DISCUSSION PAPER SERIES

No. 8867

FINITE SAMPLE PERFORMANCE OF SMALL VERSUS LARGE SCALE DYNAMIC FACTOR MODELS

Rocio Alvarez, Maximo Camacho
and Gabriel Pérez-Quirós

INTERNATIONAL MACROECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP8867.asp**

ISSN 0265-8003

FINITE SAMPLE PERFORMANCE OF SMALL VERSUS LARGE SCALE DYNAMIC FACTOR MODELS

Rocio Alvarez, Universidad de Alicante
Maximo Camacho, Universidad de Murcia
Gabriel Pérez-Quirós, Banco de España and CEPR

Discussion Paper No. 8867
February 2012

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **INTERNATIONAL MACROECONOMICS**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Rocio Alvarez, Maximo Camacho and Gabriel Pérez-Quirós

CEPR Discussion Paper No. 8867

February 2012

ABSTRACT

Finite sample performance of small versus large scale dynamic factor models*

We examine the finite-sample performance of small versus large scale dynamic factor models. Our Monte Carlo analysis reveals that small scale factor models out-perform large scale models in factor estimation and forecasting for high levels of cross-correlation across the idiosyncratic errors of series belonging to the same category, for oversampled categories and, especially, for high persistence in either the common factor series or the idiosyncratic errors. Using a panel of 147 US economic indicators, which are classified into 13 economic categories, we show that a small scale dynamic factor model that uses one representative indicator of each category yields satisfactory or even better forecasting results than a large scale dynamic factor model that uses all the economic indicator

JEL Classification: C22, E27, E32
Keywords: business cycles, output growth, time series

Rocio Alvarez
Universidad de Alicante
Carretera de San Vicente del Raspeig S/N, 03690
San Vicente del Raspeig
SPAIN

Email: rocio@merlin.fae.ua.es

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=175060

Maximo Camacho
Universidad de Murcia
Departamento de Metodos Cuantitativos
Facultad de Economia y Empresa
30100 Murcia
SPAIN

Email: mcamacho@um.es

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=160538

Gabriel Pérez-Quirós
Banco de Espana
Servicio de Estudios
Apartado de Correo, 15
Calle Alcala 50
Madrid, 28014
SPAIN

Email: gabriel.perez@bde.es

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=135792

* Camacho acknowledges financial support from MICINN (ECO2010-19830) and Fundacion Ramon Areces for financial support. Any errors are our responsibility. The views in this paper are those of the authors and do not represent the views of the Bank of Spain or the Eurosystem.

Submitted 13 February 2012

1 Introduction

Aruoba, Diebold and Scotti (2009) considered that comparative assessments of forecasts from "small data" versus "big data" dynamic factor models was a good place to develop further empirical analyses for the same economy and time period. On the one hand, small data forecasts have been computed from different enlargements of the Stock and Watson (1991) single-index small scale dynamic factor model ($SSDFM$). Recent examples of are Mariano and Murasawa (2003), Nunes (2005), Aruoba, Dieblod and Scotti (2009), Aruoba and Dieblod (2010), and Camacho and Perez Quiros (2010). The Philadelphia Fed business conditions index is also contructed using this approach. In these studies, the *strict* factor models are estimated by maximum likelihood using the Kalman filter under the assumption of having non cross-correlated idiosyncratic errors.

On the other hand, big data forecasts have been computed from various sophistications of the seminal Stock and Watson (2002a) principal components estimator, which combine the information of many predictors. Recent examples of forecasts from the so-called large scale dynamic factor models ($LSDFM$) are Forni, Hallin, Lippi and Reichlin (2005), Giannone, Reichlin and Small (2008), and Angelini et al. (2011). The Chicago Fed National Activity Index (CFNAI) is also developed under this approach.[1] The *approximate* factor models suggested in these papers lead to asymptotically consistent estimates when the number of variables and observations tends to infinity, under the assumptions of weak cross-correlation of the idiosyncratic components and that the variability of the common component is not too small.

Relatively, much more theoretical attention has recently been devoted to large scale factor models by stressing that strict factor models rely on the tight assumption that the idiosyncratic components are cross-sectionally orthogonal. However, including time series in empirical applications to compute factors from large panels frequently supposes facing non-negligible costs as well. According to Boivin and Ng (2006), the large data sets used by $LSDFM$ are typically drawn in practice from a small number of broad categories (such as industrial production, or monetary and price indicators). Since the idiosyncratic errors

[1] Notably, in November 2011 the real-time monthly average of the Philadelphia Fed index was positive (0.01) while the Chicago Fed index was negative (-0.37).

of time series belonging to a particular category are expected to be highly correlated, the assumption of weak correlation among the idiosyncratic components is more likely to fail as the number of time series of this category increases. In addition, the good asymptotic properties suggested by the theory may not hold in many empirical applications when the number of variables and observations is relatively small.[2]

The impact of this potential clash between the asymptotically good properties of $LSDFM$ suggested by the theory and their actual forecasting performance obtained in empirical applications has rarely been addressed. Among the exceptions, Stock and Watson (2002b) find deterioration in the performance of large scale (static) factor models when the degree of serial correlation and (to a lesser extent) heteroskedasticity among idiosyncratic errors are large and when serial correlation of factors is high. Boivin and Ng (2006) use large scale (static) factor models to show that including series that are highly correlated with those of the same category does not necessarily mean outperforming models that exclude these series. Boivin and Ng (2006) for the US and Caggiano, Kapetanios, and Labhard (2009) for some Euro area countries estimate large scale (static) factor models of different dimensions to show that factors extracted from pre-screened series often yield satisfactory or even better results than using larger sets of series. Notably, their preferred data sets sometimes include one-fifth of the original set of indicators. Bai and Ng (2008) find improvements over a baseline large scale (static) factor model by estimating the factors using fewer, but more informative, predictors. Banbura and Runstler (2011) use a large scale (dynamic) model to show that forecast weights are concentrated among a relatively small set of Euro area indicators. Finally, Banbura and Mondugno (2010) find that a $LSDFM$ applied to a small (14 series) dataset outperforms the forecasts obtained from medium (46 series) and large (101 series) datasets.

Of all these works, the one closest to our approach is that of Boivin and Ng (2006), although we differ from these authors in many aspects. First, our purpose is not to determine the optimal number of variables from a large dataset to be used in a large

[2] Recently, Boivin and Ng (2006) for US and Banbura and Runstler (2011) for the Euro area have shown that the predictive content of empirical large scale factor models is contained in the factors extracted from *as few as* about 40 series.

scale factor model. In contrast, we try to shed some light on the issue of which is the optimal strategy when dealing with a forecasting problem, either to start from a simple small scale factor model that reasonably selects the indicators (and which is enlarged if necessary) or to deal with a large scale factor model whose dimension can be selectively reduced to eliminate the redundant information.[3] Second, Boivin and Ng (2006) consider static models, while we compare dynamic specifications. In particular, we consider the large scale dynamic factor model of Giannone, Reichlin and Small (2008) while they use the large scale static factor model of Stock and Watson (2002a). The use of dynamic instead of static factor models is an important distinctive feature of our analysis since we address to what extent persistence in the factors and in the idiosyncratic shocks may affect the accuracy of our different factor model specifications. Third, we assess in depth the effects on factor models of using time series which are extracted from separate groups of macroeconomic indicators. Boivin and Ng (2006) use the word "categories" to refer to different sectors in the economy (prices, production, etc..) but they classify the data according to their correlation or their heteroskedastic behavior. In contrast, we concentrate on assessing the effects on the estimation of the factors and forecasting of dealing with data which are extracted from separate sectors. In addition, we examine the effects of dealing with cross-correlation across sectors and within each sector.

Within this context, our paper develops simulations in which we try to mimic different empirical forecasting scenarios. The first scenario is the case in which an analyst uses $SSDFM$ to estimate the factors and to compute the forecasts from a small number of pre-screened series which are the main (less noisy) indicators of the different categories of data. In the second scenario, the analysis is developed from a $SSDFM$ which uses a less accurate pre-screening set of indicators that includes the series exhibiting the highest averaged correlation with respect to the other series included in the same category. In the final scenario, the analysis is conducted with a $LSDFM$ that uses a large scale data set generated by including additional series in each category under the assumption that

[3] The $LSDFM$ requires a sufficiently large number of time series to achieve its statistical properties. In this sense, a $SSDFM$ cannot be viewed as a particular case of a $LSDFM$ but as a different estimation strategy.

the additional series are finer disaggregations of the main indicator with which they are correlated.

Using averaged squared errors, we propose a Monte Carlo analysis to evaluate the accuracy of these three forecasting proposals to estimate the factors and to compute out-of-sample forecasts of a target variable. We find that adding indicators that bear little information about the factor components does not necessarily lead $LSDFM$ to improve upon the forecasts of $SSDFM$. In fact, we show that when the additional time series are too correlated with the indicators already included in some categories, forecasting with many predictors performs worse than forecasting from a reasonably pre-screened dataset, especially when the categories are not highly correlated. In addition, $SSDFM$ outperform $LSDFM$ in factor estimation and forecasting for high levels of cross-correlation across the idiosyncratic errors of series from the same category, for oversampled categories and, especially, for high persistence in either the common factor series or the idiosyncratic error.

The comparative performance of small versus large scale dynamic factor models is examined using the set of 147 US monthly macroeconomic indicators suggested by Stock and Watson (2002b). The time series included in the dataset are classified by these authors into 13 economic categories such as real output, prices, and employment. In an out-of-sample exercise, we examine the accuracy of a large scale dynamic factor model that uses the 147 indicators versus a small scale dynamic factor models that uses one representative of each category to forecast the Industrial Production Index (IPI) at different short-term horizons. The empirical results obtained from actual data are in concordance with those obtained from generated data. A $SSDFM$ that uses the 13 time series exhibiting the highest averaged correlation with respect to the series of the same category yields satisfactory or even better forecasting results than a $LSDFM$ that uses the 147 economic indicators.

This paper proceeds as follows. Section 2 describes both small and large scale dynamic factor models. Section 3 presents the design details of the simulation exercise, i.e. how to generate the main series of each category and the finer disaggregations. Section 4 shows the main findings in the comparison between $SSDFM$ and $LSDFM$ for different parameter values. Section 5 describes the main results of our empirical application. Section

6 concludes.

2 Dynamic factor models

Large and small scale factor models can be represented in a similar general framework. Let y_t be a scalar time series variable to be forecasted and let $X_t = (X_{1t}, ..., X_{Nt})'$, with $t = 1, ..., T$, be the observed stationary time series which are candidate predictors of y_t. If we are interested in one-step-ahead predictions, the baseline model can be stated as

$$y_{t+1} = \alpha_0 + \alpha' X_t + \sum_{j=1}^{p} \gamma_j y_{t-j+1} + \epsilon_{yt+1}, \tag{1}$$

where $\alpha = (\alpha_1, ..., \alpha_N)'$, and ϵ_{yt+1} is a zero mean white noise.

Since estimating this expression becomes impractical as the number of predictors increases, it is standard to assume that each predictor X_{it} has zero mean and admits a factor structure:

$$X_{it} = \lambda_i' F_t + \xi_{it}, \tag{2}$$

for the ith cross-section unit at time t, $i = 1, ..., N$, $\lambda_i = (\lambda_{i1}, ..., \lambda_{ir})'$, and $t = 1, ..., T$. In this framework, the $r \times 1$ vector F_t contains the r common factors, λ_i the r factor loadings, $\chi_{it} = \lambda_i' F_t$ the common components, and ξ_{it} the idiosyncratic errors. In vector notation the model can be written as

$$X_t = \Lambda F_t + \xi_t, \tag{3}$$

where $\Lambda = (\lambda_{ij})$ is the $N \times r$ matrix of factor loadings and ξ_t is the vector of N idiosyncratic shocks. In the related literature, it is standard to assume that the vectors F_t and ξ_t are serially and cross-sectionally uncorrelated unobserved stationary processes.[4] In contrast to static factor models, the dynamics of the common factors are supposed to follow autoregressive processes. Although it is very easy to generalize, let us assume that the factors follow a simple $VAR(1)$ process

$$F_t = AF_{t-1} + u_t, \tag{4}$$

[4] In this framework the common factor is supposed to generate most of the cross-correlation between the series of the data set $\{X_{it}\}_{i=1}^{N}$.

where A is the $r \times r$ matrix of coefficients, with $E[u_t] = 0$ and $E[u_t u_t'] = \Sigma_u$. In addition, ξ_t is also assumed to follow a simple stationary $VAR(1)$ process with mean zero:

$$\xi_t = C\xi_{t-1} + v_t, \tag{5}$$

where v_t is serially uncorrelated with $E[v_t] = 0$ and $E[v_t' v_t] = \Sigma_v$.[5] Then, the target variable y_t can be forecasted through the common factors by

$$y_{t+1} = \beta_0 + \beta' F_t + \sum_{j=1}^{p} \gamma_j y_{t-j+1} + e_{yt+1}. \tag{6}$$

Finally, let us call the model a small scale dynamic factor model ($SSDFM$) when N is fixed and small and T is large, and a large scale dynamic factor model ($LSDFM$) when both N and T are large. In addition, although we leave the data to select the number of factors in the empirical exercise, let us focus the analysis in the case where there is only one factor.

2.1 Small scale dynamic factor models

The baseline model is the single-index dynamic factor model of Stock and Watson (1991) which can be written in state-space form. Accordingly, the autoregressive parameter A, the vector of the N loading factors Λ, and the $(N \times N)$ covariance matrix of the idiosyncratic shocks Σ_v, can be estimated by maximum likelihood via the Kalman filter.[6] Let h_t be the $(N+1)$ vector $h_t = (F_{t,}' \xi_t')'$, I_j be the identity matrix of dimension j, and 0_j be the vector of j zeroes. Hence, the measurement equation can be defined as

$$X_t = H h_t + e_t, \tag{7}$$

where

$$H = \begin{pmatrix} \Lambda & I_N \end{pmatrix}, \tag{8}$$

and e_t is a vector of N zeroes. In addition, the transition equation can be stated as

$$h_{t+1} = F h_t + w_t, \tag{9}$$

[5] Although assuming $VAR(p)$ dynamics for the factors and the idiosyncractic components is straightforward, it would complicates notation.

[6] For identification purposes, Σ_u is usually assumed to be one.

where the $(N+1 \times N+1)$ matrix F is

$$F = \begin{pmatrix} A & 0_N' \\ 0_N & C \end{pmatrix}, \tag{10}$$

and $w_t = (u_t, v_t')$ with zero mean and covariance matrix

$$Q = \begin{pmatrix} \Sigma_u & 0 \\ 0 & \Sigma_v \end{pmatrix}. \tag{11}$$

In the standard way, the Kalman filter also produces filtered and smoothed inferences of the common factor: $\{F^s_{t|t}\}_{t=1}^T$ and $\{F^s_{t|T}\}_{t=1}^T$. These inferences can be used in the prediction equation (6) to compute OLS forecasts of the variable y_{t+1}.

2.2 Large scale dynamic factor models

To estimate the factors in the large scale framework, we use the quasi-maximum likelihood approach suggested by Doz, Giannone and Reichlin (2007). In this method, the estimates of the parameters are obtained by maximizing the likelihood via the EM algorithm, which consists of an iterative two-step estimator. In the first step, the algorithm computes an estimate of the parameters given an estimate of the common factor. In the second step, the algorithm uses the estimated parameters to approximate the common factor by the Kalman smoother. At each iteration, the algorithm ensures higher values of the log-likelihood of the estimated common factor, so it is assumed that the process converges when the slope between two consecutive log-likelihood values is lower than a threshold.[7]

Using an initial set of time series $\{X_{it}\}_{i=1}^N$, the $(i+1)$-th iteration of the algorithm is defined as follows. Let us assume that Λ^i, A^i and Σ_x^i are known. Let F_t^i be the common factor which is the output of the Kalman filter from the i-st iteration. The updated estimates of Λ, A, and Σ_x can be obtained from

$$\Lambda^{i+1} = E[\widehat{X_t F_t^{i\prime}}](E[\widehat{F_t^i F_t^{i\prime}}])^{-1}, \tag{12}$$

$$A^{i+1} = E[\widehat{F_t^i F_{t-1}^{i\prime}}](E[\widehat{F_{t-1}^i F_{t-1}^{i\prime}}])^{-1}, \tag{13}$$

$$\Sigma_x^{i+1} = E[\widehat{X_t^i \xi_t^{i\prime}}]. \tag{14}$$

[7] In practice, we consider a threshold of 10^{-4}.

The estimates of the expectations can be obtained from

$$E[\widehat{X_t F_t'}] = \frac{1}{T}\sum_{t=1}^{T} X_t F_t^{i\prime}, \tag{15}$$

where the series $\{F_t^i\}_{t=1}^T$ is the factor estimated at the iteration i. In addition, since $E[F_t F_t'] = E[F_t F_t^{i\prime}] + E[\{F_t - F_t^{i\prime}\}\{F_t - F_t^{i\prime}\}']$, and $E[\{F_t - F_t^{i\prime}\}\{F_t - F_t^{i\prime}\}']$ is the variance of the estimated common factor, then denoting the variances by $\{V_t\}_{t=1}^T$, the expectation $E[F_t F_t']$ can be estimated by

$$E[\widehat{F_t^i F_t^{i\prime}}] = \frac{1}{T}\sum_{t=1}^{T}(F_t^i F_t^{i\prime} + V_t). \tag{16}$$

Following a similar reasoning, $E[F_t F_{t-1}'] = E[F_t F_{t-1}^{i\prime}] + E[\{F_t - F_t^{i\prime}\}\{F_{t-1} - F_{t-1}^{i\prime}\}']$, and the last expectation which we denote as $\{C_t\}_{t=2}^T$ can be estimated by the Kalman filter. Then, the expectation $E[F_t F_{t-1}']$ can be estimated by

$$E[\widehat{F_t^i F_{t-1}^{i\prime}}] = \frac{1}{T}\sum_{t=1}^{T}(F_t^i F_{t-1}^{i\prime} + C_t). \tag{17}$$

The matrix Σ_v is estimated as the diagonal matrix whose principal diagonal is given by:

$$\hat{\Sigma}_x = diag(\frac{1}{T}\sum_{t=1}^{T} X_t(X_t - \Lambda^i F_t^i)'). \tag{18}$$

These estimates can be used again in the Kalman filter to compute the factors F_t^{i+1}. The algorithm, which starts with the static principal components estimates of the common factors F_t^0 and their factor loadings Λ^0, is repeated until the quasi-maximum likelihood estimates of the parameters are obtained. These can easily be used to compute the estimates of the common factor $\{F_{t|T}\}_{t=1}^T$ using the Kalman smoother, treating the idiosyncratic errors as uncorrelated both in time and in the cross section.[8] Finally, as in the case of $SSDFM$, the forecasts of y_{t+1} are estimated by OLS regressions on (6).

3 Designing the simulation study

According to the estimation of the dynamic factor models described in the previous section, the empirical applications that use these factor models will perform worse than

[8] The algorithm requires small number of iterations to converge. In our simulations, we only required 3 or 4 iterations to converge.

theoretically expected when facing data problems that invalidate the assumptions seemingly guaranteed by the theory. In the case of $SSDFM$, the larger the covariance among idiosyncratic errors, the less accurate the estimations are expected to be. With respect to the empirical performance of $LSDFM$, the models' accuracy deteriorates when the average size of the common component falls, when the number of observations is not large on either the cross-section or on the time dimensions, and when the possibility of correlated errors increases as more series are included in the model, which is very common in practice since the data are usually drawn from a small number of broad categories.[9] In this section, we perform Monte Carlo simulations to assess the extent to which the violation of the theoretical assumptions behind $SSDFM$ and $LSDFM$ affects both the consistency of factor estimation and the accuracy of forecasts.

3.1 Forecasting scenarios

The first scenario mimics the case in which forecasters develop a reasonable pre-screening of the set of potential indicators and apply $SSDFM$ to obtain predictions from a reduced number of selected indicators. In particular, we assume that the analyst searches for the representative indicators of each economic category by screening out the noisier time series of each category. However, the analyst usually does not know which are the least noisy indicators from each category and some noisy indicators can be erroneously included to compute the forecasts. To evaluate the effects of forecasting from a less accurate pre-screened set of indicators, we also consider the forecasting scenario of computing $SSDFM$ forecasts from a small number of noisier indicators which are the series of each category that exhibits the highest average correlation with the other series included in the same category. In this case, we assume different degrees of correlation across representative series of different categories.

The second forecasting scenario mimics the case of forecasters who include a large number of indicators and apply $LSDFM$ to compute predictions. In this case, the analyst

[9] Moench, Ng and Potter (2009) develop an interesting analysis by using dynamic hierarchical factor models. The comparison between these type of models and the traditional large and small scale models used in this paper is left for further research.

does not carry out any pre-screening of the initial set of indicators which are also assumed to belong to a reduced set of different categories. In addition, the indicators that belong to each category are assumed to exhibit different degrees of correlation with the representative indicators of these categories.

3.2 Generating small data sets

To simplify the analysis, we assume that the small data set, $\{X_{it}^s\}_{i,t=1}^{N,T}$, with $N = 10$, is generated from one common factor only. First, given the parameters A and Σ_u, we generate the series of the common factor $\{F_t\}_{t=1}^T$ by

$$F_t = AF_{t-1} + u_t. \tag{19}$$

In the empirical applications, F_t usually represents the "state of the economy" or the "business cycle". In this case, $\{u_t\}_{t=1}^T$ are random numbers which are drawn from a normal distribution with zero mean and variances $\Sigma_u = 1$. To examine the dependence of the results on the persistence of the factor, we allow for different values for the parameter $A = 0.1$, 0.5, and 0.75.

Second, we assume that the idiosyncratic errors follow autoregressive processes. For particular values of the coefficient matrix C, and Σ_v, we generate the series $\xi_t = (\xi_{1t}, ..., \xi_{Nt})'$, from

$$\xi_t = C\xi_{t-1} + v_t. \tag{20}$$

In this case, $v_t = (v_{1t}, ..., v_{Nt})'$, and $\{v_{it}\}_{i,t=1}^{N,T}$ are random numbers which are drawn from a normal distribution with zero mean and variance-covariances matrix Σ_v. To simplify simulations, the autoregressive coefficients matrix C will be diagonal with two possible values $c = 0.1$ and $c = 0.75$ in all the elements of the main diagonal. In addition, to examine the effects of the errors cross-correlation, the covariance matrix will take different values across the simulations. In particular, let us consider a given value for the parameter ρ_s and generate the vector $\overrightarrow{\rho}_s = (1, \rho_s, \rho_s^2, ..., \rho_s^9)'$. Then, the matrix Σ_v can be viewed as

the Toeplitz matrix constructed from vector $\overrightarrow{\rho}_s$ as

$$\Sigma_v = \begin{pmatrix} 1 & \rho_s & \rho_s^2 & \dots & \rho_s^9 \\ \rho_s & 1 & \rho_s & \dots & \rho_s^8 \\ \rho_s^2 & \rho_s & 1 & \dots & \rho_s^7 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_s^9 & \rho_s^8 & \rho_s^7 & \dots & 1 \end{pmatrix}. \tag{21}$$

As can be deduced from this expression, parameter ρ_s represents the maximum correlation between the error terms of two series and controls the correlation across categories of data. In the simulations, the values of this parameter will be $\rho_s = 0, 0.1, 0.5$, and 0.75.

Finally, in the simulations Λ will be a column vector of N ones. Then, $\{F_t\}_{t=1}^T$, and $\{\xi_t\}_{t=1}^T$ is used in

$$X_t^S = \Lambda F_t + \xi_t, \tag{22}$$

to obtain simulations of X_t^S, with $X_t^S = \{X_{it}^s\}_{t=1}^T$, for $i = 1, ..., 10$.

Therefore, each of the ten series X_{it}^S included in X_t^S could be intuitively interpreted as ten economic sectors which depend on two components. The first component, F_t, is common to the ten categories and is usually interpreted as the business cycle, and exhibits different levels of persistence which are measured by A. The second component, ξ_{it}, refers to sectorial or idiosyncratic components which also have different levels of persistence (measured by c) and across-categories cross-correlation (measured by ρ_s).[10]

3.3 Generating large data sets

As mentioned above, for the large data set $\{X_{jt}^l\}_{j,t=1}^{M,T}$, with $M = 100$, we assume that the ten series generated in the previous section, X_t^S, represent the main indicators of each of ten different categories of data. Accordingly, we add an error term representing the idiosyncratic error of the specific series of each category to each of the ten time series $\{X_{it}^s\}_{i,t=1}^{N,T}$ for $N = 10$. These errors are called $\{w_{ikt}\}_{i,k,t=1}^{10,10,T}$ where i represents the category,

[10] For simplicity and clarity in the exposition, we present our main results with only one factor. Considering more than one factor is trivial but, although the results are of the same nature, the computation time for the simulations increases dramatically. Nevertheless, we address the possibility of estimating more than one factor in Section 4.

and k represents the series within the category. These errors are assumed to be serially correlated and cross-correlated with all the series existing within their respective category. Hence, the large data set is generated by

$$X_{ikt}^{l} = X_{it}^{s} + w_{ikt}, \tag{23}$$

where $i = 1, ..., 10$, $k = 1, ..., 10$, and $w_{it} = (w_{i1t}, ..., w_{i10t})'$ is the vector of idiosyncratic errors which is generated by

$$w_{it} = Dw_{it-1} + e_{it}^{l}. \tag{24}$$

In this expression, $\{e_{ikt}^{l}\}_{i,k,t=1}^{10,10,T}$ are random numbers drawn from a normal distribution with zero mean and covariance matrix Σ_w which is the Toeplitz matrix constructed from vector $\overrightarrow{\rho}_l$ as in (21), where $\rho_l = 0, 0.1, 0.5$, and 0.75. Therefore, parameter ρ_l controls the correlation within each of the categories of data. The autoregressive coefficients matrix D is diagonal with constant values of $d = 0.1$ and $d = 0.75$ in the main diagonal.

According to expressions (22), (23), and (24), each series of the large data set can be decomposed as follows

$$X_{ikt}^{l} = \lambda_i F_t + \xi_{ikt}^{l}, \tag{25}$$

where $\xi_{ikt}^{l} = \xi_{it} + w_{ikt}$. Then, the idiosyncratic components ξ_{ikt}^{l} are composed of a common error inside the categories, ξ_{it}, which could be cross-correlated among different categories, and a specific error term, w_{ikt}, which could be correlated with series from the same category. Finally, putting together the series along all the categories, we have the large data set

$$X_t^{l} = \left(X_{1,1,t}^{l}, X_{1,2,t}^{l}, ..., X_{1,10,t}^{l}, X_{2,1,t}^{l}, X_{2,2,t}^{l}, ..., X_{2,10,t}^{l}, ..., X_{10,1,t}^{l}, X_{10,2,t}^{l}, ..., X_{10,10,t}^{l}\right)'. \tag{26}$$

As in the case of small data sets, the generated time series can be interpreted as economic indicators that have been generated as the sum of two components: the common factor, F_t, and the idiosyncratic component, ξ_{ikt}^{l}. However, in the case of large data sets the time series also depend on the within-category cross-correlation (measured by ρ_l) and by the within category autocorrelation (measured by d).

3.4 Generating the target series

Finally, we generate the series to be predicted in a simple scenario. To simplify simulations, we consider that forecasting with factors and one lagged value of the time series is dynamically complete. Hence, the series y_t is generated from the following factor-augmented regression

$$y_{t+1} = \beta' F_t + \gamma y_t + e_{yt}, \tag{27}$$

where β is one, e_{yt} is a white noise process, with $\sigma_{e_y} = 1$. Parameter γ, which measures the autocorrelation of the target series, is assumed to take the values of 0, 0.3, 0.5 and 0.8.

4 Simulation results

In each replication, j, we estimate the small and large scale factor models and compute the accuracy of these models to infer the factor by using the Mean Squared Error over the $J = 1000$ replications

$$MSE^i = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{T}\sum_{t=1}^{T}(F_{jt} - QF^i_{jt|T})^2, \tag{28}$$

for $i = s$ in the case of the small data set and $i = l$ in the case of the large data set. In this expression, Q is the projection matrix of the true common factor on the estimated common factor.[11] In addition, we compare the out of sample forecasting accuracy of $SSDFM$ and $LSDFM$ by computing the errors in forecasting one step ahead the target series generated. Let $\widehat{\beta}$ and $\widehat{\gamma}$ be the OLS estimates of the parameters given by equation (27) using the common factor series and the values of y up to period T. Then, we construct the one-step-ahead forecast of y_{jT+1} by using the relation $\widehat{y}^i_{jT+1} = \widehat{\beta}F^i_{jt|T} + \widehat{\gamma}y_{jT}$. In this way, one can define the Mean Squared one-step-ahead Forecast Errors of model i as

$$MSFE^i = \frac{1}{J}\sum_{j=1}^{J}(y_{jT+1} - \widehat{y}^i_{jT+1})^2. \tag{29}$$

According to the forecasting scenarios described above, we call MSE^s_p, MSE^s_r, MSE^l, $MSFE^s_p$, $MSFE^s_r$, and $MSFE^l$ the mean across replications of the MSE and $MSFE$

[11] We need the projection matrix since the common factors are estimated up to a signal transformation.

which are computed from a $SSDFM$ that uses the 10 pre-screened (least noisy) series of each category (superscript s, subscript p), from a $SSDFM$ that uses 10 representative (highly correlated) series of each category (superscript s, subscript r), and from a $LSDFM$ that uses the 100 time series of the large scale simulation exercise (superscript l).

4.1 Factor estimates

Let us start the analysis of the simulations by using $MSEs$ to examine the relative accuracy of the models to infer the factors. For easier facilitate understanding, we describe how the results are presented in the tables. First, the results in Tables 1 to 3 are classified according to different values of the autoregressive coefficient of the common factor (coefficient A). This coefficient takes the value of 0.1 (low correlation) in Table 1, the value of 0.5 (medium correlation) in Table 2 and the value of 0.75 (high correlation) in Table 3. Second, each of these tables shows the accuracy of the models for different values of the cross correlation within (measured by ρ_l) and across (measured by ρ_s) categories. The first block of results refers to the case when the only cross-correlation presented in the idiosyncratic components is due to series that belong to the same category, which occurs when $\rho_s = 0$, while the following blocks of results examine the effects of progressively increasing the correlation across categories to 0.1, 0.5 and 0.75. Within each of these blocks, the tables report the models' accuracy in inferring the common factor when the correlation within categories, which is measured by ρ_l, increases from 0 to 0.1, 0.5 and 0.9. Third, the first three columns of the tables refer, respectively, to MSEs from dynamic factor models which use the set of ten less noisy indicators in a $SSDFM$ (results labelled as MSE_p^s), or the set of ten series that exhibit the highest correlation within each category in a $SSDFM$ (results labelled as MSE_r^s), or the complete set of 100 indicators in a $LSDFM$ (results labelled as MSE^l), respectively. Fourth, it is a common practice in large scale factor models to represent each category by different numbers of time series and frequently some categories might be over represented.[12] We address the effects of over sampling in the last

[12] Typically, the number of series of disaggregated industrial production indicators is somewhat higher than the number of time series included in other categories. Significant examples are Stock and Watson (2002a, 2002b), Giannone et al. (2008), and Angelini et al. (2011).

two columns of these tables. For this purpose, we simulate ten categories of data but with 20 series instead of 10 in the first category, 5 series instead of 10 in the second and third categories, and, as before, the 10 series of the other 7 categories.[13] Fifth, in Tables 1 to 3, we assume that the idiosyncratic components and the within categories errors have low serial correlation (values of $c = d = 0.1$), that the sample is small ($T = 50$), and that there is only one common factor in the estimation.[14] The robustness of the results when for higher serial correlation in errors, larger samples, and numbers of common factors selected as in Bai and Ng (2002), are analyzed in Tables A1 to A6 in the Appendix.

A small summary of the main results follows. Overall, all the tables show that the reasonably pre-screened $SSDFM$ that uses the less noisy indicators presents smaller MSE than all the other specifications ($MSE_p^s < MSE_r^s$ and $MSE_p^s < MSE^l$). This is an important result since it implies that a good preselection in the categories is very difficult to beat even if the alternatives use a lot of information from a big number of times series. This result holds for all the possible assumptions about the dynamics of the shocks, about the dynamics of the factors, about the presence of within categories correlations, and to a lesser extent about the across categories correlations. The reasonably pre-screened $SSDFM$ is only beaten when the correlation across categories is extremely high and all the other dynamic problems, such as persistence in the factor, persistence in shocks, or within categories correlation do not appear in the analysis.

Notably, the tables also show that even in the case in which the pre-screened less noisy series are not available, there are still valuable gains when the variables are pre-selected to estimate a $SSDFM$ with those series of each category that exhibit the highest correlation with the of the same category. The relative performance of $SSDFM$ with the representative highly correlated series and the $LSDFM$, show that the former improves upon the latter ($MSE_r^s < MSE^l$) when the persistence of the factor and the within categories correlation increase.

[13] The accuracy of $SSDFM$ from the least noisy indicators does not depend on the number of series that are included in each category since the model only uses the common component of each category. Hence, the tables only show MSE_r^s.and MSE^l.

[14] In their simulations, Stock and Watson (2002b) consider that T is large when it is greater than 100, that T is small when it is smaller than 50, and that T is very small when it is equal to 25.

These results are in line with some recent findings proposed in the related literature. First, our findings are in line with those of Stock and Watson (2002b). Using large scale static factor models, these authors find some deterioration on the quality of the factor estimates when the degree of serial correlation in the factor and in the idiosyncratic errors is high, even when the number of variables and observations is large. This coincides with the finding that we show in Tables 2 and 3, which report the results of increasing inertia in the simulated common factor, with A ranging from 0.1 (almost no serial correlation) in Table 1 to 0.5 (moderate correlation) in Table 2 and to 0.75 (high correlation) in Table 3. Although our results confirm the deterioration in factor estimation of all the factor models, the relative losses are not uniformly distributed across the models. When the serial correlation of the factor increases, the relative gains of pre-screening over representative series in $SSDFM$ still hold at similar rates, except for the case of very large correlation across categories, where the relative gains attenuate. Notably, the MSEs also highlight the significant losses in the relative accuracy of $LSDFM$ with respect to $SSDFM$ as the inertia of the common factor increases. In fact, when $A = 0.75$ the $SSDFM$ from the representative (highly correlated) series of each category outperforms $LSDFM$ in all scenarios.

Second, our results are in concordance with those of Boivin and Ng (2006) who suggest that the large scale (static) factor estimates are adversely affected by cross-correlation in the errors and by oversampling.[15] The $MSEs$ displayed in Tables 1 to 3 suggest that none of the two versions of $SSDFM$ is beaten by $LSDFM$ when the correlation across categories is high. In addition, the effects of using oversampled categories in factor analysis are analyzed in the last two columns of these tables, which report the $MSEs$ of estimating the factor from the representative series from each category $SSDFM$ and the large scale $LSDFM$, which uses the 10 unbalanced sets of indicators described above. Overall the $LSDFM$ with unbalanced categories performs worse than the $LSDFM$ with balanced categories, especially when the correlation across categories is small. Again, the relatively better accuracy of noisy $SSDFM$ with respect to the oversampled $LSDFM$ is

[15] Recall that our benchmarks are different. They focus on choosing the optimal number of variables in a large scale (static) factor model instead of on comparing small versus large scale dynamic factor estimation.

more evident when the low correlation across categories is combined with high correlation within categories and high persistence of the factor.

The tables that try to examine the robustness of our results to different assumptions are included in the Appendix and are labelled as Tables A1 to A6. To begin with, Tables A1 to A4 examine the effects of increasing the serial correlation of the idiosyncratic components on the factor models. In particular, the effects of having higher autocorrelations of the series specific shock (measured by d) are analyzed in Tables A1 and A2 whereas the effects of assuming higher autocorrelations of the category specific shocks (measured by c) are analyzed in Tables A3 and A4. Tables A1 and A2 show the $MSEs$ of the models when the serial correlation of the idiosyncratic component is assumed to grow from $d = 0.1$ to $d = 0.75$ in two scenarios, when the serial correlation of the factor is low ($A = 0.1$ in Table A1) and when it is high ($A = 0.75$ in Table A2). The $MSEs$ reported in the tables show that increasing the serial correlation in the idiosyncratic components contributes to deteriorating the overall performance of the models even more than when the serial correlation of the factor increases. For example, while Table 1 shows that when $\rho_l = 0$, $\rho_s = 0.75$, and $A = d = 0.1$, the MSE_p^s is 0.35, Table A1 shows that the MSE increases to 0.50 when $d = 0.75$. On comparing Table 3 and Table A2, we obtain that increasing d from 0.1 to 0.75 leads the MSE to increase from 0.40 to 0.75 when $A = 0.75$. In addition, the tables show a better accuracy of a $SSDFM$ that uses the 10 representative series versus a $LSDFM$ that uses the 100 series when there is high serial correlation in the idiosyncratic components. This result reveals that the large scale model is more negatively affected by the increase of the serial correlation than the small scale model. Finally, the tables also show that the relatively larger negative effects of increasing the correlation of the idiosyncratic components in the large scale model are magnified in the case of oversampled categories.

Tables A3 and A4 analyze the role of the serial correlation of the shock of each category, which is measured by the parameter c. This parameter is allowed to increase from $c = 0.1$ to $c = 0.75$ when the serial correlation of the factor is low ($A = 0.1$ in Table A3) and when it is high ($A = 0.75$ in Table A4). Interestingly, the $MSEs$ of the small scale models do not change significantly. However, the $MSEs$ of the large scale model exhibited relatively

better accuracy than when the serial correlation of the idiosyncratic component increases. Consequently, the representative series $SSDFM$ only outperform the large scale $LSDFM$ for high levels of serial correlation of the common factor.

The role of the number of observations in the performance of factor models under different values is examined in Tables A5 and A6. According to the theory, in the absence of the typical data problems which are accounted for by our simulations and that usually appear in empirical applications, the larger the time series the better the expected performance of $LSDFM$ with respect to $SSDFM$. This theoretical result is documented in Table A5 where the reported $MSEs$ show that under low serial correlation of the factor and low correlation of the idiosyncratic errors, the accuracy of the small scale model that uses the least noisy indicators with respect to the large scale model diminishes, and the large scale model outperforms the small scale model that uses the ten representatives that exhibit the largest correlation with the series of each category. However, the tables also show that when the serial correlation of the factor increases, $SSDFM$ clearly outperforms $LSDFM$ regarding the way in which the small set of indicators is selected. Interestingly, the tables also reveal that the relative losses in accuracy due to oversampling in $LSDFM$ are still large when the sample size increases. In fact, although Table A6 shows that the accuracy of large scale models deteriorates further when facing data problems, Table A5 reveals that the unsatisfactory empirical performance of oversampled large scale models still holds even in the absence of these data problems.

As a final remark, it is worth noting that the number of factors has been restricted to one according to the data generating process. However, the generation of time series in different categories with high within-category and across-category correlation may lead this assumption to be too restrictive.[16] To evaluate the effect of this potential restriction in the accuracy of $LSDFM$ in estimating the factor, we leave the large scale model to select the number of factors according to the procedure described in Bai and Ng (2002), where the maximum number of factor is 11. Tables A7 and A8 report the MSE^l and the averaged number of estimated factors across the 1000 replications both in the case

[16] Although the datasets have been generated from one seminal factor, estimating the model from highly correlated indicators of different categories could require more than one factor.

of balanced sets of categories and in the case of oversampled categories. According to the previous discussion, the tables reveal that the higher the correlation within categories the larger the number of estimated factors since the high correlation in each category is interpreted by the model as if the series belonging to this category shared a common factor. Notably, although selecting the number of factors increases the accuracy of $LSDFM$, the gains are not sufficiently large to qualitatively alter the results obtained in this section.

4.2 Forecasting accuracy

This section examines how close the one-step-ahead out-of-sample forecasts based on the estimated factors from small and large scale dynamic factor models are to the target series which has been generated by (27). Part of the forecast performance analysis has already been developed in the previous section since, in absence of autocorrelation in the target series (measured by γ), the forecast performance is expected to increase when the discrepancy between the actual and the estimated factors diminishes.[17] Accordingly, this section examines the effects of different values of γ ranging from 0 (no inertia) to 0.8 (high degree of time series dependence) on forecast performance. In addition, the section also addresses the effects of the data problems outlined above on the the relative forecast performance of small versus large scale dynamic factor models.

Tables 4 to 6 evaluate the ability of factor models in forecasting.[18] As in the case of factor estimates, the relative forecasting accuracy of small versus large scale dynamic factor models is examined under different scenarios and the Monte Carlo simulations allow for different degrees of cross-correlation across (ρ_s from 0 to 0.5) and within (ρ_l from 0 to 0.9) categories. Table 4 shows the $MSFE$ of the models when the factor exhibits low correlation ($A = 0.1$) while Tables 5 and 6 display the MSE of the models when the factor autocorrelation increases to medium ($A = 0.5$) and to high ($A = 0.75$), respectively.

The robustness analysis can be conducted through Tables A9 to A14 in the Appendix. Tables A9 and A10 display the $MSFE$ of the models when the autocorrelation of the series

[17] Note that the variance of the errors has been normalized $\sigma_{e_y} = 1$.

[18] To save space, the tables that show the in-sample forecast analysis were omitted. In addition, the tables that show the forecast analysis have been simplified. Larger versions of these tables are available from the authors upon request.

specific shock increases to $d = 0.75$, Tables A10 and A11 show the effects of increasing the sample size to $T = 150$, and Tables A12 and A13 analyze the forecasting accuracy when the number of common factors is selected, as Bai and Ng (2002) describe.[19]

Overall, the tables show that the typical data problems lead to similar effects on the forecasting ability of the models to those observed in the analysis of factor estimation. Hence, when the time series are over correlated, with the indicators already included in some categories, the factor or the idiosyncratic components are persistent, or some categories are oversampled, forecasting with many predictors performs worse than forecasting from a representative series dataset, especially when the categories are not highly correlated. The strategy of reasonably pre-selecting the indicators to be used by $SSDFM$ almost unambiguously outperforms $LSDFM$ and $SSDFM$ from representative chosen indicators. When the data problems become large, $SSDFM$ using representatives series of each category leads to lower $MSFE$ than $LSDFM$.

However, these results depend heavily on the magnitude of the autocorrelation of the target variable since it tends to mitigate the forecasts loses of those models which are more contaminated with data problems. That is, the models that exhibits larger deteriorations in factor estimation due to data problems present smaller increases in $MSFE$ when the autocorrelation of the target variable increases. The intuition is clear: the larger the autocorrelation of the target variable the smaller the weights of the factor in forecasting the time series and the lower the effect on forecasting of inappropriate factor estimation.

For example, Tables 1 to 3 show the sharp deterioration in factor estimation of $LSDFM$ when the inertia of the factor and the within and across categories correlation became large. In particular, if the set of parameters that measure the data problems changes from $\rho_s = 0$, $\rho_l = 0$, $A = 0.1$ to $\rho_s = 0.5$, $\rho_l = 0.9$, $A = 0.75$, the tables reveal that the accuracy of the factor estimation moves from $MSE^l = 0.12$ to $MSE^l = 0.56$ which implies a 366% increase. However, under the same change in the set of parameters, the forecast accuracy moves from $MSFE^l = 1.14$ to $MSFE^l = 1.55$ when $\gamma = 0$ which implies a 36% increase

[19] The tables that examine the effects of higher category-specific autocorrelation, measured by c are omitted to save space. The results are similar to those obtained when the series-specific autocorrelation, measured by d, increased in Tables A9 and A10.

and to $MSFE^l = 1.40$ when $\gamma = 0.8$, which implies a 23% increase only.

5 Empirical analysis

To shed some empirical light on this statement, this section examines the forecasting accuracy of small versus large scale dynamic factor models by using the dataset that comprises the 147 monthly macroeconomic indicators used in a balanced panel factor estimation by Stock and Watson (2002a) for the US economy.[20] The variables, which are available over the sample 1959:01-1998:12 are standardized and transformed to induce stationarity following their indications.

5.1 Preliminary analysis of data

According to Stock and Watson (2002a), Table 7 classifies the data in 13 different categories: (1) real output and income (series 1–19); (2) employment and hours (series 20-44); (3) retail and manufacturing trade (series 45-53); (4) consumption (series 54-58); (5) housing starts and sales (series 59-65); (6) inventories (series 66-76); (7) orders (series 77-92); (8) stock prices (series 93-99); (9) exchange rate (series 100-104); (10) interest rates (105-119); (11) money and credit (series 120-126); (12) price indexes (series 127-144); (13) Average hourly earnings (series 145-146).[21]. This table also displays the name of the categories in column 1 and the number of the series included in each category in column 2. Since there are more series from some categories than others, the problem of oversampling outlined in the simulations may apply in this example.

According to the motivation of the paper, the time series included in each category are expected to be very collinear. Hence, it would be reasonable to conjecture that dozens of variables in a large scale model, including sectorial ones, might not all be useful to improve the forecasting accuracy and that it might be worth focusing on some key variables in a small scale model. In fact, the larger the correlation within the series of the same category

[20] Although the unbalanced panel proposed by Stock and Watson (2002a) included 215 time series, we concentrate on the 147 time series that form the balanced panel.

[21] The last category, labelled *miscellaneous*, has been omitted from the empirical analysis since it included only one series.

that we find, the more likely to fail is the assumption of weak correlation across the idiosyncratic components in large scale dynamic factor models that ensured the asymptotic statistical properties to be held in this empirical exercise. To gauge the potential problem, Table 7 also shows in the third column the averaged correlation across the series of each category. Overall, the categories contains very collinear indicators which exhibit averaged correlations of more than 0.5 in the cases of housing starts and sales and exchange rates and of more than 0.4 in the cases of real output and income, consumption, stock prices, and interest rates.

Besides, the name of the series that exhibit the largest averaged correlation with the series of each category is displayed in the fourth column of Table 7. These series can be considered as the representative series of each category. The last column of Table 7 reports the magnitudes of these averaged correlations. Overall, the representative series exhibit averaged correlations with the series of the same category of more than 0.5, and in some cases the correlations rise to 0.70 in the case of exchange rates and to 0.74 in the case of housing starts. Interestingly, when finer disaggregations of sectorial data are included in a category, the representative series of the category usually refers to the total (non disaggregated) indicator.

In addition, it is of great interest in this paper to examine the correlation across the indicators of different categories. If the correlations are not absorbed by the factor, the risk of the required absence of cross-correlation across the idiosyncratic components of small scale factor models grows dramatically when the empirical correlations are very large. For this purpose, Table 8 displays the correlation across the thirteen representative series of the different categories. The high correlation coefficients reported in the table for some pairs of categories indicate that there is a high collinearity between these categories. As expected, the highest correlations appear between industrial production and employment (correlation of 0.64) and between manufacturing and trade sales and orders (correlation of 0.60).

5.2 Forecasting accuracy

In this paper we consider two real (industrial production and nonagricultural employment) and two nominal (consumer and producer price indexes) target series, each of which is called Y_t. Accordingly, we investigate the accuracy of the different specifications of dynamic factor models to forecast industrial production using the following multi-step ahead forecasting procedure described in Stock and Watson (2002a)

$$y_{t+h}^{h} = \alpha_0 + \sum_{i=0}^{m} \beta_j^{'} \widehat{F}_{t-i} + \sum_{j=0}^{v} \gamma_j z_{t-j} + \varepsilon_{t+h}^{h}. \tag{30}$$

In this equation, y_{t+h}^{h} is the h-step ahead covariance stationary transformation of the original series Y_t, where $y_{t+h}^{h} = \ln(Y_{t+h}/Y_t)$, $\widehat{F}_{t-i}$ is the i-lagged ($i = 0, 1, ..., m$) value of the ($r \times 1$) vector of estimated factors, and z_{t-j} is the j-lagged ($j = 0, 1, ..., v$) value of the 1-step ahead covariance stationary transformation of Y_t, where $z_t = \ln(Y_t/Y_{t-1})$. Expressions $\beta_j^{'}$ and γ_j refer to the standard parameters of autoregressive processes. The term ε_{t+h}^{h} is a homoskedastic martingale difference sequence with respect to the set of information at time t. Finally, in line with previous studies in forecasting with empirical factors, our model is allowed to choose values of m lying between 1 and 6 and v lying between 1 and 12 based upon the BIC selection criterion. In large scale factor specifications, r is either imposed as one or selected as Bai and Ng (2002) describe.

The pseudo real-time forecasting exercise begins with data from 1959:3-1970:1. Using this sample, m, v, and (in some cases) r are chosen, and an h period ahead forecast is formed by using values of the regressors at 1970:1 to give $y_{1970:1+h}^{h}$. Then, the sample is updated by one period, the factors and the forecasting models (including m, v, and, in some cases, r) are re-estimated, and an h-month forecast for 1970:1+h is computed (for $h = 1$ it would be 1970:2, for $h = 6$ 1970:7 and for $h = 12$ 1971:1). The forecasting procedure continues iteratively until the final forecast $y_{1998:12}^{h}$ which is made using data until 1998:11 for $h = 1$, 1998:6 for $h = 6$ and 1997:12 for $h = 12$. In each iteration, the root of the squared deviation of h-ahead forecasts from actual data are computed and the average of these figures is labeled as $RMSFE(h)$.

To investigate the benefits of forecasting with the two different versions of dynamic factor models, we consider a forecast competition of different diffusion index forecasts

from small and large scale datasets. The first competitor is a simple autoregressive model which is obtained when $\beta s_j = 0$ in (30). The second competitor is an autoregressive model that is enlarged with the factors obtained from a large scale dynamic factor model applied to the 146 economic indicators. The number of factors included in the analysis is either imposed as one or selected by using the Bai and Ng (2002) criterion. The third competitor is an autoregressive model that is enlarged with the factors obtained from a small scale dynamic factor model applied to the 13 representative indicators, which are the series of each category that exhibit the highest averaged autocorrelation. In the case of small scale factor models, the number of factors is also either imposed as one or selected by BIC.[22]

To facilitate comparisons, Tables 9 and 10 report the root mean square forecast errors relative to the autoregressive models. Hence, an entry less than one indicates that the diffusion index forecast is superior to the autoregressive univariate forecast. According to Stock and Watson (2002a), regarding the factor model and the forecasting horizon used in the analysis, the diffusion index forecasts generally improve over the benchmark univariate forecasts. However, the forecasting accuracy largely depends on the number of factors included in the analysis. For example, Table 9 shows that when only one factor is included in the diffusion index forecasts, the relative mean squared errors are always greater than 0.9, which implies that the factor forecasts are only slightly more accurate than the univariate autoregressive forecasts. To gauge this property, Figure 1 plots the h-step ahead growth of Industrial Production (IP), y^h_{t+h} over the sample 1970:01-1998:12-h. As expected, the persistence of the series increases with h, and the correlation is 0.37 when $h = 1$, and 0.96 when $h = 12$. When $h = 12$, the high persistence of the target variable is better captured by the first factor of the small scale model (correlation of 0.98) than by the first factor of the large scale model (correlation of 0.66). For easier comparisons, the first two factors of $SSDFM$ and $LSDFM$ are plotted in Figures 2 and 3.

Accordingly, the performance of factor models that determines the number of factors required in the factor estimation is much better than when the number of factor is restricted to one, especially when the forecasting horizon becomes large. Notably, Table 9

[22] In the simulation exercise, we knew that the true number of factors was one. In the empirical application, we found that the data are better characterized by using two factors.

confirms the results obtained by the simulation study conducted throughout the paper. There may be a similar forecast efficiency either when constructing the diffusion index forecasts from a small scale dataset that includes a representative (highly correlated) time series from each category or from a large scale dataset that contains larger, but redundant, information about the factors. Although none of the factor models systematically perform better than the other, the factor forecasts accuracy of the small scale model that uses 13 representative indicators is similar to (or, in many cases, better than) that obtained when the forecasts of industrial production and employment are computed from a large scale model that uses the 146 indicators

The results for nominal variables are presented in Table 10. As in the case of forecasting real variables, the diffusion index forecasts of the consumer price index and the producer price index for finished goods that are computed from small scale factor models uniformly outperform the forecasts for those nominal variables computed from large scale factor models when the number of factors is selected from the data. Regarding the forecast horizon, the small scale factor model consistently performs better than the large scale factor model when the number of factors used in the analysis are selected from the data, with relative performance improving as the horizon increases.

6 Conclusions

Two versions of dynamic factor models have received growing attention in the recent forecasting literature: the dynamic factors that use large datasets, and the dynamic factors which use a small number of indicators that has been preselected reasonably. However, the problem of systematically selecting many series from very many series that face the typical data problems associated with empirical applications is still developing.

In this paper, we propose simulations which mimic different scenarios of empirical forecasting, where the list of series, which are extracted from different economic categories, is fixed (rather than tending to infinity), and potentially greater cross-correlation and serial correlation may appear among idiosyncratic components than those warranted by the theory. Accordingly, our Monte Carlo analysis allows for indicators which belong to

different categories of data and whose idiosyncratic components show cross-correlation within and across categories in addition to serial correlation. We also allow for categories which are oversampled. Finally, the simulations examine the accuracy of small versus large data sets under different degrees of serial correlation in the factor.

To gauge the problem, we compare the forecast accuracy of a large scale factor model that uses the information provided by a large dataset with that of a small scale factor model that uses information from one representative of each category and the time series with large averaged correlation with the series of the same category. We find that adding data that have little information on the factor components does not necessarily lead large scale dynamic factor models to improve upon the forecasts of small scale dynamic factor models. In fact, we show that when the additional data are over correlated with data from some categories which are already included in factor estimation, forecasting with many predictors performs worse than forecasting from a reasonably pre-screened dataset, especially when the categories are not highly correlated. This result is stronger in the case of high persistence of the common factor, in the case of high serial correlation of the idiosyncratic components, in the case of using noisy series, and in the case of oversampled categories. In these cases, even arbitrarily selecting one time series from each category and using the resulting dataset in a small scale dynamic factor model outperforms the forecasts from large scale dynamic factor models. In these situations, our results suggest that it may be better to discard some redundant data even if such data are available. Using the 147 indicators that form the balanced panel used by Stock and Watson (2002a), we illustrate these results for US data.

References

[1] Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin., L., and Rünstler, G. 2008. Short-term forecasts of euro area GDP growth. *Econometrics Journal* 14: 25-44.

[2] Aruoba, B., Diebold, F., and Scotti, C. 2009. Real-time measurement of business conditions. *Journal of Business and Economic Statistics* 27: 417-427.

[3] Aruoba, B., and Diebold, F. 2010. Real-time macroeconomic monitoring: Real activity, inflation, and interactions. *American Economic Review* 100: 20-24.

[4] Bai, J., and Ng, S. 2002. Determining the number of Factors in approximate factor models. *Econometrica* 70: 191-221.

[5] Bai, J., and Ng, S. 2006. Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, 131: 507-537.

[6] Banbura,M., and Mondugno, M. 2010. Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data. ECB working paper 1189.

[7] Banbura, M., and Runstler, G. 2011. A look into the model factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting* 27 333-346.

[8] Boivin, J., and Ng, S. 2006. Are more data always better for factor analysis? *Journal of Econometrics* 132: 169-194.

[9] Caggiano, G., Kapetanios, G., and Labhard, V. 2011. Are more data always better for factor analysis? Results for the Euro area, the six largest Euro area countries and the UK. *Journal of Forecasting.* Forthcoming.

[10] Camacho, M., and Perez Quiros, G. 2010. Introducing the Euro-STING: Short Term INdicator of Euro Area Growth. *Journal of Applied Econometrics*, 25: 663–694.

[11] Doz, C., Giannone, D., and Reichlin, L. 2007. A quasi-maximum likelihood approach for large approximate dynamic factor models. ECB working paper 674.

[12] Forni, M., Hallin, M., Lippi, M., and Reichlin, L. 2005. The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association* 100: 830-40.

[13] Giacomini, R. and White, H. 2006. Tests of Conditional Predictive Ability. *Econometrica*, 74: 1545-1578.

[14] Giannone, D., Reichlin, L., and Small, D. 2008. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics* 55: 665-676.

[15] Mariano, R., and Murasawa, Y. 2003. A new coincident index os business cycles based on monthly and quarterly series. *Journal of Applied Econometrics* 18: 427-443.

[16] Moench, E Ng, S and Potter S. 2009 Dynamic Hierarchical Factor Models. Federal Reserve Bank of New York Staff Reports 412. December.

[17] Nunes, L. 2005. Nowcasting quarterly GDP growth in a monthly coincident indicator model. *Journal of Forecasting* 24: 575-592.

[18] Stock, J., and Watson, M. 1991. A probability model of the coincident economic indicators. In *Leading Economic Indicators: New Approaches and Forecasting Records*, edited by K. Lahiri and G. Moore. Cambridge University Press.

[19] Stock, J., and Watson, M. 2002a. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20 147-162.

[20] Stock, J., and Watson, M. 2002b. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167-1179.

Table 1. Common factor estimation (T=50, c=0.1, d=0.1, A=0.1)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation across categories ρ_s=0					
0	0.101	0.195	0.124	0.191	0.149
0.1	0.101	0.192	0.125	0.191	0.151
0.5	0.101	0.196	0.139	0.190	0.166
0.9	0.101	0.195	0.185	0.192	0.320
Correlation across categories ρ_s=0.1					
0	0.116	0.207	0.139	0.204	0.159
0.1	0.116	0.205	0.141	0.204	0.162
0.5	0.116	0.205	0.152	0.203	0.175
0.9	0.116	0.206	0.197	0.202	0.310
Correlation across categories ρ_s=0.5					
0	0.223	0.289	0.236	0.285	0.235
0.1	0.223	0.286	0.239	0.284	0.234
0.5	0.223	0.286	0.246	0.284	0.243
0.9	0.223	0.287	0.281	0.284	0.300
Correlation across categories ρ_s=0.75					
0	0.350	0.383	0.350	0.382	0.346
0.1	0.350	0.380	0.349	0.383	0.344
0.5	0.350	0.381	0.359	0.376	0.346
0.9	0.350	0.377	0.376	0.378	0.376

Notes. The values of ρs determine the cross-correlation of the idiosyncratic shocks between series from different categories, and the values of ρl determine the cross-correlation of the idiosyncratic shocks between series from the same category. T is the sample size. Parameters A and c measure the serial correlation of the factor and the idiosyncratic shocks, respectively. The Mean Squared Errors of the models uses the 10 representative series of each category, the model that uses the 10 series with higher correlation with others of each category and the model that uses all the 100 series, all of which are denoted by MSE_p^s, MSE_r^s, and MSE^l, respectively.

Table 2. Common factor estimation (T=50, c=0.1, d=0.1, A=0.5)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation across categories ρ_s=*0*					
0	0.100	0.191	0.175	0.190	0.202
0.1	0.100	0.190	0.175	0.188	0.200
0.5	0.100	0.192	0.190	0.188	0.217
0.9	0.100	0.191	0.236	0.187	0.350
Correlation across categories ρ_s=*0.1*					
0	0.115	0.204	0.191	0.201	0.207
0.1	0.115	0.203	0.191	0.201	0.208
0.5	0.115	0.204	0.206	0.200	0.229
0.9	0.115	0.203	0.250	0.199	0.340
Correlation across categories ρ_s=*0.5*					
0	0.227	0.293	0.294	0.290	0.290
0.1	0.227	0.291	0.297	0.288	0.289
0.5	0.227	0.291	0.305	0.290	0.304
0.9	0.227	0.291	0.343	0.288	0.368
Correlation across categories ρ_s=*0.75*					
0	0.372	0.399	0.414	0.403	0.409
0.1	0.372	0.400	0.415	0.405	0.415
0.5	0.372	0.407	0.430	0.402	0.422
0.9	0.372	0.400	0.450	0.402	0.448

Notes. See notes for Table 1.

Table 3. Common factor estimation (T=50, c=0.1, d=0.1, A=0.75)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation across categories ρ_s=*0*					
0	0.097	0.182	0.382	0.180	0.395
0.1	0.097	0.182	0.384	0.180	0.427
0.5	0.097	0.183	0.397	0.181	0.429
0.9	0.097	0.182	0.444	0.180	0.525
Correlation across categories ρ_s=*0.1*					
0	0.112	0.195	0.398	0.192	0.417
0.1	0.112	0.195	0.400	0.193	0.421
0.5	0.112	0.196	0.413	0.194	0.428
0.9	0.112	0.195	0.459	0.191	0.559
Correlation across categories ρ_s=*0.5*					
0	0.230	0.290	0.510	0.289	0.515
0.1	0.230	0.291	0.512	0.286	0.506
0.5	0.230	0.291	0.524	0.288	0.524
0.9	0.232	0.289	0.565	0.286	0.574
Correlation across categories ρ_s=*0.75*					
0	0.406	0.425	0.644	0.432	0.650
0.1	0.406	0.425	0.646	0.430	0.652
0.5	0.406	0.425	0.655	0.426	0.680
0.9	0.406	0.425	0.688	0.427	0.711

Notes. See notes for Table 1.

Table 4. Forecasting accuracy (T=50, c=0.1, d=0.1, A=0.1)

Correlation within categories ρ_l	Persistency of the target series γ	Same number of series in each category			Oversampling one category	
		$MSFE_p^s$	$MSFE_r^s$	$MSFE^l$	$MSFE_r^s$	$MSFE^l$
Correlation across categories ρ_s =0						
0	0	1.107	1.215	1.14	1.171	1.161
	0.3	1.101	1.202	1.161	1.218	1.199
	0.8	1.086	1.172	1.099	1.178	1.173
0.9	0	1.107	1.354	1.341	1.378	1.558
	0.3	1.101	1.146	1.129	1.166	1.286
	0.8	1.086	1.273	1.235	1.318	1.459
Correlation across categories ρ_s =0.5						
0	0	1.197	1.280	1.198	1.371	1.342
	0.3	1.200	1.248	1.237	1.324	1.321
	0.8	1.154	1.222	1.156	1.288	1.238
0.9	0	1.197	1.324	1.314	1.248	1.239
	0.3	1.200	1.320	1.300	1.441	1.425
	0.8	1.154	1.320	1.319	1.394	1.407

Notes. The estimated model is $y_{t+1} = \beta F_t + \gamma y_t + e_{yt+1}$. See notes for Table 1.

Table 5. Forecasting accuracy (T=50, c=0.1, d=0.1, A=0.5)

Correlation within categories ρ_l	Persistency of the target series γ	Same number of series in each category			Oversampling one category	
		$MSFE_p^s$	$MSFE_r^s$	$MSFE^l$	$MSFE_r^s$	$MSFE^l$
Correlation across categories ρ_s =0						
0	0	1.169	1.299	1.248	1.279	1.27
	0.3	1.121	1.203	1.184	1.289	1.292
	0.8	1.222	1.363	1.343	1.411	1.36
0.9	0	1.169	1.300	1.377	1.302	1.415
	0.3	1.121	1.313	1.335	1.306	1.413
	0.8	1.222	1.328	1.306	1.204	1.307
Correlation across categories ρ_s =0.5						
0	0	1.220	1.290	1.249	1.291	1.286
	0.3	1.299	1.357	1.349	1.433	1.399
	0.8	1.218	1.382	1.351	1.374	1.297
0.9	0	1.220	1.259	1.275	1.238	1.276
	0.3	1.299	1.395	1.397	1.397	1.445
	0.8	1.218	1.357	1.372	1.293	1.345

Notes. See notes for Tables 1 and 4.

Table 6. Forecasting accuracy (T=50, c=0.1, d=0.1, A=0.75)

Correlation within categories ρ_l	Persistency of the target series	Same number of series in each category			Oversampling one category	
	γ	$MSFE_p^s$	$MSFE_r^s$	$MSFE^l$	$MSFE_r^s$	$MSFE^l$
Correlation across categories ρ_s =0						
0	0	1.133	1.275	1.343	1.213	1.388
	0.3	1.132	1.216	1.347	1.212	1.396
	0.8	1.151	1.345	1.408	1.27	1.419
0.9	0	1.133	1.201	1.387	1.273	1.417
	0.3	1.132	1.212	1.316	1.205	1.388
	0.8	1.151	1.243	1.363	1.27	1.455
Correlation across categories ρ_s =0.5						
0	0	1.329	1.371	1.493	1.414	1.516
	0.3	1.373	1.449	1.514	1.498	1.597
	0.8	1.315	1.379	1.462	1.454	1.502
0.9	0	1.329	1.401	1.549	1.378	1.462
	0.3	1.373	1.410	1.517	1.428	1.605
	0.8	1.315	1.326	1.393	1.231	1.388

Notes. See notes for Tables 1 and 4.

Table 7. Data description

Category name	Number of series	Averaged cross-correlation	Representative series of the category	Highest averaged cross-correlation
1. Real output and income	19	0.422	Industrial production: total index	0.570
2. Employment and hours	25	0.323	Employees on nonagricultural Payrolls: total	0.475
3. Real retail, manufacturing and trade sales	9	0.381	Manufacturing & trade: total	0.623
4. Consumption	5	0.403	Personal consumption expend, total	0.640
5. Housing starts and sales	7	0.559	Housing starts: total farm & nonfarm	0.740
6. Real inventories and inventory-sales ratios	11	0.272	Manufacturing & trade inventories: total	0.426
7. Orders and unfilled orders	16	0.363	Mfg new orders: mfg industries with unfilled orders	0.435
8. Stock prices	7	0.476	S&P's common stock price index: composite	0.635
9. Exchange rates	5	0.515	United States effective exchange rate	0.701
10. Interest rates	15	0.427	Spread US treasury bills, secondary market 10-years and federal fund rate	0.517
11. Money and credit quantity aggregates	7	0.286	Money stock: M2	0.345
12. Price indexes	18	0.214	Cpi-u: all items	0.288
13. Average hourly earnings	2	0.313	Average hourly earnings of production workers: manufacturing	0.313
Total	146		13	

Notes. The dataset, the definition of the thirteen categories, and the distribution of the indicators across these categories follows Stock and Watson (2002a). The representative series of each category is the economic indicator that exhibits the largest averaged correlation with the series of the same category. The last column reports these correlations.

Table 8. Correlation across categories

	cat 1	cat 2	cat 3	cat 4	cat 5	cat 6	cat 7	cat 8	cat 9	cat 10	cat 11	cat 12	cat 13
cat 1	1.00	0.64	0.52	0.19	0.32	0.18	0.39	-0.01	0.09	0.19	-0.06	0.03	0.16
cat 2	-	1.00	0.43	0.19	0.50	0.32	0.27	-0.04	-0.01	0.09	-0.02	0.04	0.06
cat 3	-	-	1.00	0.48	0.26	0.08	0.61	0.13	0.04	0.17	-0.07	-0.01	0.03
cat 4	-	-	-	1.00	0.14	-0.11	0.23	0.17	0.03	0.14	-0.03	-0.05	-0.06
cat 5	-	-	-	-	1.00	0.20	0.20	0.01	-0.16	0.07	-0.05	0.03	0.01
cat 6	-	-	-	-	-	1.00	-0.01	-0.13	-0.06	-0.12	-0.05	-0.02	-0.01
cat 7	-	-	-	-	-	-	1.00	0.02	0.06	0.08	-0.04	0.11	0.07
cat 8	-	-	-	-	-	-	-	1.00	-0.05	0.16	0.10	-0.02	-0.02
cat 9	-	-	-	-	-	-	-	-	1.00	-0.10	-0.12	-0.04	-0.02
cat 10	-	-	-	-	-	-	-	-	-	1.00	0.01	-0.01	0.02
cat 11	-	-	-	-	-	-	-	-	-	-	1.00	0.01	0.05
cat 12	-	-	-	-	-	-	-	-	-	-	-	1.00	-0.04
cat 13	-	-	-	-	-	-	-	-	-	-	-	-	1.00

Notes. The entries refer to the correlations between pairs of representative series of each category. See notes for Table 7.

Table 9. Forecasting real variables

	Industrial production			Nonagricultural employment		
	Forecast horizon			Forecast horizon		
	h=1	*h=6*	*h=12*	*h=1*	*h=6*	*h=12*
Forecast method	*RMSFE(h)*			*RMSFE(h)*		
AR	0.007	0.031	0.049	0.002	0.009	0.017
	Relative (to the *AR*) *RMSFE(h)*			Relative (to the *AR*) *RMSFE(h)*		
LSDFM, *r*=1	0.90	0.92	0.97	0.88	0.92	0.91
SSDFM, *r*=1	0.96	0.96	0.92	0.92	0.89	0.86
LSDFM, *r**	0.87	0.66	0.52	0.84	0.79	0.65
SSDFM with *r**	0.87	0.73	0.52	0.91	0.78	0.63

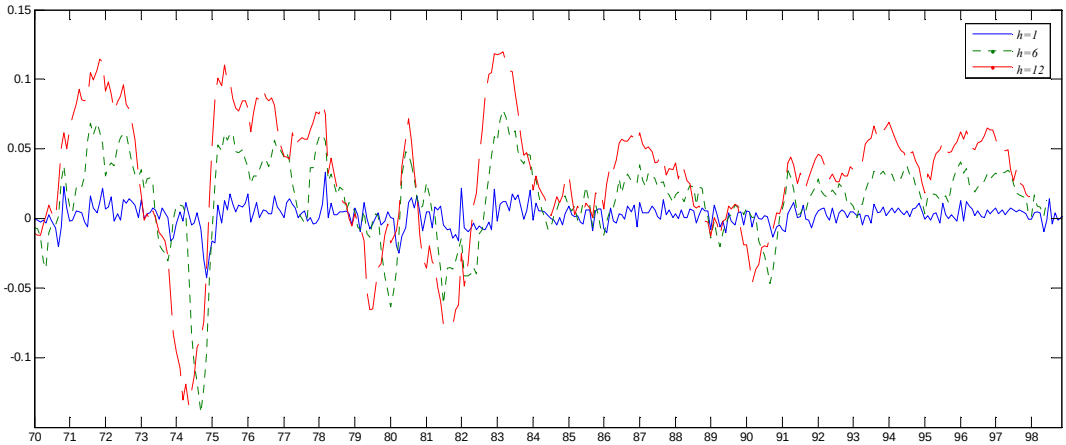
Notes. The sample period is 1959:03-1998:12 and the out-of-sample forecast period is 1971:01-1998:12. The competing models are the autoregressive model, and the autoregressive model extended with factors. The LSDFM is applied to the 146 indicators and the SSDFM is applied to the 13 representative series of each category that exhibit the largest average autocorrelation with the series of the same category. In some cases, the number of factors is restricted to r=1, while in others the optimal number of factors r^* is determined by using Bai and Ng (2002) in large scale models and by using BIC in small scale models.

Table 10. Forecasting nominal variables

	Consumer price index			Producer price index		
	Forecast horizon			Forecast horizon		
	h=1	*h=6*	*h=12*	*h=1*	*h=6*	*h=12*
Forecast method	*RMSFE(h)*			*RMSFE(h)*		
AR	0.002	0.010	0.021	0.008	0.026	0.046
	Relative (to the *AR*) *RMSFE(h)*			Relative (to the *AR*) *RMSFE(h)*		
LSDFM, *r*=1	0.98	0.81	0.75	0.87	0.87	0.90
SSDFM, *r*=1	0.99	0.80	0.75	0.94	0.91	0.90
LSDFM, *r**	1.02	0.94	0.87	1.14	1.00	0.97
SSDFM with *r**	0.99	0.92	0.86	1.00	0.95	0.88

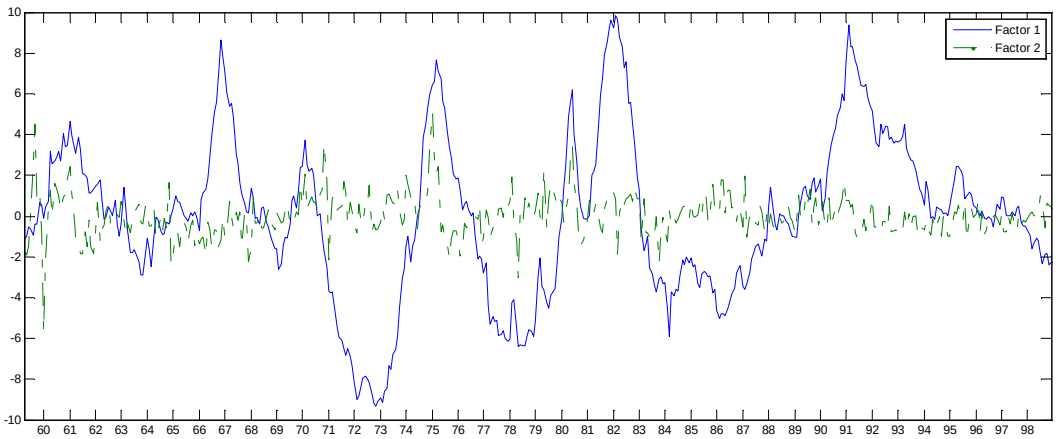
Notes. See notes for Table 9.

Figure 1. Industrial production



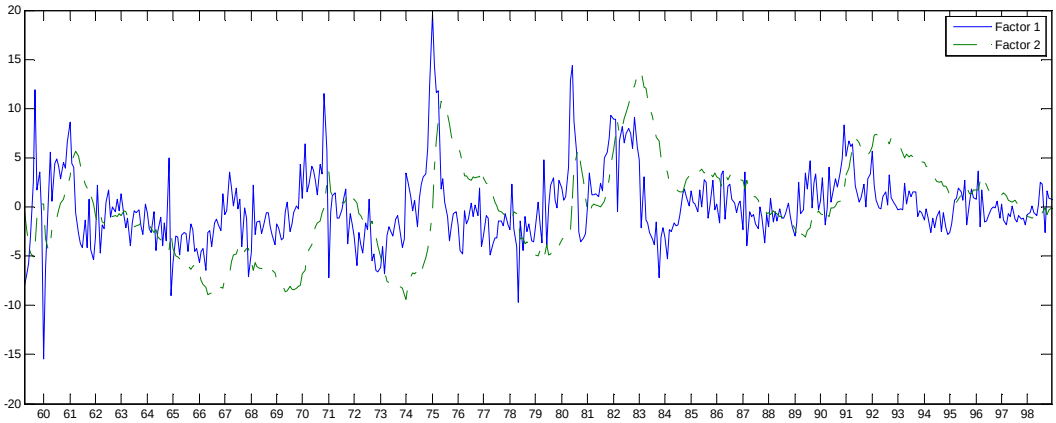
Notes. The h step ahead growth of Industrial production (IP) is $y_{t+h}^{h} = \ln(IP_{t+h} / IP_t)$. The sample is 1970:01-1998:12-h.

Figure 2. Factors estimated from the small scale model



Notes. The figure plots the first two factors obtained from *SSDFM* applied to the thirteen representative categories by using data from 1959:03 to 1998:11.

Figure 3. Factors estimated from the large scale model



Notes. The figure plots the first two factors obtained from *LSDFM* applied to the 147 indicators by using data from 1959:03 to 1998:11

APPENDIX

Table A1. Common factor estimation (T=50, c=0.1, d=0.75, A=0.1)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation across categories $\rho_s=0$					
0	0.101	0.278	0.145	0.278	0.167
0.9	0.101	0.276	0.296	0.276	0.410
Correlation across categories $\rho_s=0.75$					
0	0.346	0.421	0.356	0.421	0.349
0.9	0.346	0.422	0.426	0.422	0.431

Notes. See notes for Table 1.

Table A2. Common factor estimation (T=50, c=0.1, d=0.75, A=0.75)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation across categories $\rho_s=0$					
0	0.097	0.357	0.418	0.357	0.437
0.9	0.097	0.352	0.548	0.352	0.631
Correlation across categories $\rho_s=0.75$					
0	0.378	0.563	0.669	0.563	0.660
0.9	0.379	0.560	0.754	0.560	0.754

Notes. See notes for Table 1.

Table A3. Common factor estimation (T=50, c=0.75, d=0.75, A=0.1)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation across categories $\rho_s=0$					
0	0.169	0.265	0.317	0.272	0.382
0.9	0.172	0.265	0.518	0.263	0.604
Correlation across categories $\rho_s=0.75$					
0	0.503	0.506	0.538	0.508	0.534
0.9	0.503	0.505	0.591	0.510	0.596

Notes. See notes for Table 1.

Table A4. Common factor estimation (T=50, c=0.75, d=0.75, A=0.75)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation across categories $\rho_s=0$					
0	0.251	0.470	0.507	0.482	0.558
0.9	0.251	0.461	0.696	0.496	0.833
Correlation across categories $\rho_s=0.75$					
0	0.751	0.820	0.885	0.843	0.874
0.9	0.749	0.819	0.964	0.845	0.961

Notes. See notes for Table 1.

Table A5. Common factor estimation (T=150, c=0.1, d=0.1, A=0.1)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation across categories $\rho_s=0$					
0	0.095	0.175	0.108	0.176	0.134
0.9	0.094	0.176	0.161	0.175	0.333
Correlation across categories $\rho_s=0.75$					
0	0.350	0.376	0.340	0.375	0.333
0.9	0.350	0.377	0.370	0.375	0.364

Notes. See notes for Table 1.

Table A6. Common factor estimation (T=150, c=0.75, d=0.75, A=0.1)

Correlation within categories ρ_l	Same number of series in each category			Oversampling one category	
	MSE_p^s	MSE_r^s	MSE^l	MSE_r^s	MSE^l
Correlation error term Series of *SSDFM:* $\rho_s=0$					
0	0.092	0.168	0.195	0.168	0.218
0.9	0.092	0.169	0.252	0.169	0.314
Correlation across categories $\rho_s=0.75$					
0	0.409	0.427	0.487	0.427	0.477
0.9	0.409	0.428	0.531	0.429	0.523

Notes. See notes for Table 1.

Table A7. Common factor estimation (T=50, c=0.1, d=1, A=0.1).

Correlation within categories ρ_l	Same number of series in each category		Oversampling one category	
	$\hat{r}$	MSE^l	$\hat{r}$	MSE^l
Correlation across categories $\rho_s=0$				
0	3.33	0.119	1	0.147
0.9	10.89	0.140	1.84	0.196
Correlation across categories $\rho_s=0.75$				
0	2.60	0.326	1.20	0.350
0.9	10.89	0.288	2.84	0.363

Notes. The number of common factors is selected as in Bai and Ng (2002). The values of $\hat{r}$ are the averaged number of estimated number of factors across replications. See notes for Table 1.

Table A8. Common factor estimation (T=50, c=0.1, d=1, A=0.75).

Correlation within categories ρ_l	Same number of series in each category		Oversampling one category	
	$\hat{r}$	MSE^l	$\hat{r}$	MSE^l
Correlation across categories $\rho_s=0$				
0	2.39	0.380	1	0.404
0.9	10.88	0.403	1.89	0.455
Correlation across categories $\rho_s=0.75$				
0	2.58	0.621	1.24	0.643
0.9	10.86	0.587	2.06	0.667

Notes. See notes for Tables 1 and A7.

Table A9. Forecasting accuracy (T=50, c=0.1, d=0.75, A=0.1)

Correlation within categories ρ_l	Persistency of the target series γ	Same number of series in each category			Oversampling one category	
		$MSFE_p^s$	$MSFE_r^s$	$MSFE^l$	$MSFE_r^s$	$MSFE^l$
Correlation across categories ρ_s =0						
0	0	1.184	1.257	1.213	1.255	1.236
	0.3	1.190	1.265	1.217	1.263	1.238
	0.8	1.194	1.274	1.205	1.273	1.227
0.9	0	1.184	1.409	1.402	1.409	1.491
	0.3	1.190	1.415	1.413	1.415	1.497
	0.8	1.194	1.431	1.415	1.431	1.493
Correlation across categories ρ_s =0.5						
0	0	1.270	1.416	1.302	1.415	1.303
	0.3	1.277	1.422	1.310	1.420	1.310
	0.8	1.277	1.434	1.302	1.425	1.301
0.9	0	1.270	1.439	1.439	1.434	1.487
	0.3	1.277	1.446	1.449	1.442	1.496
	0.8	1.277	1.457	1.453	1.455	1.496

Notes. See notes for Table 4.

Table A10. Forecasting accuracy (T=50, c=0.1, d=0.75, A=0.75)

Correlation within categories ρ_l	Persistency of the target series γ	Same number of series in each category			Oversampling one category	
		$MSFE_p^s$	$MSFE_r^s$	$MSFE^l$	$MSFE_r^s$	$MSFE^l$
Correlation across categories ρ_s =0						
0	0	1.327	1.572	1.587	1.582	1.630
	0.3	1.335	1.579	1.579	1.586	1.619
	0.8	1.343	1.597	1.547	1.596	1.584
0.9	0	1.327	1.590	1.753	1.580	1.841
	0.3	1.335	1.587	1.737	1.574	1.828
	0.8	1.343	1.591	1.724	1.580	1.861
Correlation across categories ρ_s =0.5						
0	0	1.509	1.603	1.680	1.617	1.713
	0.3	1.493	1.589	1.653	1.600	1.688
	0.8	1.532	1.631	1.629	1.644	1.672
0.9	0	1.509	1.598	1.771	1.597	1.857
	0.3	1.493	1.573	1.743	1.579	1.827
	0.8	1.532	1.624	1.747	1.644	1.872

Notes. See notes for Table 4.

Table A11. Forecasting accuracy (T=150, c=0.1, d=0.1, A=0.1)

Correlation within categories ρ_l	Persistency of the target series γ	Same number of series in each category			Oversampling one category	
		$MSFE_p^s$	$MSFE_r^s$	$MSFE^l$	$MSFE_r^s$	$MSFE^l$
Correlation across categories ρ_s =0						
0	0	1.083	1.169	1.099	1.165	1.131
	0.3	1.086	1.174	1.103	1.171	1.136
	0.8	1.087	1.178	1.105	1.176	1.139
0.9	0	1.096	1.182	1.161	1.184	1.328
	0.3	1.099	1.186	1.165	1.189	1.331
	0.8	1.098	1.189	1.167	1.192	1.328
Correlation across categories ρ_s =0.5						
0	0	1.209	1.274	1.221	1.267	1.225
	0.3	1.213	1.279	1.225	1.273	1.229
	0.8	1.213	1.281	1.227	1.276	1.232
0.9	0	1.209	1.277	1.268	1.276	1.305
	0.3	1.213	1.282	1.272	1.281	1.309
	0.8	1.213	1.287	1.276	1.286	1.313

Notes. See notes for Table 4.

Table A12. Forecasting accuracy (T=150, c=0.1, d=0.1, A=0.75)

Correlation within categories ρ_l	Persistency of the target series γ	Same number of series in each category			Oversampling one category	
		$MSFE_p^s$	$MSFE_r^s$	$MSFE^l$	$MSFE_r^s$	$MSFE^l$
Correlation across categories ρ_s =0						
0	0	1.019	1.101	1.085	1.117	1.100
	0.3	1.021	1.101	1.082	1.118	1.097
	0.8	1.028	1.107	1.092	1.123	1.105
0.9	0	1.088	1.179	1.205	1.182	1.267
	0.3	1.091	1.185	1.206	1.187	1.267
	0.8	1.098	1.195	1.210	1.195	1.276
Correlation across categories ρ_s =0.5						
0	0	1.228	1.300	1.315	1.297	1.319
	0.3	1.231	1.303	1.311	1.301	1.316
	0.8	1.237	1.317	1.318	1.314	1.322
0.9	0	1.228	1.305	1.365	1.303	1.386
	0.3	1.231	1.312	1.363	1.310	1.382
	0.8	1.237	1.322	1.372	1.320	1.393

Notes. See notes for Table 4.

Table A13. Forecasting accuracy (T=50, c=0.1, d=0.1, A=0.1).

Correlation within categories ρ_l	Persistency of the target series γ	Same number of series in each category	Oversampling one category
		$MSFE^l$	$MSFE^l$
Correlation across categories ρ_s =0			
0	0	1.335	1.252
	0.3	1.328	1.246
	0.8	1.303	1.229
0.9	0	1.515	1.339
	0.3	1.518	1.335
	0.8	1.510	1.332
Correlation across categories ρ_s =0.5			
0	0	1.426	1.367
	0.3	1.418	1.361
	0.8	1.415	1.359
0.9	0	1.732	1.439
	0.3	1.721	1.432
	0.8	1.696	1.419

Notes. The number of common factors is selected as in Bai and Ng (2002). See notes for Table 4.

Table A14. Forecasting accuracy (T=150, c=0.1, d=0.1, A=0.75).

Correlation within categories ρ_l	Persistency of the target series γ	Same number of series in each category	Oversampling one category
		$MSFE^l$	$MSFE^l$
Correlation across categories ρ_s =0			
0	0	1.491	1.446
	0.3	1.470	1.423
	0.8	1.448	1.395
0.9	0	2.023	1.608
	0.3	1.986	1.586
	0.8	1.931	1.556
Correlation across categories ρ_s =0.5			
0	0	1.632	1.540
	0.3	1.609	1.518
	0.8	1.595	1.513
0.9	0	1.928	1.610
	0.3	1.881	1.589
	0.8	1.857	1.584

Notes. See notes for Tables 4 and A13.