Factor models for large and incomplete data sets with unknown group structure *

Maximo Camacho† German Lopez-Buenache‡
University of Murcia Catholic University of Murcia

November 19, 2021

Abstract

Most economic applications rely on a large number of time series, which are typically drawn with a remarkable clustering structure and available over different spans. To handle these databases, this paper combines the Expectation-Maximization (EM) algorithm outlined by Stock and Watson (2002) and the estimation algorithm for large factor models with an unknown number of group structures and unknown membership advocated by Ando and Bai (2016, 2017). Several Monte Carlo experiments are used to show the good performance of the proposal to determine the correct number of clusters, to provide the right number of group-specific factors, to identify error-free group membership and to perform accurate estimates of the unobserved missing data. In addition, we find that our proposal substantially outperforms the standard EM algorithm when the data exhibit grouped factor structure. Using the Federal Reserve Economic Data FRED-QD, our procedure detects two distinct groups of macroeconomic indicators: the real activity indicators and the nominal indicators. We illustrate the usefulness of our group-specific factor model for the study of business cycle chronology and for forecasting purposes.

Keywords: Big data, Factor models, Clustering, Missing data, Business cycles, Forecasting.
JEL Classification: C32, C33, C55, C82, E32.

*M. Camacho is grateful for the support of grants PID2019-107192GB-I00 (AEI/10.13039/501100011033) and 19884/GERM/15. We thank the two reviewers and the associate editor in charge of our manuscript for their valuable comments. All remaining errors are our responsibility. Data and codes that replicate our results are available from the authors' websites.

†Department of Quantitative Analysis, University of Murcia. Campus de Espinardo 30100 Murcia (Spain). E-mail: mcamacho@um.es

‡Department of Business Administration, Catholic University of Murcia. Avenida de los Jeronimos, 135, 30107 Murcia (Spain). E-mail: glopez@ucam.es

1 Introduction

Most economic empirical analyses are based on peculiar data structures. As information technology improves, empirical economists face data sets that have hundreds of economic indicators, which frequently involves dealing with an information overload without a clear way to organize the data. Since these data are usually very collinear, factor analysis has become one of the most appealing econometric approaches for managing these large dimensional data sets because they condense the common dynamics of the time series in a relatively small number of unobserved factors. This class of models has been successfully applied in economics, as documented by the surveys of Bai and Ng (2008) and Stock and Watson (2011).

A second characteristic of economic data sets is that they are usually collected with incomplete statistical information. In these data sets, some time series are available over a diminished time span and show missing observations at the beginning of the sample since data collection by their sources started at different dates. This so-called ragged edge problem also arises at the end of the sample due to the differences in publication lags among the variables that characterizes the flow of economic information in real time.

To handle unbalanced panels, Stock and Watson (2002) introduce an important computational contribution by extending the factor model with an iterative Expectation-Maximization (EM) algorithm. In sum, the algorithm is initialized with an estimate of the factors from a balanced panel where missing data are replaced by initial guesses. Then, the factors are used to provide an estimate of the missing observations. The process is iterated until the estimates do not change substantially. In practice, the EM algorithm can also handle outliers if they are replaced by missing values in the data set. In empirical applications, Bernanke and Boivin (2003), Angelini, Henry and Marcellino (2006), Schumacher and Breitung (2008), and Marcellino and Schumacher (2010) deal with missing observations by applying an EM algorithm in a principal components framework satisfactorily.[1]

The last characteristic of the economic data sets is that they are typically drawn with a remarkable clustering structure. For example, group structures appear in macroeconomic indicators because, usually, either they contain sectoral splits of different categories or they are classified into real and nominal data. Groups structures also appear in international data sets, which usually contain country regional-specific data that exhibit strong cross-correlation among the series in the same region and separated from the cross-correlation observed in other regions.

Within this framework, it is reasonable to think that a model that uses group-specific factors will describe group comovements better than a model that uses only global common factors. Factor models with block structures have been studied, first, from data sets organized into blocks using a priori information. Typically, these proposals consider a panel of variables with a set of common pervasive factors, which affect all the variables in all groups, and group-specific pervasive factors, which affect the variables only in a specific group. Examples are Gregory et al. (1997), Kose et al. (2003), Crucini et al. (2011), Hallin and Liška (2011), Ando and Bai (2015), Breitung and Eickmeier (2016), Rodriguez-Caballero and Ergemen (2017) and Choi, et al. (2018), among

[1] In recent contributions, McCracken and Ng (2021) and Jin, Miao and Su (2021) propose extensions of the EM algorithm on principal component frameworks. A comparison of these approaches with Stock and Watson (2002) is left for further research.

others. However, Ando and Bai (2016) contribute to this literature by developing a new procedure that is designed to explore group structures in large factor models, with the distinctive feature of endogenously assigning the variables to groups, determining the number of groups, and estimating the group-specific factors.[2] Ando and Bai (2017) extend this proposal to allow for both global common factors and group-specific factors.[3]

In spite of the aforementioned advances to handle economic data sets, the existing literature does not provide yet any comprehensive approach to deal with large panels with group structure and unknown membership containing missing observations, which is the norm rather than the exception in empirical applications. This limits the potential of factor models to characterize the co-movement of economic variables considerably. On the one hand, taken into account the missing observations in standard factor models with an EM algorithm as in Stock and Watson (2002) would imply that the iterative updated estimates of the missing observations for a variable i that belongs to group g is provided by the expectation conditional on only global factors, which is clearly misleading under the premise of group heterogeneity.

On the other hand, the algorithms of Ando and Bai (2016, 2017) require data sets where all the panel members are observed every time period. Thus, taken into account group membership implies removing the observations for which there is no full set of data. This entails computing inferences from shorter cross sections (when removing complete columns of the panel) and/or shorter time series (when removing complete rows of the panel). This could potentially lead to high costs either due to loss of efficiency, when omitting data with highly valuable information, or due to bias, when the missing are not at random. In addition, handling missing data is particularly challenged for real-time forecasters as the real-time data routinely exhibit end-of-sample ragged edge problems since some variables are observed with longer delays than others.

Aware of these limitations, we contribute to the existing literature on factor models by bringing together these two strands of the literature. In particular, we develop an algorithm that deals with unbalanced panels of data that exhibit cluster structure and potentially contain outliers. Our proposal relies on the Stock-Watson EM algorithm to account for missing data in conjunction with the Ando-Bai algorithm that determines the number of groups, the group membership, the number of global factors and the number of group-specific factors automatically. In particular, we rely on a C_P-type criterion for model selection that is minimized through the implementation of a recursive algorithm, which endogenously deals with missing data by using a group-specific EM algorithm. Based on this criterion, we develop a method to choose among three factor model alternatives: only global factors as in Stock and Watson (2002), global and group-specific factors as in Ando and Bai (2017), and only group-specific factors as in Ando and Bai (2016).

By means of several Monte Carlo experiments, which are designed to capture some basic data problems that characterize economic data sets, we evaluate the ability of the proposed algorithm to choose models with the appropriate number of groups, the true number of group-specific factors, the convenient group membership and a suitable inference of the missing data. Among the data

[2] Ando and Bai (2016) do not consider common factors affecting all the variables. However, they allow for a large number of potential observable factors.

[3] Although not pursued in this paper, Alonso et al. (2020) propose an alternative procedure to build dynamic factor models with a cluster structure that allow for common global factors and group-specific factors.

generating processes, we include non-homoskedastic errors with cross sectional dependence, errors that have some serial correlations and factors with temporal dependence. Overall, our results indicate that, despite the inclusion of missing observations to simulate unbalanced data sets, there are no significant differences between the cluster abilities of our proposal and the group factor models of Ando and Bai (2016, 2017), which requires balanced data sets. In addition, we show that our group-specific inferences of missing data outperforms the imputations of the missing data conditional on the full set of common factors as in the standard EM algorithm of Stock and Watson (2002).

Furthermore, we perform an empirical application that illustrates how the method developed in this paper can be used by practitioners seeking to model an unbalanced large set of observed variables in terms of a smaller number of underlying latent factors exhibiting group structures with unknown membership. For this purpose, we employ the novel quarterly FRED-QD data set of McCracken and Ng (2021), which comprises a set of 248 economic indicators of the United States. The data set aims to provide researches access to a regularly updated version of the Stock and Watson (2012) data, which is the quarterly version of the monthly data used in Stock and Watson (2002). As McCracken and Ng (2021) acknowledge, the data set contains missing values in 38 out of 248 cases, which implies that about 15% of the indicators are incomplete because their samples start late. In addition, these authors point out 30 outliers in the FRED-QD data set, most of which are found in bank reserves variables.

Our results indicate that the quarterly FRED-QD data set is better characterized by a factor model that admits only group-specific factors than by factor models that additionally have some pervasive factors common to all variables or by a factor model with only common factors. In particular, we find two distinctive groups of economic indicators that we interpret as the real activity group and the nominal group. The selected factors explain fifty percent and almost forty per cent of the variance of the indicators that belong to each group, respectively. We document the high empirical reliability of the group-specific EM algorithm at filling in the gaps of the missing observations of both nominal and real activity indicators. In addition, our results suggest a very promising role of the estimated factors in the study of the US business cycle chronology.

Finally, we evaluate the forecasting performance of the method outlined in this paper through an out-of-sample forecasting exercise from 2006:1 to 2019:4. In this experiment, the group-specific factor model has resulted in meaningful forecasting improvements over the Stock-Watson factor model. The improvements are uniformly observed over the forecasts of the 248 economic indicators, and they become substantially large for many real economic activity indicators. In spite of this promising out-of-sample results, which omits data revisions, the forecast performance might be reassessed in the future as more real-time data sets become available and real-time assessments become feasible.

The rest of the paper is organized as follows. Section 2 describes the model selection algorithm and the estimation process. Section 3 provides a description of the Monte Carlo experiment to analyze the ability of the proposal to select the correct number of groups, the group membership, the number of factors and the inference on missing data. Section 4 examines the ability of this new algorithm to transform the information content of the FRED-QD data set into a smaller number of

group-specific factors, which are used to compute business cycle inferences and to perform forecasts. Section 5 concludes and proposes some lines of further research.

2 Grouped factor models with missing data

2.1 Model formulation

Let $t = 1, ..., T$ be the time index, S be the number of groups, $N_1, ..., N_S$ be the number of indicators in each group observed for a sample of T observations, and $N = \sum_{s=1}^{S} N_s$ be the total number of indicators, with $p_s = N_s/N$ being the proportion of indicators in each group. Let us assume that the number of groups is finite and independent of N and T, and that each group contains a minimum p_{min} and maximum p_{max} proportion of indicators, with $0 < p_{min} < p_s < p_{max} < 1$ and $s = 1, \ldots, S$.

The value of the i-th indicator x_{it}, observed at time t, belonging to group $g_i \in \{1, ..., S\}$ from a collection of stationary time series $\{x_{it}\}_{i=1,...,N}$ that admits a grouped factor representation, is expressed as follows

$$x_{it} = f'_{ct}\lambda_{ci} + f'_{g_it}\lambda_{g_ii} + \epsilon_{it}. \tag{1}$$

The $r \times 1$ vector λ_i collects the factor loadings that measure the unknown sensitivity of x_i to the unobservable common factors r factors collected in the $r \times 1$ vector f_{ct}, which affect all indicators in all groups. The $r_{g_i} \times 1$ vector λ_{g_ii} collects the factor loadings that measure the unknown sensitivity of x_i to the unobservable group-specific r_{g_i} factors collected in the $r_{g_i} \times 1$ vector f_{g_it}, which affect the indicators only of group g_i.[4] The common factors and the group-specific factors are orthogonal, whereas correlations between factors of different groups is allowed, although they cannot be perfectly correlated. In this expression, ϵ_{it} is the unit-specific error, which is independent of f_{ct} and f_{g_it} for all i and t. To begin with, we assume that $E(\epsilon_{it}) = 0$ and $var(\epsilon_{it}) = \sigma_i^2$, and that it is independent over i and t.[5]

Although the group membership is unknown, the time series are labeled through the unobservable state indicator g_i in the whole sequence of realizations, which are collected in $G = (g_1, \cdots, g_N)$. For example, $g_i = s$ indicates that the time series x_i belongs to group s. It is useful to represent the conditional factor representation of the T observations of this i-th indicator in matrix notation

$$x_i = F_c\lambda_{ci} + F_s\lambda_{si} + \epsilon_i, \tag{2}$$

where $x_i = (x_{i1}, x_{i2}, \cdots, x_{iT})'$; $F_c = (f_{c1}, f_{c2}, \cdots, f_{cT})'$ and $\lambda_{ci} = \left(\lambda_{ci}^1, \lambda_{ci}^2, \ldots, \lambda_{ci}^r\right)'$ are the $T \times r$ matrix of common factors and the $r \times 1$ vector of common factor loadings; $F_s = (f_{s1}, f_{s2}, \cdots, f_{sT})'$ and $\lambda_{si} = \left(\lambda_{si}^1, \lambda_{si}^2, \ldots, \lambda_{si}^{r_s}\right)'$ are the $T \times r_s$ matrix of group-specific factors and the $r_s \times 1$ vector of group-specific factor loadings for group s; and $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{iT})'$.

Given the group membership G and the common factor structures $F_c\lambda_{ci}$, we define the variable $z_i^s = x_i - F_c\lambda_{ci}$ for each group $s = 1, \ldots, S$. If we collect the N_s variables z_i^s that belongs to group

[4] To eliminate scale effects, the time indicators are standardized to have zero mean and unit variance.

[5] In the simulations, we examine the potential effects of cross-sectional dependent and serially correlated errors as well as the effects of serially correlated factors.

s in the $T \times N_s$ matrix $Z_s = (z_1, \ldots, z_{N_s})$, we can obtain the principal components' estimate of the group-specific factors, $\hat{F}_s$, subject to the normalization $F_s'F_s/T = I_{r_s}$, as $\sqrt{T}$ times the eigenvectors corresponding to the r_s largest eigenvalues of the $T \times T$ matrix Z_sZ_s'. If we collect the group-specific factor loadings of this group in the $r_s \times N_s$ matrix $\Lambda_s = (\lambda_{s1}, \ldots, \lambda_{sN_s})$, the estimates of the factor loadings can be obtained as $\hat{\Lambda}_s = \hat{F}_s'Z_s/T$. In addition, given the group membership G and the group-specific factor structures $F_s\lambda_{si}$, we define the variable $z_i^c = x_i - F_s\lambda_{si}$ for each group $s = 1, \ldots, S$. If we collect the N variables z_i^c in the $T \times N$ matrix $Z_c = (z_1, \ldots, z_N)$, we can obtain the principal components' estimate of the common factors, $\hat{F}_c$, subject to the normalization $F_c'F_c/T = I_r$, as $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix Z_cZ_c'. If we collect the common factor loadings in the $r \times N$ matrix $\Lambda_c = (\lambda_{c1}, \ldots, \lambda_{cN})$, the estimates of the factor loadings can be obtained as $\hat{\Lambda}_c = \hat{F}_c'Z_c/T$.

To consider missing values in the data, we assume that not all the observations are available in x_i. Then, we call x_i^{obs} the $T^{obs} \times 1$ vector that contains the observations available for variable i, which is only a subset of x_i when $T > T^{obs}$. As we will show below, a key step during the EM iterations is the mapping from the full set x_i to the observed data x_i^{obs}. For a known $T^{obs} \times T$ matrix A_i, this relationship is stated by

$$x_i^{obs} = A_i x_i, \tag{3}$$

where A_i is the identity matrix of size T for which the $T - T^{obs}$ rows corresponding to the missing observations in x_i are removed. When x_i is fully observed, A_i is simply the identity matrix.

2.2 Model estimation

This section presents an estimation method for the model that combines the EM algorithm proposed by Stock and Watson (2002) to deal with missing observations and the algorithm advocated by Ando and Bai (2017) to identify group membership and estimate of common and group-specific factors.[6] Given the number of groups, S, the number of common factors, r, and the number of group-specific factors, r_s, with $s = 1, \ldots, S$, the least squares estimator of $F = \{F_c, F_1, \ldots, F_S\}$ and $\Lambda = \{\Lambda_c, \Lambda_1, \ldots, \Lambda_S\}$ is defined as the minimizer of the objective function

$$V(F, \Lambda) = \sum_{i=1}^{N} \sum_{t=1}^{T} I_{it} \left(x_{it} - f_{ct}'\lambda_{ci} - f_{g_it}'\lambda_{g_ii}\right)^2, \tag{4}$$

where $I_{it} = 1$ if x_{it} is observable at t and 0 otherwise.

Our algorithm is described in Algorithm 1. The first step requires initial (naive) guesses of the missing observations.[7] These guesses together with the available data comprise a balanced data set from which we estimate initial factors and loadings using standard principal components for a fixed number of common factors. Now, the standard version of the EM algorithm can be applied to the balanced panel to provide initial estimates of the missing values of the time series and a first guess of the balanced data set $x^{(0)}$ ignoring the group-specific factor structures. Next,

[6] It is worth emphasizing that, in contrast to Ando and Bai (2017), we do not allow for observed factors.

[7] Initially, we replace missing observations by random draws from the standard normal distribution.

Algorithm 1 Model estimation: minimization of $V(F, \Lambda)$

Set S, r, $\{r_s\}_{s=1}^S$ and a naive guess of missing values

Initialization $x_{it}^{(0)}$, $G^{(0)}$, $\{F_c^{(0)}, \Lambda_c^{(0)}\}$, and $\{F_s^{(0)}, \Lambda_s^{(0)}\}_{s=1}^S$

while $m \leq M$ **do** (*)

Update membership $G^{(m)}$ given $x_{it}^{(m-1)}$, $\{F_c^{(m-1)}, \Lambda_c^{(m-1)}\}$ and $\{F_s^{(m-1)}, \Lambda_s^{(m-1)}\}_{s=1}^S$ using (5)

for $j \leq J$ **do** (*)

Expectation of $x_{it}^{(m)}$ given $G^{(m)}$, $\{F_c^{(m-1)}, \Lambda_c^{(m-1)}\}$ and $\{F_s^{(m-1)}, \Lambda_s^{(m-1)}\}_{s=1}^S$ using (6)

Maximization 1 PC of $\{F_c^{(m)}, \Lambda_c^{(m)}\}$ given $G^{(m)}$, $x_{it}^{(m)}$ and $\{F_s^{(m-1)}, \Lambda_s^{(m-1)}\}_{s=1}^S$

Maximization 2 PC of $\{F_s^{(m)}, \Lambda_s^{(m)}\}_{s=1}^S$ given $G^{(m)}$, $x_{it}^{(m)}$ and $\{F_c^{(m)}, \Lambda_c^{(m)}\}$

end for

end while

(*) Stop when convergence is achieved. PC refers to Principal Components.

we require an initialization of the group membership G. For the sake of simplicity, we use the K-means algorithm, which separates the data set into S clusters and provides an initial membership $G^{(0)} = \left(g_1^{(0)}, \cdots, g_N^{(0)}\right)$. Finally, we obtain starting values of $\{F_c^{(0)}, \Lambda_c^{(0)}\}$ by principal components on x_{it}, and $\{F_s^{(0)}, \Lambda_s^{(0)}\}$ for $s = 1, \ldots, S$ by group-specific principal components on $x_{it} - F_c\lambda_{ci}$.

In a given iteration (m), the second step consists on providing an optimal assignment for each indicator $G^{(m)} = \left(g_1^{(m)}, \cdots, g_N^{(m)}\right)$ given the common and group-specific factor structure of the previous iteration, $\{F_c^{(m-1)}, \Lambda_c^{(m-1)}\}$ and $\{F_s^{(m-1)}, \Lambda_s^{(m-1)}\}$ for $s = 1, \ldots, S$. For this purpose, we assign each indicator to the group with best in-sample fit

$$g_i^{(m)} = arg \min_{s \in \{1,\ldots,S\}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - f_{ct}^{\prime(m-1)} \lambda_{ci}^{(m-1)} - f_{st}^{\prime(m-1)} \lambda_{si}^{(m-1)}\right)^2. \tag{5}$$

This implies that the time series x_i is assigned to group s if it minimizes the sum of squared residuals among the S possible groups.

In the third step, given the group membership $G^{(m)}$, the common factor structures, $\{F_c^{(m)}, \Lambda_c^{(m)}\}$, and group-specific factor structures, $\{F_s^{(m)}, \Lambda_s^{(m)}\}$ for $s = 1, \ldots, S$, are updated with an iterative group-specific EM algorithm by alternating the expectation of $x_i^{(m)}$ with respect to $\{F_c^{(m)}, \Lambda_c^{(m)}\}$ and $\{F_s^{(m)}, \Lambda_s^{(m)}\}$ and the optimization of $\{F_c^{(m)}, \Lambda_c^{(m)}\}$ and $\{F_s^{(m)}, \Lambda_s^{(m)}\}$ given $x_i^{(m)}$. In the expectation step of the j-th iteration, missing observations for each variable are updated by the expectation of x_{it} conditional on the observations available for i, and the common and group-specific factors and loadings from the previous iteration. If we define $F\lambda_i^{(m(j-1))} = F_c^{(m(j-1))}\lambda_{ci}^{(m(j-1))} + F_s^{(m(j-1))}\lambda_{si}^{(m(j-1))}$, the missing observations are updated by using

$$x_i^{(mj)} = F\lambda_i^{(m(j-1))} + A_i'(A_iA_i')^{-1}\left(X_i^{obs} - A_iF\lambda_i^{(m(j-1))}\right), \tag{6}$$

where $s = 1, \ldots, S$. Then, the elements of the estimated balanced panel are constructed as $\hat{x}_{it}^{(mj)} =$

x_{it} if x_{it} is observed and $\hat{x}_{it}^{(mj)} = f_{ct}^{\prime(m(j-1))}\lambda_{ci}^{(m(j-1))} + f_{g_it}^{\prime(m(j-1))}\lambda_{g_ii}^{(m(j-1))}$ otherwise.

With the updated balanced data set in hand, the maximization step consists on reestimating the common factors and loadings, $\{F_c^{(mj)}, \lambda_{ci}^{(mj)}\}$, by principal components on $\hat{x}_{it}^{(mj)} - f_{g_it}^{\prime(m(j-1))}\lambda_{g_ii}^{(m(j-1))}$. In addition, the maximixation step reestimates the group-specific factor and loadings, $\{F_s^{(mj)}, \lambda_{si}^{(mj)}\}$ by principal components on $\hat{x}_{it}^{(mj)} - f_{ct}^{\prime(mj)}\lambda_{ci}^{(mj)}$ for each group $s = 1, \ldots, S$. The expectation and maximization steps are iterated until the maximum percentage change of the variables' estimates is smaller than a convergence tolerance δ_{EM}. This algorithm provides both, estimates of the missing values in the time series and estimates of the common and group-specific factors and loadings.

To obtain the minimizer of $V(F, \Lambda)$, we iterate the second and third steps until convergence. In practical implementations, we stop the iterations once

$$100 * \left(V(F^{(m)}, \Lambda^{(m)}) - V(F^{(m-1)}, \Lambda^{(m-1)}) \right) / V(F^{(m-1)}, \Lambda^{(m-1)}) < \delta_V \tag{7}$$

for some absolute convergence tolerance δ_V.

Although we does not account for observed factors, this multi-level factor structure that admits common and group specific factors generalizes Ando and Bai (2017) to deal with missing observations. In a similar vein, our proposal can be simplified easily to generalize Ando and Bai (2016), who advocated a model with completely separated group-specific factor structures in different clusters without pervasive factors common to all variables. In this simplified algorithm, which we call Algorithm 1′, the term $F_c\lambda_{ci}$ does not appear in expression (4). In addition, the estimation method outlined in Algorithm 1 is still valid, although Algorithm 1′ does not include neither $\{F_c, \Lambda_c\}$ nor the step Maximization 1.

2.3 Model selection

We have assumed so far that the number of groups S, the number of common factors, r, and the number of group-specific factors, $\{r_1, \ldots, r_S\}$ are known. In practice, we are faced with the problem of estimating these quantities from the data. In this paper, we adapt the approach put forward by Ando and Bai (2017) to deal with this model specification uncertainty.

Following these authors, we propose to choose these unknown parameters as the minimizers of the Panel Information Criterion (PIC)

$$\begin{aligned} PIC(s, r, r_1, \ldots, r_s) = & \frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} I_{it} \left(x_{it} - f_{ct}^{\prime}\lambda_{ci} - f_{g_it}^{\prime}\lambda_{g_ii} \right)^2 \\ & + \sigma^2 r \left(\frac{T+N}{TN} \right) log(TN) \\ & + \sigma^2 \sum_{s=1}^{S} r_s \left(\frac{N_s}{N} \right) \left(\frac{T+N_s}{TN_s} \right) log(TN_s). \end{aligned} \tag{8}$$

Thus, the model selection consists on choosing the minimizers of the distances between the actual data and the factor estimation, measured by the sum squared error over the observed data, $V(F, \Lambda)$, which is penalized by redundant model flexibility. In particular, the second and third terms penalize the overestimation on the number of common and group-specific factors, where σ^2

provides scaling for the penalty term.[8]

In practice, the number of groups, the number of common factors and the number of group-specific factors can be determined with the following model search algorithm, as Algorithm 2 describes. The first step requires fixing a maximum number of groups, S_{max}, a maximum number of common factors, r_{max}, and maximum number of group-specific factors, $r_{max,j}$. Thus, the potential number of groups are $\{S_1, \ldots, S_{max}\}$, the potential number of common factors are $\{r_1, ..., r_{max}\}$, and the potential number of group-specific factors for group s are $\{r_{1,s}, ..., r_{max,s}\}$, with $s = 1, \ldots, S$ and $S = S_1, \ldots, S_{max}$.[9] For the first value of the number of groups, S_1, optimize the number of common and group specific factors by minimizing PIC. Now, repeat the second step for $\{S_2, \ldots, S_{max}\}$ and choose the number of groups (and its corresponding number of common and group-specific factors) that minimizes the value of PIC.

Algorithm 2 Model selection: minimization of $PIC(s, r, r_1, \ldots, r_s)$

Set $S_{max}, r_{max}, \{r_{max,s}\}_{s=1}^{S}$ for $S = S_1, \ldots, S_{max}$
while $S \leq S_{max}$ **do**
  for $r \leq r_{max}$ and $r_s \leq r_{max,s}$ **do**
    Run Algorithm 1
  end for
  Select r^* and $\{r_s^*\}_{s=1}^{S}$ that minimize (8). Store the resulting $PICs$ for each $S \leq S_{max}$
end while
Choose S^* that minimize the stored $PICs$

This algorithm generalizes the model selection algorithm proposed by Ando and Bai (2017) to deal with missing observations. Again, our proposal can be simplified easily to generalize the model selection algorithm advocated by Ando and Bai (2016), which is designed for a factor model that admits only group-specific factors. The simplified algorithm, which we call Algorithm 2′, omits the second term in the right side of expression (8) and expressions r_{max} and r^* are dropped from the algorithm.

In empirical applications, the framework described above is useful to choose among three alternative factor structures: only common factors, as in Stock and Watson (2002), common and group-specific factors, as in Ando and Bai (2017), or only group specific-factors, as in Ando and Bai (2016). For a given data set, the discriminating method consists on running Algorithm 1 and Algorithm 2, which estimate a factor model that admits common and group-specific factors, leading to an optimal value of the model selection criterion, $PIC_c(S_c^*, r_c^*, r_{c1}^*, \ldots, r_{cS_c^*}^*)$. In addition, the method implies running the modifications Algorithm 1′ and Algorithm 2′, which do not allow for common factors, achieving an optimal value of the model selection criterion, $PIC_g(S_g^*, r_{g1}^*, \ldots, r_{gS_g^*}^*)$.

If the optimal number of groups is $S_c^* = 1$, then we infer that the data generating process does not admit a grouped factor structure because the factors are common to all the time series. If $S_c^* > 1$, then we infer that the data generating process admits common and group-specific factors

[8] In practical implementations, σ^2 can be obtained as the mean squared error under the maximum number of groups and the maximum number of common and group-specific factors.

[9] In the applications developed in this paper, we set $S_1 = 1$, which implies that $S \in \{1, 2, \ldots, S_{max}\}$.

when $PIC_c(S_c^*, r_c^*, r_{c1}^*, \ldots, r_{cS_c^*}^*) < PIC_g(S_g^*, r_{g1}^*, \ldots, r_{gS_g^*}^*)$. However, we propose a factor structure with factors that are common only within some clusters of variables excluding pervasive factors common to all variables when $PIC_g(S_g^*, r_{g1}^*, \ldots, r_{gS_g^*}^*) < PIC_c(S_c^*, r_c^*, r_{c1}^*, \ldots, r_{cS_c^*}^*)$.

3 Simulation experiments

In this section, we set up several Monte Carlo experiments to examine the small sample performance of our proposal to determine the correct number of clusters, to provide the right number of common and group-specific factors, to achieve error-free group membership and to perform accurate estimates of the unobserved missing data over that of the Stock and Watson (2002) EM algorithm. To facilitate comparisons, we design the data-generating processes following Ando and Bai (2016, 2017).

In our simulation experiments, we generate a total of 500 sets of N_s idiosyncratic components ϵ_i of length T in each group $s = 1, 2, \ldots, S$. Following Ando and Bai (2017), we set the number of groups $S = 3$, only one common factor, $r = 1$, the same number of group-specific factors in each group, $r_1 = r_2 = r_3 = 3$, and the same number of time series in each group, $N_1 = N_2 = N_3 = N$. The common factor f_{ct} is drawn from a uniform distribution on [0,1], and each element of the factor-loading matrix Λ_c follows uniform [-2,2]. Each element of the $r_s \times 1$ group-specific factors f_{st} is drawn from $N(0, 1)$, while, to generate group heterogeneity, each element of the $r_s \times 1$ vector of group-specific factor loadings λ_{si} is drawn from $N(0.5 \times s, 1)$, for $s = 1, 2, 3$.

To examine the relative performance of the model to the number of variables and the number of observations, we generate times series with grouped factor structure $x_i = F_c\lambda_{ci} + F_s\lambda_{si} + \epsilon_i$ for several combinations of (N, T): $(50, 150)$, $(100, 150)$, $(50, 250)$, $(100, 250)$. These settings are similar to the standard data sets used in economics and includes a combination that is close to the cross-section and time-series dimensions of our empirical application.

Although we first generate the data from the processes described above, we replace a certain fraction of the data by missing values. To examine the effect of the magnitude of these missing values in the model's performance, we consider that fractions $P = 10\%$ and $P = 20\%$ of the time series present a certain amount of missing data. For these series, we contemplate amounts of $M = 50$ and $M = 83$ missing data (out of a total of T observations).[10] Again, these settings replicate standard situations in economic applications, including our empirical example.

For the sets of simulations of the different data generating processes, we apply the grouped factor model with missing data outlined in this paper, allowing for a maximum number of four groups and a maximum number of eight common factors and eight group-specific factors. The iterative method that computes successive approximations to the final model estimates stops when the magnitude of the percentage difference between two consecutive values of $V(F, \Lambda)$ is below the convergence tolerance $\delta_V = 10^{-3}$. In the same way, the group-specific EM algorithm stops when

[10] In the (N, T, P, M) combinations $(50, 150, 10, 50)$, $(100, 150, 10, 50)$, $(50, 250, 10, 83)$ and $(100, 250, 10, 83)$, the missing data represent 3.3% of the total data. This percentage rises to 6.7% in the combinations $(50, 150, 20, 50)$, $(100, 150, 20, 50)$, $(50, 250, 20, 83)$ and $(100, 250, 20, 83)$. Following a reviewer's suggestion, we enlarged the proportion of missing values and we find a slight deterioration for up to 20% of missing values and a moderate deterioration for 20% of missing values. The detailed results are available from the authors upon request.

the maximum percentage change of the variables' estimates is smaller than $\delta_{EM} = 10^{-3}$.

Our Monte Carlo experiments are designed to focus on the effects of heteroskedasticity, cross-correlated and serially-correlated errors and on the effects of allowing dynamics on the factors. However, for comparative purposes, in the first data-generating process, the noise term is homoskedastic and serially uncorrelated. Thus, each element ϵ_{it} is drawn independently from a normal distribution with mean 0 and variance $\sigma^2 = 1$. Panel A of Table 1 reports the results of the simulations for the combinations of (N, T, P, M), which are outlined in the first four columns.

The next two columns of the table display the percentages of under- (U) and correct (C) identification of the number of groups. The figures reported in the table indicate that the clustering ability of the panel data model with factor structure does not deteriorates significantly when the data sets contain missing observations because the high percentages of times that the model determines the correct number of clusters are comparable to those obtained by Ando and Bai (2016, 2017). Remarkably, the model accuracy diminishes only slightly when the number of missing data increases from 3.3% to 6.7% of the total observations, although the proposed method continues providing very good results.

The figures displayed in the next six columns of the table show the percentages of under- and correct identification of the true number of factors, r_s, with $s = 1, 2, 3$, and are calculated under the condition that the number of groups is correctly selected. As shown in the table, the proposed method provides accurate results selecting the true number of common and group-specific factors when there are missing observations. Again, we only find a small deterioration when the percentage of time series with missing data increases, but, even in this case, the reported figures show that the true number of group-specific factors is estimated very well.[11]

The last six columns evaluate the model accuracy to perform a correct classification of the generated time series into the three groups. The outcome of interest used to judge the effectiveness of the model is the proportion of time series for each group that are classified incorrectly. Thus, for a given group, we compute the Over (Under) classification rate as 100 times the number of time series that exceeds (are lower than) the correct number of series for this group along the 500 simulations over the total number of time series that has been generated for this group. A quick glance of the table indicates that all the misclassification rates lie within a very tight range of 0.7% and 2.2%. This result shows that our method identifies group membership with large accuracy.

Before moving to more complex data generating processes, we asses the performance of the model estimation and model selection algorithms presented in the paper to infer from the data if the data generating process admits common and group-specific factors or only group-specific factors. For this purpose, we set $S = 3$ and we generate 500 data sets from $x_i = F_c\lambda_{ci} + F_s\lambda_{si} + \epsilon_i$ and 500 data sets from $x_i = F_s\lambda_{si} + \epsilon_i$. For these two groups of data sets, we run the two pairs of algorithms Algorithm 1 and Algorithm 2 (common and group-specific factors) and Algorithm $1'$ and Algorithm $2'$ (only group-specific factors) that we describe in Section 2. We obtain that $PIC_c(3, r_c^*, r_{c1}^*, r_{c2}^*, r_{c3}^*) < PIC_g(3, r_{g1}^*, r_{g2}^*, r_{g3}^*)$, which means that we select a factor model with common and group-specific factors, in 91% of the simulations that allow for common and group-

[11] Following a reviewer's suggestion, we also tested for the performance of the methodology when the missing data are randomly placed and we found that the results were robust. These results are available from the authors upon request.

specific factors. In addition, for all the data generating processes that allow for factor models with only group-specific factors we obtain $PIC_g(3, r^*_{g1}, r^*_{g2}, r^*_{g3}) < PIC_c(3, r^*_c, r^*_{c1}, r^*_{c2}, r^*_{c3})$, which implies that we select a factor model with only group-specific factors. These results suggest that our procedure is able to select the true data generating process from the observed data with very high accuracy.

Next, we investigate the performance of the methodology when the noise term is nonhomoskedastic and cross-sectionally correlated. In this case, the data-generating process for the errors is $\epsilon_{it} = 0.9e^1_{it} + \delta_t 0.9e^2_{it}$, where $\delta_t = 1$ if t is odd and zero if t is even. The $N \times 1$ vectors $e^1_t = (e^1_{it}, \ldots, e^1_{Nt})'$ and $e^2_t = (e^2_{it}, \ldots, e^2_{Nt})'$ are independent and follow multivariate normal distributions $N(0, \Sigma)$, with $\sigma_{ij} = 0.3^{|i-j|}$. The figures reported in Panel B of Table 1 suggest that nonhomoskedastic errors and cross-sectionally correlation introduce some performance deterioration for all the combinations (N, T, P, M), although the performance losses are not significant, especially in the case of reasonably large numbers of time series.

The third data-generating process contains idiosyncratic errors that exhibit some serial and cross-sectional correlations. To this end, the noise term is now generated as $\epsilon_{it} = e_{it} + 0.2\epsilon_{it-1}$, where $t = 1, \ldots, T$. The $N \times 1$ vector $e_t = (e_{it}, \ldots, e_{Nt})'$ follows a multivariate normal distribution $N(0, \Sigma)$, with $\sigma_{ij} = 0.3^{|i-j|}$. Panel C of Table 1 shows that the performance also deteriorates somewhat, although the numbers reported in the table confirm that our method continues performing well regardless of whether we focus on the determination of the number of groups, on factor extraction or on group membership.

The fourth data generating process focuses on allowing for temporal dependence of the factors. In particular, the common factor is assumed to be an autoregressive process of order one, $f_{ct} = \phi f_{ct-1} + u_t$, where u_t is an independent Gaussian error term with mean 0 and variance 1. In addition, each element of the $r_s \times 1$ group-specific factors is also assumed to follow an autoregressive processes of order one, $f^j_{st} = \phi f^j_{st-1} + u^j_{st}$, where u^j_t is an independent Gaussian error term with mean 0 and variance 1 and $j = 1, \ldots, r_s$. In the simulations, we set $\phi = 0.3$. The figures reported in Panel D of Table 1 suggest that serially correlated factors implies a modest finite-sample performance deterioration, which is similar to the case of idiosyncratic errors that exhibit serial correlations.[12]

In sum, these results suggest that the procedure described in this paper to deal with unbalanced panels of data sets that exhibit a group factor structure is robust to the data problems that are typically faced in empirical applications, especially for reasonably large data sets.

Finally, to assess the performance of our proposal to estimate the unobserved missing data, we also carry out the following simulation experiment. For each combination (N, T, P, M), we use the data-generating model with homoskedastic and serially uncorrelated noise terms to simulate data from a factor model with group structure and replace M observations by missing values in a P proportion of the series. Then, we impute the unobserved data both with the standard EM advocated by Stock and Watson (2002) and with the group-specific extension of the EM algorithm that we propose in this paper. For each simulation, we compute the mean squared differences between these two estimates of the missing values and the true data that were removed once

[12]Stock and Watson (2002) also find some deterioration in the performance of factor models for data generating processes that allow for serial correlation in the factor process. However, they find that this has little effect on the quality of the estimators and forecasts.

simulated.

For each combination (N, T, P, M), Table 2 displays in the last column, which has been labeled as *better*, the percentage of times (out of the total simulations) that Algorithm 1 and Algorithm 2 reached lower mean squared error than the standard Stock and Watson (2002) EM algorithm. The figures reported in the table indicate that our proposal outperforms the standard EM algorithm substantially when the data exhibit grouped factor structure. This indicates that our algorithm takes advantage of the group similarity when imputing the values of the unobserved data points.[13]

4 Empirical application

In this section, we assess the effectiveness of our proposal to handle a data set of time series that are available over different spans, that contain outliers and that exhibit a remarkable clustering structure. For this purpose, we focus on the Federal Reserve Economic Data (FRED-QD), which updates the Stock and Watson (2012) data set in real time. The data consists of 248 US quarterly economic indicators compiled in an easily downloadable way by the Research division of the Federal Reserve Bank of St. Louis and available from its website.[14]

4.1 In-sample analysis

The effective sample of the latest available vintage ranges from the third quarter of 1959 to the last quarter of 2019. However, 38 out of the 248 indicators (15.3%) are not available for the entire sample, which implies 2,102 missing observations out of the 60,264 potential figures (3.5%). In contrast to Ando and Bai (2016, 2017), we do not require removing indicators with missing observations or shortening the sample period. In addition, McCracken and Ng (2021) documented 30 outliers that we also treat as missing.[15]

Prior to enter into the model, each series was transformed to be approximately integrated of order zero using the transformation codes provided by McCracken and Ng (2021). Typically, real activity variables were transformed to growth rates, interest rates were transformed to first differences, and prices were transformed to first differences of rates of inflation. Then, we standardize all series to have zero mean and unit variance.

We apply the proposed model selection criteria outlined in the pairs of algorithms Algorithm 1 and Algorithm 2 and Algorithm $1'$ and Algorithm $2'$ to decide first whether the data admits a standard factor structure as in Stock and Watson (2002), a grouped factor with common and group-specific factors or a grouped factor with only group-specific factors. In addition, the algorithms allow us to choose the number of groups, from a maximum of 5 groups, and the number of factors, from a maximum of 8 factors.

Figure 1 shows the behavior of PIC as a function of the number of groups when the grouped

[13] This section focuses on group factor models that allow for common and group-specific factors. We repeated the analysis for group factor models that allow for only group-specific factors and we obtained that the results were qualitatively similar. These results are available from the authors upon request.

[14] The data are available at https://research.stlouisfed.org/econ/mccracken/fred-databases/

[15] As in Stock and Watson (2002), McCracken and Ng (2021), we define an outlier as an observation that deviates from the sample median by more than ten interquartile ranges.

factor model admits common and group-specific factors (red line, Algorithm 1 and Algorithm 2) and when the grouped factor model admits only group-specific factors (black line, Algorithm $1'$ and Algorithm $2'$). The two pairs of algorithms show that the model selection criterion is minimized for only two groups. However, the minimum PIC achieved for the model that admits common and group-specific factors is 0.797 whereas that for the model that admits only group-specific factors is 0.785. This result suggests that a factor model that allows for only group-specific factors is better suited for the Federal Reserve Economic Data FRED-QD than a factor model that allows for both common and group-specific factors.

The number of time series in these two clusters are 103 and 145, respectively. The estimated numbers of group-specific factors are three for the first group and four for the second group. In this context, it is standard to analyze to what extent the factors are able to capture the variability of the original series using the estimated eigenvalues. In our study, we find that the three factors estimated for the first group account for almost forty per cent of the variance of the indicators that belong to that group. In the second group, the four selected factors account for more than fifty per cent of the variance of the data.

To provide the group membership with economic meaning, it is worth mentioning that McCracken and Ng (2021) classified the indicators into 14 categories: National Income and Product Accounts (NIPA); Industrial Production; Employment and Unemployment; Housing; Inventories, Orders, and Sales; Prices; Earnings and Productivity; Interest Rates; Money and Credit; Household Balance Sheets; Exchange Rates; Other; Stock Markets; and Non-Household Balance Sheets. The classification of the economic indicators into one of the two groups emerges very distinctively from Figure 2. The figure displays the two-group membership, which is represented as floating dots in the ordinate axis, against the 248 economic indicators, which are grouped in the abscissa axis by category and ordered as in the Appendix of McCracken and Ng (2021).

Overall, our method classifies categories 4, 6, 10, 11, 12, 13, and the monetary indicators of category 9 into the first group. These categories refer to Housing, Prices, Household Balance Sheets; Exchange Rates; Other and Stock Markets. According to this classification, we call this group of economic indicators the nominal group.

We further analyze the relationship between this group and the nominal economic conditions by regressing the 248 indicators against the first factor of this group, which accounts for almost twenty percent of the variance of the first group. The top panel of Figure 3 displays the coefficients of determination of these regressions. Unequivocally, this factor loads mainly on prices (category 6). The bottom panel of this figure points out the high correlation between this this factor and the stationary transformation of CPI.

By contrast, categories 1, 2, 3, 5, 7, and 8 fall into the second group. These categories include real output, employment, real earnings and productivity. Quite interestingly, interest rate, and the credit variables of the ninth category belong to this group as well. As these indicators typically capture the broad movements in economic activity, we refer to this group as the real activity group.

Following the same reasoning as in the case of the nominal group, the top panel of Figure 4 plots the coefficients of determination of the regressions of the first factor of the real activity group against the entire panel of indicators. This points out the high correlation between the factor

and the NIPA indicators, industrial production, employment, unemploment, inventories, orders and sales. A visual inspection of the bottom panel of this figure reveals the high correspondence between the first factor of the second group and GDP growth rate, which constitutes an assessment of the US broad economic performance.

This result agrees with McCracken and Ng (2016), who suggest reorganizing the factor estimates into two groups: one for real activity, and one for nominal activity.[16] Posterior to factor estimation, these authors construct a diffusion index for the real activity by computing the common variations at low frequencies of the factor associated with the industrial production and employment series. They show that the evolution of this index is closely related to the NBER-referenced business cycle. In the same way, they construct a diffusion index for the nominal activity from the factors related with price, spread and interest rate variables. They show that the index seems to line up with price pressure inflation expectations in the past five decades.

In a similar vein, previous empirical works have pre-classified economic data sets into nominal and real data to extract indexes of US inflation and real activity from two separate factor models. While not claiming to be exhaustive, Aruoba and Diebold (2010) use five indicators of real activity and six inflation indicators to compute real activity and inflation indexes, which are used to examine real activity and inflation together over the cycle. Leiva-Leon (2014) extends the previous approach by considering a unified factor model where the two separate factors are extracted from the same two sets of real and nominal indicators. Our results are consistent with this classification although, in contrast to these authors, the two separate groups and the indicators in our large data set are assigned to each group have been selected in a systematic and statistically optimal manner.

Following McCracken and Ng (2016), we further examine the business cycle information that can be extracted from the first factor of the real activity group. For this purpose, let us assume that there is a regime switch in the index f_{1t}^1 itself and that its switching mechanism at time t is controlled by an unobservable state variable, v_t, that follows a first-order Markov chain. Thus, we specify a simple switching model (Hamilton, 1989) as:

$$f_{1t}^1 = \mu_{v_t} + u_t, \tag{9}$$

where u_t is an independent Gaussian perturbation with zero mean and variance σ_u^2. The nonlinear behavior of the time series is governed by the mean of the process, which is allowed to change within each of the two distinct regimes $v_t = 0$ and $v_t = 1$. As we impose that $\mu_0 > \mu_1$, the first regime refers to expansions and the second regime refers to recessions. The Markov-switching assumption implies that the transition probabilities are independent of the information set at $t-1$, $I_{t-1} = \{f_{11}^1, f_{12}^1, \ldots, f_{1t-1}^1\}$, and of the business cycle states prior to $t-1$

$$p(v_t = i|v_{t-1} = j, v_{t-1} = l, \ldots, I_{t-1}) = p(v_t = i|v_{t-1} = j) = p_{ij}. \tag{10}$$

The black line reported in Figure 5 shows the smoothed probabilities of recession inferred from the first real factor, $p(v_t = 1|I_T)$, and shaded areas that correspond to the NBER-referenced recessions. The figure indicates that the probabilities of recession are in striking accord with the

[16]These authors use the monthly version of the database used in this paper (FRED-MD).

professional consensus as to the history of US business cycles. The probabilities are close to either zero or one, emphasizing that the first factor of the real group captures well the underlying pattern of the dichotomous shifts between expansions and recessions. Remarkably, there are no instances in which an NBER recessions are not matched by high recession probabilities. In addition, we find that our factor model can also be used to date the reference cycle because the spikes in the probabilities of recession occur at about the NBER turning points.

We further assess the accuracy of the first real activity factor at capturing the US business cycle fluctuations by comparing the smoothed probabilities inferred from the estimated factor with those inferred from observable information. In particular, we use the smoothed probabilities inferred from GDP as it is a meaningful observable measure of the overall economic conditions. These probabilities are also plotted in the red line that appears in Figure 5. The comparison of both estimates shows clear gains in the characterization of the US business cycle recessions based on the higher informational content of the factor that summarizes the real activity data.[17] Using the simple metric suggested by Hamilton (1989) to consider a recession, which implies $p(v_t = 1|I_T) > 0.5$, only the factor-based recession probabilities would have identified the recession at the beginning of the 70s and the 2001 mild recession.

Finally, we provide some insights to assess the reliability of the values that fill in the gaps of the missing observations through the group-specific EM algorithm. For illustrative purposes, we compare the dynamics of some indicators that are available for the whole sample period with other indicators of the same category that contain missing values that have been imputed by the model. The top panel of Figure 6 plots the evolution of the growth rates of Real Output in Manufacturing Sector, which is not available from 1959:3 to 1987:1 and is observable only since 1987:2. Thus, this series has been enlarged back to 1959:3 with the EM estimates for the first 111 missing observations, with the cutoff signaled with a vertical dotted line.

The figure shows that the dynamics of Real Output in Manufacturing Sector in the unobserved period is similar to that of the observed period, with the reduction in the variance that characterize the US economic activity since the mid-eighties (the Great Moderation). To facilitate comparisons, the figure also plots the rate of growth of GDP, which is available for the whole sample period. As expected, these two indicators are highly synchronized in the observed second part of the sample. Interestingly, we also find a very similar pattern between the imputed vales of the Real Output in Manufacturing Sector and the observed values of GDP in the first part of the sample.

To provide a comparable example with indicators of the nominal group, the bottom panel of Figure 6 displays the S&P's Common Stock Price Index, which is not available from 1959:3 to 1971:1 (first 47 observations are missing), with the cutoff also signaled with a vertical dotted line. Again, the imputed values of the time series in the first part of the sample shows the dynamic patterns of the observed data in the second part of the sample. For comparison purposes, the figure also plots the NASDAQ composite index, which is observed for the whole sample period. Regardless of the part of the sample that we consider, the figure shows that the S&P's Common Stock Price Index is highly synchronized with the observed NASDAQ composite index.

[17] Within recessions, the Brier score loss is 0.06 for the probabilities inferred with the factor and 0.17 for the probabilities obtained with GDP.

4.2 Out-of-sample analysis

The model is evaluated in terms of its forecasting ability to perform 1- and 4-quarters ahead forecasts, with a special interest in evaluating the improvements of forecasting with a group-specific factor model over the forecasts conducted with the Stock-Watson factor model.

In each case, to compute the h-step-ahead forecast of the indicator $x_{i,t}$, we use the direct multistep forecasting equation

$$x_{i,t}^h = \alpha_{i,h} + \sum_{j=0}^{q-1} \beta_{i,hj}' \hat{f}_{g_i,t-h-j} + \sum_{l=0}^{p-1} \gamma_{i,hl} x_{i,t-h-l} + \epsilon_{i,t}^h, \tag{11}$$

where $\hat{f}_{g_i,t-j}$ is the $(h+j)$-lagged value of the $r_{g_i} \times 1$ vector of estimated group-specific factors and $\epsilon_{i,t}^h$ is an independent error term for the i-th indicator. Following McCracken and Ng (2016), we fix the number of autoregressive lags p at 4. In addition, we consider the same number of lags $q = 4$ for the group-specific factors.

Although interesting, we are precluded from using real-time vintages because they are available only from May 2018, making the real-time forecasting analysis unfeasible at the moment. Instead, we perform an out-of-sample exercise that tries to mimic a real-time analysis. The method consists of computing forecasts from successive enlargements of a partition of the latest available data set. It begins with data from the beginning of the sample until 2005:4. Using this sample, the data were screened for outliers and standardized, classification was performed and the group-specific number of factors are chosen. With these choices, the forecasting equation was used to compute forecasts $x_{2005:4+h}^h$ (for $h = 1$, it would be the forecast for 2006:1; for $h = 4$, it would be the forecast for 2006:4).

Then, the sample is updated by one period, classification and the factors are estimated, the forecasting models are re-estimated and the forecasts for $x_{2006:1+h}$ are computed. The forecasting procedure continues iteratively until the forecast for 2019:4 are computed, which are made using data until 2019:3 for $h = 1$ and using data until 2018:4 for $h = 4$. In each iteration, the squared deviation of h-ahead forecasts from actual data are computed and the average of these figures for each indicator, labeled as MSE, is stored.

For comparison purposes, similar h-ahead forecasts are performed using the factors extracted as in Stock and Watson (2002) and the resulting MSEs are also stored. Figure 7 reports the results of this forecasting exercise for the 248 series by plotting the MSEs of the group-specific factor model relative to the MSE of the forecasts of the Stock-Watson factor model. Hence, bars smaller than one signify that group-specific factors produce more accurate forecasts than common factors.

Regardless of the forecasting horizon, the figure shows that the improvements of the forecasts of group-specific factor models over the forecasts of standard Stock-Watson factor model are spread over all categories. More precisely, the group-specific model performs better for 75% of the 1-quarter-ahead forecasts and this figure rises to almost 80% in the case of 4-quarter-ahead forecasts. Noteworthy, the highest performance gains appear in interest rates, although substantial improvements also emerge for NIPA, industrial production and employment.

Finally, we pay special attention to the 1- and 4-quarter-ahead forecasts of the eight target

variables selected by McCracken and Ng (2021). In particular, the consider four real activity indicators, Real GDP, Industrial Production (INDPRO), the Unemployment rate (UNRATE) and the Federal Funds Rate (FFR), and four nominal indicators, the Consumer Prices Index (CPI), the Personal Consumption Expenditures Price Index (PCEPI), the GDP deflator (GDPPI), and the Production Price Index (PPI). To perform this comparison, Figure 8 displays the relative MSEs of the group-specific factor model forecasts over the Stock-Watson factor model forecasts for each target variable and forecasting horizon. With the unique exception of unemployment, the figure shows that the former outperforms the latter, being the gains in forecasting GDP and FFR remarkably large.[18]

5 Conclusion

To deal with large sets of unbalanced panels that contain cluster structure and potential outliers, this paper merge two strands from the econometric literature. First, the expectation-maximization (EM) algorithm combined with the factor estimator-based principal component analysis (PCA), as introduced by Stock and Watson (2002). Second, the grouped factor structures with unknown group membership and number of groups, as proposed by Ando and Bai (2016, 2017).

We examine the benefits of this contribution to the factor model literature through several Monte Carlo simulations. These experiments show that the proposed method works very well. In particular, we failed to find significant differences between the cluster abilities of our method, which deals with missing data in a simple way, and the factor models with unknown group membership proposed by Ando and Bai (2016, 2017), which require balanced data sets. The model performance does not deteriorate significantly when the simulations include non-homoskedastic errors with cross sectional dependence, errors that have some serial correlations and factors with temporal dependence. In addition, the simulations indicate that our group-specific EM algorithm outperforms the standard EM algorithm proposed by Stock and Watson (2002) and provides accurate inferences of the missing data.

Using the novel repository FRED-QD maintained by the St. Louis Fed and documented by McCracken and Ng (2021), we examine the empirical performance of our proposal to determine the number of clusters, to provide group-specific factors, to achieve accurate estimates of the unobserved missing data and to perform out-of-sample forecasts the US economic indicators. Our results suggest that a factor model that allows for only group-specific factors is better suited for this data set than a factor model that allows for both common and group-specific factors. We identify two distinctive groups of economic indicators: the group of nominal indicators and the group of real activity indicators. We find it promising to use the estimated factors for the identification the US business chronology. Finally, we find substantial improvements in forecasting with group factor models over forecasting with factor models that do not deal with grouped structure, as in Stock

[18]For these target variables, we also compared the forecast performance of our proposal over several alternative models. The first one was filling the missing data with the last valid observation and using the group factor model of Ando and Bai (2016). The second one was a factor model that omits the indicators with missing data. The third one was the group factor model of Ando and Bai (2016) that uses the balanced panel of indicators (excluding the indicators with missing values). We found that our model outperforms all of these alternatives. These results are available from the authors upon request.

and Watson (2002), or forecasting with factor models that require removing the indicators with missing observations, as in Ando and Bai (2016).

According to our results, we consider that the proposed framework is also a very promising tool for handling large and unbalanced data sets with group structures. In fact, we look forward to future work addressing the following issues. First, we see a natural extension of the factor model for large and incomplete data sets with unknown group structure advocated in this paper to consider additional explanatory variables. This extension can consider homogeneous coefficients over all cross-sectional units (Ando and Bai, 2016), heterogeneous group-specific coefficients (Ando and Bai, 2016) or coefficients that vary over all the cross-section units (Ando and Bai, 2017).

Second, the research division of the Federal Reserve Bank of St. Louis also provides a monthly frequency companion to the data set used in our empirical application, so-called FRED-MD. This raises the possibility that further empirical gains can be realized by extending the factor model introduced in this paper to handle mixed frequencies. Although we do not take advantage of this feature here, mixing frequencies deserves to be pursued further as it could combine the economic information provided by FRED-MD and FRED-QD at the same time.

Third, the real-time vintages of the data set used in the empirical application are only available since 2018. According to this data limitation, our assessment of the out-of-sample forecasting performance relies on fractions of the finally revised data. As new vintages will become available, real-time reassessments will provide new insights on the model's forecasting accuracy.

Fourth, our approach focuses on dealing with missing data using the Stock and Watson (2002) EM algorithm. However, some modifications has been adopted in the recent literature. For example, McCracken and Ng (2021) initialize the algorithm to the unconditional sample mean based on the non-missing values and the mean and variance of the data are re-calculated in each iteration. In a recent paper, Jin, Miao and Su (2021) replace the missing observations by zeros and conduct the usual PC analysis for a scaled version of the data matrix where the scale is determined by the percentage of observed values in the data. Following the EM algorithm, they replace the missing observations by such initial estimators of the common components and obtain updated PC estimators, iterating this procedure until convergence. We leave the comparison of these approaches for further research.

Finally, the performance of the estimation procedure may be specifically evaluated when applied to data sets showing different structures of practical relevance. In particular, we could extend our approach to handle seasonal patterns or long memory in the set of economic indicators.

References

[1] Alonso, A., Galeano, P., and Peña, D. 2020. A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics* 216: 35-52.

[2] Ando, T. and Bai, J. 2015. Asset pricing with a general multifactor structure. *Journal of Financial Econometrics* 13: 556-604.

[3] Ando, T. and Bai, J. 2016. Panel data models with grouped factor structure under unknown group membership *Journal of Applied Econometrics* 31: 163-191.

[4] Ando, T. and Bai, J. 2017. Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association* 112: 1182-1198.

[5] Angelini, E., Henry, J., and Marcellino, M. 2006. Interpolation and backdating with a large information set. *Journal of Economic Dynamics and Control* 30: 2693-2724.

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

[7] Bai, J., and Ng, S. 2008. Large dimensional factor analysis. *Foundations and Trends in Econometrics* 3: 89-163.

[8] Bernanke, B., and Boivin, J. 2003. Monetary policy in a data-rich environment. *Journal of Monetary Economics* 50: 525-546.

[9] Breitung, J., and Eickmeier, S. 2016. Analyzing international business and financial cycles using multi-level factor models: A comparison of alternative approaches. In *Dynamic factor models*. Emerald Group Publishing Limited.

[10] Choi, I., Kim, D., Kim, Y., and Kwark, N. 2018. A multilevel factor model: Identification, asymptotic theory and applications. *Journal of Applied Econometrics* 33: 355-377.

[11] Crucini, M., Kose, M., Otrok, C. 2011. What are the driving forces of international business cycles? *Review of Economic Dynamics* 14: 156-175.

[12] Gregory, A., Head, A., and Raynauld, J. 1997. Measuring world business cycles. *International Economic Review* 38: 677-701.

[13] Hallin, M., and Liška, R. 2011. Dynamic factors in the presence of blocks. *Journal of Econometrics* 163: 29-41.

[14] Hamilton, J., 1989. A new approach to the economic analysis of nonstationary time series and the business cycles. *Econometrica* 57: 357-384.

[15] Jin, S., Miao, K., and Su, L. 2021. On factor models with random missing: EM estimation, inference, and cross validation. *Journal of Econometrics* 222: 745-777.

[16] Kose, M., Otrok, C., and Whiteman, C. 2003. International business cycles: World, region, and country-specific factors. *American Economic Review* 93: 1216-1239.

[17] Leiva-Leon, D. 2014. Real vs. nominal cycles: a multistate Markov-switching bi-factor approach. *Studies in Nonlinear Dynamics and Econometrics* 18: 557-580.

[18] Marcellino, M., and Schumacher, Ch. 2010. Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics* 72: 518-550.

[19] McCracken, M., and Ng, S. 2016. FRED-MD: A Monthly database for macroeconomic Rrsearch. *Journal of Business and Economic Statistics* 34: 574-589.

[20] McCracken, M., and Ng, S. 2021. FRED-QD: A quarterly database for macroeconomic research. *Federal Reserve Bank of St. Louis Review* 103: 1-44.

[21] Rodriguez-Caballero, C., and Ergemen, Y. 2017. *Estimation of a dynamic multilevel factor model with possible long-range dependence.* Technical report, Universidad Carlos III de Madrid. Departamento de Estadística.

[22] Schumacher, Ch., and Breitung, J. 2008. Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. *International Journal of Forecasting* 24: 386-398.

[23] Stock, J., and Watson, M. 2002. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics* 20: 147-162.

[24] Stock, J., and Watson, M. 2011. Dynamic factor models. In M. Clements and D. Hendry (eds.), *The Oxford Handbook of Economic Forecasting*. Oxford: Oxford University Press.

[25] Stock, J., and Watson, M. 2012. Disentangling the Channels of the 2007-2009 Recession. *Brookings Papers on Economic Activity*, Spring: 81-156.

Table 1: Simulation: model's accuracy

N	T	P	M	S		r		r_1		r_2		r_3		N_1		N_2		N_3	
				U	C	U	C	U	C	U	C	U	C	U	O	U	O	U	O
Panel A: Errors are homoskedastic and serially uncorrelated																			
50	150	10	50	5.8	86.8	0	100	0	82.2	0	84.8	0	82.0	1.4	1.3	1.2	1.3	1.4	1.3
50	150	20	50	6.5	84.8	0	100	0	81.8	0	83.0	0	77.4	2.0	2.1	1.8	2.0	2.1	1.8
100	150	10	50	1.6	94.2	0	100	0	92.8	0	92.4	0	93.4	0.8	0.9	0.7	0.7	0.8	0.7
100	150	20	50	2.2	93.2	0	100	0	91.2	0	90.4	0	90.4	1.3	1.1	1.2	1.3	1.0	1.2
50	250	10	84	6.8	84.8	0	100	0	89.1	0	83.2	0	81.8	1.9	1.7	2.1	1.5	1.9	1.7
50	250	20	84	8.6	82.2	0	100	0	80.1	0	80.6	0	80.1	2.1	2.2	2.0	1.9	2.2	2.1
100	250	10	84	1.8	93.2	0	100	0	93.8	0	92.8	0	91.8	0.9	0.9	0.8	0.8	0.9	0.8
100	250	20	84	3.2	91.0	0	100	0	89.4	0	90.2	0	90.0	1.2	1.2	1.2	1.2	1.1	1.3
Panel B: Errors are nonhomoskedastic and cross-sectionally correlated																			
50	150	10	50	6.8	81.6	0	100	0	81.8	0	80.6	0	80.6	1.8	1.7	2.2	2.1	1.9	1.8
50	150	20	50	7.2	80.2	0	100	0	80.4	0	78.8	0	80.1	1.9	2.1	2.1	1.9	2.1	1.9
100	150	10	50	2.8	90.6	0	100	0	91.6	0	91.8	0	90.2	1.2	0.9	0.8	1.1	0.9	1.1
100	150	20	50	3.2	89.4	0	100	0	89.4	0	89.6	0	89.8	1.6	1.1	1.3	1.5	1.2	1.3
50	250	10	84	7.1	79.8	0	100	0	80.1	0	80.6	0	80.1	1.9	1.8	2.1	2.1	1.9	1.9
50	250	20	84	8.6	76.8	0	100	0	76.4	0	78.6	0	78.8	2.1	2.1	1.9	2.1	2.1	1.9
100	250	10	84	4.6	87.8	0	100	0	89.6	0	90.2	0	89.6	1.3	1.2	1.1	0.8	1.1	0.2
100	250	20	84	6.2	84.8	0	100	0	87.8	0	88.6	0	86.8	1.3	1.3	1.2	1.9	1.3	1.6
Panel C: Errors have some serial and cross-sectional correlations																			
50	150	10	50	7.2	79.8	0	100	0	80.8	0	78.8	0	80.1	1.9	1.9	2.1	2.2	1.9	2.0
50	150	20	50	8.1	77.8	0	100	0	78.6	0	78.2	0	77.5	2.1	2.2	2.3	2.1	2.2	1.9
100	150	10	50	3.6	87.2	0	100	0	90.1	0	90.4	0	90.6	1.1	1.1	0.9	0.8	1.2	1.1
100	150	20	50	4.1	83.4	0	100	0	87.6	0	88.2	0	90.2	1.3	1.2	1.4	1.5	1.1	1.4
50	250	10	84	8.6	78.8	0	100	0	78.8	0	81.2	0	80.2	1.9	2.0	2.0	2.1	2.1	1.9
50	250	20	84	9.4	72.8	0	100	0	75.2	0	74.8	0	76.2	2.2	2.1	2.0	2.2	1.9	2.1
100	250	10	84	5.2	83.6	0	100	0	90.1	0	89.2	0	89.8	1.2	1.1	1.2	0.9	1.9	1.1
100	250	20	84	6.4	81.2	0	100	0	86.8	0	88.6	0	84.6	1.4	1.4	1.1	1.1	1.4	1.8
Panel D: Dynamic factors																			
50	150	10	50	7.8	79.4	0	100	0	80.2	0	79.4	0	80.0	1.8	2.1	2.0	2.1	1.9	1.9
50	150	20	50	8.2	76.4	0	100	0	77.8	0	77.2	0	77.4	2.2	2.1	2.4	2.2	2.1	2.0
100	150	10	50	3.8	86.8	0	100	0	90.4	0	90.4	0	90.6	0.9	1.2	1.8	0.9	1.3	1.0
100	150	20	50	4.2	82.2	0	100	0	87.6	0	89.4	0	87.2	1.2	1.4	1.5	1.6	1.2	1.3
50	250	10	84	8.6	78.6	0	100	0	80.2	0	80.1	0	80.0	1.8	2.1	2.1	1.9	1.9	1.9
50	250	20	84	9.6	72.4	0	100	0	77.2	0	74.2	0	74.6	2.3	2.2	2.1	1.9	1.8	2.1
100	250	10	84	5.8	82.2	0	100	0	89.6	0	88.2	0	88.8	1.1	1.3	1.3	1.2	0.9	1.3
100	250	20	84	6.2	80.6	0	100	0	87.2	0	87.2	0	87.6	1.5	1.6	1.2	1.1	1.3	1.4

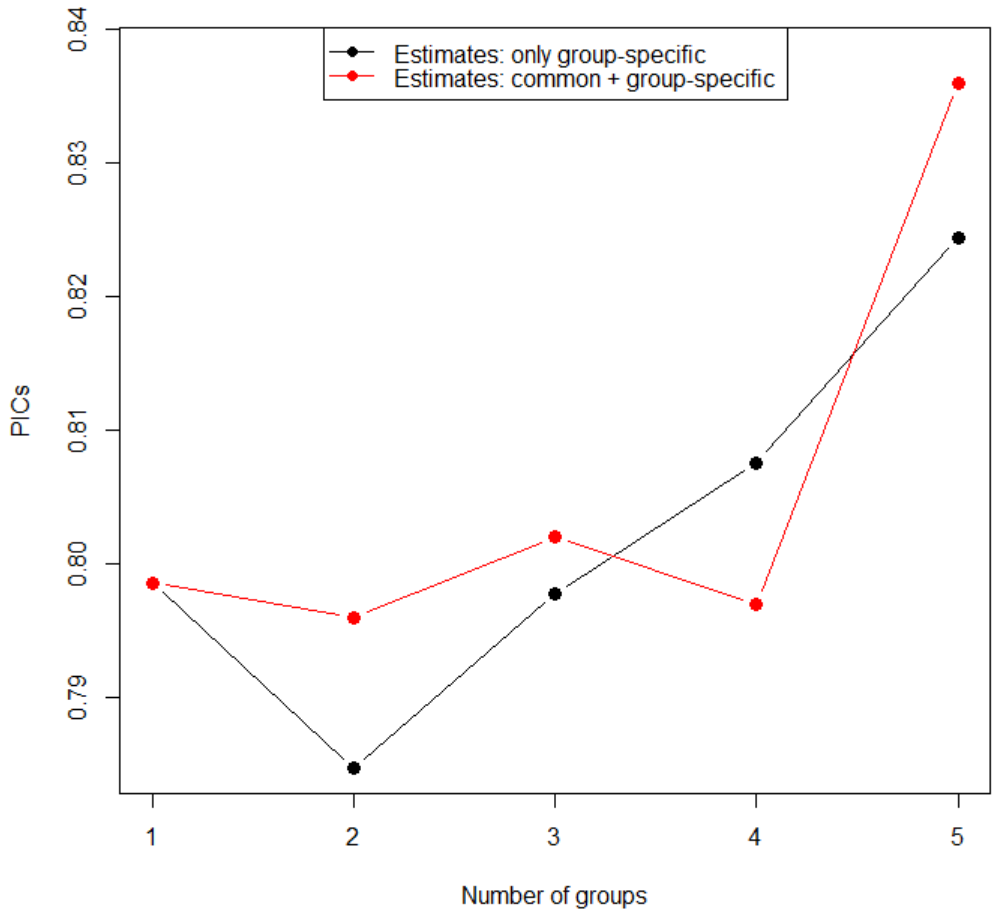
Notes. In the first four columns, N is the number of time series in each group, T is the time dimension, P is the percentage of time series with missing data and M is the number of missing data in each of these series. In the next eight columns, the figures refer to the percentages of under- (U) or correct (C) identification of the number of groups (S) and the number of common (r) and group-specific factors (r_i, $i = 1, 2, 3$). The last six columns refer to Underclassification (U) or Overclassification (O) rates.

Table 2: Comparison with the standard EM algorithm

N	T	P	M	*Better*
50	150	10	50	88.2
50	150	20	50	82.4
100	150	10	50	96.8
100	150	20	50	95.4
50	250	10	84	85.4
50	250	20	84	81.2
100	250	10	84	94.7
100	250	20	84	92.8

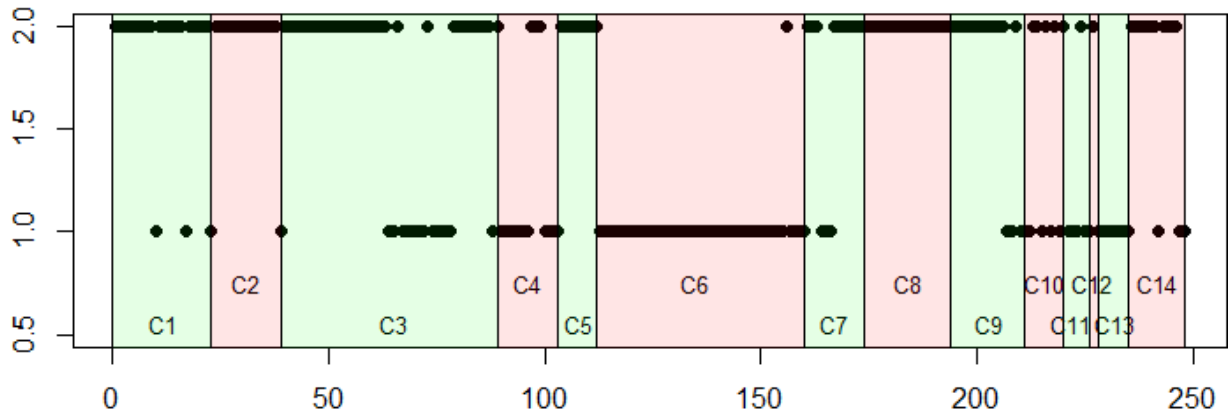
Notes. See notes of Table 1. The last column shows the percentage of times (out of the total simulations) that our grouped factor model with EM algorithm exhibits lower mean squared error over the standard Stock and Watson (2002) EM algorithm at filling in missing data.

Figure 1: Model selection



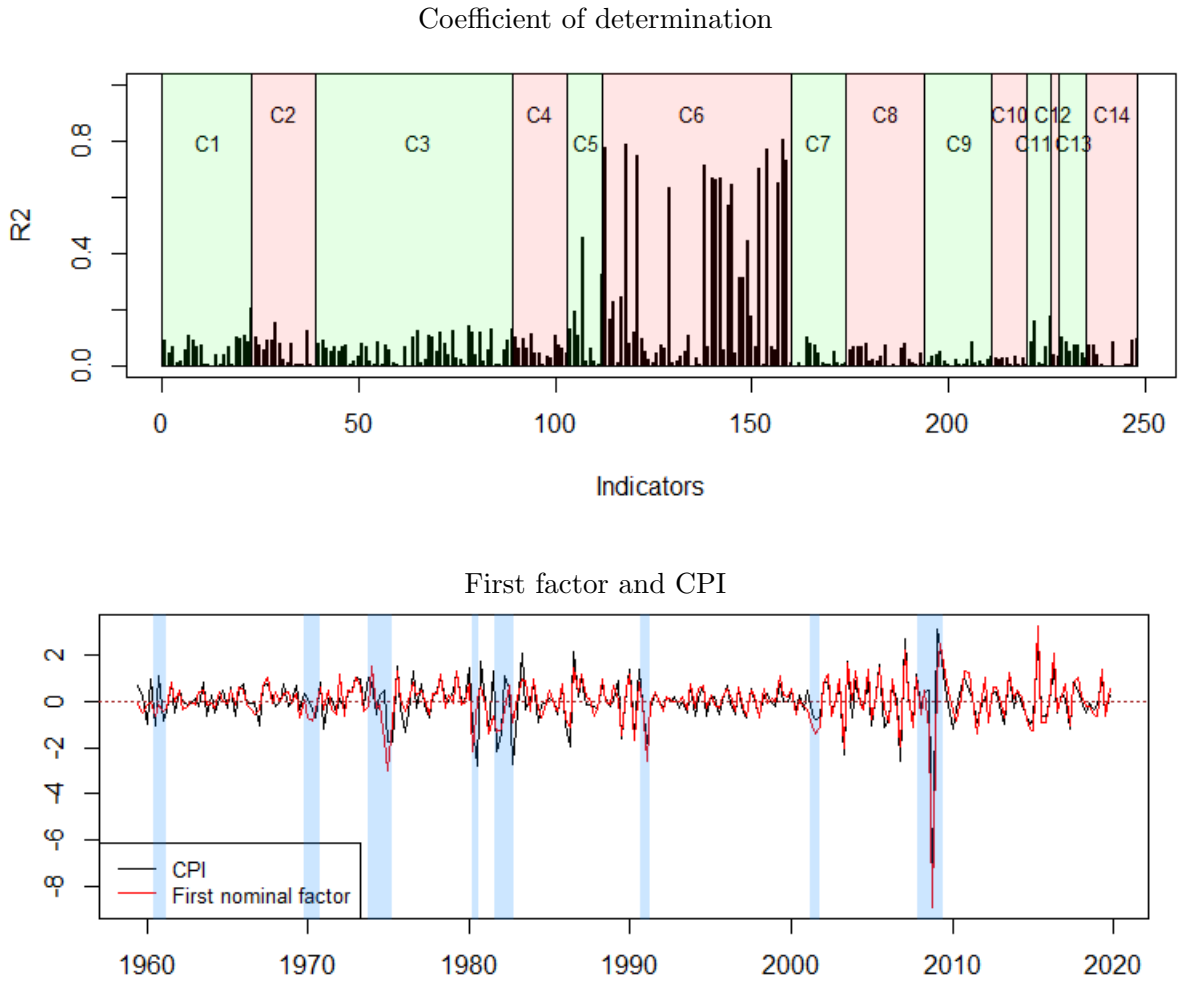
Notes. The figure shows the value of $PICs$ as a function of the number of groups. The red line refers to a group factor model that allows for both common and group-specific factors and the black line refers to a group factor model that allows for only group-specific factors.

Figure 2: Classification $G = (g_1, \ldots, g_N)$



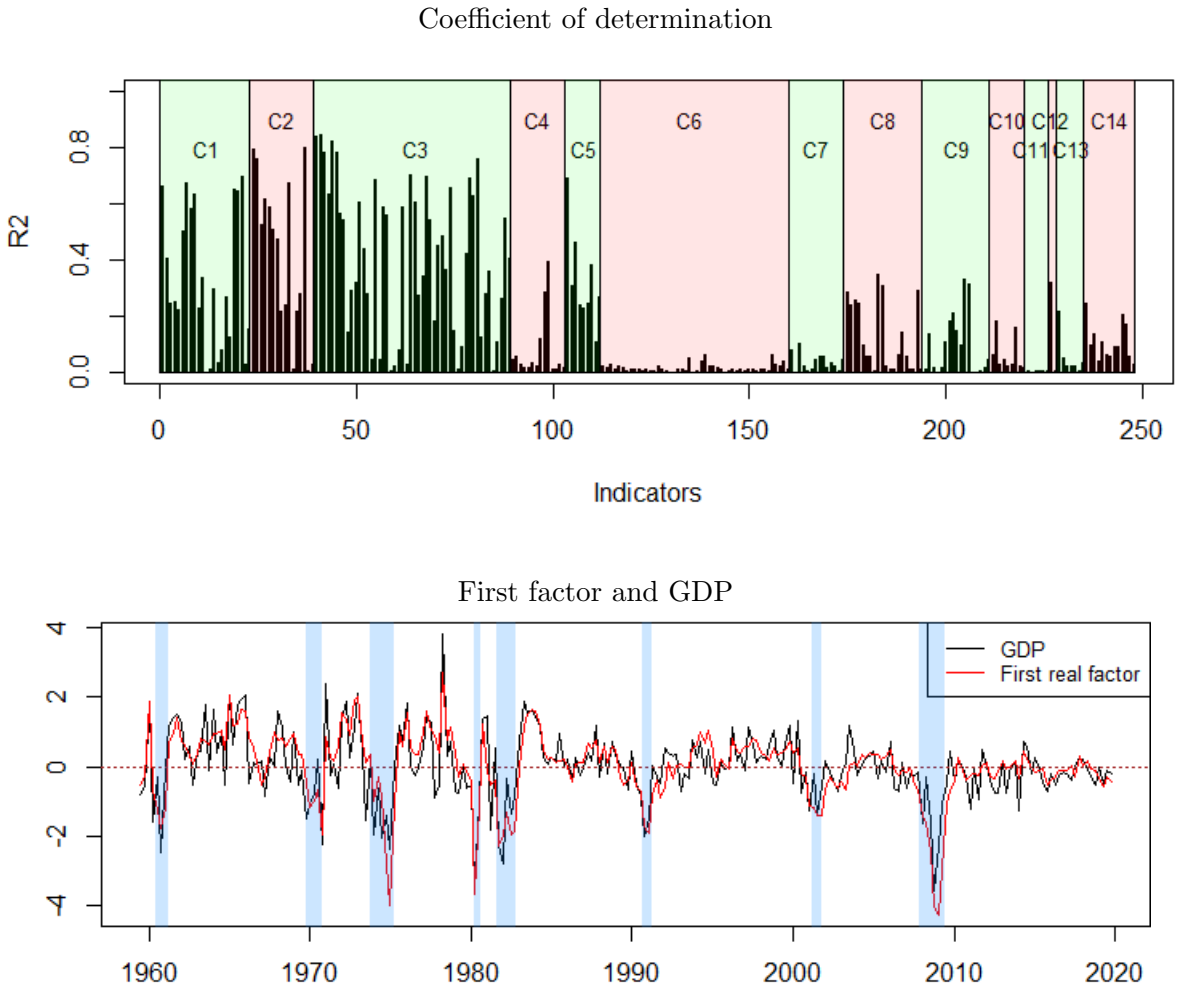
Notes. Classification of the $N = 248$ indicators provided by FRED-QD into the two identified groups. Following McCracken and Ng (2020), the indicators are grouped by categories labeled as follows: National Income and Product Accounts (C1), Industrial Production (C2), Employment and Unemployment (C3), Housing (C4), Inventories, Orders, and Sales (C5), Prices (C6), Earnings and Productivity (C7), Interest Rates (C8), Money and Credit (C9), Household Balance Sheets (C10), Exchange Rates (C11), Other (C12), Stock Markets (C13), Non-Household Balance Sheets (C14).

Figure 3: Nominal group: first factor and economic indicators



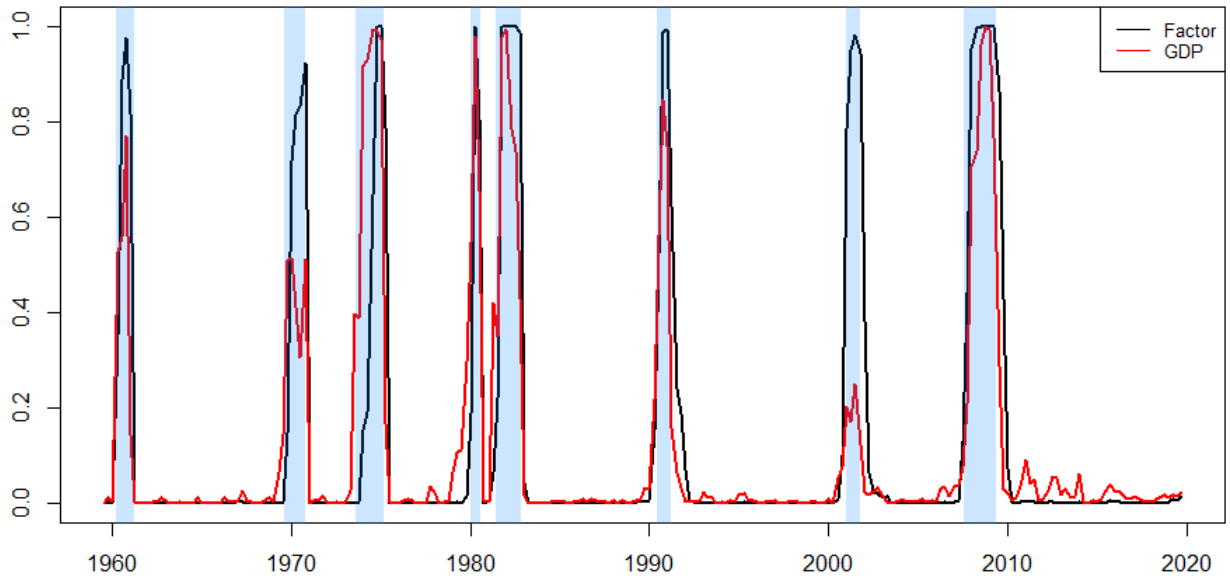
Notes. Top panel displays the coefficient of determination between the first factor of the nominal group and the economic indicators, grouped by categories (C1 to C14) following McCracken and Ng (2020) (See notes of Figure 2). The factor and the stationary transformation of CPI are plotted (normalized to have zero mean and unit variance) in the bottom panel. The shaded areas refer to the NBER-referenced recessions.

Figure 4: Real activity group: first factor and economic indicators



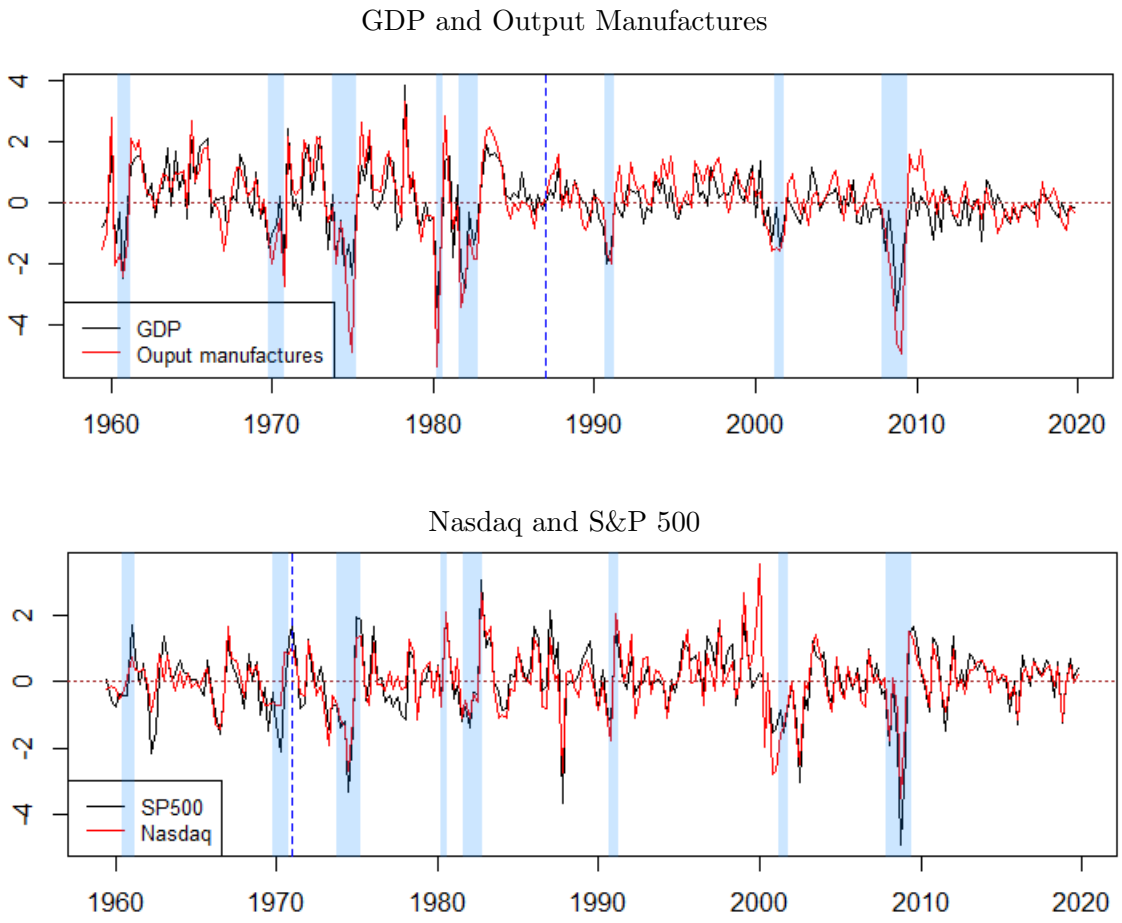
Notes. Top panel displays the coefficient of determination between the first factor of the real activity group and the economic indicators, grouped by categories (C1 to C14) following McCracken and Ng (2020) (See notes of Figure 2). The factor and GDP growth rate are plotted (normalized to have zero mean and unit variance) in the bottom panel. The shaded areas refer to the NBER-referenced recessions.

Figure 5: Smoothed probabilities



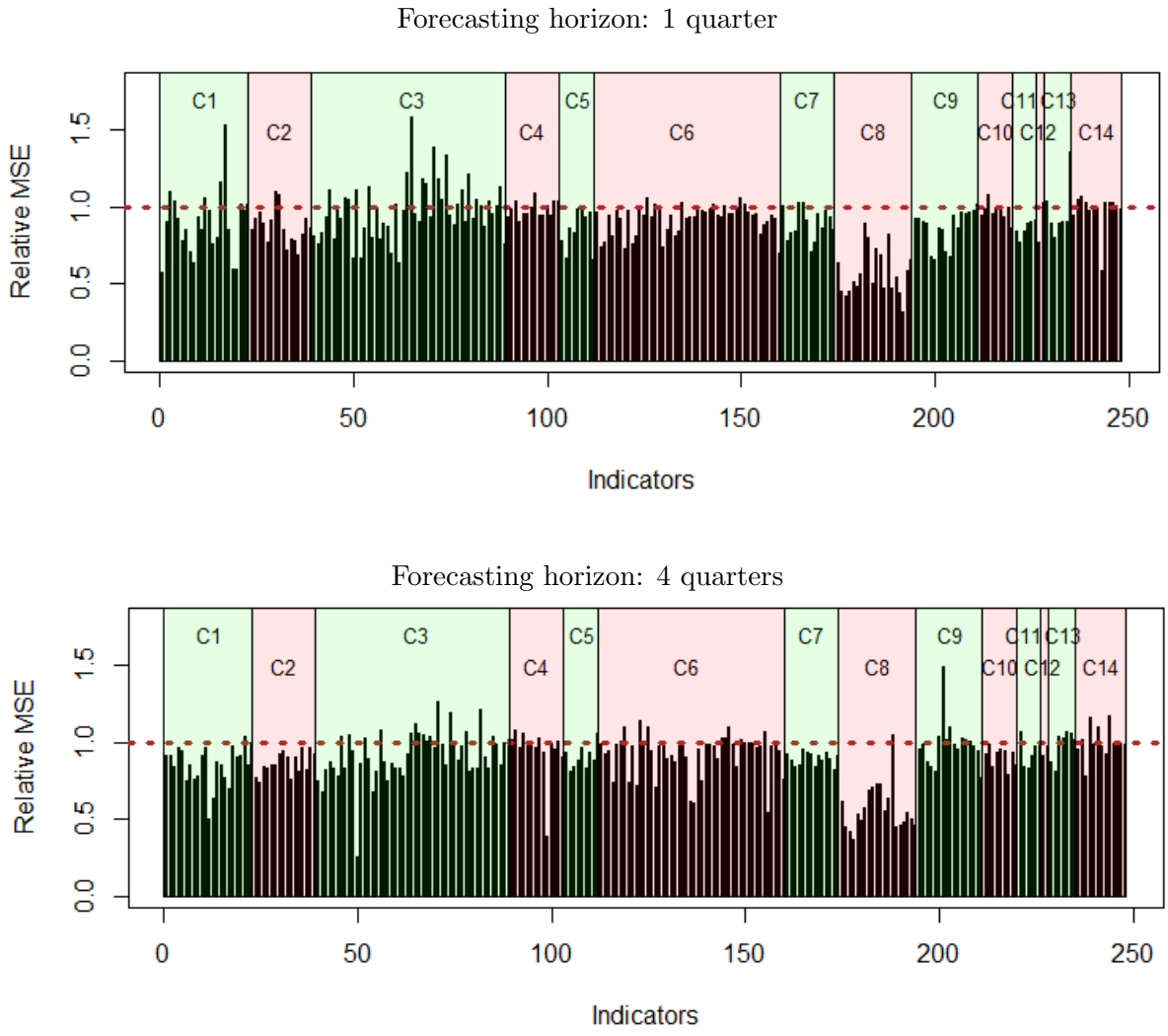
Notes. The figure shows the smoothed probability estimates for the recession regime based on the real activity factor estimates (solid line) and GDP (dashed line). The shaded areas refer to the NBER-referenced recessions.

Figure 6: Fill in the missing observations



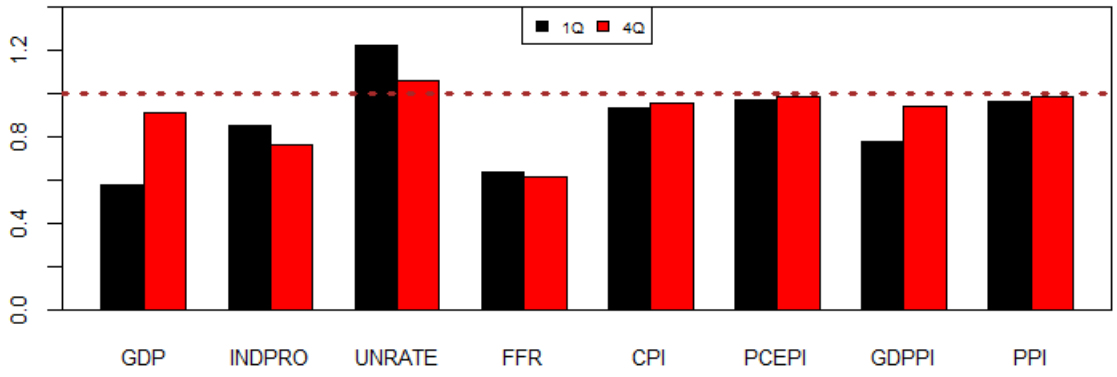
Notes. The top panel displays the growth rates of GDP and Real Output in Manufacturing Sector (first 111 observations are missing). The bottom panel displays the stationary transformations of the NASDAQ Composite Index and the S&P's Common Stock Price Index (first 47 observations are missing). The vertical lines refer to dates of the last missing observations. To facilitate comparisons, the series been normalized to have zero mean and unit variance. The shaded areas refer to the NBER-referenced recessions.

Figure 7: Relative MSE 1- and 4-quarter horizons



Notes. The figure shows the relative MSEs of the forecasts from the group-specific factor model over the Stock-Watson factor model across the 248 macroeconomic indicators, which are grouped by categories (C1 to C14) following McCracken and Ng (2020) (See notes of Figure 2).

Figure 8: Relative MSE for some key indicators



Notes. The figure shows the relative MSEs of the forecasts from the group-specific factor model over the Stock-Watson factor model for the eight target variables selected in McCracken and Ng (2020): Real GDP (GDP), Industrial Production (INDPRO), the Unemployment rate (UNRATE), the Federal Funds Rate (FFR)), the Consumer Price Index (CPI), the Personal Consumption Expenditures Price Index (PCEPI), the GDP deflator (GDPPI), and the Production Price Index (PPI). Black bars correspond to 1-quarter-ahead forecasts. Red bars correspond to 4-quarter-ahead forecasts.