ARTICLE IN PRESS
International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

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 macroeconometric forecasting.

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 cavailable in the

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

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

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

$$X = F\Lambda' + e, \tag{1}$$

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):

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

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

[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.

Please cite this article in press as: Kim, H. H., & Swanson, N. R., Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2016.02.012