Forecasting Macroeconomic Variables Using Linear and Nonlinear Models

Hyun Hak Kim and Norman R. Swanson
Rutgers University

January 2010

Abstract

Numerous studies have recently been undertaken in the empirical forecasting literature in order to assess the usefulness of recent diffusion index methodologies introduced by Bai (2003), Bai and Ng (2002, 2005, 2006a,b,c,d), Forni, Hallin, Lippi, and Reichlin (2000, 2005), Forni and Reichlin (1996, 1998), Stock and Watson (1996, 1998, 1999, 2002a,b,2004a,b, 2005) and others. In particular, many studies have endeavored to ascertain whether the principle components constructed using diffusion index methods contain marginal predictive content, over and above that already contained in observable economic and financial time series. In this paper, we add to the extant literature on this topic by assessing the predictive accuracy of a large group of nonlinear models, all of which are constructed using both principle components as well as using observable time series. We design a "horse-race" in which mean-square-forecast-error "best" models are selected, in the context of a variety of model specification methods, forecast horizons, sample periods, and "target variables" to be predicted. In addition to pure common factor prediction models, the forecast model specification methods that we analyze include bagging, boosting, Bayesian model averaging, ridge regression, least angle regression, elastic net and non-negative garotte as well as univariate autoregressive and autoregressive-exogenous model as benchmarks. One aspect of this paper is that we provide a survey of the above methods. For a number of target variables, we find that various of these models perform better than benchmark linear autoregressive forecasting models constructed using only observable variables, hence suggesting that the diffusion index methodology based models offer a convenient way to filter large-scale economic datasets prior to their use in forecast model construction.

Keywords: prediction, bagging, boosting, Bayesian model averaging, ridge regression, least angle regression, elastic net and non-negative garotte.
JEL Classification: G1.

---

* Hyun Hak Kim, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, hykim@econ.rutgers.edu. Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, nswanson@econ.rutgers.edu. The authors also owe many thanks to seminar participants at Rutgers University, and to Nii Armah, Gary Koop, John Landon-Lane and Hiroki Tsurumi for numerous useful suggestions on earlier version of this paper. The authors gratefully acknowledge financial support from a Rutgers University Research Council grant.

1 Introduction

In macroeconometrics and financial economic forecasting, the number of variables (N) is occasionally greater than the number of observations (T) used for model estimation. Moreover, even when this is not the case, there are still often hundreds of explanatory variables available for use in prediction model construction, many of which are highly collinear. This has led to the development of many data shrinkage techniques, and one of the most widely applied of these is diffusion index methodology. One reason for the importance of this methodology is the recent explosion in data availability, which has made it increasingly common to have larger numbers of variables available than observational time periods, hence precluding the use of many standard modelling approaches when constructing and selecting amongst prediction models. Another reason is that diffusion index methodology offers a simple and sensible approach for extracting common factors that underlie the dynamic evolution of large numbers of variables. To be more specific, let y_t be a time series vector of dimension $(T \times 1)$ and let X_t be a time-series predictor matrix of dimension $(T \times N)$ .Then, define the following dynamic factor model, where F_t denotes r unobserved common factors that can be extracted from X_t. Namely,

$$X_t = \Lambda F_t + e_t, \tag{1}$$

where e_t is an $N \times 1$ vector of disturbances and Λ is a $T \times r$ coefficient vector. Using common factors extracted from the above model, we consider linear forecasting models of the form:

$$Y_{t+h} = \beta'_F F_t + \beta_W W_t + \varepsilon_{t+h}, \tag{2}$$

where h is the forecast horizon and W_t is a $p \times 1$ vector of observed variables, including lags of Y_t,, etc. Therefore, we first estimate the unobserved factors, F_t, and then forecast Y_{t+h} with observed variables and $\hat{F}_t$, where $\hat{F}_t$ is an estimator of F_t. Even though factor model are now widely used, many issues remain outstanding, such a the determination of number of factors to be implemented in subsequent prediction model construction. The papers cited above discuss these any many related issues.

In this paper we investigates several methods for forecasting using many predictors, and compare their predictive accuracy using mean-square-forecast-error (MSFE) loss. The variables that we predict include a variety of macroeconomic variables that are useful for evaluating economic policy actions taken by federal policy setting authorities. Our impetus for choice of these variables stems from the following statement made in Federal

Reserve Bank of New York's website: "*In formulating the nation's monetary policy, the Federal Reserve considers a number of factors, including the economic and financial indicators which follow, as well as the anecdotal reports compiled in the Beige Book. Real Gross Domestic Product (GDP); Consumer Price Index (CPI); Nonfarm Payroll Employment Housing Starts; Industrial Production/Capacity Utilization; Retail Sales; Business Sales and Inventories; Advance Durable Goods Shipments, New Orders and Unfilled Orders; Lightweight Vehicle Sales; Yield on 10-year Treasury Bond; S&P 500 Stock Index; M2*" (see http://www.newyorkfed.org/education/bythe.html). Our target predictor variables are largely taken from the above set of variables (see Armah and Swanson (2008) for further discussion). More specifically, forecasts are constructed for fourteen series, including: the unemployment rate, personal income less transfer payments, the 10 year Treasury-bond yield, the consumer price index, the producer price index, non-farm payroll employment, housing starts, industrial production, M2, the S&P 500 index, gross domestic product,[1] retail sales, business sales and inventory and advanced durable goods shipment, and new orders and unfilled orders.

Recently, Stock and Watson (2005a) surveyed several methods for shrinkage that are based on factor augmented autoregression models. We add to their discussion by additionally considering several other models in our "horse-race". In particular, and as mentioned above, the models that we analyze include, in addition to pure common factor prediction models, bagging, boosting, Bayesian model averaging, ridge regression, least angle regression, elastic net and non-negative garotte. Additionally, we specify various linear benchmark models, including univariate and multivariate autoregressive models, and numerous forecasting averaging and combination models as benchmarks. In order to assess forecasting accuracy, we construct Diebold-Mariano (1995) predictive accuracy tests.

For a number of our target variables, we find that various of these models, and in particular component-wise boosting, have lower MSFEs than benchmark linear autoregressive forecasting models constructed using only observable variables, hence suggesting that models that incorporate common factors constructed using diffusion index methodology offer a convenient way to filter the information contained in large-scale economic datasets. We also find that forecasts constructed as a simple average of all individual model based forecasts yield the MSFE-best model for various variables and forecast horizons. This result is not surprising, given the large body of research establishing the usefulness of such forecasts in empirical

[1] Gross domestic product is usuall observed quaterly. We interpolate these data to a monthly frequency following Chow and Lin (1971),

settings. However, we do find that there are many variable - forecast horizon combinations for which neither the linear not the mean-forecast models are MSFE-best. In addition, we find that there is little to choose between using rolling versus recursive estimation strategies when constructing real-time prediction models, although models constructed using recursive strategies appear to perform better more frequently, when comparing point MSFEs. Finally, we find that there are many instances where our more sophisticated forecasting models yield predictions that are significantly more accurate than those of benchmark linear models, based on examination of Diebold and Mariano (1995) predictive accuracy test results, such as when forecasting unemployment, money growth and GDP growth.

The results presented in this paper are meant to add not only to the diffusion index literature, but also to the extraordinary collection of papers on forecasting that Clive W.J. Granger wrote between 1958 and 2009. Indeed, Clive W.J. Granger is in many respects that father of time series forecasting in the field of economics, and we salute his innumerable contributions in areas from predictive accuracy testing, model selection analysis, and forecast combination to forecast loss function analysis, forecasting using nonstationary data, and nonlinear forecasting model specification. We will sorely miss him.

The rest of the paper is organized as follows. In the next section we provide a brief overview of some of the techniques used in this paper. The following section (Section 3) provides a more formal introduction to dynamic factor models. Thereafter, we outline the forecasting models that we implement. Estimation and data issues are discussed in Section 5, and empirical results are presented in Section 6. Concluding remarks are given in Section 7.

2 Background

Recent forecasting studies using large-scale datasets and pseudo out-of-sample forecasting include: Artis et al. (2002), Boivin and Ng (2005, 2006), Forni et al. (2005), Stock and Watson (1999, 2002, 2003, 2005a,b, 2006). Stock and Watson (2006) discuss in some detail the literature on the use of diffusion indices for forecasting.

In this paper, we consider a variety of "robust" estimation techniques, including bagging, boosting, ridge regression, least angle regression, elastic net, non-negative garotte and Bayesian model averaging, all in the context of diffusion index models. In the following discussion, we briefly summarize some of the key literature on these methods.

Bagging, which was introduced by Breiman (1996), is a machine based learning algo-

rithm whereby outputs from different predictors are combined in order to improve overall forecasting accuracy. Bühlmann and Yu (2002) use bagging in order to improve forecast accuracy when data are i.i.d.. Inoue and Kilian (2005) and Stock and Watson (2005a) extend bagging to time series models. Stock and Watson (2005a) considers "bagging" as a form of shrinkage, when constructing prediction models. In this paper, we follow the same approach. This allows us to avoid time intensive bootstrap computation done elsewhere in the bagging literature. Boosting, a close relative of bagging, is another statistical learning algorithm, and was originally designed for classification problems in the context of Probability Approximate Correct (PAC) learning (see Schapire (1990)). The method was implemented in Freund and Schapire (1997) (using the algorithm called 'AdaBoost.M1'). Hastie et al. (2001) argue that "boosting" is one of the most powerful learning algorithms currently available, and the method has been extended to regression problems in Ridgeway et al. (1999) and Shrestha and Solomatine (2006). In the economics literature, Bai and Ng (2008a) use a boosting algorithm for selecting the predictors in factor augmented autoregressions. We follow the approach of Bai and Ng (2008a) when implementing boosting.

The "least absolute shrinkage and selection operator" (i.e. the so-called 'Lasso') was introduced by Tibshirani (1996), and is another attractive technique for variable selection using high-dimensional datasets, especially when the number of variables (N) is greater than the number of observations (T). One version of the Lasso, "Least Angle Regression" (i.e. the so-called LAR), is introduced in Efron et al. (2004), and is a method for choosing a linear model using the same set of data as that used to evaluate and implement the model. LAR is based on well known model-selection methods known as "forward-selection", which have been extensivey used to examine cross-sectional data. (For further details, see page 408 in Efron et al. (2004)). Bai and Ng (2008b) show how to apply LARs and Lassos using time series models, and Gelper and Croux (2008) extend Bai and Ng (2008b)'s work to time series forecasting with many predictors. A related method is the so-called "Elastic net", proposed by Zou and Hastie (2005), which is similar to the Lasso, as it simultaneously carries out automatic variable selection and continuous shrinkage. Its name comes from the notion that it is similar in structure to a stretchable fishing net that retains 'all the big fish'. LARs-EN is proposed by Zou and Hastie (2005) for computing entire elastic net regularization paths using only a single least squares model, for the case where the number of variables is greater than the number of observations. Bai and Ng (2008b) apply the elastic net method to time series contexts. We follow Gelper and Croux (2008)'s algorithm for constructing the LARs estimator, and Zou and Hastie (2005) for constructing the LARs-

Elastic net estimator. Another method that we consider is the so-called, "non-negative garotte", originally introduced by Breiman (1995). This method is a scaled version of the least square estimator with shrinkage factors. Yuan and Lin (2007) develop an efficient garrotte algorithm and prove consistency in variable selection. As far as we know, this method has hithertofore not been used in the econometrics literature. We follow Yuan and Lin (2007) and apply it to time series forecasting. Yet another method that we consider is ridge regression, which is a well known linear regerssion shrinkage method. It is essentially a classical linear regression technique which modifies sum of square residual computations to include a penalty for inclusion of larger numbers of parameters. Finally, we consider Bayesian model averaging (henceforth, BMA), as it is one of the most attractive methods of model selection currently available (see Fernandez et al. (2001b) and Koop and Potter (2004), Ravazzolo et al. (2008)). The concept of Bayesian model averaging can be described with simple probability rules. If we consider R different models, each model has a parameter vector and is represented by its prior probability, likelihood function and posterior probability. Given this information, using Bayesian inference, we can obtain model averaging weights based on the posterior probabilities of the alternative models. Koop and Potter (2004) consider BMA in the context of many predictors and evaluate its performance. We follow their approach.

Technical details about the above methods are described in Section 4. First, however, we describe the factor models that form the basis of our empirical analysis.

3 Factor Models

In the following discussion of diffusion index methodology, we follow Stock and Watson (2002).

3.1 Basic Framework

Let X_{it} be the observed data for the ith cross-section unit at time t, for $i = 1, ..., N$ and $t = 1, .., T$. Consider the following model:

$$X_{it} = \lambda_i' F_t + e_{it}, \tag{3}$$

where F_t is a vector of common factors, λ_i is a vector of factor loadings associated with F_t, and e_{it} is the idiosyncratic component of X_{it}. The product $\lambda_i F_t$ is called the common component of X_{it}. This is the dimension reducting factor representation of the data. Many economic analyses fit naturally into the above framework. For example, Stock and Watson

(1999) consider inflation forecasting with diffusion indices constructed from a large number of macroeconomic variables. For example, consider the forecasting equation:

$$Y_{t+h} = \beta_F' F_t + \beta_W' W_t + \varepsilon_{t+h}, \tag{4}$$

for $i = 1, 2..., N$, where h is forecast horizon and W_t is a $p \times 1$ vector of observed variables, including, among others, lags of Y_t. Following Bai and Ng (2002), the whole panel of data $X = (X_1, ..., X_N)$ can be represented as (3). Connor and Korajczyk (1986, 1988, 1993) note that the factors can be consistently estimated by principal components as $N \to \infty$, even if e_{it} is weakly cross-sectionally correlated. Similarly, Forni et al. (2005) and Stock and Watson (2002) discuss consistent estimation of the factors when $N, T \to \infty$. In a predictive context, Ding and Hwang (1999) analyze the properties of forecasts constructed from principal components when N and T are large. They perform their analysis under the assumption that the error processes $\{e_{it}, \varepsilon_{t+h}\}$ are cross sectionally and serially i.i.d. We work with high-dimensional factor models that allow both N and T to tend to infinity, and in which e_{it} may be serially and cross-sectionally correlated so that the covariance matrix of $e_t = (e_{1t}, ..., e_{Nt})$ does not have to be a diagonal matrix. We will also assume $\{F_t\}$ and $\{e_{it}\}$ are two groups of mutually independent stochastic variables. Furthermore, it is well known that if $\Lambda = (\lambda_1, ..., \lambda_N)$ for $\Lambda F_t = \Lambda Q Q^{-1} F_t$, a normalization is needed in order to uniquely define the factors, where Q is a nonsingular matrix. Now, assuming that $(\Lambda'\Lambda/N) \to I_r$, we restrict Q to be orthonormal, for example. This assumption, together with others noted in Stock and Watson (2002) and Bai and Ng (2002), enables us to identify the factors up to a change of sign and consistently estimate them up to an orthonormal transformation.

Forecasts of Y_{T+h} based on (4) involve a two step procedure because both the regressors and coefficients in the forecasting equations are unknown. The data sample $\{X_t\}_{t=1}^T$ are first used to estimate the factors, $\{\tilde{F}_t\}_{t=1}^T$ by means of principal components. With the estimated factors in hand, we obtain the estimators $\hat{\beta}_F$ and $\hat{\beta}_W$ by regressing Y_{t+h} on the estimates of F_t and the observable variables in W_t. In this paper, we try different methods for estimating $\hat{\beta}_F$ and $\hat{\beta}_W$ and then compare predictability of resultant prediction models. Of note is that if $\sqrt{T}/N \to 0$, then the generated regressor problem does not arise, in the sense that least squares estimates of $\hat{\beta}_F$ and $\hat{\beta}_W$ are $\sqrt{T}$ consistent and asymptotically normal (see Bai and Ng (2008b)).

The problem of obtaining the necessary estimates in (3) would be simplified if we knew F. Then λ_i could be estimated via least squares by setting $\{X_{it}\}_{t=1}^T$ to be the dependent

variable and $\{F_t\}_{t=1}^{T}$ to be the explanatory variable. On the other hand, if Λ were known, F_t could be estimated by regressing $\{X_{it}\}_{i=1}^{N}$ on $\{\lambda_i\}_{i=1}^{N}$. Since the common factors are not observed, in the regression analysis of (3), we replace F_t by $\tilde{F}_t$, estimates that span the same space as F_t when $N, T \to \infty$. Estimation of these common factors from large panel data sets of macroeconomic variables can be carried out using principal component analysis. We refer the reader to Stock and Watson (2002, 2005a,b, 1999) and Bai and Ng (2002, 2008a,b) for a detailed explanation of this procedure, and to Connor and Korajczyk (1986, 1988, 1993), Forni et al. (2005) and Armah and Swanson (2008) for further detailed discussion of diffusion models, in general.

3.2 Factor Estimation

Factor analysis can be done via the use of simple principal component analysis as outlined in Johnson and Wichern (2002). Drawing from the discussion in this paper consider the following linear combinations:

$$P_i = \mathbf{a}_i'\mathbf{X} = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{iN}X_N, \tag{5}$$

and obtain

$$\begin{aligned} Var\left(P_i\right) &= \mathbf{a}_i'\Sigma\mathbf{a}_i \qquad i = 1, 2, ..., N & (6)\\ Cov(P_i, P_k) &= \mathbf{a}_i'\Sigma\mathbf{a}_j \qquad i \text{ and } j = 1, 2, ..., N, & (7) \end{aligned}$$

where Σ is the covariance matrix associated with the random vector $\mathbf{X} = [X_1, X_2, ..., X_N]$. The principal components are those uncorrelated linear combinations, $P_1, P_2, ...P_N$, whose variances in (6) are as large as possible, given unobservable a_i. The first principal component is the linear combination with maximum variance ; $Var\left(P_i\right) = \mathbf{a}_i'\Sigma\mathbf{a}_i$ and which clearly can be increased by multiplying any $\mathbf{a}_i$ by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length. Therefore, the $i-$th principal component can be defined as linear combination, $\mathbf{a}_i'\mathbf{X}$, that maximizes $Var\left(\mathbf{a}_i'\mathbf{X}\right)$, subject to $\mathbf{a}_i'\mathbf{a}_i = 1$ and $Cov\left(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_j'\mathbf{X}\right) = 0$ for $j < i$. Let Σ have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_N, e_N)$, where $\lambda_1 \succeq \lambda_2 \succeq \cdots \succeq \lambda_N \succeq 0$. Then, the i th principal component is given by :

$$P_i = e_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{iN}X_N. \tag{8}$$

With these choices,

$$Var(P_i) = e_i'\Sigma e_i = \lambda_i \qquad i = 1, 2, ..., N \tag{9}$$

$$Cov(P_i, P_j) = e_i'\Sigma e_k = 0 \qquad i \neq j \tag{10}$$

Then,

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{NN} = \sum_{i=1}^{N} Var\left(X_i\right) = \lambda_1 + \lambda_2 + \cdots + \lambda_N = \sum_{i=1}^{N} Var\left(P_i\right). \tag{11}$$

Thus, the total population variance is:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{NN} = \lambda_1 + \lambda_2 + \cdots + \lambda_N \tag{12}$$

and consequently, the proportion of total variance due to the ith principal component is

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_N} \qquad i = 1, 2, ..., N. \tag{13}$$

Sample principal components are, in general, not invariant with respect to changes in scale. Variables measured on different scales or on a common scale with widely differing ranges are often standardized. One often sets

$$Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{ii}}}, \tag{14}$$

where $E\left(Z_i\right) = 0$ and $Var\left(Z_i\right) = \rho_i$. The principal component of Z may be obtained from the eigenvectors of the correlation ρ_i of X_i. In practice, the principal component is specified in terms of the sample covariance matrix of the predictors, i.e. the X_t's, and the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_N, e_N)$, where $\lambda_1 \succeq \lambda_2 \succeq \cdots \succeq \lambda_N \succeq 0$.

More specifiacally, let $r < N$ be the arbitrary number of common factors, $\Lambda^r = (\lambda_1, ..., \lambda_r)$ be an $T \times r$ factor loading matrix, and F^r be an $r \times N$ factor matrix. One solution for (3) can be found by solving the following optimization problem:

$$V(r) = \min_{\Lambda^r, F^r} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(X_{it} - \Lambda_i^r F_t^r\right)^2 \tag{15}$$

$$\text{subject to } \Lambda^{r'}\Lambda^r/N = I_r \text{ or } F^{r'}F^r/T = I_r$$

This optimization problem is also equivalent to maximizing $tr\left(F^{k\prime}\left(XX'\right)F^k\right)$ and this problem can be solved by setting the solution, $\hat{F}^k$, to be the matrix of the r eigenvectors of $X'X$ that correspond to the r largest eigenvalues of $X'X$. $\hat{\Lambda}^r$ is constructed as $\sqrt{N} \times \nu_{(r)}$, where

$\nu_{(r)}$ are the eigenvectors corresponding to the r largest eigenvalues of $X'X$. Then normalization such that $\hat{\Lambda}^{r\prime}\hat{\Lambda}^r/N = I_r$ implies that $\hat{F}^r = X\hat{\Lambda}^r/N$. Moreover, a rescaled estimator of the factors is defined as $\bar{F}^r = \hat{F}^r\left(\hat{F}^{r\prime}\hat{F}^r/T\right)^{1/2}$, and its asymptotic properties are given in Theorem 1 of Bai and Ng (2002).

Before implementing principal component analysis in the context of factor models, there remains the question of how many components to retain. Bai and Ng (2002) provide one solution to the problem of how to choose the number of factors. They establish convergence rates for factor estimates under consistent estimation of the number of factors, r, and propose panel criteria consistently estimate the number of factors. Begin with an arbitrary number r ($< \min[N,T]$) and let λ_i^r and F_t^r be the r factors included in the estimation via solving (15). As with model selection criteria, the assumption that $r+1$ factors yield a model that is less efficient than when r factors are specified. Let F^r be a matrix of r factors and

$$V(k,F^r) = \min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\Lambda_i^{r\prime}F_t^r\right)^2 \tag{16}$$

be the sum of squared residuals from regression of X_i on the r factors for all i. Then without loss of generality, we can set

$$V\left(k,\hat{F}^r\right) = \min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\lambda_i^{r\prime}\hat{F}_t^r\right)^2 \tag{17}$$

Along these lines Bai and Ng (2002) define selection criteria of the form $PC(r) = V\left(k,\hat{F}^r\right) + kh(N,T)$, where $h(\cdot)$ is a penalty function. In this paper, the following is used (for discussion of its relatively superior performance, see Bai and Ng (2002),Armah and Swanson (2008)):

$$BIC_3(r) = V\left(k,\hat{F}^k\right) + k\hat{\sigma}^2\left(\frac{(N+T-k)\ln(NT)}{NT}\right). \tag{18}$$

Our consistent estimate of the true number of factors is thus:

$$\hat{r} = \arg\min_{0\leq r\leq r_{\max}} BIC_r(r)$$

Afterward, we use this criteria for choosing number of factors in the estimation.

4 Forecasting Methods

Now that we have a framework for estimating factors for use in construction of prediction models, we turn to the specification of prediction models.

4.1 Bagging

Bagging, which is short for 'bootstrap aggregation', was introduced by Breiman (1996) as a device for reducing the prediction error of learning algorithms. Bagging starts by drawing bootstrap samples from the training sample (i.e. in-sample), applying the learning algorithm (prediction model) to each bootstrap sample, and averaging or voting the resulting prediction rules; that is, averaging or otherwise voting the predicted values for test observations. Consider the regression problem with the training sample $Z = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), ..., (\mathbf{X}_T, Y_T)\}$, where $\mathbf{X}_i = (X_{i1}, X_{i2}, ...X_{iN})$, and where we form prediction values, $\hat{Y}(\mathbf{X})$. Bagging averages this prediction value over bootstrap sample so that it can reduce the variance of prediction. For each sample, among B bootstrap samples, $Z_b^*, b = 1, 2, ..., B$, we regress Y_b^* on $\mathbf{X}_b^*$ and get fitted value $\hat{Y}_b^*(\mathbf{X}_b^*)$. Then the bagging estimator is:

$$\hat{Y}_{bagging} = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_b^*(\mathbf{X}_b^*) \tag{19}$$

Let φ be the empirical distribution of original data set. The true bagging estimator is $E_{\varphi}\hat{Y}^*(\mathbf{X}^*)$ and (19) is a Monte Carlo estimate of the true estimator (see Hastie et al. (2001)).

Breiman (1996) presents empirical evidence that bagging can indeed reduce prediction error and that averaging over bootstrap samples reduces the variance component of the prediction error. Because of the above arguments, bagging is good for the data which is unstable. Bühlmann and Yu (2002) consider bagging with a fixed number of strictly exogenous regressors and i.i.d. errors, and show that, asymptotically, the bagging estimator can be represented in shrinkage form. Namely:

$$\bar{Y}_{T+1|T} = \sum_{t=1}^{n}\psi(\kappa t_i)\hat{\delta}_i P_{T+1} + o_p(1), \tag{20}$$

where $\bar{Y}_{T+1|T}$ is the forecast of Y_{T+1} made using data through time T, $\hat{\delta}_i = T^{-1}\Sigma_{t=1}^{T}P_{it}Y_{t+1}$ is the least squares estimator of δ_i(the i-th element of δ), $t_i = \sqrt{T}\hat{\delta}_i/s_e$, with $s_e^2 = \Sigma_{t=1}^{T}(Y_{t+1} - \hat{\delta}'P_t)^2/(T-n)$, the factor κ is user specified, and ψ is a function specific to the forecasting method. In the current context, ψ is:

$$\psi^{Bagging}(t) = 1 - \Phi(t+c) + \Phi(t-c) + t^{-1}[\phi(t-c) - \phi(t+c)] \tag{21}$$

where c is the pre-test critical value. Further, ϕ is the standard normal density and Φ is the standard normal CDF. Stock and Watson (2005a) follow this approach in the context of many predictors. We also follow this shrinkage representation, thus avoiding generating huge

numbers of bootstrap sample (see Stock and Watson (2005a) for details). The t-statistics used for shrinkage factors are computed using LS with Newey-West standard errors and the pretest critical value for bagging in this paper is set at $c = 1.96$.

4.2 Boosting

The motivation for boosting is a procedure that combines the outputs of many "weak" learners (model) to produce a powerful "committee." In this sense, boosting bears a resemblance to bagging and other committee-based approaches. Boosting was first proposed by Freund and Schapire (1997), using the so-called 'AdaBoost' algorithm and AdaBoost and other boosting algorithms have attracted a lot of attention due to their great success in data modeling. Conceptually, the boosting method builds on a user-determined many weak learner and uses it repeatedly on modified data which are typically outputs from previous iterations of the algorithm. The final boosted procedure takes the form of linear combinations of weak learners. Consider AdaBoost. Namely, consider a two class problem with response variable $Y \in \{-1, 1\}$ and initialize the observation weights $w_i = 1/N$, for $i = 1, 2, ..., N$. Given an explanatory variable X, a classifier or learner, $G(X)$, is defined to produce two cases, -1 or 1. At the m-th iteration, we get an error rate

$$\hat{e}_m = \frac{1}{N} \frac{\sum_{i=1}^{N} w_i I\left(Y_i \neq G_m\left(X_i\right)\right)}{\sum_{i=1}^{N} w_i} \tag{22}$$

where I is the indicator function. Now, compute $\alpha_m = \log\left(\left(1 - e_m\right)/e_m\right)$ and set

$$w_i' = w_i \cdot \exp\left[\alpha_m \cdot I\left(Y_i \neq G\left(X_i\right)\right)\right] \tag{23}$$

and at M-th step, the output is

$$G(X) = sign\left[\sum_{i=1}^{M} \alpha_m G_m(X)\right] \tag{24}$$

That is, at every step $G_m(X)$ gives output with weighted samples, and the final output is weighted combination of each 'weak' learners. Friedman et al. (2000) extend this AdaBoost to Real AdaBoost which returns real-valued predictions. We can generalize this algorithm to the basic regression case with data $(Y, \mathbf{X})$, where Y is a $T \times 1$ vector and $\mathbf{X} = \{X_1, X_2, ..., X_N\}$, where X_i is a $T \times 1$ vector including lags of Y. Let $\hat{G}(\mathbf{X})$ be a function(learner) (defined

on $\mathbb{R}^n$), and let $L(Y_t, G(X_t))$ be the loss function that penalizes the deviation of $\hat{G}(\mathbf{X})$ from Y, at time t. The objective is to estimate the $G(\cdot)$ that minimizes the expected loss, $E\left[L\left(Y_t, \hat{G}(\mathbf{X}_t)\right)\right]$. If we use a quadratic loss function, it is easy to get an optimal solution. The appropriate algorithm is:

Algorithm 1 *Boosting*

1. Initialize : $\hat{G}_0(X_t) = \bar{Y}$ for each t.
2. For $m = 1, ..., M$, compute residuals $u_t = Y_t - \hat{G}_{m-1}(X_t)$ under the quadratic loss function and fit the base learner to the current residuals. The fit is denoted by $\hat{g}_m = X_t'\hat{\beta}$ where $\hat{\beta} = \arg\min \Sigma_{t+1}^{T} (u_t - X_t'\beta)^2 + \lambda \|\beta\|^2$, for some λ if with regularized regression.
3. Update $G_m(X) = G_{m-1}(X) + \nu \hat{g}_m(X)$, where $0 \leq \nu \leq 1$ is the step length. Repeat above **second** step until some stopping value for the number of iteration is reached.

Therefore, $G(\cdot)$ is the sum of weak learners, $G_m(\cdot)$ and $G(X) = G_0(X) = \nu \Sigma_{m=1}^{M} \hat{g}_m(X)$. Boosting will have different results according to the particular selection of learners. Popular ones for regression problems include the smoothing spline, kernel regression, or least squares. There is also clearly dependence upon loss function. In step two, the algorithm is similar to stagewise forward regression[2], which is also similar to LARS (discussed in section 4.4). Step three represents boosting have 'ensemble scheme' that aggregates many function estimates from different sample from training data and this is why boosting looks similar to bagging algorithm.

Friedman (2001) introduce L_2Boosting, which takes the simple form of refitting base learners to residuals of previous iterations. Bühlmann and Yu (2003) suggest fitting learners using one predictor at one time when large number of predictors exists and also call the boosting algorithm that minimizes quadratic loss as Component-Wise L_2Boosting. Bai and Ng (2008a) modify this algorithm to handle time-series. We follow this algorithm in the sequel. Namely, we consider:

Algorithm 2 *Component-Wise L_2Boosting*

1. Initialize : $\hat{G}_0(X_t) = \bar{Y}_t$ for each t.

[2]Stagewise forward regression begins with $\hat{\mu} = 0$ and define a vector of 'current correlations': $\hat{c} = c(\hat{\mu}) = X'(Y - \hat{\mu})$ then find the variable index $j = argmax\,|\hat{c}_j|$ and update $\hat{\mu}_m = \hat{\mu}_{m-1} + \varepsilon \cdot sign(\hat{c}_j) X_j$ where ε is a small constant.

2. For $m = 1, ..., M$

   a. for $t = 1, ..., T$, let $u_t = Y_t - \hat{G}_{t,m-1}$ be the 'current residual'.

   b. for each $i = 1, .., N$, regress the current residual vector u on $X_{.,i}$(the i-th variable) to obtain $\hat{\beta}_i$. Compute the $\hat{d}_{.,i} = u - X_{.,i}\hat{\beta}_i$, and sum of square residual, $SSR_i = \hat{d}'_{.,i}\hat{d}_{.,i}$;

   c. let i^*_m be such that $SSR_{i^*} = \min_{i\in[1,...,N]} SSR_i$

   d. let $\hat{g}_m = z_{.,i^*_m}\hat{\beta}_{i^*_m}$

3. For $t = 1, ..., T$, update $\hat{G}_m(\cdot) = \hat{G}_{m-1}(\cdot) + \nu\hat{g}_m(\cdot)$, where $0 \leq \nu \leq 1$ is the step length.

In Algorithm 2 (b), $X_{.,i}$ is the vector of T observations for i-th explanatory variable and variable i^*_m is the smallest sum of square residual among all candidate variables at the m-th step. In the set of variables, $\mathbf{X}$, in the above algorithm, we include lags of Y like W in (4).

Since boosting algorithm has no criteria to end the procedure, we may encounter a problem of over-fitting if we iterate component-wise boosting algorithm too many times. Therefore, selecting the number of iterations is crucial. Bai and Ng (2008a) define the stopping parameter M using an information criteria of the form:

$$IC(m) = \log\left(\hat{\sigma}^2_m\right) + \frac{A_T \cdot df_m}{T} \tag{25}$$

where $\hat{\sigma}^2_m = \Sigma^T_{t=1}\left(Y_t - \hat{G}_{t,m}\right)^2$. Then

$$M = \arg\min_m IC(m). \tag{26}$$

Also degrees of freedom is defined as $df_m = trace(B_m)$, where $B_m = B_{m-1} + \nu P^{(m)}(I_T - B_{m-1}) = I_T - \Pi^m_{j=0}\left(I_T - \nu P^{(j)}\right)$, with $P^{(m)} = X_{.,i^*_m}\left(X'_{.,i^*_m}X_{.,i^*_m}\right)^{-1}X_{.,i^*_m}$. Starting values of B_m are given as $B_0 = \frac{1}{\nu}P^{(0)} = \mathbf{1}'_T\mathbf{1}_T/T$, where $\mathbf{1}_T$ is $T \times 1$ vector of 1's.

4.3 Least Absolute Shrinkage Selection Operator

The least absolute shrinkage selection operator (Lasso) was introduced by Tibshirani (1996). One reason for its popularity is that it provides an efficient computational tool for handling more predictor variables than observations. Moreover, it produces parsimonious models that are easy to interpret. (Meinshausen and Yu (2009)). Bai and Ng (2008b) use the Lasso estimator for model selection in the context of time series data. The Lasso can be interpreted as a constrained version of ordinary least squares. Given X $(T \times N)$ and Y $(T \times 1)$, it can be

assumed that the covariates have mean zero and unit length, under certain standardizations, so that the response variable also has mean zero. Namely:

$$\sum_{i=1}^{T} Y_i = 0, \quad \sum_{i=1}^{T} X_{ij} = 0 \text{ and } \sum_{i=1}^{T} X_{ij}^2 = 1, \quad \text{for } j = 1, 2, ...N \tag{27}$$

And the predictor is:

$$\hat{\boldsymbol{\mu}} = \sum_{j=1}^{N} X_j \hat{\beta}_j = X\hat{\boldsymbol{\beta}} \tag{28}$$

where $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_N\right)$ is the candidate solution for the Lasso estimator. Total square error is

$$SST\left(\hat{\boldsymbol{\beta}}\right) = \|Y - \hat{\boldsymbol{\mu}}\|^2 = \sum_{i=1}^{T} \left(Y_i - \hat{\mu}_i\right)^2$$

Let $N\left(\hat{\boldsymbol{\beta}}\right)$ be the absolute norm of $\hat{\boldsymbol{\beta}}$,

$$N\left(\hat{\boldsymbol{\beta}}\right) = \sum_{j=1}^{N} \left|\hat{\beta}_j\right|$$

Then the solution to the Lasso is:

$$\hat{\boldsymbol{\beta}}_{Lasso} = \min SST\left(\hat{\boldsymbol{\beta}}\right), \quad \text{subject to a bound, } t, \text{ on } N\left(\hat{\beta}\right)$$

The Lasso is the special case of the algorithm which will be explained in next section, and hence further discussion of it is omitted for the sake of brevity.

4.4 Least Angle Regression

Least Angle Regressions (LARs) shares similar properties with the model-selection method known as stagewise forward regression as explained in section 4.2. It is also very similar to the Lasso in equation (27) and (28). Here is a brief look of the algorithm. The method starts with $\hat{\mu}_j = 0$, where $\hat{\mu}_j$ is the current estimator of Y, with j being the number of predictors. Let $c\left(\hat{\mu}\right) = \hat{c} = X'\left(Y - \hat{\mu}_j\right)$ be the 'current correlations' where X is all of predictors. Then, we consider $\hat{c}_j$ proportional to the correlation between covariate X_j and current residual. Therefore, there exists a j such that $|\hat{c}_j|$ is maximized and the updating rule is $\hat{\mu}_{j+1} = \hat{\mu}_j + \hat{\kappa} sign\left(\hat{c}_j\right) X_{j.}$, where $\hat{\kappa}$ is some small constant. Here, "small" is critical because if it is not small, then we are actually implementing another model selection method,

namely forward selection[3].

More specifically, like stagewise regression, start at $\hat{\mu}_r = 0$ and build up $\hat{\boldsymbol{\mu}}$ in steps. Consider $\hat{\mu}_k = 0$ to be the current LARs estimator at a certain step so that we can set $c(\hat{\mu}) = \hat{c} = X'(Y - \hat{\mu}_M)$ to be the current correlations. We can define a set M as a set including variables which correspond to covariates with the largest absolute correlations; so that we can define

$$\hat{C} = \max_j \{\hat{c}_j\} \text{ and } M = \left\{ j : |\hat{c}_j| = \left| \hat{C} \right| \right\}$$

by letting $s_j = sign(\hat{c}_j)$, for $j \in M$ and defining the active matrix corresponding to M as

$$X_M = (s_j x_j)_{j \in M} \tag{29}$$

Let $G_M = X_M' X_M$ and $A_M = \left(\mathbf{1}_M' G_M^{-1} \mathbf{1}_M\right)^{-1/2}$, where $\mathbf{1}_M$ is a vector of ones equaling the rank of M. A unit equiangular vector with columns of the active set matrix X_M can be defined as

$$u_M = X_M w_M, \quad w_M = A_M G_M^{-1} \mathbf{1}_M, \quad a_M = X' w_M \tag{30}$$

so that $X_M' u_M = A_M \mathbf{1}$. LARs then updates $\hat{\mu}$ as

$$\hat{\mu}_{M^+} = \hat{\mu} + \hat{\kappa} u_M \tag{31}$$

where

$$\hat{\kappa} = \min_{j \in M^c}{}^+ \left(\frac{\hat{C} - \hat{c}_j}{A_M - a_j} \right) \left(\frac{\hat{C} + \hat{c}_j}{A_M + a_j} \right). \tag{32}$$

Efron et al. (2004) show that the Lasso is in fact a special case of LARs that imposes the sign restriction. Let $\hat{\beta}$ be a Lasso solution as in equation (28), with $\hat{\mu} = X\hat{\beta}$. Using Lemma 8 in Efron et al. (2004), we see that the sign of any non-zero coordinate $\hat{\beta}_j$ must agree with the sign, s_j, of the current correlation, $\hat{c}_j$. That is, $sign\left(\hat{\beta}_j\right) = sign(\hat{c}_j) = s_j$. (Some modification to LARs leads to the above result.) Consider an active set, M, and corresponding LARs estimator, $\hat{\mu}_M$, corresponding to a Lasso estimator $\hat{\mu} = X\hat{\beta}$. Define $\hat{d}$ to be an $N \times 1$ vector equal to $s_j w_{M_k}$, for $j \in M$, and zero elsewhere. Then we can see that:

$$\boldsymbol{\mu}(\kappa) = X\beta(\kappa), \quad \text{where} \quad \beta_j(\kappa) = \hat{\beta}_j + \kappa \hat{d}_j$$

[3]Forward selection is a naive version of stagewise forward regression. First, let $\mu = E(Y|X)$ then start with a constant model $\hat{\mu} = 0$ assuming that Y is centered and X_i are standardized. Then given a set of covariates, select X_i with the largest absolute correlation with Y. Next step is fitting a linear model with X_i and update $\hat{\mu}_m = \hat{\mu}_{m-1} + \hat{\beta}_i X_i$ then get a residual vector $r = Y - \hat{\mu}_m$ and project other X_i orthogonally to $X_{\text{-}i}$.and repeat this selection process.

Here, $\beta_j(\kappa)$ will change its sign at $\kappa_j = -\frac{\hat{\beta}_j}{\hat{d}_j}$ and this change happens at $\tilde{\kappa} = \min_{\kappa_j > 0}\{\kappa_j\}$ at first. If $\tilde{\kappa} < \hat{\kappa}$, stop the LARS step at $\kappa = \tilde{\kappa}$ and remove $\tilde{j}$ from the calculation of next equiangular direction, so that

$$\hat{\mu}_{M^+} = \hat{\mu}_M + \tilde{\kappa} u_M \quad \text{and} \quad M^+ = M - \left\{\tilde{j}\right\}$$

Then this solution is equivalent to Lasso solution $\hat{\mu}$.

For applying LARs to time series data, Gelper and Croux (2008) revise the basic algorithm described here. They start by fitting an autoregressive model to response variable, excluding predictor variables, using least squares. The corresponding residual series is retained and its standardized version is denoted z_0. The time-series LARs (henceforth, TS-LARs) procedure ranks the predictors according to how much they contribute to improving upon autoregressive fit. The following is algorithm of Gelper and Croux (2008) used starts with centering response variable and standardizing explanatory variables.

Algorithm 3 *Time Series LARs*

1. Fit an autoregressive model to the response variable without predictors using LS and retain the corresponding residual, z_0.

2. For $k = 1, 2, ..., N$

   (a) for $j = 1, ..., k$, find the first ranked predictor, X_{j^*}, which has the highest R^2 measure $R^2\left(z_{k-1} \sim \bar{X}_{j^*}\right)$, where $\bar{X}_{j^*}$ is a matrix of lagged X_{j^*} variables and, R^2 is measure of least square regression fit. The predictor with highest R^2 is denoted $X_{(k)}$.

   (b) $X_{(k)}$ denotes the k-th ranked predictor and the active set A contains $X_{(1)}, X_{(2)}, ..., X_{(k)}$, with $k \geq 2$.

   (c) The H matrix corresponding to the active predictor, $H_{(k)}$ is the projection matrix on the space spanned by the columns of $\bar{X}_{j^*}$. That is, $H_{(k)} = \bar{X}_{(j^*)}\left(\bar{X}'_{(j^*)}\bar{X}_{(j^*)}\right)^{-1}\bar{X}_{(j^*)}$.

   (d) Let $\tilde{X}_{(k)}$ be the standardized vector of fitted values $H_{(k)}z_{k-1}$. Then find equiangular vector u_k, where $u_k = \left(\tilde{X}_{(1)}, \tilde{X}_{(2)}, ..., \tilde{X}_{(k)}\right)w_k$ with $w_k = \frac{R_k^{-1}\mathbf{1}_k}{\sqrt{\mathbf{1}'_k R_k^{-1}\mathbf{1}_k}}$, and where R_k is the $(k \times k)$ correlation matrix computed from $\tilde{X}_{(1)}, \tilde{X}_{(2)}, ..., \tilde{X}_{(k)}$ and $\mathbf{1}_k$ is a vector of ones of length of k.

3. Update the response $z_k = z_{k-1} - \gamma_k u_k$, where γ_k is the smallest positive solution such that for a predictor X_j which is not in the active set so far, the following holds: $R^2\left(z_{k-1} - \gamma u_k \sim \tilde{X}_{(k)}\right) = R^2\left(z_{k-1} - \gamma u_k \sim \bar{X}_j\right)$.

   Then the associated predictor, $X_{(k+1)}$ is added to the active set and the new response is standardized and denoted by z_k. See page 10 of Gelper and Croux (2008) for further computational details.

After ranking the predictors, $X_{(k+1)}$, the highest ranked will be included in the final model. Now the only choice remaining is how many "highest" predictors to include in the model. This number, k is chosen using the Bayesian Information Criteria, as done in Gelper and Croux (2008).

4.5 Elastic Net

Zou and Hastie (2005) point out that the Lasso has some limitations under certain scenarios, such as when T is greater than N or when there is a group of variables among which the pairwise correlations are very high. They develop a new regularization method, so called the "elastic net", that works as well as Lasso and remedies the above problems. The algorithm is like a stretchable fishing net that retains all the big fish and that's why it is named. The method starts similarly to the Lasso estimator explained in previous section. With some transformation, it is also assumed that responses are centered and explanatory variables also standardized, as in equation (27). For any fixed non-negative λ_1 and λ_2, the naive elastic net criterion is defined as:

$$L(\lambda_1, \lambda_2, \beta) = |Y - \mathbf{X}\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1, \tag{33}$$

where $|\beta|^2 = \sum_j^N \beta_j^2$ and $|\beta|_1 = \sum_j^N |\beta_j|$. The naive elastic net estimator is $\hat{\beta} = \arg\min_\beta \{L(\lambda_1, \lambda_2, \beta)\}$. This problem is equivalent to the optimization problem:

$$\hat{\beta} = \arg\min_\beta |Y - \mathbf{X}\beta|^2, \quad \text{subject to } (1-\alpha)|\beta|_1 + \alpha|\beta|^2, \tag{34}$$

where $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$. The term $(1-\alpha)|\beta|_1 + \alpha|\beta|^2$ is called the elastic net penalty, and leads to the Lasso or ridge estimator, depending on the value of α.(If $\alpha = 1$, it becomes ridge regression, if $a = 1$, it is the Lasso, and if $\alpha \in (0.1)$, it has properties of both methods.) The solution to the naive elastic net solution begins with defining new data set (X^+, Y^+), where

$$\mathbf{X}^{+}_{(T+N)\times N} = (1+\lambda_2)^{-1/2}\begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_N \end{pmatrix} \qquad Y^{+}_{(T+N)\times 1} = \begin{pmatrix} Y \\ \mathbf{0}_N \end{pmatrix}. \tag{35}$$

Then we can rewrite the naive elastic new criterion as:

$$L\left(\frac{\lambda_1}{\sqrt{1+\lambda_2}}, \beta\right) = L\left(\frac{\lambda_1}{\sqrt{1+\lambda_2}}, \beta^{+}\right) = \left|Y^{+} - \mathbf{X}^{+}\beta^{+}\right|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\left|\beta^{+}\right|_1 \tag{36}$$

If we let

$$\hat{\beta}^{+} = \arg\min_{\beta^{+}} L\left(\frac{\lambda_1}{\sqrt{1+\lambda_2}}, \beta^{+}\right) \tag{37}$$

the the naive elastic net estimates $\hat{\beta}_{NEN}$ is

$$\hat{\beta}_{NEN} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^{+} \tag{38}$$

In this orthogonal setting, the naive elastic net can be represented as combination of ordinary squares and parameters (λ_1, λ_2):

$$\hat{\beta}_{i,NEN} = \frac{\left(\left|\hat{\beta}_{i,LS}\right| - \lambda_1/2\right)_{+}}{1+\lambda_2} sign\left\{\hat{\beta}_{i,LS}\right\}. \tag{39}$$

Here, $+$ denotes the positive part of parentheses, which is itself if it is bigger than zero and zero otherwise. The ridge estimator can be written as

$$\hat{\beta}_{ridge} = \frac{\hat{\beta}_{LS}}{1+\lambda_2} \tag{40}$$

and the Lasso is

$$\hat{\beta}_{Lasso} = \left(\left|\hat{\beta}_{i,LS}\right| - \lambda_1/2\right)_{+} sign\left\{\hat{\beta}_{i,LS}\right\}, \tag{41}$$

and depends on the value of λ_1 and λ_2.

In the above naive elastic net, there is double shrinkage, which does not help to reduce the variance and may lead unnecessary bias. Zou and Hastie (2005) propose the elastic net, in which this double shrinkage is corrected. Given equation (35), the naive elastic net solves the Lasso-type problem of the type:

$$\hat{\beta}^{+} = \arg\min_{\beta^{+}} \left|Y^{+} - \mathbf{X}^{+}\beta^{+}\right|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\left|\beta^{+}\right|_1. \tag{42}$$

In this context, the elastic net estimator, $\hat{\beta}_{EN}$, is defined as:

$$\hat{\beta}_{EN} = \sqrt{1+\lambda_2}\hat{\beta}^{+}. \tag{43}$$

Thus ,

$$\hat{\beta}_{EN} = (1 + \lambda_2)\, \hat{\beta}_{NEN}. \tag{44}$$

By this rescaling, this estimator preserves the properties of naive elastic net. Moreover, by Theorem 2 in Zou and Hastie (2005), is can be seen that the elastic net is a stabilized version of Lasso. Namely,

$$\hat{\beta}_{EN} = \arg\min_{\beta} \beta' \left(\frac{\mathbf{X}'\mathbf{X} + \boldsymbol{\lambda}_2 \mathbf{I}}{1 + \lambda_2} \right) \beta - 2Y'\mathbf{X}\beta + \lambda_1 |\beta|_1 \tag{45}$$

and

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \beta' \left(\mathbf{X}'\mathbf{X} \right) \beta - -2Y'\mathbf{X}\beta + \lambda_1 |\beta|_1 \tag{46}$$

Zou and Hastie (2005) propose an algorithm called the LARs-EN to estimate $\hat{\beta}_{EN}$ using the LARs of previous section. Namely, with fixed λ_2, the elastic net problem is equivalent to the Lasso problem on the augmented data set $(\mathbf{X}^+, Y^+)$. This is the approach that we follow in this paper. In summary, LARs can be used to create the "entire elastic net solution path" and provide an efficient EN estimator.

4.6 Non-Negative Garotte

The non-negative garotte estimator was introduced by Breiman (1995), who showed that their proposed estimation algorithm is a stable variable selection method that often outperforms its competitors like ridge regression, in the content of cross-sectional data. Since the non-negative garotte (henceforth, NNG) estimator is a scaled version of the least square estimator, consider a standard linear regression model. Consider the simple regression equation

$$Y = X\beta + \varepsilon, \tag{47}$$

where all variables are orthogonalized as in equation(27). There are given shrinkage factors, $s(\lambda) = (s_1(\lambda), s_2(\lambda), ..., s_N(\lambda))'$, and the objective is to minimize:

$$\frac{1}{2} \|Y - \hat{\mu}s\|^2 + n\lambda \sum_{j=1}^{N} s_j, \qquad \text{subject to } s_j > 0, \forall j, \tag{48}$$

where $\hat{\mu}$ is defined as in equation (28). Here $\lambda > 0$ is the tuning parameter. The NNG estimate of the regression coefficient is subsequently defined as :

$$\hat{\beta}_j^{NNG}(\lambda) = s_j(\lambda)\, \hat{\beta}_j^{LS}, \quad j = 1, 2, ..., N. \tag{49}$$

where $\hat{\beta}_j^{LS}$ is the least square estimator of β_j. The minimizer of expression (48) has the following explicit form:

$$s_j(\lambda) = \left(1 - \frac{\lambda}{\hat{\beta}_j^{LS}}\right)_+, \quad j = 1, 2, ..., N. \tag{50}$$

The disadvantage of NNG is its dependence on the ordinary least squares estimator, which can be especially problematic in small samples. Accordingly, Yuan and Lin (2007) consider Lasso, ridge regression, and the elastic net as alternatives for providing and initial estimate use in the NNG; and they prove that if the initial estimate is consistent regardless of methods, the non-negative garotte remains a consistent estimator, given that the tuning parameter, λ, is chosen appropriately. However, Zou (2006) show that the original non-negative garotte with ordinary least square is also consistent, if N is fixed, as $T \to \infty$. Our approach is to start the algorithm with the least square estimator. Yuan (2007) gives an algorithm for non-negative garotte as follows, and this is the algorithm that we use.

Algorithm 4 *Non-Negative Garotte*

1. Set $d^0 = 0$, $l = 1$, $\theta^0 = Y$

2. Compute the current active set

$$C_1 = \arg\max_i \left(\hat{\mu}_i' \theta^{l-1}\right)$$

   where $\hat{\mu}_i = X_i \hat{\beta}_i^{LS}$

3. Compute the current direction γ, which is a N dimensional vector defined by $\gamma_{c_l} = 0$ and

$$\gamma_{C_l} = \left(\hat{\mu}_{C_l}' \hat{\mu}_{C_l}'\right)^{-1} \hat{\mu}_{C_l}' \theta^{l-1}$$

4. For every $i \notin C_l$, compute how far the group non-negative garotte will progress in direction γ before X_i enters the active set. This can be measured by a α_i such that

$$\hat{\mu}_i' \left(\theta^{l-1} - \alpha_i \hat{\mu} \gamma\right) = \hat{\mu}_j' \left(\theta^{l-1} - \alpha_i \hat{\mu} \gamma\right)$$

   where j is arbitrary chosen from C_l.

5. For every $i \in C_l$, compute $\alpha_i = \min(\beta_i, 1)$ where $\beta_i = -d_i^{l-1}/\gamma_i$, if nonnegative, measures how far the group non-negative garotte will progress before d_i becomes zero.

6. If $\alpha_i \leq 0$, $\forall i$ or $\min_{i,\alpha_i>0}\{\alpha_i\} > 1$, set $\alpha = 1$. Otherwise, denote $\alpha = \min_{i,\alpha_i>0}\{\alpha_i\} \equiv \alpha_{i^*}$. Set $d^l = d^{l-1} + \alpha\gamma$. If $i^* \notin C_l$, update $C_{l+1} = C_l \cup \{i^*\}$; else update $C_{l+1} = C_l - \{i^*\}$.

7. Set $\theta^l = Y - Zd^l$ and $l = l + 1$. Go back to step (3) until $\alpha = 1$.

4.7 Bayesian Model Averaging

Bayesian Model Averaging (BMA) has received considerable attention in recent years (see. e.g. Hoeting et al. (1999) and Koop and Potter (2004)). The basic idea of BMA starts with supposing researchers are interested in R possible models, denoted by $M_1, ..., M_R$, and that the quantity to be forecast is y_{t+h}.. Each model has a parameter vector, θ_r, that depends on its prior, likelihood and posterior. If we denote κ as a parameter vector that is common to all possible models, then κ is a function of θ_r, for all of r. We can define the following probability:

$$p(\kappa|Data) = \sum_{r=1}^{R} p(\kappa|Data, M_r)\, p(M_r|Data). \tag{51}$$

If $g(\kappa)$ is a function of κ, the conditional expectation is given as:

$$E[g(\kappa)|Data] = \sum_{r=1}^{R} E[g(\kappa)|Data, M_r]\, p(M_r|Data). \tag{52}$$

Then:

$$E(y_{T+h}|Data) = \sum_{r=1}^{R} p(M_r|Data)\, E(y_{T+h}|Data, M_r). \tag{53}$$

By Bayesian inference, we can obtain these results for every model and average them, where the weights used in averaging are posterior probabilities. However, implementing Bayesian model averaging can be difficult because the number of models considered can be very large. For example, if we have 15 potential variables, then we have 2^{15} possible models. This means that we must estimate $32,768$ models at every forecasting horizon and prior to the construction of each new prediction if recursive or rolling estimation methods are used, as in this paper. This leads to algorithms which do not require us to consider every possible models. The most popular one is MC^3, which takes draws from the posterior distribution of models and MCMC draws from the posterior distribution of parameters.

The likelihood function for each model is based on the normal linear regression model as

in (47):

$$p\left(y|\beta_r,\sigma_r^2\right)=\frac{1}{(2\pi)^{N/2}}\left\{\sigma\exp\left[-\frac{\sigma_r^2}{2}\left(\beta_r-\hat{\beta}_r\right)'X_rX_r'\left(\beta_r-\hat{\beta}_r\right)\right]\right\}\left\{\sigma_r^v\right\}\exp\left[-\frac{\sigma_r^2v_r}{2s_r^{-2}}\right], \tag{54}$$

where $v_r=T-K_r$, $\hat{\beta}_r=\left(X_r'X_r\right)^{-1}X_ry$, $s_r^2=\frac{\left(y-X_r\hat{\beta}_r\right)'\left(y-X_r\hat{\beta}_r\right)}{v_r}$ and X_r is a $T\times K_r$ matrix containing some or all columns of X.

The selection of prior is quite crucial in model averaging. Here, we use a Normal-Gamma natural conjugate prior following Fernandez et al. (2001a) and Koop and Potter (2004). Since we assume

$$\varepsilon\sim N\left(0,\sigma^2I_T\right), \tag{55}$$

we can use a natural conjugate prior:

$$\beta_r|\sigma^2\sim N\left(\underline{\beta}_r,\sigma^{-2}\underline{V}_r\right), \tag{56}$$

and

$$\sigma^{-2}\sim G\left(\underline{s}^{-2},\underline{d}\right), \tag{57}$$

where $G\left(\underline{s}^{-2},\underline{d}\right)$ denotes the Gamma distribution with mean $\underline{s}^{-2}$ and degrees of freedom $\underline{d}$. When we have many potential explanatory variables, we may suspect that many of variables might be irrelevant, and so we can set $\underline{\beta}_r=0$. For choosing $\underline{V}_r$, we use the so-called g-prior, which was first introduced by Zellner (1986). Namely:

$$\underline{V}_r=\left[g_rX_r'X_r\right]^{-1}. \tag{58}$$

This prior states that the prior covariance of β_r is proportional to the comparable data-based quantity. Summarizing, we can set:

$$\beta_r|\sigma^2\sim N\left(0,\sigma^{-2}\left[g_rX_r'X_r\right]^{-1}\right). \tag{59}$$

For the models considered, $p\left(M_r|Data\right)$ can be estimated for $r=1,...,R$. It is also straightforward to calculate $E\left(y_{T+h}|Data,M_r\right)$ for every model. An alternative to Bayesian model averaging is Bayesian model selection, which involves simply choosing M^* which has the maximum value for $p\left(M_r|Data\right)$. Point forecasts can then be based on $E\left(y_{T+h}|Data,M^*\right)$. We consider only model averaging. BMA has been studied with regression models with many variables. In theory, if you treat models as random variables, model averaging is the correct thing to do, in the sense that equation (53) follows from probability rules. Moreover,

Min and Zellner (1993) and Raftery et al. (1997) show how model averaging is optimal for forecasting in decision theory problems. Koop and Potter (2004) point out that BMA does not suffer from the criticisms associated with sequential testing procedures, since it formally includes model uncertainty in the statistical procedure.

We follow the methodology of Koop and Potter (2004), as applied to factor modelling based on Stock and Watson (2002). In particular, in (47), we assume that:

$$\varepsilon \sim N\left(0, \sigma^2 I_T\right). \tag{60}$$

We use a natural conjugate prior,

$$\beta|\sigma^{-2} \sim N\left(\beta, \sigma^2 \bar{B}\right) \tag{61}$$

and

$$\sigma^{-2} \sim G\left(\bar{s}^{-2}, \bar{\nu}\right), \tag{62}$$

where $G\left(\bar{s}^{-2}, \bar{\nu}\right)$ denotes the Gamma distribution with mean s^{-2} and degrees of freedom $\bar{\nu}$. According to the algorithm in Section 3 of Clyde (1999), Koop and Potter (2004) use an orthogonal transformation of (47) of the sort used in the factor analysis literature. The posterior for parameter β_r, β for r model, has a multivariate t distribution with mean

$$E\left(\beta_r|y, M_r\right) \equiv \bar{\beta}_r = \bar{V}_r X_r' y, \tag{63}$$

where X_r are the explanatory variables included in model r, with covariance matrix

$$var(\beta_r|y, M_r) = \frac{\bar{\upsilon}\bar{s}_r^2}{\bar{\upsilon} - 2}\bar{V}_r \tag{64}$$

and $\bar{\upsilon} = N$ degrees of freedom, with

$$\bar{V}_r = \left[(1+g)X_r' X_r\right]^{-1} \tag{65}$$

and

$$\bar{s}_r = \frac{\frac{1}{g_r+1} y' P_x y + \frac{g_r}{g_r+1}\left(y - \bar{y}\mathbf{1}_n\right)'\left(y - \bar{y}\mathbf{1}_n\right)}{\bar{\upsilon}}, \tag{66}$$

where

$$P_x = I_n - X_r\left(X_r' X_r\right)^{-1} X_r' \tag{67}$$

and $\mathbf{1}_n$ is an $nx1$ row vector whose components are all ones.

Using $g-prior$, the marginal likelihood for model r is :

$$p\left(y|M_r\right) \propto \left(\frac{g_r}{g_r+1}\right)^{\frac{k_r}{2}} \left[\left(\frac{1}{g_r+1}\right) y' P_{X_r} y + \left(\frac{g_r}{g_r+1}\right)\left(y-\bar{y}\mathbf{1}_n\right)'\left(y-\bar{y}\mathbf{1}_n\right)\right]^{-\frac{N-1}{2}}, \quad (68)$$

where k_r is the number of potential explanatory variables in model r. So, the posterior model probabilities can be found from $p\left(M_r|y\right) = cp\left(y|M_r\right)p\left(M_r\right)$, where c is a constant. The common way to allocate is using equal prior model probability, such as setting $p\left(M_r\right) = 1/R$. Since R and c are constants, this cancels out in all related equation.

If we define $Z = XW$, where W is a nonsingular $K \times K$ matrix chosen so that the columns of Z are orthogonal,: we can write (47) as

$$y = Z\alpha + \varepsilon, \quad (69)$$

where $\alpha = W^{-1}\beta$. Now, this is thus the framework that we set up in Section 2. The prior for σ^2 is unaffected by the transformation and the prior for the regression coefficient becomes

$$\alpha|\sigma^2 \sim N\left(\bar{\alpha}, \sigma^2\bar{A}\right), \quad (70)$$

where $\bar{\alpha} = W^{-1}\beta$ and $\bar{A} = W^{-1}\bar{B}\left(W^{-1}\right)'$.

A standard choice of the prior model probability, $p\left(M_r\right)$ is:

$$p\left(\gamma\right) = \prod_{j=1}^{K} \theta_j^{\gamma_j}\left(1-\theta_j\right)^{r_j}, \quad (71)$$

where θ_j (for $j = 1, ..., K$) is the prior probability that each potential explanatory variable enters the model. In this paper, a popular noninformative benchmark case, $\theta_j = \frac{1}{2}$ is chosen, which implies that $p\left(M_r\right) = \frac{1}{R}$, for $r = 1, ..., R$. As done in Koop and Potter (2004), we tried another standard method. Namely, we set these values proportional to the eigenvalues of $X'X$. Now, we need to choose $\beta, \bar{B}, \bar{s}^{-2}, \bar{\nu}$ and W. Koop and Potter (2004) follow what Fernandez et al. (2001a) suggests and use a noninformative prior for σ^{-2} (namely, $\bar{\nu} = 0$ and $\bar{s}^{-2}$ does not enter the marginal likelihood or posterior). Following Fernandez et al. (2001a), the prior for these regression coefficients over zero and use a g-prior for $\bar{B}$. That is,

$$\bar{B} = \left(gX'X\right)^{-1} \quad (72)$$

The g-prior cannot be used if the original number of explanatory variables exceeds T, since then $X'X$ is singular. Koop and Potter (2004) solve this problem by working directly with

a g-prior for (69) and, thus:

$$\bar{A} = (gZ'Z)^{-1}. \tag{73}$$

In further discussion of the specification of g, Fernandez et al. (2001a) investigate the property of many possible g-priors. We follow Koop and Potter (2004) and focus on two different cases :

$$g = \frac{1}{T} \tag{74}$$

and

$$g = \frac{1}{K^2} \tag{75}$$

In implementing BMA, we also set $\theta_j = 0.5$ for all j, so that all models in our BMA have a same prior probability.

5 Data, Estimation, and Benchmark Forecasting Models

5.1 Data

The data that we use in our large-scale dataset are monthly observations on 146 U.S. macroeconomic time series for the period 1960:01 - 2007:12 ($N = 146, T = 576$). Forecasts are constructed for fourteen series, including : the unemployment rate, personal income less transfer payments, the 10 year Treasury-bond yield, the consumer price index, the producer price index, non-farm payroll employment, housing starts, industrial production, M2, the S&P 500 index, gross domestic product, retail sales, business sales and inventory and advanced durable goods shipment, and new orders and unfilled orders.[4] As discussed above, when formulating the nation's monetary policy, the Federal Reserve takes values of these variables into account.

All 146 series, excluding the selected target variables to be predicted are used in factor proxy construction for each series. For example, when we forecast the unemployment rate, the other thirteen series listed above enter into our large-scale dataset. The last three variables in our list of fourteen are only available for the period 1992:01~2007:12. These series are extracted from Global Data Insight and added to Stock and Watson (2005a)'s dataset, which includes 131 series up to 2003, which are in turn updated by us through 2007:12. Later observations are not added in order to mitigate the effects of real-time data revision on our

[4]Note that gross domestic product is reported quaterly. We interpolate these data to a monthly frequency following Chow and Lin (1971),

empirical findings. Table 1 lists the fourteen forecasted series. The third row of the table gives the transformation of the variable used in order to induce stationarity. In general, logarithms were taken for all nonnegative series that were not already in rates (see Stock and Watson (2002, 2005a) for complete details). Using the transformed data set, denoted above by $\mathbf{X}$, the factors are estimated by the method of principal components. Thereafter, the alternative methods outlined in the previous sections were used to form predictions. Finally, note that a full list of predictor variables is provided in the appendix.

5.2 Estimation

Pseudo out-of-sample forecasts are calculated for each variable and method for prediction horizons $h = 1, 3, 6$ and 12. All estimation is done anew, at each point in time, prior to the construction of each new prediction, using both recursive and rolling estimation strategies. Note that at each estimation period, the number of factors included will be different, according to the results from the test discussed in Section 2.2. Note also that lags of the target predictor variables are also included in the set of explanatory variables, in all cases. Selection of the number of lagged variables to include is done using the Bayesian Information Criteria (BIC). Once factors are estimated, at each point in time, as a second step we estimate the coefficients used in equation (4) via implementation of the various shrinkage and other methods discussed above. Out of sample forecasts begin after 10 years (i.e. the initial in-sample estimation period is R =120 observations). For example, when forecasting the unemployment rate, when $h = 12$, the first forecast will be $\hat{Y}_{136} = \hat{\beta}_F \tilde{F}_{124} + \hat{\beta}_W W_{124}$. In our rolling estimation scheme, the in-sample estimation period used to calibrate our prediction models is of length 10 years. The recursive estimation scheme begins with an in-sample period of 10 years, but a new observation is added to this sample prior to the re-estimation and construction of each new forecast, as we iterate through the ex-ante prediction period. Note that the actual observations being predicted as well as the number of predictions in our ex-ante prediction period remains fixed, regardless of forecast horizon, in order that we may compare predictive accuracy across forecast horizons as well as models.

Forecast performance is evaluated using mean square forecast error (MSFE), defined as:

$$MSFE_{i,h} = \sum_{t=R-h+2}^{T-h+1} \left(Y_{t+h} - \hat{Y}_{i,t+h} \right)^2 \tag{76}$$

where $\hat{y}_{i,t+h}$ is $i-$th method's forecast for horizon h. Forecasting experiments are carried out for $h = 1, 3, 6$ and 12, and the forecast accuracy of the models is evaluated using the predictive

accuracy test of Diebold and Mariano (1995), which is implemented using quadratic loss (i.e. MSFE), which has a null hypothesis that two models being compared have equal predictive accuracy, and finally which yields statistics with an asymptotic $N(0,1)$ limiting distribution. Namely, the null hypothesis of the test is:

$$H_0 : E\left[L\left(\varepsilon^1_{t+h|t}\right)\right] - E\left[L\left(\varepsilon^2_{t+h|t}\right)\right] = 0, \tag{77}$$

where $\varepsilon^i_{t+h|t}$ is $i-$th method's error for $h-$step forecast and $L(\cdot)$ is assumed to be the quadratic function. Thus, If the statistic is negative and significantly different from zero, then model 2 is preferred over model 1. The actual statistic in this case is constructed as: $DM = P^{-1}\sum_{i=1}^{P} d_t/\sigma_{\overline{d}}$, where $d_t = \left(\widehat{\varepsilon^1_{t+h|t}}\right)^2 - \left(\widehat{\varepsilon^2_{t+h|t}}\right)^2$, $\overline{d}$ is the mean of d_t, P denotes the length of the ex-ante prediction period, $\sigma_{\overline{d}}$ is a heteroskedasticity and autocorrelation robust estimator of the standard deviation of $\overline{d}$, and $\widehat{\varepsilon^1_{t+h|t}}$and $\widehat{\varepsilon^2_{t+h|t}}$ are estimates of the true prediction errors $\varepsilon^1_{t+h|t}$and $\varepsilon^2_{t+h|t}$ for Models 1 and 2, constructed by subtracting the actual from the predicted values of the target variable, respectively.

5.3 Benchmark Forecasting Models

In addition to the various prediction models discussed above, we form predictions using the following benchmark models.

Univariate Autoregression: Forecasts from a univariate AR(p) model are computed as $\hat{Y}^{AR}_{t+h} = \hat{\alpha} + \hat{\phi}(L) Y_t$, where coefficients are estimated via least squares (LS) regression, and lags are selected via use of the BIC.

Multivariate Autoregression: Forecasts from an ARX(p) model are computed as $Y^{ARX}_{t+h} = \hat{\alpha} + \hat{\beta} Z_t + \hat{\phi}(L) Y_t$, where Z_t is a set of exogenous (X) predictors selected using the SIC, lags are selected using the SIC, and parameters are estimated via LS regression. Selection of the exogenous predictors includes choosing up to six variables prior to the construction of each new prediction model that is in turn estimated prior to the construction of each new prediction, as the recursive or rolling sample iterates forward over time.

Principal Component Regression: Forecasts from principal component regression are computed as $\hat{Y}^{PCR}_{t+h} = \hat{\alpha} + \hat{\gamma}\tilde{F}_t$, where $\tilde{F}_t$ is estimated via principal components using $\{X_i\}_{i=1}^{T}$, as in equation (4), and coefficients are estimated via LS regression of Y^h_{t+h} onto $\left(1, \tilde{F}_{1t}, ..., \tilde{F}_{rt}\right)$

Factor Augmented Autoregression: Based on equations (4), forecasts are computed as $Y^h_{t+h} = \hat{\alpha} + \hat{\beta}_F \tilde{F}_t + \hat{\beta}_W(L) Y_t$. This equation is a direct combination of an AR(p) model,

with lags selected via use of the BIC, with the above principal component regression model.

Combined Bivariate ADL Model: As done in Stock and Watson (2005a), we implement a combined bivariate autoregressive distributed lag (ADL) model. Forecasts are constructed by combining individual forecasts computed from bivariate autoregressive distributed lag (ADL) models. The i-th ADL model includes $p_{i,x}$ lags of $X_{i,t}$, and $p_{i,y}$ lags of Y_t, and has the form $\hat{Y}_{t+h}^{ADL} = \hat{\alpha} + \hat{\beta}_i(L) X_{i,t} + \hat{\phi}_i(L) Y_t$,where coefficients are estimated via LS regression. The combined forecast is

$$\hat{Y}_{T+h|T}^{Comb,h} = \sum_{t=1}^{n} w_i \hat{Y}_{T+h|T}^{ADL,h}$$

Here, we set $(w_i = 1/n)$, where $n = 146$.

There are a number of studies that compare the performance of combining methods in controlled experiments, including: Clemen (1989), Diebold and Lopez (1996), Newbold and Harvey (2002), and Timmermann (2005) and, in the literature on factor models, Figlewski (1983), Figlewski and Urich (1983), Kitchen and Monaco (2003), and Stock and Watson (2003, 2004, 2005a, 2006). In this literature, combination methods typically outperform individual forecasts. This stylized fact is called the "forecast combining puzzle."

Mean Forecast Combination: To further examine the issue of forecast combination, we construct one forecast as the average of forecasts constructed according to the thirteen forecasting models summarized above and listed in Table 2.

6 Empirical Results

In this section, we discuss the results of our prediction experiments. The forecasting models used are summarized in Table 2, and variable mnemonics are given in Table 1. Details of the data and estimation procedures used to construct the sequences of recursive and rolling ex-ante h-step ahead forecasts reported on are outlined in the previous section.

Results are gathered in Table 3 to 6. Tables 3 and 4 report MSFEs for all alternative forecasting models, using recursive (Table 3) and rolling (Table 4) estimation strategies. Recall that all forecasting models, including lags and parameters, are re-estimated at each point in time, prior to the construction of each new prediction. Panels (a) - (d) reports results for 1, 3, 6 and 12 month ahead predictions, respectively. In each panel, first row reports the MSFE of our AR(SIC) model, and all other rows report MSFEs relative to the AR(SIC) value. Thus, entries greater than unity imply MSFEs greater than that of our AR(SIC) model, etc. Bold entries denote lowest MSFEs for a particular target variable.

Notice that in Panel (a) of Table 3, every forecast method yields a lower MSFE than the AR(SIC) model, when predicting the unemployment rate (UR). This result hold for all forecast horizons (see Panels (a)-(d) of the table). Moreover, for most variables, there are various models that have lower point MSFEs that the AR(SIC) model, regardless of forecast horizon. The exception is CPI when $h = 1$ and $h = 12$. Additionally, note that there are no models that uniformly yield lowest MSFEs, across both forecast horizons and variables. However, various models perform quite well, including component-wise boosting. This suggests that models that incorporate common factors constructed using diffusion index methodology offer a convenient way to filter the information contained in large-scale economic datasets. We also find that forecasts constructed as a simple average of all individual model based forecasts also yields the MSFE-best model for various variables and forecast horizons. This result is not surprising, given the large body of research establishing the usefulness of such forecasts in empirical settings. In addition, comparison of the results in Tables 3 and 4 suggests that there is little to choose between using rolling versus recursive estimation strategies when constructing real-time prediction models, although models constructed using recursive strategies appear to perform better more frequently, when comparing point MSFEs. Finally, note that Diebold Mariano (DM: 1995) test statistics are reported in Table 5 for point MSFEs in Table 3, and in Table 6 for point MSFEs in Table 4. As discussed above, these test statistics have a standard normal distribution under the null hypothesis of equal predictive accuracy. In all cases, we set the AR(SIC) to be the benchmark model. Thus, values that are significant and negative in these tables indicate that the model listed in the first column of the tables is predictively more accurate than the AR(SIC) model (see previous section for further details). As should be expected, given that the different forecasting models variously outperform and do not outperform the AR(SIC) when comparing point MSFEs, our DM test statistics are both negative and positive. More importantly, notice that the null of equal predictive accuracy is rejected in many cases (e.g. statistic values are great than 1.96 (5% level) or 1.67 (10% level) in absolute value). In summary, we find that there are many instances where our more sophisticated forecasting models yield predictions that are significantly more accurate than those of benchmark linear model, based on examination of DM predictive accuracy test results, such as when forecasting unemployment, money growth and GDP growth. Notice also that the AR(SIC) performs relatively better as the forecast horizon is increased, which should be expected, given that long run forecasts of economic time series are notoriously difficult to accurately construct, in which case more parsimonious models such as our linear benchmark tend to increase in relative accuracy.

Another observation of interest is that in most cases the MSFE-best model does not change as we move from recursive to rolling estimation strategies (compare Tables 3 and 4). This result is not particularly surprising given that, as already noted, there is really very little to choose between the estimation strategies.

There are many additional results that are apparent upon examination of the tables, but that are less relevant to the overall conclusion of this paper that our sophisticated models perform quite well against simple linear alternatives. For example, note that when forecasting the Standard & Poors 500 index (SNP), retail sales (RS) and advanced durable goods shipments, new orders and unfilled orders (DSNU), the AR(SIC) model is not often beaten, expect for selected horizons (e.g. Boosting for SNP using recursive samples when, $h = 3$, ARX(SIC) for RS using rolling samples when $h = 6$ and NNG for DSNU using rolling samples when $h = 1$). Additionally, when predicting IPX, no method yields lower point MSFE than the mean forecast except when $h = 3$ using recursive samples, in which case the ARX(SIC) is MSFE-best. NNG is MSFE-better than all other models for multi-step ahead forecasting of BSI. Finally, boosting performs well for various variables, including IPX, M2 and PPI.

7 Concluding Remarks

We present the results of a "horse-race" in which mean-square-forecast-error "best" models are selected, in the context of a variety of model specification methods, forecast horizons, sample periods, and "target variables" to be predicted. In addition to pure common factor prediction models, the forecast model specification methods that we analyze include bagging, boosting, Bayesian model averaging, ridge regression, least angle regression, elastic net and non-negative garotte as well as univariate autoregressive and autoregressive-exogenous model as benchmarks. In order to assess predictive accuracy based on the use of forecasting models constructed using these various methods, we construct Diebold-Mariano (1995) predictive accuracy tests. For a number of target variables, we find that various of these models, and in particular component-wise boosting, perform better than benchmark linear autoregressive forecasting models constructed using only observable variables, hence suggesting that the diffusion index methodology based models offer a convenient way to filter large-scale economic datasets prior to their use in forecast model construction. We also find that forecasts constructed as a simple average of all individual model based forecasts yield the MSFE-best model for various variables and forecast horizons. This result is not surprising, given the large body of research establishing the usefulness of such forecasts in empirical settings.

What is perhaps interesting is that we do find that there are many variable - forecast horizon combinations for which neither the linear not the mean-forecast models are MSFE-best. In addition, we find that there is little to choose between using rolling versus recursive estimation strategies when constructing real-time prediction models, although models constructed using recursive strategies appear to perform better more frequently, when comparing point MSFEs. Finally, we find that there are many instances where our more sophisticated forecasting models yield predictions that are significantly more accurate than those of benchmark linear models, based on examination of Diebold and Mariano (1995) predictive accuracy test results, such as when forecasting unemployment, money growth and GDP growth.

References

Armah, N. A. and Swanson, N. R. (2008). Seeing inside the black box: Using diffusion index methodology to construct factor proxies in large scale macroeconomic time series environments. Working Papers 08-25, Federal Reserve Bank of Philadelphia.

Artis, M. J., Banerjee, A., and Marcellino, M. (2002). Factor forecasts for the uk. CEPR Discussion Papers 3119, C.E.P.R. Discussion Papers.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2008a). Boosting diffusion indices. In *Journal of Applied Econometrics*.

Bai, J. and Ng, S. (2008b). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Boivin, J. and Ng, S. (2005). Understanding and comparing factor-based forecasts. *International Journal of Central Banking*, 1(3):117–152.

Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, 132(1):169–194.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Breiman, L. (1996). Bagging predictors. In *Machine Learning*, pages 123–140.

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

Bühlmann, P. and Yu, B. (2003). Boosting with the l_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339.

Chow, G. C. and Lin, A.-l. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The Review of Economics and Statistics*, 53(4):372–75.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583.

Clyde, M. (1999). Bayesian model averaging and model search strategies. In J. M. Bernardo, J. O. Berger, A. P. D. and Smith, A., editors, *Bayesian Statistics 6*, pages 157–185. Oxford University Press.

Connor, G. and Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory : A new framework for analysis. *Journal of Financial Economics*, 15(3):373–394.

Connor, G. and Korajczyk, R. A. (1988). Risk and return in an equilibrium apt : Application of a new test methodology. *Journal of Financial Economics*, 21(2):255–289.

Connor, G. and Korajczyk, R. A. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance*, 48(4):1263–91.

Diebold, F. X. and Lopez, J. A. (1996). Forecast evaluation and combination. NBER Technical Working Papers 0192, National Bureau of Economic Research, Inc.

Ding, A. A. and Hwang, J. T. G. (1999). Prediction intervals, factor analysis models, and high-dimensional empirical linear prediction. *Journal of the American Statistical Association*, 94(446):446–455.

Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

Fernandez, C., Ley, E., and Steel, M. F. J. (2001a). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Fernandez, C., Ley, E., and Steel, M. F. J. (2001b). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576.

Figlewski, S. (1983). Optimal price forecasting using survey data. *The Review of Economics and Statistics*, 65(1):13–21.

Figlewski, S. and Urich, T. (1983). Optimal aggregation of money supply forecasts: Accuracy, profitability and market efficiency. *Journal of Finance*, 38(3):695–710.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100:830–840.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Gelper, S. and Croux, C. (2008). Least angle regression for time series forecasting with many predictors, working paper. Technical report, Katholieke Universiteit Leuven.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14:382–417.

Inoue, A. and Kilian, L. (2005). How useful is bagging in forecasting economic time series? a case study of us cpi inflation. CEPR Discussion Papers 5304, Centre for Economic Policy Research.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis (Fifth Edition)*. Upper Saddle River, NJ: Prentice Hall.

Kitchen, J. and Monaco, R. (2003). Real- time forecastingin practiee; the u.s. treasury staff's real-time gdp forecast system. *Business Economics*, 38(4):10–19.

Koop, G. and Potter, S. (2004). Forecasting in dynamic factor models using bayesian model averaging. *Econometrics Journal*, 7(2):550–565.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270.

Min, C.-k. 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(1-2):89–118.

Newbold, P. and Harvey, D. (2002). Forecast combination and encompassing. In Clements, M. and Hendry, D., editors, *A Companion to Economic Forecasting*, pages 268–283. Blackwell Press: Oxford.

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.

Ravazzolo, F., Paap, R., van Dijk, D., and Franses, P. H. (2008). *Bayesian Model Averaging in the Presence of Strutural Breaks*, chapter 15. Frontier of Economics and Globalization.

Ridgeway, G., Madigan, D., and Richardson, T. (1999). Boosting methodology for regression problems. In *The Seventh International Workshop on Artificial Intelligence and Statistics (Uncertainty '99*, pages 152–161. Morgan Kaufmann.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227.

Shrestha, D. L. and Solomatine, D. P. (2006). Experiments with adaboost.rt, an improved boosting scheme for regression. *Neural Computation*, 18(7):1678–1710.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293–335.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179.

Stock, J. H. and Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, 41(3):788–829.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Stock, J. H. and Watson, M. W. (2005a). An empirical comparison of methods for forecasting using many predictors, manuscript. Technical report, Harvard University and Princeton University.

Stock, J. H. and Watson, M. W. (2005b). Implications of dynamic factor models for var analysis. NBER Working Papers 11467, National Bureau of Economic Research, Inc.

Stock, J. H. and Watson, M. W. (2006). Forecasting with many predictors. In Elliott, G., Granger, C., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 10, pages 515–554. Elsevier.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Timmermann, A. G. (2005). Forecast combinations. CEPR Discussion Papers 5361, C.E.P.R. Discussion Papers.

Yuan, M. (2007). Nonnegative garrote component selection in functional anova models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 660–666. JMLR Workshop and Conference Proceedings.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society*, 69(2):143–161.

Zellner (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions,. In Goel, P. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*. Armsterdam: North-Holland.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal Of The Royal Statistical Society Series B*, 67(2):301–320.

Table 1: List of Forecasted Variables

Series	Abbreviation	Y_{t+h}
Unemployment Rate	UR	$Z_{t+1}-Z_t$
Personal Income less transfer payments	PILT	$\ln(Z_{t+1}/Z_t)$
10-Year Treasury Bond	TB10Y	$Z_{t+1}-Z_t$
Consumer Price Index	CPI	$\ln(Z_{t+1}/Z_t)$
Producer Price Index	PPI	$\ln(Z_{t+1}/Z_t)$
Nonfarm Payroll Employment	NPE	$\ln(Z_{t+1}/Z_t)$
Housing Starts	HS	$\ln(Z_t)$
Industrial Production	IPX	$\ln(Z_{t+1}/Z_t)$
M2	M2	$\ln(Z_{t+1}/Z_t) - \Delta\ln(Z_t)$
S&P 500 Index	SNP	$\ln(Z_{t+1}/Z_t)$
Gross Domestic Product	GNP	$\ln(Z_{t+1}/Z_t)$
Retail Sales	RS	$\ln(Z_{t+1}/Z_t)$
Business Sales and Inventory	BSI	$\ln(Z_{t+1}/Z_t)$
Advanced Durable Goods Shipments, New Orders and Unfilled Orders	DSNU	$\ln(Z_{t+1}/Z_t)$

Table 2: List of forecasting methods

Method	Description
AR(SIC)	Autoregressove model with lags selected by SIC
ARX	Autoregressive-exogenous model
Combined-ADL	Combined autoregressive distributed lag model
FAAR	Factor Augmented Autoregressive model using OLS
PCR	Principal Component regression only with factors
Bagging	Bagging with shrinkage representation using $c = 1.96$
Boosting	Component Boosting with $M = 50$
BMA($1/T$)	Bayesian Model averaging with g-prior $= 1/T$
BMA($1/N^2$)	Bayesian Model averaging with g-prior $= 1/N^2$
Ridge	Ridge regression
LARS	Least Angle Regression
EN	Elastic Net
NNG	Non-negative garotte
Mean	Mean of every thirteen forecasted

Table 3: Mean Square Forecast Errors Using Recursive Samples and Flexible Lag Structures

a. Recursive sample, h = 1

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
AR(SIC)	12.647	1.992	40.439	**0.384**	1.798	0.206	2.420	2.849	0.465	75.005	0.401	0.937	0.192	1.572
ARX(SIC)	0.887	**0.834**	1.055	1.116	1.066	0.984	0.932	0.902	0.973	1.080	**0.754**	1.165	0.922	1.148
Combined-ADL	0.966	0.972	0.993	1.001	**0.998**	0.971	0.980	0.974	0.978	0.991	0.962	**0.954**	0.979	0.986
FAAR	0.816	0.981	0.967	1.091	1.028	0.940	0.909	0.920	0.893	0.993	0.892	1.157	1.044	1.072
PCR	0.859	1.046	1.092	1.157	1.438	2.007	3.072	0.997	1.026	1.021	2.532	1.198	1.701	1.099
Bagging	0.849	1.071	1.431	1.091	1.105	1.037	1.495	1.183	1.012	1.058	0.899	1.050	**0.887**	1.162
C-Boosting	0.819	1.022	0.970	1.012	1.053	1.510	0.949	0.969	0.893	0.982	1.491	1.018	1.452	1.047
BMA(1/n)	0.807	1.021	0.973	1.015	1.012	1.215	0.919	0.980	0.907	0.991	0.931	1.143	1.299	1.104
BMA($1/k^2$)	**0.806**	1.021	0.970	1.017	1.013	1.221	0.922	0.976	0.901	0.989	0.928	1.157	1.284	1.098
Ridge	0.823	0.971	0.965	1.075	1.032	**0.930**	0.909	0.910	**0.888**	0.976	0.885	1.063	1.016	1.025
LARS	0.920	0.999	0.977	1.080	1.029	0.939	0.931	0.908	0.909	0.992	0.886	1.149	1.031	1.025
EN	0.922	1.001	0.982	1.087	1.036	0.948	0.938	0.908	0.912	0.999	0.897	1.167	1.041	1.040
NNG	0.967	1.002	0.995	1.016	1.000	0.985	0.970	0.971	0.985	0.975	0.969	1.062	1.016	0.989
Mean	0.833	0.916	**0.959**	1.019	1.013	0.940	**0.875**	**0.879**	0.890	**0.961**	0.829	1.032	0.984	**0.978**

b. Recursive sample , h = 3

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
AR(SIC)	12.587	2.149	47.161	0.438	2.645	0.264	4.639	3.251	0.523	81.160	0.886	0.927	0.179	1.300
ARX(SIC)	0.932	**0.852**	1.016	1.000	1.000	**0.788**	1.000	**0.825**	1.067	1.074	1.000	1.000	1.142	1.359
Combined-ADL	0.985	0.973	1.005	1.002	1.001	0.964	0.961	0.970	0.997	0.996	0.960	**0.987**	0.989	**0.994**
FAAR	0.961	1.004	1.054	1.055	1.033	0.887	**0.763**	0.915	0.978	1.031	0.883	1.110	1.018	1.102
PCR	0.967	0.992	1.036	1.033	1.014	1.571	1.926	0.933	0.978	1.024	1.206	1.165	1.869	1.272
Bagging	0.981	0.986	1.040	1.038	1.018	1.087	1.274	1.335	**0.956**	**0.976**	0.895	1.023	0.981	1.218
C-Boosting	0.937	0.967	**0.978**	**0.986**	**0.992**	1.319	0.802	0.930	0.983	1.004	1.109	1.039	1.570	1.092
BMA(1/n)	0.939	0.971	0.992	0.987	0.996	1.148	0.766	0.921	0.984	1.017	1.144	1.041	1.389	1.057
BMA($1/k^2$)	0.939	0.972	1.001	0.994	0.995	1.128	0.776	0.924	0.980	1.016	1.147	1.044	1.348	1.063
Ridge	0.943	0.984	1.014	1.008	1.004	0.872	0.763	0.897	0.970	1.004	**0.875**	1.065	1.004	1.087
LARS	0.969	1.011	1.178	1.170	1.228	0.883	0.763	0.891	1.051	1.071	1.458	1.213	0.979	1.884
EN	0.983	1.016	1.193	1.177	1.235	0.902	0.776	0.903	1.056	1.083	1.475	1.224	0.986	1.905
NNG	0.960	1.008	1.133	1.144	1.480	0.976	0.958	0.983	1.065	1.046	1.504	1.109	**0.937**	1.779
Mean	**0.910**	0.911	1.008	1.017	1.011	0.852	0.773	0.850	0.969	0.999	0.892	1.039	1.011	1.155

Table 3 continued
c. Recursive sample , h = 6

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
AR(SIC)	13.213	2.006	46.479	0.429	2.727	0.348	9.020	3.446	0.519	81.126	0.851	0.893	0.207	1.422
ARX(SIC)	1.103	**0.960**	1.027	1.000	1.000	0.947	1.000	1.029	1.052	1.059	1.000	1.000	0.993	1.145
Combined-ADL	0.984	0.978	**0.999**	0.999	1.003	0.941	0.959	0.963	0.998	**0.997**	0.977	**0.999**	0.987	**0.982**
FAAR	0.945	1.052	1.035	1.046	1.039	0.806	0.736	0.948	1.035	1.033	0.978	1.022	1.087	1.160
PCR	0.936	1.115	1.021	1.044	1.004	1.249	1.260	0.967	1.035	1.032	1.321	1.138	1.615	1.283
Bagging	0.926	1.034	1.045	1.035	1.040	0.971	0.960	1.167	1.036	1.028	**0.952**	1.101	1.030	1.101
C-Boosting	0.923	1.067	0.996	1.011	0.981	1.167	0.748	0.923	**0.997**	0.998	1.209	1.054	1.422	1.034
BMA(1/n)	0.928	1.083	0.998	1.001	**0.979**	1.022	0.740	0.971	1.007	1.001	1.192	1.077	1.351	1.018
BMA($1/k^2$)	0.927	1.085	1.002	1.006	0.984	1.011	0.740	0.971	1.010	1.004	1.194	1.080	1.339	1.024
Ridge	0.931	1.033	1.009	1.010	1.001	0.784	0.733	0.930	1.011	1.008	0.963	1.023	1.061	1.113
LARS	0.990	1.043	1.074	1.188	1.274	0.803	**0.711**	0.936	1.117	1.097	1.366	1.172	1.059	1.902
EN	0.995	1.052	1.081	1.198	1.280	0.819	0.724	0.950	1.121	1.107	1.375	1.183	1.068	1.925
NNG	1.014	1.007	1.090	1.133	1.354	0.990	1.008	1.062	1.092	1.049	1.397	1.137	**0.935**	1.706
Mean	**0.913**	0.966	**0.993**	1.018	1.006	**0.770**	0.718	**0.880**	1.002	0.999	0.982	1.027	1.004	1.161

d. Recursive sample , h = 12

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
AR(SIC)	13.609	2.029	45.775	**0.410**	2.881	0.432	16.996	3.408	0.534	80.688	0.943	0.816	0.232	1.640
ARX(SIC)	5.575	0.976	1.003	1.001	1.000	0.921	1.000	1.027	1.045	1.076	1.000	1.000	1.000	1.000
Combined-ADL	0.989	0.981	1.001	1.001	0.997	0.923	0.971	0.967	1.000	**0.999**	0.966	**0.982**	0.981	**0.980**
FAAR	0.966	0.979	1.029	1.039	1.072	0.772	0.772	0.919	1.052	1.049	0.916	1.161	1.002	1.082
PCR	0.965	1.018	1.017	1.091	1.048	1.026	0.884	0.940	1.031	1.049	1.176	1.257	1.554	1.087
Bagging	0.970	1.158	1.022	1.084	1.026	0.823	0.817	1.381	1.029	1.023	**0.887**	1.237	0.952	0.982
C-Boosting	**0.942**	1.011	1.002	1.034	**0.986**	1.007	0.756	0.928	0.994	1.005	1.158	1.216	1.421	1.015
BMA(1/n)	0.951	0.983	**0.993**	1.027	0.993	1.049	0.767	0.933	**0.990**	1.008	1.236	1.219	1.543	1.029
BMA($1/k^2$)	0.950	0.967	0.995	1.025	0.998	1.043	0.766	0.930	0.995	1.014	1.232	1.230	1.564	1.030
Ridge	0.947	0.960	1.006	1.034	1.018	0.753	0.758	0.900	1.009	1.018	0.898	1.102	1.037	1.040
LARS	1.044	0.970	1.142	1.018	1.158	0.827	0.766	0.927	1.106	1.069	1.343	1.284	0.951	1.335
EN	1.064	0.975	1.146	1.022	1.161	0.854	0.800	0.957	1.111	1.074	1.365	1.315	0.951	1.333
NNG	1.102	1.033	1.151	1.010	1.203	1.156	1.164	1.114	1.078	1.044	1.417	1.175	**0.886**	1.294
Mean	0.957	**0.916**	1.003	1.006	0.997	**0.746**	**0.736**	**0.883**	1.007	1.010	0.946	1.117	1.019	1.053

NOTES: Mean Square Forecast Errors (MSFEs) are reported for target variables being forecasted, as well as for different forecasting methods. In the first row, AR(SIC) reports actual MSFEs. All other rows of numerical entries are the ratio of the AR(SIC) MSFE to the MSFE of the model constructed using the alternative method reported in the first column. For a decription of the variables used, see Table 2. Results are reported for 1,3,6 and 12 step ahead forecasts. Recursive sample begin in January 1960 and end in December 2007 for the first 11 variables, and begin in January 1992 for the last 3 variables. the first in-sample period used in model estimation includes 10 years of data. GDP is extrapolated to monthly using the method of Chow and Lin (1971).

Table 4: Mean Square Forecast Errors Using Rolling Samples and Flexible Lag Structures

a. Rolling sample with 10 years , h = 1

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
AR(SIC)	12.780	2.243	40.839	0.372	1.850	0.215	2.467	2.952	0.461	75.311	0.391	0.938	0.194	1.562
ARX(SIC)	1.002	1.261	1.183	1.120	1.197	1.027	1.688	0.984	1.228	1.257	0.965	1.187	1.133	1.259
Combined-ADL	0.961	0.964	0.989	**0.997**	**0.998**	0.975	0.985	0.970	0.983	0.993	0.960	**0.939**	0.978	0.983
FAAR	0.809	0.987	1.017	1.098	1.089	1.009	0.987	0.963	0.949	1.040	0.981	1.140	1.046	1.133
PCR	0.845	0.943	1.175	1.191	1.420	1.908	3.709	0.969	1.089	1.080	2.703	1.209	1.587	1.183
Bagging	0.903	0.970	1.558	1.517	1.453	1.042	1.807	1.064	1.205	1.080	1.695	1.089	**0.921**	1.253
C-Boosting	0.800	0.897	1.012	1.010	1.060	1.534	1.373	0.941	0.961	0.997	1.469	1.017	1.345	1.015
BMA(1/n)	0.795	0.917	1.022	1.017	1.039	1.274	1.010	0.961	0.967	1.006	1.062	1.104	1.230	1.089
BMA(1/k^2)	**0.790**	0.921	1.015	1.026	1.042	1.266	1.044	0.963	0.962	1.003	1.064	1.118	1.218	1.092
Ridge	0.812	0.953	0.999	1.056	1.079	0.988	0.990	0.935	0.938	0.996	0.964	1.061	1.015	1.070
LARS	0.954	1.006	1.024	1.090	1.093	0.996	0.957	0.961	0.961	1.032	0.950	1.135	1.030	1.079
EN	0.959	1.011	1.031	1.096	1.109	1.006	0.962	0.967	0.968	1.042	0.966	1.149	1.040	1.093
NNG	0.964	0.989	0.996	1.040	1.015	0.981	0.979	0.974	0.984	0.992	0.977	1.068	1.004	**0.956**
Mean	0.822	**0.882**	**0.971**	1.006	1.017	**0.955**	**0.931**	**0.882**	**0.928**	**0.987**	**0.906**	1.019	0.987	1.003

b. Rolling sample with 10 years , h = 3

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
AR(SIC)	12.643	2.254	47.418	**0.439**	2.649	0.266	4.652	3.277	0.515	81.814	0.977	0.932	0.181	1.307
ARX(SIC)	1.686	**0.882**	1.165	1.050	1.026	**0.764**	0.963	0.853	1.207	1.254	1.000	1.000	1.351	1.302
Combined-ADL	0.988	0.968	1.011	1.001	1.003	0.970	0.967	0.971	**0.997**	**0.993**	0.960	**0.981**	0.986	**0.994**
FAAR	0.995	1.034	1.135	1.051	1.075	0.906	0.815	0.934	1.044	1.077	0.957	1.088	1.030	1.104
PCR	1.014	0.985	1.100	1.027	1.045	1.579	2.232	0.947	1.040	1.051	1.141	1.117	1.625	1.318
Bagging	1.022	0.953	1.130	1.060	1.039	0.981	1.357	1.020	1.051	1.018	0.883	1.088	0.960	1.071
C-Boosting	0.930	0.931	**0.977**	1.003	**0.983**	1.381	1.031	0.914	1.005	1.001	1.053	1.037	1.501	1.069
BMA(1/n)	0.946	0.959	1.005	1.001	0.989	1.220	0.814	0.915	1.006	1.005	1.099	1.032	1.368	1.050
BMA(1/k^2)	0.944	0.959	1.010	1.006	0.991	1.207	0.836	0.915	1.005	1.013	1.098	1.036	1.353	1.051
Ridge	0.965	0.978	1.033	**1.000**	1.009	0.885	0.816	0.903	1.005	1.016	0.936	1.043	0.999	1.091
LARS	0.995	1.025	1.258	1.320	1.363	0.885	**0.778**	0.945	1.246	1.096	1.390	1.192	0.968	1.955
EN	1.004	1.031	1.280	1.335	1.364	0.886	0.784	0.933	1.260	1.103	1.399	1.203	0.972	1.981
NNG	0.981	0.975	1.156	1.276	1.479	0.973	0.938	1.002	1.227	1.057	1.294	1.130	0.926	1.771
Mean	**0.913**	0.904	1.039	1.011	1.015	0.848	0.790	**0.846**	1.009	1.002	**0.888**	1.036	1.002	1.119

Table 4 continued
c. Rolling sample with 10 years , h = 6

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
AR(SIC)	13.61441	2.155	46.813	0.441	2.722	0.349	9.062	3.489	0.533	81.554	0.883	0.922	0.207	1.448
ARX(SIC)	1.390	1.048	1.177	1.055	1.030	0.844	1.000	0.970	1.195	1.139	1.000	**0.985**	1.049	1.567
Combined-ADL	0.981	0.978	1.001	1.001	1.006	0.949	0.957	0.968	1.000	**0.997**	0.981	0.995	0.981	**0.981**
FAAR	0.977	1.119	1.073	1.094	1.074	0.840	0.813	1.004	1.072	1.076	1.008	1.029	1.080	1.178
PCR	0.944	1.093	1.061	1.049	1.030	1.363	1.408	1.025	1.043	1.065	1.320	1.060	1.314	1.255
Bagging	1.020	1.003	1.098	1.076	1.112	0.919	0.959	0.996	1.071	1.043	1.094	1.191	0.919	1.046
C-Boosting	**0.925**	1.039	**0.999**	1.008	0.999	1.178	0.820	0.943	**0.985**	1.012	1.193	1.017	1.317	0.991
BMA(1/n)	0.938	1.106	1.002	**0.994**	**0.994**	1.058	0.780	0.988	**0.985**	1.013	1.114	1.041	1.300	1.041
BMA($1/k^2$)	0.935	1.108	1.010	1.002	0.995	1.054	0.794	0.990	0.988	1.019	1.115	1.042	1.291	1.047
Ridge	0.941	1.062	1.018	1.017	1.007	0.825	0.813	0.972	1.003	1.025	1.005	1.012	1.030	1.120
LARS	0.994	1.067	1.116	1.327	1.337	0.811	0.764	1.016	1.223	1.121	1.298	1.121	1.079	1.834
EN	1.008	1.072	1.129	1.336	1.340	0.832	0.779	1.029	1.231	1.130	1.308	1.123	1.101	1.860
NNG	0.992	0.977	1.106	1.213	1.376	0.985	0.996	1.096	1.188	1.077	1.275	1.131	0.916	1.666
Mean	0.929	**0.975**	1.005	1.028	1.008	**0.791**	**0.735**	**0.905**	1.015	1.007	**0.981**	1.008	0.971	1.151

d. Rolling sample with 10 years , h = 12

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
AR(SIC)	13.764	2.049	45.989	0.413	2.959	0.432	16.979	3.409	0.547	82.170	0.979	0.837	0.242	1.646
ARX(SIC)	8.074	1.029	1.081	1.080	1.004	1.007	1.000	1.104	1.157	1.276	1.000	1.000	1.000	1.002
Combined-ADL	0.988	0.979	1.002	1.004	0.994	0.939	0.963	0.974	1.005	**0.996**	0.969	**0.974**	0.965	**0.976**
FAAR	0.983	1.004	1.070	1.058	1.084	0.776	0.884	0.926	1.117	1.080	0.967	1.130	0.979	1.073
PCR	0.971	1.049	1.047	1.126	1.026	1.072	0.939	0.958	1.057	1.066	1.147	1.198	1.329	1.078
Bagging	0.998	1.138	1.020	1.224	1.145	0.790	0.864	1.094	1.166	1.033	0.973	1.299	0.918	1.045
C-Boosting	0.953	1.006	**0.983**	1.006	**0.980**	1.023	0.810	0.938	**0.977**	0.999	1.109	1.144	1.282	1.054
BMA(1/n)	0.944	0.989	0.988	1.013	1.004	1.020	0.821	0.941	1.002	1.003	1.114	1.160	1.373	1.048
BMA($1/k^2$)	**0.943**	0.983	0.989	1.014	1.010	1.023	0.823	0.939	1.005	1.013	1.117	1.167	1.375	1.051
Ridge	0.944	0.963	1.010	1.040	1.008	**0.773**	0.858	0.912	1.024	1.020	0.944	1.065	1.008	1.032
LARS	1.059	1.047	1.205	1.023	1.171	0.818	0.871	0.933	1.268	1.101	1.248	1.252	0.951	1.352
EN	1.072	1.054	1.214	1.030	1.179	0.840	0.912	0.942	1.284	1.108	1.263	1.279	0.968	1.368
NNG	1.101	1.077	1.165	1.003	1.176	1.160	1.158	1.151	1.182	1.043	1.273	1.200	**0.849**	1.346
Mean	0.977	**0.937**	1.008	**0.993**	0.989	0.792	**0.764**	**0.902**	1.016	1.014	**0.931**	1.092	0.969	1.061

NOTES: See notes to Table 3.

Table 5: Diebold and Mariano Test Results

a. Recursive sample , h = 1

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
ARX(SIC)	1.586	1.772	-0.918	-0.631	-1.057	0.132	0.702	1.018	0.383	-1.104	3.247	-0.671	0.484	-0.961
Combined-ADL	4.620	3.460	2.716	-0.224	0.628	3.687	3.496	4.852	3.961	2.775	5.471	2.502	1.987	0.958
FAAR	2.695	0.687	0.818	-0.954	-0.840	1.056	1.254	1.481	2.309	0.138	1.375	-1.741	-0.503	-0.594
PCR	2.017	-0.516	-0.944	-1.478	-2.985	-6.600	-6.742	0.044	-0.319	-0.316	-4.957	-1.629	-2.322	-0.572
Bagging	2.297	-0.736	-2.360	-1.211	-1.898	-0.438	-3.164	-1.526	-0.183	-0.845	1.490	-0.777	0.749	-0.611
C-Boosting	2.969	-0.265	0.832	-0.550	-0.930	-4.582	0.918	0.567	2.809	0.400	-3.880	-0.393	-1.859	-0.665
BMA(1/n)	2.845	-0.242	0.764	-1.031	-0.616	-2.730	1.313	0.339	2.164	0.168	0.858	-1.128	-1.512	-0.925
BMA(1/k^2)	2.908	-0.251	0.848	-0.950	-0.655	-2.772	1.232	0.389	2.298	0.213	0.894	-1.205	-1.459	-0.831
Ridge	2.871	1.048	0.939	-1.045	-0.846	1.357	1.249	1.782	2.600	0.542	1.476	-1.276	-0.218	-0.258
LARS	1.195	0.036	0.596	-1.265	-0.861	1.172	1.029	1.732	2.081	0.172	1.500	-1.729	-0.370	-0.221
EN	1.087	-0.044	0.422	-1.259	-0.971	0.960	0.878	1.671	1.917	0.029	1.279	-1.756	-0.451	-0.330
NNG	2.679	-0.177	0.676	-0.859	0.033	0.882	2.751	2.314	2.666	2.001	3.357	-1.145	-0.667	0.545
Mean	3.279	1.960	1.201	-0.552	-0.474	1.253	2.156	2.827	3.169	0.999	2.752	-0.552	0.188	0.290

b. Recursive sample, h = 3

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
ARX(SIC)	0.666	2.582	-1.397	0.330	-0.048	1.833	-0.699	1.641	-1.041	-1.504	0.254	0.402	-0.747	-1.422
Combined-ADL	2.468	4.501	-1.157	-0.834	-0.563	3.856	4.841	4.047	1.049	1.293	4.595	0.768	1.313	0.544
FAAR	0.646	-0.140	-0.986	-1.594	-1.295	1.268	2.336	1.163	0.487	-0.731	1.742	-1.232	-0.193	-0.760
PCR	0.474	0.126	-0.536	-0.837	-0.357	-4.088	-3.892	0.750	0.481	-0.561	-1.771	-1.396	-2.701	-1.274
Bagging	0.322	0.188	-0.975	-1.103	-0.597	-0.867	-2.134	-2.610	1.410	0.631	1.881	-0.155	0.140	-1.083
C-Boosting	1.341	0.526	1.263	0.807	0.436	-3.145	2.232	1.187	0.709	-0.143	-1.085	-0.466	-2.261	-0.818
BMA(1/n)	1.121	0.447	0.439	0.808	0.276	-1.563	2.501	1.126	0.595	-0.459	-1.269	-0.490	-1.897	-0.502
BMA(1/k^2)	1.123	0.440	-0.047	0.349	0.301	-1.392	2.405	1.079	0.677	-0.419	-1.306	-0.531	-1.791	-0.542
Ridge	1.095	0.627	-0.418	-0.368	-0.199	1.565	2.323	1.479	0.903	-0.131	1.871	-1.084	-0.055	-0.688
LARS	0.450	-0.306	-1.193	-2.943	-2.603	1.316	2.446	1.427	-0.802	-1.297	-2.697	-1.466	0.226	-2.450
EN	0.237	-0.409	-1.209	-2.951	-2.556	1.006	2.210	1.207	-0.868	-1.454	-2.724	-1.487	0.144	-2.403
NNG	1.493	-0.365	-1.300	-2.327	-3.293	1.053	2.133	0.869	-1.428	-1.278	-3.354	-0.885	1.016	-2.571
Mean	1.924	2.735	-0.184	-0.975	-0.462	2.413	3.123	2.772	1.124	0.034	1.742	-0.547	-0.140	-1.139

Table 5 continued

c. Recursive sample, h = 6

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
ARX(SIC)	-0.870	0.701	-1.082	-0.031	-0.264	0.698	0.266	-0.258	-1.650	-1.347	0.933	-0.537	0.040	-0.953
Combined-ADL	2.696	3.783	0.250	0.394	-0.881	5.154	4.407	5.089	0.921	1.252	4.151	0.112	1.283	1.147
FAAR	0.672	-0.920	-0.883	-1.582	-1.351	1.512	2.440	0.494	-1.518	-0.836	0.369	-0.436	-0.797	-1.108
PCR	0.816	-1.380	-0.439	-1.404	-0.108	-1.626	-1.484	0.276	-1.055	-0.729	-3.413	-0.818	-2.286	-1.556
Bagging	1.092	-0.318	-0.675	-1.362	-1.634	0.235	0.435	-1.444	-1.270	-0.716	1.150	-0.467	-0.213	-0.604
C-Boosting	1.413	-0.913	0.193	-0.886	0.981	-1.426	2.645	0.844	0.206	0.073	-2.731	-0.441	-1.810	-0.427
BMA(1/n)	1.037	-1.107	0.118	-0.135	1.104	-0.170	2.486	0.238	-0.486	-0.036	-2.492	-0.630	-1.569	-0.225
BMA(1/k^2)	1.055	-1.132	-0.154	-0.523	0.864	-0.086	2.518	0.239	-0.795	-0.127	-2.494	-0.666	-1.547	-0.292
Ridge	1.032	-0.657	-0.349	-0.655	-0.050	1.790	2.471	0.701	-0.724	-0.272	0.675	-0.374	-0.662	-1.166
LARS	0.109	-0.743	-1.129	-2.697	-2.446	1.596	2.783	0.614	-1.728	-1.465	-3.175	-0.852	-0.513	-2.726
EN	0.057	-0.836	-1.150	-2.723	-2.568	1.348	2.499	0.446	-1.762	-1.544	-3.171	-0.887	-0.559	-2.737
NNG	-0.308	-0.413	-1.418	-2.325	-2.348	0.203	-0.165	-1.266	-1.423	-0.948	-4.000	-0.707	1.171	-2.539
Mean	1.488	0.792	0.427	-1.051	-0.203	2.744	3.694	1.548	-0.101	0.046	0.476	-0.266	-0.040	-1.706

d. Recursive sample, h = 12

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
ARX(SIC)	-0.723	1.483	-0.168	-0.912	0.209	1.808	-0.118	-0.398	-1.080	-2.004	0.104	0.782	0.019	0.412
Combined-ADL	2.429	4.138	-0.329	-0.978	1.334	5.610	4.721	4.998	-0.249	0.725	4.110	1.015	1.340	1.701
FAAR	0.502	0.460	-1.487	-2.447	-1.480	2.066	2.950	1.118	-2.036	-1.300	1.302	-1.067	-0.017	-1.185
PCR	0.490	-0.269	-0.785	-2.621	-0.988	-0.228	1.347	0.740	-1.105	-1.415	-1.712	-1.644	-2.926	-1.236
Bagging	0.529	-1.403	-0.524	-2.015	-0.791	1.612	2.245	-3.353	-1.035	-0.669	2.007	-1.559	0.557	0.195
C-Boosting	1.195	-0.187	-0.104	-1.996	0.638	-0.069	3.588	1.166	0.453	-0.305	-1.805	-1.995	-2.533	-0.306
BMA(1/n)	0.798	0.332	0.400	-1.903	0.229	-0.443	2.987	0.867	0.695	-0.509	-2.365	-1.881	-2.705	-0.589
BMA(1/k^2)	0.821	0.676	0.305	-2.024	0.056	-0.393	3.041	0.924	0.394	-0.673	-2.326	-1.893	-2.731	-0.633
Ridge	0.931	0.971	-0.461	-2.291	-0.594	2.417	3.255	1.456	-0.628	-0.908	1.658	-1.059	-0.462	-1.033
LARS	-0.499	0.408	-1.928	-0.426	-1.880	1.532	2.793	0.974	-2.407	-1.367	-3.210	-1.778	0.254	-1.863
EN	-0.691	0.328	-1.969	-0.522	-1.923	1.190	2.208	0.539	-2.444	-1.411	-3.255	-1.860	0.240	-1.844
NNG	-1.269	-0.626	-2.074	-0.239	-2.114	-1.964	-1.852	-1.756	-2.026	-1.192	-3.993	-1.782	0.909	-2.033
Mean	0.703	2.419	-0.139	-0.585	0.117	3.552	4.750	2.367	-0.473	-0.498	1.162	-1.331	-0.348	-1.351

NOTES: See notes to Table 3. We test the null hypothesis of equal predictive accuracy. Namely, we test whether each indididual set of forecasts from a given Method is different from those of our "benchmark" model (i.e. AR(SIC)). See Diebold and Mariano (1995) for complete details of the test of equal predictive accuracy that we report on. The statistic is distributed as a standard normal random variable, under the null, assuming that parameter estimation error vanishes asymptotically, and that the models are nonnested. Positive statistic values indicate that the MSFE of the AR(SIC) is higher, and so rejection of the null at a given level of confidence using standard normal critical values indicates that the alternative model is preferred, if the statistic is positive. The converse holds if the statistic is negative.

Table 6: Diebold and Mariano Test Results

a. Rolling sample with 10 years , h = 1

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
ARX(SIC)	-0.029	-0.446	-1.659	-1.101	-2.079	-0.209	-0.662	0.095	-1.554	-2.890	0.240	-0.719	-0.834	-1.147
Combined-ADL	4.800	3.270	2.581	0.581	0.548	2.947	3.172	4.936	3.135	1.847	5.807	2.583	2.029	1.013
FAAR	2.526	0.309	-0.230	-1.232	-1.668	-0.149	0.183	0.593	1.039	-0.619	0.214	-1.538	-0.541	-1.139
PCR	2.182	0.583	-1.416	-1.565	-2.531	-5.162	-8.191	0.432	-0.868	-1.073	-5.588	-1.503	-2.143	-1.014
Bagging	1.320	0.291	-2.139	-3.107	-2.799	-0.377	-4.035	-0.706	-1.969	-1.233	-4.007	-0.879	0.451	-0.916
C-Boosting	3.172	1.162	-0.231	-0.287	-0.968	-4.017	-3.281	1.074	1.026	0.073	-3.942	-0.497	-1.557	-0.192
BMA(1/n)	2.788	0.927	-0.379	-0.849	-1.293	-2.421	-0.172	0.633	0.975	-0.122	-0.746	-1.028	-1.260	-0.917
BMA($1/k^2$)	2.860	0.859	-0.266	-1.046	-1.315	-2.366	-0.717	0.603	1.076	-0.056	-0.776	-1.126	-1.213	-0.931
Ridge	2.839	1.022	0.017	-1.014	-1.414	0.200	0.138	1.157	1.468	0.098	0.421	-1.056	-0.201	-0.760
LARS	0.682	-0.181	-0.391	-1.500	-1.784	0.070	0.700	0.720	0.871	-0.574	0.651	-1.547	-0.369	-0.753
EN	0.557	-0.311	-0.502	-1.574	-1.906	-0.099	0.578	0.565	0.660	-0.696	0.410	-1.572	-0.469	-0.828
NNG	2.077	0.714	0.301	-0.920	-0.908	0.996	1.698	1.372	1.319	0.491	1.440	-1.051	-0.169	1.336
Mean	3.427	2.146	0.585	-0.187	-0.405	0.744	1.192	2.618	2.010	0.345	1.543	-0.377	0.144	-0.037

b. Rolling sample with 10 years , h = 3

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
ARX(SIC)	-0.747	1.537	-1.717	-1.635	-1.525	1.878	1.559	1.142	-2.306	-2.949	-0.115	0.219	-1.166	-0.909
Combined-ADL	1.364	3.849	-1.292	-0.324	-1.314	2.880	3.842	2.920	0.634	1.694	5.094	0.963	1.384	0.508
FAAR	0.071	-1.014	-1.731	-1.222	-2.100	0.999	1.832	0.838	-0.901	-1.443	0.564	-0.845	-0.270	-0.737
PCR	-0.157	0.209	-0.947	-0.507	-0.847	-3.855	-5.116	0.626	-0.693	-0.966	-1.243	-1.038	-2.291	-1.335
Bagging	-0.269	0.617	-1.384	-1.160	-0.805	0.166	-2.501	-0.249	-0.938	-0.373	1.892	-0.831	0.241	-0.307
C-Boosting	1.449	1.036	1.054	-0.098	0.590	-3.311	-0.275	1.306	-0.142	-0.025	-0.641	-0.498	-2.097	-0.684
BMA(1/n)	0.997	0.603	-0.148	-0.031	0.411	-2.015	1.984	1.081	-0.185	-0.144	-1.020	-0.470	-1.702	-0.648
BMA($1/k^2$)	1.036	0.614	-0.286	-0.151	0.328	-1.900	1.744	1.074	-0.170	-0.381	-1.004	-0.530	-1.662	-0.637
Ridge	0.611	0.686	-0.669	0.016	-0.392	1.346	1.812	1.413	-0.139	-0.409	0.984	-0.774	0.010	-0.671
LARS	0.069	-0.387	-1.217	-2.265	-2.947	1.128	2.313	0.645	-2.545	-1.460	-1.963	-1.236	0.310	-2.456
EN	-0.049	-0.475	-1.225	-2.342	-2.884	1.097	2.093	0.767	-2.599	-1.511	-2.022	-1.249	0.257	-2.439
NNG	0.525	0.475	-1.170	-1.949	-3.033	0.726	2.535	-0.051	-2.438	-1.236	-2.211	-0.959	1.034	-2.505
Mean	1.862	2.364	-0.584	-0.285	-0.429	2.154	2.926	2.730	-0.262	-0.050	2.013	-0.530	-0.027	-0.827

Table 6 continued

c. Rolling sample with 10 years , h = 6

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
ARX(SIC)	-2.013	-0.485	-2.602	-1.562	-1.270	1.415	-0.305	0.362	-4.495	-1.384	0.336	0.765	-0.229	-1.138
Combined-ADL	2.531	3.150	-0.277	-0.201	-1.406	4.890	4.536	4.288	0.071	0.895	2.832	0.365	1.522	1.069
FAAR	0.231	-1.512	-1.213	-1.090	-1.770	1.174	1.553	-0.038	-1.782	-1.719	-0.175	-0.561	-0.617	-1.076
PCR	0.606	-0.769	-0.908	-0.725	-0.590	-1.932	-1.994	-0.204	-0.996	-1.449	-3.766	-0.448	-1.474	-1.262
Bagging	-0.267	-0.019	-1.027	-1.099	-2.154	0.671	0.442	0.037	-1.629	-0.953	-1.568	-0.839	0.719	-0.203
C-Boosting	1.239	-0.474	0.027	-0.213	0.018	-1.307	1.427	0.686	0.761	-0.498	-2.860	-0.160	-1.573	0.127
BMA(1/n)	0.810	-1.362	-0.066	0.194	0.185	-0.388	2.010	0.111	0.819	-0.560	-1.870	-0.401	-1.407	-0.583
BMA($1/k^2$)	0.825	-1.388	-0.304	-0.055	0.171	-0.358	1.899	0.092	0.684	-0.738	-1.882	-0.441	-1.383	-0.641
Ridge	0.764	-1.298	-0.502	-0.441	-0.234	1.393	1.547	0.285	-0.151	-0.892	-0.131	-0.227	-0.307	-1.131
LARS	0.068	-0.809	-1.435	-2.438	-2.696	1.488	2.059	-0.141	-2.611	-2.067	-2.328	-0.638	-0.564	-2.453
EN	-0.078	-0.901	-1.472	-2.479	-2.756	1.221	1.836	-0.250	-2.686	-2.122	-2.313	-0.649	-0.665	-2.419
NNG	0.206	0.367	-1.446	-2.450	-2.381	0.317	0.089	-1.436	-2.305	-1.588	-2.627	-0.668	1.341	-2.395
Mean	1.100	0.444	-0.233	-0.632	-0.225	2.229	3.194	1.340	-0.511	-0.259	0.625	-0.086	0.311	-1.426

d. Rolling sample with 10 years , h = 12

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	RS	BSI	DSNU
ARX(SIC)	-0.737	-0.439	-2.643	-1.885	-0.154	-0.123	-0.065	-1.444	-1.730	-3.266	1.216	0.169	1.282	-0.185
Combined-ADL	2.291	3.322	-0.709	-1.706	1.742	5.645	5.201	4.518	-0.973	1.010	4.238	1.204	2.192	1.537
FAAR	0.290	-0.081	-2.160	-1.846	-1.329	2.480	1.287	1.065	-1.489	-2.266	0.567	-0.857	0.147	-0.953
PCR	0.520	-0.659	-1.341	-2.269	-0.343	-0.662	0.554	0.549	-0.932	-1.756	-1.678	-1.279	-2.194	-1.037
Bagging	0.047	-1.457	-0.325	-3.099	-1.830	2.536	1.648	-1.087	-2.347	-0.884	0.449	-1.979	0.795	-0.425
C-Boosting	1.369	-0.102	0.667	-0.286	0.505	-0.265	2.356	1.174	0.723	0.084	-1.518	-1.476	-2.089	-0.933
BMA(1/n)	1.402	0.183	0.540	-0.951	-0.085	-0.202	2.057	0.927	-0.062	-0.185	-1.530	-1.554	-2.193	-0.961
BMA($1/k^2$)	1.393	0.290	0.493	-1.033	-0.212	-0.240	2.059	0.935	-0.139	-0.733	-1.560	-1.536	-2.174	-1.023
Ridge	1.276	0.930	-0.511	-1.487	-0.198	2.876	1.673	1.488	-0.512	-0.904	1.129	-0.692	-0.087	-0.815
LARS	-0.820	-0.578	-2.403	-0.447	-1.694	1.888	1.312	0.959	-2.983	-2.015	-2.264	-1.491	0.225	-1.815
EN	-0.950	-0.644	-2.465	-0.579	-1.735	1.529	0.809	0.775	-3.119	-2.057	-2.312	-1.562	0.136	-1.820
NNG	-1.324	-1.291	-1.983	-0.061	-1.776	-1.927	-1.877	-1.947	-2.056	-1.112	-2.590	-1.688	1.142	-2.059
Mean	0.337	1.559	-0.312	0.450	0.318	3.398	3.915	2.189	-0.395	-0.692	1.747	-1.056	0.440	-1.532

NOTE : See notes for Table 5

Appendix. Description of Data
1 = level of the series
2 = first difference
3 = second deference
4 = log of the series
5 = first difference of the log
6 = second difference of log

Code	Short	Long Description	Tcode
001	IPS10	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX	5
002	IPS11	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL	5
003	IPS299	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS	5
004	IPS12	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS	5
005	IPS13	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS	5
006	IPS18	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS	5
007	IPS25	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT	5
008	IPS32	INDUSTRIAL PRODUCTION INDEX - MATERIALS	5
009	IPS34	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS	5
010	IPS38	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS	5
011	IPS43	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)	5
012	IPS307	INDUSTRIAL PRODUCTION INDEX - RESIDENTIAL UTILITIES	5
013	IPS306	INDUSTRIAL PRODUCTION INDEX - FUELS	5
014	A0M051	Personal income less transfer payments (AR, bil. chain 2000 $)	5
015	PMP	NAPM PRODUCTION INDEX (PERCENT)	1
016	LHEL	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)	2
017	LHELX	EMPLOYMENT: RATIO; HELP-WANTED ADS:NO. UNEMPLOYED CLF	2
018	LHEM	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)	5
019	LHNAG	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)	5
020	LHUR	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%,SA)	2
021	LHU680	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)	2
022	LHU5	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)	5
023	LHU14	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)	5
024	LHU15	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)	5
025	LHU26	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)	5
026	LHU27	UNEMPLOY.BY DURATION: PERSONS UNEMPL.27 WKS + (THOUS,SA)	5
027	CES002	EMPLOYEES ON NONFARM PAYROLLS - TOTAL PRIVATE	5
028	CES003	EMPLOYEES ON NONFARM PAYROLLS - GOODS-PRODUCING	5
029	CES006	EMPLOYEES ON NONFARM PAYROLLS - MINING	5
030	CES011	EMPLOYEES ON NONFARM PAYROLLS - CONSTRUCTION	5
031	CES015	EMPLOYEES ON NONFARM PAYROLLS - MANUFACTURING	5
032	CES017	EMPLOYEES ON NONFARM PAYROLLS - DURABLE GOODS	5
033	CES033	EMPLOYEES ON NONFARM PAYROLLS - NONDURABLE GOODS	5
034	CES046	EMPLOYEES ON NONFARM PAYROLLS - SERVICE-PROVIDING	5
035	CES048	EMPLOYEES ON NONFARM PAYROLLS - TRADE, TRANSPORTATION, AND UTILITIES	5
036	CES049	EMPLOYEES ON NONFARM PAYROLLS - WHOLESALE TRADE	5
037	CES053	EMPLOYEES ON NONFARM PAYROLLS - RETAIL TRADE	5
038	CES088	EMPLOYEES ON NONFARM PAYROLLS - FINANCIAL ACTIVITIES	5
039	CES140	EMPLOYEES ON NONFARM PAYROLLS - GOVERNMENT	5
040	CES151	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFAR	2
041	CES155	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFAR	2
042	PMEMP	NAPM EMPLOYMENT INDEX (PERCENT)	2
043	HSFR	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SA	4
044	HSNE	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.	4
045	HSMW	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.	4
046	HSSOU	HOUSING STARTS:SOUTH (THOUS.U.)S.A.	4

Appendix continued

Code	Short	Long Description	Tcode
047	HSWST	HOUSING STARTS:WEST (THOUS.U.)S.A.	4
048	HSBR	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)	4
049	HSBNE	HOUSES AUTHORIZED BY BUILD. PERMITS:NORTHEAST(THOU.U.)S.A	4
050	HSBMW	HOUSES AUTHORIZED BY BUILD. PERMITS:MIDWEST(THOU.U.)S.A.	4
051	HSBSOU	HOUSES AUTHORIZED BY BUILD. PERMITS:SOUTH(THOU.U.)S.A.	4
052	HSBWST	HOUSES AUTHORIZED BY BUILD. PERMITS:WEST(THOU.U.)S.A.	4
053	PMI	PURCHASING MANAGERS' INDEX (SA)	2
054	PMNO	NAPM NEW ORDERS INDEX (PERCENT)	2
055	PMDEL	NAPM VENDOR DELIVERIES INDEX (PERCENT)	2
056	PMNV	NAPM INVENTORIES INDEX (PERCENT)	2
057	FM1	MONEY STOCK: M1(CURR,TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL$,SA)	6
058	FM2	MONEY STOCK:M2(M1+O'NITE RPS,EURO$,G/P&B/D MMMFS&SAV&SM TIME DEP(BIL$,	6
059	FMFBA	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL$,SA)	6
060	FMRRA	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL$,SA)	6
061	FMRNBA	DEPOSITORY INST RESERVES:NONBORROWED,ADJ RES REQ CHGS(MIL$,SA)	6
062	CCINRV	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)	6
063	FSPCOM	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)	5
064	FSPIN	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)	5
065	FSDXP	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)	2
066	FSPXE	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)	5
067	FYFF	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)	2
068	FYGM3	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)	2
069	FYGM6	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)	2
070	FYGT1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)	2
071	FYGT10	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)	2
072	FYGT5	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)	2
073	FYAAAC	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)	2
074	FYBAAC	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)	2
075	FYAC	BOND YIELD: MOODY'S A CORPORATE (% PER ANNUM)	2
076	EXRUS	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)	5
077	EXRUK	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)	5
078	EXRCAN	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)	5
079	PWFSA	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)	6
080	PWFCSA	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)	6
081	PWIMSA	PRODUCER PRICE INDEX:INTERMED MAT.SUPPLIES & COMPONENTS(82=100,SA)	6
082	PWCMSA	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)	6
083	PMCP	NAPM COMMODITY PRICES INDEX (PERCENT)	2
084	PUNEW	CPI-U: ALL ITEMS (82-84=100,SA)	6
085	PU83	CPI-U: APPAREL & UPKEEP (82-84=100,SA)	6
086	PU84	CPI-U: TRANSPORTATION (82-84=100,SA)	6
087	PU85	CPI-U: MEDICAL CARE (82-84=100,SA)	6
088	PUC	CPI-U: COMMODITIES (82-84=100,SA)	6
089	PUCD	CPI-U: DURABLES (82-84=100,SA)	6
090	PUS	CPI-U: SERVICES (82-84=100,SA)	6
091	PUXF	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)	6
092	PUXHS	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)	6
093	PUXM	CPI-U: ALL ITEMS LESS MEDICAL CARE (82-84=100,SA)	6
094	CES275	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO	6
095	CES277	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO	6
096	CES278	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO	6
097	exrus	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)	5
098	exruk	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)	5
099	exrcan	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)	5
100	ccinrv	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)	6
101	pmcp	NAPM COMMODITY PRICES INDEX (PERCENT)	1

Appendix continued

Code	Short	Long Description	Tcode
102	pwfsa	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)	6
103	pwfcsa	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)	6
104	pwimsa	PRODUCER PRICE INDEX:INTERMED MAT.SUPPLIES & COMPONENTS(82=100,SA)	6
105	pwcmsa	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)	6
106	punew	CPI-U: ALL ITEMS (82-84=100,SA)	6
107	pu83	CPI-U: APPAREL & UPKEEP (82-84=100,SA)	6
108	pu84	CPI-U: TRANSPORTATION (82-84=100,SA)	6
109	pu85	CPI-U: MEDICAL CARE (82-84=100,SA)	6
110	puc	CPI-U: COMMODITIES (82-84=100,SA)	6
111	pucd	CPI-U: DURABLES (82-84=100,SA)	6
112	pus	CPI-U: SERVICES (82-84=100,SA)	6
113	puxf	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)	6
114	puxhs	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)	6
115	puxm	CPI-U: ALL ITEMS LESS MEDICAL CARE (82-84=100,SA)	6
116	exrsw	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)	5
117	exrjan	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)	5
118	PSCCOM	SPOT MARKET PRICE INDEX:BLS & CRB: ALL COMMODITIES(1967=100)	6
119	hhsntn	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS(BCD-83)	1
120	sfygm3	fygm3-fyff	1
121	sfygm6	fygm6-fyff	1
122	sfygt1	fygt1-fyff	1
123	sfygt10	fygt10-fyff	1
124	sfygt5	fygt5-fyff	1
125	sfyaaac	fyaaac-fyff	1
126	sfybaac	fybaac-fyff	1
127	sfybac	fyac-fyff	1
128	CCIPY	CONSUMER INSTAL CREDIT TO PERSONAL INCOME, RATIO (%,SA)(BCD-95)	5
129	hmob	Mobile homes: manufacturers shipments (thous. of units, saar)	5
130	mocmq	New orders (net)-consumer goods & materials, 1992 dollars (bci)	5
131	msondq	New orders, nondefense capital goods, in 1992 dollars (bci)	1
132	fcls	Loans & sec @ all coml banks: total (bils, sa)	5
133	fcsgv	Loans & sec @ all coml banks: U.S. govt securities (bil$, sa)	5
134	fclre	Loans & sec @ all coml banks: real estate loans (bil$, sa)	5
135	fclin	Loans & sec @ all coml banks: loans to individuals (bil$, sa)	5
136	fste	U.S. mdse exports: total exports (f.a.s. value) (mil.$, s.a.)	5
137	UTL11	CAPACITY UTILIZATION - MANUFACTURING (SIC),PERCENT OF CAPACITY, SA, FRB	2
138	UTL15	CAPACITY UTILIZATION - NONMETALLIC MINERAL PRODUCT NAICS=327, PERCENT OF CAPACITY, SA, FRB	2
139	UTL17	CAPACITY UTILIZATION - FABRICATED METAL PRODUCT NAICS=332, PERCENT OF CAPACITY, SA, FRB	2
140	UTL21	CAPACITY UTILIZATION - MOTOR VEHICLES AND PARTS NAICS=3361-3, PERCENT OF CAPACITY, SA, FRB	1
141	UTL22	CAPACITY UTILIZATION - AEROSPCE and MISCELLANEOUS TRANSPORTATION EQ., PERCENT OF CAPACITY, SA, FRB	2
142	UTL29	CAPACITY UTILIZATION - PAPER NAICS=322, PERCENT OF CAPACITY, SA, FRB	2
143	A0M007	Mfr's New Orders, Durable Goods Industries (Bil. Chain 2000$)	1
144	GDP	Gross Domestic Product extrapolated under CPI	5
145	RS*	Total current-dollar sales for retail sectors	5
146	BSI*	Total current-dollar sales and inventories for the manufacturing, wholesale	5
147	DSNU*	"Shipments, new orders, and unfilled orders, expressed in current dollars, for advance durable goods"	5

NOTES: From Bureau of Census, Department of Commerce. Rest of data from Global Insight Database.
* Data starts from 1992:01 and rest of data start from 1960:01 through 2007:12.