Forecasting Financial and Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence

Hyun Hak Kim and Norman R. Swanson
Rutgers University

October 2010

Abstract

In this paper, we empirically assess the predictive accuracy of a large group of models based on the use of principle components and other shrinkage methods, including Bayesian model averaging and various bagging, boosting, LASSO and related methods Our results suggest that model averaging does not dominate other well designed prediction model specification methods, and that using a combination of factor and other shrinkage methods often yields superior predictions. For example, when using recursive estimation windows, which dominate other "windowing" approaches in our experiments, prediction models constructed using pure principal component type models combined with shrinkage methods yield mean square forecast error "best" models around 70% of the time, when used to predict 11 key macroeconomic indicators at various forecast horizons. Baseline linear models (which "win" around 5% of the time) and model averaging methods (which win around 25% of the time) fare substantially worse than our sophisticated nonlinear models. Ancillary findings based on our forecasting experiments underscore the advantages of using recursive estimation strategies, and provide new evidence of the usefulness of yield and yield-spread variables in nonlinear prediction specification.

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

−−−−−−−−−−−−−−−−−−−−−−−−−−−−−

[*] Hyun Hak Kim (hykim@econ.rutgers.edu) and Norman R. Swanson (nswanson@econ.rutgers.edu), Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA. This paper has been prepared for the Sir Clive W.J. Granger Memorial Conference held at Nottingham University in May 2010. We are very grateful to the organizers, Rob Taylor and David Hendry, for hosting this special event. The authors also owe many thanks to seminar participants at the conference, as well as at the Bank of Canada and Rutgers University, for useful comments. Additional thanks are owed to Nii Armah, Valentina Corradi, David Hendry, Gary Koop, John Landon-Lane, Fuchun Li, Greg Tkacz, Hiroki Tsurumi, and Hal White for numerous useful suggestions on an earlier version of this paper. The authors gratefully acknowledge financial support from a Rutgers University Research Council grant.

1 Introduction

Technological advances over the last five decades have led to impressive gains in not only computational power, but also in the quantity of available financial and macroeconomic data. Indeed, there has been something of a race going on in recent years, as technology, both computational and theoretical, has been hard pressed to keep up with the ever increasing mountain of data available for empirical use. From a computational perspective, this has helped spur the development of data shrinkage techniques, for example. In economics, one of the most widely applied of these is diffusion index methodology. Diffusion index techniques offer 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 be a time series vector of dimension $(T \times 1)$ and let X be a time-series predictor matrix of dimension $(T \times N)$, and define the following dynamic factor model, where F_t denotes a $1 \times r$ vector of unobserved common factors that can be extracted from X_t. Namely, let $X_t = F_t\Lambda' + e_t$, where e_t is an $1 \times N$ vector of disturbances and Λ is an $N \times r$ coefficient matrix. Using common factors extracted from the above model, we follow Stock and Watson (2002a,b) and Bai and Ng (2006a) and consider forecasting models of the form:

$$Y_{t+h} = W_t\beta_W + F_t\beta_F + \varepsilon_{t+h}, \tag{1}$$

where h is the forecast horizon, Y_t is the scalar valued "target" variable to be forecasted, W_t is a $1 \times s$ vector of observable variables, including lags of Y_t, ε_t is a disturbance term, and the β's are parameters to be estimated, defined conformably. In one of the approaches followed in this paper, we first estimate the unobserved factors, F_t, and then forecast Y_{t+h} using observed variables and $\hat{F}_t$, where $\hat{F}_t$ is an estimator of F_t. Even though factor models are now widely used, many issues remain outstanding, such as the determination of the number of factors to be used in subsequent prediction model construction (see e.g. Bai and Ng (2002, 2006b, 2008)). In light of this, and in order to add functional flexibility, we additionally implement versions of (1) where the numbers and functions of factors to be used is subsequently selected using a variety of additional shrinkage methods. Various other related methods, including targeted regressor selection based on shrinkage, are also implemented. In this sense, we add to the recent work of Stock and Watson (2005a) as well as Bai and Ng (2008a,b), who survey several methods for shrinkage that are based on factor augmented autoregression models. Shrinkage methods considered in this paper include bagging, boosting, Bayesian model averaging, simple model averaging, ridge regression, least angle regression, elastic net and the non-negative garotte. We also evaluate various linear models, and hence add to the recent work of Pesaran et al. (2010), who carry out a broad examination of factor-augmented vector autoregression models.

In summary, the purpose of this paper is to empirically assess the predictive accuracy of various linear models; pure principal component type models; principal components models constructed using subsets of variables selected based on the elastic net and other shrinkage techniques; principle components models where the factors to be used in prediction are directly selected using shrinkage methods such as ridge regression and bagging; models constructed by directly applying shrinkage methods (other than principle components) to the data; and a number of model averaging methods. The horse-race that we carry out using all of the above approaches allows us to provide new evidence on the usefulness of factors in general as well as on various related issues such as whether model averaging still "wins" rather ubiquitously.

The variables that we predict include a variety of macroeconomic variables that are useful for evaluating the state of the economy. More specifically, forecasts are constructed for eleven series, including: the unemployment rate, personal income less transfer payments, the 10 year Treasury-bond yield, the consumer price index, the producer price index, nonfarm payroll employment, housing starts, industrial production, M2, the S&P 500 index, and gross domestic product. These variables constitute 11 of the 14 variables (for which long data samples are available) that the Federal Reserve takes into account, when formulating the nation's monetary policy. In particular, as noted in Armah and Swanson (2010b) and on the 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 finding can be summarized as follows. First, as might be expected, for a number of our target variables, we find that various sophisticated models, such as component-wise boosting, have lower mean square forecast errors (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. More specifically, models constructed using pure principal component type models combined with shrinkage methods yield MSFE-"best" models around 70% of the time, across multiple forecast horizons, and for various prediction periods. Moreover, a small subset of combined factor/shrinkage type models "win" approximately 50% of the time, including c-boosting, ridge regression, least angle regression, elastic net and the non-negative garotte, with c-boosting the clear overall "winner". Baseline linear models (which "win" around 5% of the time) and model averaging methods (which "win" around 25% of the time) fare substantially worse than our sophisticated nonlinear models. Ancillary findings based on our forecasting experiments underscore the advantages of using recursive estimation windowing strategies[1], and provide new evidence of the usefulness of yield and yield-spread variables in nonlinear prediction specification.

Although we leave many important issues to future research, such as the prevalence of structural breaks other than level shifts, and the use of even more general nonlinear methods for describing the data series that we examine, we believe that results presented in this paper add not only to the diffusion index literature, but also to the extraordinary collection of papers on forecasting that Clive W.J. Granger wrote during his decades long research career. Indeed, as we and others have said many times, we believe that Clive W.J. Granger is in many respects the father of time series forecasting, 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.

The rest of the paper is organized as follows. In the next section we provide a brief survey of dynamic factor models. In Section 3, we survey the robust shrinkage estimation methods used in our prediction experiments. Data, forecasting methods, and baseline forecasting models are discussed in Section 4, and empirical results are presented in Section 5. Concluding remarks are given in Section 6.

2 Diffusion Index Models

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), and Stock and

[1] For further discussion of estimation windows and the related issue of structural breaks, see Pesaran and Timmermann (2007).

Watson (1999, 2002, 2005a,b, 2006). Stock and Watson (2006) discuss in some detail the literature on the use of diffusion indices for forecasting. In the following brief discussion of diffusion index methodology, we follow Stock and Watson (2002).

2.1 Factor Models: Basic Framework

Let X_{tj} be the observed datum for the $j-$th cross-sectional unit at time t, for $t = 1, ..., T$ and $j = 1, ..., N$. Recall that we shall consider the following model:

$$X_{tj} = F_t \Lambda_j' + e_{tj}, \tag{2}$$

where F_t is a $1 \times r$ vector of common factors, Λ_j is an $1 \times r$ vector of factor loadings associated with F_t, and e_{tj} is the idiosyncratic component of X_{tj}. The product $F_t \Lambda_j'$ is called the common component of X_{tj}. This is the dimension reducing 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. Recall also that our basic forecasting equation is:

$$Y_{t+h} = W_t \beta_W + F_t \beta_F + \varepsilon_{t+h}, \tag{3}$$

where h is the forecast horizon and W_t is a $1 \times s$ 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 (2). Connor and Korajczyk (1986, 1988, 1993) note that the factors can be consistently estimated by principal components as $N \to \infty$, even if e_{tj} 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 $\{e_{tj}, \varepsilon_{t+h}\}$ are cross-sectionally and serially *iid* We work with high-dimensional factor models that allow both N and T to tend to infinity, and in which e_{tj} may be serially and cross-sectionally correlated so that the covariance matrix of $e_t = (e_{t1}, ..., e_{tN})$ does not have to be a diagonal matrix. We will also assume $\{F_t\}$ and $\{e_{tj}\}$ are two groups of mutually independent stochastic variables. Furthermore, it is well known that if $\Lambda = (\Lambda_1, ..., \Lambda_N)'$ for $F_t \Lambda' = F_t Q Q^{-1} \Lambda'$, a normalization is needed in order to uniquely define the factors, where Q is a nonsingular matrix. Assuming that $(\Lambda' \Lambda / N) \to I_r$, we restrict Q to be orthonormal. 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 (3) involve a two step procedure because both the regressors and coefficients in the forecasting equations are unknown. The data X_t are first used to estimate the factors, $\hat{F}_t$, 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 $\hat{F}_t$ and W_t. 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 (2008)). In this paper, we try different methods for estimating $\hat{\beta}_F$ and then compare the predictive accuracy of the resultant forecasting models.[2]

2.2 Factor Models: Estimation

Before implementing principal components analysis in the context of factor models, there remains the question of how many components to use. Bai and Ng (2002) provide one solution to the problem of choosing the number of factors. They establish convergence rates for factor estimates under consistent estimation of the number of factors, r, and propose panel criterion to consistently estimate the number of factors. Begin with an arbitrary number r ($< \min[N, T]$) and let Λ and F be the coefficient vector and r factors included in the estimation via solving (2). Let F be a matrix of r factors and

$$V(r, F) = \min_{\Lambda} \frac{1}{NT} \sum_{t=1}^{T} \sum_{j=1}^{N} \left(X_{tj} - F_t \Lambda_j' \right)^2 \tag{4}$$

be the sum of squared residuals from regression of X_j on the r factors for all j. Then we can estimate number of factors, r, can be determined using loss function $V(r, F) + kg(N, T)$ where $g(N, T)$ is the penalty function. Also, without loss of generality, we can set

$$V\left(r, \hat{F}\right) = \min_{\Lambda} \frac{1}{NT} \sum_{t=1}^{T} \sum_{j=1}^{N} \left(X_{tj} - \hat{F}_t \Lambda_j' \right)^2 \tag{5}$$

Along these lines Bai and Ng (2002) define selection criteria of the form $PC(r) = V\left(r, \hat{F}\right) + rh(N, T)$, where $h(\cdot)$ is a penalty function. In this paper, the following version is used (for

[2] We refer the reader to Stock and Watson (1999, 2002, 2005a,b) and Bai and Ng (2002, 2008, 2009) for a detailed explanation of this procedure, and to Connor and Korajczyk (1986, 1988, 1993), Forni et al. (2005) and Armah and Swanson (2010a) for further detailed discussion of generic diffusion models.

discussion, see Bai and Ng (2002) and Armah and Swanson (2010a)):

$$SIC(r) = V\left(r, \hat{F}\right) + r\hat{\sigma}^2 \left(\frac{(N+T-r)\ln(NT)}{NT} \right). \tag{6}$$

Our consistent estimate of the true number of factors is thus $\hat{r} = \arg \min_{0 \leq r \leq r_{\max}} SIC(r)$. We use this criteria for choosing the number of factors in the sequel, in cases where factors are included in our prediction models.

3 Robust Estimation Techniques

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. 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 algorithm 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 *iid.*. Inoue and Kilian (2005) and Stock and Watson (2005a) extend bagging to time series models. Stock and Watson (2005a) consider "bagging" as a form of shrinkage, when constructing prediction models. In this paper, we use the same algorithm that they do when constructing bagging estimators. 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)) and is implemented in Freund and Schapire (1997) using the algorithm called "AdaBoost.M1". Hastie et al. (2001) apply it to classification, and argue that "boosting" is one of the most powerful learning algorithms currently available. 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 (2009) use a boosting for selecting the predictors in factor augmented autoregressions. We implement a boosting algorithm that mirrors that used by these authors.

The "least absolute shrinkage and selection operator" (LASSO) was introduced by Tibshirani (1996), and is another attractive technique for variable selection using high-dimensional datasets, especially when N is greater than T. One version of the LASSO, "Least Angle Regression" (LARs), 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. LARs is based on a well known model-selection approach known as "forward-selection", which has been extensively used to examine cross-sectional data (for further details, see Efron et al. (2004)). Bai and Ng (2008) show how to apply the LARs and LASSO methods in the context of time series data, and Gelper and Croux (2008) extend Bai and Ng (2008)'s work to time series forecasting with many predictors. We implement Gelper and Croux (2008)'s algorithm when constructing the LARs estimator. 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-Elastic Net (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 (2008) apply the elastic net method to time series using the approach of Zou and Hastie (2005). We also follow their approach when implementing the elastic net.

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 previously 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 regression shrinkage method which modifies sum of square residual computations to include a penalty for inclusion of larger numbers of parameters. Ridge regression has been used widely, and hence we omit discussion of further details in the sequel.

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), Koop and Potter (2004) and 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.

The following sub-sections provide summary details on the implementation of the above methods in contexts where in a first step we estimate factors using the approach discussed above, while in a second step we select factor weights using a shrinkage method. Approaches in which we first directly implement shrinkage methods to select an informative set of variables for: (i) direct use in prediction model construction; or (ii) use in a second step where factors are constructed for subsequent use in prediction model construction follow immediately.

3.1 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 involves drawing bootstrap samples from the training sample (i.e. in-sample), applying a learning algorithm (prediction model) to each bootstrap sample, and averaging the predicted values for test observations. Consider the regression problem with the training sample $\{Y, X\}$. Generate B bootstrap samples from dataset form predictions, $\hat{Y}_b^*(X_b^*)$, say, using each bootstrap sample, $b = 1, ..., B$. Bagging simply averages the B prediction values, and can reduce prediction variance. Bühlmann and Yu (2002) consider bagging with a fixed number of strictly exogenous regressors and *iid* errors, and show that, asymptotically, the bagging estimator can be represented in shrinkage form. Namely:

$$\hat{Y}_{T+h}^{Bagging} = \sum_{j=1}^{N} \psi(t_j)\hat{\beta}_j \mathbf{P}_{Tj} + o_p(1), \tag{7}$$

where $\hat{Y}_{T+h}^{Bagging}$ is the forecast of Y_{T+h} made using data through time T, $\hat{\beta}_j = T^{-1}\Sigma_{t=1}^{T}\mathbf{P}_{tj}Y_{t+h}$ is the least squares estimator of β_j, $\mathbf{P}_t$ is the $1 \times N$ vector of the orthonormal predictors, $t_j = \sqrt{T}\hat{\beta}_j/s_e$, with $s_e^2 = \Sigma_{t=1}^{T}(Y_{t+h} - \mathbf{P}_t\hat{\beta}')^2/(T-N)$, where $\hat{\beta} = \left(\hat{\beta}_1, ..., \hat{\beta}_N\right)'$, and ψ is a function specific to the forecasting method. In the current context we set:

$$\psi(t) = 1 - \Phi(t+c) + \Phi(t-c) + t^{-1}[\phi(t-c) - \phi(t+c)], \tag{8}$$

where c is the pre-test critical value, ϕ is the standard normal density and Φ is the standard normal CDF. Further, following Stock and Watson (2005a) define

$$\hat{Y}_{T+h}^{Bagging} = W_T\hat{\beta}_W + \sum_{j=1}^{r} \psi(t_j)\hat{\beta}_{Fj}\hat{F}_{Tj} \tag{9}$$

where $\hat{\beta}_W$ is the LS estimator of β_W, W_t is a vector of lags of Y_T as in (3), and $\hat{\beta}_{Fj}$ is estimated using residuals, $Z_{T+h} = Y_{T+h} - W_T\hat{\beta}_W$. 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 $c = 1.96$.

3.2 Boosting

Boosting (see Freund and Schapire (1997)) is a procedure that combines the outputs of many "weak learners" (model) to produce a powerful "committee" (prediction). In this sense, boosting bears a resemblance to bagging and other committee-based approaches. Conceptually, the boosting method builds on a user-determined set of many weak learners (for example, least square estimators) and uses the set 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.

The AdaBoost algorithm proposes a method to estimate the response variable $Y \in \{-1, 1\}$. First, initialize weights associated with each observation, say $w_t = 1/T$, for $t = 1, ..., T$. Given a particular explanatory variable set, X, a classifier or learner, $\hat{\mu}$, is used to produce one of two outcomes, -1 or 1. At the i^{th} algorithm iteration, fit a learner $\hat{\mu}^i$ to the date using weights w_t, then construct the error rate

$$\hat{e}^i = \frac{\sum_{t=1}^{T} w_t \mathbf{I}(Y_t \neq \hat{\mu}_t^i)}{\sum_{t=1}^{T} w_t} \tag{10}$$

where $I(\cdot)$ is the indicator function. Now, compute $\alpha^i = \log\left(\left(1 - \hat{e}^i\right)/\hat{e}^i\right)$, then update

Friedman (2001) introduce "L_2Boosting", which takes the simple approach of refitting base learners to residuals from previous iterations (i.e. they use no penalty $\lambda \|\beta\|$). Bühlmann and Yu (2003) suggest another boosting algorithm fitting learners using one predictor at one time when large numbers of predictors exist. Bai and Ng (2009) modify this algorithm to handle time-series. We use their "Component-Wise L_2Boosting" algorithm in the sequel. Let $Z = Y - \hat{Y}^W$, which we obtain as a first step by fitting an autoregressive model to response variable using W_t as regressors. Then, using estimated factors: (i) Initialize : $\hat{\mu}^0(F_t) = \bar{Z}$, for each t. (ii) For $i = 1, ..., M$ iterations, carry out the following procedure. For $t = 1, ..., T$, let $u_t = Z_t - \hat{\mu}^{i-1}(D_t)$ be the "current residual". For each $j = 1, .., r$, regress the current $T \times 1$ residuals, u on $\hat{F}_j$ (the j-th factor) to obtain $\hat{\beta}_j$. (iii) Compute $\hat{d}_j = u - \hat{F}_j\hat{\beta}_j$ for $j = 1, .., r$,and the sum of square residuals, $SSR_j = \hat{d}_j'\hat{d}_j$. Let j_*^i denote the column selected at the i^{th} iteration, say, such that $SSR_{j_*^i} = \min_{j \in [1,...,r]} SSR_j$ then let $\hat{g}_*^i(F) = \hat{F}_{j_*^i}\hat{\beta}_{j_*^i}$. (iv) For $t = 1, ..., T$, update $\hat{\mu}^i = \hat{\mu}^{i-1} + \nu\hat{g}_*^i$, where $0 \leq \nu \leq 1$ is the step length.

We may encounter a problem of over-fitting if we iterate this algorithm too many times. Therefore, selecting the number of iterations is crucial. Bai and Ng (2009) define the stopping parameter M using an information criterion of the form: $IC\left(i\right)=\log\left[\hat{\sigma}^{i^2}\right]+\frac{A_T\cdot df^i}{T}$, where $\hat{\sigma}^{i^2}=\Sigma_{t=1}^{T}\left(Y_t-\hat{\mu}^i\left(\hat{F}_t\right)\right)^2$ and $A_T=\log(T)$. Then $M=\arg\min_i IC\left(i\right)$. Here, the degrees of freedom is defined as $df^i=trace\left(B^i\right)$, where $B^i=B^{i-1}\nu\mathbf{P}^{(i)}\left(I_T-B_{i-1}\right)=I_T-\Pi_{h=0}^{i}\left(I_T-\nu\mathbf{P}^{(h)}\right)$, with $\mathbf{P}^{(i)}=\hat{F}_{j_*^i}\left(\hat{F}'_{j_*^i}\hat{F}_{j_*^i}\right)^{-1}\hat{F}_{j_*^i}$. Starting values for B^i 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. Our boosting estimation uses this criterion. Finally, we have

$$\hat{Y}_{t+h}^{Boosting}=W_t\hat{\beta}_W+\hat{\mu}^M\left(\hat{F}_t\right) \tag{11}$$

where $\hat{\beta}_W$ is defined above and M denotes the final number of iterations of the above boosting algorithm.

3.3 Least Angle Regression (LARs)

Like many other stagewise regression approaches, we start with $\hat{\mu}^0=0$, work with the residuals after fitting W_t to the target variable, and construct a first estimate $\hat{\mu}=X_t\hat{\beta}$, in steps, using standardized data. Define $\hat{\mu}_{\mathcal{G}}$ to be the current LARs estimator, where $\mathcal{G}$ is a set of variables that is incrementally increased according to the relevance of each variable examined. As with stagewise regression, we define $c\left(\hat{\mu}_{\mathcal{G}}\right)=\hat{c}=X'\left(Y-\hat{\mu}_{\mathcal{G}}\right)$, where X is the "current" set of regressors, to be the "current correlation" vector of length N. In particular, define the set $\mathcal{G}$ to be the set including covariates with the largest absolute correlations; so that we can define $\hat{C}=\max_j\left\{\hat{c}_j\right\}$ and $\mathcal{G}=\left\{j:\left|\hat{c}_j\right|=\left|\hat{C}\right|\right\}$, by letting $s_j=sign\left(\hat{c}_j\right)$ equals ± 1, for $j\in\mathcal{G}$ and defining the active matrix corresponding to $\mathcal{G}$ as $\mathcal{X}_{\mathcal{G}}=\left(...s_jX_j...\right)_{j\in\mathcal{G}}$. Let

$$\mathcal{D}_{\mathcal{G}}=\mathcal{X}'_{\mathcal{G}}\mathcal{X}_{\mathcal{G}} \text{ and } A_{\mathcal{G}}=\left(\mathbf{1}'_{\mathcal{G}}\mathcal{D}_{\mathcal{G}}^{-1}\mathbf{1}_{\mathcal{G}}\right)^{-\frac{1}{2}}, \tag{12}$$

where $\mathbf{1}_{\mathcal{G}}$ is a vector of ones equal in length to the rank of $\mathcal{G}$. A unit equiangular vector with columns of $\mathcal{X}_{\mathcal{G}}$ can be defined as $u_{\mathcal{G}}=\mathcal{X}_{\mathcal{G}}w_{\mathcal{G}}$, where $w_{\mathcal{G}}=A_{\mathcal{G}}\mathcal{D}_{\mathcal{G}}^{-1}\mathbf{1}_{\mathcal{G}}$ so that $\mathcal{X}'_{\mathcal{G}}u_{\mathcal{G}}=A_{\mathcal{G}}\mathbf{1}_{\mathcal{G}}$. LARs then updates $\hat{\mu}$ as

$$\hat{\mu}_{\mathcal{G}+}=\hat{\mu}_{\mathcal{G}}+\hat{\gamma}u_{\mathcal{G}} \tag{13}$$

where

$$\hat{\gamma}=\min_{j\in\mathcal{G}^c}{}^{+}\left(\frac{\hat{C}-\hat{c}_j}{A_{\mathcal{G}}-a_j}\right)\left(\frac{\hat{C}+\hat{c}_j}{A_{\mathcal{G}}+a_j}\right), \tag{14}$$

with $a_j = X'w_j$ for $j \in \mathcal{G}^c$. Efron et al. (2004) show that the lasso is in fact a special case of LARs that imposes a specific sign restriction.

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 the dependent variable, excluding predictor variables, using least squares. The corresponding residual series is retained and its standardized version is denoted Z. The time-series LARs (henceforth, TS-LARs) procedure ranks the predictors according to how much they contribute to improving upon autoregressive fit. Using estimated factors as regressors, the following is the "LARs" algorithm of Gelper and Croux (2008): (i) Fit an autoregressive model to the dependent variable without factors and retain the corresponding residuals. The objective is to forecast these residuals. Begin by setting $\hat{\mu}^0 = \hat{\mu}^0 \left(\hat{F}\right) = \bar{Z}$, as done in the boosting algorithm above, except that the data used in this algorithm are standardized. (ii) For $i = 1, 2, ..., r$:(a) Pick j_*^i from $j = 1, 2, ..., r$ $(\leq N)$ which has the highest R^2 value, $R^2 \left(\hat{\mu}^{i-1} \sim \hat{F}_j\right)$, where R^2 is a measure of least square regression fit, and where "$\backsim$ " denotes horizontal concatenation. The predictor with highest R^2 is denoted $\hat{F}_{(i)} = \hat{F}_{j_*^i}$, and this predictor will be included in the active set $\mathcal{G}^i$. That is, $\hat{F}_{(i)}$ denotes the i^{th} ranked predictor, the active set $\mathcal{G}^i$ will contain $\hat{F}_{(1)}, \hat{F}_{(2)}, ..., \hat{F}_{(i)}$, and j_*^i is excluded in next procedure. (b) Denote the hat-matrix corresponding to the i^{th} ranked active predictor by $H_{(i)}$, which is the projection matrix on the space spanned by the columns of $\hat{F}_{(i)}$. That is, $H_{(i)} = \hat{F}_{(i)} \left(\hat{F}'_{(i)} \hat{F}^{-1}_{(i)}\right) \hat{F}'_{(i)}$.(c) Let $\tilde{F}_{(i)} = H_{(i)} \hat{\mu}^{i-1}$ be the $T \times 1$ standardized vector of values $\hat{F}$ at i^{th} iteration. Then find equiangular vector u^i, where $u^i = \left(\tilde{F}_{(1)}, \tilde{F}_{(2)}, ..., \tilde{F}_{(i)}\right) w^i$, $w^i = \frac{\mathcal{D}^{-1}_{\mathcal{G}^i} \mathbf{1}_i}{\sqrt{\mathbf{1}'_i \mathcal{D}^{-1}_{\mathcal{G}^i} \mathbf{1}_i}}$, $\mathcal{D}_{\mathcal{G}^i} = \mathcal{F}'_{\mathcal{G}^i} \mathcal{F}_{\mathcal{G}^i}$, $\mathcal{F}_{\mathcal{G}^i} = \left(...s_j \hat{F}^j ...\right)_{j \in \mathcal{G}^i}$, $s_j = sign\left(\hat{c}_j\right)$ and $\hat{c} = \hat{F}'\left(\bar{Z} - \hat{\mu}^i\right)$. (iii) Update the response $\hat{\mu}^i = \hat{\mu}^{i-1} - \hat{\gamma}^i u^i$, where $\hat{\gamma}^i$ is the smallest positive solution for a predictor $\hat{F}_j$ which is not already in the active set, and is defined in (14).Then go back to Step 2, where $\hat{F}_{(i+1)}$ is added to the active set and the new response is standardized and denoted by $\hat{\mu}^{i+1}$ (see Gelper and Croux (2008) for further computational details).

After ranking the predictors, $\hat{F}_{(i+1)}$, the highest ranked will be included in the final model. Now, the only choice remaining is how many predictors to include in the model. This number, i^* is chosen using the Schwarz Information Criterion (SIC), as done in Gelper and Croux (2008). Finally, construct

$$\hat{Y}^{LARs^+}_{t+h} = W_t \hat{\beta}_W + \hat{\mu}^{i^*}(\hat{F}_t) \tag{15}$$

where $\hat{\mu}^{i^*}(\hat{F}_t)$ is the optimal value of $\hat{\mu}^i$ selected using the SIC, and evaluated at time t. The final predictor of Y is formed by adding back the mean to $\hat{Y}_{t+h}^{LARs^+}$.

3.4 LARs - Elastic Net (LARs-EN)

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, the so-called the "Elastic Net", that they claim remedies the above problems. The method is similar to the LASSO estimator discussed above and in Bai and Ng (2008b).

In order to motivate the LARs-EN, we begin with a generic discussion of the "naïve elastic net" (NEN). Assume again that we are interested in X and Y, and that the variables in X are standardized as above. For any fixed non-negative η_1 and η_2, the naive elastic net criterion is defined as:

$$L(\eta_1, \eta_2, \beta) = |Y - D\beta|^2 + \eta_2 |\beta|^2 + \eta_1 |\beta|_1, \tag{16}$$

where $|\beta|^2 = \sum_j^N (\beta_j)^2$ and $|\beta|_1 = \sum_j^N |\beta_j|$. The naive elastic net estimator is $\hat{\beta}^{NEN} = \arg\min_{\beta} \{L(\eta_1, \eta_2, \beta)\}$. This problem is equivalent to the optimization problem:

$$\hat{\beta}^{NEN} = \arg\min_{\beta} |Y - X\beta|^2, \ \ \text{subject to} \ \ (1-\alpha)|\beta|_1 + \alpha |\beta|^2, \tag{17}$$

where $\alpha = \frac{\eta_2}{\eta_1+\eta_2}$. 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 $\alpha = 0$, 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

$$X^+_{(T+N)\times N} = (1+\eta_2)^{-1/2} \begin{pmatrix} X \\ \sqrt{\eta_2}\mathbf{I}_N \end{pmatrix}, \qquad Y^+_{(T+N)\times 1} = \begin{pmatrix} Y \\ \mathbf{0}_N \end{pmatrix}. \tag{18}$$

Then, we can rewrite the naive elastic criterion as:

$$L\left(\frac{\eta_1}{\sqrt{1+\eta_2}}, \beta\right) = L\left(\frac{\eta_1}{\sqrt{1+\eta_2}}, \beta^+\right) = \left|Y^+ - D^+\beta^+\right|^2 + \frac{\eta_1}{\sqrt{1+\eta_2}}\left|\beta^+\right|_1. \tag{19}$$

If we let

$$\hat{\beta}^+ = \arg\min_{\beta^+} L\left(\frac{\eta_1}{\sqrt{1+\eta_2}}, \beta^+\right) \tag{20}$$

then the NEN estimator $\hat{\beta}^{NEN}$ is:

$$\hat{\beta}^{NEN} = \frac{1}{\sqrt{1+\eta_2}}\hat{\beta}^{+} \tag{21}$$

In this orthogonal setting, the naive elastic net can be represented as combination of ordinary least squares and the parameters (η_1, η_2). Namely:

$$\hat{\beta}^{NEN} = \frac{\left(\left|\hat{\beta}^{LS}\right| - \eta_1/2\right)_{pos}}{1+\eta_2} sign\left\{\hat{\beta}^{LS}\right\}. \tag{22}$$

where $\hat{\beta}^{LS}$ is the least square estimator of β and $sign\left(\cdot\right)$ equals ±1. Here, "*pos*" denotes the positive part of the term in parentheses. Using these expression, the ridge estimator can be written as

$$\hat{\beta}^{ridge} = \frac{\hat{\beta}^{LS}}{1+\eta_2} \tag{23}$$

and the LASSO estimator is

$$\hat{\beta}^{lasso} = \left(\left|\hat{\beta}^{LS}\right| - \eta_1/2\right)_{pos} sign\left\{\hat{\beta}^{LS}\right\}. \tag{24}$$

Zou and Hastie (2005), in the context of above naive elastic net, point out that there is double shrinkage, which does not help to reduce the variance and may lead to unnecessary bias so that they propose the elastic net, in which this double shrinkage is corrected. Given equation (18), the naive elastic net solves the LASSO problem of the type:

$$\hat{\beta}^{+} = \arg\min_{\beta^{+}} \left|Y^{+} - X^{+}\beta^{+}\right|^2 + \frac{\eta_1}{\sqrt{1+\eta_2}}\left|\beta^{+}\right|_1. \tag{25}$$

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

$$\hat{\beta}^{EN} = \sqrt{1+\eta_2}\hat{\beta}^{+}. \tag{26}$$

Thus ,

$$\hat{\beta}^{EN} = \left(1+\eta_2\right)\hat{\beta}^{NEN}. \tag{27}$$

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,

$$\hat{\beta}^{EN} = \arg\min_{\beta}\beta'\left(\frac{X'X + \eta_2\mathbf{I}_N}{1+\eta_2}\right)\beta - 2Y'X\beta + \eta_1\left|\beta\right|_1, \tag{28}$$

which we use in the forecasting model given as (3) to construct predictions.

Zou and Hastie (2005) propose an algorithm called the LARs-EN to estimate $\hat{\beta}^{EN}$ using the LARs discussed above. With fixed η_2, the elastic net problem is equivalent to the LASSO problem on the augmented data set (X^{+}, Y^{+}), so that LARs can create an entire elastic net solution path efficiently by letting $\mathcal{D}_{\mathcal{G}}$ in (12) be $\frac{1}{1+\eta_2}\left(\mathcal{X}_{\mathcal{G}}'\mathcal{X}_{\mathcal{G}} + \eta_2 \mathbf{I}_{\mathcal{G}}\right)$ for any active set $\mathcal{G}$. Then the LARs-EN algorithm updates the elastic net estimator sequentially. Choosing tuning parameters, η_1 and η_2, is a critical issue here. Hastie et al. (2001) discuss some popular ways to choose tuning parameters, and Zou and Hastie (2005) use tenfold cross-validation (CV). Since there are two tuning parameters, it is necessary to cross-validate on two dimensions. At first, we pick a small grid of values for η_2, say $(0, 0.01, 0.1, 1, 10, 100)$. Then, for each η_2, LARs-EN produces the entire solution path of the elastic net. η_2 and i, the number of iterations in the LARs algorithm is selected by tenfold CV. The selected η_2 value will be the one giving the smallest CV error. We follow the above approach when using the EN method.

3.5 Non-Negative Garotte (NNG)

The NNG estimator of Breiman (1995) estimator is a scaled version of the least squares estimator. As in the previous section, we begin by considering generic X and Y. Assume that the following shrinkage factors are given: $q(\zeta) = (q_1(\zeta), q_2(\zeta), ..., q_N(\zeta))'$. The objective is to choose the shrinkage factors in order to minimize:

$$\frac{1}{2}\|Y - Gq\|^2 + T\zeta\sum_{j=1}^{N} q_j, \qquad \text{subject to } q_j > 0,\ j = 1, .., N, \tag{29}$$

where $G = (G_1, .., G_N)'$, $G_j = X_j\widehat{\beta}_j^{LS}$, and $\widehat{\beta}^{LS}$ is the least squares estimator. Here $\zeta > 0$ is the tuning parameter. The NNG estimator of the regression coefficient vector is defined as $\hat{\beta}_j^{NNG}(\zeta) = q_j(\zeta)\hat{\beta}_j^{LS}$, and the estimate of Y is defined as $\widehat{\mu} = X\hat{\beta}^{NNG}(\zeta)$. Assuming, for example, that $X'X = I$, the minimizer of expression (29) has the following explicit form: $q_j(\zeta) = \left(1 - \frac{\zeta}{(\hat{\beta}_j^{LS})^2}\right)_+$, $j = 1, ..., N$. This ensures that the shrinking factor may be identically zero for redundant predictors. The disadvantage of the 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 an initial estimate for use in the NNG; and they prove that if the initial estimate is consistent, the non-negative garotte is a consistent estimator, given that the tuning parameter, ζ, is chosen appropriately. Zou (2006) shows that the original non-negative garotte with ordinary least squares is also consistent, if N is fixed, as $T \to \infty$. Our approach is to start the algorithm with the least squares estimator, as in Yuan (2007), who outline the following algorithm for the non-negative garotte that we use in the sequel: (i) First, set $i = 1$, $q^0 = 0$, $\hat{\mu}^0 = \bar{Z}$. Then compute the current active set

$$\mathcal{G}^i = \arg\max_j \left(G_j' \hat{\mu}^{i-1} \right),$$

where $G_j = \hat{F}_j \hat{\beta}_j$, is the j^{th} element of the $T \times r$ vector G and the initial $\hat{\beta}$ is obtained by regressing $\hat{F}$ on Z using least squares. (ii) Compute the current direction γ, which is a r dimensional vector defined by $\gamma_{(\mathcal{G}^i)^c} = 0$ and

$$\gamma_{\mathcal{G}^i} = \left(G_{\mathcal{G}^i}' G_{\mathcal{G}^i}' \right)^{-1} G_{\mathcal{G}^i}' \hat{\mu}^{i-1}$$

(iii) For every $j' \notin \mathcal{G}^i$, compute how far the non-negative garotte will progress in direction γ before $\hat{F}_j$ enters the active set. This can be measured by a α_j such that

$$G_{j'}' \left(\hat{\mu}^{i-1} - \alpha_j G' \gamma \right) = G_j' \left(\hat{\mu}^{i-1} - \alpha_j G' \gamma \right)$$

where j is arbitrary chosen from $\mathcal{G}^i$. And for every $j \in \mathcal{G}^i$, compute $\alpha_j = \min \left(\beta_j, 1 \right)$, where $\beta_j = -q_j^{i-1} / \gamma_j$, if nonnegative, measures how far the group non-negative garotte will "progress" before d_j becomes zero. (iv) If $\alpha_j \leq 0, \forall j$ or $\min\limits_{j, \alpha_j > 0} \{\alpha_j\} > 1$, set $\alpha = 1$. Otherwise, denote $\alpha = \min\limits_{j, \alpha_j > 0} \{\alpha_j\} \equiv \alpha_{j^*}$ Set $q^i = q^{i-1} + \alpha' \gamma$. If $j^* \notin \mathcal{G}^i$, update $\mathcal{G}^{i+1}$ by adding j^* to the set $\mathcal{G}^i$; else update $\mathcal{G}^{i+1}$ by taking out j^* from the set $\mathcal{G}^i$. (v) Set $\hat{\mu}^i = Y - G' q^i$ and $i = i + 1$. Go back to Step 1 repeat until $\alpha = 1$, yielding $\hat{\mu}^{final} = \hat{\mu}^{NNG}$. Finally, form

$$\hat{Y}_{t+h}^{NNG^+} = W_t \hat{\beta}_W + \hat{\mu}^{NNG}, \tag{30}$$

and construct the prediction $\hat{Y}_{t+h}^{NNG}$ by adding back the mean to $\hat{Y}_{t+h}^{NNG^+}$.

3.6 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)). Assume that we are interested in Q possible models, denoted by $M_1, ..., M_Q,$. The posterior probability of interest is,

$$p\left(\omega | Data\right) = \sum_{q=1}^{Q} p\left(\omega | Data, M_q\right) p\left(M_q | Data\right), \tag{31}$$

where ω is a vector of coefficients. If $g(\omega)$ is a function of ω, the conditional expectation is given as:

$$E\left[g(\omega)|Data\right] = \sum_{q=1}^{Q} E\left[g(\omega)|Data, M_q\right] p\left(M_q|Data\right). \tag{32}$$

Accordingly Bayesian model averaging involves obtaining results for all candidate models and averaging them with weights determined by the posterior model probabilities. However, if we have 20 potential variables, then we have 2^{20} possible models. This means that we must estimate $1,048,576$ 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. Thus there is a need for algorithms which do not require us to consider every possible model. The most popular one is MC^3, which takes draws from the posterior distribution of the models and MCMC draws from the posterior distributions of the parameters. In this paper, we use the related algorithm of Koop and Potter (2004) which follows Clyde (1999).

To implement Bayesian model averaging, we require a slightly different setup for that discussed above, in order to handle lagged dependent variables, which are included in most of our models. Chipman et al. (2001) suggest to integrating them out using non-informative priors. Specifically, let our forecasting model be

$$Y^*_{t+h} = \beta^* F^*_t + \varepsilon^*_t, \tag{33}$$

where $Y^*_{t+h} = \left[I_T - W_t\left(W_t'W_t\right)W_t'\right]Y_{t+h}$, $F^*_t = \left[I_T - W_t\left(W_t'W_t\right)W_t'\right]\hat{F}_t$, $W_t, \hat{F}_t$ is defined in (3), and $\varepsilon_{t+h} \sim N\left(0, \sigma^2\right)$. We use a natural conjugate prior (i.e. $\beta^*|\sigma^{-2} \sim N\left(\underline{\beta}^*, \sigma^2\underline{V}\right)$ and $\sigma^{-2} \sim G\left(\underline{s}^{-2}, \underline{\varpi}\right)$, where $G\left(\underline{s}^{-2}, \underline{\varpi}\right)$ denotes the Gamma distribution with mean $\underline{s}^{-2}$ and degrees of freedom $\underline{\varpi}$.

Each candidate models is described with U which is an $r \times 1$ vector which shows whether each column of explanatory variables is included in current model with a one or a zero. According to Koop and Potter (2004), $p(U|Y^*)$ is drawn directly, since our explanatory variables are orthogonal. We set $p\left(Y^*|U, \sigma^2\right)$ to be the marginal likelihood for the normal regression model defined by U, and derive $P\left(U|Y^*, \sigma^2\right)$, given a prior, $p(U)$ and $p\left(\sigma^2|Y^*, U\right)$ takes the inverted-Gamma form as usual. Next step is specifying the prior model probability, $p\left(M_q\right)$ or equivalently, a prior for $p(U)$:

$$p(U) = \prod_{j=1}^{R} \upsilon_j^{U_j}\left(1 - \upsilon_j\right)^{U_j}, \tag{34}$$

where υ_j is the prior probability that each potential factor enters the model. A common benchmark case sets $\upsilon_j = \frac{1}{2}$, equivalently, $P(M_q) = \frac{1}{Q}$ for $q = 1, ..., Q$. Other choices are also possible. For example, we can allow υ_j to depend on the j-th largest eigenvalue of $\hat{F}_t\hat{F}$.

Using the strategy described in Fernandez et al. (2001a) and Kass and Raftery (1995), we use a noninformative, improper, prior over parameters for lagged variables to all models and Koop and Potter (2004) suggest a noninformative prior for σ^{-2}, that is, if $\underline{\varpi} = 0$, s^{-2} does not enter the marginal likelihood or posterior). Following Fernandez et al. (2001a), we set $\underline{\beta}^* = \mathbf{0}_R$ and use a g-prior form for $\underline{V}$ by setting

$$\underline{V}_r = [g_r F_r^{*\prime} F_r^*]^{-1} \tag{35}$$

(see Fernandez et al. (2001a) and Zellner (1986) for more details about the use of g-priors). Then, next issue is specifying g. Fernandez et al. (2001a) studied about the properties of many possible choices for g and Koop and Potter (2004), in an objective Bayesian spirit, focus on some values for g as $g = \frac{1}{T}$ and $g = \frac{1}{R^2}$. Finally, we will have

$$\hat{Y}_{t+h}^{*,BMA} = \hat{\beta}_F F_t^* \tag{36}$$

and our forecast, $\hat{Y}_{t+h}^{BMA}$ is redefined as $[I_T - W_t (W_t' W_t) W_t']^{-1} \hat{Y}_{t+h}^{*,BMA}$.

4 Data, Forecasting Methods, and Baseline Forecasting Models

4.1 Data

The data that we use are monthly observations on 146 U.S. macroeconomic time series for the period 1960:01 - 2009:5 ($N = 144, T = 593$)[3]. Forecasts are constructed for eleven variables, 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, and gross domestic product.[4]. Table 1 lists the eleven variables. 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). Note that a full list of predictor variables is provided in

[3] This is an updated and expanded version of the Stock and Watson (2005a,b) dataset.

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

the appendix to an earlier working paper version which is available upon request from the authors.

4.2 Forecasting Methods

Using the transformed data set, denoted above by X, the factors are estimated by the method of principal components. Thereafter, the alternative methods outlined in the previous sections are used to form forecasting models and predictions. In particular, we consider three specification types when constructing shrinkage based prediction models.

Specification Type 1: Principal components are first constructed, and then prediction models are formed using the shrinkage methods of Section 3 to select functions of and weights for the factors to be used in our prediction models of the type given in (3).

Specification Type 2: Principal component models of the type given in (3) are constructed using subsets of variables from the largescale dataset that are first selected via application of the shrinkage methods of Section 3. This is different from the above approach of estimating factors using all of the variables.

Specification Type 3: Prediction models are constructed using only the shrinkage methods discussed in Section 3, without use of factor analysis at any stage.

We specify various linear as well as pure principal component type prediction models (see the subsequent section for details of these models).

In our prediction experiments, pseudo out-of-sample forecasts are calculated for each variable and method, for prediction horizons $h = 1, 3$, and 12. All estimation, including lag selection, shrinkage method application, and factor construction 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, following the testing approach discussed in Section 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 variable to include is done using the SIC. Out-of-sample forecasts begin after 13 years (e.g. the initial in-sample estimation period is R =156 observations, and the out-of-sample period consists of P observations, for $h = 1$). For example, when forecasting the unemployment rate, when $h = 12$, the first forecast will be $\hat{Y}_{168} = \hat{\beta}_W W_{156} + \hat{\beta}_F \tilde{F}_{156}$. In our rolling estimation scheme, the in-sample estimation period used to calibrate our prediction models is fixed at length 10 years. The recursive estimation scheme begins with the same 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:

$$MSFE_{i,h} = \sum_{t=R-h+2}^{T-h+1} \left(Y_{t+h} - \hat{Y}_{i,t+h} \right)^2 \tag{37}$$

where $\widehat{Y}_{i,t+h}$ is $i-$th method's forecast for horizon h. Forecast accuracy is evaluated using point MSFEs as well as the predictive accuracy test of Diebold and Mariano (DM: 1995), which is implemented using quadratic loss, and which has a null hypothesis that the two models being compared have equal predictive accuracy. DM test statistics have asymptotic $N(0,1)$ limiting distributions, under the assumption that parameter estimation error vanishes as $T, P, R \to \infty$, and assuming that each pair of models being compared is nonnested. Namely, the null hypothesis of the test is $H_0 : E\left[l\left(\varepsilon^1_{t+h|t}\right)\right] - E\left[l\left(\varepsilon^2_{t+h|t}\right)\right] = 0$,where $\varepsilon^i_{t+h|t}$ is $i-$th model's prediction error and $l(\cdot)$ is the quadratic loss function. The actual statistic in this case is constructed as: $DM = P^{-1}\sum_{i=1}^{P} d_t/\hat{\sigma}_{\overline{d}}$, 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, $\hat{\sigma}_{\overline{d}}$ 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 $\varepsilon^2_{t+h|t}$. Thus, if the statistic is negative and significantly different from zero, then Model 2 is preferred over Model 1.

4.3 Baseline Forecasting Models

In conjunction with the various forecast model specification approaches discussed above, we form predictions using the following benchmark models, all of which are estimated using least squares.

Univariate Autoregression: Forecasts from a univariate AR(p) model are computed as $\hat{Y}^{AR}_{t+h} = \hat{\alpha} + \hat{\phi}(L)Y_t$, with lags , p, selected using of the SIC.

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 lagged predictor variables selected using the SIC. Dependent variable lags are also selected using the SIC. Selection of the exogenous predictors includes choosing up to six variables prior to the construction of each new prediction model, as the recursive or rolling samples iterate forward over time.

Principal Component Regression: Forecasts from principal component regression are computed as $\hat{Y}_{t+h}^{PCR} = \hat{\alpha} + \hat{\gamma}\hat{F}_t$, where $\hat{F}_t$ is estimated via principal components using $\{X_t\}_{t=1}^{T}$, as in equation (3).

Factor Augmented Autoregression: Based on equations (3), forecasts are computed as $Y_{t+h}^{h} = \hat{\alpha} + \hat{\beta}_F\hat{F}_t + \hat{\beta}_W(L)Y_t$. This model combines an AR(p) model, with lags selected using the SIC, with the above principal component regression model.

Combined Bivariate ADL Model: As 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 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 $\hat{Y}_{t+h}^{ADL} = \hat{\alpha} + \hat{\beta}_i(L)X_{i,t} + \hat{\phi}_i(L)Y_t$. 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, Stock and Watson (2004, 2005a, 2006), and the references cited therein. In this literature, combination methods typically outperform individual forecasts. This stylized fact is sometimes called the "forecast combining puzzle."

Mean Forecast Combination: To further examine the issue of forecast combination, we form forecasts as the simple average of the thirteen forecasting models summarized in Table 2.

5 Empirical Results

In this section, we discuss the results of our prediction experiments. The variable mnemonics are given in Table 1, and forecasting models used are summarized in Table 2. 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 Section 4. For the case where models are estimated using recursive data windows, our results are gathered in Tables 3 to 6 and some summary results are presented in Tables 7 and 8.

Tables 3-6 report MSFEs and the results of DM predictive accuracy tests for all alternative forecasting models, using Specification Type 1 without lags (Table 3), Specification Type 1 with lags (Table 4), Specification Type 2 (Table 5), and Specification Type 3 (Table 6). Panels A-C reports results for h =1, 3 and 12 month prediction horizons, respectively. In each panel, the first row of entries 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 point MSFEs greater than that of our AR(SIC) model, etc. Entries in bold denote point-MSFE-"best" models for a given variable,forecast horizon, and specification type. In each table, dot-circled entries denote cases for which the MSFE-best model of the specification type reported on in a table (using recursive estimation) yields a lower MSFE than that based on using rolling estimation. Circled entries denote models that are MSFE-best across all specification types and estimation types (i.e. rolling and recursive). Boxed entries denote cases where models are "winners" across all specification types, when only viewing recursively estimated models (i.e. cases where the MSFE-best model uses rolling estimation (rolling window results are not reported here, for the sake of brevity). The results from DM predictive accuracy tests, for which the null hypothesis is that of equal predictive accuracy between the benchmark model (defined to be the AR(SIC) model), and the model listed in the first column of the table, as described in Section 4.2, are reported with single starred entries denoting rejection at the 10% level, and double starred entries denoting rejection at the 5% level.

Various results are apparent, upon inspection of tables. For example, for Specification Type 1, 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 holds for all forecast horizons except one (see Panels A-C of the table). Indeed, for most variables, there are various models that have lower point MSFEs that the AR(SIC) model, regardless of forecast horizon. However, there are exceptions. For example, for TB10Y, there are few models that yield lower MSFEs that the AR(SIC) model, and these are all for prediction horizons other than $h = 1$. In particular, for $h = 3$, only the Combined-ADL model is MSFE-"better", while for $h = 12$, Combined -ADL as well as a very few others (including C-Boosting, BMA, LARS and EN) "beat" AR(SIC), and they are only marginally superior, at that (see Tables 3 and 4). Additionally, comparison of the results in Tables 3 and 4 suggests that there is little advantage to using lags of factors when constructing predictions in our context. Instead, it appears that the more important determinant of model performance is the type of combination factor/shrinkage type model employed when constructing forecasts. Further evidence of this will be discussed in some detail shortly.

There are no models that uniformly yield lowest MSFEs, across both forecast horizon and variable. However, various models perform quite well, including in particular FAAR and PCR models. This supports the oft reported result that models that incorporate common factors offer a convenient way to filter the information contained in large-scale economic datasets.

When comparing results across Specification Types 1, 2, and 3 (compare Tables 3-6), we find that forecasts constructed using our model averaging specifications (Combined-ADL, BMA, and Mean) yield MSFE-best predictions for 4/11, 5/11, and 3/11 variables when using only recursive estimation, and for 0/11, 5/11, and 3/11 variables when using both recursive and rolling estimation windows (see Panel A of Table 7). This result is quite interesting, given the plethora of recent evidence indicating the superiority of model averaging methods in a variety of forecasting contexts; and is accounted for in part by our use of various relatively complicated combined factor/shrinkage models. In particular, note in the right hand section of Panel A in Table 7 that C-Boosting, Ridge, LARS, EN, and NNG "win" in 15/33 cases, when considering MSFE-best models across all specification and estimation window types. Moreover, the majority of these "wins" are accounted for by Specifications 1 and 2, suggesting that our shrinkage type methods perform best when coupled with factor analysis. In contrast, pure factor models (FAAR and PCR) yield "wins" in 8/33 cases, model averaging methods yield "wins" in 8/33 cases, and our non-factor and non-shrinkage based models "win" in 2/33 cases. Thus, the dominant model type is the combination factor/shrinkage type model. Moreover, models that involve factors, in aggregate, "win" in 23/33 cases; model averaging fares quite poorly; and pure linear models are almost never MSFE-best.

As evidenced in Panel B of Table 7, MSFE-best recursively estimated models dominate MSFE-best models estimated using rolling windows around 70% of the time, regardless of forecast horizon. (Complete results for rolling estimation-type models are available upon request from the authors.) This is not surprising, given the number of times that our more complicated combination factor/shrinkage type models are MSFE-best across all specification and estimation types; and suggests that structural breaks play a secondary role to parameter estimation error in determining the MSFE-"best" models.[5]

It should also be noted that DM test statistics yield ample evidence that a variety of models are statistically superior to our simple linear benchmark model, including many of our more sophisticated shrinkage based models. Such models are denoted as starred entries in the tables (see Section 4.2 for further details).

Finally, turning to the results in Table 8, notice that for a single forecast horizon, $h = 1$, results have been re-calculated for sub-samples corresponding to all of the NBER-dated expansionary periods in our sample, and to the combination of all recessionary and all expan-

[5] In lieu of this finding, the experiments carried out in this paper were replicated using the approach proposed by Clements and Hendry for addressing level shifts in the underlying data generating processes of our target variables (for details, refer to Clements and Hendry (1994), Clements and Hendry (1995), Clements and Hendry (2008)). Adjusting for level shifts by using differences of differences did not lead to notably improved prediction performance, however. (Complete results are available upon request from the authors.)

sionary periods. Although results apparent upon inspection of this table are largely in accord with those reported above, one additional noteworthy finding is worth stressing. Namely, in Panel A of the table, note that, when MSFE-best models are tabulated by specification type, our model averaging specifications perform quite well, particularly for Specification Types 2 and 3. This conforms to the results that can be observed by individually looking at each of Tables 3-6 (i.e. compare the bolded MSFE-best models in each individual table). However, notice that when results are summarized across all specification types (see Panel B of the table), then the model averaging type specifications yield MSFE-best predictions in around 30% instead of 40% of the cases. This is because Specification Type 1, where model averaging clearly "wins" the least, is the predominant winner when comparing the three specification types, as mentioned previously. Namely, the model building approach whereby we first construct factors and thereafter use shrinkage methods to estimate functions of and weights for factors to be used in our prediction models is the dominant specification type. This result serves to further stress that when more complicated specification methods are used, model averaging methods fare worse, and combination factor/shrinkage based approaches fare better. Put differently, we have evidence that when simpler linear models are specified, model averaging does worse than when more sophisticated nonlinear models are specified. Additionally, pure factor type models also perform well, particularly for the long expansion period from 1982-1990.

Given the importance of factors in our MSFE-best forecasting models, it would seem worthwhile to examine which variables contribute to the estimated factors used in our MSFE-best models, across all specification and estimation window types. This is done in Figure 1, where we report the ten most frequently selected variables for a variety of MSFE-best models and forecast horizons. Keeping in mind that factors are re-estimated at each point in time, prior to each new prediction being constructed, a 45 degree line denotes cases for which a particular variables is selected every time. For example, in Panels A and B, the BAA Bond Yield - Federal Funds Rate spread is the most frequently selected predictor when constructing factors to forecast the Producer Price Index and Housing Starts, respectively. For Specification Type 1, variables are selected based on the $A(j)$ and $M(j)$ statistics following Bai and Ng (2006a) and Armah and Swanson (2010a), and for Specification Type 2, we directly observe variables which are selected by shrinkage methods and then used to construct factors, prior to the construction of each new forecast. The list of selected variables does not vary much, for Specification Type 1. On the other hand, in Panels D and F, we see that the most frequently selected variables are not selected all the time. For example, in Panel

D, CPI:Apparel is selected over all periods and the 3 month Treasury bill yield is selected continuously, after 1979. Of further note is that interest-rate related variables (i.e. Treasury bills rates, Treasury bond rates, and spreads with Federal Funds Rate) are frequently selected, across all specification type, estimation window types, and forecast horizons. This confirms that in addition to their well established usefulness in linear models, yields and spreads remain important in nonlinear modelling contexts.

6 Concluding Remarks

This paper surveys factor models and shrinkage techniques, and presents the results of a "horse-race" in which mean-square-forecast-error (MSFE) "best" models are selected, in the context of a variety of forecast horizons, estimation schemes and sample periods. 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 plus exogenous variables models. For the majority of the target variables that we forecast, we find that various of these shrinkage methods, when combined with factor analysis (e.g. component-wise boosting), perform better than all other models. This suggests that diffusion index methodology is particularly useful when combined with other shrinkage methods, thus adding to the extant evidence of this finding (see Bai and Ng (2008a,b) and Stock and Watson (2005a). We also find that model averaging methods perform surprisingly poorly. Given the rather extensive empirical evidence to the contrary, when specifying linear prediction models, this is taken as further evidence of the usefulness of our more sophisticated nonlinear modelling approach.

References

Armah, N. A. and Swanson, N. R. (2010a). Seeing inside the black box: Using diffusion index methodology to construct factor proxies in large scale macroeconomic time series environments. *Econometric Reviews*, 29:476–510.

Armah, N. A. and Swanson, N. R. (2010b). Some variables are more worthy than others: New diffusion index evidence on the monitoring of key economic indicators. *Applied Financial Economics*, forthcoming.

Artis, M. J., Banerjee, A., and Marcellino, M. (2002). Factor forecasts for the UK. Discussion Paper 3119, Center for Economic Policy Research.

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

Bai, J. and Ng, S. (2006a). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74:1133–1150.

Bai, J. and Ng, S. (2006b). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, 131:507–537.

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

Bai, J. and Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, 4:607–629.

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

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

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

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.

Chipman, H., George, E. I., and Mcculloch, R. E. (2001). The practical implementation of bayesian model selection. *Lecture Notes -Monograph Series*, 38:65–134.

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

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

Clements, M. P. and Hendry, D. F. (1994). Towards a theory of economic forecasting. In Hargreaves, C., editor, *Non-stationary time series analyses and cointegration*, pages 9–52. Oxford University Press.

Clements, M. P. and Hendry, D. F. (1995). Macro-economic forecasting and modelling. *Economic Journal*, 105:1001–1003.

Clements, M. P. and Hendry, D. F. (2008). Intercept corrections and structural change. Working paper, Oxford University.

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: Oxford.

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

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

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

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

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*, 94:446–455.

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*, 100:381–427.

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

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 Associ-*

ation, 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*, 55:119–139.

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

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

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

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. Discussion Paper 5304, Centre for Economic Policy Research.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–95.

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

Newbold, P. and Harvey, D. (2002). Forecast combination and encompassing. In Clements, M. and Hendry, D., editors, *A Companion to Economic Forecasting.* pp268-283, Blackwell Press: Oxford.

Pesaran, M. H., Pick, A., and Timmermann, A. (2010). Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics*, forthcoming.

Pesaran, M. H. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137:134–161.

Ravazzolo, F., Paap, R., van Dijk, D., and Franses, P. H. (2008). Bayesian model averaging in the presence of strutural breaks. In M, W. and D, R., editors, *Forecasting in the Presence of Structural Breaks and Model Uncertainty.* pp 561-594, Elsevier: Amsterdam.

Ridgeway, G., Madigan, D., and Richardson, T. (1999). Boosting methodology for regression problems. In *The Seventh International Workshop on Artificial Intelligence and Statistics*, pages 152–161. Morgan Kaufmann: New York.

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

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

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

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. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23:405–430.

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

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

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: Armsterdam:.

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. Discussion Paper 5361, Center for Economic Policy Research.

Yuan, M. (2007). Nonnegative garrote component selection in functional anova models. In *Proceedings of The Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2, pp 660-666.

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

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*. North-Holland: Armsterdam.

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*, 67:301–320.

Table 1: Target Variables For Which Forecasts Are Constructed*

Series	Abbreviation	Y_{t+h}
Unemployment Rate	UR	Z_{t+1}-Z_t
Personal Income Less Transfer Payments	PI	ln(Z_{t+1}-Z_t)
10 Year Treasury Bond Yield	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	NNE	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)
S&P 500 Index	SNP	ln(Z_{t+1}-Z_t)
Gross Domestic Product	GNP	ln(Z_{t+1}-Z_t)

* Notes : Data used in model estimation and prediction construction are monthly U.S. figures for the period 1960:1-2009:5. The transformation used in forecast model specification and forecast construction is given in the last column of the table. See Section 4.1 for complete details.

Table 2: Models and Methods Used In Real-Time Forecasting Experiments*

Method	Description
AR(SIC)	Autoregressive model with lags selected by the SIC
ARX	Autoregressive model with exogenous regressors
Combined-ADL	Combined autoregressive distributed lag model
FAAR	Factor augmented autoregressive model
PCR	Principal components regression
Bagging	Bagging with shrinkage, c = 1.96
Boosting	Component boosting, 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	Arithmetic mean

* Notes: This table summarizes the model specification methods used in the construction of prediction models. In addition to the above pure linear, factor and shrinkage based methods, three different combined factor and shrinkage type prediction model specification methods are used in our forecasting experiments, including: Specification Type1 - Principal components are first constructed, and then prediction models are formed using the above shrinkage methods (ranging from bagging to NNG) to select functions of and weights for the factors to be used in our prediction moels. Specification Type 2 - Principal component models are constructed using subsets of variables from the large-scale dataset that are first selected via application of the above shrinkage methods (ranging from bagging to NNG). This is different from the above approach of estimating factors using all of the variables. Specification Type 3 - Prediction models are constructed using only the above shrinkage methods (ranging from bagging to NNG), without use of factor analysis at any stage. See Sections 3 and 4.3 for complete details.

Table 3. Relative Mean Square Forecast Errors: Recursive Estimation, Specification Type 1 (no lags)*

Panel A: Recursive, h = 1

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	12.713	0.009	40.975	0.003	0.012	0.001	2.477	0.021	0.004	0.573	0.008
ARX(SIC)	0.897	0.974	1.038	0.939	1.031	0.989	**0.900**	0.874	1.120	1.104	**0.916**
Combined-ADL	0.957**	1.052	0.987	1.030	1.019	0.938	0.977**	0.944	1.101*	1.002	1.093**
FAAR	**0.780****	0.902	0.950	0.916	0.969	**0.811***	0.961	0.804**	0.953	1.023	0.965
PCR	0.830**	**0.870**	1.019	**0.875**	**0.943**	0.922	1.764**	**0.800****	1.43**	1.018	0.973
Bagging	1.025	1.062	0.977	1.341*	1.167**	0.913	1.084	1.080	0.985	1.019	0.958
C-Boosting	0.902*	0.969	0.953	0.963	0.989	0.875**	0.949	0.848**	0.958	0.978	1.006
BMA(1/T)	0.899	0.965	0.954	0.954	0.991	0.873**	0.960	0.851**	0.972	0.989	1.018
BMA(1/N^2)	0.892*	0.969	0.947	0.954	0.991	0.866**	0.949	0.839**	0.969	0.987	1.012
Ridge	0.887**	0.964	**0.940**	0.963	1.000	0.885*	0.938	0.816**	0.969	1.006	0.996
LARS	0.913*	0.968	0.972**	0.977	0.984	0.954**	0.981	0.949**	0.977	0.982	0.995
EN	0.913*	0.969	0.972**	0.977	0.984	0.954**	0.981	0.95**	0.977	0.982	0.995
NNG	0.966**	0.98**	0.994	0.979*	0.984	0.95**	0.989	0.984*	0.989**	0.985	0.991
Mean	0.859**	0.933**	0.942**	**0.910	0.953	0.841**	0.910**	0.845**	**0.939****	**0.976**	0.940**

Panel B: Recursive, h = 3

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	12.857	0.009	47.642	0.004	0.014	0.001	5.173	0.023	0.005	0.620	0.009
ARX(SIC)	0.988	0.902	1.016*	0.981	0.945	0.940	1.000	0.895	1.000	1.028	1.032
Combined-ADL	0.977**	1.058	**0.998**	1.059*	1.045	0.948	0.955**	0.948	1.233**	1.010	1.109
FAAR	0.915	0.867**	1.026	**0.929**	0.936	**0.818****	0.895	0.866	1.006	1.052	1.058
PCR	**0.912**	**0.865****	1.004	0.930	**0.909**	0.835*	1.447**	0.859	1.164*	1.043	1.020
Bagging	1.062	1.071	1.013	1.168**	1.096	1.016	0.899	0.938	1.017	1.004	1.025
C-Boosting	0.935	0.924*	1.004	0.977	0.984	0.883*	0.852*	0.880	0.988	1.005	0.983
BMA(1/T)	0.946	0.935	1.006	0.992	0.983	0.868*	**0.852***	0.888	0.996	1.006	0.994
BMA(1/N^2)	0.932	0.920	1.008	0.988	0.984	0.861*	0.854*	0.881	0.994	1.011	0.996
Ridge	0.919	0.893**	1.012	0.982	0.991	0.866*	0.891	0.865	0.993	1.017	0.994
LARS	0.977	0.977**	1.003	0.992	0.993	0.984	0.926*	0.963	0.997	0.994	0.974
EN	0.977	0.977**	1.003	0.992	0.993	0.984	0.926*	0.963	0.996	**0.993**	0.974
NNG	0.980*	0.992*	1.005	0.990	0.990	0.989	0.984**	0.987*	0.996	1.003	0.985*
Mean	0.920*	0.898**	1.000	0.947	0.938**	0.858**	0.862**	**0.849****	**0.977**	0.998	**0.955**

Panel C: Recursive, h = 12

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	14.951	0.009	46.773	0.004	0.014	0.002	20.916	0.026	0.006	**0.620**	0.009
ARX(SIC)	1.014	0.993	1.001	1.004	1.006	0.991	1.000	0.995	1.000	1.046	1.000
Combined-ADL	0.980**	1.064	**0.996**	1.043	1.037	0.966	0.952**	0.952	1.212**	1.010	1.172**
FAAR	0.956	1.009	1.032	**0.886****	0.939	**0.874**	**0.818****	0.972	0.989	1.022	1.045
PCR	0.958	1.003	1.021	0.929	0.948	0.887	0.956	0.962	1.061	1.023	1.034
Bagging	1.072**	0.968	1.035	0.895**	0.993	1.178**	0.932	1.052*	0.982	1.003	1.008
C-Boosting	0.950	0.986	1.005	0.901**	0.955*	0.909	0.85**	0.954	0.989	1.007	1.010
BMA(1/T)	0.960	1.000	1.002	0.901*	0.955	0.922	0.852**	0.956	0.994	1.003	1.015
BMA(1/N^2)	0.959	0.997	1.004	0.903*	0.955	0.908	0.854**	0.955	0.995	1.005	1.020
Ridge	**0.939**	0.988	1.007	0.896**	0.954	0.892	0.875**	0.949	0.991	1.007	1.021
LARS	0.959	0.981	1.005	0.983**	0.985**	0.932	0.909**	0.936	0.993	1.008	1.001
EN	0.960	0.980	1.004	0.983**	0.985**	0.932	0.909**	0.936	0.992	1.008	1.001
NNG	0.975**	0.988*	1.010	0.992**	0.991**	0.975**	0.981**	0.967**	0.992	1.011	1.000
Mean	0.942	**0.955**	1.005	0.894**	**0.939****	0.875**	0.853**	**0.918**	**0.957****	1.001	**0.999**

*Notes: See notes to Tables 1 and 2. Numerical entries in this table are mean square forecast errors (MSFEs) based on the use of various recursively estimated prediction models. Forecasts are monthly, for the period 1974:3-2009:5. Models and target variables are predicted in Tables 1 and 2. Forecast horizons reported on include h=1,3 and 12. Entries in the first row, corresponding to our benchmark AR(SIC) model, are actual MSFEs, while all other entries are relative MSFEs, such that numerical values less than unity constitute cases for which the alternative model has lower point MSFE than the AR(SIC) model. Entries in bold denote point-MSFE "best" models for a given variable and forecast horizon. Dot-circled entries denote cases for which the Specification Type 1 (no lags) MSFE-best model using recursive estimation yields a lower MSFE than that based on using rolling estimation. Circled entries denote models that are MSFE-best across all specification types and estimation types (i.e. rolling and recursive). Boxed entries denote cases where models are "winners" across all specification types, when only viewing recursively estimated models. The results from Diebold and Mariano (19 accuracy tests, for which the null hypothesis is that of equal predictive accuracy between the benchmark model (defined to be the AR(SI model), and the model listed in the first column of the table, are reported with single starred entries denoting rejection at the 10% level, and double starred entries denoting rejection at the 5% level. See Sections 4 and 5 for complete details.

Table 4. Relative Mean Square Forecast Errors: Recursive Estimation, Specification Type 1 (with lags)*

Panel A: Recursive, h = 1

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	12.713	0.009	40.975	0.003	0.012	0.001	2.477	0.021	0.004	0.573	0.008
ARX(SIC)	0.897	0.974	1.038	0.939	1.031	0.989	0.900	0.874	1.120	1.104	0.916
Combined-ADL	0.957**	1.052	0.987	1.030	1.019	0.938	0.977**	0.944	1.101*	1.002	1.093**
FAAR	0.850*	0.926	1.044	0.888	1.008	1.005	1.079	0.851	0.968	1.095	1.050
PCR	0.908	0.888	1.058	0.864	1.002	0.999	1.646**	0.855	1.292**	1.091	1.076
Bagging	1.287**	1.017	1.069*	2.566**	1.545**	2.160**	1.851**	1.304**	1.028	1.131**	0.962
C-Boosting	0.903	0.968	0.961	0.951	1.002	0.910	0.945	0.827**	0.963	0.975	1.005
BMA(1/T)	0.910	0.972	0.988	0.942	1.018	0.904	0.956	0.804**	0.959	1.012	1.019
BMA(1/N^2)	0.907	0.962	0.996	0.955	1.023	0.904	0.954	0.816**	0.947	1.002	1.022
Ridge	0.911	0.959	0.988	0.919	1.014	0.944	0.992	0.821**	0.977	1.048	1.040
LARS	0.975**	0.977*	0.981	0.988	0.988	0.967*	0.974	0.948**	0.972*	0.989	0.995
EN	0.977**	0.978**	0.982	0.988	0.988	0.969*	0.975	0.949**	*0.970	0.989	0.992
NNG	0.972**	0.990	0.994	0.984	0.996	0.975	0.989	0.964**	0.993	0.993	0.994
Mean	0.867**	0.922**	0.955	0.889**	0.944	0.879**	0.922*	0.821**	**0.930	0.977	0.948*

Panel B: Recursive, h = 3

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	12.857	0.009	47.642	0.004	0.014	0.001	5.173	0.023	0.005	0.620	0.009
ARX(SIC)	0.988	0.902	1.016*	0.981	0.945	0.940	1.000	0.895	1.000	1.028	1.032
Combined-ADL	0.977**	1.058	0.998	1.059*	1.045	0.948	0.955**	0.948	1.233**	1.010	1.109
FAAR	1.014	0.931	1.106	0.907	0.992	0.886	0.898	0.925	1.069	1.117*	1.144
PCR	0.999	0.928	1.092	0.906	0.975	0.898	1.404**	0.921	1.249**	1.107	1.115
Bagging	1.174**	1.017	1.141**	1.339**	1.204*	1.295**	1.050	1.010	0.995	1.007	1.087**
C-Boosting	0.951	0.914*	1.010	0.946	0.969	0.832**	0.879	0.868	1.006	1.007	0.967
BMA(1/T)	0.944	0.932	1.020	0.943	0.982	0.818**	0.903	0.851	1.027	1.030	0.990
BMA(1/N^2)	0.954	0.942	1.011	0.953	0.981	0.836*	0.889	0.862	1.020	1.011	0.979
Ridge	0.944	0.917	1.047	0.933	0.992	0.844	0.891	0.869	1.046	1.064	1.033
LARS	0.979	0.973**	0.992	0.984	0.982	0.968	0.951**	0.962	0.996	1.000	0.969
EN	0.973*	0.975**	0.991	0.983	0.986	0.963	0.965**	0.962	0.996	1.000	0.969**
NNG	0.980	0.986*	1.001	0.991	0.995	0.963**	0.977**	0.967**	0.993	0.993	*0.970
Mean	0.924	0.891**	0.988	0.901**	0.928**	0.84**	0.851**	0.838**	0.977	0.997	0.962

Panel C: Recursive, h = 12

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	14.951	0.009	46.773	0.004	0.014	0.002	20.916	0.026	0.006	0.621	0.009
ARX(SIC)	1.014	0.993	1.001	1.004	1.006	0.991	1.000	0.995	1.000	1.046	1.000
Combined-ADL	0.980**	1.064	0.997	1.043	1.037	0.966	0.952**	0.952	1.212**	1.010	1.172**
FAAR	0.985	1.070	1.087	0.938	0.951	0.932	0.841*	1.082	1.049	1.081*	1.145**
PCR	0.983	1.069	1.081	0.932	0.942	0.924	1.020	1.071	1.116*	1.081*	1.132**
Bagging	1.003	**1.050	1.053	1.137*	1.078	1.174**	0.900	1.104**	0.971	1.034	1.001
C-Boosting	0.913	0.985	0.988	0.89**	0.947	0.896	0.846**	0.947	0.941	0.999	1.031
BMA(1/T)	0.930	1.007	1.002	0.908	0.935*	0.888	0.853**	0.975	0.981	1.006	1.031
BMA(1/N^2)	0.936	0.997	0.999	0.909*	0.952	0.907	0.833**	0.964	0.982	1.002	1.019
Ridge	0.926	1.005	1.029	0.897	0.931	0.867	0.887*	1.006	1.001	1.029	1.067
LARS	0.968**	0.973	0.992	0.974**	0.988	0.974*	0.923**	0.963*	0.973**	0.995	1.004
EN	0.969**	0.971	0.992	0.972**	0.989	0.963**	0.929**	0.965*	0.975**	0.994	1.003
NNG	0.979**	0.985	1.002	0.993	1.007	0.975**	0.967**	0.978*	0.994	0.998	0.999
Mean	0.902**	0.956	0.995	0.888**	0.927**	**0.860	0.829**	0.925	0.943**	0.999	1.010

*Notes: See notes to Table 3. Dot-circled entries denote cases for which the Specification Type 1 (lags) MSFE-best model using recursive estimation yields lower MSFE than using rolling estimation. Circled entries denote models that are MSFE-best across all specification types and estimation types (i.e. rolling and recursive). Boxed entries denote cases where models are "winners" across all specification types, when only viewing recursively estimated models.

Table 5. Relative Mean Square Forecast Errors: Recursive Estimation, Specification Type 2*

Panel A: Recursive, h = 1

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	12.713	0.009	40.975	0.003	0.012	0.001	2.477	0.021	0.004	0.573	0.008
C-Boosting	**0.891***	0.962	0.971	0.961	1.024	0.887	0.961	0.906	1.047	1.011	**0.865****
BMA(1/T)	0.896*	**0.956**	1.005	0.968	0.989	**0.870****	0.990	**0.864****	0.960	0.995	1.013
BMA(1/N^2)	0.900*	0.962	0.986	**0.945**	0.983	0.899*	**0.942**	0.893**	**0.926**	1.019	1.012
LARS	0.914**	0.994	0.972**	0.998	1.008	0.916**	0.978	0.996	0.982**	0.983	0.876**
EN	1.149*	1.217	1.118	3.646**	1.464**	2.804**	11.041**	1.186**	4.340**	1.092**	1.308**
NNG	0.993**	0.996*	0.997	0.999	1.000	0.991**	1.001*	0.997*	1.000	1.001	1.000
Mean	0.907**	0.963**	**0.968**	0.960	**0.979**	0.886**	0.953**	0.902**	0.951*	**0.984**	0.93**

Panel B: Recursive, h = 3

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	12.857	0.009	47.642	0.004	0.014	0.001	5.173	0.023	0.005	0.620	0.009
C-Boosting	**0.934**	**0.902**	1.028	**0.946**	1.020	**0.847****	**0.780**	**0.819***	1.016	1.017	0.985
BMA(1/T)	0.959	0.920	1.011	0.996	1.023	0.903	0.882	0.902	0.994	1.009	0.991
BMA(1/N^2)	0.946	0.937	1.006	1.005	1.011	0.912	0.871	0.890**	1.001	1.010	1.027
LARS	0.983	0.982**	1.000	0.996	1.005	0.968**	0.937	0.960*	0.990	0.998	0.994
EN	1.136**	1.206**	**0.961**	2.678**	1.280**	2.166**	5.287**	1.103*	3.488**	1.010	**1.240
NNG	0.997**	0.996**	1.000	0.997	0.998	0.995**	1.000	0.999	0.999	1.001	0.998**
Mean	0.943	0.922**	1.005	0.966	**0.994**	0.887**	0.827**	0.871**	**0.976**	**0.997**	**0.966**

Panel C: Recursive, h = 12

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	14.951	0.009	46.773	0.004	0.014	0.002	20.916	0.026	0.006	**0.620**	0.009
C-Boosting	0.936	0.976	1.031	0.907	0.972	**0.845***	**0.786****	0.940	0.962	1.016	1.006
BMA(1/T)	0.947	1.000	1.003	**0.902****	0.991	0.930	0.887*	0.959	0.997	1.004	1.011
BMA(1/N^2)	0.938	1.007	1.003	0.917*	0.975	0.920	0.881**	0.993	0.981	1.007	1.024
LARS	0.957	0.979	1.002	0.970**	0.979**	0.966**	0.910**	0.912**	**0.959****	1.006	**0.981**
EN	0.977	1.19**	**0.979**	2.497**	1.251**	1.242**	1.307**	0.977	3.206**	1.010	1.226**
NNG	0.997**	0.999	1.001	0.997**	0.997**	0.995**	0.997**	0.995**	0.999	1.002	0.999
Mean	**0.933**	**0.965**	1.004	0.913**	**0.966****	0.892**	0.846**	**0.925***	0.961*	1.004	0.994

*Notes: See notes to Table 3. Dot-circled entries denote cases for which the Specification Type 2 MSFE-best model using recursive estimation yields lower MSFE than using rolling estimation. Circled entries denote models that are MSFE-best across all specification types and estimation types (i.e. rolling and recursive). Boxed entries denote cases where models are "winners" across all specification types, when only viewing recursively estimated models.

Table 6. Relative Mean Square Forecast Errors: Recursive Estimation, Specification Type 3*

Panel A: Recursive, h = 1

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	12.713	0.009	0.000	0.003	0.012	0.001	2.477	0.021	0.004	0.573	0.008
ARX(SIC)	**0.897**	0.974	1.038	0.939	1.031	0.989	**0.900**	**0.873**	1.120	1.104	0.916
Combined-ADL	0.957**	1.052	**0.987**	1.030	1.019	0.938	0.977**	0.944	1.101*	1.002	1.093**
C-Boosting	0.944	0.965*	0.992	0.962	0.975	0.910	0.924*	0.936	1.010	**0.988**	0.915**
BMA(1/T)	1.012	1.137	1.059	1.541	1.223**	1.685**	1.250	0.980	1.193	1.231**	0.933
BMA(1/N^2)	0.933	0.985	1.028	1.018	1.089	1.042	1.066	0.891	1.131	1.077	0.911
Ridge	1.668**	1.575**	1.424**	1.547**	1.643**	1.743**	1.795**	1.789**	1.430**	1.688**	1.388**
LARS	1.952**	0.993	1.797**	0.998	1.008	0.914**	2.02**	1.008	0.978**	1.975**	0.875**
EN	1.057	0.994	1.116	0.998	1.008	0.916**	1.082	0.996	0.982**	1.258**	0.876**
NNG	0.993**	0.996*	0.997	0.999	1.000	0.991**	1.001*	0.997*	1.000	1.001	1.000
Mean	0.924	0.943*	0.995	**0.933**	**0.956**	0.826**	0.910	0.875**	**0.977**	1.045	0.873**

Panel B: Recursive, h = 3

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	12.857	0.009	47.642	0.004	0.014	0.001	5.173	0.023	0.005	**0.62**	0.009
ARX(SIC)	0.988	**0.902**	1.016*	0.981	**0.945**	0.940	1.000	0.895	1.000	1.028	1.032
Combined-ADL	0.977**	1.058	**0.998**	1.059*	1.045	0.948	0.955**	0.948	1.233**	1.010	1.109
C-Boosting	**0.943**	0.951*	1.010	0.999	1.016	0.899**	0.820**	0.886**	0.980	1.014	**0.974**
BMA(1/T)	1.154	1.022	1.241**	1.092	1.094	1.109	1.041	1.076	1.089	1.158*	1.168**
BMA(1/N^2)	0.969	0.922	1.025	1.047	1.034	0.877	0.941	**0.881**	1.063	1.034	1.011
Ridge	1.873**	1.517**	1.743**	1.362**	1.479**	1.675**	1.133	1.811**	1.447**	1.813**	1.95**
LARS	2.183**	0.977**	1.923**	0.997	1.006	0.962**	1.299	0.958**	0.989	2.099**	1.255**
EN	1.169	0.982**	1.319**	0.996	1.005	0.968**	0.828	0.96**	0.990	1.243**	0.994
NNG	0.997**	0.996**	1.000	0.997	0.998	0.995**	1.001	0.999	0.999	1.000	0.998**
Mean	0.991	0.911**	1.070	0.926*	0.953	0.859**	0.723**	0.881**	0.938*	1.033	0.992

Panel C: Recursive, h = 12

Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
AR(SIC)	14.951	0.009	46.773	0.004	0.014	0.002	20.916	0.026	0.006	**0.62**	0.009
ARX(SIC)	1.014	0.993	1.001	1.004	1.006	0.991	1.000	0.995	1.000	1.046	1.000
Combined-ADL	0.980**	1.064	**0.996**	1.043	1.037	0.966	0.952**	0.952	1.212**	1.010	1.172**
C-Boosting	**0.926**	**0.961**	1.015	0.934*	0.971	0.862**	0.874**	0.934	0.969	1.007	0.995
BMA(1/T)	1.233	1.073	1.152**	1.298**	1.199	1.760	**1.760	1.164	1.366**	1.082	1.254**
BMA(1/N^2)	1.019	1.009	1.039	1.127	1.106	1.447	1.618**	0.958	1.163*	1.017	1.074
Ridge	1.555**	1.807**	1.752**	1.382**	1.677*	1.859*	1.087	1.936**	1.316**	1.794**	1.925**
LARS	1.858**	0.979	1.983**	0.975*	0.979*	1.123	1.312	2.212**	0.957**	2.226**	0.983
EN	1.207	0.978	1.327**	0.97**	0.979**	0.966**	0.803**	**0.889**	0.959**	1.283**	**0.981**
NNG	0.997**	0.999	1.001	0.997**	0.997**	0.995**	0.997**	0.995**	0.999	1.002	0.999
Mean	0.960	0.966	1.076*	0.899**	**0.953**	0.885	0.840**	0.925	**0.910	1.047	1.011

*Notes: See notes to Table 3. Dot-circled entries denote cases for which the Specification Type 3 MSFE-best model using recursive estimation yields lower MSFE than using rolling estimation. Circled entries denote models that are MSFE-best across all specification types and estimation types (i.e. rolling and recursive). Boxed entries denote cases where models are "winners" across all specification types, when only viewing recursively estimated models.

Table 7. Forecast Experiment Summary Results*

Panel A: Summary of MSFE-"best" Models Across All Specification Types

	Recursive Estimation Window			Recursive and Rolling Estimation Windows		
	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
AR(SIC)	0	0	0	1	0	0
ARX(SIC)	1	0	0	1	0	0
Combined-ADL	0	0	0	0	0	0
FAAR	2	0	1	2	0	1
PCR	4	3	0	3	2	0
Bagging	0	0	0	0	0	0
C-Boosting	2	1	2	3	2	3
BMA(1/T)	0	1	0	0	0	0
BMA(1/N^2)	0	0	0	0	2	0
Ridge	1	0	0	1	0	0
LARS	0	0	1	0	0	1
EN	0	1	3	0	1	3
NNG	0	1	1	0	1	0
Mean	1	4	3	0	3	3

Panel B: Summary of MSFE-"best" Models

	Winners by Estimaton Window Type			Winners by Specification Type		
	h = 1	h = 3	h = 12	h = 1	h = 3	h = 12
Specification Type 1						
Rolling	2	2	3	7	4	5
Recursive	9	9	8			
Specification Type 2						
Rolling	5	9	4	4	6	5
Recursive	6	2	7			
Specification Type 3						
Rolling	3	2	2	0	1	1
Recursive	8	9	9			

*Notes: See notes to Table 3. Specification types are defined as follows. Specification Type1 - Principal components are first constructed, and then prediction models are formed using the above shrinkage methods (ranging from bagging to NNG) to select functions of and weights for the factors to be used in our prediction model. Specification Type 2 - Principal component models are constructed using subsets of variables from the largescale dataset that are first selected via application of the above shrinkage methods (ranging from bagging to NNG). This is different from the above approach of estimatiing factors using all of the variables. Specification Type 3 - Prediction models are constructed using only the above shrinkage methods (ranging from bagging to NNG), without use of factor analysis at any stage.

Table 8. Forecast Experiment Summary Results: Various Subsamples*

Panel A: Wins by Specification Type
h = 1, Recursive Estimation

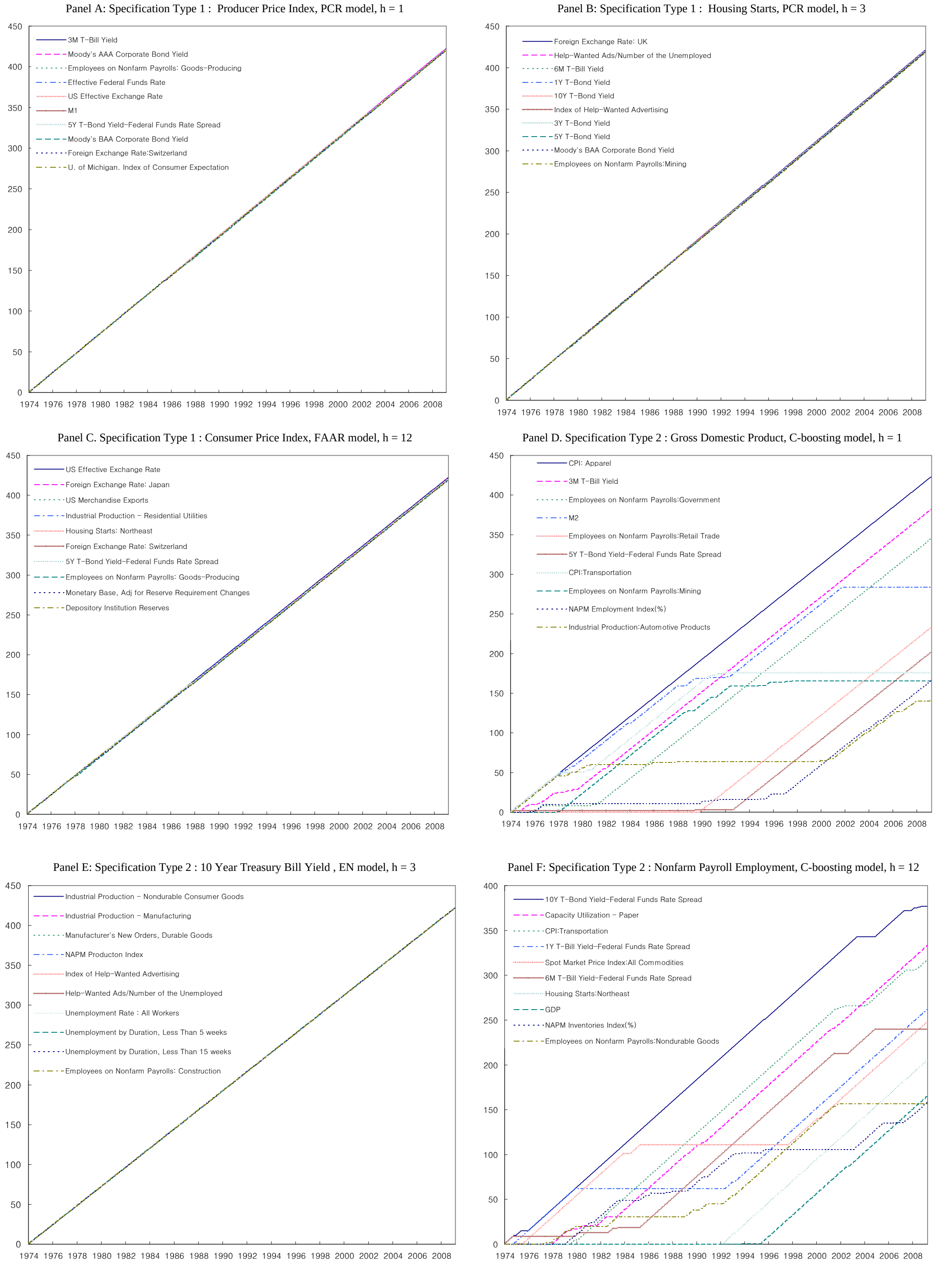
	Specification Type 1				Specification Type 2				Specification Type 3			
Subsample	Mean	Linear Factor	Nonlinear Factor	Other	Mean	Linear Factor	Nonlinear Factor	Other	Mean	Linear Factor	Nonlinear Factor	Other
75:03 ~ 79:12	3	1	5	2	4	0	6	1	3	0	5	3
80:07 ~ 81:06	1	4	2	4	5	0	5	1	6	0	2	3
82:11 ~ 90:06	1	8	2	0	8	0	3	0	4	0	4	3
91:03 ~ 01:02	5	2	2	2	6	0	5	0	8	0	1	2
01:11 ~ 07:11	5	0	4	2	6	0	5	0	5	0	2	4
Non Recession	1	6	2	2	8	0	3	0	7	0	3	1
Recession	3	5	1	2	5	0	6	0	7	0	2	2

Panel B: Wins Across All Specification Types
h = 1, Recursive Estimation

	Specification Type 1				Specification Type 2				Specification Type 3			
Subsample	Mean	Linear Factor	Nonlinear Factor	Other	Mean	Linear Factor	Nonlinear Factor	Other	Mean	Linear Factor	Nonlinear Factor	Other
75:03 ~ 79:12	2	1	3	1	0	0	2	0	1	0	1	0
80:07 ~ 81:06	0	4	0	2	1	0	3	0	1	0	0	0
82:11 ~ 90:06	1	8	0	0	1	0	0	0	1	0	0	0
91:03 ~ 01:02	3	1	2	0	0	0	0	0	4	0	0	1
01:11 ~ 07:11	0	4	3	1	1	0	0	0	2	0	0	0
Non Recession	1	5	2	0	0	0	1	0	2	0	0	0
Recession	1	3	0	1	0	0	1	0	4	0	1	0

*Notes: See notes to Tables 3 and 7. In the above table, "Mean" includes the following models: BMA, Combined-ADL and Mean. "Linear Factor" includes the following models: FAAR and PCR. "Nonlinear Factor" includes the following models: all shrinkage/factor combination models (i.e. Specification Types 1 and 2). Finally, "Other" includes our linear AR(SIC) and ARX(SIC) models. See Section 4.3 for further details.

Figure 1: Most Frequently Selected Variables by Various Specification Types*



Panel A: Specification Type 1 : Producer Price Index, PCR model, h = 1

Panel B: Specification Type 1 : Housing Starts, PCR model, h = 3

Panel C. Specification Type 1 : Consumer Price Index, FAAR model, h = 12

Panel D. Specification Type 2 : Gross Domestic Product, C-boosting model, h = 1

Panel E: Specification Type 2 : 10 Year Treasury Bill Yield , EN model, h = 3

Panel F: Specification Type 2 : Nonfarm Payroll Employment, C-boosting model, h = 12

3M T-Bill Yield
Moody's AAA Corporate Bond Yield
Employees on Nonfarm Payrolls: Goods-Producing
Effective Federal Funds Rate
US Effective Exchange Rate
M1
5Y T-Bond Yield-Federal Funds Rate Spread
Moody's BAA Corporate Bond Yield
Foreign Exchange Rate:Switzerland
U. of Michigan. Index of Consumer Expectation
Foreign Exchange Rate: UK
Help-Wanted Ads/Number of the Unemployed
6M T-Bill Yield
1Y T-Bond Yield
10Y T-Bond Yield
Index of Help-Wanted Advertising
3Y T-Bond Yield
5Y T-Bond Yield
Moody's BAA Corporate Bond Yield
Employees on Nonfarm Payrolls:Mining
US Effective Exchange Rate
Foreign Exchange Rate: Japan
US Merchandise Exports
Industrial Production - Residential Utilities
Housing Starts: Northeast
Foreign Exchange Rate: Switzerland
5Y T-Bond Yield-Federal Funds Rate Spread
Employees on Nonfarm Payrolls: Goods-Producing
Monetary Base, Adj for Reserve Requirement Changes
Depository Institution Reserves
CPI: Apparel
3M T-Bill Yield
Employees on Nonfarm Payrolls:Government
M2
Employees on Nonfarm Payrolls:Retail Trade
5Y T-Bond Yield-Federal Funds Rate Spread
CPI:Transportation
Employees on Nonfarm Payrolls:Mining
NAPM Employment Index(%)
Industrial Production:Automotive Products
Industrial Production - Nondurable Consumer Goods
Industrial Production - Manufacturing
Manufacturer's New Orders, Durable Goods
NAPM Producton Index
Index of Help-Wanted Advertising
Help-Wanted Ads/Number of the Unemployed
Unemployment Rate : All Workers
Unemployment by Duration, Less Than 5 weeks
Unemployment by Duration, Less Than 15 weeks
Employees on Nonfarm Payrolls: Construction
10Y T-Bond Yield-Federal Funds Rate Spread
Capacity Utilization - Paper
CPI:Transportation
1Y T-Bill Yield-Federal Funds Rate Spread
Spot Market Price Index:All Commodities
6M T-Bill Yield-Federal Funds Rate Spread
Housing Starts:Northeast
GDP
NAPM Inventories Index(%)
Employees on Nonfarm Payrolls:Nondurable Goods
0
50
100
150
200
250
300
350
400
450
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008

***Notes: Panels in this figure depict the 10 most commonly selected variables for use in factor construction, across the entire prediction period from 1974:3-2009:5, where factors are re-estimated at each point in time, prior to each new prediction being constructed. 45 degree lines denote cases for which a particular variables is selected every time. All models reported on are MSFE-best models, across Specification Types 1 and 2, and estimation window types. For example, in Panels A and B, the BAA Bond Yield - Federal Funds Rate spread is the most frequently selected predictor when constructing factors to forecast the Producer Price Index and Housing Starts, respectively. Note that in Panel E, the 10 most commonly selected variables by EN are picked at every point in time.