Forecasting with dynamic factor models

Abstract The validity of previous findings that dynamic factor models are useful for macroeconomic forecasting is of great importance for subsequent studies which use these models not only as a starting point for further developments but also as a benchmark for the evaluation of the forecasting performance of these further developments. Reanalyzing a standard macroeconomic dataset, we do not find any evidence corroborating the usefulness of dynamic factor models. We therefore explore two possible ways for improvement. First, we try to find those factors which have the greatest predictive power and then we try to utilize frequency-domain information. Our empirical results indicate that only the latter attempt is promising. Focusing on the low-frequency components of the macroeconomic time series and disregarding the high-frequency components can actually improve the forecasting performance.

Keywords Dynamic Factor Models, Forecasting, Model Selection, Shrinkage, Band Regression

JEL C52; C53; C55; E17.

1 Introduction

Factor models have become increasingly popular for the efficient extraction of information from a large number of macroeconomic variables. To investigate the forecasting performance of these models, Eickmeier and Ziegler (2008) conducted a meta-analysis of 52 studies and obtained mixed results that depended on the region, the category of the variable to be predicted, the size of the data set, and the estimation technique. This is aggravated by the facts that it is a priori not clear how many factors should be included (Bai and Ng 2002, 2006, 2008b) and that the findings change noticeably when different sub-periods (states of the business cycle) are considered (Kim and Swanson 2014). Accordingly, many efforts have been made to improve the standard factor-augmented forecast which is based on lagged values of the variable of interest and a small number of factors. Two approaches are of particular interest. The first approach by Bai and Ng (2008a) allows the factors to be used as predictors to depend on the variable to be predicted. As pointed out by the authors, the obvious procedure to include the factors in their natural order fails to take their predictive power into account, hence it could possibly be improved by using fewer but more informative predictors (targeted-predictors).

The second approach applies high-dimensional methods to large sets of factors. Noticing the difficulty in comparing these high-dimensional methods theoretically because of differences in the modeling assumptions and empirically because of differences in the datasets and implementations,
Stock and Watson (2012) provided a simple shrinkage representation that covers pretest and information criterion methods, Bayesian model averaging, empirical Bayes methods, as well as bagging and examined in an empirical analysis of a large macroeconomic dataset whether the shrinkage methods can outperform the standard factor-augmented forecast (based on those five factors with the largest variances). They found that this was not the case. However, factor proponents might take some small comfort in the fact that the standard forecast appeared to improve upon a simple autoregressive benchmark for a group of variables that included the major measures of economic activity.

In general, it is difficult to assess the significance of any further development of an existing method because the improved method is usually much more complex and depends on a larger number of tuning parameters which increases the risk of a data-snooping bias. Moreover, often the same data set (possibly with a slightly extended investigation period) is used to evaluate the further development that has already been used for the evaluation of the original method as well as for the design of the further development. Of course, it is not helping when there are doubts about the usefulness of the original method. In the case of the standard factor-augmented forecast, Stock and Watson (2012) had to select the data set and the investigation period, the variables to be predicted, the transformations (e.g., taking logarithms and/or differencing) to be applied to achieve stationarity, the number of lagged values, and the number of factors. Their claim of the outperformance of the standard forecast over a simple autoregressive forecast encourages not only the exploration of much more sophisticated further developments (see, e.g., Kim and Swanson, 2014) but also the use of the standard forecast as a benchmark for the performance of new forecasts (see, e.g., Cheng and Hansen, 2015). Clearly, the results of studies using this benchmark will be severely compromised if their claim is not justified. After all, what is the point of beating a bad benchmark?

There are two major goals of this paper. The first is to scrutinize the usefulness of the standard factor-augmented forecast. To that end, we reanalyze the dataset used by Stock and Watson (2012) with a focus on the appropriateness of the dataset and the continuous evaluation of the forecasting performance throughout the whole investigation period. Our second major goal is to explore options for possible improvements. Taking up the idea that relationships between variables may exist only in certain frequency bands (Hannan, 1963; Engle, 1974; Reschenhofer and Chudy, 2015), we examine whether the use of frequency-domain information can improve forecasts based on factor models. We also address the central issue of how to select the factors. We may only try to determine the number of factors to be included and then just use the first factors (in their natural order) or, alternatively, find that subset of factors which has the greatest predictive power. It is only in the first case that conventional model selection criteria such as AIC and BIC are adequate. In the
second case, criteria specially designed for nonnested models (Foster and George, 1994; Tibshirani and Knight, 1999; George and Foster, 2000; Reschenhofer, 2015) should be used.

The rest of the paper is organized as follows. Section 2 discusses the models, the model-identification methods, and the forecasting techniques. In Section 3, the data as well as the data transformations are described and the empirical results are presented. Section 4 concludes.

2 Methods

2.1 Simple linear forecasts

Let \( Y_1, \ldots, Y_n \) be the time series of interest and \( X_{1h}, \ldots, X_{nh}, h=1, \ldots, H \), a large number of possible predictor series. Following Stock and Watson (2012), the series are first subjected to suitable transformations in order to achieve stationarity. Then, autoregressive dynamics are partialed out by regressing \((Y_2, \ldots, Y_n)'\) on \((1, \ldots, 1)'\), \((Y_1, \ldots, Y_n-1)'\), \ldots, \((Y_2, \ldots, Y_{n-4})'\) and for each \( h \), \((X_{1h}, \ldots, X_{nh})'\) on \((1, \ldots, 1)'\), \((Y_1, \ldots, Y_n)'\), \ldots, \((Y_2, \ldots, Y_{n-3})'\), where it is assumed that the observations \( Y_0, Y_1, Y_2 \) are available. Denoting the estimated parameter vector of the autoregression by \( \hat{\phi} = (\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_H) \), the residual vector by \( y = (y_2, \ldots, y_n)' \), and the residual vectors of the other regressions by \( x^*_h = (x_{1h}, \ldots, x_{nh})' \), \( h=1, \ldots, H \), a simple forecast of \( y_{n+1} \) is given by

\[
\hat{y}_{n+1} = \sum_{k \in M} \hat{\beta}_k x_{nk},
\]

where the OLS estimates \( \hat{\beta}_k \) are obtained by regressing \( y \) on \( x_k = (x_{1k}, \ldots, x_{(n-1)k})' \), \( k \in M \subseteq \{1, \ldots, H\} \), and the associated forecast of \( Y_{n+1} \) is given by

\[
\hat{Y}_{n+1} = \hat{\phi}_0 + \hat{\phi}_1 Y_n + \ldots + \hat{\phi}_3 Y_{n-3} + \hat{y}_{n+1}.
\]

The separate estimation of the parameters \( \phi_k \) and \( \beta_l \) can be justified by orthogonality arguments.

2.2 Selecting the predictors in the case of nonnested models

Model selection criteria try to balance the tradeoff between the goodness-of-fit of a model and its complexity. For example, the FPE (Rothman, 1968; Akaike, 1969) uses the residual sum of squares and the number of predictors for the quantification of these conflicting objectives and selects that model which minimizes the product of the residual sum of squares and a penalty term, which increases as the number of predictors increases. The penalty term of the FPE is constructed so that the product is an unbiased estimator of the mean squared prediction error. If a predictor is included
that is actually dispensable, it will still explain some random fluctuations and thereby reduce the sum of squared residuals. Clearly, this reduction will be much greater if this predictor is not fixed a priori but is found by data snooping, i.e., by trying different predictors and choosing the one which fits best. The FPE-penalty term just neutralizes the effect of the inclusion of a number $h$ of fixed predictors, hence its penalization will not be harsh enough if the “best” $h$ predictors are chosen from a set of $H > h$ predictors. A data-snooping bias can only be avoided by using a penalty term that depends on both $h$ and $H$. However, the two most widely used model selection criteria, namely AIC (Akaike, 1973), which is asymptotically equivalent to FPE in linear regression models, and BIC (Schwarz, 1978), take only the number $h$ of actually included predictors into account. Thus, the (asymptotic) unbiasedness of AIC as well as the consistency of BIC are guaranteed only in the case of nested models where there is only one candidate model for each model dimension.

In the case of nonnested models (with orthogonal predictors), unbiasedness can be achieved by replacing in the multiplicative FPE-penalty term $(n+h)/(n-h)$ the number of predictors $h$, which coincides with the expected value of the sum of $h$ $\chi^2(1)$-distributed random variables, by the expected value $\zeta(h,H)$ of the sum of the $h$ largest of $H$ i.i.d. $\chi^2(1)$ random variables (Reschenhofer, 2004; for related criteria see Tibshirani and Knight, 1999; George and Foster, 2000; for tables of $\zeta(h,H)$ see Reschenhofer, 2010). However, the resulting criterion $\text{FPE}_{\text{sub}}$ suffers from important shortcomings. Firstly, its usefulness is limited by the fact that the values of $\zeta(h,H)$ are not readily available in software packages and must be looked up in tables. Secondly, the penalty term $(n+\zeta(h,H))/(n-\zeta(h,H))$ may quickly become numerically unstable as $h$ and $H$ increase. Thirdly, the increase from $\zeta(h,H)$ to $\zeta(h+1,H)$ seems to be too small to prevent the inclusion of an unneeded predictor when there are $h$ dominant predictors that are certain to be included. In this case, it would be more appropriate to regard the first $h$ predictors as fixed and the next predictor as the best fitting of the remaining $H-h$ predictors rather than as the worst fitting of the best $h+1$ predictors.

Luckily, we can deal with all three issues at the same time by taking a stepwise approach (STP), according to which model dimension $h+1$ should be preferred over model dimension $h$ if

$$\frac{n+h+\zeta_1(1,H-h)}{n-h-\zeta_1(1,H-h)} \frac{\text{RSS}(h+1)}{\text{RSS}(h)} \leq \frac{n+h}{n-h}$$

(Reschenhofer et al., 2012, 2013; for a related but not stepwise approach see Foster & George, 1994). Here we need only the expected value of the maximum, which can be approximated by

$$\hat{\zeta}(1,H) = 2 \log(H) - \log(\log(H))$$

(see Reschenhofer, 2004), hence no tables are needed. Moreover, numerical problems because of small denominators occur only when $h$ is close to $n$. 

\[ \text{(3) } \]

\[ \text{(4) } \]
2.3 Using factors for prediction

A widely used method to extract \( h \) common factors from a large number of available macroeconomic and financial variables is to use the first \( h \) principal components, which are, under suitable conditions, consistent estimates of the factors (Connor and Korajczyk 1986, 1988, 1993). Forni et al. (2000) provided conditions for the identification and consistent estimation of a predetermined number of factors. Their conditions are purely technical (e.g., stationarity, boundedness of spectral densities) and are therefore of little use for improving the precision of the estimates. In contrast, the estimates discussed in 2.4 are based on concrete ideas about the shapes of the spectral densities.

Using principal components as predictors offers big advantages over using the original variables. Firstly, principal components can be ordered according to the size of the associated eigenvalues which allows us to consider nested models and to use conventional criteria such as AIC and BIC for choosing the best model. Secondly, principal components are orthogonal which makes the problem of finding the best model for each model dimension computationally tractable and therefore allows us to study also nonnested models and to choose the overall best model with the help of the stepwise procedure (3) for orthogonal predictors, which has been discussed in the previous subsection. Clearly, these advantages are purely technical and do not imply superior forecasting performance in practice.

The forecast of \( y_{n+1} \) based on a subset of principal components is given by

\[
\hat{y}_{n+1} = \sum_{k \in M} \hat{\delta}_k f_{nk},
\]

where \( M \subseteq \{1, \ldots, K\} \), \( K < H \), \( f_k = (x_1, \ldots, x_H) v_k \) is the \( k \)'th principal components with \( v_k \) denoting the eigenvector associated with the \( k \)-th largest eigenvalue of the sample covariance matrix of the (standardized) predictors \( x_1, \ldots, x_H \), and the OLS estimate \( \hat{\delta}_k \) is obtained by regressing \( y \) on \( f_k \). Usually, we consider only \( K < H \) principal components to avoid numerical problems with the smallest eigenvalues.

2.4 Adjusted band regression

In some economic applications, it may be useful to focus on certain frequency bands (e.g., the neighborhood of frequency zero when we are looking for cointegrating relationships; see Phillips, 1991) and disregard others (e.g., narrow bands around all seasonal frequencies when we are analyzing not seasonally adjusted time series). In the case of forecasting (quarterly) macroeconomic
times series, we could make use of the fact that these series are typically dominated by their low-frequency components. For that purpose, Reschenhofer and Chudy (2015) imposed the restriction

$$\bar{X}'\bar{y} = 0,$$  \hspace{1cm} (6)

on the representation

$$\hat{\beta} = (X'X)^{-1}X'y = \left((X + \bar{X})'(X + \bar{X})\right)^{-1}\left((X + \bar{X})'(\bar{y} + \bar{y})\right)$$

$$= \left(X'\bar{X} + \bar{X}'\bar{X}\right)^{-1}(X'\bar{y} + \bar{X}'\bar{y}),$$  \hspace{1cm} (7)

denotes the conventional OLS estimator, where $X$ and $y$ are obtained by projecting $X$ and $y$ onto the span of the columns of the matrix

$$G = \sqrt{\frac{2}{n}} \begin{pmatrix} \cos(\omega_1, 1) & \sin(\omega_1, 1) & \cdots & \cos(\omega_1, 1) & \sin(\omega_1, 1) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \cos(\omega_{(n-1)} & \sin(\omega_{(n-1)}) & \cdots & \cos(\omega_{(n-1)}) & \sin(\omega_{(n-1)}) \end{pmatrix},$$  \hspace{1cm} (8)

and

$$\bar{y} = y - \bar{y}, \quad \bar{X} = X - \bar{X},$$  \hspace{1cm} (9)

are the first $r_{<m}=[n/2]$ Fourier frequencies. The resulting estimator

$$\tilde{\beta} = (X'\bar{X} + \bar{X}'\bar{X})^{-1}X'\bar{y}$$  \hspace{1cm} (11)

is a compromise between the band-regression estimator

$$\hat{\beta} = (X'X)^{-1}X'y = (X'GG'X)^{-1}X'GG'y$$  \hspace{1cm} (12)

(see Hannan, 1963; Engle, 1974), which is suitable for the low-frequency components, and the trivial estimator $\beta_0 = 0$, which is suitable for the high-frequency components (provided that the latter do not contain any useful information). Accordingly, we also study forecasts of the form

$$\hat{y}_{n+1} = \sum_{k \in M} \tilde{\beta}_k x_{nk}$$  \hspace{1cm} (13)

and

$$\hat{y}_{n+1} = \sum_{k \in M} \tilde{\delta}_k f_{nk},$$  \hspace{1cm} (14)

respectively. The adjusted-band-regression estimator (11) may be regarded as a shrinkage version of the band-regression estimator (12). Clearly, the larger estimates obtained via (12) are not appropriate for (13) or (14) because they would give too much weight to the high-frequency components of the predictors.
Restriction (6) expresses our belief that there are no important linear relationships between the high-frequency components. However, we do not want to restrict the applicability of the adjusted-band-regression estimator (11) by detailed specifications of the correlation patterns and/or spectral shapes. Despite the fact that the construction of this estimator is straightforward and intuitively appealing, it would be extremely difficult to specify assumptions which allow the derivation of meaningful optimality properties. Modeling a steep decline after frequency $\omega r$ as a discontinuity is an option but requires dealing with piecewise continuous spectra, the asymptotic estimation theory of which is not yet fully developed (see, e.g., Taniguchi, 2008). We also do not try to find the smallest possible $r$ such that restriction (6) is plausible. In view of the typical shapes of univariate spectral densities and squared coherence functions of quarterly macroeconomic time series, it seems that $[0.4 m]$ is a safe choice for $r$.

3 Empirical results

3.1 Evidence against previous findings

We use the same macroeconomic dataset as Stock and Watson (2012). This dataset consists of 143 quarterly U.S. time series from 1960:II to 2008:IV and can be downloaded from Mark Watson’s website (http://www.princeton.edu/~mwatson/). Of the 34 aggregated series in the dataset, only the GDP is used in the empirical analysis. All $H=109$ disaggregated series, which include the 10 GDP components, are used as possible predictors. In order to achieve stationarity, Stock and Watson (2012) transformed the predictors by taking logarithms and/or differencing. However, they ignored possible structural breaks such as the end of the Bretton Woods system in 1971, the slowdown in growth after the oil price shock in 1973, or the decrease in volatility starting in the 1980s (Great Moderation). Clearly, the impact on the forecasting performance of a factor model depends on the magnitude of the instabilities (see, e.g., Sock and Watson, 2009; Chen et al., 2014) and, in the case of a rolling analysis, on the length of the estimation window. Also disputable is their final transformation, namely the replacement of possible outliers with the median values of the preceding five observations, which may even turn an extremely large positive/negative value into a negative/positive value. Shrinking outliers towards the current mean/median would certainly be more plausible. However, we made no effort to improve the data transformations because we did not want to compromise comparability with previous studies.

Another point of criticism of previous studies is their use of a single measure, e.g., the sum of squared prediction errors, for the assessment of the forecasting performance. More informative is a plot of the cumulative absolute or squared prediction errors, which allows a continuous assessment...
over the whole evaluation period. In the rest of this section, we will compare various forecasts of the differenced log GDP that are based on its lags, the 109 (transformed) disaggregated series and the factors estimated from them. In a rolling analysis, each subsample of \( n=100 \) successive quarters is used to partial out the autoregressive dynamics, estimate the principal components from the residuals, and compute forecasts of the form (2) using either (1), (5), (13), or (14) (see Section 2). The autoregression with four lags (AR4) is used as a benchmark.

Figure 1.A examines the performance of the OLS forecasts using the first five, the second five, the third five, … principal components. Apparently, only the forecast using the first five principal components (PC1-5) can compete with the benchmark. Looking only at the total sum of absolute prediction errors (or the total sum of squared prediction errors) we might conclude that this forecast indeed outperforms the benchmark. However, the plots of the cumulative errors show that there is practically no difference for almost the whole observation period. The PC1-5 forecast outperforms the benchmark only because of its slightly better performance in the very last quarters, which is certainly not a proof of its superiority. Moreover, this negligible outperformance depends on the inclusion of a small subset of predictors. Figure 1.B shows that the PC1-5 forecast fails to beat the benchmark when the GDP components are removed from the dataset.
Figure 1: Cumulative sums of absolute prediction errors relative to AR4 benchmark for
(A) OLS forecasts based on 1st five, 2nd five, … principal components obtained from all predictors
(B) OLS forecasts based on 1st five, 2nd five, … principal components obtained from all predictors except for the GDP components
(C) OLS forecasts based on 1st $h$ principal components ($h$ chosen by AIC, BIC, FPE$_{sub}$, STP)
(D) OLS forecasts based on $h$ best fitting principal components ($h$ chosen by AIC, BIC, FPE$_{sub}$, STP)
3.2 Selecting the predictors
Instead of using a fixed number of principal components, we might try to choose the optimum number with the help of a model selection criterion. AIC and BIC choose the first 3 or 4 principal components while $\text{FPE}_\text{sub}$ and STP most of the time select not a single one. Overall, the performance of all four criteria is very close to that of the benchmark because the corresponding four lines in Figure 1.C are roughly parallel to the x-axis. When not the $h$ first principal components (ordered according to the size of the eigenvalues) are chosen for a model of dimension $h$ but rather the $h$ best fitting of the $K=90$ first principal components, AIC and BIC are no longer suitable. In the former case (nested models), there is only one model for each $h$ whereas in the latter case (nonnested models), there are $K!/(h!(K-h)!)$ models for each $h$, from which the best fitting model is selected. The orthogonality of the principal components allows us to find the best fitting model for each model dimension $h$ just by running $K$ regressions with only a single principal component and choosing the $h$ best fitting principal components. Despite the computational simplicity of this procedure, the chosen models must still be regarded as the best of a large number of models of the same dimension, hence there is a huge danger of a data-snooping bias, which must be taken care of. Not surprisingly, AIC and BIC fail to do so and therefore always select a much too large model dimension and consequently perform very poorly. In contrast, $\text{FPE}_\text{sub}$ and STP still perform reasonably well (see Figure 1.D).

3.3 Using frequency-domain information
Figure 2.A shows that the first ten principal components are more useful than the second ten, the second ten are more useful than the third ten, … ($p=1/5!<0.01$). Thus it seems that, in general, the leading principal components are more informative than the ones following behind, a possible explanation of which is provided by Figure 3. Since quarterly macroeconomic series as well as the relationships between them are typically dominated by their low frequency components, we may expect that the first principal components in their effort to explain as much variation as possible focus primarily on the lower frequencies while the other principal components must deal with the rest. Indeed, the periodograms of the first and second principal component have a peak close to frequency zero (see Figures 3.A and 3.B) while the periodograms of the 89th and 90th principal component are featureless and resemble periodograms obtained from white noise (see Figures 3.C and 3.D). The importance of the lower frequencies is also corroborated by the fact that the use of adjusted band regression (see Section 2.4) instead of OLS yields a further improvement (compare each line in Figure 2.A with its counterpart in Figure 2.B; $p=0.55^5<0.05$). The tuning parameter $r$ was set so that just the lowest 40% of the Fourier frequencies are used in the calculations.
Perhaps the most compelling evidence against the use of many predictors is obtained when we focus just on a small number of predictors which are closest to the variable of interest. Retaining only the ten GDP components and omitting all the other predictors, we obtain a forecast that clearly outperforms the benchmark as well as all factor-augmented forecasts (see Figure 2.C). Despite the small number of predictors, we may still switch to factors/principal components in order to benefit from their orthogonality properties and further reduce the model dimension without compromising the precision of the forecast. It turns out that the use of only three principal components instead of all ten GDP components improves the forecasting performance (see Figure 2.C). A further improvement can be achieved by using adjusted band-regression instead of OLS (see Figure 2.D).

4 Discussion
Reanalyzing the macroeconomic dataset used by Stock and Watson (2012), we have not found any evidence corroborating their claim that the standard factor-augmented forecast, which is based on lagged values of the variable of interest and a small number of factors, outperforms a simple autoregressive benchmark for macroeconomic variables like the GDP. In combination with their disappointing results obtained by retaining a large number of factors and applying a suitable shrinkage procedure, this outcome questions the usefulness of factor models altogether. Thus, a key take-away from our study is that the standard factor-augmented forecast should not be used as a benchmark in future studies exploring the performance of more sophisticated factor models.

The standard factor-augmented forecast appears to outperform the benchmark only when we use a single measure like the sum of absolute (or squared) prediction errors for the evaluation of the forecasting performance. Looking at the cumulative prediction errors, we see that the difference is negligible. Also, automatic criteria best suited for the selection of a subset of predictors refuse to include any factors. Moreover, the marginal outperformance of the standard factor-augmented forecast turns into an underperformance when a small number of predictors are removed from the dataset.

Apart from criticizing the use of factor models for forecasting, we have also tried to improve their performance by focusing on the low-frequency components and disregarding the high-frequency components. Technically, this can be achieved by dismissing OLS regression in favor of adjusted band-regression, which can be regarded as a shrinkage version of band regression. Although the results obtained in this way are promising, they have still to be corroborated by a future study using a different and/or newer dataset.
Figure 2: Cumulative sums of absolute prediction errors relative to AR4 benchmark for:

(A) OLS forecasts based on 1\textsuperscript{st} ten, 2\textsuperscript{nd} ten, … principal components

(B) 40\% adjusted band regression forecasts based on 1\textsuperscript{st} ten, 2\textsuperscript{nd} ten, … principal components

(C) OLS forecasts based on best subsets of principal components obtained from GDP components

(D) 40\% adjusted band-regression forecasts based on best subsets of principal components obtained from GDP components
Forecasting with dynamic factor models

Figure 3: Periodograms of the $k$'th principal component for (A) $k=1$, (B) $k=2$, (C) $k=89$, (D) $k=90$
References


Rothman D (1968) Letter to the editor. Technometrics 10:432


