



Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods



Hyun Hak Kim^a, Norman R. Swanson^{b,*}

^a The Bank of Korea, 55 Namdaemunno, Jung-Gu, Seoul 100-794, Republic of Korea

^b Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA

ARTICLE INFO

Keywords:

Prediction
Independent component analysis
Sparse principal component analysis
Bagging
Boosting
Bayesian model averaging
Ridge regression
Least angle regression
Elastic net and non-negative garotte

ABSTRACT

A number of recent studies in the economics literature have focused on the usefulness of factor models in the context of prediction using “big data” (see Bai and Ng, 2008; Dufour and Stevanovic, 2010; Forni, Hallin, Lippi, & Reichlin, 2000; Forni et al., 2005; Kim and Swanson, 2014a; Stock and Watson, 2002b, 2006, 2012, and the references cited therein). We add to this literature by analyzing whether “big data” are useful for modelling low frequency macroeconomic variables, such as unemployment, inflation and GDP. In particular, we analyze the predictive benefits associated with the use of principal component analysis (PCA), independent component analysis (ICA), and sparse principal component analysis (SPCA). We also evaluate machine learning, variable selection and shrinkage methods, including bagging, boosting, ridge regression, least angle regression, the elastic net, and the non-negative garotte. Our approach is to carry out a forecasting “horse-race” using prediction models that are constructed based on a variety of model specification approaches, factor estimation methods, and data windowing methods, in the context of predicting 11 macroeconomic variables that are relevant to monetary policy assessment. In many instances, we find that various of our benchmark models, including autoregressive (AR) models, AR models with exogenous variables, and (Bayesian) model averaging, do not dominate specifications based on factor-type dimension reduction combined with various machine learning, variable selection, and shrinkage methods (called “combination” models). We find that forecast combination methods are mean square forecast error (MSFE) “best” for only three variables out of 11 for a forecast horizon of $h = 1$, and for four variables when $h = 3$ or 12. In addition, non-PCA type factor estimation methods yield MSFE-best predictions for nine variables out of 11 for $h = 1$, although PCA dominates at longer horizons. Interestingly, we also find evidence of the usefulness of combination models for approximately half of our variables when $h > 1$. Most importantly, we present strong new evidence of the usefulness of factor-based dimension reduction when utilizing “big data” for macroeconomic forecasting.

© 2016 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

1. Introduction

In recent years, a considerable amount of research has focused on the analysis of “big data” in economics. This in turn has resulted in considerable attention being paid to the rich variety of methods that are available in the

* Corresponding author.

E-mail addresses: khdoube@bok.or.kr (H.H. Kim), nswanson@econ.rutgers.edu (N.R. Swanson).

<http://dx.doi.org/10.1016/j.ijforecast.2016.02.012>

0169-2070/© 2016 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

areas of machine learning, data mining, variable selection, dimension reduction, and shrinkage. In this paper, we utilize various of these methods in order to add to the discussion of the usefulness of “big data” for forecasting macroeconomic variables such as unemployment, inflation and GDP. From the perspective of dimension reduction, we construct diffusion indices, and add to the discussion of the usefulness of such indices for macroeconomic forecasting.¹ In particular, when constructing diffusion indices, we implement principal component analysis (PCA), independent component analysis (ICA) and sparse principal component analysis (SPCA).² We also evaluate machine learning, variable selection and shrinkage methods, including bagging, boosting, ridge regression, least angle regression, the elastic net, and the non-negative garotte. Finally, we combine various dimension reduction techniques with these machine learning and shrinkage methods, and evaluate the usefulness of these approaches for forecasting.

In order to assess all of the above techniques, we carry out a large number of real-time out-of-sample forecasting experiments. Our venue for this “horse-race” is the prediction of 11 key macroeconomic variables that are relevant to monetary policy assessment. These variables include 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 was noted by Kim and Swanson (2014a), they 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, as well as the anecdotal reports compiled in the Beige Book”.

The idea of a diffusion index involves the use of appropriately “distilled” latent common factors that have been extracted from a large number of variables as inputs in the specification of subsequent parsimonious (yet “information rich”) 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 . Specifically, let

$$X = F\Lambda' + e, \quad (1)$$

¹ A small sample of recent forecasting studies using large-scale datasets and pseudo out-of-sample forecasting includes those by Armah and Swanson (2010a,b), Artis, Banerjee, and Marcellino (2005), Boivin and Ng (2005, 2006), Forni, Hallin, Lippi, and Reichlin (2005), and Stock and Watson (1999, 2002a, 2005, 2006, 2012). In addition, Stock and Watson (2006) discuss the literature on the use of diffusion indices for forecasting in some detail.

² There is a vast (and growing) body of literature in this area. A few of the relevant papers, addressing both empirical and theoretical issues, include those by Armah and Swanson (2010a,b), Artis et al. (2005), Bai and Ng (2002, 2006b, 2008), Banerjee and Marcellino (2008), Boivin and Ng (2005, 2006), Ding and Hwang (1999); Dufour and Stevanovic (2010), and Stock and Watson (2002a, 2005, 2006, 2012).

The above papers consider PCA. However, there is also a small and growing body of literature that examines ICA in the context of macroeconomic forecasting (see e.g. Moneta, Entner, Hoyer, & Coad, 2013; Tan & Zhang, 2012; Yau, 2004). We were unable to find any papers to date that have examined the use of SPCA in our context. However, the method has been applied empirically in various other fields. For example, see Carvalho et al. (2008) and Mayrink and Lucas (2013) in the context of gene expression genomics.

where e is a disturbance matrix and Λ is an $N \times r$ coefficient matrix. Once F has been extracted using one of the estimation methods examined in this paper, we construct the following forecasting model based on the work of Bai and Ng (2006a), Kim and Swanson (2014a) and Stock and Watson (2002a,b):

$$Y_{t+h} = W_t\beta_W + F_t\beta_F + \varepsilon_{t+h}, \quad (2)$$

where Y_t is the target variable to be predicted, h is the prediction horizon, W_t is a $1 \times s$ vector of “additional” explanatory variables, 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 here, we begin by estimating r unobserved (latent) factors, say \hat{F} , from the N observable predictors, X . In order to achieve useful dimension reduction, r is assumed to be much less than N (i.e., $r \ll N$). Then, parameter estimates, $\hat{\beta}_W$ and $\hat{\beta}_F$, are constructed using an 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.

Kim and Swanson (2014a) use principal component analysis (PCA) to obtain estimates of the latent factors, called principal components. PCA yields “uncorrelated” latent principal components via the use of data projection in the direction of the maximum variance, and principal components (PCs) are ordered naturally in terms of their variance contributions. The first PC defines the direction that captures the maximum variance possible, the second PC defines the direction of the maximum variance in the remaining orthogonal subspace, and so forth. Perhaps because PCs are easy to derive through the use of singular value decompositions, this is the method that is used most frequently in factor analysis (for details, see e.g. Bai & Ng, 2002, 2006b; Stock & Watson, 2002a). As was discussed above, this paper also implements ICA and SPCA for the estimation of latent factors. These methods are used in a variety of contexts in the statistics discipline. However, economists are yet to explore the usefulness of SPCA in forecasting contexts, and few empirical investigations of the usefulness of ICA have been reported in economics (see above for examples from this small body of literature). Notably, ICA (see e.g. Comon, 1994; Lee, 1998) uses so-called “negentropy”, which is a measure of the entropy, to construct independent factors. SPCA is designed to uncover *uncorrelated* components and ultimately factors, just like PCA. However, the method also searches for components with factor loading coefficient matrices that are “sparse” (i.e., the matrices can contain zeros). Since PCA yields nonzero loadings for the entire set of variables, their practical interpretation is more difficult than in contexts where the factors are characterized by sparsity. Note that the importance of sparsity has been discussed not only in the context of forecasting (see e.g. Bai & Ng, 2008), but also recently in a number of papers in the financial econometrics literature (see e.g. Fan, Rigollet, & Wang, 2015). For further discussions of this and related issues, see Jolliffe, Trendafilov, and Uddin (2003), Vines (2000), and Zou, Hastie, and Tibshirani (2006).

In order to add functional flexibility to our forecasting models, we also implement versions of Eq. (2) where

the numbers and functions of factors used are specified via the implementation of a variety of machine learning, variable selection and shrinkage methods, as discussed above. One key feature of many of these methods is that they are used for targeted regressor and factor selection. Related research that focuses on forecast combination methods is discussed by [Aiolfi and Timmermann \(2006\)](#), [Bai and Ng \(2008\)](#), and [Stock and Watson \(2012\)](#); and our discussion is meant to add to the recent work reported by [Kim and Swanson \(2014a\)](#) and [Stock and Watson \(2012\)](#), who survey and analyze several machine learning, variable selection and shrinkage methods that are based on factor augmented autoregression models of the variety given in Eq. (2). Finally, our experiments also consider various linear benchmark forecasting models, including autoregressive (AR) models, AR models with exogenous variables, and combined autoregressive distributed lag models.

Our main findings can be summarized as follows. First, models that are specified using factors almost always dominate all other models, in terms of mean square forecast errors (MSFE). In addition, ICA and SPCA are preferred to PCA for estimating the factors in our MSFE-“best” models when the forecasting horizon is $h = 1$, although the more standard approach of using PCA “wins” at all other forecasting horizons. One reason for this switch between PCA and the other methods may be that the PCA is “more” robust to structural breaks at the one-step-ahead horizon. [Stock and Watson \(2008\)](#) note that, in some cases, factors may play the same “averaging” role as “pooling” forecasts does, particularly in the face of intercept breaks in forecasting models. This argument derives in part from the fact that all of the factor loadings in principal components are nonzero. SPCA induces sparseness in the factor loadings, and thus may not offer this beneficial feature. Moreover, given the increasing inability of forecast model regression coefficients to adapt swiftly to structural change as the forecast horizon increases, this feature may account in part for our finding that PCA dominates at longer horizons, but not at the one-step-ahead horizon. However, any further empirical and theoretical analysis of this finding is left for future research.

Second, our benchmark AR-type models are never MSFE-best, and both model averaging techniques that include the use of arithmetic mean forecasts and Bayesian model averaging yield MSFE-best models for only about one third of the 11 variables, regardless of the forecast horizon. This is because pure factor-type models, machine learning, variable selection and shrinkage models, along with “combination models” that combine dimension reduction via the use of factors with machine learning and shrinkage, are the MSFE-best models for most of our variables, across all forecast horizons. In many cases, though, the key to “beating” model averaging methods involves the use of our combination models.

Third, even though combination models are important, pure machine learning, variable selection, and shrinkage methods almost never deliver MSFE-best models. Rather, they are most useful when combined with factor analysis methods, as was discussed above.

Fourth, recursive estimation strategies clearly dominate rolling strategies when constructing one-step-ahead

forecasts. However, rolling estimation methods are preferred to recursive methods at longer forecast horizons. This may be due in part to the presence of structural breaks, although further empirical and theoretical investigations are left to future research.

Overall, our findings suggest that, when analyzing “big data”, the dimension reduction associated with the specification and estimation of factors, along with machine learning and shrinkage methods, are very useful for forecasting macroeconomic variables.

The rest of the paper is organized as follows. In the next section, we provide a survey of dynamic factor models, independent component analysis, and sparse principal component analysis. In Section 3, we survey the machine learning, variable selection and shrinkage methods used in our prediction experiments. Data, forecasting methods, and baseline forecasting models are discussed in Section 4, and empirical results are presented in Section 5. Concluding remarks are given in Section 6.

2. Diffusion index models

2.1. Principal component analysis

In this section, we outline the factor and forecasting models that we use, and also provide a brief overview of PCA. For a detailed discussion of principal component analysis, see [Bai and Ng \(2002, 2008, 2009\)](#), [Kim and Swanson \(2014a\)](#), [Stock and Watson \(1999, 2002a, 2005, 2012\)](#), and the references cited therein.

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

$$X_{ij} = \Lambda_j' F_t + e_{ij}, \quad (3)$$

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_{ij} is the idiosyncratic component of X_{ij} .³ The product $\Lambda_j' F_t$ is called the common component of X_{ij} . This is the dimension reducing factor representation of the data. More specifically, for $r < N$, a factor analysis model has the form:

$$X_1 = \lambda_{11}F_1 + \dots + \lambda_{1r}F_r + e_1 \quad (4)$$

$$X_2 = \lambda_{21}F_1 + \dots + \lambda_{2r}F_r + e_2$$

$$\vdots$$

$$X_N = \lambda_{N1}F_1 + \dots + \lambda_{Nr}F_r + e_N.$$

Here, F is a vector of $r < N$ underlying latent variables or factors, λ_{ij} is an element of an $N \times r$ matrix, Λ , of factor loadings, and the e are uncorrelated zero-mean disturbances. Many economic analyses fit naturally into this 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:

$$Y_{t+h} = W_t \beta_W + F_t \beta_F + \varepsilon_{t+h}, \quad (5)$$

³ In what follows, we assume that all variables are standardized, as is customary in this literature.

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 $X = (X_1, \dots, X_N)$ can be represented as in Eq. (3). We then estimate the factors F_t using principal components analysis, independent component analysis, or sparse principal component analysis. In particular, forecasts of Y_{t+h} based on Eq. (5) 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 . Once we have the estimated factors, we can then obtain the estimators $\hat{\beta}_F$ and $\hat{\beta}_W$ by regressing Y_{t+h} on \hat{F}_t and W_t . Notably, if $\sqrt{T}/N \rightarrow 0$, 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 & Ng, 2008](#)). In this paper, we try different methods for estimating $\hat{\beta}_F$, then compare the predictive accuracies of the resulting forecasting models.⁴

In the following sections, we provide brief overviews of ICA and SPCA, underscoring the differences between these methods and PCA.

2.2. Independent component analysis

Independent component analysis (ICA) is predicated on the idea of “opening” the black box in which principal components often reside. A few uses of ICA include mobile phone signal processing, brain imaging, voice signal extraction and stock price modeling. In each case, there is a large set of individual signals observed, and it is assumed that each signal depends on several factors, which are unobserved.

The starting point for ICA is the very simple assumption that the components, F , are statistically independent in Eq. (3). The key is the measurement of this independence between components. The method can be depicted graphically as is shown in [Fig. 1](#).

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

$$X = S\Omega. \quad (6)$$

We can rewrite Eq. (6) as

$$\begin{aligned} X_1 &= \omega_{11}S_1 + \dots + \omega_{1N}S_N \\ X_2 &= \omega_{21}S_1 + \dots + \omega_{2N}S_N \\ &\vdots \\ X_N &= \omega_{N1}S_1 + \dots + \omega_{NN}S_N, \end{aligned} \quad (7)$$

⁴ We refer the reader to the studies by [Bai and Ng \(2002, 2008, 2009\)](#) and [Stock and Watson \(1999, 2002a, 2005, 2012\)](#) for a detailed explanation of this procedure, and to [Armah and Swanson \(2010b\)](#), [Connor and Korajczyk \(1986, 1988, 1993\)](#) and [Forni et al. \(2005\)](#) for further detailed discussions of generic diffusion index models.

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

$$F = X\Psi$$

or

$$F = S\Omega\Psi.$$

As we assume that the mixing matrix Ω is square, Ψ is also square, and $\Psi = \Omega^{-1}$, so that F is exactly the 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. The independent components F are latent variables, just the same as principal components, meaning that they cannot be observed directly. Also, the mixing matrix Ω is assumed to be unknown. All that we observe is data, X , and both Ω and S must be estimated using X . Only then can we estimate the demixing matrix, Ψ , and the independent components, F . However, Eq. (7) is not identified unless several assumptions are made. The first assumption is that the sources, S , are statistically independent. Since various sources of information (for example, consumers' behaviors, political decisions, etc.) may impact the values of macroeconomic variables, this assumption is not strong. The second assumption is that the signals are stationary. For further details, see [Tong, Liu, Soon, and Huang \(1991\)](#).

ICA under Eq. (7) assumes that N components of F exist. However, we can simply construct factors using up to r ($<N$) components, without loss of generality. In practice, we can construct r independent components by preprocessing with r principal components. See chapters 6 and 10 of [Stone \(2004\)](#) for further details. In general, the above model would be more realistic if noise terms were added. However, for the sake of simplicity, noise terms are omitted; and indeed, the estimation of the noise-free model is already computationally difficult (see [Hyvärinen & Oja, 2000](#), for a discussion of the noise-free model, and [Hyvärinen, 1998, 1999](#), for a discussion of the model with noise added). For a detailed explanation of the estimation algorithm that we use, we refer to the working paper version of this paper (i.e., see [Kim & Swanson, 2014b](#)).

2.2.1. Comparison with principal component analysis

As is evident from [Fig. 1](#), ICA is exactly the same as PCA if we let the demixing matrix be the factor loading coefficient matrix that is associated with the principal components analysis. The key difference between ICA and PCA is in the properties of the factors obtained. Principal components are uncorrelated and have descending variances, so that they are easy to order in terms of their variances. Moreover, the components that explain the largest shares of the variance are often assumed to be the ones that are “relevant” for subsequent use in diffusion index forecasting. In particular, the first principal component captures the maximum variance possible, the second component also captures the maximum variance but in an orthogonal subspace, and is thus uncorrelated with the first component, and so on.

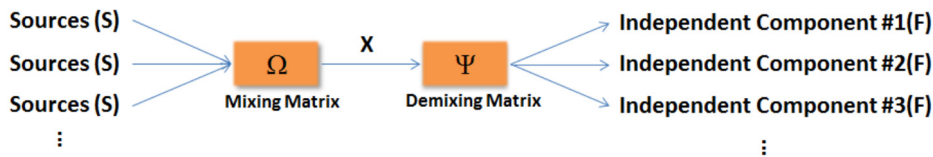


Fig. 1. Schematic representation of ICA.

For the sake of 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

$$E(F_1 F_2) = E(F_1) E(F_2). \quad (8)$$

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

$$E[F_1^{*p} F_2^{*q}] = E[F_1^{*p}] E[F_2^{*q}] \quad (9)$$

for every positive integer value of p and q . That is, the condition holds for all moments.

Evidently, PCA estimation is much simpler than ICA, since it only involves finding a linear transformation of components which are uncorrelated. Moreover, PCA ranks components using their variances or correlations, so that the components associated with higher variances or correlations are assumed to have more explanatory power than those with lower variances or correlations. On the other hand, ICA is unable to determine the variance associated with each independent component, since both S and Ω in Eq. (6) are unknown, so that any scalar multiplier in one of the sources, S_j , could be cancelled by dividing the corresponding mixing vector ω_j by the same scalar. Therefore, we can change the order of X in Eq. (6) randomly, which means that we cannot determine the order of the independent components. From a forecasting perspective, this is probably a good thing, since there is no *a priori* reason to believe that “largest variance” PCA components will be the most relevant for predicting any particular target variable. Moreover, this feature of ICA is the reason for using PCA for pre-processing in ICA algorithms. For further details about preprocessing, see Appendix F of Stone (2004).

2.3. Sparse principal component analysis

In the paper in which they develop SPCA, Zou et al. (2006) note that, under PCA, the factor loading coefficients are all typically nonzero, making the interpretation of the estimated components difficult. They address this issue by proposing a modified PCA method (i.e., SPCA) in which the lasso (elastic net) is used to construct principal components with sparse loadings. This is done by first reformulating PCA as a regression-type optimization problem, then using a lasso (elastic net) on the coefficients in a suitably constrained regression model.

Since the seminal paper by Zou et al. (2006), many authors have proposed variants of SPCA. For example, Jolliffe (1995) modified the loadings to be values such as

1, -1 and 0. Another approach is to set thresholds for the absolute values of the loadings, below which the loadings are set to zero. Jolliffe et al. (2003) suggested using the so-called “SCoTLASS” (simplified component technique lasso) to construct modified principal components with possible zero loadings, λ , by solving

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

for some tuning parameter φ . The absolute value threshold results in (various) zero loadings, which induces sparseness. However, the SCoTLASS constraint does not ensure convexity, and therefore the approach may be expensive computationally. As an alternative, Zou et al. (2006) develop a regression optimization framework; namely, they assume that X are dependent variables, F are explanatory variables, and the loadings are coefficients. They then use the lasso (and elastic net) to derive a sparse loading matrix. Other recent approaches include those discussed by Guo, James, Levina, Michailidis, and Zhu (2010) and Leng and Wang (2009), both of which are based on the work of Zou et al. (2006). We follow the approach of Zou et al. (2006), and readers are referred to Sections 3.3–3.5 of their paper for full details. As in the case of ICA, we again refer the reader to Kim and Swanson (2014b) for a detailed discussion of the estimation procedures that we implemented in order to use SPCA in our forecasting experiments.

2.4. Selecting the number of factors

The selection of the number of factors to use when applying PCA in our experiments is an important issue, since the number of factors used in our forecasting models may have an impact on the predictive performance of the models. In some contexts, such as in macroeconomics, the factors and numbers of factors could conceivably be chosen based on theoretical arguments. Of course, empirical analyses are often used for selecting the number of factors as well. Indeed, there are several empirical approaches for determining the appropriate number of factors for PCA. Well-known methods include the scree plot and the evaluation of percentages of cumulative variances. These methods are straightforward, and Neto, Jackson, and Somers (2005) provide a nice survey of them. Cross-validation is also feasible, but can be computationally expensive in big data environments. In light of this, Josse and Husson (2012) suggest using general cross-validation (GCV) to approximate leave-one-out cross-validation-based estimation of the number of factors. In the econometrics literature, Bai and Ng (2002) suggest that the number of factors be chosen using a selection criterion of the form $PC(r) = V(r, \hat{F}) + rh(N, T)$, where

$h(\cdot)$ is a penalty function, $V(\cdot)$ minimizes the Euclidian distance between the variables in the dataset and their factor projection, and r is the number of factors (see Kim & Swanson, 2014a, for further details).⁵ With regard to ICA, note that Li, Adali, and Calhoun (2007) propose the use of information criteria for selecting the number of factors in ICA. There is no specific research that we are aware of on the selection of the number of factors in SPCA. Our approach is to simply use the number of factors chosen based on Bai and Ng (2002) in all cases. We leave the analysis of the trade-offs associated with using alternative estimates of r for future research.

3. Machine learning, variable selection, and shrinkage methods

Our forecasting experiments consider a variety of machine learning, variable selection and shrinkage methods. The methods considered include bagging, boosting, ridge regression, least angle regression, the elastic net, the non-negative garotte and Bayesian model averaging. Here, we provide brief summaries of a number of these methods, and relevant citations to detailed discussions of them.

Bagging, which was introduced by Breiman (1996), is a machine-based learning algorithm whereby the outputs of different predictors of bootstrap samples are combined in order to improve the overall forecasting accuracy. Bühlmann and Yu (2002) use bagging to improve the forecast accuracy when the data are *i.i.d.* Inoue and Kilian (2008) and Stock and Watson (2012) extend bagging to time series models. Stock and Watson (2012) consider “bagging” as a form of shrinkage, when constructing prediction models. In this paper, we use the same algorithm that they use when constructing bagging estimators. This allows us to avoid the time-intensive bootstrap computation done elsewhere in the bagging literature. Boosting, a close relative of bagging, is another statistical learning algorithm, which was originally designed for classification problems in the context of probability approximate correct (PAC) learning (see Schapire, 1990), and is implemented by Freund and Schapire (1997) using the AdaBoost.M1 algorithm. Hastie, Tibshirani, and Friedman (2009) apply it to classification, and argue that “boosting” is one of the most powerful learning algorithms available at present. The method was also extended to regression problems by Ridgeway, Madigan, and Richardson (1999) and Shrestha and Solomatine (2006). In the economics literature, Bai and Ng (2009) use boosting to select the predictors in factor augmented autoregressions. We implement a boosting algorithm which mirrors that used by these authors.

The other shrinkage methods implemented here are penalized regression methods. One such method that we consider is ridge regression, which is a well-known linear method where the minimization of the sum of square residuals is modified to include a penalty that is a function of the number of parameters. Conveniently,

ridge regression uses a quadratic penalty term, and thus has a closed-form solution. We also implement the “least absolute shrinkage and selection operator” (lasso) that was introduced by Tibshirani (1996), which is another attractive technique for variable selection using high-dimensional datasets, especially when N is greater than T . This method is similar to the ridge regression, but uses an L_1 penalty function instead of ridge’s L_2 penalty, thus allowing for sparsity. Third, we examine “least angle regression” (LARS), which was introduced by Efron, Hastie, Johnstone, and Tibshirani (2004), and can be interpreted as an algorithm which finds a solution path for the lasso. Moreover, LARS is based on a well-known model-selection approach known as “forward-selection”, which has been used extensively for examining cross-sectional data (for further details, see Efron et al., 2004). Bai and Ng (2008) show how to apply LARS and the lasso in the context of time series data, and Gelper and Croux (2008) extend Bai and Ng’s (2008) work to time series forecasting with many predictors. We implement Gelper and Croux’s (2008) algorithm when constructing the LARS estimator. A related method that we consider is called the “elastic net”, which was proposed by Zou and Hastie (2005) and is also similar to the lasso, as it carries out automatic variable selection and continuous shrinkage simultaneously, via the use of penalized regression with both L_1 and L_2 penalty functions. Its name comes from the notion that it is similar in structure to a stretchable fishing net that retains “all of the big fish”. 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 “non-negative garotte”, which was 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 garotte algorithm and prove consistency in variable selection. We follow Yuan and Lin (2007) in what follows.

In addition to the above machine learning and shrinkage methods, we also consider Bayesian model averaging (henceforth BMA), as it is one of the most attractive methods for model selection that is available at present (see Fernandez, Ley, & Steel, 2001; Koop & Potter, 2004; Ravazzolo, Paap, van Dijk, & Franses, 2008). The concept of Bayesian model averaging can be described using simple probability rules.⁶ 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, we can use Bayesian inference to 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 the ways in which they are used in our forecasting framework.

For a comprehensive discussion of the above methods and estimation algorithms, the reader is referred to the work of Kim and Swanson (2014b).

⁵ Other recent approaches for selecting the number of factors include those of Chen, Huang, and Tu (2010) and Onatski (2009), and the references cited therein.

⁶ We also consider simple arithmetic model averaging.

4. Data, forecasting methods, and baseline forecasting models

4.1. Data

The data that we use are monthly observations on 144 US. macroeconomic time series for the period 1960:01–2009:5 ($N = 144$, $T = 593$).⁷ Forecasts are constructed for eleven variables, namely 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.⁸ Table 1 lists the eleven variables. The third column of the table gives the transformations of the variables that are used in order to induce stationarity. In general, logarithmic differences were taken for all non-negative series that were not already in rates (see Stock & Watson, 2002a, 2012, for complete details). Note that a full list of all 144 predictor variables is provided in an online appendix (see Appendix A).

4.2. Forecasting methods

Using the transformed dataset, denoted by X , factors are estimated using the techniques discussed above. Thereafter, the estimation methods outlined in the previous section are used to form forecasting models and predictions. In our experiments, we consider three specification types, as follows.

Specification type 1 (SP1): Factors are first constructed using the large-scale dataset and each of PCA, ICA, and SPCA. Prediction models are then constructed by using the machine learning and shrinkage methods in Section 3 to select functions of and weights for the factors to be used in prediction models of the variety given in Eq. (5). This specification type is estimated both with and without lags of factors.

Specification type 2 (SP2): Factors are first constructed using subsets of variables from the large-scale dataset and each of PCA, ICA, and SPCA. The variables used in the factor calculations are pre-selected via the application of the machine learning and shrinkage methods discussed in Section 3. Thereafter, prediction models of the variety given in Eq. (5) are estimated. This differs from the approach above of estimating factors using all of the variables. Note that the forecasting models are estimated both with and without lags of factors.

Using the notation in Eq. (5), SP1 and SP2 each include F_t , while the “lags of factors” versions, called SP1L and SP2L, include F_t and F_{t-1} . The reason why only one lag was

utilized is that no additional forecast improvement was found when more than one lag was included.

Specification type 3 (SP3): Prediction models are constructed using only the machine and shrinkage methods discussed in Section 3, without the use of factor analysis at any stage.

Specification type 4 (SP4): Prediction models are constructed using only machine learning and shrinkage methods, and with only variables which have nonzero coefficients, as is specified via pre-selection using SPCA.

Note that specification types 3 and 4 do not use factor augmented autoregressions (FAAR) and pure factor-based models (such as principal component regression; see the next subsection for complete details) as candidate forecasting models, since models with these specification types do not include factors of any type.

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, machine learning and shrinkage method implementation, 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 the number of factors included will be different in each estimation period, following the re-estimation of r . Note also that lags of the target predictor variables are included in the set of explanatory variables as well in all cases. The numbers of lags (of both variables and factors) to include are selected using the SIC. Out-of-sample forecasts begin after 13 years (e.g., for $h = 1$, 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). 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 for $h = 1$, the first forecast will be $\hat{Y}_{157}^{h=1} = \hat{\beta}_W W_{156} + \hat{\beta}_F \tilde{F}_{156}$, while for $h = 12$, 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 a length of 12 years. The recursive estimation scheme begins with the same in-sample period of 12 years (when $h = 12$), but in this case a new observation is added to the sample prior to the re-estimation and construction of each new forecast, as we iterate through the ex-ante prediction period. Thus, note that both the actual observations being predicted and the number of predictions in our ex-ante prediction period remain the same, regardless of the forecast horizon, in order to facilitate comparisons across forecast horizons as well as models.

Forecast performances are evaluated using mean square forecast errors (MSFE), defined as:

$$MSFE_{i,h} = \sum_{t=R-h+2}^{T-h+1} \left(Y_{t+h} - \hat{Y}_{i,t+h} \right)^2, \quad (10)$$

where $\hat{Y}_{i,t+h}$ is the forecast for horizon h . The forecast accuracy is evaluated using both the above point MSFE measure and the predictive accuracy test statistic (called

⁷ This is an updated and expanded version of Stock and Watson's (2002b; 2012) dataset, although data definitional changes and series discontinuities prevent us from updating the database to a more current date. We leave a discussion of the use of alternative data samples to future research.

⁸ Note that the gross domestic product is reported quarterly. We interpolate these data to a monthly frequency following Chow and Lin (1971).

Table 1

Target forecast variables.

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	TB	$Z_{t+1} - Z_t$
Consumer price index	CPI	$\ln(Z_{t+1}/Z_t)$
Producer price index	PPI	$\ln(Z_{t+1}/Z_t)$
Nonfarm payroll employment	NPE	$\ln(Z_{t+1}/Z_t)$
Housing starts	HS	$\ln(Z_t)$
Industrial production	IPX	$\ln(Z_{t+1}/Z_t)$
M2	M2	$\ln(Z_{t+1}/Z_t)$
S&P 500 index	SNP	$\ln(Z_{t+1}/Z_t)$
Gross domestic product	GNP	$\ln(Z_{t+1}/Z_t)$

Notes: The data used in model estimation and prediction construction are monthly US figures for the period 1960:1–2009:5. The data transformations used in the prediction experiments are given in the last column of the table. See Section 4 for further details.

“DM” hereafter) of Diebold and Mariano (1995), which is implemented using quadratic loss, and has a null hypothesis that the two models being compared have equal predictive accuracies (for details describing the importance of accounting for parameter estimation error and nonnest- edness in the DM and related predictive accuracy tests, see Clark & McCracken, 2001; McCracken, 2000, 2004, 2007).⁹ In the simplest case, the DM test statistic has an asymptotic $N(0, 1)$ limiting distribution, under the assumption that parameter estimation error vanishes (i.e., $P/R \rightarrow 0$ as $T, P, R \rightarrow \infty$), and assuming that each pair of models being compared is nonnested. The null hypothesis of the test is $H_0 : E[l(\varepsilon_{t+h|t}^1)] - E[l(\varepsilon_{t+h|t}^2)] = 0$, where $\varepsilon_{t+h|t}^i$ is the i th model’s prediction error and $l(\cdot)$ is the quadratic loss function. The actual statistic in this case is constructed as $DM = P^{-1} \sum_{i=1}^P d_t / \hat{\sigma}_{\bar{d}}$, where $d_t = (\widehat{\varepsilon_{t+h|t}^1})^2 - (\widehat{\varepsilon_{t+h|t}^2})^2$, \bar{d} is the mean of d_t , $\hat{\sigma}_{\bar{d}}$ is a heteroskedasticity and autocorrelation robust estimator of the standard deviation of \bar{d} , and $\widehat{\varepsilon_{t+h|t}^1}$ and $\widehat{\varepsilon_{t+h|t}^2}$ are estimates of the true prediction errors $\varepsilon_{t+h|t}^1$ and $\varepsilon_{t+h|t}^2$. Thus, if the statistic is negative and significantly different from zero, then Model 2 is preferred to Model 1.

4.3. Baseline forecasting models

In addition to the various forecast models discussed above (see specification types 1–4), we also 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 $\hat{Y}_{t+h}^{AR} = \hat{\alpha} + \hat{\phi}(L) Y_t$, with the number of lags p selected using the SIC.

Multivariate autoregression: Forecasts from an $ARX(p)$ model are computed as $Y_{t+h}^{ARX} = \hat{\alpha} + \hat{\beta} Z_t + \hat{\phi}(L) Y_t$, where Z_t is a set of lagged predictor variables selected using

the SIC.¹⁰ Dependent variable lags are also selected using the SIC. The selection of the exogenous predictors involves 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 components regression: Forecasts from a principal component regression are computed as $\hat{Y}_{t+h}^{PCR} = \hat{\alpha} + \hat{\gamma} \hat{F}_t$, where \hat{F}_t is estimated via principal components using X , as in Eq. (5).

Factor augmented autoregression: Based on Eq. (5), forecasts are computed as $Y_{t+h}^h = \hat{\alpha} + \hat{\beta}_F \hat{F}_t + \hat{\beta}_W(L) Y_t$. This model combines an $AR(p)$ model, where the number of lags is selected using the SIC, with the above principal component regression (PCR) model. PCR and factor augmented autoregressive (FAAR) models are estimated using ordinary least squares. The factors in the models above are constructed using PCA, ICA and SPCA.

Combined bivariate ADL model: Following Stock and Watson (2012), we implement a combined bivariate autoregressive distributed lag (ADL) model. The 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 , for $i = 1, \dots, N$. Thus, the model is specified as: $\hat{Y}_{t+h}^{ADL} = \hat{\alpha} + \hat{\beta}_i(L) X_{i,t} + \hat{\phi}_i(L) Y_t$. The combined forecast is $\hat{Y}_{t+h|T}^{Comb,h} = \sum_{i=1}^N w_i \hat{Y}_{t+h|T}^{ADL,h}$. Here, we set $(w_i = 1/N)$ and $N = 144$. In each model, first $p_{i,x}$ and then $p_{i,y}$ are selected, each using the SIC. There are a number of studies that have compared the performances of combining methods in controlled experiments, including those by Clemen (1989), Diebold and Lopez (1996), Newbold and Harvey (2002), and Timmermann (2006); and, in the literature on factor models, Stock and Watson (2004, 2006, 2012) and the references cited therein. This literature has typically found combination methods to outperform individual forecasts. This stylized fact is sometimes called the “forecast combining puzzle”.

⁹ The experiments carried out in this paper do not consider so-called real-time data. However, it is worth noting that the use of real-time datasets in macroeconometrics, and in forecasting and policy analysis in particular, has received a considerable amount of attention in the literature in recent years. For a discussion of DM and related tests using real-time data, the reader is referred to Clark and McCracken (2009).

¹⁰ For this model, autoregressive lags are first selected using the SIC, as in the case of the estimation of the $AR(p)$ model. Thereafter, the first lag of each variable in the entire dataset is added to the model sequentially, where we assuming for simplicity that $h = 1$. If the adjusted R^2 increases by more than 0.01, then the variable is retained, and the search continues until the first lag of each variable has been tried. This process is repeated until six lags of each variable have been tried.

Table 2
Models used in the forecasting experiments.

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
Bagging	Bagging with shrinkage, $c = 1.96$
Boosting	Component boosting, $M = 50$
BMA1	Bayesian model averaging with $g\text{-prior} = 1/T$
BMA2	Bayesian model averaging with $g\text{-prior} = 1/N^2$
Ridge	Ridge regression
LARS	Least angle regression
EN	Elastic net
NNG	Non-negative garrote
Mean	Arithmetic mean

Notes: This table summarizes all of the models used in the forecasting experiments. In addition to estimating the above pure linear and factor models (i.e., AR, ARX, CADL, FAAR, PCR), we also consider the various machine learning and shrinkage methods above, as well as various combined factor and machine learning/shrinkage methods, when implementing our forecasting experiments. Complete details for all models, other than the pure linear models, are given in Section 4.2, where we discuss specification types 1–4, and the various strategies for factor estimation, machine learning and shrinkage method implementation are outlined. For further details, see also Sections 3 and 4.

Mean forecast combination: To examine the issue of forecast combination further, and in addition to the Bayesian model averaging methods discussed previously, we form forecasts as the arithmetic average of the thirteen forecasting models summarized in Table 2, which include all of those outlined in this and previous sections.

Finally, we note that all of the above benchmark models remain unchanged, regardless of the specification type, as we define specification types only in the context of dimension reduction, machine learning, variable selection and shrinkage methods. The only exceptions are our mean forecast combinations, since they combine all of the benchmark forecasts as well as all other models. As was stated above, our entire set of models is listed in Table 2.

5. Empirical results

This section summarizes the results of our prediction experiments. The variables (and transformations thereof) that we forecast are listed in Table 1. There are six different specification “permutations”. 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 specifications. Adding specification types 3 and 4, and multiplying by two (for recursive and rolling window strategies), yields a total of $(12 + 2) \times 2 = 28$ specification types for each target variable and forecast horizon. The forecast models that we use in our experiments are summarized in Table 2. For the sake of brevity, we focus on our key findings and results, rather than reporting the entirety of our experimental findings. Complete details are available from the authors upon request.

Panel A of Table 3 contains the lowest relative MSFEs from amongst all models, for specification types 1–4 and for factors estimated using PCA, ICA, and SPCA, where the

word “relative” is used to mean MSFEs relative to that of the AR(SIC) benchmark. The models to which each of these MSFEs correspond are given in Panel A of Table 4. Thus, the MSFE value of 0.780 given in the upper right corner of Panel A of Table 3 is that obtained using our FAAR model (see the upper right corner of Panel A of Table 4) under specification type 1, with the factors estimated using PCA. Panel A of Tables 3 and 4 correspond to experiments run using recursive estimation. For the corresponding results based on a rolling estimation, refer to Panel B of Tables 3 and 4. The tables are partitioned into three sets of results, corresponding to forecast horizons of $h = 1, 3$, and 12. Entries in bold denote the lowest relative MSFEs from amongst all models for a given model specification (i.e., specification types 1–4). Since the benchmark models, including AR(SIC), ARX, etc., are included as candidate models when selecting the lowest relative MSFEs from amongst all models, it is possible for the MSFE entries in Table 3 to be unity for specification types 1–4 and for factors estimated using PCA, ICA, and SPCA, which would indicate that the AR(SIC) model was MSFE-“best” for a given specification type and factor estimation method. In addition, for a given specification type, the entries for multiple factor estimation methods might be the same. This can occur because the candidate models include some that contain no factors (e.g., the ARX model), and the same ARX model might “win” not only against models estimated using PCA, say, but also against models estimated using SPCA. These cases are made clear by an inspection of the model (in Table 4) to which the MSFE (in Table 3) pertains. For example, in the case of specification type 1 and $h = 1$, the GDP MSFEs are 0.916 for all three factor estimation methods. This is because ARX, one of benchmark models, yields a lower MSFE than any other model for specification type 1, regardless of whether the models that contain factors are estimated using PCA, ICA, or SPCA. Finally, since specification types 3 and 4 do not involve the use of factors, there are no bold entries in the rows that correspond to these specification types. That is, there is no “choice” to be made across factor estimation methods for these specification types.

An inspection of Tables 3–4 leads to a number of clear-cut conclusions.

First, although there are a limited number of exceptions, most of the entries in Table 3 are less than unity, which indicates that our autoregressive model is dominated by other models in virtually all of our experiments. For example, note that, under recursive estimation, the relative MSFE value for IPX when using specification type 1 (SP1) and $h = 1$ is 0.268 (see the third entry in the “IPX” column in Table 3, Panel A). This entry is clearly less than unity. In addition, note that this entry is starred, which indicates that the model corresponding to the entry (i.e., the FAAR model, with factors estimated using SPCA, see the third entry in the “IPX” column in Table 4, Panel A) is statistically significantly different from the AR(SIC) model, at a 90% confidence level, based on an application of the DM test discussed above.

Second, the MSFE-best models, from across all specification types and factor estimation methods, are usually specified using specification type 1 (i.e., SP1), when $h = 1$.

Table 3

Lowest point MSFEs by forecast estimation and factor specification method.

Forecast horizon	Factor spec. mtd.	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP	
Panel A: Recursive window estimation													
h = 1	SP1	PCA	0.780^a	0.870	0.940	0.875	0.943	0.811	0.900	0.800	0.939^a	0.976	0.916^a
		ICA	0.897	0.920	0.931	0.840^a	0.843^a	0.802	0.901	0.574	0.965	0.920	0.916^a
		SPCA	0.827	0.789^a	0.409^a	0.870	0.858	0.706^a	0.542^a	0.268^a	0.969	0.897^a	0.916^a
	SP1L	PCA	0.850^a	0.889^a	0.955^a	0.865^a	0.945^a	0.879^a	0.901^a	0.804^a	0.930^a	0.976^a	0.916^a
		ICA	0.897	0.966	0.978	0.939	0.960	0.918	0.901^a	0.861	0.991	1.002	0.916^a
		SPCA	0.897	0.954	0.987	0.939	0.972	0.881	0.901^a	0.826	0.954	0.998	0.916^a
	SP2	PCA	0.861^a	0.950^a	0.965^a	0.933	0.968	0.854^a	0.901^a	0.833^a	0.942	0.985^a	0.871
		ICA	0.897	0.959	0.971	0.939	0.965^a	0.861	0.901^a	0.874	0.959	0.991	0.867^a
		SPCA	0.897	0.959	0.976	0.939	0.966	0.860	0.901^a	0.873	0.940^a	0.986	0.873
	SP2L	PCA	0.861^a	0.950^a	0.965^a	0.933	0.968	0.854^a	0.901^a	0.833^a	0.942^a	0.985^a	0.871^a
		ICA	0.864	0.957	0.975	0.923^a	0.967	0.862	0.901^a	0.840	0.961	0.993	0.871^a
		SPCA	0.868	0.961	0.974	0.939	0.963^a	0.859	0.901^a	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^a	0.866^a	0.998	0.929	0.910^a	0.819	0.852^a	0.850	0.977	0.994^a	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^a	0.940	0.775^a	0.862	0.816^a	0.942^a	0.997	0.944^a
	SP1L	PCA	0.925^a	0.892^a	0.988	0.901^a	0.929^a	0.818^a	0.852^a	0.838^a	0.978^a	0.993^a	0.963^a
		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^a	0.895^a	0.992^a	0.888	0.945	0.827^a	0.783	0.809^a	0.967	0.995	0.954^a
		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^a	0.993	0.963
	SP2L	PCA	0.916^a	0.895^a	0.992^a	0.888	0.945	0.827^a	0.783	0.809^a	0.967^a	0.995	0.954^a
		ICA	0.916^a	0.902	0.998	0.903	0.945	0.827^a	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^a	0.939^a	0.874	0.818^a	0.919^a	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^a
		SPCA	0.933^a	0.940^a	0.992^a	0.928	0.950	0.845^a	0.841	0.932	0.950^a	0.996	0.993
	SP1L	PCA	0.903^a	0.956^a	0.988^a	0.888^a	0.927^a	0.860^a	0.829^a	0.926^a	0.942^a	0.995^a	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^a	0.891^a	0.950^a	0.816^a	0.749	0.916^a	0.930	0.995^a	0.982^a
		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^a	0.948^a	0.997	0.935	0.957	0.843	0.910	0.919	0.916^a	0.997	0.992
	SP2L	PCA	0.926	0.949^a	0.992^a	0.891^a	0.950^a	0.816^a	0.749	0.916^a	0.930^a	0.995	0.982^a
		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^a	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													
h = 1	SP1	PCA	0.787^a	0.909^a	0.944^a	0.843^a	0.971	0.831^a	0.841^a	0.803^a	0.863^a	0.998^a	0.940^a
		ICA	0.871	1.014	0.977	0.876	0.973	0.918	0.841^a	0.910	0.918	0.998^a	0.948
		SPCA	0.871	1.023	0.977	0.883	0.996	0.877	0.841^a	0.875	0.869	1.007	0.945
	SP1L	PCA	0.852^a	0.989^a	0.954^a	0.850^a	0.973	0.871^a	0.841^a	0.845^a	0.845^a	1.002	0.943
		ICA	0.871	1.004	0.982	0.883	0.985	0.924	0.841^a	0.877	0.908	1.008	0.941^a
		SPCA	0.871	1.081	0.992	0.883	1.003	0.911	0.841^a	0.851	0.880	1.008	0.989
	SP2	PCA	0.871^a	1.085	0.963^a	0.849	0.936^a	0.869^a	0.841^a	0.858^a	0.889	0.998^a	0.915^a
		ICA	0.871^a	1.114	0.977	0.849	0.941	0.884	0.841^a	0.858^a	0.908	1.006	0.915^a
		SPCA	0.871^a	1.087	0.979	0.844^a	0.949	0.877	0.841^a	0.892	0.888^a	1.007	0.927
	SP2L	PCA	0.871^a	1.088	0.964^a	0.850	0.948^a	0.865^a	0.841^a	0.833^a	0.886^a	0.997^a	0.905^a
		ICA	0.871^a	1.100	0.977	0.843	0.953	0.880	0.841^a	0.841	0.909	1.004	0.905^a
		SPCA	0.871^a	1.095	0.979	0.840^a	0.957	0.879	0.841^a	0.864	0.910	1.004	0.915

(continued on next page)

This can be seen from an inspection of the block of entries in Table 3 (Panel A) that are associated with $h = 1$, and

by noting that the lowest “bold” MSFE falls under SP1 for each of the 11 columns of entries in this block of MSFEs,

Table 3 (continued)

Forecast horizon	Factor spec. mtd.	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
$h = 3$	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
	PCA	0.882^a	0.872^a	1.002	0.861^a	0.937	0.786^a	0.769^a	0.835^a	0.914^a	0.997^a	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
	PCA	0.904^a	0.889^a	0.981^a	0.848^a	0.920	0.807^a	0.744^a	0.820^a	0.908^a	0.988	0.953^a
	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
	PCA	0.895^a	0.883^a	0.998	0.875	0.941	0.814^a	0.740	0.833^a	0.912^a	0.989	0.929^a
	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
	PCA	0.889	0.886^a	0.988^a	0.864	0.942	0.792^a	0.738	0.823^a	0.911^a	0.985^a	0.938^a
	ICA	0.888^a	0.901	0.998	0.865	0.941	0.792^a	0.806	0.838	0.921	0.985^a	0.947
	SPCA	0.927	0.919	1.002	0.861^a	0.936	0.843	0.772	0.858	0.929	0.985^a	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
	PCA	0.897	0.935^a	0.997^a	0.812	0.891	0.729	0.723	0.884^a	0.896^a	1.007	1.010
	ICA	0.930	0.944	0.997^a	0.863	0.949	0.779	0.741	0.909	0.937	0.996	0.999
	SPCA	0.879^a	0.953	0.997^a	0.781	0.920	0.720^a	0.715^a	0.890	0.904	1.006	0.997^a
	PCA	0.864^a	0.946^a	0.997	0.819	0.902	0.737	0.726	0.898^a	0.899^a	1.000	0.996
	ICA	0.908	0.951	0.997	0.872	0.962	0.730^a	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^a
	PCA	0.893^a	0.929^a	0.997^a	0.818^a	0.912^a	0.692	0.637	0.880^a	0.884	0.994	0.994
	ICA	0.911	0.932	0.997^a	0.833	0.915	0.691^a	0.726	0.902	0.888	0.994	0.993
	SPCA	0.901	0.935	0.997^a	0.819	0.921	0.692	0.693	0.896	0.879	0.991	0.991^a
	PCA	0.883^a	0.927^a	0.997^a	0.816^a	0.903^a	0.714^a	0.624	0.888^a	0.880^a	0.993	0.996
	ICA	0.895	0.929	0.997^a	0.835	0.917	0.719	0.695	0.898	0.897	0.994	0.993
	SPCA	0.888	0.935	0.997^a	0.836	0.910	0.722	0.768	0.897	0.905	0.994	0.991^a
$h = 12$	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 the notes to Tables 1 and 2. The numerical entries in this table are the lowest (relative) mean square forecast errors (MSFEs), based on the use of models estimated with recursive (Panel A) and rolling (Panel B) data windowing methods, and three different factor estimation methods (PCA, ICA and SPCA; see Section 2 for further discussion), for six different specification types (SP1, SP1L, SP2, SP2L, SP3, and SP4; see Section 4 for details). The prediction models and target variables are described in Tables 1 and 2. The forecasts are monthly, for the period 1974:3–2009:5. Forecasts are reported for horizons $h = 1, 3$, and 12. Tabulated relative MSFEs are calculated such that numerical values less than unity constitute cases in which the alternative model has a lower point MSFE than the AR(SIC) model. Entries in bold denote point-MSFE “best” models among the three factor estimation methods, for a given specification type, estimation window and forecast horizon. For a listing of these MSFE “best” models, compare Panel A of Table 3 with Panel A of Table 4. See Section 5 for further details.

^a In addition, bold entries with “^a” indicate instances in which the AR(SIC) model is statistically inferior to the model yielding the stated “best” MSFE.

with two exceptions (i.e., CPI and GDP). Furthermore, when comparing all 11 of these so-called “overall winners” for the $h = 1$ case, we see that PCA is the chosen factor estimation method for only two variables (i.e., UR and M2), while SPCA “wins” for six variables, and ICA “wins” for two.¹¹ This supports the use of SPCA and ICA at the $h = 1$ forecast horizon. Note also that the fact that SP1 “wins” over SP1L (recall that SP1 uses no lags of factors, while SP1L is exactly the same, except that factor lags are included) indicates that additional factor lags, over and above the single lag used under SP1 to ensure that our experiments are *ex ante*, are not useful when $h = 1$.

Third, note that the above conclusion with regard to the performances of the SPCA and ICA factor estimation methods, for $h = 1$, is based solely upon an inspection of the results associated with recursive estimation. Interestingly, this conclusion largely remains intact when the MSFEs associated with the “overall winners” discussed above are

compared with those of the “overall winners” under rolling estimation (see Tables 3 and 4, Panel B). Indeed, for this case (i.e., $h = 1$), rolling estimation yields only one “overall winner” that is lower than the corresponding “overall winner” for the recursive estimation case. This is for the case of CPI, where the rolling estimation under specification type 4 yields a lower MSFE than that of any other method under any other modelling permutation. This finding clearly supports the use of recursive estimation when $h = 1$. To simplify further discussion, let us define the “globally best” MSFE as the lowest “overall winner” when comparing results under both the rolling and recursive estimation strategies. All of the results discussed in the context of our “globally best” models are summarized in Table 5. In particular, for each forecast horizon, one can read from Table 5 the “globally best” specification type, estimation window type, factor estimation method, and model.

Fourth, although recursive estimation yields the “globally best” MSFEs for all 11 variables for $h = 1$, such is not the case for $h = 3$ or $h = 12$. Indeed, for $h = 3$, the lowest MSFEs across all specification type, factor es-

¹¹ For CPI, the “overall winner” does not incorporate factors, and hence the sum of these “wins” is only 10.

Table 4

The forecast models that correspond to the lowest point MSFEs reported in Table 3.

Forecast horizon		Factor spec. mtd.	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
Panel A: Recursive window estimation													
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	Mean	BMA1	Mean	Mean	ARX
h = 3	SP1	PCA	PCR	PCR	CADL	FAAR	PCR	FAAR	Boost	Mean	Mean	LARS	Mean
		ICA	FAAR	ARX	PCR	FAAR	ARX	FAAR	LARS	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	LARS	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	LARS	Boost	ARX	Boost	Boost	Boost	Mean	Mean	Mean
		SPCA	Mean	ARX	CADL	Mean	ARX	Mean	Boost	Mean	Boost	LARS	Mean
	SP2L	PCA	Boost	Mean	EN	Boost	ARX	Boost	Boost	Mean	Mean	Mean	Mean
		ICA	Boost	ARX	CADL	Boost	ARX	Boost	Boost	LARS	Mean	Mean	Mean
		SPCA	Mean	ARX	BMA2	Mean	ARX	Mean	Boost	LARS	Boost	Mean	Mean
	SP3		Boost	ARX	CADL	Mean	ARX	Mean	Mean	BMA2	Mean	AR	Boost
		SP4	Mean	ARX	Mean	Mean	ARX	Mean	Mean	NNG	Mean	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	LARS	Mean
	SP1L	PCA	Mean	Mean	Boost	Mean	Mean	Mean	Mean	Mean	Boost	LARS	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	LARS	LARS
		ICA	Mean	Mean	CADL	Boost	Mean	EN	Boost	Mean	Mean	LARS	Mean
		SPCA	Boost	Mean	CADL	Mean	Mean	EN	Boost	Mean	Mean	LARS	Mean
	SP2L	PCA	Mean	Mean	Mean	BMA1	Mean	Boost	Boost	Mean	Mean	LARS	LARS
		ICA	Mean	Mean	BMA2	Boost	Mean	Boost	Boost	Mean	Mean	Mean	LARS
		SPCA	Boost	Mean	Mean	Mean	Mean	Mean	Boost	Mean	Mean	BMA2	LARS
	SP3		Boost	Boost	CADL	Mean	Mean	Boost	EN	EN	Mean	AR	EN
	SP4	Mean	Mean	CADL	Mean	Mean	Mean	Mean	Boost	Mean	Mean	AR	Mean
Panel B: Rolling window estimation													
h = 1	SP1	PCA	FAAR	PCR	Mean	FAAR	Mean	FAAR	ARX	PCR	FAAR	LARS	Mean
		ICA	ARX	AR	Mean	Mean	Mean	Mean	ARX	ARX	Mean	NNG	Mean
		SPCA	ARX	AR	Mean	Mean	ARX	LARS	Mean	ARX	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	LARS
	SP2	PCA	ARX	AR	Mean	Mean	LARS	Mean	ARX	Boost	Mean	EN	EN
		ICA	ARX	AR	Mean	Mean	LARS	Mean	ARX	Boost	Mean	AR	EN
		SPCA	ARX	AR	Mean	Boost	LARS	Mean	ARX	Mean	Mean	AR	LARS
	SP2L	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	LARS

(continued on next page)

timization method and data window choices (i.e., the “globally best” MSFEs) are obtained via rolling estimation for six of the 11 variables. Moreover, for $h = 12$, the “globally best” MSFEs are obtained via rolling estimation for nine of

the 11 variables. When comparing factor estimation methods for these “globally best” models, we see that PCA dominates for eight of the 11 variables when $h = 3$, and for seven when $h = 12$. Thus, the choice between using PCA or

Table 4 (continued)

Forecast horizon	Factor spec. mtd.	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
$h = 3$	PCA	ARX	AR	Mean	Mean	EN	Mean	ARX	BMA2	Mean	LARS	LARS
	SP2L ICA	ARX	AR	Mean	Mean	EN	Mean	ARX	Boost	Mean	AR	LARS
	SPCA	ARX	AR	Mean	Mean	Mean	Mean	ARX	Boost	Mean	AR	LARS
	SP3	ARX	AR	CADL	Boost	AR	Boost	ARX	Boost	LARS	AR	EN
	SP4	ARX	AR	Boost	BMA2	Mean	Mean	ARX	Mean	Boost	AR	Mean
	PCA	Mean	PCR	AR	Mean	Mean	PCR	Boost	Mean	FAAR	LARS	Boost
	SP1 ICA	Mean	Mean	PCR	Mean	Mean	Mean	Bagg	Mean	Bagg	AR	Mean
	SPCA	Mean	Mean	BMA2	BMA1	Mean	Mean	Mean	Mean	Mean	AR	Mean
	PCA	Mean	Mean	LARS	Mean	Mean	Mean	Boost	Mean	Mean	Mean	Mean
	SP1L ICA	Mean	Mean	AR	BMA2	Boost	Mean	Boost	Mean	Mean	AR	Mean
	SPCA	Mean	Mean	AR	BMA2	NNG	Mean	Mean	Mean	Mean	AR	LARS
	PCA	Mean	Mean	NNG	Mean	Mean	BMA2	Boost	Mean	EN	NNG	LARS
	SP2 ICA	Mean	Mean	BMA2	Mean	Mean	Mean	Boost	Mean	EN	NNG	Mean
	SPCA	Boost	Mean	BMA1	BMA2	Mean	Mean	Boost	Mean	Mean	NNG	Mean
	PCA	Boost	Mean	BMA1	Mean	Mean	Boost	BMA2	Mean	Mean	NNG	Mean
	SP2L ICA	Boost	Mean	BMA2	Mean	Mean	Boost	Boost	Boost	Boost	NNG	Mean
	SPCA	Mean	Mean	AR	Mean	Mean	Boost	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	LARS
	PCA	Mean	Mean	CADL	Mean	PCR	FAAR	Boost	Mean	Mean	AR	AR
	SP1 ICA	Mean	Mean	CADL	Ridge	Mean	Mean	FAAR	Mean	Mean	Bagg	Mean
	SPCA	Mean	Mean	CADL	BMA2	Mean	Mean	Mean	Mean	Mean	AR	Mean
	PCA	Mean	Mean	CADL	Mean	Mean	Mean	Mean	Mean	Mean	AR	NNG
	SP1L ICA	Mean	Mean	CADL	Mean	Mean	Mean	Mean	Mean	Mean	AR	Bagg
	SPCA	Mean	NNG	NNG	BMA2	Boost	Mean	Mean	LARS	LARS	AR	LARS
$h = 12$	PCA	Mean	Mean	CADL	Mean	Mean	EN	Boost	Mean	Boost	NNG	Mean
	SP2 ICA	Mean	Mean	CADL	Mean	Mean	EN	Boost	Mean	Boost	NNG	Mean
	SPCA	Mean	Mean	CADL	Mean	Mean	EN	Boost	Mean	Boost	LARS	Mean
	PCA	Mean	Mean	CADL	Mean	Mean	Boost	Boost	Mean	Mean	BMA2	Mean
	SP2L ICA	Mean	Mean	CADL	Mean	Mean	Boost	Boost	Mean	Boost	NNG	Mean
	SPCA	Mean	Mean	CADL	Mean	LARS	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

Notes: See the notes to Tables 1–3. Panels A and B report the MSFE “best” models, based on the results reported in Table 3.

Table 5

Summary of winning methods and models 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	LARS
$h = 12$	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	LARS	Bagg

Notes: See the notes to Tables 1–3. This table contains details of the winning forecast model/method for each forecast horizon. The entries correspond to the lowest (bold) MSFE entries in Table 3 for each forecast horizon, across all specification types, for each variable. In summary, the “winning” method (i.e., specification type, estimation windowing method, factor estimation method – when factors enter into the “best” model, and model – as given in Table 2) is summarized for each forecast horizon and target variable.

one of our other factor estimation methods becomes more difficult at longer forecast horizons, and on average it is better to use PCA.

Fifth, as was discussed above, Table 4 shows which forecast models (see the list of models in Table 2) have the lowest relative MSFEs, as reported in Table 3, for

each target variable, and for each specification type, factor estimation method, and forecast horizon (Panel A summarizes the results for recursive estimation, while Panel B does the same for rolling estimation). For example, the upper-leftmost three entries in Panel A of Table 4 show that, for unemployment, the FAAR, ARX, and FAAR

models result in the MSFE-best predictions, under SP1 and for each of PCA, ICA, and SPCA, respectively, given recursive estimation. The corresponding MSFEs for these models, as reported in Table 3 (Panel A), are 0.780, 0.897 and 0.827, respectively. Again as discussed above, bold entries in Table 4 denote the forecasting models that yield the MSFE-best predictions for a given specification type, forecast horizon, and target variable. When comparing only the “globally best” models across Table 4 (Panels A and B), which we have defined to be the MSFE-best models for each variable across all specification and modeling permutations, we see that for $h = 1$, FAAR wins four times, PCR wins twice, Mean or BMA wins three times, and Boost wins twice. Here, Boost is estimated under specification type 1, indicating that it involves the use of estimated factors. Thus, overall, 10 of the 11 “globally best” models are factor-based models, since Mean also uses factor-type models.¹² Moreover, Mean only wins twice. This evidence is strongly in favor of the use of factor models for forecasting macroeconomic variables when $h = 1$, and provides evidence against the oft-noted success of Bayesian averaging and arithmetic mean combinations, since Mean only “wins” twice.

Sixth, when the above model assessment is carried out for $h = 3$ and $h = 12$, we see that the following “wins” obtain. For $h = 3$: PCR (3), Mean (4), Boosting (1), LARS (1), NNG (1), and Ridge (1). For $h = 12$: PCR (1), Mean or BMA (4), Boosting (4), LARS (1), and Bagging (1). At both of these horizons, nine of the 11 “winning” models incorporate factors, which supports our conclusion above concerning the usefulness of factor models. Interestingly, our machine learning and shrinkage type models fare much better at the longer forecast horizons, being critical for four of the 11 variables when $h = 3$, and again for six of the 11 variables when $h = 12$. This feature of our results might be explained in part by the presence of structural breaks that are more “damaging” to predictions made at longer forecast horizons (see the introduction for further discussion, including a discussion of why ICA and SPCA might be preferred to PCA for $h = 1$ but not for $h > 1$). All of the above results in the context of our “globally best” models are summarized in Table 5. In particular, the “globally best” specification type, estimation window type, factor estimation method, and model for each forecast horizon can be read from this table.

Overall, our findings suggest that the dimension reduction associated with the specification and estimation of factors, as well as with machine learning and shrinkage methods, are very useful for forecasting macroeconomic variables, when analyzing “big data”. The exact method and model that it is best to use is case-specific, as might be expected, although dimension reduction methods seem useful at all forecast horizons, while the machine learning and shrinkage methods are more useful at longer forecast horizons. Finally, there is substantial evidence to suggest that SPCA and ICA offer interesting alternatives to the use of PCA when estimating factors, particularly for one-step-ahead prediction.

6. Concluding remarks

In this paper, we outline and discuss a number of interesting new forecasting methods that have been developed recently in the statistics and econometrics literatures. We focus in particular on the examination of a variety of factor estimation methods, including principal components analysis (PCA), independent component analysis (ICA), and sparse principal component analysis (SPCA), together with hybrid forecasting methods that use these factor estimation methods in conjunction with various types of machine learning, variable selection and shrinkage methods, including bagging, boosting, least angle regression, the elastic net, and the nonnegative garrote, for example. We analyze all models and methods by carrying out a series of real-time prediction experiments, in the context of predicting 11 key macroeconomic indicators at various forecast horizons. We find that simple time series models and model averaging methods do not dominate hybrid methods that couple factor estimation methods with machine learning and shrinkage methods. We also find that SPCA and ICA are useful alternatives to PCA, perhaps due to their sparseness features. Overall, we find strong new evidence of the usefulness of the dimension reduction associated with the specification and estimation of factors, and find that combining such dimension reductions with learning and shrinkage methods yields promising results when forecasting macroeconomic variables.

Acknowledgments

The authors owe many thanks to the editor, Michael McCracken, an associate editor, two anonymous referees, Nii Armah, Valentina Corradi, David Hendry, Gary Koop, John Landon-Lane, Fuchun Li, Greg Tkacz, Hiroki Tsurumi and Halbert White for useful comments on earlier drafts of this paper. Additional thanks are owed to participants at the Sir Clive W.J. Granger Memorial Conference held at Nottingham University in May 2010, the Korean International Economics Association Conference, the Eastern Economic Association Conference and the International Symposium on Forecasting, as well as to participants at seminars at the Bank of Canada, Bank of Korea, European Central Bank, Rutgers University, Yonsei University, and Peking University. The views stated herein are those of the authors, and are not necessarily those of the Bank of Korea.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/j.ijforecast.2016.02.012>.

References

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

¹² The “winning” CPI model is BMA, estimated under specification type 4, and hence, factors enter into only 10 of the 11 models.

- Armah, N. A., & 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.
- Artis, M. J., Banerjee, A., & Marcellino, M. (2005). Factor forecasts for the UK. *Journal of Forecasting*, 24(4), 279–298.
- Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191–221.
- Bai, J., & Ng, S. (2006a). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4), 1133–1150.
- Bai, J., & Ng, S. (2006b). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, 131(1–2), 507–537.
- Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2), 304–317.
- Bai, J., & Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, 24(4), 607–629.
- Banerjee, A., & Marcellino, M. (2008). Factor-augmented error correction models. CEPR Discussion Papers 6707, C.E.P.R. Discussion Papers.
- Boivin, J., & Ng, S. (2005). Understanding and comparing factor-based forecasts. *International Journal of Central Banking*, 1(3), 117–152.
- Boivin, J., & 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., & Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, 30, 927–961.
- Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103, 1438–1456.
- Chen, Y.-P., Huang, H.-C., & Tu, I.-P. (2010). A new approach for selecting the number of factors. *Computational Statistics and Data Analysis*, 54, 2990–2998.
- Chow, G. C., & 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), 372–375.
- Clark, T., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105, 85–110.
- Clark, T., & McCracken, M. W. (2009). Tests of equal predictive ability with real-time data. *Journal of Business and Economic Statistics*, 27, 441–454.
- Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4), 559–583.
- Comon, P. (1994). Independent component analysis – a new concept? *Signal Processing*, 36, 287–314.
- Connor, G., & 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., & 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., & Korajczyk, R. A. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance*, 48(4), 1263–1291.
- Diebold, F. X., & Lopez, J. A. (1996). Forecast evaluation and combination. NBER Technical Working Papers 0192, National Bureau of Economic Research, Inc.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3), 253–263.
- Ding, A. A., & 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., & Stevanovic, D. (2010). Factor-augmented VARMA models: Identification, estimation, forecasting and impulse responses. Working paper, McGill University.
- Efron, B., Hastie, T., Johnstone, L., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32, 407–499.
- Fan, J., Rigollet, P., & Wang, W. (2015). Estimation of functionals of sparse covariance matrices. *Annals of Statistics*, 43, 2706–2737.
- Fernandez, C., Ley, E., & Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5), 563–576.
- Forni, M., Hallin, M., Lippi, M., & 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., & Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100, 830–840.
- Freund, Y., & 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.
- Gelper, S., & Croux, C. (2008). *Least angle regression for time series forecasting with many predictors*. Technical Report. Katholieke Universiteit Leuven.
- Guo, J., James, G., Levina, E., Michailidis, G., & Zhu, J. (2010). Principal component analysis with sparse fused loadings. *Journal of Computational and Graphical Statistics*, 19(4), 947–962.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning* (2nd ed.). Springer.
- Hyvärinen, A. (1998). Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22, 49–67.
- Hyvärinen, A. (1999). Gaussian moments for noisy independent component analysis. *IEEE Signal Processing Letters*, 6(6), 145–147.
- Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4–5), 411–430.
- Inoue, A., & Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of US CPI inflation. *Journal of the American Statistical Association*, 103(482), 511–522.
- Jolliffe, I. T. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22, 29–35.
- Jolliffe, I., Trendafilov, N., & Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12, 531–547.
- Josse, J., & Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximation. *Computational Statistics and Data Analysis*, 56, 1869–1879.
- Kim, H. H., & Swanson, N. R. (2014a). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, 178(2), 352–367.
- Kim, H. H., & Swanson, N. R. (2014b). Mining big data using parsimonious factor and shrinkage methods. Working paper, Rutgers University.
- Koop, G., & 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* (1st ed.). Boston, Massachusetts: Springer.
- Leng, C., & Wang, H. (2009). On general adaptive sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 18(1), 201–215.
- Li, Y.-O., Adali, T., & Calhoun, V. D. (2007). Estimating the number of independent components for functional magnetic resonance imaging data. *Human Brain Mapping*, 28(11), 1251–1266.
- Mayrink, V. D., & Lucas, J. E. (2013). Sparse latent factor models with interactions: Analysis of gene expression data. *The Annals of Applied Statistics*, 7, 799–822.
- McCracken, M. W. (2000). Robust out-of-sample inference. *Journal of Econometrics*, 99, 195–223.
- McCracken, M. W. (2004). Parameter estimation error and tests of equal forecast accuracy between non-nested models. *International Journal of Forecasting*, 20, 503–514.
- McCracken, M. W. (2007). Asymptotics for out-of-sample tests of Granger causality. *Journal of Econometrics*, 140, 719–752.
- Moneta, A., Entner, D., Hoyer, P., & Coad, A. (2013). Causal inference by independent component analysis with applications to micro- and macroeconomic data. *Oxford Bulletin of Economics and Statistics*, 75, 705–730.
- Neto, P., Jackson, D., & Somers, K. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, 49, 974–997.
- Newbold, P., & Harvey, D. I. (2002). Forecast combination and encompassing. In M. P. Clements, & D. F. Hendry (Eds.), *A companion to economic forecasting* (pp. 268–283). Oxford: Blackwell Press.
- Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77, 1447–1479.
- Ravazzolo, F., Paap, R., van Dijk, D., & Franses, P. H. (2008). *Frontier of economics and globalization, Bayesian model averaging in the presence of structural breaks*. (chapter 15).
- Ridgeway, G., Madigan, D., & Richardson, T. (1999). Boosting methodology for regression problems. In *The seventh international workshop on artificial intelligence and statistics*, (Uncertainty '99). (pp. 152–161). Morgan Kaufmann.
- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2), 197–227.
- Shrestha, D. L., & Solomatine, D. P. (2006). Experiments with adaboost.rt, an improved boosting scheme for regression. *Neural Computation*, 18(7), 1678–1710.
- Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2), 293–335.

- Stock, J. H., & 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., & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20(2), 147–162.
- Stock, J. H., & 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., & Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. NBER Working Papers 11467, National Bureau of Economic Research, Inc.
- Stock, J. H., & Watson, M. W. (2006). Forecasting with many predictors. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting*, Vol. 1 (pp. 515–554). Elsevier, (chapter 10).
- Stock, J. H., & Watson, M. W. (2008). Forecasting in dynamic factor models subject to structural instability. In J. Castle, & N. Shephard (Eds.), *The methodology and practice of econometrics, a festschrift in honour of Professor David F. Hendry*. Oxford University Press.
- Stock, J. H., & Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics*, 30(4), 481–493.
- Stone, J. V. (2004). *Independent component analysis*. MIT Press.
- Tan, L., & Zhang, H. (2012). Forecast of employment based on independent component analysis. In *Information computing and applications, third international conference, ICICA 2012, vol. Part I, CCIS 307*. (pp. 373–381). Springer-Verlag.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267–288.
- Timmermann, A. G. (2006). Forecast combinations. In G. Elliott, & A. Timmermann (Eds.), *Handbook of economic forecasting*, vol. 1 (pp. 135–196). Elsevier, (chapter 4).
- Tong, L., Liu, R.-w., Soon, V., & 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.
- Yau, R. (2004). Macroeconomic forecasting with independent component analysis. Econometric Society 2004 Far Eastern Meetings, 741.
- Yuan, M., & Lin, Y. (2007). On the non-negative garotte estimator. *Journal of the Royal Statistical Society*, 69(2), 143–161.
- Zou, H., & 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., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2), 262–286.

Hyun Hak Kim is a recent graduate of Rutgers University, and now holds a post at the Bank of Korea. His primary research interests include econometric theory, financial and macro econometrics, time series analysis, and forecasting. He has published in leading econometrics journals, including the *Journal of Econometrics* and various other field journals.

Norman R. Swanson was educated at the University of Waterloo and the University of California, San Diego. He is currently Professor and Director of Graduate Studies in the Economics Department at Rutgers University. He has held previous positions at Pennsylvania State University (1994–1999), Texas A&M University (1999–2001), and Purdue University (2001–2002). His primary research interests include econometric theory, financial and macro econometrics, time series analysis, and forecasting. He is a fellow of the *Journal of Econometrics*, he is currently an associate editor of various econometrics journals, and he has served as guest editor for journals ranging from *Journal of Econometrics* to *Journal of Business and Economic Statistics*. Swanson has recent scholarly publications in leading economics and statistics journals including *Econometrica*, *Journal of Econometrics*, *Review of Economics and Statistics*, *Journal of Business and Economic Statistics*, and the *Journal of the American Statistical Association*, among others; and is or has acted as ongoing visiting scholar to various central banks including The Federal Reserve Bank of Philadelphia and the Bank of Canada.