, 2, 5, 100\}$ for the standardized data. Again the prior treats the predictors as independent. The prior probability for each forecasting model is $1/M$, as in the equally weighted forecast combination. For $\phi = 0$, the BMA method of forecast combination reduces to the equally weighted method. Table 4 presents the outcome of a grid search over ϕ.

We find that, as reported by Wright (2003b), the BMA method is clearly superior to the equally weighted forecast combination method. It also is superior to the median forecast and the AFTER forecast, and about equal to the ARM forecast. Nevertheless, with a ratio of 90% at $h = 1$ and 69% at $h = 12$, this BMA method in our application is less accurate than the BA and CBA methods or for that matter the other shrinkage methods.

3.4.2 Forecast Combination: Randomly Chosen Subsets of Extra Predictors.

Forecasts Based on Bayesian Model Averaging. As the previous discussion illustrated, articles on forecast combination methods for inflation typically restrict the forecasting models under consideration to include only one indicator of real economic activity at a time. There is no reason for this approach to generate the forecast with the lowest possible PMSE regardless of whether we use equal weights or data-based weights, because the search is restricted to a subset of the possible combinations of the extra predictors. A complete Bayesian solution to this problem would involve averaging over all possible forecasting 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 $2^{30} = 1{,}073{,}741{,}824$ possible combinations of predictors to consider. 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. (Also see George and McCulloch 1993 for an alternative stochastic search variable selection algorithm.)

The random selection is based on a Markov chain Monte Carlo (MCMC) algorithm that moves through the forecasting model space. Unlike Wright's method, this algorithm involves simulation of the posterior distribution and is quite computationally demanding. Our results are based on 5,000 draws from the posterior distribution at each point in time.

MATLAB code for the algorithm of Raftery et al. is publicly available at *http://www.spatial-econometrics.com*. We modified Raftery et al.'s approach to ensure that the benchmark model includes 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 method of Wright (2003b). For the remaining parameters of the forecast prior the algorithm involves a Gaussian prior with mean 0 and hyperparameters $\nu = 2.58$, $\lambda = .28$, and $\phi \in \{0, .01, .05, .1, .2, .3, .4, .5, 1, 2, 5, 100\}$ in the notation used by Raftery et al., where ϕ denotes the prior standard deviation of the standardized predictor data (see Raftery et al. 1997 for further details). We report the best empirical results in the last column of Table 5. 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 those for our preferred values of ϕ.

This version of BMA clearly produces more accurate results than the restricted version involving only one extra predictor at a time. Compared with BMA based on one extra predictor at a time, at the 1-month horizon, the PMSE ratio for the best BMA predictor falls from 90% to 82% and at the 1-year horizon from 69% to 62%. Thus the forecast accuracy of the method of Raftery et al. is comparable to that of bagging methods and other shrinkage methods.

ARM Forecasts. It is natural to broaden the set of candidate models underlying the ARM method along similar lines. Therefore, we also applied the ARM method based on 5,000 randomly chosen subsets of the predictors. Table 5 shows that this greatly improves the accuracy of the 1-month-ahead forecasts, but not of the 1-year-ahead forecasts. In the former case, ARM is about as accurate as Raftery et al.'s BMA method or the BA method, whereas in the latter case it is noticeably less accurate.

AFTER Forecasts. The same modification also may be applied to the AFTER method. Table 5 shows that the accuracy of the AFTER forecasts improves at both horizons, but remains below that of Raftery et al.'s method at the 1-month horizon. Whereas the 1-year forecast is comparable to that of Raftery et al.'s method and even slightly better than BA and CBA, the 1-month-ahead forecast has a noticeably larger recursive PMSE than many of the leading competitors.

4. CONCLUSION

In this article we considered the widely studied question of whether the inclusion of indicators of real economic activity lowers the PMSE of forecasting models of U.S. consumer price inflation. We proposed three new variants of the bagging algorithm specifically designed for this type of forecasting problem. Over a 20-year period, we compared the accuracy of simulated out-of-sample forecasts based on these bagging methods to that of alternative forecasting methods for U.S. consumer price 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 pretests, forecasts from estimated factor models, forecasts from models estimated by various shrinkage estimators, unweighted combination forecasts, median and trimmed mean forecasts, ARM forecasts, AFTER forecasts, and finally forecasts obtained by Bayesian model averaging.

Consumer price inflation is not only one of the key macroeconomic variables of interest to businesses and policymakers, but it is also more difficult to predict than real economic growth, for example. This is especially true for the period since the 1980s, on which we focus, a period during which evidence that real economic indicators help predict inflation has become weaker. The empirical evidence presented in this article shows that bagging can achieve large reductions in PMSEs, even in challenging applications such as inflation forecasting. The gains in accuracy compare favorably with the benchmark model and with results reported in previous studies. At the 1-month horizon, regression-based bagging methods achieve gains relative to the inflation-only model of 16–18%. At the 1-year horizon, the gains increase to 35–36%.

But bagging is not the only method capable of achieving such gains. We found that the Bayesian shrinkage predictor, the ridge regression predictor, the iterated LASSO predictor, and the Bayesian model average predictor based on random subsets of extra predictors are about as accurate as the bagging predictor. ARM and AFTER predictors based on random subsets of extra predictors also performed well in some cases, but their performance was more uneven across horizons. The high accuracy of the ridge regression predictor in particular suggests that similar accuracy is feasible at much lower computational cost than required for bagging or for Bayesian model averaging based on random subsets of predictors. But recently proposed asymptotic approximations to bagging methods for orthogonalized predictors such as CBA and BA^F may eliminate the need for computer simulation in bagging (see Stock and Watson 2006). It will be of interest to see how these asymptotic approximations compare with the full-fledged bagging approach used in this article.

The fact that several methods are about equally accurate in this application does not mean that the choice of forecasting method does not matter. We showed that other methods including equally weighted forecasts, median forecasts, ARM forecasts, AFTER forecasts, and Bayesian forecast averages based on one extra predictor at a time do not perform well in this application. We observed that forecast combination methods based on one extra predictor at a time may unduly constrain the predictor set and demonstrated that relaxing this constraint may greatly improve forecast accuracy.

A final question of interest is which of the three variants of bagging is preferred. We showed that there is no clear ranking between bagging methods based on untransformed regressors and bagging methods based on orthogonalized regressors. Both perform well. Although bagging methods based on factor models also may perform well in some cases, their accuracy is sensitive to the choice of rank. In our application, the BA^F predictor was not as robust as the BA and CBA predictors. Nevertheless, we found that compared to the unrestricted factor model of rank r, the BA^F predictor tended to improve forecast accuracy for common choices of r.

Because in our application the set of potential predictors is only moderately large, the principal components may be estimated too imprecisely for the factor model forecast or its bagging equivalent to work well. An interesting avenue for future research will be the use of factor bagging methods on panels with larger cross-sections. Some preliminary evidence has been provided by Stock and Watson (2006) and Edelstein (2006).

APPENDIX: DATA SOURCES

These indicators of real economic activity were obtained from *http://www.economagic.com*:

INDPRO	Industrial production
HOUST	Housing starts
HSN1F	House sales
ISM	ISM index of manufacturing activity
HELPWANT	Help wanted index
TCU	Capacity utilization
UNRATE	Unemployment rate
PAYEMS	Nonfarm payroll employment
CIVPART	Civilian participation rate
AWHI	Aggregate weekly hours, private nonfarm payrolls
MORTG	Mortgage rate
MPRIME	Prime rate
CD1M	1-month CD rate
FEDFUND	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
TOTASS_AUSA	Total number of motor vehicle assemblies
TCM20Y–TBSM3M	Spread of 10-year T-bond rate over 3-month T-bill rate
UEMP15OV	Number of civilians unemployed for more than 15 weeks
UEMPLT5	Number of civilians unemployed for less than 5 weeks

[Received March 2005. Revised January 2007.]

REFERENCES

Avramov, D. (2002), "Stock Return Predictability and Model Uncertainty," *Journal of Financial Economics*, 64, 423–458.

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

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

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

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

Cecchetti, S., Chu, R., and Steindel, C. (2000), "The Unreliability of Inflation Indicators," *Federal Reserve Bank of New York Current Issues in Economics and Finance*, 6, 1–6.

Cremers, K. J. M. (2002), "Stock Return Predictability: A Bayesian Model Selection Perspective," *Review of Financial Studies*, 15, 1223–1249.

Edelstein, P. (2006), "Commodity Prices, Inflation Forecasts, and Monetary Policy," mimeo, University of Michigan, Dept. of Economics.

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

——— (2003), "Do Financial Variables Help Forecasting Inflation and Real Activity in the Euro Area," *Journal of Monetary Economics*, 50, 1243–1255.

——— (2005), "The Generalized Factor Model: One-Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100, 830–840.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–890.

Gonçalves, S., and Kilian, L. (2004), "Bootstrapping Autoregressions With Conditional Heteroskedasticity of Unknown Form," *Journal of Econometrics*, 123, 89–120.

Gonçalves, S., and White, H. (2004), "Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models," *Journal of Econometrics*, 119, 199–220.

Hall, P., and Horowitz, J. L. (1996), "Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators," *Econometrica*, 64, 891–916.

Inoue, A., and Kilian, L. (2006), "On the Selection of Forecasting Models," *Journal of Econometrics*, 130, 273–306.

Inoue, A., and Shintani, M. (2006), "Bootstrapping GMM Estimators for Time Series," *Journal of Econometrics*, 133, 531–555.

Koop, G., and Potter, S. (2004), "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *Econometrics Journal*, 7, 550–565.

Lee, T.-H., and Yang, Y. (2006), "Bagging Binary and Quantile Predictors for Time Series," *Journal of Econometrics*, 135, 465–497.

Litterman, R. B. (1986), "Forecasting With Bayesian Vector Autoregressions: Five Years of Experience," *Journal of Business & Economic Statistics*, 4, 25–38.

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

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

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

Min, C., and Zellner, A. (1993), "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts With Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56, 89–118.

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

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

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

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

——— (2002a), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

——— (2002b), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162.

——— (2003), "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature*, 41, 788–829.

——— (2006), "Shrinkage Methods for Forecasting Using Many Predictors," mimeo, Harvard University, Dept. of Economics.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

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

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

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

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

Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96, 574–809.

——— (2003), "Regression With Multiple Candidate Models: Selecting or Mixing?" *Statistica Sinica*, 13, 783–809.

——— (2004), "Combining Forecasting Procedures: Some Theoretical Results," *Econometric Theory*, 20, 176–222.