ECONOMETRIC ESSAYS ON NONLINEAR METHODS AND DIFFUSION INDEX FORECASTING

by

HYUN HAK KIM

A dissertation submitted to the

Graduate School - New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Economics

Written under the direction of

Norman Rasmus Swanson

and approved by

New Brunswick, New Jersey
May 2012

ABSTRACT OF THE DISSERTATION

Econometric Essays in Essays on Nonlinear Methods and Diffusion Index Forecasting

By HYUN HAK KIM

Dissertation Director:

Norman Rasmus Swanson

This dissertation comprises two essays in macroeconomic forecasting. The first essay empirically examines approaches to combining factor models and robust estimation, 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 window schemes and sample periods. For the majority of the target variables that we forecast, it is found that variety of these shrinkage methods, when combined with simple factors formed using principal component analysis (e.g. component-wise boosting), perform better than all other models. It is also found that model averaging methods perform surprisingly poorly, given our prior that they would "win" in most cases. The second essays outlines and discusses a number of interesting new forecasting methods that have recently been developed in the statistics and econometrics literature. It focuses in particular on the examination of a variety of factor modeling methods, including principal components, independent component analysis (ICA) and sparse principal component analysis (SPCA). Further, it outlines a number of approaches for creating hybrid forecasting models that use these factor modeling approaches in conjunction with various type of shrinkage methods. The results show that pure factor modeling approaches alone are not enough to lead to our overall finding that simple linear econometric models as well as models based on various forecast combination strategies are dominated by more complicated (factor/shrinkage) type models.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Norman Swanson, for his dedicated guidance and financial support through my research career. Dr. Swanson created an enabling environment for me to complete my essays in a timely manner. He was extremely helpful in the successful completion of my graduate career.

Many thanks are owed to Roger Klein for his patience and for assiduously explaining to me the basic concept in Econometrics. At a crucial point in my graduate career, Dr. Klein made me believe in myself again. I would like to thank John Landon-Lane for his great advices about Bayesian prospective in my research. Moreover, Dr. Landon-Lane gave me various intuitions in macroeconomics forecasting.

I also would like to thank Hiroki Tsurumi, Nii Ayi Armah, Bruce Mizrach for helpful comments as well as insightful discussions at various stages of my dissertation. I also thank Hyunjoong Kim and Dongjun Chung for introducing statistical learning theory. Special thanks to Dorothy Rinaldi for actively managing my graduate career. Ms. Rinaldi helped me even on the last day of her duty. She really saw to my welfare.

Finally, I would like to thank my parents for their immense sacrifices and for believing in me. I also thank to my brother for his unwavering support.

Dedication

To my family

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation considers the forecasting performance of various macroeconomic time series models using diffusion index models and shrinkage methods. The second chapter empirically assesses 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, least angle regression and related methods. 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.

The third chapter begins by summarizing a number of recent studies in the econometrics literature which have focused on the usefulness of factor models in the context of prediction (see Bai and Ng (2008), Dufour and Stevanovic (2010), Forni et al. (2000, 2005), Kim and Swanson (2010), Stock and Watson (2002b, 2005a, 2006), and the references cited therein).

This chapter adds to the literature by examining a number of novel factor estimation methods within the framework of diffusion index forecasting. In particular, it is considered the use of independent component analysis (ICA) and sparse principal component analysis (SPCA), coupled with a variety of other factor estimation as well as data shrinkage methods, including bagging, boosting, and the elastic net, among others. A number of forecasting experiments are carried out, involving the estimation of 28 different baseline model types, each constructed using a variety of specification approaches, estimation approaches, and benchmark econometric models; and all used in the prediction of 11 key macroeconomic variables relevant for monetary policy assessment. It is found that various our benchmarks, including autoregressive (AR) models, AR models with exogenous variables, and (Bayesian) model averaging, do not dominate more complicated nonlinear methods, and that using a combination of factor and other shrinkage methods often yields superior predictions. For example, SPCA yields mean square forecast error "best" (MSFE-best) prediction models in most cases, in the context of short forecast horizons. Indeed, benchmark econometric models in this chapter are never found to be MSFE-best, regardless of the target variable being forecast, and the forecast horizon. This is somewhat contrary to the oft reported finding that model averaging usually yields superior predictions when forecasting the types of aggregate macroeconomic variables that we examine. Additionally, pure shrinkage type prediction models and standard (linear) regression models, never MSFE-dominate models based on the use of factors constructed using either principal component analysis, independent component analysis or sparse component analysis. This result provides strong new evidence of the usefulness of factor based forecasting, although it should be stressed that principal component analysis alone does not yield this clear-cut result. Rather, it is usually ICA and SPCA type factor estimation approaches, often coupled with shrinkage, that yield the "best" models. Ancillary findings include the following: (i) Recursive estimation window strategies only dominate rolling strategies at the 1-step ahead forecast horizon. (ii) Including lags in factor model approaches does not generally yield improved predictions.

Chapter 2

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

2.1 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 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, Stock and Watson (2002a,b) as well as Bai and Ng (2006a) examine linear autoregressive type forecasting models augmented by the inclusion of common factors.

In this paper, use the forecasting models of Stock and Watson (2002a,b) as a starting point in an analysis of diffusion index and other shrinkage methods. In particular, 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. However, 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 specification (see e.g. Bai and Ng (2002, 2006b, 2008)). In light of this, and in order to add functional flexibility, we additionally implement prediction models 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 (2008, 2009), 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. (2011), 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, non-farm 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 (2011) 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."*

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 factor models. In Section 3, we survey the robust shrinkage estimation methods used in our prediction experiments. Data, forecasting methods, and benchmark forecasting models are discussed in Section 4, and empirical results are presented in Section 5. Concluding

---

[1] For further discussion of estimation windows and the related issue of structural breaks, see Pesaran et al. (2011).

remarks are given in Section 6.

2.2 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 Watson (1999, 2002a, 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 (2002a).

Let X_{tj} be the observed datum for the $j-$th cross-sectional unit at time t, for $t = 1, ..., T$ and $j = 1, ..., N$. We begin with the following model:

$$X_{tj} = F_t \Lambda_j' + e_{tj}, \tag{2.1}$$

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 a useful dimension reducing factor representation of the data, particularly when $r << N$, as is usually assumed to be the case in the empirical literature. Following Bai and Ng (2002), the whole panel of data $X = (X_1, ..., X_N)$ can be represented as (2.1). 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 (2002a) discuss consistent estimation of the factors when $N, T \to \infty$. 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 $F_t\Lambda' = F_tQQ^{-1}\Lambda'$, 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 (2002a) 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.

With regard to choice of r, note that 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. 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 discussion, see Bai and Ng (2002) and Armah and Swanson (2010b)):

$$SIC(r) = V\left(r, \hat{F}\right) + r\hat{\sigma}^2\left(\frac{(N+T-r)\ln(NT)}{NT}\right). \tag{2.2}$$

A consistent estimate of the true number of factors is $\hat{r} = \arg \min_{0 \leq r \leq r_{\max}} SIC(r)$. In a number of our models, we use this criteria for choosing the number of factors. However, as discussed above, we also use a variety of shrinkage methods to specify numbers and functions of factors to be used alternative prediction models. These shrinkage models, including bagging and other methods outlined in the introduction are also directly applied to our panel of data, without constructing factors.

The basic structure of the forecasting models examined in this paper is the same as that examined in Artis et al. (2002), Bai and Ng (2002, 2006a,b, 2008, 2009), Boivin and Ng (2005) and Stock and Watson (2002a, 2005a,b, 2006). In particular, we consider models of the following generic form:

$$Y_{t+h} = W_t \beta_W + F_t \beta_F + \varepsilon_{t+h}, \tag{2.3}$$

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 estimated using least squares. 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*. Forecasts of Y_{t+h} based on (2.3) 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 $\sqrt{T}/N \rightarrow 0$, 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].

2.3 Robust Estimation Techniques

We consider a variety of "robust" estimation techniques including statistical learning algorithms (bagging and boosting), as well as various penalized regression methods including ridge regression, least angle regression, elastic net, and the non-negative garotte. We also consider forecast combination in the form of Bayesian model averaging.

The following sub-sections provide summary details on implementation of the above

---

[2] We refer the reader to Stock and Watson (1999, 2002a, 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 (2010b) for further detailed discussion of generic diffusion models.

methods in contexts where in a first step we estimate factors using the principal components analysis, while in a second step we select factor weights using shrinkage. Approaches in which we first directly implement shrinkage 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. Note that all variables are assumed to be standardized in the sequel. Algorithms for the methods outlined below are given in the originating papers cited as well as discussed in some detail in Kim and Swanson (2011).

2.3.1 Statistical Learning (Bagging and Boosting)

2.3.1.1 Bagging

Bagging, which is a short for "bootstrap aggregation", was introduced by Breiman (1996). Bagging involves first drawing bootstrap samples from in-sample "training" data, and then constructing predictions, which are later combined. Thus, if a bootstrap sample based predictor is defined as $\hat{Y}_b^* = \hat{\beta}_b^* X_b^*$, where $b = 1, ..., B$ denotes the b-th bootstrap sample drawn from the original dataset, then the bagging predictor is $\hat{Y}^{Bagging} = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}_b^*$. In this paper, we follow Bühlmann and Yu (2002) and Stock and Watson (2005a) who note that that, asymptotically, the bagging estimator can be represented in shrinkage form. Namely:

$$\hat{Y}_{t+h}^{Bagging} = W_t \hat{\beta}_W + \sum_{j=1}^{r} \psi\left(t_j\right) \hat{\beta}_{Fj} \hat{F}_{t,j} \tag{2.4}$$

where $\hat{Y}_{t+h}^{Bagging}$ is the forecast of Y_{t+h} made using data through time t, and $\hat{\beta}_W$ is the least squares (LS) estimator from a regression of Y_{t+h} on W_t, where W_t is a vector of lags of Y_t as in (2.3) including a vector of ones, $\hat{\beta}_{Fj}$ is a LS estimator from a regression of residuals, $Z_t = Y_{t+h} - W_t \hat{\beta}_W$ on $\hat{F}_{T-h,j}$, and t_j is the t-statistic associated with $\hat{\beta}_{Fj}$, defined as $\sqrt{T}\hat{\beta}_{Fj}/s_e$, where s_e, is a Newey-West standard error, and ψ is a function specific to the forecasting method. In the current context we set:

$$\psi(t) = 1 - \Phi(t+c) + \Phi(t-c) + t^{-1}[\phi(t-c) - \phi(t+c)], \tag{2.5}$$

where c is the pretest critical value, ϕ is the standard normal density and Φ is the standard normal CDF. In this paper, we follow Stock and Watson (2005a), and set the pretest critical value for bagging, c to be 1.96.

2.3.1.2 Boosting

Boosting (see Freund and Schapire (1997)) is a procedure that builds on a user-determined set of functions (e.g. least square estimators), often called "learners" and uses the set repeatedly on filtered data which are typically outputs from previous iterations of the learning algorithm. The output of a boosting algorithm generally takes the form:

$$\hat{Y}^M = \sum_{m=1}^{M} \kappa_m f(X; \beta_m),$$

where the κ_m can be interpreted as weights, and $f(X; \beta_m)$ are function of the panel dataset, X. Friedman (2001) introduce "L_2Boosting", which takes the simple approach of refitting "base learners" to residuals from previous iterations.[3] Bühlmann and Yu (2003) a boosting algorithm fitting "learners" using one predictor at a, in contexts where a large numbers of predictors are available, in the context of *iid* data. Bai and Ng (2009) modify this algorithm to handle time-series. We use their "Component-Wise L_2Boosting" algorithm in the sequel, with least squares "learners".

As an example, consider the case where boosting is done on the original W_t data as well

---

[3] Other extensions of the original boosting problem discussed by Friedman (2001) are given in Ridgeway et al. (1999) and Shrestha and Solomatine (2006).

as factors, $\hat{F}_t$, constructed using principal components analysis; and denote the output of the boosting algorithm as $\hat{\mu}^M\left(\hat{F}_t\right)$. Then, predictions are constructed using the following model:

$$\hat{Y}_{t+h}^{Boosting} = W_t\hat{\beta}_W + \hat{\mu}^M\left(\hat{F}_t\right). \tag{2.6}$$

Evidently, when shrinkage is done directly on X_t, then $\hat{F}_t$ in the above expression is suitably replaced with X_t.

2.3.2 Penalized Regression (Least Angle Regression, Elastic Net, and Non-Negative Garotte)

Ridge regression, which was introduced by Hoerl and Kennard (1970), is likely the most well known penalized regression method (see Kim and Swanson (2011)) for further discussion. Recent advances in penalized regression have centered to some extent on the penalty function. Ridge regression is characterized by an L_2 penalty function. More recently, there has been much research examining the properties of L_1 penalty functions, using the so called Lasso (least absolute shrinkage and selection operator) regression method, as introduced by Tibshirani (1996), and various hybrids and generalizations thereof. Examples of these include least angle regression , the elastic net, and the non-negative garotte, all of which are implemented in our prediction experiments.

2.3.2.1 Least Angle Regression (LAR)

Least Angle Regression (LAR), as introduced by Efron et al. (2004), is based on a model-selection approach known as forward stage-wise regression, which has been extensively used to examine cross-sectional data (for further details, see Efron et al. (2004) and Bai and Ng (2008)). Gelper and Croux (2008) extend Bai and Ng (2008) to time series forecasting with many predictors. We implement the algorithm of Gelper and Croux (2008) when constructing the LAR estimator.

Like many other stagewise regression approaches, start with $\hat{\mu}^0 = \bar{Y}$, the mean of the target variable, use the residuals after fitting W_t to the target variable, and construct a first estimate, $\hat{\mu} = X_t\hat{\beta}$, in stepwise fashion, using standardized data, and in M iterations, say. Possible explanatory variables are incrementally examined, and their added to the estimator function, $\widehat{\mu}$, according to their explanatory power. Following the same notation as used above, in the case where shrinkage is done solely on common factors, the objective is to construct predictions,

$$\hat{Y}_{t+h}^{LAR} = W_t\hat{\beta}_W + \hat{\mu}^M(\hat{F}_t).$$

2.3.2.2 Elastic Net (EN)

Zou and Hastie (2005) point out that the lasso has undesirable properties when T is greater than N or when there is a group of variables amongst which all pairwise correlations are very high. They develop a new regularization method that they claim remedies the above problems. The so-called elastic net (EN) 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". Zou and Hastie (2005) In this paper, we use the algorithm of Bai and Ng (2008), who modify the naive EN to use time series rather than cross sectional data. To fix ideas, assume again that we are interested in X and Y, and that variables are standardized. For any fixed non-negative η_1 and η_2, the elastic net criterion is defined as:

$$L(\eta_1, \eta_2, \beta) = |Y - X\beta|^2 + \eta_2 |\beta|^2 + \eta_1 |\beta|_1, \tag{2.7}$$

where $|\beta|^2 = \sum_j^N (\beta_j)^2$ and $|\beta|_1 = \sum_j^N |\beta_j|$. The solution to this problem is the so-called naive elastic net, given as:

$$\hat{\beta}^{NEN} = \frac{\left(\left|\hat{\beta}^{LS}\right| - \eta_1/2\right)_{pos}}{1+\eta_2} sign\left\{\hat{\beta}^{LS}\right\}. \tag{2.8}$$

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. Zou and Hastie (2005), in the context of above naive elastic net, point out that there is double shrinkage in this criterion, which does not help to reduce the variance and may lead to additional bias so that they propose a version of the elastic net in which this double shrinkage is corrected. In this context, the elastic net estimator, $\hat{\beta}^{EN}$, is defined as:

$$\hat{\beta}^{EN} = \left(1+\eta_2\right)\hat{\beta}^{NEN}, \tag{2.9}$$

where η_2 is a constant, usually "optimized" via cross validation methods. Zou and Hastie (2005) propose an algorithm called "LAR-EN" to estimate $\hat{\beta}^{EN}$ using the LAR algorithm implemented in this paper.[4] In the current context, $\hat{\beta}^{EN}$ is either the coefficient vector associated with the $\hat{F}_t$ in a forecasting model of the variety given in (2.3), assuming that $\psi\left(\cdot\right) = 1$, or is a coefficient vector associated directly with the panel dataset, X_t.

2.3.2.3 NON-NEGATIVE GAROTTE (NNG)

The non-negative garotte (NNG), was introduced by Breiman (1995). This method is a scaled version of the least square estimator with shrinkage factors, and is closely related to the EN and LAR. 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. As usual, we begin by considering standardized X and Y. Assume that the following shrinkage factor is given: $q\left(\zeta\right) = \left(q_1\left(\zeta\right), q_2\left(\zeta\right), ..., q_N\left(\zeta\right)\right)'$, where $\zeta > 0$ is a

---

[4] We use their algorithm, which is discussed in more detail in Kim and Swanson (2011).

tuning parameter. The objective is to choose the shrinkage factor in order to minimize:

$$\frac{1}{2}\left\|Y-Gq\right\|^2+T\zeta\sum_{j=1}^{N}q_j, \qquad \text{subject to } q_j>0,\ j=1,..,N, \tag{2.10}$$

where $G=(G_1,..,G_N)'$, $G_j=X_j\widehat{\beta}_j^{LS}$, and $\widehat{\beta}^{LS}$ is the least squares estimator. The NNG estimator of the regression coefficient vector is defined as $\hat{\beta}_j^{NNG}=q_j\left(\zeta\right)\hat{\beta}_j^{LS}$, and the estimate of Y is defined as $\widehat{\mu}=X\hat{\beta}^{NNG}\left(\zeta\right)$, so that predictions can be formed in a manner that is analogous to that discussed in the previous subsections. Assuming, for example, that $X'X=I$, the minimizer of expression (2.10) has the following explicit form: $q_j\left(\zeta\right)=\left(1-\frac{\zeta}{(\hat{\beta}_j^{LS})^2}\right)_+$, $j=1,...,N$. 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. However, Zou (2006) shows that the NNG with ordinary least squares is also consistent, if N is fixed, as $T\rightarrow\infty$. Our approach is to start the algorithm with the least squares estimator, as in Yuan (2007), who outline a simple algorithm for the non-negative garotte that we use in the sequel.

2.3.3 Bayesian Model Averaging

In recent years, Bayesian Model Averaging (BMA) has been applied to many forecasting problems, and has been frequently shown to yield improved predictive accuracy, relative to approaches based on the use of individual models. For this reason, we include BMA in our prediction experiments; and we view it as one of our benchmark modeling approaches. For further discussion of BMA in a forecasting context, see Koop and Potter (2004), Wright (2008, 2009), and Kim and Swanson (2011)

In addition, for a concise discussion of general BMA methodology, see Hoeting et al. (1999) and Chipman et al. (2001). The basic idea of BMA starts with supposing interest focuses on Q possible models, denoted by $M_1,...,M_Q$, say. In forecasting contexts, BMA involves averaging target predictions, Y_{t+h} from the candidate models, with weights appropriately chosen. In a very real sense, thus, it resembles bagging. The key difference is that BMA puts little weight on implausible models, as opposed to other varieties of shrinkage discussed above that operate directly on regressors. The algorithm that we use for implementation of BMA follows closely Chipman et al. (2001), Fernandez et al. (2001a), and Koop and Potter (2004). For complete details, the reader is referred to Kim and Swanson (2011).

2.4 Data

Following a long tradition in the diffusion index literature, we examine monthly data observations on 144 U.S. macroeconomic time series for the period 1960:01 - 2009:5 ($N = 144, T = 593$)[5]. 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.[6]. 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, as noted in Armah and Swanson (2011), Kim and Swanson (2011), and on the Federal Reserve Bank of New York's website. 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 (2002a, 2005a) for complete details). Note that a full list of predictor variables is provided in the appendix to an earlier working paper version which is available upon request from the authors.

---

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

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

2.5 Forecasting Methodology

Using the transformed dataset, denoted above by X, factors are estimated by the method of principal component analysis discussed in Section 2. In Kim and Swanson (2011), factors are additionally estimated using independent component analysis and sparse principal component analysis. After estimating factors, 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 (2.3). *Specification Type 2:* Principal component models of the type given in (2.3) 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.

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, 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 windows. 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 lags 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 $P = T - R = 593 - 156 = 437$ observations, for $h = 1$). Moreover, the initial in-sample estimation period is adjusted so that the ex ante prediction sample length, P, remains fixed, regardless of the forecast horizon. For example, when forecasting the unemployment rate, when $h = 1$, the first forecast will be $\hat{Y}_{157}^{h=1} = \hat{\beta}_W W_{156} + \hat{\beta}_F \tilde{F}_{156}$, while in the case where $h = 12$, the first forecast will be $\hat{Y}_{157}^{h=12} = \hat{\beta}_W W_{145} + \hat{\beta}_F \tilde{F}_{145}$ In our rolling estimation scheme, the in-sample estimation period used to calibrate our prediction models is fixed at length 12 years. The recursive estimation scheme begins with the same in-sample period of 12 years (when $h = 12$), 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, thus, 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 to facilitate comparison across forecast horizons as well as models.

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

$$MSFE_{i,h} = \sum_{t=R-h+2}^{T-h+1} \left(Y_{t+h} - \hat{Y}_{i,t+h} \right)^2, \tag{2.11}$$

where $\widehat{Y}_{i,t+h}$ is the forecast for horizon h for the $i-$th model. 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.

In concert 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 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 $\hat{Y}^{PCR}_{t+h} = \hat{\alpha} + \hat{\gamma}\hat{F}_t$, where $\hat{F}_t$ is estimated via principal components using $\{X_t\}^T_{t=1}$, as in equation (2.3).

Factor Augmented Autoregression: Based on equations (2.3), forecasts are computed as $Y^h_{t+h} = \hat{\alpha} + \hat{\beta}_F \hat{F}_t + \hat{\beta}_W (L) Y_t$. 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 $\hat{Y}^{ADL}_{t+h} = \hat{\alpha} + \hat{\beta}_i (L) X_{i,t} + \hat{\phi}_i (L) Y_t$. The combined forecast is $\hat{Y}^{Comb,h}_{T+h|T} = \sum\limits_{t=1}^{n} w_i \hat{Y}^{ADL,h}_{T+h|T}$. 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.

2.6 Empirical Results

In this section, we discuss the results of our prediction experiments. For the case where models are estimated using recursive data windows, our results are gathered in Tables 3 to 6. Detailed results based on rolling estimation are omitted for the sake of brevity, although they are available upon request from the authors. Summary statistics based upon both estimation window types are contained 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 contain results for h =1, 3 and 12 month ahead 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 those of our AR(SIC) model. Entries in bold denote MSFE-"best" models for a given variable,forecast horizon, and specification type. For example, in Panel C of Table 3, the MSFE-best model for unemployment (UR), when h=12, is ridge regression, with a MSFE of 0.939. Recalling that all reported entries in Tables 3-6 are for recursively estimated models, note that in each table, dot-circled entries denote cases for which the MSFE-best model yields a lower MSFE than that based on using rolling estimation, under the same specification type. For example, the ridge regression MSFE of 0.939 discussed above (i.e. see Table 3, UR, $h = 12$) is not dot-circled because one of the models, under rolling window estimation, yields a lower MSFE, under Specification Type 1 without lags. However, the MSFE value for UR of 0.780 in Table 3, under $h = 1$ is dot-circled, denoting that no model yields a lower MSFE under rolling window estimation, for Specification Type 1 with no lags. This method of reporting allows us to compare rolling window estimation results without having to actually report rolling type MSFEs in our tables. Boxed entries denote cases where models are MSFE "winners" across all specification types (i.e. across Tables 3-6), when only viewing recursively estimated models. For example, in Panel A of Table 3, the MSFE-best value for HS is ARX(SIC), and the value is boxed, denoting the fact that this MSFE value is the lowest across all 4 tables (i.e. across all specification types), under recursive estimation. Note that we do not draw a box around ARX(SIC) in other specification since it is redundant, as the ARX(SIC) in each specification type is identical (only factor methods change across specifications types; benchmark linear models remain the same). However, the fact that the entry is not also dot-circled indicates that a lower MSFE arises for one of the models when estimated using rolling windows of data, for the specification reported on in this particular table. Finally, circled entries denote models that are MSFE-best across all specification and estimation window types. For example, in Panel A of Table 3 it is apparent that FAAR in the "universal" MSFE-best model for UR, at horizon $h = 1$.That is, if one method in recursive estimation "wins", we circle it and do not put a box around it, as this would be redundant information.

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 tables, 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 model yields a lower MSFE than the AR(SIC) model except bagging, when predicting the unemployment rate (UR), regardless of forecast horizon, with one exception (i.e. for $h = 12$, the ARX(SIC) model also has higher MSFE than the AR(SIC) model). Indeed, for most variables, there are various models that have lower point MSFEs than 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, other than when $h = 1$, regardless of specification type (compare Table 3-6). Still, even in this case, there are some models that outperform the AR(SIC) model, even for horizons other than $h = 1$, including the Combined-ADL model under Specifications 1 and 2 (see Tables 3-5, Panels B and C), and LAR or EN under Specification 3 (see Table 6, Panels B and C). 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. Evidence of this will is discussed in some detail below.

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.

Turning to Table 7, notice that the results reported in Panel A summarize findings of Tables 1-3. In particular, “wins" are reported across all specification types, so that each row of entries in the panel sum to 11, the number of target variables in our experiments. When comparing results for $h = 1, 3$, and 12, we see that forecasts constructed using our model averaging specifications (Combined-ADL, BMA, and Mean) yield MSFE-best predictions for

1/11 ($h = 1$), 5/11 ($h = 3$), and 3/11 ($h = 12$) variables when using only recursive estimation, and for 0/11 ($h = 1$), 5/11 ($h = 3$), and 3/11 ($h = 12$) variables, when using both recursive and rolling estimation windows. 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, when combining "wins" across all three forecast horizons in the right hand section of Panel A in Table 7, note that C-Boosting, Ridge, LAR, EN, and NNG "win" in 15/33 cases. 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. Finally, 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. This is perhaps 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.[7]

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

---

[7] 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, 1995, 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.)

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 expansionary periods. Although results drawn from 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 far fewer 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 forecasting experiments, 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 (2010b), 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.

2.7 Concluding Remarks

This paper empirically examines approaches to combining factor models and robust estimation, 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 window 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, the elastic net and the 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 simple factors formed using principal component 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 (2008, 2009), and Stock and Watson (2005a)).

We also find that model averaging methods perform surprisingly poorly, given our prior that they would “win” in most cases. Given the rather extensive empirical evidence suggesting the usefulness of model averaging when specifying linear prediction models, this is taken as further evidence of the usefulness of more sophisticated nonlinear modelling approaches.

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 (1995) 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, 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

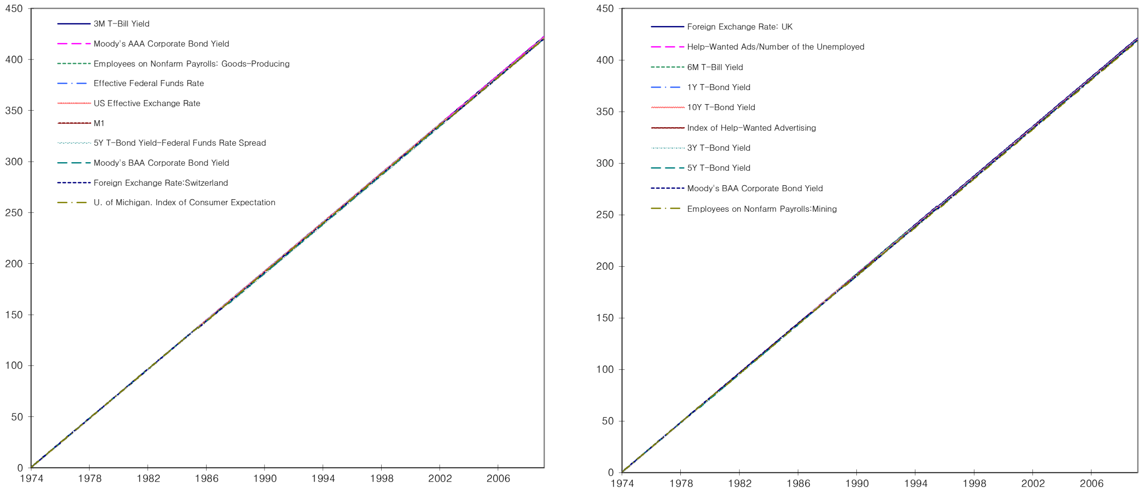
Panel B: Wins Across All Specification Types

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

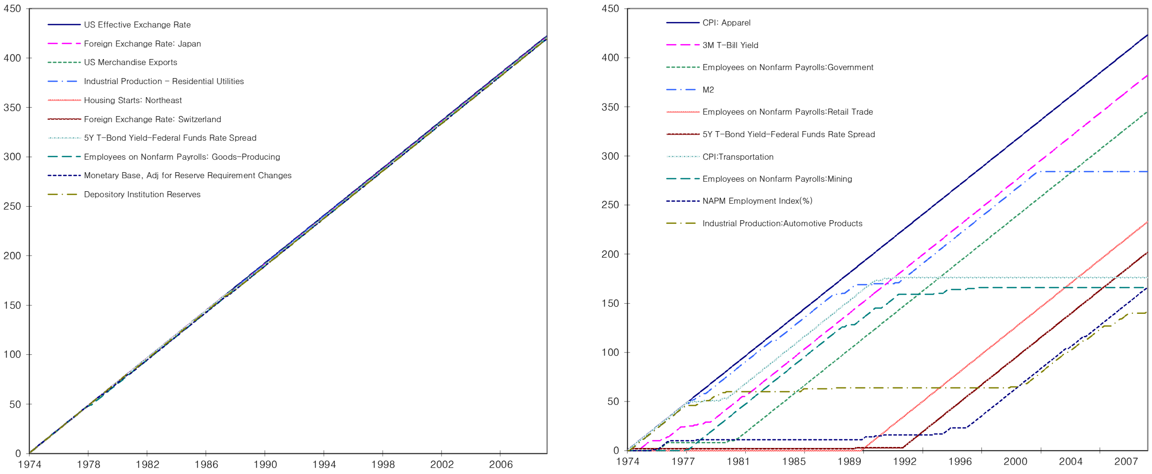
*Notes: See notes to Tables 3 and 7. In the above table, "Mean" includes the following models: BMA, Combined-ADL and Mean. "Linear Factor" includes the following models: FAAR and PCR. "Nonlinear Factor" includes the following models: all shrinkage/factor combination models (i.e. Specification Types 1 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*



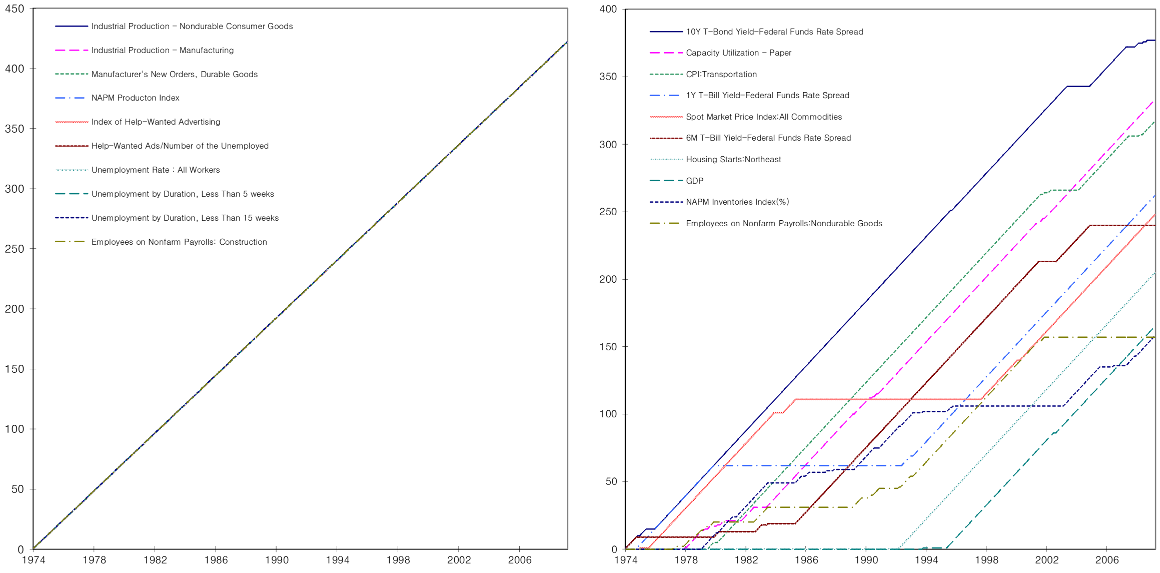
Panel A: Specification Type 1 : Producer Price Index, PCR model, h = 1

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

*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

Chapter 3

Forecasting Using Parsimonious Factor and Shrinkage Methods

3.1 Introduction

In macroeconomics and financial economics, researchers benefit greatly from a wealth of available information. However, available datasets are sometimes so large as to make dimension reduction an important consideration, both theoretical as well as empirical contexts. One dimension reduction technique, involving the construction of diffusion indices, has received consideration attention in the recent econometrics literature, particularly in the context of forecasting (see e.g. Armah and Swanson (2010a,b), Artis et al. (2002), Bai and Ng (2002, 2006b, 2008), Boivin and Ng (2005, 2006), Ding and Hwang (1999), Stock and Watson (2002a, 2005a,b, 2006)). Other recent important papers which extend the discussion in the above papers to vector and error-correction type models include Banerjee and Marcellino (2008), Dufour and Stevanovic (2010).

In this paper, we add to the extant literature on diffusion index forecasting by examining a number of novel factor estimation methods within the framework of diffusion index forecasting. In particular, we consider the use of independent component analysis (ICA) and sparse principal component analysis (SPCA), coupled with a variety of other factor estimation as well as data shrinkage methods, including bagging, boosting, least angle regression, the elastic net, and the nonnegative garotte. Our primary objective is the evaluation of the above estimation and shrinkage methods in the context of a large number of real-time out-of-sample forecasting experiments; and our venue for this "horse-race" is the prediction of 11 key macroeconomic variables relevant for monetary policy assessment. These variables include the unemployment, personal income, 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; and as noted in Kim and Swanson (2010) are discussed on the Federal Reserve Bank of New York's website, where it is stated that "In formulating the nation's monetary policy, the Federal Reserve considers a number of factors, including the economic and financial indicators <above>, as well as the anecdotal reports compiled in the Beige Book."

The notion of a diffusion index is to use appropriately "distilled" latent common factors extracted from a large number of variables in subsequent parsimonious models. More specifically, let X be an $T \times N$-dimensional matrix of observations, and define an $T \times r$-dimensional matrix of dynamic factors, F. Namely, let

$$X = F\Lambda' + e \tag{3.12}$$

where e is a disturbance matrix and Λ is an $N \times r$ coefficient matrix. Once F is extracted using one of the estimation methods examined in this paper, construct the following forecasting model based on Stock and Watson (2002a,b), Bai and Ng (2006a) and Kim and Swanson (2010). Namely, let Y_{t+h}, be an h-step ahead target variable to be predicted, and specify:

$$Y_{t+h} = W_t\beta_W + F_t\beta_F + \varepsilon_{t+h}, \tag{3.13}$$

where W_t is a $1 \times s$ vector and F_t is a $1 \times r$ vector of factors, extracted from F. The parameters, β_W and β_F are defined conformably, and ε_{t+h} is a disturbance term. In empirical contexts such as that considered herein, we first extract r unobserved factors, $\hat{F}$, from the N observable predictors, X. To achieve dimension reduction, r is assumed to be lower than N. (i.e. $r \leq N$) Then, parameter estimates, $\hat{\beta}_W$ and $\hat{\beta}_F$ are constructed using and in-sample dataset with Y_{t+h}, W_t, and $\hat{F}_t$. Finally, ex-ante forecasts based on rolling or recursive estimation schemes are formed. Although our approach is to consider various different specifications of the above model, the two issues that we primarily focus on include: (i) Which method is most useful for estimating the factors in the above model. In Kim and Swanson (2010), for example, principal component analysis (PCA) is used in obtaining estimates of the latent factors. PCA yields "uncorrelated" latent principal components via the use of data projection in the direction of the maximum variance; and principal components (PCs) are naturally ordered in terms of their variance contribution. The first PC defines the direction that captures the maximum variance possible, the second PC defines the direction of maximum variance in the remaining orthogonal subspace, and so forth. Perhaps because derivation of PCs is easily done via use of singular value decompositions, it is the most frequently used method in factor analysis (see e.g. Bai and Ng (2002, 2006b) and Stock and Watson (2002a) for details). In this paper, we additionally adopt two nonlinear methods for deriving latent factors, including ICA and SPCA. These nonlinear methods are used in the statistics research in a variety of contexts, although econometricians have yet to explore their usefulness in forecasting contexts, to the best of our knowledge.

ICA (see e.g. Comon (1994), Lee (1998)) uses a measure of entropy, so-called "negentropy" to construct independent factors. SPCA is designed to uncover *uncorrelated* components and ultimately factors, just like PCA. However, the method also searches for components whose factor loading coefficient matrices are "sparse" (i.e., the matrices can contain zeros). Since PCA yields nonzero loadings for entire set of variables, practical interpretation thereof is difficult, and estimation efficiency may become an issue. SPCA addresses these issues, yielding more parsimonious factors models. For further discussion, see Vines (2000), Jolliffe et al. (2003), and Zou et al. (2006), whose approach we follow in this paper. Further modifications of the approach in Zou et al. (2006) are discussed in Leng and Wang (2009) and Croux et al. (2011).

In order to add functional flexibility to our forecasting models, we additionally implement versions of (3.13) where the numbers and functions of factors used are specified via implementation of a variety of shrinkage methods, including those methods mentioned above. The key feature of our shrinkage methods is that they are used for targeted regressor and factor selection. Related research that focuses on shrinkage and related forecast combination methods is discussed in Stock and Watson (2005a), Aiolfi and Timmermann (2006), and Bai and Ng (2008). Moreover, our discussion and examination of shrinkage adds to the recent work of Stock and Watson (2005a) and Kim and Swanson (2010) who survey several methods for shrinkage that are based on factor augmented autoregression models like (3.13). In our experiments, we also consider various linear benchmark forecasting models including autoregressive (AR) models, AR models with exogenous variables, and combined autoregressive distributed lag models. Finally, we consider predictions at various different forecast horizons.

Our finding can be summarized as follows. Simple benchmark approaches based on the use of various AR type models, including Bayesian model averaging, do not dominate more complicated nonlinear methods that involve the use of factors, particularly when the factors are constructed using nonlinear estimation methods including ICA and SPCA. Moreover, in many cases, these nonlinear methods, when coupled with various shrinkage strategies yield overall mean square forecast error "best" (MSFE-best) prediction models. For example, SPCA yields mean square forecast error "best" (MSFE-best) prediction models in most cases, in the context of short forecast horizons. Indeed, our benchmark econometric models are never found to be MSFE-best, regardless of the target variable being forecast, and the forecast horizon. Recalling that Bayesian model averaging is one of our benchmarks, this finding is somewhat contrary to the oft reported finding that model averaging usually yields superior predictions when forecasting the types of aggregate macroeconomic variables that we examine. It is also noteworthy that pure shrinkage-based prediction models never MSFE-dominate models based on the use of factors constructed using either principal component analysis, independent component analysis or sparse component analysis. This result provides strong new evidence of the usefulness of factor based forecasting, although it should be stressed that principal component analysis alone does not yield this clear-cut result. Rather, it is usually ICA and SPCA type factor estimation approaches, often coupled with shrinkage, that yield the "best" models. Ancillary findings include the following: (i) Recursive estimation window strategies only dominate rolling strategies at the 1-step ahead forecast horizon. (ii) Including lags in factor model approaches does not generally yield improved predictions.

The rest of the paper is organized as follows. In the next section we provide a survey of dynamic factor models with independent component analysis and sparse component analysis. 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.

3.2 Diffusion Index Models

Recent forecasting studies using large-scale datasets and pseudo out-of-sample forecasting include: Armah and Swanson (2010a,b), Artis et al. (2002), Boivin and Ng (2005, 2006), Forni et al. (2005), and Stock and Watson (1999, 2002a, 2005a,b, 2006). Stock and Watson (2006) discuss in some detail the literature on the use of diffusion indices for forecasting. In this section, we begin by outlining the basic factor model framework which we use (see e.g. Stock and Watson (2002a,b) and Kim and Swanson (2010)). Thereafter we discuss independent component analysis and sparse principal component analysis.

3.2.1 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:

$$X_{tj} = \Lambda_j' F_t + e_{tj}, \tag{3.14}$$

where F_t is a $r \times 1$ vector of common factors, Λ_j is an $r \times 1$ vector of factor loadings associated with F_t, and e_{tj} is the idiosyncratic component of X_{tj}. The product $\Lambda_j' F_t$ is called the common component of X_{tj}. This is the dimension reducing factor representation of the data. More specifically, With $r < N$, a factor analysis model has the form

$$\begin{aligned} X_1 &= \lambda_{11} F_1 + \cdots + \lambda_{1r} F_r + e_1 \qquad (3.15) \\ X_2 &= \lambda_{21} F_1 + \cdots + \lambda_{2r} F_r + e_2 \\ &\vdots \\ X_N &= \lambda_{N1} F_1 + \cdots + \lambda_{Nr} F_r + e_N \end{aligned}$$

Here F is a vector of $r < N$ underlying latent variables or factors, λ_{ij} is a element of $N \times r$ matrix, Λ of factor loadings, and the ε are uncorrelated zero-mean disturbances. 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 generic forecasting equation is:

$$Y_{t+h} = W_t\beta_W + F_t\beta_F + \varepsilon_{t+h}, \tag{3.16}$$

where h is the forecast horizon, W_t is a $1 \times s$ vector (possibly including lags of Y), and F_t is a $1 \times r$ vector of factors, extracted from F. The parameters, β_W and β_F are defined conformably, and ε_{t+h} is a disturbance term. Following Bai and Ng (2002, 2006b, 2008, 2009), the whole panel of data $X = (X_1, ..., X_N)$ can be represented as (3.14). We then estimate the factors, F_t, via principal components analysis, independent component analysis and sparse principal component analysis. In particular, forecasts of Y_{t+h} based on (3.16) involve a two step procedure because both the regressors and the coefficients in the forecasting equation are unknown. The data, X_t, are first used to estimate the factors, yielding $\hat{F}_t$. 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 usual 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.[8]

In following sections, we introduce ICA and SPCA and underscore the difference between these methods and PCA. We omit detailed discussion of principal component analysis, given the extensive discussion thereof in the literature (see e.g. Stock and Watson (1999, 2002a, 2005a,b), Bai and Ng (2002, 2008, 2009), and Kim and Swanson (2010)).[9]

---

[8] We refer the reader to Stock and Watson (1999, 2002a, 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 (2010b) for further detailed discussion of generic diffusion index models.

[9] In the sequel, we assume that all variables are standardized, as is customary in this literature.

3.2.2 Independent Component Analysis

Independent Component Analysis (ICA) is of relevance in a variety of disciplines, since it is predicated on the idea of "opening" the black box in which principal components are often reside. A few uses of ICA include mobile phone signal processing, brain imaging, voice signal extraction and stock price modeling. In all cases, there is a large set of observed individual signals, and it is assumed that each signal depends on several factors, which are unobserved.

The starting point for ICA is the very simple assumptions that the components, F, are statistically independent in equation (3.14). The key is the measurement of this independence between components. The method can be graphically depicted as follows:

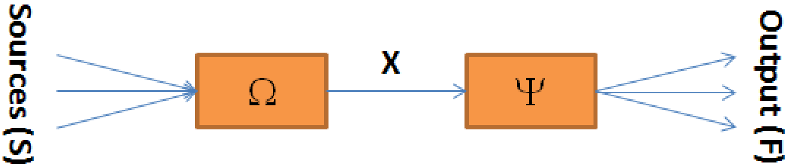


Figure 1: Schematic representation of ICA

More specifically, ICA begins with statistical independent source data, S, which are mixed according to Ω; and X which is observed, is a mixture of S weighted by Ω. For simplicity, we assume that the unknown mixing matrix, Ω, is square, although this assumption can be relaxed (see Hyvärinen and Oja (2000)). Using matrix notation, we have that

$$X = S\Omega \tag{3.17}$$

We can rewrite (3.17) as follows,

$$\begin{aligned} X_1 &= \omega_{11}S_1 + \cdots + \omega_{1N}S_N \qquad (3.18) \\ X_2 &= \omega_{21}S_1 + \cdots + \omega_{2N}S_N \\ &\vdots \\ X_N &= \omega_{1N}S_1 + \cdots + \omega_{NN}S_N \end{aligned}$$

where ω_{ij} is the (i,j) element of Ω. Since Ω and S are unobserved, we have to estimate the demixing matrix Ψ which transforms the observed X into the independent components F. That is,

$$F = X\Psi$$

or

$$F = S\Omega\Psi$$

Since we assume that mixing matrix, Ω is square, Ψ is also square, and $\Psi = \Omega^{-1}$, so that F is exactly same as S, and perfect separation occurs. In general, it is only possible to find Ψ such that $\Omega\Psi = PD$ where P is a permutation matrix and D is a diagonal scaling matrix (see Tong et al. (1991)).

The independent components, F are latent variables, just the same as principal components, meaning that they cannot be directly observed. Also, the mixing matrix, Ω is assumed to be unknown. All we observe is data, X, and we must estimate both Ω and S using it. Only then can we estimate the demixing matrix Ψ, and the independent components, F.

However (3.18) are not identified unless several assumptions are made. The first assumption is that the sources, S, are statistically independent. Since various sources (for example, consumer's behavior, political decisions, etc.) may impact macroeconomic variables, this assumption is not strong. The second assumption is that the signals are stationary. For further details, see Tong et al. (1991).

ICA under (3.18) assumes that N components of F exist. However, we can simply construct factors using only up to r $(< N)$ components, without loss of generality. In practice, we can construct r independent components by preprocessing with r principal components. See chapter 6 and 10 of Stone (2004) for further details. In general, the above model would be more realistic that there were noise terms added. For simplicity, however, noise terms are omitted; and indeed the estimation of the noise-free model is already computationally difficult (see Hyvärinen and Oja (2000) for a discussion of the noise-free model, and Hyvärinen (1998, 1999a) for a discussion of the model with noise added).

3.2.2.1 Comparison with Principal Component Analysis

As is evident from Figure 1, ICA is exactly the same as PCA, if we let demixing matrix be the factor loading coefficients associated with principal components analysis. The key difference between independent component analysis ICA and principal component analysis PCA is in the properties of the factors obtained. Principal components are uncorrelated and have descending variance so that they can easily be ordered in terms of their variances. Moreover, those components explaining the largest share of the variance are often assumed to be the "relevant" ones for subsequent use in diffusion index forecasting. In particular, the first principal component captures the maximum variance possible, the second component also capture the maximum variance but in an orthogonal subspace, and is thus uncorrelated with the first component.

For simplicity, consider two observables, $X = (X_1, X_2)$. PCA finds a matrix which transforms X into uncorrelated components $F = (F_1, F_2)$, such that the uncorrelated components have a joint probability density function, $p_F(F)$ with

$$E(F_1 F_2) = E(F_1) E(F_2). \tag{3.19}$$

On the other hand, ICA finds a demixing matrix which transforms the observed $X = (X_1, X_2)$ into independent components $F^* = (F_1^*, F_2^*)$, such that the independent components have a joint pdf $p_{F^*}(F^*)$ with

$$E\left[F_1^{*p} F_2^{*q}\right] = E\left[F_1^{*p}\right] E\left[F_2^{*q}\right], \tag{3.20}$$

for every positive integer value of p and q. That is, it works for any moments.

Evidently, PCA estimation is much simpler than ICA, since it just involves finding a linear transformation of components which are uncorrelated. Moreover, PCA ranks components using their variances or correlation so that components associated with higher variance or correlation are considered more explanation power than those with lower variance or correlation. On the other hand, ICA is unable to find the variance associated with each independent component since both S and Ω in (3.17) are unknown so that any scalar multiplier in one of the source, S_j could be cancelled by dividing the corresponding mixing vector, ω_j by the same scalar. Therefore, we change randomly the order of X, in (3.17) so that we cannot determine the order of the independent components. From the perspective of forecasting, this is probably a good thing, since there is no a prior reason to believe that “largest variance” PCA components are the most relevant for predicting any particular target variable. Moreover, this feature of ICA is the reason that PCA for pre-processing in ICA algorithms. For further details about preprocessing, see Appendix F of Stone (2004).

3.2.2.2 Estimation of ICA

Estimation of independent components is done by estimating the demixing matrix iteratively, systematically increasing the degree of independence of the components. As noted above, uncorrelated components are not independent (except under Gaussianity). However, there is no direct measure for independence. The standard approach is instead to use so-called “nongaussianity” as a measure of independence. In contrast, Gaussian variables cannot produce independent components. This is straightforward since the distribution of any orthogonal transformation of two independent and Gaussian random variables, say X_1 and X_2, has the same distribution as that of X_1 and X_2, in turn implying that the mixing matrix, Ω cannot be identified.

For simplicity, let all independent components the same distribution. For the first independent component, consider a linear combination of X_j, $j = 1, ..., N$ so that $F_j = X\Psi_j$, where Ψ_j is a vector to be estimated. If Ψ_j were one of the rows of the inverse of Ω, this linear combination would equal one of the independent components. In practice, it is not possible to obtain such a Ψ_j exactly since matrix Ω is not observed. Instead let $\Xi = \Psi\Omega$. Then, we can express F_j as a linear combination of the unobserved source S because

$$F_j = X\Psi_j = S\Omega\Psi_j = S\Xi_j$$

and weights of combination are given by Ξ_j . Note also that a sum of two independent random variables is in a concrete sense more Gaussian than the original variables, given a central limit theorem. Therefore, $S\Xi_j$ is more Gaussian than any of the $S's$. In practice, the objective is to extract Ψ_j as a vector maximizing the nongaussianity of $X\Psi_j$. This in turn implies that $X\Psi_j = S\Xi_j$ is an independent component.

Measuring Nongaussianity In this section we discuss how we measure nongaussianity. The easiest way is via use of kurtosis.

1. Kurtosis: $kurt(F) = E\left[F^4\right] - 3\left(E\left[F^2\right]\right)^2$, which is zero under Gaussianity. However, this measure is very sensitive to outliers, and so is not particularly useful for measuring nongaussianity.

2. Entropy−Negentropy: Another way of measuring nongaussianity or independence is

entropy. The differential entropy H of a random variable F with pdf, p_F is defined as

$$H(F) = -\int p_F(f) \ln p_F(f)\, dF \tag{3.21}$$

Note that a moment of a pdf can be expressed as an expectation, and (3.21) can thus be expressed as

$$H(F) = -E\left[\ln p_F(f)\right]. \tag{3.22}$$

A fundamental result of information theory is that a Gaussian variable has the largest entropy among all random variables of equal variance. This supports the use of entropy as a measure of nongaussianity. Moreover, entropy tends to be smaller when the distribution is dense around a certain value. Based on these results, one often uses a modified version of entropy, so called negentropy, N,where:

$$N(F) = H(F_{gauss}) - H(F), \tag{3.23}$$

where F_{gauss} is a Gaussian random variable with the same covariance matrix as F. This negentropy, $N(\cdot)$, as a measure of nongaussianity, is zero for a Gaussian variable and always nonnegative. Comon (1994), Hyvärinen (1999b) and Hyvärinen and Oja (2000) note that negentropy has additional interesting properties, including noting that it is invariant for invertible linear transformations.

3. Mutual Information: This measure is of the amount of information each variable contains about each other variable. Namely, it is the difference between the sum of individual entropies and the joint entropy of two variables, and is defined as follows:

$$I(F) = \sum_{i=1}^{n} H(F_i) - H(F), \tag{3.24}$$

for n random variables. The quantity $I(F)$ is equivalent to the Kullback-Leiber distance between density $g(F)$ of F and its independence version $\prod_{i=1}^{n} g_i(F_i)$, where $g_i(F_i)$ is the marginal density of F_i. The mutual information becomes zero if the variables are statistically independent. This is somewhat similar to negentropy. If we have an invertible linear transformation $F = X\Psi$, then

$$I(F) = \sum_{i=1}^{n} H(F_i) - H(X) - \ln|\det \Psi| \tag{3.25}$$

becomes

$$I(F) = \sum_{i=1}^{n} H(F_i) - H(X). \tag{3.26}$$

Finding Ψ to minimize $I(F) = I(X\Psi)$ involves looking for the orthogonal transformation that leads to the most independence between its components; and this is equivalent to minimizing the sum of the entropies of the separate components of F. That is, minimizing of mutual information is equivalent to finding directions where negentropy is maximized.

Estimation of Entropy Negentropy is well known and understood in statistics literature, and is the optimal estimator of nongaussianity in contexts such as that considered here. A classical approximation of negentropy using higher-order moments is the following:

$$N(F) \approx \frac{1}{12} E\left[F^3\right]^2 + \frac{1}{48} kurt(F)^2. \tag{3.27}$$

Another approximation from Hyvärinen (1998) is based on the maximum-entropy principle, does not explicitly include a measure of kurtosis, and is defined as follows:

$$N(F) \approx \Sigma_j k_j \left[E\{G_j(F)\} - E\{G_j(Z)\}\right]^2, \tag{3.28}$$

where k_i are positive constants, Z is a standardized Gaussian variable, F is standardized, and the functions G_i are some nonquadratic functions. Note that (3.28) can be used consistently, in the sense that it is always non-negative, and equals zero if F has a Gaussian distribution. Simple version of this approximation use only one nonquadratic function, G, leading to:

$$N(F) \propto \left[E\left\{G(F)\right\} - E\left\{G(\nu)\right\}\right]^2. \tag{3.29}$$

This equation is a generalization of (3.27), when F is symmetric. If one sets G as the quartic, (3.29) becomes (3.27). Therefore, choosing an appropriate G function is important. If we pick non-fast growing G, we may have more robust estimators. Hyvärinen and Oja (2000) suggest two Gs, and they show that these functions yield good approximations. They are:

$$G_1(y) = \frac{1}{a_1} \log\cosh a_1 y \tag{3.30}$$

and

$$G_2(y) = -\exp\left(-u^2/2\right), \tag{3.31}$$

where $1 \leq a_1 \leq 2$ is some suitable constant.

3.2.2.3 ICA Algorithm: FastICA

ICA implementation involves finding a direction for a unit vector, Ψ_j, such that the component projection matrix, $X\Psi_j$, maximized nongaussianity. In this paper, we estimate negentropy to measure nongaussianity via the "FastICA" algorithm which efficiently minimize negentropy. The FastICA is a popular ICA algorithm which is based on a fixed-point scheme for tracking maximal nongaussianity of the projection matrix, $X\Psi = \{X\Psi_1, ..., X\Psi_n\}$ where Ψ_j is the column vectors of Ψ, which are not correlated with each other. Simply put, the algorithm finds a unit vector Ψ_j such that $X\Psi_j$ maximizes nongaussianity which is measured by negentropy, as given by (3.29). Note that the variance of $X\Psi_j$ is constrained to be unity, so that the norm of Ψ_j is constrained to be unity, since we use standardized data.

Let g be the derivative of the nonquadratic function G used in (3.30) and (3.31). FastICA is the fixed-point algorithm which maximizes (3.29), and maxima are obtained at optima $E\left\{G\left(F\right)\right\} = E\left\{G\left(X\Psi\right)\right\}$. Let us explain with one unit for simplicity. Using Kuhn-Tucker conditions, the optima of $E\left\{G\left(X\Psi_j\right)\right\}$ under the constraint $E\left\{G\left(X\Psi_j\right)^2\right\} = \|\Psi_j\| = 1$ can be obtained at $E\left\{G\left(X\Psi_j\right)\right\} - \lambda\Psi_j = 0$. One can solve this equation using the Newton-Rhapson method. See Hyvärinen and Oja (2000) for computational details. Thus, we have the following iterative procedure:

$$\Psi^* = \Psi - \frac{\left[E\left\{Xg\left(X\Psi_j\right)\right\} - \lambda\Psi_j\right]}{\left[E\left\{g\left(X\Psi_j\right)\right\} - \lambda\right]} \tag{3.32}$$

Multiplying both sides by $\lambda - E\left\{g'\left(X\Psi_j\right)\right\}$ yields

$$\Psi^* = E\left\{Xg\left(X\Psi_j\right)\right\} - E\left\{g'\left(X\Psi_j\right)\right\}\Psi_j \tag{3.33}$$

Here is the basic form of FastICA algorithm.

1. Choose an initial weight vector Ψ.

2. For $j = 1, ..., r$, find mixing vectors yielding components with minimized negentropy. Let $\Psi_j^* = E\left\{Xg\left(X\Psi_j\right)\right\} - E\left\{g'\left(X\Psi_j\right)\right\}\Psi_j$.

3. Set $\Psi_j^+ = \Psi_j^* / \left\|\Psi_j^*\right\|$. If convergence is not achieved, go back to Step 2.

4. To decorrelate j independent components, for $j \geq 2$, set (a) $\Psi_j^+ = \Psi_j^+ - \Sigma_{h=1}^{j-1}\Psi_j^{+\prime}\Psi_h\Psi_h$ and then (b) $\Psi_j^+ = \Psi_j^+ / \sqrt{\left\|\Psi_j^{+\prime}\Psi_j^+\right\|}$.

The initial vector Ψ is given from the loading of the r ordinary principal components (Penny et al. (2001), Stone (2004)) Once the final Ψ is estimated, $X\Psi$ are the independent components. In this paper, we choose G as in (3.30) and accordingly g is defined as $\tanh(u)$ if we set $a_1 = 1$.

3.2.3 Sparse Principal Component Analysis

As was explained in the previous section, principal components are linear combinations of variables that are ordered by covariance contributions, and selection is of a small number of components which maximize the variance that is explained. However, factor loading coefficients are all typically nonzero, making interpretation of estimated components difficult. SPCA aids in the interpretation of principal components by placing (zero) restrictions on various factor loading coefficients.

For example, Jolliffe (1995) modifies loadings to be a values such as 1, -1 or 0, for example. Another approach is setting thresholds for the absolute value of the loadings, below which loadings are set to zero. Jolliffe et al. (2003) suggests using so-called "SCoTLASS" to construct modified principal components with possible zero loadings, λ, by solving

$$\max \lambda'(X'X)\lambda, \text{ subject to } \sum_{j=1}^{N} |\lambda_j| \leq \varphi, \ \lambda'\lambda = 1,$$

for some tuning parameter φ. The absolute value threshold results in (various) zero loadings, hence inducing sparseness. However, the SCoTLASS constraint does not ensure convexity, and therefore the approach may be computationally expensive. As an alternative, Zou et al. (2006) develop a regression optimization framework. Namely, they consider X as a dependent variables, F as explanatory variables, and the loadings as coefficients. They then use of the lasso (and elastic net) to derive a sparse loading matrix. Other recent approaches include those discussed in Leng and Wang (2009), Guo et al. (2010), all of which are based on Zou et al. (2006). We follow the approach of Zou et al. (2006), and readers are referred to Section 3.3-3.5 of the paper for complete details. As an introduction to the method, the following paragraphs draw on some of the key methods of the paper.

3.2.3.1 Estimation of Sparse Principal Components

Suppose we derive principle components (PCs), F via ordinary PCA. In particular, our standardized data matrix, X is identical to UDV' by the singular value decomposition. The PCs, F, are defined as UD, and V are the factor coefficient loadings. Then, let the estimated j-th principal component , F_j be the dependent variable and X be the independent variables. Suppose that $\hat{\lambda}_j^{Ridge}$ is the ridge estimator[10] of the loading for j-th principal component, we have to solve a following problem to get the ridge estimator,

$$\hat{\lambda}_j^{Ridge} = \arg\min_{\lambda_j} \|F_j - X\lambda_j\|^2 + \eta \|\lambda_j\|^2 \tag{3.34}$$

Note that after normalization, the coefficients are independent of η, therefore the ridge penalty term, $\eta \|\lambda_j\|^2$, is not used to penalize the regression coefficients but rather in the construction of the principal components. Add an L_1 penalty to (3.34) and solve the following optimization problem; namely, solve the so-called naïve elastic net (NEN) (see Section 3.4 for details on the NEN), as follows:

$$\hat{\lambda}_j^{NEN} = \arg\min_{\lambda_j} \|F_j - X\lambda_j\|^2 + \eta \|\lambda_j\| + \eta_1 \|\lambda_j\|_1 , \tag{3.35}$$

where $\|\lambda_j\|_1 = \sum_{i=1}^{N} |\lambda_{ij}|$. Thus, $X\hat{\lambda}_j$ is the j-th principal component. In this problem, large enough η_1 guarantees a sparse λ, and hence a sparse loading matrix. With a fixed value of η, the problem (3.35) can be solved using the LAR-EN algorithm[11] proposed by Zou and Hastie (2005). Zou et al. (2006) modify this idea to a more general lasso regression type problem. In particular, they use a two-stage analysis in which they first estimate the

---

[10] See Section 3.3 for further details about the ridge estimator.

[11] See Section 3.5 for details about the LAR-EN algorithm.

principal components by the ordinary PCA, and thereafter find the sparse loadings using (3.35). This type of SPCA is predicated on the fact that PCA can be written as a penalized regression problem[12], and thus the lasso, or the elastic net, can be directly integrated into the regression criterion such that the resulting modified PCA produces sparse loadings.

Continuing the above discussion, note that Zou et al. (2006) suggest using the following penalized regression type criterion. Let X_t denote the t-th row vector of the matrix X. For any positive value of η, let

$$\left(\hat{\delta}_j, \hat{\lambda}_j\right) = \arg\min_{\delta_j, \lambda_j} \sum_{t=1}^{T} \left\| X_t - \delta_j \lambda_j' X_t \right\|^2 + \eta \left\| \lambda_j \right\|^2 . \tag{3.36}$$

$$\textit{subject to} \quad \left\| \delta_j \right\|^2 = 1$$

Then, $\hat{\lambda}_j$ becomes the approximation to the j-th factor loadings, λ_j. If we let λ equal δ, then $\sum_{t=1}^{T} \left\| X_t - \delta_j \lambda_j' X_t \right\|^2 = \sum_{t=1}^{T} \left\| X_t - \delta_j \delta_j' X_t \right\|^2$. Therefore, $\hat{\lambda}(= \hat{\delta})$ becomes the j-th ordinary principal component's loading. (See Hastie et al. (2009) for details.) (3.36) can be easily extended to derive the whole sequence of PCs. Let there be r components. Set Δ and Λ to be $N \times r$ matrices. For any positive value of η, let

$$\left(\hat{\Delta}, \hat{\Lambda}\right) = \arg\min_{\Delta, \Lambda} \sum_{t=1}^{T} \left\| X_t - \Delta \Lambda' X_t \right\|^2 + \eta \sum_{j=1}^{r} \left\| \lambda_j \right\|^2 \tag{3.37}$$

$$\textit{subject to} \quad \Delta' \Delta = I_r.$$

Here, Λ is an $N \times r$ matrix with column λ_j and Δ is also an $N \times r$ orthonormal constraint, so that $\hat{\lambda}_j$ is the approximation to the j-th factor loadings, λ_j, for $j = 1, ..., r$. As we see in the above expression, by setting Δ and Λ to be equal, $\hat{\Lambda}$ becomes the exact r factor loadings of ordinary principal components. (3.37) is the generalized derivation of principal components and enables us to obtain sparse loadings by modifying the original PCA problem. The

---

[12] See Section 3.2 of Kim and Swanson (2010) for penalized regression

penalty parameter in the above expression is applied for all variables, and so we do not yet have sparse loadings, however. To construct sparsity, add the lasso penalty into the problem (3.37), and consider the following penalized regression problem,

$$\left(\hat{\Delta},\hat{\Lambda}\right)=\arg\min_{\Delta,\Lambda}\sum_{t=1}^{T}\left\|X_t-\Delta\Lambda' X_t\right\|^2+\eta\sum_{j=1}^{r}\left\|\lambda_j\right\|^2+\sum_{j=1}^{r}\eta_{1,j}\left\|\lambda_j\right\|_1 \tag{3.38}$$
$$\textit{subject to}\ \ \Delta'\Delta=I_r.$$

Here, $\hat{\Lambda}$ is the approximation to the factor loadings. This problem has two penalties; the first term, η is applied to all possible r components, and the second term, $\eta_{1,j}$ is applied to individual components to penalize their loadings. As in the estimation of a single component in (3.36), if we set $\Lambda=\Delta$, then we have $\sum_{t=1}^{T}\left\|X_t-\Delta\Lambda' X_t\right\|^2=\sum_{t=1}^{T}\left\|X_t-\Delta\Delta' X_t\right\|^2$, and so $\hat{\Lambda}(=\hat{\Delta})$ becomes the ordinary principal component's loading. Since (3.38) is not jointly convex for Δ and Λ, two steps to solve this problem are apparent. The first one involves fixing Δ then minimizing over Λ,which leads to a problem involving r elastic nets. In particular, since Δ is orthonormal, let $\Delta^{\dagger}$ be any orthonormal matrix such that $\left[\Delta;\Delta^{\dagger}\right]$ is $r\times r$ orthonormal matrix. Then we have

$$\begin{aligned}\sum_{t=1}^{T}\left\|X_t-\Delta\Lambda' X_t\right\|^2 &= \left\|X-X\Lambda\Delta'\right\|^2\\ &= \left\|X\Delta^{\dagger}\right\|^2+\left\|X\Delta-X\Lambda\right\|^2\\ &= \left\|X\Delta^{\dagger}\right\|^2+\sum_{j=1}^{r}\left\|X\delta_j-X\lambda_j\right\|^2\end{aligned}$$

That is, let Δ be given; and the optimal solution for Λ is based on minimizing

$$\arg\min_{\Lambda}\sum_{j=1}^{r}\left[\left\|X\delta_j-X\lambda_j\right\|^2+\eta\left\|\lambda_j\right\|^2+\eta_{1,j}\left\|\lambda_j\right\|_1\right]. \tag{3.39}$$

This is equivalent to r independent elastic net problems. If we rewrite (3.39) for a single loading, we have

$$\hat{\lambda}_j = \arg\min_{\lambda_j} \left\| F_j^* - X\lambda_j \right\|^2 + \eta \left\| \lambda_j \right\|^2 + \eta_{1,j} \left\| \lambda_j \right\|_1 , \tag{3.40}$$

where $F_j^* = X\delta_j$. And (3.40) is identical to

$$\left(\delta_j - \lambda_j \right)' X'X \left(\delta_j - \lambda_j \right) + \eta \left\| \lambda_j \right\|^2 + \eta_{1,j} \left\| \lambda_j \right\|_1 . \tag{3.41}$$

Here, we only need to calculate the correlation matrix, since we already standardized X. In the end, we solve these elastic nets efficiently via the LAR-EN algorithm discussed below.

The next step involves minimizing (3.38) over Δ, with fixed Λ. Then penalty term in this problem is now meaningless, and so the problem to be solved by minimizing

$$\sum_{t=1}^{T} \left\| X_t - \Delta\Lambda' X_t \right\|^2 \tag{3.42}$$
$$\textit{subject to} \;\; \Delta'\Delta = I_r .$$

This problem can be solved by the so called Procrustes transformation. (see Chapter 14.5 of Hastie et al. (2009) for details). Since $\sum_{t=1}^{T} \left\| X_t - \Delta\Lambda' X_t \right\|^2 = \left\| X - X\Lambda\Delta' \right\|^2$, using an appropriate transformation, we have the following singular value decomposition

$$X'X\Lambda = UDV'$$

where $\hat{\Delta} = UV'$. In practice, we let Δ be the factor loading of ordinary PCs, then we estimate Λ as a sparse factor loading. In this variant of the problem, the LAR-EN algorithm discussed below delivers a whole sequence of sparse approximations for each PC and the corresponding values of $\eta_{1,j}$.

3.2.3.2 SPCA algorithm

The numerical solution for the SPCA criterion to obtain sparse principal components is given as the following:

1. Let Δ be the loadings of the first r ordinary principal components.
2. Given Δ, solve the following problem for $j = 1, 2, ..., r$.

$$\lambda_j = \arg\min_{\lambda} \left(\delta_j - \lambda\right)' X'X \left(\delta_j - \lambda\right) + \eta \left\|\lambda\right\|^2 + \eta_{1,j} \left\|\lambda\right\|_1 .$$

3. For each fixed $\Lambda = [\lambda_1, ..., \lambda_r]$, do the singular vector decomposition on $X'X\Lambda = UDV'$, then update $\Delta^* = UV'$.
4. Repeat steps 2-3, until convergence.

In practice, the choice of η does not change the result much. Especially, in the case of full rank matrix of X, where zero is a reasonable value to use. Moreover, one may try to pick η_1 via cross-validation, or a related method. However, the LAR-EN algorithm efficiently solves this problem for all possible values of η_1. See Zou and Hastie (2005) or Kim and Swanson (2010) for computation details. Since the tuning parameter, η_1, affects the sparsity and variance of the components simultaneously, the algorithm is designed to give more weight to variance.

Note that if $\hat{F}$, are factors estimated by ordinary PCA, then they are uncorrelated so that we can compute the total variance explained by $\hat{F}$ as $tr\left(\hat{F}'\hat{F}\right)$. However, two conditions for principal components, uncorrelatedness and orthogonality, are not guaranteed in the case of sparse principal components. Still, it is necessary to derive the total variance in order to explain how much the components explain, even when the above two conditions are not satisfied. Zou et al. (2006) proposes a new way to compute the variance explained by the components, accounting for any correlation among the components. Since variance is given as $tr\left(\hat{F}'\hat{F}\right)$ for total variance if sparse principal components are already uncorrelated, this formula can be used more generally to compute the total variance of sparse principal components. Let $\tilde{F} = \left[\tilde{F}_1, ..., \tilde{F}_r\right]$ be the r components constructed via sparse principal component analysis. Denote $\hat{r}_j$ as the residual after regressing $\tilde{F}_j$ on $\tilde{F}_1, ..., \tilde{F}_{j-1}$, so that

$$\hat{r}_j = \tilde{F}_j - \mathbf{P}_{1,...,j-1}\tilde{F}_j,$$

where $\mathbf{P}_{1,...,j-1}$ is the projection matrix on $\tilde{F}_1, ..., \tilde{F}_{j-1}$. Then, the adjusted variance of a single component is $\|\hat{r}_j\|^2$ and the total variance is $\sum_{j=1}^{r} \|\hat{r}_j\|^2$. In practice, computation is easily done by QR factorization. If we let $\tilde{F} = QR$, then $\|\hat{r}_j\|^2 = R_{jj}^2$, so that total variance is $\sum_{j=1}^{r} R_{jj}^2$. Since the above computation is sequential, the order of components matters. However, in the current paper, we derive sparse PCs based on ordinary PCs, which are in turn already ordered by the size of the variance.

3.3 Robust Estimation Techniques

We consider a variety of "robust" estimation techniques in our forecasting experiments. The methods considered include bagging, boosting, ridge regression, least angle regression, the elastic net, the non-negative garotte and Bayesian model averaging. In Kim and Swanson (2010), we surveyed these methods, whereas in the current paper we implement them in the context of the factor model methodology discussed above. Here, we briefly summarize the shrinkage methods, and provide relevant citations to detailed discussions thereof.

Bagging, which was introduced by Breiman (1996), is a machine based learning algorithm whereby outputs from different predictors of bootstrap samples 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. (2009) 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 following methods are basically regression with regression coefficient penalization. First, consider 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. Conveniently, ridge regression uses a quadratic penalty term, and has a closed form solution. Second, 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. This method doesn't yield a closed form solution, and it needs to be estimated numerically. Third, "Least Angle Regression" (LAR), which is introduced in Efron et al. (2004), is a method for choosing a linear model using the same set of data as that used to evaluate and implement the model, and can be interpreted as the algorithm which finds a solution path for the lasso. LAR 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 LAR and lasso 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 LAR estimator. Fourth, is a related method called the "Elastic Net", which is proposed by Zou and Hastie (2005), which is also 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". LAR-Elastic Net (LAR-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. Finally, we consider 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. We follow Yuan and Lin (2007) in the sequel.

In addition to the above shrinkage methods, 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.

In the following subsections, we explain the intuition behind the above methods, and how they are used in our forecasting framework.

3.3.1 Bagging

Bagging, which is short for "bootstrap aggregation", was introduced by Breiman (1996) as a device for reducing prediction error in 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. Consider the regression problem with the training sample $\{Y, X\}$. Generate B bootstrap samples from the dataset and form predictions, $\hat{Y}_b^*(X_b^*)$, say, using each bootstrap sample, $b = 1, ..., B$. Bagging averages these predictions across bootstrap samples in order to reduce prediction variation. In particular, for each bootstrap sample, $\{Y_b^*, X_b^*\}$, regress Y_b^* on X_b^* and construct the fitted value $\hat{Y}_b^*(X_b^*)$. The bagging predictor is defined as follows:

$$\hat{Y}^{Bagging} = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_b^*(X_b^*) \tag{3.43}$$

Inoue and Kilian (2005) apply this bagging predictor in a time series context. 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:

$$\hat{Y}_{T+h}^{Bagging} = \sum_{j=1}^{N}\psi(\omega_j)\hat{\beta}_j X_{Tj} + o_p(1), \tag{3.44}$$

where $\hat{Y}_{T+h}^{Bagging}$ is the forecast of Y_{t+h} made using data through time T, $\hat{\beta}_j$ is the least squares estimator of β_j under $Y = X\beta$ and $\omega_j = \sqrt{T}\hat{\beta}_j/s_e$, with $s_e^2 = \Sigma_{t=1}^{T}(Y_{t+h} - X_t\hat{\beta}^{'})^2/(T-$

N), where $\hat{\beta} = \left(\hat{\beta}_1, ..., \hat{\beta}_N\right)'$. Also, ψ is

$$\psi(\omega) = 1 - \Phi(\omega + c) + \Phi(\omega - c) + \omega^{-1}[\phi(\omega - c) - \phi(\omega + c)], \tag{3.45}$$

where c is the pre-test critical value, ϕ is the standard normal density and Φ is the standard normal CDF.

Now, following Stock and Watson (2005a) define the forecasting model using bagging as follows:

$$\hat{Y}_{T+h}^{Bagging} = W_T\hat{\beta}_W + \sum_{j=1}^{r} \psi(\omega_j)\,\hat{\beta}_{Fj}\hat{F}_{Tj}, \tag{3.46}$$

where $\hat{\beta}_W$ is the LS estimator of β_W, W_t is a vector of observed variables (e.g. lags of Y) as in (3.16), and $\hat{\beta}_{Fj}$ is estimated using residuals, $Y_{T+h} - W_T\hat{\beta}_W$. The t-statistics used for shrinkage (i.e. the ω_j) are computed using least squares and Newey-West standard errors. Further, the pretest critical value for bagging in this paper is set at $c = 1.96$.

3.3.2 Boosting

Boosting (see Freund and Schapire (1997)) is a procedure that combines the outputs of many "weak learners" (models) to produce a "committee" (prediction). In this sense, boosting bears a resemblance to bagging and other "committee-based" shrinkage 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. Typically this output comes from minimizing a loss function averaged over training data. In this sense, boosting has something in common with forward stagewise regression. The final boosted procedure takes the form of linear combinations of weak learners. Freund and Schapire (1997), proposed the so-called "adaBoost" algorithm. AdaBoost and other boosting algorithms have attracted a lot of attention due to their success in data modeling.

Friedman et al. (2000) extend AdaBoost to "Real AdaBoost", which focuses on the construction of real-valued predictions. Suppose that we have a training sample of data, (Y, X), and let $\hat{\mu}(X)$ be a function (learner) defined on $\mathbb{R}^n$. Also, let $L(Y_t, \mu(X_t))$ be the loss function that penalizes deviations of $\hat{\mu}(X)$ from Y, at time t. The objective is to estimate $\mu(\cdot)$ that minimizes expected loss, $E[L(Y_t, \hat{\mu}(X_t))]$. Popular "learners" include smoothing splines, kernel regressions and least squares. Additionally, in AdaBoost, an exponential loss function is used.

Friedman (2001) introduces "L_2Boosting", which takes the simple approach of refitting base learners to residuals from previous iterations under quadratic loss. 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$_2$Boosting" algorithm in the sequel.

Boosting Algorithm Let $Z = Y - \hat{Y}^W$, which is obtained in a first step by fitting an autoregressive model to response variable using W_t as regressors. Then, using estimated factors:

1. Initialize : $\hat{\mu}^0(F_t) = \bar{Z}$, for each t.

2. 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 $\hat{\beta}_j$.

3. Compute $\hat{d}_j = u - \hat{F}_j \hat{\beta}_j$ for $j = 1, .., r$, and the sum of squared residuals, $SSR_j = \hat{d}_j' \hat{d}_j$. Let j_*^i denote the column selected at the i^{th} iteration, say, such that $SSR_{j_*^i} = \min_{j \in [1,...,r]} SSR_j$, and let $\hat{g}_*^i(F) = \hat{F}_{j_*^i} \hat{\beta}_{j_*^i}$.

4. 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.

Over-fitting may arise if this algorithm is iterated too many times. Therefore, selecting the number of iterations, M is crucial. Bai and Ng (2009) define the stopping parameter M using an information criterion of the form:

$$IC\left(i\right) = \log\left[\hat{\sigma}^{i^2}\right] + \frac{A_T \cdot df^i}{T} \tag{3.47}$$

where $\hat{\sigma}^{i^2} = \Sigma_{t=1}^{T}\left(Y_t - \hat{\mu}^i\left(\hat{F}_t\right)\right)^2$ and $A_T = \log(T)$. Evidently,

$$M = \arg\min_i IC\left(i\right). \tag{3.48}$$

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 a $T \times 1$ vector of 1's. Our boosting estimation uses this criterion. Finally, we have:

$$\hat{Y}_{t+h}^{Boosting} = W_t\hat{\beta}_W + \hat{\mu}^M\left(\hat{F}_t\right), \tag{3.49}$$

where $\hat{\beta}_W$ is defined above.

3.3.3 Ridge Regression

In the following three subsections, we discuss penalized regression approaches, including ridge regression, least angle regression, the elastic net and the nonnegative garotte. These methods shrink regression coefficients by retaining a only a subset of potential predictor variables. Ridge regression, as introduced by Hoerl and Kennard (1970), is the classical penalized regression method, and is introduced here in order to place the methods discussed thereafter in context. Consider explanatory variables that are stacked in an $T \times N$ matrix, and a univariate response or target variable, Y. Coefficients are estimated by minimizing a penalized residual sum of squares criterion. Namely, define:

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left[\sum_{t=1}^{T} \left(Y_t - \sum_{i=1}^{N} X_{ti}\beta_i \right)^2 + \eta \sum_{i=1}^{N} \beta_i^2 \right] \tag{3.50}$$

where η is a positive penalty parameter. The larger is η, the more we penalize coefficients, and the smaller the eventual subset of possible predictors that is used. The ridge regression estimator of (3.50) can be restated in the context of constrained regression, as follows:

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left[\sum_{t=1}^{T} \left(Y_t - \sum_{i=1}^{N} X_{ti}\beta_i \right)^2 \right], \tag{3.51}$$

$$\text{subject to } \sum_{i=1}^{N} \beta_i^2 \leq m,$$

where m is a positive number which corresponds to η. (Note that all observable predictors are standardized here, as elsewhere in this paper.). The ridge criterion (3.50) picks coefficients to minimize the residual sum of squares, and can conveniently be written in matrix form, as follows:

$$RSS\left(\eta\right) = \left(Y - X\beta\right)' \left(Y - X\beta\right) + \eta\beta'\beta, \tag{3.52}$$

where RSS denotes the residual sum of squares. Thus,

$$\hat{\beta}^{Ridge} = \left(X'X + \eta\mathbf{I}\right)^{-1} X'Y, \tag{3.53}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. In our experiments, we use the following model for forecasting:

$$\hat{Y}_{t+h}^{Ridge} = W_t\hat{\beta}_W + \hat{F}_t\hat{\beta}_F^{Ridge}. \tag{3.54}$$

Note that there is another penalized regression method that is similar to ridge regression, which is called the lasso (i.e. least absolute shrinkage selection operator). The key difference between two methods is the penalty function. The lasso estimator is defined as follows:

$$\hat{\beta}^{Lasso} = \arg\min_{\beta} \left[\sum_{t=1}^{T} \left(Y_t - \sum_{i=1}^{N} X_{ti}\beta_i \right)^2 \right], \tag{3.55}$$

$$\text{subject to } \sum_{i=1}^{N} |\beta_i| \leq m$$

That is, the L_2 ridge penalty is replaced by an L_1 lasso penalty. Accordingly, the lasso does not have a closed form solution like the ridge estimator. Although we report findings based upon ridge regression type models, we no not estimate the lasso, as it can be interpreted as a special case of least angle regression, which is discussed in the next sub section.

3.3.4 Least Angle Regression (LAR)

Least Angle Regression (LAR) is proposed in Efron et al. (2004), and can be viewed as an application of forward stagewise regression. In forward stagewise regression, predictor sets are constructed by adding one new predictor at a time, based upon the explanatory context of each new candidate predictor in the context of a continually updated least squares estimator. For details, see Efron et al. (2004).

Like many other stagewise regression approaches, start with $\hat{\mu}^0 = \bar{Y}$, the mean of the target variable, use the residuals after fitting W_t to the target variable, and construct a first estimate, $\hat{\mu} = X_t\hat{\beta}$, in stepwise fashion, using standardized data. Define $\hat{\mu}_{\mathcal{G}}$ to be the current LAR estimator, where $\mathcal{G}$ is a set of variables that is incrementally increased according to the relevance of each variable examined. 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 $\hat{C} = \max_j \{\hat{c}_j\}$ and $\mathcal{G} = \left\{ j : |\hat{c}_j| = \left|\hat{C}\right| \right\}$, by letting $s_j = sign\left(\hat{c}_j\right)$ (i.e. ± 1), for $j \in \mathcal{G}$, and defining the active matrix corresponding to $\mathcal{G}$ as $\mathcal{X}_{\mathcal{G}} = (...s_j X_j...)_{j \in \mathcal{G}}$. the objective is to find the predictor, X_j, that is most highly correlated with the residual. Let

$$\mathcal{D}_{\mathcal{G}} = \mathcal{X}_{\mathcal{G}}' \mathcal{X}_{\mathcal{G}} \text{ and } A_{\mathcal{G}} = \left(\mathbf{1}_{\mathcal{G}}' \mathcal{D}_{\mathcal{G}}^{-1} \mathbf{1}_{\mathcal{G}}\right)^{-\frac{1}{2}}, \tag{3.56}$$

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 $\mathcal{X}_{\mathcal{G}}' u_{\mathcal{G}} = A_{\mathcal{G}} \mathbf{1}_{\mathcal{G}}$. LAR then updates $\hat{\mu}$ as

$$\hat{\mu}_{\mathcal{G}+} = \hat{\mu}_{\mathcal{G}} + \hat{\gamma} u_{\mathcal{G}} \tag{3.57}$$

where

$$\hat{\gamma} = \min_{j \in \mathcal{G}^c}{}^{+} \left(\frac{\hat{C} - \hat{c}_j}{A_{\mathcal{G}} - a_j} \right) \left(\frac{\hat{C} + \hat{c}_j}{A_{\mathcal{G}} + a_j} \right), \tag{3.58}$$

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 LAR that imposes specific sign restrictions. In summary, LAR is a procedure that simply seeks new predictors that have the highest correlation with the current residual.

In order to apply LAR to time series data, Gelper and Croux (2008) revise the basic algorithm described here. They start by fitting an autoregressive model to the target variable, excluding predictor variables, using least squares. The corresponding residual series is retained and its standardized version is denoted by Z. The time-series LAR (henceforth, TS-LAR) procedure ranks the predictors according to how much they contribute to improving upon autoregressive fit. Using estimated factors as regressors, the following is the "LAR" algorithm of Gelper and Croux (2008):

LAR Algorithm

1. 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, and using standardized data.

2. 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 iteration.

   (b) Denote the 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, $H_{(i)} = \hat{F}_{(i)}\left(\hat{F}'_{(i)}\hat{F}_{(i)}^{-1}\right)\hat{F}'_{(i)}$.

   (c) Let $\tilde{F}_{(i)} = H_{(i)}\hat{\mu}^{i-1}$ be the $T \times 1$ standardized vector of values, $\hat{F}$, at the i^{th} iteration. Then, find the equiangular vector u^i, where $u^i = \left(\tilde{F}_{(1)}, \tilde{F}_{(2)}, ..., \tilde{F}_{(i)}\right) w^i$, $w^i = \frac{\mathcal{D}_{\mathcal{G}^i}^{-1}\mathbf{1}_i}{\sqrt{\mathbf{1}_i'\mathcal{D}_{\mathcal{G}^i}^{-1}\mathbf{1}_i}}$, $\mathcal{D}_{\mathcal{G}^i} = \mathcal{F}'_{\mathcal{G}^i}\mathcal{F}_{\mathcal{G}^i}$, $\mathcal{F}_{\mathcal{G}^i} = \left(...s_j\hat{F}^j...\right)_{j\in\mathcal{G}^i}$, $s_j = sign\left(\hat{c}_j\right)$, and $\hat{c} = \hat{F}'\left(\bar{Z} - \hat{\mu}^i\right)$.

3. (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 (3.58).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, $\hat{F}$, the highest ranked will be included in the final model. Now, the only choice remaining is how many predictors to include in the model. Finally, construct

$$\hat{Y}_{t+h}^{LAR^{+}} = W_t\hat{\beta}_W + \hat{\mu}^{LAR}(\hat{F}_t) \tag{3.59}$$

where $\hat{\mu}^{LAR}(\hat{F}_t)$ is the optimal value of the LAR estimator. The final predictor of Y is formed by adding back the mean to $\hat{Y}_{t+h}^{LAR^{+}}$.

3.3.5 Elastic Net (EN)

The elastic net (EN) is proposed by Zou and Hastie (2005), who point out various limitations of the lasso. Since it is a modification of the lasso, it can be viewed as a type of LAR, and indeed, their algorithm is sometimes called "LAR-EN". In order to motivate the LAR-EN algorithm, 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. For any fixed non-negative η_1 and η_2, the naive elastic net criterion is defined as:

$$L(\eta_1, \eta_2, \beta) = |Y - X\beta|^2 + \eta_2 |\beta|^2 + \eta_1 |\beta|_1 , \tag{3.60}$$

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

$$\hat{\beta}^{NEN} = \arg\min_{\beta} |Y - X\beta|^2 , \text{ subject to } (1-\alpha)|\beta|_1 + \alpha |\beta|^2 , \tag{3.61}$$

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 naïve elastic involves defining a new dataset (X^{+}, Y^{+}), where

$$X^+_{(T+N)\times N} = (1+\eta_2)^{-1/2}\left(\begin{array}{c} X \\ \\ \sqrt{\eta_2}\mathbf{I}_N \end{array}\right), \qquad Y^+_{(T+N)\times 1} = \left(\begin{array}{c} Y \\ \\ \mathbf{0}_N \end{array}\right). \tag{3.62}$$

Then, we can rewrite the naive elastic criterion as:

$$L\left(\frac{\eta_1}{\sqrt{1+\eta_2}}, \beta\right) = L\left(\frac{\eta_1}{\sqrt{1+\eta_2}}, \beta^+\right) = \left|Y^+ - D^+\beta^+\right|^2 + \frac{\eta_1}{\sqrt{1+\eta_2}}\left|\beta^+\right|_1. \tag{3.63}$$

If we let

$$\hat{\beta}^+ = \arg\min_{\beta^+} L\left(\frac{\eta_1}{\sqrt{1+\eta_2}}, \beta^+\right), \tag{3.64}$$

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

$$\hat{\beta}^{NEN} = \frac{1}{\sqrt{1+\eta_2}}\hat{\beta}^+. \tag{3.65}$$

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

$$\hat{\beta}^{NEN} = \frac{\left(\left|\hat{\beta}^{LS}\right| - \eta_1/2\right)_{pos}}{1+\eta_2} sign\left\{\hat{\beta}^{LS}\right\}, \tag{3.66}$$

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

$$\hat{\beta}^{Ridge} = \frac{\hat{\beta}^{LS}}{1+\eta_2} \tag{3.67}$$

and the lasso estimator is

$$\hat{\beta}^{Lasso} = \left(\left|\hat{\beta}^{LS}\right| - \eta_1/2\right)_{pos} sign\left\{\hat{\beta}^{LS}\right\}. \tag{3.68}$$

Zou and Hastie (2005), in the context of the 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, and they propose the elastic net, in which this double shrinkage is corrected. Given equation (3.62), the naive elastic net solves the regularization problem of the type:

$$\hat{\beta}^{+} = \arg\min_{\beta^{+}} \left|Y^{+} - X^{+}\beta^{+}\right|^{2} + \frac{\eta_1}{\sqrt{1+\eta_2}} \left|\beta^{+}\right|_{1}. \tag{3.69}$$

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

$$\hat{\beta}^{EN} = \sqrt{1+\eta_2}\hat{\beta}^{+}. \tag{3.70}$$

Thus ,

$$\hat{\beta}^{EN} = (1+\eta_2)\,\hat{\beta}^{NEN}. \tag{3.71}$$

Via this rescaling, the 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,

$$\hat{\beta}^{EN} = \arg\min_{\beta} \beta' \left(\frac{X'X + \eta_2 \mathbf{I}_N}{1+\eta_2}\right) \beta - 2Y'X\beta + \eta_1 \left|\beta\right|_{1}, \tag{3.72}$$

which is the estimator that we use in the forecasting model given as (3.16) when carrying out our prediction experiments.

Zou and Hastie (2005) propose an algorithm called the LAR-EN to estimate $\hat{\beta}^{EN}$ using LAR, as discussed above. With fixed η_2, the elastic net problem is equivalent to the lasso problem on the augmented dataset (X^{+}, Y^{+}), where $\mathcal{D}_{\mathcal{G}}$ in (3.56) is equal to $\frac{1}{1+\eta_2}\left(\mathcal{X}_{\mathcal{G}}'\mathcal{X}_{\mathcal{G}} + \eta_2 \mathbf{I}_{\mathcal{G}}\right)$ for any active set $\mathcal{G}$. Then the LAR-EN algorithm updates the elastic net estimator sequentially.

Choosing tuning parameters, η_1 and η_2, is a critical issue in the current context. Hastie et al. (2009) 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. We do this by picking a small grid of values for η_2 value, say $(0, 0.01, 0.1, 1, 10, 100)$. LAR-EN selects the η_2 value that yields the smallest CV error. We follow this approach when implementing LAR-EN.

3.3.6 Non-Negative Garotte (NNG)

The NNG estimator of Breiman (1995) 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 shrinkage factors in order to minimize:

$$\frac{1}{2}\|Y - Gq\|^2 + T\zeta\sum_{j=1}^{N} q_j, \qquad \text{subject to } q_j > 0,\ j = 1, .., N, \tag{3.73}$$

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 (3.73) 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:

Non-Negative Garotte Algorithm

1. 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 $T \times r$ matrix G; and the initial $\hat{\beta}$ is obtained by regressing $\hat{F}$ on Z, using least squares.

2. Compute the current direction γ, which is an r dimensional vector defined by $\gamma_{(\mathcal{G}^i)^c} = 0$ and

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

3. 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 arbitrarily chosen from $\mathcal{G}^i$. Now, 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 q_j becomes zero.

4. If $\alpha_j \leq 0, \forall j$ or $\min_{j, \alpha_j > 0} \{\alpha_j\} > 1$, set $\alpha = 1$. Otherwise, denote $\alpha = \min_{j, \alpha_j > 0} \{\alpha_j\} \equiv \alpha_{j^*}$ 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 $\mathcal{G}^i$.

5. 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

$$\hat{Y}_{t+h}^{NNG^+} = W_t\hat{\beta}_W + \hat{\mu}^{NNG}, \tag{3.74}$$

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

3.3.7 Bayesian Model Averaging (BMA)

Bayesian Model Averaging (BMA) has received considerable attention in recent years in forecasting literature (see e.g. Koop and Potter (2004), and Wright (2008, 2009)) For consise discussion of BMA implementation, see Hoeting et al. (1999) and Chipman et al. (2001). The basic idea of BMA starts with supposing interest focuses on Q possible models, denoted by $M_1, ..., M_Q$, say. In forecasting contexts, BMA involves averaging target predictions, Y_{t+h} from the candidate models, with weights appropriately chosen. In a very real sense, thus, it resembles bagging. One might also selecting one model by choosing M_{q^*} which maximizes $p\left(M_q|Data\right)$, but model averaging is generally preferred. If we denote ω as particular parameter vector, then BMA begins by noting that:

$$p\left(\omega|Data\right) = \sum_{q=1}^{Q} p\left(\omega|Data, M_q\right) p\left(M_q|Data\right). \tag{3.75}$$

If $g\left(\omega\right)$ is a function of ω, then without loss of generality, the conditional expectation is given as:

$$E\left[g\left(\omega\right)|Data\right] = \sum_{q=1}^{Q} E\left[g\left(\omega\right)|Data, M_q\right] p\left(M_q|Data\right). \tag{3.76}$$

This mean that we can compute the variance of the parameter for quadratic g. Accordingly, BMA involves obtaining results for all candidate models and averaging them with weights determined by the posterior model probabilities. That is, BMA, puts little weight on implausible models, as opposed to other varieties of shrinkage discussed above that operate directly on regressors. As we have 144 variables in our empirical work, we have 2^{144} possible models. This means that we must estimate OVER 10^{43} models at every forecasting horizon, and prior to the construction of each new prediction in this paper. Though there has been a quantum leap in computing technology in recent years, it would take several years to do this. Koop and Potter (2004) use Clyde (1999) approach to dealing with this problem, and take posterior draws of the parameters and associated variance using Gibbs sampling. This algorithm they use is somewhat different from the popular Markov Chain Monte Carlo algorithm in that draws are taken directly from the conditional probability of the parameters given the data and the variance. In this paper, we use the algorithm given in Koop and Potter (2004).

To implement Bayesian model averaging, we require a slightly different setup for that discussed above, in order to handle observable variables, W_t in (3.13). Chipman et al. (2001) suggest integrating them out using non-informative priors. Specifically, we transform our forecasting framework to be:

$$Y^*_{t+h} = \beta^* F^*_t + \varepsilon^*_t, \tag{3.77}$$

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 (3.16), and $\varepsilon_{t+h} \sim N\left(0, \sigma^2\right)$. This assumption leads 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 model 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. In this sense, U is similar to the current set in penalized regression. Moreover, U gives the prior model probability, $p\left(M_q\right)$, as the prior for U is equivalent to $p\left(M_q\right)$. According to Koop and Potter (2004), $p\left(U | Y^*\right)$ 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, $p\left(U\right)$ and $p\left(\sigma^2|Y^*,U\right)$ takes the inverted-Gamma form as usual. The next step involves specifying the prior model probability, $p\left(M_q\right)$ or equivalently, a prior for $p\left(U\right)$:

$$p\left(U\right) = \prod_{j=1}^{R} \upsilon_j^{U_j} \left(1-\upsilon_j\right)^{U_j}, \tag{3.78}$$

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, $P\left(M_q\right) = \frac{1}{Q}$ for $q = 1, ..., Q$. Other choices are also possible. For example, we could allow υ_j to depend on the j-th largest eigenvalue of $\hat{F}'\hat{F}$.

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 in all models; and in particular we follow Koop and Potter (2004), who suggest a noninformative prior for σ^{-2}. Namely, 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

$$\underline{V}_r = \left[g_r F_r^{*\prime} F_r^*\right]^{-1} \tag{3.79}$$

(see Fernandez et al. (2001a) and Zellner (1986) for more details on the use of g-priors). Finally, we are left with the issue of specification of g. Fernandez et al. (2001a) examine the properties of many possible choices for g and Koop and Potter (2004), in an objective Bayesian spirit, focus on values for g including $g = \frac{1}{T}$ and $g = \frac{1}{Q^2}$. We specifiy the same functions for g. Using the above approach, we form:

$$\hat{Y}_{t+h}^{*,BMA} = \hat{\beta}_F F_t^* \tag{3.80}$$

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

3.4 Data, Forecasting Methods, and Baseline Forecasting Models

3.4.1 Data

The data that we use are monthly observations on 144 U.S. macroeconomic time series for the period 1960:01 - 2009:5 ($N = 144, T = 593$)[13]. 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.[14] 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, logarithmic differences were taken for all nonnegative series that were not already in rates (see Stock and Watson (2002a, 2005a) for complete details). Note that a full list of the 144 predictor variables is provided in an appendix to an earlier version of this paper which is available upon request from the authors.

3.4.2 Forecasting Methods

Using the transformed dataset, denoted by X, factors are estimated using linear and nonlinear principal components methods, as discussed above. Thereafter,the robust estimation methods outlined in the previous sections are used to form forecasting models and predictions. In particular, we consider three specification types, as follows.

Specification Type 1: Linear and nonlinear principal components are first constructed using the large-scale dataset; and then prediction models are formed using the shrinkage

---

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

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

methods of Section 3 to select functions of and weights for the factors to be used in prediction models of the variety given in (3.16). This specification type is estimated with and without lags of factors.

Specification Type 2: Linear and nonlinear principal components are first constructed using subsets of variables from the large-scale dataset that are pre-selected via application of the robust shrinkage methods discussed in Section 3. Thereafter, prediction models of the variety given in (3.16) are estimated. This is different from the above approach of estimating factors using all of the variables. Note that forecasting models are estimated with and without lags of factors.

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

Specification Type 4: Prediction models are constructed using only shrinkage methods, and only with variables which have nonzero coefficients, as specified via pre-selection using SPCA.

In Specification Types 3 and 4, factor augmented autoregressions (FAAR) and pure factor based models (such as principal component regression - see next subsection for complete details) are not used as candidate forecasting models, since models with these specification types are not based on principal or independent components.

In our prediction experiments, pseudo out-of-sample forecasts are calculated for each variable, model variety, and specification type, for prediction horizons $h = 1, 3$, and 12. All estimation, including lag selection, shrinkage method application, and factor selection is done anew, at each point in time, prior to the construction of each new prediction, using both recursive and rolling data window 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 $P = T - R = 593 - 156 = 437$ observations, for $h = 1$). Moreover, the initial in-sample estimation period is adjusted so that the ex ante prediction sample length, P, remains fixed, regardless of the forecast horizon. For example, when forecasting the unemployment rate, when $h = 1$, the first forecast will be $\hat{Y}_{157}^{h=1} = \hat{\beta}_W W_{156} + \hat{\beta}_F \tilde{F}_{156}$, while in the case where $h = 12$, the first forecast will be $\hat{Y}_{157}^{h=12} = \hat{\beta}_W W_{145} + \hat{\beta}_F \tilde{F}_{145}$ In our rolling estimation scheme, the in-sample estimation period used to calibrate our prediction models is fixed at length 12 years. The recursive estimation scheme begins with the same in-sample period of 12 years (when $h = 12$), 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, thus, 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 to facillitate comparison across forecast horizons as well as models.

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

$$MSFE_{i,h} = \sum_{t=R-h+2}^{T-h+1} \left(Y_{t+h} - \hat{Y}_{i,t+h} \right)^2 \tag{3.81}$$

where $\widehat{Y}_{i,t+h}$ is the forecast for horizon h. Forecast accuracy is evaluated using the above point MSFE measure as well as the predictive accuracy test of Diebold and Mariano (1995), which is implemented using quadratic loss, and which has a null hypothesis that the two models being compared have equal predictive accuracy. See Kim and Swanson (2010) for details. 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 $H_0 : E\left[l\left(\varepsilon_{t+h|t}^1\right)\right] - E\left[l\left(\varepsilon_{t+h|t}^2\right)\right] = 0$,where $\varepsilon_{t+h|t}^i$ 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.

3.4.3 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 $\hat{Y}^{PCR}_{t+h} = \hat{\alpha} + \hat{\gamma}\hat{F}_t$, where $\hat{F}_t$ is estimated via principal components using X, as in equation (3.16).

Factor Augmented Autoregression: Based on equations (3.16), forecasts are computed as $Y^h_{t+h} = \hat{\alpha} + \hat{\beta}_F\hat{F}_t + \hat{\beta}_W(L)\,Y_t$. This model combines an AR(p) model, with lags selected using the SIC, with the above principal component regression (PCR) model. PCR and factor augmented autoregressive (FAAR) models are estimated using ordinary least square. Factors in the above models are constructed using PCA, ICA and SPCA.

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 $\hat{Y}_{t+h}^{ADL} = \hat{\alpha}+\hat{\beta}_i\left(L\right)X_{i,t}+\hat{\phi}_i\left(L\right)Y_t$. 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 $\left(w_i = 1/N\right)$, where $N = 144$. 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, and in addition to the Bayesian model averaging methods discussed in the previous section, we form forecasts as the simple average of the thirteen forecasting models summarized in Table 2.

3.5 Empirical Results

In this section, we summarize the results of our prediction experiments. Target variable mnemonics are given in Table 1, and forecasting models used are summarized in Panel A of Table 2. There are 6 different specifications types. Specification Types 1 and 2 (estimated with and without lags) are estimated via PCA, ICA and SPCA, so that there $4 \times 3 = 12$ permutations of these two types. Adding Specification Types 3 and 4, and multiplying by two (for recursive and rolling windowing strategies) yields a total of $(12+2)\times 2 = 28$ specification types for each target variable and each forecast horizon. Forecast modelling methods are summarized in Panel B of Table 2. For the sake of brevity, we eschew reporting the entirety of our experimental findings, instead focusing on key findings and results. Complete details are available upon request from the authors.

Table 3 summarizes point MSFEs for the “best” models, relative to the AR(SIC) model, where the AR(SIC) MSFE is normalized to unity. Results are reported in two panels, with the first panel summarizing findings across recursively estimated prediction models, and the second panel likewise reporting findings based on models estimated using rolling windows of data. Entries in bold denote point-MSFE "best" models among three principal component methods, for a given specifications, estimation window and forecast horizon. Entries in bold denote point-MSFE "best" models among three principal component methods, for a given specifications, estimation window and forecast horizon. Dot-circled entries denote cases for which each specification’s MSFE-best model using recursive window estimation yields a lower MSFE than that based on using rolling window estimation. Boxed entries denote cases where models are "winners" across all principal component methods, when only viewing models estimated using both recursive and rolling data windows, for a given forecasting horizon and specification type. Thus, there is only one “boxed” entry for each target variable / forecast horizon permutation, *across both panels of the table.* Since the benchmark models, including AR(SIC), ARX, etc. are included as candidate models in each specification type / principal component method permutation, there are some cases where the lowest relative MSFEs are same across principal component (PC) methods, for a given specification type. For example, in the case of Specification Type 1 and $h = 1$, GDP MSFEs are 0.916 for all three PC methods. This is because ARX, one of benchmark models, yields a lower MSFE than any other model used in conjunction with any of the PC methods. Moreover, since Specification Types 3 and 4 do not involve use of a principal component method, there are no bold entries in rows corresponding to these specification types.

Although there are a limited number of exceptions, most of the entries in Table 3 are less than unity, indicating that our factor based forecasting models dominate the autoregressive model. For example, note that the relative MSFE value for IPX when using SP1 and $h = 1$, is 0.268. Other bold entries can be seen to range from the low 0.80s to the mid 0.90s. Almost all of these entries are associated with models in which the DM null hypothesis of equal predictive accuracy is rejected.

Entries in the Table 4 show which forecast modelling method from Panel A of Table 2 has the lowest relative MSFE for each target variable, and for each specification type, principal component method, and forecast horizon, by estimation window (Panel A summarizes results for recursive window estimation, and Panel B does the same for rolling window estimation). These entries, thus, report the forecast modelling methods associated with each MSFE value given in Table 3. For example, in the leftmost three entries of Panel A of Table 3, we see that for unemployment, the FAAR, ARX, and FAAR methods resulted in the least MSFE predictions, under SP1 and for each of PCA, ICA, and SPCA, respectively, where these MSFEs, as reported in Table 3, are 0.780, 0.897 and 0.827, respectively. Bold entries in Panels A and B of the table denote forecasting method yielding the MSFE-best predictions, for a given specification type, forecast horizon, and target variable. Panel C of Table 4 summarizes the number of forecast method “wins” across 6 main specification types for the 11 target variables, by forecast horizon (i.e. reports the number of bold entries by forecast modelling method in Panels A and B). Note that FAAR and PCR are methods that are not used in Specification Types 3 and 4, since these specifications have no forecast modelling methods based on the principal component methods. Accordingly, mean forecasts in Specification Types 3 and 4 are constructed using the arithmetic mean of all forecast modelling methods except these two.

Notice also, in Table 4, that ARX appears in multiple entries. For example, for HS and $h = 1$, ARX appears as the “winner" in numerous cases. The reason for this is that each specification type has the same ARX model as one of the baseline models, and so correct interpretation of this finding is that the *same* ARX model dominates for a couple of variables (i.e. HS and GDP), when $h = 1$, regardless of PC method used for specification of factor models. However, note that for HS, the FAAR model that “wins” under SP1 and SPCA for $h = 1$, and has a relative MSFE (from Table 3) of 0.542, which is substantially lower than the value of 0.901 that applies to all of the cases where ARX “wins”. Thus, care must be taken when interpreting the results of Table 4; inasmuch as the ARX model is much less dominant than may appear to be the case upon cursory inspection of entries.

Interestingly, boosting and LAR perform well in several specifications and forecast horizons. This is particularly true for higher forecast horizons, where the only method to “win” more frequently involves simply using the arithmetic mean.

Entries in Panel A of Table 5 report which principal component method yields the lowest MSFE for each specification type, forecast horizon and target variable, when models are estimated using recursive data windows. (Since Specification Types 3 and 4 do not involve principal components, they are excluded in this table.) Panel B is the same as Panel A, except that results are for models estimated using rolling windows of data. Panel C of the table summarizes the result in Panel A across target variables, thus reporting counts of the number of times each principal component method “wins” by specification type, forecast horizon, and estimation window method. For example, upon inspection of Panel C, we see that for Specification Type 2 without lags, PCA, ICA and SPCA win 7, 2 and 1 times, respectively, for $h = 1$. Notice that SPCA performs very well under Specification 1, when $h = 1$, although PCA “wins” the most across all other specification types, regardless of forecast horizon. Moreover, ICA performs much worse than either other principal component estimation method. However, this result does not directly imply that PCA is a better method for factor analysis, since these results are based on complex hybrid forecasting modelling strategies coupling principal component methods with shrinkage and other regression modelling strategies.

Entries in Panel A of Table 6 report which estimation window method yields the lowest

MSFE for each specification type, principal component method, forecast horizon and target variable. Since Specification Types 3 and 4 do not involve principal components, they are excluded in this table. Panel B of Table 5 summarizes the result in Panel A across principal component methods and target variables. Evidently, the entries in this table correspond to the dot-circled entries in Panels A and B of Table 3. Here, 'Recur' stands for recursive window estimation and 'Roll' for rolling window estimation. Recursive window estimation "wins" in 93 out of 154 cases, when $h = 1$. On the other hand, it is interesting to note that rolling window estimation dominates at the $h = 12$ horizon, winning in 119 of 154 cases. Thus, the trade-off between using less data (and hence increasing parameter uncertainty while adjusting more quickly to structural breaks) and using more data (and hence failing to account for breaks), appears to depend on forecast horizon. For further discussion of data windowing, including a discussion of window combination, see Clark and McCracken (2009).

Panels A, B, and C of Table 7 summarize the results reported in Table 3 and 4. Entries in Panel A correspond to the dot-circled MSFEs in Table 3. and are the "best" MSFEs for each target variable, by specification type and forecast horizon. Bold entries in this panel are the "best" MSFEs for each target variable and forecast horizon, but across all specification types. Further, the window estimation scheme / principal component method / winning model combinations associated with each bold entry in Panel A are given in Panel B of the table. Finally, the specification type / window estimation scheme / principal component method / winning model combinations associated with each bold MSFE entry in Panel A are given in Panel C of the table.

In Panel A, note that SP1 and SP1L yields the MSFE-best prediction models in 15 of 33 possible cases, across forecast horizon (i.e. count up the bold entries in the table), with more than one half of these "wins" arising for the case where $h = 1$. Thus, just as estimation window selection seems to require differentiating across forecast horizon, so to does forecast horizon make a difference when ranking our specification types. However, recall from the results reported in Table 5 that although PCA “wins” quite frequently, the “wins” accorded to ICA and SPCA arise rather uniformly across forecast horizon.

Upon inspection of Panel B, the following conclusions emerge. First, of the window estimation scheme / principal component method / winning model combinations, recursive windowing “wins” 17 of 33 times. Thus, over all permutations and variables, the evidence suggests that there is little to choose between the two schemes. This points to even further evidence of the potential usefulness of the methods discussed in Clark and McCracken (2009). Second, the PCA principal components methods actually only “wins” in 14 of 33 possible cases, overall. This suggests that although PCA “wins” in many more cases when disaggreagting our findings, as reported earlier, when we actually summarize across the very best models, it wins less than one half of the time. Given the clear dominance of principal component methods in general, though (note that Specifications 3 and 4 “win” very infrequently), we have strong evidence that ICA and SPCA are very useful factor modelling tools; and in particular, we have seen from earlier discussion that SPCA is the clear winner from amongst non-PCA factor methods. Thus, as discussed in numerous papers, imposing parsimony on our factor modelling methods is quite useful. This in turn points to the fact that there is much remaining to be done in the area of parsimonious diffusion index modelling, given the novel nature and relative youth of these methods in the literature. Third, we see that the arithmetic mean forecasting model “wins” in only 9 of 33 cases. This is rather surprising new evidence that simple model averaging does not necessarily yield MSFE-best predictions. However, in order to “beat” model averaging methods, including arithmetic mean and Bayesian averaging approaches, we have needed to introduce into our horse-race numerous complex new models. Indeed, we see from further inspection of this table that most of the winning models involve combining complicated principal component methods with interesting new forms of shrinkage. It is really the combination of factor models and shrinkage that is delivering our results that model averaging does not always “win”.

Finally, turning to Panel C of Table 7, note that hybrid methods including factor methodology with shrinkage “win” in 9 of 33 cases, while simpler factor modelling approaches that do not additionally use shrinkage “win” in 10 of 33 cases. Pure shrinkage methods (i.e. SP3 and SP4 with shrinkage) “win” in 3 cases, while Bayesian model averaging and simpler arithmetic mean combination methods “win” the remaining 11 cases. Finally, simple linear autoregressive type models never win. We take these final results as further evidence of the usefulness of new methods in factor modelling and shrinkage, when the objective is prediction of macroeconomic time series variables.

3.6 Concluding Remarks

In this paper we outline and discuss a number of interesting new forecasting methods that have recently been developed in the statistics and econometrics literature. We focus in particular on the examination of a variety of factor modelling methods, including principal components as discussed by Stock and Watson (2002a,b) and others, independent component analysis (ICA) and sparse principal component analysis (SPCA). Further, we outline a number of approaches for creating hybrid forecasting models that use these factor modelling approaches in conjunction with various type of shrinkage, including boosting, bagging, and other methods. Finally, we carry out a series of real-time prediction experiments evaluating all of these methods against a number of benchmark linear models and forecast combination approaches. Our experiments are carried out in the context of predicting 11 key macroeconomic indicators at various forecast horizons.

We find that the simplest principal components type models “win” around 40% of the time. Interestingly, ICA and SPCA type models also “win” around 40% of the time. Thus, non factor modelling approaches only “win” around 20% of the time. Moreover, hybrid methods including factor approaches coupled with shrinkage “win” around 1/3 of the time, so that pure factor modelling approaches alone are not enough to lead to our overall finding that simple linear econometric models as well as models based on various forecast combination strategies are dominated by more complicated (factor/shrinkage) type models. Indeed, simple linear autoregressive type models never “win” in our experiments. We take these results as evidence of the usefulness of new methods in factor modelling and shrinkage, when the objective is prediction of macroeconomic time series variables.

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 prediction model specification and prediction 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 and Specification Type*

Panel A: Models and Methods Used in Real-Time Forecasting

Method	Description
AR(SIC)	Autoregressive model with lags selected by the SIC
ARX	Autoregressive model with exogenous regressors
CADL	Combined autoregressive distributed lag model
FAAR	Factor augmented autoregressive model
PCR	Principal components regression
Bagg	Bagging with shrinkage, $c = 1.96$
Boost	Component boosting, $M = 50$
BMA1	Bayesian model averaging with g-prior $= 1/T$
BMA2	Bayesian model averaging with g-prior $= 1/N^2$
Ridge	Ridge regression
LAR	Least angle regression
EN	Elastic net
NNG	Non-negative garotte
Mean	Arithmetic mean of the above forecasting method

Panel B: Specification Details

Estimation Window	Specification Type	Lags Included	Principal Component	Abbreviation
Recursive/ Rolling	1	No	PCA	SP1
			ICA	
			SPCA	
		Yes	PCA	SP1L
			ICA	
			SPCA	
	2	No	PCA	SP2
			ICA	
			SPCA	
		Yes	PCA	SP2L
			ICA	
			SPCA	
	3	No	Not Applicable	SP3
	4	No	Not Applicable	SP4

* 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, four different combined factor and shrinkage type prediction model estimation 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. Specification Type 4 - Prediction models are constructed using subsets of variables from the large-scale dataset that are first selected via application of the sparse princinpal component method. Then prediction models are estimated using 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: Point MSFEs Summarized By Principal Component Methods and Specification Type*

Panel A: Recursive Window Estimation

Forecast Horizon	Specification Method		UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
h = 1	SP1	PCA	**0.780**	0.870	0.940	0.875	0.943	0.811	0.900	0.800	**0.939**	0.976	**0.916**
		ICA	0.897	0.920	0.931	**0.840**	**0.843**	0.802	0.901	0.574	0.965	0.920	**0.916**
		SPCA	0.827	**0.789**	**0.409**	0.870	0.858	**0.706**	**0.542**	**0.268**	0.969	**0.897**	**0.916**
	SP1L	PCA	**0.850**	**0.889**	**0.955**	**0.865**	**0.945**	**0.879**	**0.901**	**0.804**	**0.930**	**0.976**	**0.916**
		ICA	0.897	0.966	0.978	0.939	0.960	0.918	**0.901**	0.861	0.991	1.002	**0.916**
		SPCA	0.897	0.954	0.987	0.939	0.972	0.881	**0.901**	0.826	0.954	0.998	**0.916**
	SP2	PCA	**0.861**	**0.950**	**0.965**	**0.933**	0.968	**0.854**	**0.901**	**0.833**	0.942	**0.985**	0.871
		ICA	0.897	0.959	0.971	0.939	**0.965**	0.861	**0.901**	0.874	0.959	0.991	**0.867**
		SPCA	0.897	0.959	0.976	0.939	0.966	0.860	**0.901**	0.873	**0.940**	0.986	0.873
	SP2L	PCA	**0.861**	**0.950**	**0.965**	0.933	0.968	**0.854**	**0.901**	**0.833**	**0.942**	**0.985**	0.871
		ICA	0.864	0.957	0.975	**0.923**	0.967	0.862	**0.901**	0.840	0.961	0.993	**0.871**
		SPCA	0.868	0.961	0.974	0.939	**0.963**	0.859	**0.901**	0.874	0.950	0.991	0.879
	SP3		0.897	0.944	0.987	0.933	0.956	0.826	0.901	0.874	0.977	0.989	0.873
	SP4		0.897	0.964	0.979	0.939	0.962	0.865	0.901	0.829	0.971	0.986	0.916
h = 3	SP1	PCA	**0.913**	**0.866**	0.998	0.929	**0.910**	0.819	**0.852**	0.850	0.977	**0.994**	0.956
		ICA	0.914	0.902	**0.975**	0.922	0.945	0.819	0.917	0.834	0.969	1.002	0.976
		SPCA	0.916	0.892	0.988	**0.895**	0.940	**0.775**	0.862	**0.816**	**0.942**	0.997	**0.944**
	SP1L	PCA	**0.925**	**0.892**	0.988	**0.901**	**0.929**	**0.818**	**0.852**	**0.838**	**0.978**	**0.993**	**0.963**
		ICA	0.963	0.902	0.998	0.967	0.945	0.927	0.948	0.895	0.997	1.007	0.979
		SPCA	0.951	0.902	**0.984**	0.968	0.945	0.924	0.912	0.887	0.990	0.997	0.988
	SP2	PCA	**0.916**	**0.895**	**0.992**	**0.888**	**0.945**	**0.827**	**0.783**	**0.809**	0.967	0.995	**0.954**
		ICA	0.941	0.902	0.995	0.959	**0.945**	0.859	0.824	0.821	0.980	0.997	0.963
		SPCA	0.943	0.902	0.998	0.975	**0.945**	0.894	0.793	0.873	**0.964**	**0.993**	0.963
	SP2L	PCA	0.916	**0.895**	**0.992**	**0.888**	**0.945**	**0.827**	**0.783**	0.809	**0.967**	0.995	**0.954**
		ICA	**0.916**	0.902	0.998	0.903	**0.945**	0.827	0.854	0.812	0.979	0.997	0.967
		SPCA	0.950	0.902	0.994	0.972	**0.945**	0.889	0.803	**0.812**	0.974	**0.993**	0.962
	SP3		0.943	0.902	0.998	0.926	0.945	0.860	0.723	0.881	0.939	1.001	0.975
	SP4		0.950	0.902	0.986	0.979	0.945	0.898	0.937	0.872	0.990	0.988	0.978
h = 12	SP1	PCA	0.939	0.956	0.997	**0.886**	**0.939**	0.874	**0.818**	**0.919**	0.958	1.002	0.999
		ICA	0.948	0.944	0.997	0.960	0.977	0.907	0.844	0.952	0.960	1.001	**0.986**
		SPCA	**0.933**	**0.940**	**0.992**	0.928	0.950	**0.845**	0.841	0.932	**0.950**	**0.996**	0.993
	SP1L	PCA	**0.903**	**0.956**	**0.988**	**0.888**	**0.927**	**0.860**	**0.829**	**0.926**	**0.942**	**0.995**	1.000
		ICA	0.943	0.969	0.997	0.961	0.981	0.912	0.912	0.939	0.964	1.002	**0.981**
		SPCA	0.912	0.977	0.997	0.945	0.970	0.879	0.832	0.937	0.981	1.001	0.997
	SP2	PCA	0.926	0.949	**0.992**	**0.891**	**0.950**	**0.816**	**0.749**	**0.916**	0.930	**0.995**	**0.982**
		ICA	0.941	0.949	0.997	0.909	0.960	0.843	0.901	0.942	0.933	0.999	0.991
		SPCA	**0.916**	**0.948**	0.997	0.935	0.957	0.843	0.910	0.919	**0.916**	0.997	0.992
	SP2L	PCA	0.926	**0.949**	**0.992**	**0.891**	**0.950**	**0.816**	**0.749**	**0.916**	**0.930**	0.995	**0.982**
		ICA	0.933	0.953	**0.992**	0.894	0.964	0.853	0.883	0.944	0.942	0.998	0.985
		SPCA	**0.914**	0.950	0.996	0.958	0.968	0.872	0.880	0.938	0.961	**0.994**	0.989
	SP3		0.926	0.961	0.997	0.899	0.953	0.862	0.804	0.890	0.910	1.002	0.982
	SP4		0.926	0.963	0.997	0.943	0.962	0.855	0.886	0.927	0.976	1.001	0.990

Panel B. Rolling Window Estimation

Forecast Horizon	Specification	Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
h = 1	SP1	PCA	**0.787**	**0.909**	**0.944**	**0.843**	**0.971**	**0.831**	**0.841**	**0.803**	**0.863**	**0.998**	**0.940**
		ICA	0.871	1.014	0.977	0.876	0.973	0.918	**0.841**	0.910	0.918	0.998	0.948
		SPCA	0.871	1.023	0.977	0.883	0.996	0.877	**0.841**	0.875	0.869	1.007	0.945
	SP1L	PCA	**0.852**	**0.989**	**0.954**	**0.850**	**0.973**	**0.871**	**0.841**	**0.845**	**0.845**	**1.002**	0.943
		ICA	0.871	1.004	0.982	0.883	0.985	0.924	**0.841**	0.877	0.908	1.008	**0.941**
		SPCA	0.871	1.081	0.992	0.883	1.003	0.911	**0.841**	0.851	0.880	1.008	0.989
	SP2	PCA	**0.871**	**1.085**	**0.963**	0.849	**0.936**	**0.869**	**0.841**	**0.858**	0.889	**0.998**	**0.915**
		ICA	**0.871**	1.114	0.977	0.849	0.941	0.884	**0.841**	0.858	0.908	1.006	0.915
		SPCA	**0.871**	1.087	0.979	**0.844**	0.949	0.877	**0.841**	0.892	**0.888**	1.007	0.927
	SP2L	PCA	**0.871**	**1.088**	**0.964**	0.850	**0.948**	**0.865**	**0.841**	**0.833**	**0.886**	**0.997**	**0.905**
		ICA	**0.871**	1.100	0.977	0.843	0.953	0.880	**0.841**	0.841	0.909	1.004	0.905
		SPCA	**0.871**	1.095	0.979	**0.840**	0.957	0.879	**0.841**	0.864	0.910	1.004	0.915
	SP3		0.871	1.114	0.992	0.858	1.000	0.924	0.841	0.841	0.916	1.008	0.930
	SP4		0.871	1.091	0.977	0.828	0.946	0.872	0.841	0.867	0.899	1.008	0.945
h = 3	SP1	PCA	**0.882**	**0.872**	1.002	**0.861**	**0.937**	**0.786**	**0.769**	**0.835**	**0.914**	**0.997**	**0.937**
		ICA	0.923	0.925	0.996	0.890	0.941	0.833	0.839	0.854	0.978	1.004	0.957
		SPCA	0.926	0.913	**0.993**	0.870	0.944	0.847	0.807	0.869	0.941	1.003	0.969
	SP1L	PCA	**0.904**	**0.889**	**0.981**	**0.848**	**0.920**	**0.807**	**0.744**	**0.820**	**0.908**	**0.988**	**0.953**
		ICA	0.936	0.925	1.001	0.900	0.951	0.876	0.854	0.877	0.976	1.008	0.957
		SPCA	0.957	0.903	1.002	0.905	0.945	0.905	0.840	0.884	0.981	1.001	0.972
	SP2	PCA	**0.895**	**0.883**	0.998	0.875	0.941	**0.814**	**0.740**	**0.833**	**0.912**	**0.989**	**0.929**
		ICA	0.912	0.899	**0.995**	0.875	**0.939**	0.838	0.743	0.850	0.915	0.989	0.950
		SPCA	0.919	0.914	0.997	**0.863**	0.941	0.846	0.785	0.857	0.927	0.989	0.947
	SP2L	PCA	0.889	**0.886**	**0.988**	0.864	0.942	**0.792**	**0.738**	**0.823**	**0.911**	**0.985**	**0.938**
		ICA	**0.888**	0.901	0.998	0.865	0.941	0.792	0.806	0.838	0.921	0.985	0.947
		SPCA	0.927	0.919	1.002	**0.861**	**0.936**	0.843	0.772	0.858	0.929	0.985	0.943
	SP3		0.911	0.903	1.002	0.906	0.960	0.839	0.683	0.844	0.950	1.002	0.970
	SP4		0.930	0.903	1.002	0.842	0.925	0.831	0.806	0.858	0.942	0.994	0.960
h = 12	SP1	PCA	0.897	**0.935**	**0.997**	0.812	**0.891**	0.729	0.723	**0.884**	**0.896**	1.007	1.010
		ICA	0.930	0.944	**0.997**	0.863	0.949	0.779	0.741	0.909	0.937	**0.996**	0.999
		SPCA	**0.879**	0.953	**0.997**	**0.781**	0.920	**0.720**	**0.715**	0.890	0.904	1.006	**0.997**
	SP1L	PCA	**0.864**	**0.946**	0.997	0.819	**0.902**	0.737	0.726	**0.898**	**0.899**	**1.000**	0.996
		ICA	0.908	0.951	0.997	0.872	0.962	**0.730**	0.773	0.902	0.942	1.003	0.987
		SPCA	0.869	0.983	**0.992**	**0.816**	0.938	0.759	**0.712**	0.943	0.960	1.002	**0.984**
	SP2	PCA	**0.893**	**0.929**	**0.997**	**0.818**	**0.912**	0.692	**0.637**	**0.880**	0.884	0.994	0.994
		ICA	0.911	0.932	0.997	0.833	0.915	**0.691**	0.726	0.902	0.888	0.994	0.993
		SPCA	0.901	0.935	0.997	0.819	0.921	0.692	0.693	0.896	**0.879**	**0.991**	**0.991**
	SP2L	PCA	**0.883**	**0.927**	**0.997**	**0.816**	**0.903**	**0.714**	**0.624**	**0.888**	**0.880**	**0.993**	0.996
		ICA	0.895	0.929	0.997	0.835	0.917	0.719	0.695	0.898	0.897	0.994	0.993
		SPCA	0.888	0.935	0.997	0.836	0.910	0.722	0.768	0.897	0.905	0.994	**0.991**
	SP3		0.903	0.971	0.997	0.799	0.947	0.690	0.551	0.940	0.891	1.001	0.998
	SP4		0.882	0.937	0.997	0.804	0.912	0.702	0.616	0.886	0.902	0.997	0.985

*Notes: See notes to Tables 1 and 2. Numerical entries in this table are the lowest mean square forecast errors (MSFEs) based on the use of various recursively estimated (Panel A) and rolling estimated (Panel B) prediction models using three different types of pricinpal component methods (PCA, ICA and SPCA) for six different specification types. Prediction models and target variables are described in Tables 1 and 2. See Section 3.2 for factor discussion. Forecasts are monthly, for the period 1974:3-2009:5. Forecast horizons reported on include h=1,3 and 12. 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 among three principal component methods, for a given specifications, estimation window and forecast horizon. Dot-circled entries denote cases for which each specification's MSFE-best model using recursive window estimation yields a lower MSFE than that based on using rolling window estimation. Boxed entries denote cases where models are "winners" across all principal component methods, when only viewing models estimated using both recursive and rolling data windows, for a given forecasting horizon and specification type.

Table 4: Summary of MSFE-"Best" Models*

Panel A: Recursive Window Estimation

Forecast Horizon	Specification Method		UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
h = 1	SP1	PCA	**FAAR**	PCR	Ridge	PCR	PCR	FAAR	ARX	PCR	**Mean**	Mean	**ARX**
		ICA	ARX	FAAR	FAAR	**FAAR**	**FAAR**	Ridge	ARX	FAAR	Mean	Boost	**ARX**
		SPCA	FAAR	**PCR**	**PCR**	BMA1	BMA2	**Mean**	**FAAR**	**FAAR**	Mean	**Boost**	**ARX**
	SP1L	PCA	**FAAR**	**PCR**	**Mean**	**PCR**	**Mean**	**Mean**	**ARX**	**BMA1**	**Mean**	**Boost**	**ARX**
		ICA	ARX	Mean	Mean	ARX	Mean	Mean	**ARX**	Mean	Mean	AR	**ARX**
		SPCA	ARX	Mean	CADL	ARX	Mean	Boost	**ARX**	Mean	Mean	Mean	**ARX**
	SP2	PCA	**Boost**	**Mean**	**Mean**	**Boost**	Mean	**Mean**	**ARX**	**BMA1**	BMA2	**Mean**	Boost
		ICA	ARX	Mean	Mean	ARX	**Mean**	Mean	**ARX**	ARX	EN	Mean	**Boost**
		SPCA	ARX	Mean	Mean	ARX	Mean	Mean	**ARX**	BMA1	**Boost**	Mean	Boost
	SP2L	PCA	**Boost**	**Mean**	**Mean**	Boost	Mean	**Mean**	**ARX**	**BMA1**	**BMA2**	**Mean**	Boost
		ICA	Boost	Mean	Mean	**Boost**	Mean	Mean	**ARX**	Boost	EN	Mean	**Boost**
		SPCA	Boost	Mean	Mean	ARX	**Mean**	Mean	**ARX**	ARX	Boost	Mean	Boost
	SP3		ARX	Mean	CADL	Mean	Mean	Mean	ARX	ARX	Mean	Boost	Mean
	SP4		ARX	Mean	Mean	ARX	Mean	Mean	ARX	BMA1	Mean	Mean	ARX
h = 3	SP1	PCA	**PCR**	**PCR**	CADL	FAAR	**PCR**	FAAR	**Boost**	Mean	Mean	**LAR**	Mean
		ICA	FAAR	ARX	**PCR**	FAAR	ARX	FAAR	LAR	Mean	Bagg	AR	Mean
		SPCA	Mean	PCR	Mean	**FAAR**	Mean	**Ridge**	Mean	**FAAR**	**Mean**	NNG	**Mean**
	SP1L	PCA	**Mean**	**Mean**	Mean	**Mean**	**Mean**	**BMA1**	**Mean**	**Mean**	**Mean**	**NNG**	**Mean**
		ICA	Mean	ARX	CADL	Mean	ARX	Mean	LAR	ARX	NNG	AR	Mean
		SPCA	Mean	ARX	**Mean**	Mean	ARX	BMA2	Mean	Mean	NNG	NNG	NNG
	SP2	PCA	**Boost**	**Mean**	**EN**	**Boost**	**ARX**	**Boost**	**Boost**	**Mean**	Mean	Mean	**Mean**
		ICA	Mean	ARX	LAR	Boost	**ARX**	Boost	Boost	Boost	Mean	Mean	Mean
		SPCA	Mean	ARX	CADL	Mean	**ARX**	Mean	Boost	Mean	**Boost**	**LAR**	Mean
	SP2L	PCA	Boost	**Mean**	**EN**	**Boost**	**ARX**	**Boost**	**Boost**	Mean	**Mean**	Mean	**Mean**
		ICA	**Boost**	ARX	CADL	Boost	**ARX**	Boost	Boost	LAR	Mean	Mean	Mean
		SPCA	Mean	ARX	BMA2	Mean	**ARX**	Mean	Boost	**LAR**	Boost	**Mean**	Mean
	SP3		Boost	ARX	CADL	Mean	ARX	Mean	Mean	BMA2	Mean	AR	Boost
	SP4		Mean	ARX	Mean	Mean	ARX	Mean	Mean	Mean	NNG	Mean	Mean
h = 12	SP1	PCA	Ridge	Mean	CADL	**FAAR**	**FAAR**	FAAR	**FAAR**	**Mean**	Mean	AR	Mean
		ICA	Mean	Mean	CADL	Mean	Mean	Mean	FAAR	CADL	Mean	AR	**Bagg**
		SPCA	**Mean**	**Mean**	**NNG**	Mean	Mean	**Mean**	Mean	Mean	**Mean**	**LAR**	Mean
	SP1L	PCA	**Mean**	**Mean**	**Boost**	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**	**Boost**	**LAR**	AR
		ICA	Mean	Bagg	CADL	Mean	Mean	Mean	FAAR	Bagg	Mean	AR	**Bagg**
		SPCA	Mean	Mean	CADL	Mean	BMA2	Mean	Mean	Mean	Mean	AR	Mean
	SP2	PCA	Mean	Mean	**Mean**	**BMA1**	**Mean**	**Boost**	**Boost**	**Mean**	Mean	**LAR**	**LAR**
		ICA	Mean	Mean	CADL	Boost	Mean	EN	Boost	Mean	Mean	LAR	Mean
		SPCA	**Boost**	**Mean**	CADL	Mean	Mean	EN	Boost	Mean	**Mean**	LAR	Mean
	SP2L	PCA	Mean	**Mean**	**Mean**	**BMA1**	**Mean**	**Boost**	**Boost**	**Mean**	**Mean**	LAR	**LAR**
		ICA	Mean	Mean	**BMA2**	Boost	Mean	Boost	Boost	Mean	Mean	Mean	LAR
		SPCA	**Boost**	Mean	Mean	Mean	Mean	Mean	Boost	Mean	Mean	**BMA2**	LAR
	SP3		Boost	Boost	CADL	Mean	Mean	Boost	EN	EN	Mean	AR	EN
	SP4		Mean	Mean	CADL	Mean	Mean	Mean	Boost	Mean	Mean	AR	Mean

Panel B. Rolling Window Estimation

Forecast Horizon	Specification Method		UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
h = 1	SP1	PCA	**FAAR**	**PCR**	**Mean**	**FAAR**	**Mean**	**FAAR**	**ARX**	**PCR**	**FAAR**	**LAR**	**Mean**
		ICA	ARX	AR	Mean	Mean	Mean	Mean	**ARX**	ARX	Mean	NNG	Mean
		SPCA	ARX	AR	Mean	ARX	LAR	Mean	**ARX**	Mean	Mean	AR	Mean
	SP1L	PCA	**Mean**	**PCR**	**Mean**	**Mean**	**Mean**	**Mean**	**ARX**	**Mean**	**Mean**	**AR**	Mean
		ICA	ARX	AR	Mean	ARX	Mean	Mean	**ARX**	Mean	Mean	AR	**Mean**
		SPCA	ARX	AR	CADL	ARX	AR	Mean	**ARX**	Mean	Mean	AR	LAR
	SP2	PCA	**ARX**	**AR**	**Mean**	Mean	**LAR**	**Mean**	**ARX**	**Boost**	Mean	**EN**	**EN**
		ICA	**ARX**	AR	Mean	Mean	LAR	Mean	**ARX**	Boost	Mean	AR	EN
		SPCA	**ARX**	AR	Mean	**Boost**	LAR	Mean	**ARX**	Mean	**Mean**	AR	LAR
	SP2L	PCA	**ARX**	**AR**	**Mean**	Mean	**EN**	**Mean**	**ARX**	**BMA2**	**Mean**	**LAR**	**LAR**
		ICA	**ARX**	AR	Mean	Mean	EN	Mean	**ARX**	Boost	Mean	AR	LAR
		SPCA	**ARX**	AR	Mean	**Mean**	Mean	Mean	**ARX**	Boost	Mean	AR	LAR
	SP3		ARX	AR	CADL	Boost	AR	Boost	ARX	Boost	LAR	AR	EN
	SP4		ARX	AR	Boost	BMA2	Mean	Mean	ARX	Mean	Boost	AR	Mean
h = 3	SP1	PCA	**Mean**	**PCR**	AR	**Mean**	**Mean**	**PCR**	**Boost**	**Mean**	**FAAR**	**LAR**	**Boost**
		ICA	Mean	Mean	PCR	Mean	Mean	Mean	Bagg	Mean	Bagg	AR	Mean
		SPCA	Mean	Mean	**BMA2**	BMA1	Mean	Mean	Mean	Mean	Mean	AR	Mean
	SP1L	PCA	**Mean**	**Mean**	**LAR**	**Mean**	**Mean**	**Mean**	**Boost**	**Mean**	**Mean**	**Mean**	**Mean**
		ICA	Mean	Mean	AR	BMA2	Boost	Mean	Boost	Mean	Mean	AR	Mean
		SPCA	Mean	Mean	AR	BMA2	NNG	Mean	Mean	Mean	Mean	AR	LAR
	SP2	PCA	**Mean**	**Mean**	NNG	Mean	Mean	**BMA2**	**Boost**	**Mean**	**EN**	**NNG**	**LAR**
		ICA	Mean	Mean	**BMA2**	Mean	**Mean**	Mean	Boost	Mean	EN	NNG	Mean
		SPCA	Boost	Mean	BMA1	**BMA2**	Mean	Mean	Boost	Mean	Mean	NNG	Mean
	SP2L	PCA	Boost	**Mean**	**BMA1**	Mean	Mean	**Boost**	**BMA2**	**Mean**	**Mean**	**NNG**	**Mean**
		ICA	**Boost**	Mean	BMA2	Mean	Mean	Boost	Boost	Boost	Boost	NNG	Mean
		SPCA	Mean	Mean	AR	**Mean**	**Mean**	Mean	Boost	Mean	Boost	NNG	Mean
	SP3		Boost	Boost	AR	Boost	NNG	Boost	Boost	Boost	Boost	AR	Boost
	SP4		Mean	Mean	AR	Mean	Mean	Boost	Mean	Boost	Boost	Mean	LAR
h = 12	SP1	PCA	Mean	**Mean**	**CADL**	Mean	**PCR**	FAAR	Boost	**Mean**	**Mean**	AR	AR
		ICA	Mean	Mean	**CADL**	Ridge	Mean	Mean	FAAR	Mean	Mean	**Bagg**	Mean
		SPCA	**Mean**	Mean	**CADL**	**BMA2**	Mean	**Mean**	**Mean**	Mean	Mean	AR	**Mean**
	SP1L	PCA	**Mean**	**Mean**	CADL	Mean	**Mean**	Mean	Mean	**Mean**	**Mean**	**AR**	NNG
		ICA	Mean	Mean	CADL	Mean	Mean	**Mean**	Mean	Mean	Mean	AR	Bagg
		SPCA	Mean	NNG	**NNG**	**BMA2**	Boost	Mean	**Mean**	LAR	LAR	AR	**LAR**
	SP2	PCA	**Mean**	**Mean**	**CADL**	**Mean**	**Mean**	EN	**Boost**	**Mean**	Boost	NNG	Mean
		ICA	Mean	Mean	CADL	Mean	Mean	**EN**	Boost	Mean	Boost	NNG	Mean
		SPCA	Mean	Mean	CADL	Mean	Mean	EN	Boost	Mean	**Boost**	**LAR**	**Mean**
	SP2L	PCA	**Mean**	**Mean**	**CADL**	**Mean**	**Mean**	**Boost**	**Boost**	**Mean**	**Mean**	**BMA2**	Mean
		ICA	Mean	Mean	CADL	Mean	Mean	Boost	Boost	Mean	Boost	NNG	Mean
		SPCA	Mean	Mean	CADL	Mean	LAR	Boost	Boost	Mean	Boost	NNG	**Mean**
	SP3		Boost	Boost	CADL	EN	EN	Boost	Boost	Boost	Boost	AR	NNG
	SP4		Mean	Mean	CADL	Boost	Mean	Mean	Boost	Mean	Mean	NNG	EN

Panel C: Summary of Panel A and B

h=1

	Recursive Window Estimation							Rolling Window Estimation						
	SP1	SP1L	SP2	SP2L	SP3	SP4	Total	SP1	SP1L	SP2	SP2L	SP3	SP4	Total
AR	0	1	0	0	0	0	1	3	6	5	5	3	2	24
ARX	6	10	8	5	3	4	36	7	7	6	6	2	2	30
CADL	0	1	0	0	1	0	2	0	1	0	0	1	0	2
FAAR	10	1	0	0	0	0	11	4	0	0	0	0	0	4
PCR	6	2	0	0	0	0	8	2	1	0	0	0	0	3
Bagg	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Boost	2	2	6	10	1	0	21	0	0	3	2	3	2	10
BMA1	1	1	2	1	0	1	6	0	0	0	0	0	0	0
BMA2	1	0	1	1	0	0	3	0	0	0	1	0	1	2
Ridge	2	0	0	0	0	0	2	0	0	0	0	0	0	0
LAR	0	0	0	0	0	0	0	2	1	4	4	1	0	12
EN	0	0	1	1	0	0	2	0	0	3	2	1	0	6
NNG	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Mean	5	15	15	15	6	6	62	14	17	12	13	0	4	60

h=3

	SP1	SP1L	SP2	SP2L	SP3	SP4	Total	SP1	SP1L	SP2	SP2L	SP3	SP4	Total
AR	1	1	0	0	1	0	3	3	4	0	1	2	1	11
ARX	2	5	5	5	2	2	21	0	0	0	0	0	0	0
CADL	1	1	1	1	1	0	5	0	0	0	0	0	0	0
FAAR	7	0	0	0	0	0	7	1	0	0	0	0	0	1
PCR	5	0	0	0	0	0	5	3	0	0	0	0	0	3
Bagg	1	0	0	0	0	0	1	2	0	0	0	0	0	2
Boost	1	0	10	10	2	0	23	2	3	4	9	8	3	29
BMA1	0	1	0	0	0	0	1	1	0	1	1	0	0	3
BMA2	0	1	0	1	1	0	3	1	2	3	2	0	0	8
Ridge	1	0	0	0	0	0	1	0	0	0	0	0	0	0
LAR	2	1	2	2	0	0	7	1	2	1	0	0	1	5
EN	0	0	1	1	0	0	2	0	0	2	0	0	0	2
NNG	1	5	0	0	0	1	7	0	1	4	3	1	0	9
Mean	11	18	14	13	4	8	68	19	21	18	17	0	6	81

h=12

	SP1	SP1L	SP2	SP2L	SP3	SP4	Total	SP1	SP1L	SP2	SP2L	SP3	SP4	Total
AR	2	3	0	0	1	1	7	3	3	0	0	1	0	7
ARX	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CADL	3	2	2	0	1	1	9	3	2	3	3	1	1	13
FAAR	5	1	0	0	0	0	6	2	0	0	0	0	0	2
PCR	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Bagg	1	3	0	0	0	0	4	1	1	0	0	0	0	2
Boost	0	2	6	7	3	1	19	1	1	6	8	6	2	24
BMA1	0	0	1	1	0	0	2	0	0	0	0	0	0	0
BMA2	0	1	0	2	0	0	3	1	1	0	1	0	0	3
Ridge	1	0	0	0	0	0	1	1	0	0	0	0	0	1
LAR	1	1	4	4	0	0	10	0	3	1	1	0	0	5
EN	0	0	2	0	3	0	5	0	0	3	0	2	1	6
NNG	1	0	0	0	0	0	1	0	3	2	2	1	1	9
Mean	19	20	18	19	3	8	87	20	19	18	18	0	6	81

*Notes: See notes to Tables 1, 2 and 3. Entries in Panels A and B are the forecasting method yielding the lowest MSFEs, corresponding to the entries in Table 3. Entries in Panel C denote the number of MSFE "wins" by forecasting method across all entries in Panel A and B. Forecasting methods are described in Table 2.

Table 5: Numerical Summary of MSFE-"Best" Principal Component Method*

Panel A: Recursive Window Estimation

Specification	Horizon	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
SP1	h=1	PCA	SPCA	SPCA	ICA	ICA	SPCA	SPCA	SPCA	PCA	SPCA	ALL
	h=3	PCA	PCA	ICA	SPCA	PCA	SPCA	PCA	SPCA	SPCA	PCA	SPCA
	h=12	SPCA	SPCA	SPCA	PCA	PCA	SPCA	PCA	PCA	SPCA	SPCA	ICA
SP1L	h=1	PCA	PCA	PCA	PCA	PCA	PCA	ALL	PCA	PCA	PCA	PCA
	h=3	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
	h=12	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	ICA
SP2	h=1	PCA	PCA	PCA	PCA	ICA	PCA	ALL	PCA	SPCA	PCA	ICA
	h=3	PCA	PCA	PCA	PCA	ALL	PCA	PCA	PCA	SPCA	SPCA	PCA
	h=12	SPCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	SPCA	PCA	PCA
SP2L	h=1	PCA	PCA	PCA	ICA	SPCA	PCA	ALL	PCA	PCA	PCA	ICA
	h=3	PCA	PCA	PCA	PCA	ALL	ICA	PCA	SPCA	PCA	SPCA	PCA
	h=12	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	SPCA	PCA

Panel B: Rolling Window Estimation

Specification	Horizon	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
SP1	h=1	PCA	PCA	PCA	PCA	PCA	PCA	ALL	PCA	PCA	PCA	PCA
	h=3	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
	h=12	SPCA	PCA	PCA	SPCA	PCA	SPCA	SPCA	PCA	PCA	ICA	SPCA
SP1L	h=1	PCA	PCA	PCA	PCA	PCA	PCA	ALL	PCA	PCA	PCA	ICA
	h=3	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
	h=12	PCA	PCA	SPCA	SPCA	PCA	ICA	SPCA	PCA	PCA	PCA	SPCA
SP2	h=1	PCA	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	SPCA	PCA	PCA
	h=3	PCA	PCA	ICA	SPCA	ICA	PCA	PCA	PCA	PCA	PCA	PCA
	h=12	PCA	PCA	PCA	PCA	PCA	ICA	PCA	PCA	SPCA	SPCA	SPCA
SP2L	h=1	PCA	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
	h=3	ICA	PCA	PCA	SPCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA
	h=12	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	SPCA

Panel C: Summary of MSFE-best by PC method

	Recursive Window Estimation									Rolling Window Estimation								
	h=1			h=3			h=12			h=1			h=3			h=12		
	PCA	ICA	SPCA	PCA	ICA	SPCA	PCA	ICA	SPCA	PCA	ICA	SPCA	PCA	ICA	SPCA	PCA	ICA	SPCA
SP1	2	2	6	5	1	5	4	1	6	10	0	0	10	0	1	5	1	5
SP1L	10	0	0	10	0	1	10	1	0	9	1	0	11	0	0	6	1	4
SP2	7	2	1	8	0	2	8	0	3	9	0	2	8	2	1	7	1	3
SP2L	7	2	1	7	1	2	9	0	2	10	0	1	8	1	2	10	0	1

* Notes: Entries in Panels A and B of this table show which principal component method yields the lowest MSFE predictions. If a benchmark model (AR,ARX and CADL) is MSFE-"better" than PCA, ICA and SPCA in Table 4, the entry is the "ALL", otherwise, entries correspond to MSFE-best principal component methods reported in Table 4. In Panel C, entries are counts of each pricipal component method "wins" from first two panels of the table. Since there is no column for ALL, sum of counts across one row of entries in a box doesn't need to be eleven, equallying the number of target variables.

Table 6: Numerical Summary of MSFE-"Best" Estimation Windows

Panel A. MSFE-'best' Models by Estimation Scheme Across Specification

Specification Type	Forecast Horizon	PC Method	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
SP1	h=1	PCA	Recur	Recur	Recur	Roll	Recur	Recur	Roll	Recur	Roll	Recur	Recur
		ICA	Roll	Recur	Recur	Recur	Recur	Recur	Roll	Recur	Roll	Recur	Recur
		SPCA	Recur	Recur	Recur	Recur	Recur	Recur	Recur	Recur	Roll	Recur	Recur
	h=3	PCA	Roll	Recur	Recur	Roll	Recur	Roll	Roll	Roll	Roll	Recur	Roll
		ICA	Recur	Recur	Recur	Roll	Roll	Recur	Roll	Recur	Recur	Recur	Roll
		SPCA	Recur	Recur	Recur	Roll	Recur	Recur	Roll	Recur	Roll	Recur	Recur
	h=12	PCA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Recur
		ICA	Roll	Recur	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur
		SPCA	Roll	Recur	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Recur
SP1L	h=1	PCA	Recur	Recur	Roll	Roll	Recur	Roll	Roll	Recur	Roll	Recur	Recur
		ICA	Roll	Recur	Recur	Roll	Recur	Recur	Roll	Recur	Roll	Recur	Recur
		SPCA	Roll	Recur	Recur	Roll	Recur	Recur	Roll	Recur	Roll	Recur	Recur
	h=3	PCA	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll
		ICA	Roll	Recur	Recur	Roll	Recur	Roll	Roll	Roll	Roll	Recur	Roll
		SPCA	Recur	Recur	Recur	Roll	Recur	Roll	Roll	Roll	Roll	Recur	Roll
	h=12	PCA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Roll
		ICA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Recur
		SPCA	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Recur	Roll	Recur	Roll
SP2	h=1	PCA	Recur	Recur	Roll	Roll	Roll	Recur	Roll	Recur	Roll	Recur	Recur
		ICA	Roll	Recur	Recur	Roll	Roll	Recur	Roll	Roll	Roll	Recur	Recur
		SPCA	Roll	Recur	Recur	Roll	Roll	Recur	Roll	Recur	Roll	Recur	Recur
	h=3	PCA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Recur	Roll	Roll	Roll
		ICA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Recur	Roll	Roll	Roll
		SPCA	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll
	h=12	PCA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur
		ICA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur
		SPCA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll
SP2L	h=1	PCA	Recur	Recur	Roll	Roll	Roll	Recur	Roll	Recur	Roll	Recur	Recur
		ICA	Recur	Recur	Recur	Roll	Roll	Recur	Roll	Recur	Roll	Recur	Recur
		SPCA	Recur	Recur	Recur	Roll	Roll	Recur	Roll	Roll	Roll	Recur	Recur
	h=3	PCA	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Roll	Roll	Roll
		ICA	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Roll	Roll	Roll
		SPCA	Roll	Recur	Recur	Roll	Roll	Roll	Roll	Recur	Roll	Roll	Roll
	h=12	PCA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur
		ICA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur
		SPCA	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Recur
SP3	h=1		Roll	Recur	Recur	Roll	Recur	Recur	Roll	Roll	Roll	Recur	Recur
	h=3		Roll	Recur	Recur	Roll	Recur	Roll	Roll	Roll	Recur	Recur	Roll
	h=12		Roll	Recur	Recur	Roll	Roll	Roll	Roll	Recur	Roll	Roll	Recur
SP4	h=1		Roll	Recur	Roll	Roll	Roll	Recur	Roll	Recur	Roll	Recur	Recur
	h=3		Roll	Recur	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Roll
	h=12		Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll

Panel B: MSFE-Best Models By Estimation Window and Specification Type

	h = 1		h=3		h=12	
	Recur	Roll	Recur	Roll	Recur	Roll
SP1	26	7	19	14	10	23
SP1L	20	13	9	24	8	25
SP2	17	16	5	28	5	28
SP2L	19	14	5	28	7	26
SP3	6	5	5	6	4	7
SP4	5	6	3	8	1	10
Total	93	61	46	108	35	119

* Note: See notes to Tables 1,2 and 3. Entries in Panel A show which estimation type yields the MSFE-best model. "Recur" refers recursive window estimation and "Roll" refers to rolling window esimation. Entries correspond to dot-circled entries in Panels A and B in Table 3. Panel B is a summary of "wins" by specification type and forecasting horizon.

Table 7: Best MSFEs By Specification Type and Forecasting Horizon

Panel A: Lowes MSFEs By Specification Type

Forecast Horizon	Specification Type	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
h = 1	SP1	**0.780**	**0.789**	**0.409**	0.840	**0.843**	**0.706**	**0.542**	**0.268**	0.863	**0.897**	0.916
	SP1L	0.850	0.889	0.954	0.850	0.945	0.871	0.841	0.804	**0.845**	0.976	0.916
	SP2	0.861	0.950	0.963	0.844	0.936	0.854	0.841	0.833	0.888	0.985	**0.867**
	SP2L	0.861	0.950	0.964	0.840	0.948	0.854	0.841	0.833	0.886	0.985	0.871
	SP3	0.871	0.944	0.987	0.858	0.956	0.826	0.841	0.841	0.916	0.989	0.873
	SP4	0.871	0.964	0.977	**0.828**	0.946	0.865	0.841	0.829	0.899	0.986	0.916
h = 3	SP1	**0.882**	**0.866**	**0.975**	0.861	**0.910**	**0.775**	0.769	0.816	0.914	0.994	0.937
	SP1L	0.904	0.889	0.981	0.848	0.920	0.807	0.744	0.820	**0.908**	0.988	0.953
	SP2	0.895	0.883	0.992	0.863	0.939	0.814	0.740	**0.809**	0.912	0.989	**0.929**
	SP2L	0.888	0.886	0.988	0.861	0.936	0.792	0.738	0.809	0.911	**0.985**	0.938
	SP3	0.911	0.902	0.998	0.906	0.945	0.839	**0.683**	0.844	0.939	1.001	0.970
	SP4	0.930	0.902	0.986	**0.842**	0.925	0.831	0.806	0.858	0.942	0.988	0.960
h = 12	SP1	0.879	0.935	0.992	**0.781**	**0.891**	0.720	0.715	0.884	0.896	0.996	0.986
	SP1L	**0.864**	0.946	**0.988**	0.816	0.902	0.730	0.712	0.898	0.899	0.995	**0.981**
	SP2	0.893	0.929	0.992	0.818	0.912	0.691	0.637	**0.880**	**0.879**	**0.991**	0.982
	SP2L	0.883	**0.927**	0.992	0.816	0.903	0.714	0.624	0.888	0.880	0.993	0.982
	SP3	0.903	0.961	0.997	0.799	0.947	**0.690**	**0.551**	0.890	0.891	1.001	0.982
	SP4	0.882	0.937	0.997	0.804	0.912	0.702	0.616	0.886	0.902	0.997	0.985

Panel B: MSFE-Best Window/PC/Model Combination By Forecasting Horizon and Sepcification

Forecast Horizon	Specification Type	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
h=1	SP1	**Recur**	**Recur**	**Recur**	Recur	**Recur**	**Recur**	**Recur**	**Recur**	Roll	**Recur**	Recur
		PCA	**SPCA**	**SPCA**	ICA	**ICA**	**SPCA**	**SPCA**	**SPCA**	PCA	**SPCA**	PCA
		FAAR	**PCR**	**PCR**	FAAR	**FAAR**	**Mean**	**FAAR**	**FAAR**	FAAR	**Boost**	ARX
	SP1L	Recur	Recur	Roll	Roll	Recur	Roll	Roll	Recur	**Roll**	Recur	Recur
		PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	**PCA**	PCA	PCA
		FAAR	PCR	Mean	Mean	Mean	Mean	ARX	BMA1	**Mean**	Boost	ARX
	SP2	Recur	Recur	Roll	Roll	Roll	Recur	Roll	Recur	Roll	Recur	**Recur**
		PCA	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	SPCA	PCA	**ICA**
		Boost	Mean	Mean	Boost	LAR	Mean	ARX	BMA1	Mean	Mean	**Boost**
	SP2L	Recur	Recur	Roll	Roll	Roll	Recur	Roll	Recur	Roll	Recur	Recur
		PCA	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
		Boost	Mean	Mean	Mean	EN	Mean	ARX	BMA1	Mean	Mean	Boost
	SP3	Roll	Recur	Recur	Roll	Recur	Recur	Roll	Roll	Roll	Recur	Recur
		ARX	Mean	CADL	Boost	Mean	Mean	ARX	Boost	LAR	Boost	Mean
	SP4	Roll	Recur	Roll	**Roll**	Roll	Recur	Roll	Recur	Roll	Recur	Recur
		ARX	Mean	Boost	**BMA2**	Mean	Mean	ARX	BMA1	Boost	Mean	ARX
h=3	SP1	**Roll**	**Recur**	**Recur**	Roll	**Recur**	**Recur**	Roll	Recur	Roll	Recur	Roll
		PCA	**PCA**	**ICA**	PCA	**PCA**	**SPCA**	PCA	SPCA	PCA	PCA	PCA
		Mean	**PCR**	**PCR**	Mean	**PCR**	**Ridge**	Boost	FAAR	FAAR	LAR	Boost
	SP1L	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll	**Roll**	Roll	Roll
		PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	**PCA**	PCA	PCA
		Mean	Mean	LAR	Mean	Mean	Mean	Boost	Mean	**Mean**	Mean	Mean
	SP2	Roll	Roll	Recur	Roll	Roll	Roll	Roll	**Recur**	Roll	Roll	**Roll**
		PCA	PCA	PCA	SPCA	ICA	PCA	PCA	**PCA**	PCA	PCA	**PCA**
		Mean	Mean	EN	BMA2	Mean	BMA2	Boost	**Mean**	EN	NNG	**LAR**
	SP2L	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur	Roll	**Roll**	Roll
		ICA	PCA	PCA	SPCA	SPCA	PCA	PCA	PCA	PCA	**PCA**	PCA
		Boost	Mean	BMA1	Mean	Mean	Boost	BMA2	Mean	Mean	**NNG**	Mean
	SP3	Roll	Recur	Recur	Roll	Recur	Roll	**Roll**	Roll	Recur	Recur	Roll
		Boost	ARX	CADL	Boost	ARX	Boost	**Boost**	Boost	Mean	AR	Boost
	SP4	Roll	Recur	Recur	**Roll**	Roll	Roll	Roll	Roll	Roll	Recur	Roll
		Mean	ARX	Mean	**Mean**	Mean	Boost	Mean	Boost	Boost	Mean	LAR
h=12	SP1	Roll	Roll	Recur	**Roll**	**Roll**	Roll	Roll	Roll	Roll	Recur	Recur
		SPCA	PCA	SPCA	**SPCA**	**PCA**	SPCA	SPCA	PCA	PCA	SPCA	ICA
		Mean	Mean	NNG	**BMA2**	**PCR**	Mean	Mean	Mean	Mean	LAR	Bagg
	SP1L	**Roll**	Roll	**Recur**	Roll	Roll	Roll	Roll	Roll	Roll	Recur	**Recur**
		PCA	PCA	**PCA**	SPCA	PCA	ICA	SPCA	PCA	PCA	PCA	**ICA**
		Mean	Mean	**Boost**	BMA2	Mean	Mean	Mean	Mean	Mean	LAR	**Bagg**
	SP2	Roll	Roll	Recur	Roll	Roll	Roll	Roll	**Roll**	**Roll**	**Roll**	Recur
		PCA	PCA	PCA	PCA	PCA	ICA	PCA	**PCA**	**SPCA**	**SPCA**	PCA
		Mean	Mean	Mean	Mean	Mean	EN	Boost	**Mean**	**Boost**	**LAR**	LAR
	SP2L	Roll	**Roll**	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur
		PCA	**PCA**	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
		Mean	**Mean**	Mean	Mean	Mean	Boost	Boost	Mean	Mean	BMA2	LAR
	SP3	Roll	Recur	Recur	Roll	Roll	**Roll**	**Roll**	Recur	Roll	Roll	Recur
		Boost	Boost	CADL	EN	EN	**Boost**	**Boost**	EN	Boost	AR	EN
	SP4	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Roll
		Mean	Mean	CADL	Boost	Mean	Mean	Boost	Mean	Mean	NNG	EN

* Notes: See notes to Table 1, 2, 3 and 4. Entries in Panel A of this table report lowest relative MSFE model types across principal component method and estimation type. Thus, entries resort and report on the dot-circled entries in Panels A and B of Table 3, sorted by forecasting horizon. Bold entries in Panel A are lowest relative MSFEs by forecasting horizon. These are the final "winners", by forecasting horizon. For Specification Type 1 without (SP1) and with lags (SP1L), and Specification Type 2 without (SP2) and with lags (SP2L), each three rows of entries for each forecasting horizon in Panel B are the corresonding estimation type, principal component method and forecasting methods corresponding to the entries in Panel A. For Specification Types 3 and 4, each two rows of entries for each forecasting horizon in Panel B are the corresonding estimation type and forecasting methods corresponding to the entries in Panel A.

Panel C: Summary of MSFE-"Best" Specification Type/Window/PC/Model Combination By Forecast Horizon*

Forecast Horizon	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
h=1	SP1	SP1	SP1	SP4	SP1	SP1	SP1	SP1	SP1L	SP1	SP2
	Recur	Recur	Recur	Roll	Recur	Recur	Recur	Recur	Roll	Recur	Recur
	PCA	SPCA	SPCA	N/A	ICA	SPCA	SPCA	SPCA	PCA	SPCA	ICA
	FAAR	PCR	PCR	BMA2	FAAR	Mean	FAAR	FAAR	Mean	Boost	Boost
h=3	SP1	SP1	SP1	SP4	SP1	SP1	SP3	SP2	SP1L	SP2L	SP2
	Roll	Recur	Recur	Roll	Recur	Recur	Roll	Recur	Roll	Roll	Roll
	PCA	PCA	ICA	N/A	PCA	SPCA	N/A	PCA	PCA	PCA	PCA
	Mean	PCR	PCR	Mean	PCR	Ridge	Boost	Mean	Mean	NNG	LAR
h=12	SP1L	SP2L	SP1L	SP1	SP1	SP3	SP3	SP2	SP2	SP2	SP1L
	Roll	Roll	Recur	Roll	Roll	Roll	Roll	Roll	Roll	Roll	Recur
	PCA	PCA	PCA	SPCA	PCA	N/A	N/A	PCA	SPCA	SPCA	ICA
	Mean	Mean	Boost	BMA2	PCR	Boost	Boost	Mean	Boost	LAR	Bagg

* Notes: See notes to Table 1,2,3,4, and 6. Entries in this table show the lowest relative MSFE model across all specification type, estimation type, principal method, and forecasting method, corresponding to the bold entries in Table 6. Each Four rows of entries for each forecasting horizon include specification type, estimation type, principal component method and forecasting method. Since Specification Type 3 and 4 are not carried out using factors, the third rows of entries are reported as "N/A" when either of these two specifcation types win, overall. Benchmark models such as AR and ARX models are never MSFE-best across all specification types, for a given forecasting horizon and variables.

References

Aiolfi, M. and Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, 135(1-2):31–53.

Armah, N. A. and Swanson, N. R. (2010a). Diffusion index models and index proxies: Recent results and new direction. *European Journal of Pure and Applied Mathematics*, 3:478–501.

Armah, N. A. and Swanson, N. R. (2010b). 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. (2011). Some variables are more worthy than others: New diffusion index evidence on the monitoring of key economic indicator. *Applied Financial Economics*, 21:43–60.

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

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

Bai, J. and Ng, S. (2006a). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Bai, J. and Ng, S. (2006b). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, 131(1-2):507–537.

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

Bai, J. and Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, 24(4):607–629.

Banerjee, A. and Marcellino, M. (2008). Factor-augmented error correction models. CEPR Discussion Papers 6707, C.E.P.R. Discussion Papers.

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

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

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):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. In *Institute of Mathematical Statistics*, pages 65–134.

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

Clark, T. E. and McCracken, M. W. (2009). Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, 50(2):363–395.

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

Comon, P. (1994). Independent component analysis - a new concept? *Signal Processing*, 36:287–314.

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

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

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

Croux, C., Filzmoser, P., and Fritz, H. (2011). Robust sparse principal component analysis, working paper no.1113. Technical report, Catholic University of Leuven Department of Decision Science.

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

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

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

Dufour, J.-M. and Stevanovic, D. (2010). Factor-augmented varma models: Identification, estimation, forecasting and impulse responses. Working paper, McGill University.

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

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

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

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82(4):540–554.

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

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

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

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

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

Guo, J., James, G., Levina, E., Michailidis, G., and Zhu, J. (2010). Principal component analysis with sparse fused loadings. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning*. Springer, 2nd edition.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

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

Hyvärinen, A. (1998). Independent component analysis in the presence of gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22:49–67.

Hyvärinen, A. (1999a). Gaussian moments for noisy independent component analysis. *IEEE Signal Processing Letters*, 6(6):145–147.

Hyvärinen, A. (1999b). Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430.

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

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12:531–547.

Jolliffe, I. T. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22:29–35.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kim, H. H. and Swanson, N. R. (2010). Forecasting macroeconomic variables using linear and nonlinear models. Working paper, Rutgers University.

Kim, H. H. and Swanson, N. R. (2011). Diffusion indices sing nonlinear factor methods. Working paper, Rutgers University.

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

Lee, T.-W. (1998). *Independent Component Analysis - Theory and Applications*. Springer, Boston, Massachusetts, 1 edition.

Leng, C. and Wang, H. (2009). On general adaptive sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 18(1):201–215.

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

Penny, W., Robert, S., and Everson, R. (2001). Ica: Model order selection and dynamic source models. In Roberts, S. and Everson, R., editors, *Independent Component Analysis: Principles and Practice*, pages 299–314. Cambridge University Press, Cambridge, UK.

Pesaran, M. H., Pick, A., and Timmermann, A. (2011). Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics*, 164:173–187.

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

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

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

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

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

Stock, J. H. and Watson, M. W. (2002a). 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. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–62.

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

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

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

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

Stone, J. V. (2004). *Independent Component Analysis*. MIT Press.

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

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

Tong, L., Liu, R.-w., Soon, V., and Huang, Y.-F. (1991). Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38:499–509.

Vines, S. (2000). Simple principal components. *Applied Statistics*, 49:441–451.

Wright, J. H. (2008). Bayesian model averaging and exchange rate forecasting. *Journal of Econometrics*, 146:329–341.

Wright, J. H. (2009). Forecasting u.s. inflation by bayesian model averaging. *Journal of Forecasting*, 28:131–144.

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

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

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

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

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

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):262–286.

Curriculum Vitae

Hyun Hak Kim

EDUCATION	**Ph.D. in Economics,** May 2012 Rutgers University, New Brunswick, New Jersey, United States **M.A. in Economics**, May 2008 Rutgers University, New Brunswick, New Jersey, United States **B.B.A. and B.A. in Statistics**, Feb. 2004 Yonsei University, Seoul, Korea
WORK EXPERIENCE	Daishin Securities Inc., Dec. 2003 – Sep. 2004 DfA Capital Management, Inc. Jun. – Aug. 2009
RESEARCH EXPERIENCE	**Research Assistant,** *Department of Economics, Rutgers University,* 2007–2008, **Research Assistant,** *Department of Economics, Rutgers University,* 2009–Present,
TEACHING EXPERIENCE	**Instructor***, Department of Economics, Rutgers University,* 2008–2011 **Teaching Assistant,** *Department of Economics, Rutgers University,* 2008–2010
LANGUAGES	Korean (native), English (fluent), French (basic)