UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working Papers
Statistics and Econometrics
18-03
ISSN 2387-0303
Junio 2018

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

Estimation of the common component in Dynamic Factor Models.

Angela Caro[a], Daniel Peña[a,b]

Abstract

One of the most effective techniques that allows a low-dimensional representation of Big Datasets is the Dynamic Factor Model (DFM). We analyze the finite sample performance of the well-known Principal Component estimator for the *common component* under different scenarios. Simulation results show that for data samples with large number of observations and small time series dimension, the variance-covariance matrix specification with lags provides better estimations than the classic variance-covariance matrix. However, in high-dimension data samples the classic variance-covariance matrix performs better no matter the sample size. Second, we apply the Principal Component estimator to obtain estimates of the business cycles of the Euro Area and its country members. This application, together with a cluster analysis, studies the phenomenon known as the *Two-Speed Europe* with two groups of countries not geographically related.

Keywords: Time series, Factor Models, Principal Components, Canonical Correlations.

[a] Department of Statistics, Universidad Carlos III de Madrid.
[b] Institute UC3M-BS of Financial Big Data

Acknowledgements: the first author acknowledge financial support from the Spanish Ministry of Education, Culture and Sport for the Training of University Teachers, and the second author from the Spanish Ministry of Education and Science, research project ECO2015-66593-P

Estimation of the common component in Dynamic Factor Models

Angela Caro*
Department of Statistics, University Carlos III of Madrid
and
Daniel Peña†
Department of Statistics, Institute UC3M-BS of Financial Big Data
University Carlos III of Madrid

June 25, 2018

Abstract

One of the most effective techniques that allows a low-dimensional representation of Big Datasets is the Dynamic Factor Model (DFM). We analyze the finite sample performance of the well-known Principal Component estimator for the *common component* under different scenarios. Simulation results show that for data samples with large number of observations and small time series dimension, the variance-covariance matrix specification with lags provides better estimations than the classic variance-covariance matrix. However, in high-dimension data samples the classic variance-covariance matrix performs better no matter the sample size. Second, we apply the principal component estimator to obtain estimates of the business cycles of the Euro Area and its country members. This application, together with a cluster analysis, studies the phenomenon known as the *Two-Speed Europe* with two groups of countries not geographically related.

Keywords: Time series, Factor Models, Principal Components, Canonical Correlations.

*Supported by the Spanish Ministerio de Educación, Cultura y Deporte under grant FPU15/03983.
†Partially supported by the Spanish Ministerio de Economía y Competitividad under grant ECO2015-66593-P.

1 Introduction

Nowadays, information and communication technologies have achieved a big improvement and all knowledge fields, including Economics, Mathematics and Statistics, have taken advantage of these developments in dealing with high-dimensional datasets. One highlighted improvement has been the reduction of the processing and storage costs inherent to large data banks, the so-called Big Data. These databases include information that grows exponentially day by day and can be easily shared all over the world through Internet. In Peña (2014) one can find some of the main implications when analyzing these enormous datasets. As a consequence of these developments, the evolution of specific subareas such as time series econometrics, multivariate analysis, non-parametric methods or Bayesian estimation, has made possible the estimation of more complex statistical models which help to analyze and predict large number of macroeconomic and financial variables. It is known between researchers that classical economic and multivariate time series models present serious limitations when the number of variables to consider is high. One of the main limitations is that the estimated number of parameters grows with the square of the time series dimension. Therefore, finding simplified structures or factors that reduce this number of parameters has became a must when applying these models to real data.

Two well-known solutions that face the problem of dimensionality presented in macroeconomic and financial time series are (1) to assume that many coefficients are zero and apply some regularization method that allows an efficient estimation of the model, and (2) to reduce the time series vector's dimension. The former is the approach of the autoregressive Bayesian models, where regularization is established as a priori information, see, e.g., Doan et al. (1984) or more recently Bańbura et al. (2010). Another possible form of regularization is through LASSO and Ridge regression methods, which also have a Bayesian interpretation, see for example Belloni et al. (2012). The second solution consists in finding linear combinations of the time series which represent determinant features. Some techniques that pursue this goal are classical principal components and its application to augmented observations as in Ku et al. (1995), the Scalar Component Model (SCM) introduced in Tiao and Tsay (1989), the reduced-rank models of Ahn and Reinsel (1990), the Dynamic Principal Components introduced by Brillinger and generalized by Peña and Yohai (2016), and the well-known Dynamic Factor Model (DFM).

This work focuses on DFM, which has been considered one of the most effective techniques when dealing with the problem of high dimensionality present in macroeconomic and financial time series datasets. This model was originally proposed by Geweke (1977) and Sargent et al. (1977)

as an extension of the classical *static* factor model for macroeconomic time series. These models are well-known in macroeconomic (comovements of macroeconomic aggregates, cross-country variation, forecasting with diffusion indexes) and finance applications (asset returns, risk management, portfolio allocation, arbitrage pricing theory), and also are widely applied in different areas of research, such as management (demand analysis, aggregate implications of microeconomic behavior), medicine, and environment. The main idea in the DFM is that the comovements of a N-dimensional vector of time series y_t can be explained by the sum of two mutually orthogonal unobserved components: the *common component* which have a pervasive effect over all the variables in y_t, and the *idiosyncratic component* or noise, which is specific to each time series variable. This work studies the estimation of the factors, and factor loadings in DFM. Particulary, we analyze the performance of two variance-covariance matrix specifications when estimating the *common component* by principal components. Our interest is to find out in which scenarios it would be more beneficial to consider each one of the two specifications: the classic variance-covariance matrix which includes contemporaneous information, and the one proposed by (Peña and Box, 1987) which includes past information. In order to analyze this performance, we carry out a simulation study in which different scenarios are considered.
The article is organized as follows; Section 2 introduces a brief summary about the DFM and the Principal Component estimator. Section 3 presents the data-generating processes considered in the the simulation exercise, introduces the measure we apply in order to evaluate the performance of the estimations and shows the simulation results. Section 4 gives an example of a macroeconomic application about the Euro Area business cycles. Finally, some concluding remarks and potential extensions are given in Section 5. Tables and figures are available in the Appendix and upon request.

2 Dynamic Factor Model

The state-of-the-art about DFM has distinguished two representations in terms of the dynamic behaviour of the latent common factors. On one hand, the standard representation, which is known as the *static* or *stacked* representation, introduces the latent factors, $\mathbf{f}_t$, in Equation 1 contemporaneously. On the other hand the *dynamic* representation takes into account the current effect, as well as, lags of the common factors, see Stock and Watson (2016) for an in depth explanation of the relationship between both representations. The idea behind DFM is that the comovements of a N-dimensional vector of time series variables, $\mathbf{y}_t$, are explained by the sum of two latent components: $\Lambda \mathbf{f}_t$ and $\mathbf{e}_t$, where $\Lambda \mathbf{f}_t$ is the *common component*, $\mathbf{f}_t$ is a $r \times 1$ vector of common factors and Λ is a $N \times r$ matrix of factor loadings, and $\mathbf{e}_t$, is the *idiosyncratic component*, a $N \times 1$ vector of idiosyncratic disturbances or errors. Moreover, the factors follow time series processes, which has been generally assumed to be a vector autoregression, VAR(p), where p is the degree of the polynomial matrix $\Phi(L) = (I - \phi_1 L - ... - \phi_p L^p)$. When the model admits an infinite lag order and assumes that the idiosyncratic components are nonorthogonal it is known as the *Generalized* DFM, proposed in Forni et al. (2000). Finally, η_t is a $r \times 1$ Gaussian white noise vector with positive and finite covariance matrix Σ_η, which is independent of the idiosyncratic errors $\mathbf{e}_t$, that is, $E\mathbf{e}_t\eta'_{t-k} = 0$ for all k.

$$\mathbf{y}_t = \Lambda \mathbf{f}_t + \mathbf{e}_t \tag{1}$$

$$\Phi(L)\mathbf{f}_t = \eta_t \tag{2}$$

In Equation 1 just the left-hand-side is observed whereas the remaining information must be estimated with the information contained in the N-dimensional vector y_t.

It is important to consider the following general assumptions about the factors $\mathbf{f}_t$, the factor loadings in Λ matrix, and the idiosyncratric errors, $\mathbf{e}_t$, from Stock and Watson (2002). In order to avoid the problem of *identification*, given that for any nonsingular matrix A, $\Lambda \mathbf{f}_t = \Lambda A A^{-1} \mathbf{f}_t$, we assume that:

A.1 $(\Lambda'\Lambda/N) \rightarrow I_r$

A.2 $E(\mathbf{f}_t\mathbf{f}'_t) = \Sigma_{\mathbf{ff}}$

A.3 $\mid \lambda_i \mid \leq \bar{\lambda} < \infty$

A.4 $T^{-1}\sum_t \mathbf{f}_t\mathbf{f}'_t \xrightarrow{p} \Sigma_{\mathbf{ff}}$

Where $\Sigma_{\mathbf{ff}}$ is a diagonal matrix with elements $\sigma_{ii} > \sigma_{ij} > 0$ for $i < j$, which means that factors may present autocorrelation. The factors will be identified up to a change of sign given that A matrix is restricted to be diagonal with elements of ± 1.

Different characterizations of the idiosyncratic component give rise to various versions of DFM such as, *exact* or *approximate*, among others. The main difference between the *approximate* and the *classic* or *exact* factor model is that in the latter the idiosyncratic errors, $\mathbf{e}_t$, are assumed to be cross-sectionally and serially uncorrelated, whereas in the *approximate* specification the errors are allowed to present serial and weak cross-sectional correlation. Let e_{it} indicates the ith element of $\mathbf{e}_t$, then the assumptions about the errors $\mathbf{e}_t$ in the exact DFM would be:

A.5 $E(\mathbf{e}_t'\mathbf{e}_{t+u}/N) = 0$

A.6 $E(e_{it}e_{jt}) = 0$

and in the approximate DFM:

A.7 $E(\mathbf{e}_t'\mathbf{e}_t/N) = \gamma_{N,t}(u)$, and
$\lim_{n\to\infty} sup_t \sum_{u=-\infty}^{\infty} \mid \gamma_{N,t}(u) \mid < \infty$

A.8 $E(e_{it}e_{jt}) = \tau_{ij}$, $\lim_{n\to\infty} sup_t N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \mid \tau_{ij,t} \mid < \infty$

A.9 $\lim_{n\to\infty} sup_t N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \mid cov(e_{is}e_{it}, e_{js}e_{jt}) \mid < \infty$

These assumptions in the *approximate* DFM are consistent with macroeconomic data and have been widely considered in the literature about DFM.
Furthermore, as it is said in Hallin and Lippi (2015), most of these models have the nature of *statistical models*, in the sense that they make assumptions on the underlying data-generating process. Traditionally, some of these models have assumed stationarity such as in Peña and Box (1987), Stock and Watson (2002), Stock and Watson (1988), Bai and Ng (2002) and Lam and Yao (2012), between others. Whereas, Peña and Poncela (2006) assume nonstationarity for integrated process, Pan and Yao (2008) for generalized process and Motta et al. (2011) and Motta and Ombao (2012), assume locally stationary process. It is recommended to see Bai et al. (2008), Stock and Watson (2011) and Stock and Watson (2016) for an overview of the different DFM specifications, estimation methodologies, and empirical applications.

2.1 Principal Component Estimator

Two of the main tasks that researches face when dealing with DFM is to estimate the number of common factors r, as well as, the estimation of the factor loading space $\mathcal{M}(\Lambda)$ and the common latent factors $\mathbf{f}_t$. In the present work we are interested in the estimation of the *common* component and we assume that the number of common factors r is known. Specifically, we pay attention to one of the most applied methodologies in dimension-reduction problems, Principal Component Analysis (PCA). Some well-known references about the consistency of the Principal Component estimator are Connor and Korajczyk (1986), Forni and Reichlin (1998), Forni et al. (2000), Bai

(2003) and Bai and Ng (2006). Furthermore, we recommend to see Stock and Watson (2011) which summarize the different methodologies within the time-domain estimation of DFM in three generations. In summary, the first generation applied Gaussian Maximum Likelihood (MLE) and the Kalman filter to estimate low-dimensional parametric models. The second generation considered cross-sectional averaging methods, mainly PC, to estimate high-dimensional nonparametric models. The third generation combines both, using the consistent nonparametric estimates of the factors (second generation) in the estimation of the state-space model (first generation), obtaining the parameter estimates.

For the cross-sectional averaging methods, the vector $\mathbf{f}_t$ is considered a r-dimensional parameter to be estimated using the cross-sectional averaging of $\mathbf{y}_t$. Therefore, the estimator of $\mathbf{f}_t$, $\widehat{\mathbf{f}}_t$, is obtained as the weighted average of $\mathbf{y}_t$ using a nonrandom matrix of weights W, which is normalized such that $WW'/N = I_r$. The principal component estimator sets $W = \widehat{\Lambda}$, where $\widehat{\Lambda}$ is the matrix of scaled eigenvectors associated with the r largest eigenvalues of the sample covariance matrix M described bellow, and the factors are computed as $\widehat{\mathbf{f}}_t = \widehat{\Lambda}'\mathbf{y}_t$, the scaled first r principal components of $\mathbf{y}_t$ and this estimator is consistent under general error structure as shown in (Stock and Watson, 2011). As we mentioned in the introduction, our interest is to find out which variance-covariance matrix specification, the classic one $\Gamma_y(0)$, or the one proposed in Peña and Box (1987), provides the accuratest estimation of $\widehat{\Lambda}$, in terms of sample size T, time series dimension N, as well as, the number of lags k_0 considered in the M_k matrix in Equation 3. The idea behind M_k is to accumulate the information from different time lags which makes sense when dealing with time series data. Furthermore, it is shown in Lam and Yao (2012) that when the number of time observations T is small, such specification is particularly useful. The variance-covariance matrix which includes lags is defined by the following equation:

$$M_k = \sum_{k=1}^{k_0} \Gamma_y(k)\Gamma_y(k)' \quad \text{for} \quad k \geq 1 \tag{3}$$

In both specifications, $\Gamma_y(k) = cov(y_{t+k}, y_t)$ and the subscript k in M_k means the number of lags k_0 considered in the sum of Equation 3. What we need to estimate the factor loading space $\mathcal{M}(\Lambda)$ is to implement an eigen decomposition on:

$$\widehat{\Gamma}_y(0) = \frac{1}{T}\sum_{t=k+1}^{T} Y_t Y_t' \tag{4}$$

$$\widehat{M}_k = \sum_{k=1}^{k_0} \widehat{\Gamma}_y(k)\widehat{\Gamma}_y(k)' \quad \text{for} \quad k \geq 1 \tag{5}$$

where $\widehat{\Gamma}_y(k) = \frac{1}{T}\sum_{t=k+1}^{T} Y_t Y'_{t-k}$ is the sample variance-covariance matrix of y_t at lag k, $(k = 0, ..., k_0)$.

The covariance matrix of y_t, $\Gamma_y(k)$ is:

$$\Gamma_y(k) = \Lambda\Gamma_f(k)\Lambda' + \Gamma_e(k) \quad \text{for} \quad k \geq 0 \tag{6}$$

and

$$\begin{aligned} \Gamma_y(k)\Gamma_y(k)' &= [\Lambda\Gamma_f(k)\Lambda' + \Gamma_e(k)]\,[\Lambda\Gamma_f(k)\Lambda' + \Gamma_e(k)'] \\ &= \Lambda\Gamma_f(k)^2\Lambda' + \Gamma_e(k)\Gamma_e(k)' + \Gamma_e(k)\Lambda\Gamma_f(k)\Lambda' + \Lambda\Gamma_f(k)\Lambda'\Gamma_e(k)' \end{aligned} \tag{7}$$

Then the sum of k matrices:

$$\begin{aligned} M_k = \sum_{k=1}^{k_0} \Gamma_y(k)\Gamma_y(k)' = \Lambda \sum_{k=1}^{k_0} \Gamma_f(k)^2\Lambda' + \\ + \sum_{k=1}^{k_0} \Gamma_e(k)^2 + 2\Lambda \sum_{k=1}^{k_0} (\Gamma_f(k)\Lambda'\Gamma_e(k)) \end{aligned} \tag{8}$$

We distinguish four cases depending on the assumptions about the errors $\mathbf{e}_t$: (1) DFM with identical error structure, this is the simplest situation where the errors do not present serial correlation nor cross-correlation. The covariance matrix $\Gamma_e(0) = \sigma^2 I$ is diagonal and $\Gamma_e(k) = 0$ for $k > 0$. Then, it is easy to see that $\Gamma_y(0)$ and $\Gamma_y(k)$ share the same r eigenvectors which are the columns of the loading matrix Λ. (2) DFM with non scalar error structure, where the errors do not present serial correlation neither serial cross-correlation. $\Gamma_e(0)$ is a full rank non diagonal matrix and $\Gamma_e(k) = 0$ for $k > 0$. Then it is easy to see that $\Gamma_y(0)$ and $\Gamma_y(k)$ have different eigenvectors but all the matrices $\Gamma_y(k)$ for $k > 0$ will have the same eigenvectors. (3) DFM where the errors present serial correlation and instantaneous cross-correlation. The error covariance matrix $\Gamma_e(0)$ is non diagonal and has full rank, and $\Gamma_e(k)$ for $k > 0$ is a diagonal matrix. Therefore, all the matrices $\Gamma_y(k)$ for $k \geq 0$, will have different eigenvectors. (4) DFM where erros are serially and cross-sectionally correlated. The covariance matrices $\Gamma_e(k)$ for $k \geq 0$ are full rank and sparse. Then, all matrices $\Gamma_y(k)$ for $k \geq 0$, will have different eigenvectors.

Next section provides a simulation study of cases (1) and (2), where the parameters of interest are the sample size T, the time series dimension N and the number of static common factors r.

3 Simulation study

3.1 Data-generating processes

We consider three data generating processes DGP_1, DGP_2, and DGP_3 with $r = 1, 2, 3$ common latent factors, respectively. Under a stationary framework and in line with the model in Equation 1, each observation is generated by:

$$y_{it} = f_t'\lambda_i + e_{it}, \quad \text{for} \quad i = 1, ..., N \quad \text{and} \quad t = 1, ..., T \tag{9}$$

where the idiosyncratic errors follow Gaussian White Noise processes, such as, $e_{it} \sim WN(0, \sigma_e)$ with $\sigma_e = 2$ under case (1), where errors do not present serial correlation nor cross-correlation. In case (2), the idiosyncratic errors present instantenous cross-correlation and they are generated following the model:

$$e_{it} = \alpha e_{i-1,t} + \sigma_i \varepsilon_{it}, \quad \text{for} \quad i = 1, ..., N \quad \text{and} \quad t = 1, ..., T \tag{10}$$

where the element σ_i has Uniform (1,10) distribution, the noise ε_{it} has Normal (0,1) distribution, and the parameter α takes values from the Uniform (0, 0.7) distribution.
The r-dimensional common factors, f_t, are generated as a vector autoregressive process of order 1 given by the following equation:

$$f_t = \phi f_{t-1} + \eta_t \tag{11}$$

with $\phi = (0.8, 0.5, 0.2)'$. We assume that the errors η_t are independent of the idiosyncratic errors e_t, such that $E(\eta_t e_t') = 0$. Each element λ_i of the factor-loading matrix Λ corresponding to f_{1t} has Uniform (0,1) distribution. The ones corresponding to f_{2t} have Uniform (0,1) distribution when $i = 1, ..., N/2$ and negative Uniform (0,1) when $i = N/2 + 1, ..., N$. The ones associated to f_{3t} have Normal (0,1) distribution when $i = 1, ..., N/2$, and when $i = N/2 + 1, ..., N$, λ_i equals zero.
In the simulation exercise, for DGP_i with $i = 1, 2, 3$, we run 500 iterations for each one of the 16 combinations (N, T) and within each combination the number of lags in $\widehat{M_k}$ matrix are $k_0 = 1, 2, ..., 10, 15, 20$. The sample sizes are, $N = 10, 20, 50, 100$, and the time observations, $T = 50, 100, 200, 500$.

3.2 The angle between subspaces: Canonical Correlations

Our objective is to find out which one of the two variance-covariance matrices $\widehat{\Gamma}_y(0)$ and $\widehat{M_k}$ provides the most accurate estimation taking into account different sample sizes, number of time observations, as well as, the parameter k_0. Note that we have assume a scalar variance-covariance matrix for the noise of the form $\Gamma_e(0) = \sigma^2 I$, so that in theory both procedures would be equivalent, as the eigenvectors of the matrices $\Gamma_y(0)$ and M_k are identical. However, in finite samples we may find differences and this is the situation we will consider first. In a companion paper we analyze the case in which the covariance matrices of the noise have a more general form. We evaluate the performance of each variance-covariance matrix specification using the measure of the angle between the subspaces generated by each estimated $\widehat{\Lambda}$ matrix and the original loading matrix Λ, which is equivalent to the Canonical Correlations (CC) between the estimated matrix of eigenvectors $\widehat{\Lambda}$, and Λ, see Equation 11. Briefly, the canonical correlations represent the relationship of dependence between the subspaces generated by two sets of variables. Let X and Y be the $(n \times p)$ and $(n \times q)$ corresponding matrices of eigenvectors Λ and $\widehat{\Lambda}$, respectively. Then, we seek a linear combination of the X variables (eigenvectors of Λ), which is the most correlated with a linear combination of the Y variables (eigenvectors of $\widehat{\Lambda}$). The eigen decomposition of $PXPY$, where PX and PY define the projector onto the corresponding column-space of X and Y, give the canonical correlations (square roots of the eigenvalues) and the coefficients of linear combinations that define the canonical variates (eigenvectors). In other words, the best loading matrix estimation $\widehat{\Lambda}$ will be the one with larger canonical correlation δ^2. From a geometric point of view, the maximal canonical correlation is equivalent to the cosine of the angle formed by X^* and Y^*, the subspaces generated by X and Y. Then,

$$\cos\theta^2 = \delta^2 = \frac{(\alpha' S_{12} \beta)^2}{(\alpha' S_{11} \alpha)(\beta' S_{22} \beta)} \tag{12}$$

Where S_{ij} is the ML estimation of the matrix V_{ij} being $V_{11} = E[XX']$, $V_{22} = E[YY']$ and $V_{12} = E[XY']$. The vectors α and β are the eigenvectors linked to the largest eigenvalues of matrices $\widehat{A}$ and $\widehat{B}$, respectively, in:

$$\widehat{A}_{p\times p} = S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$$

$$\widehat{B}_{p\times p} = S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}$$

3.3 Simulation results

Let $\widehat{\Lambda}_0$ and $\widehat{\Lambda}_k$ be the estimated Λ matrices with no lags and with lags $k = k_0$, respectively. We write CC_0 to define the maximum canonical correlation for $(\widehat{\Lambda}_0, \Lambda)$ and CC_k the one for $(\widehat{\Lambda}_k, \Lambda)$. For each DGP$_i$, $i = 1, 2, 3$, we have sixteen scenarios in terms of N and T. In case (1) in which idiosyncratic errors do not present serial correlation nor cross-correlation, the analysis just considers the cases when $k_0 = 1$ given that $\Gamma_y(0)$ and $\Gamma_y(k)$ share the same r eigenvectors. Although we expect similar performances under both covariance matrices specifications, the estimate $\widehat{\Lambda}_0$ provides slightly larger canonical correlation coefficients for the most part of scenarios. Nevertheless, there are some scenarios in which the estimate $\widehat{\Lambda}_k$ achieves superior performance. Figures 1 to 3 depict the CC_0 (solid line) and the CC_k (dotted line) for DGP$_1$, DGP$_2$ and DGP$_3$, respectively. The abscissas axis represents the sample size T and from left to right, each plot has a fixed number of time series, $N = 10, 20, 50, 100$. The first plot in Figure 1 (when $N = 10$) shows that when T increases, the estimation for Λ improves, as we expected. This patterns is detected in all plots in Figures 1 to 3. Continuing with the example for $N = 10$ in Figure 1, for $T > 100$ the estimate $\widehat{\Lambda}_k$ achieves better results than the estimate with no lags. This finding shows up when $N = 20$ too (see second plot in Figure 1). Nevertheless, when the data set includes larger number of time series (third and fourth plot in Figure 1), $N \geq 50$, the estimate $\widehat{\Lambda}_0$ would be better. The latter behavior is corroborated also for DGP$_2$ in Figure 2 and DGP$_3$ in Figure 3. Moreover, for DGP_2 and DGP_3, when the dimension of time series is small ($N = 10, 20$ third and fourth plots) and a large number of observation is available ($T \geq 150$ and $T \geq 350$, respectively) it is more recommendable to consider the estimate $\widehat{\Lambda}_k$, as we saw previously for DGP_1 in Figure 1.

In case (2), idiosyncratic errors present a more realistic structure and, as we specified in Section 2, $\Gamma_y(0)$ and M_k have different eigenvectors. Figures 4 to 6 present the simulation results when $\Gamma_e(0)$ is a full rank non diagonal matrix, for DGP$_1$, DGP$_2$ and DGP$_3$, respectively. In Figure 4, for DGP_1 and $k_0 = 2$, the estimator $\widehat{M_k}$ provides the largest canonical correlation coefficients, meaning that the estimates $\widehat{\Lambda}_k$ are better than $\widehat{\Lambda}_0$. This pattern is also observe in Figures 5 for $k_0 = 5$, and Figure 6 for $k_0 = 4$ when $N = 10, 20, 50$ and the number of time observations available is $T > 100$. Bottom right plots in Figures 5 and 6 show that when N increases up to 100, CC_0 are a little larger than CC_k, this is reasonable since $\widehat{\Gamma}_k$ contains less information because of the lags considered.

4 Macroeconomic application

In 1962 the country-members of the European Economic Community (EEC) contemplated the creation of an economic and monetary union in the Marjolin Memorandum. From then, the country-members have followed several steps to achieve the current monetary cooperation known as the European Monetary Union (EMU) or Euro Area (EA). During the last decade, the Financial Crisis in 2008 and the European Debt Sovereign Crisis in 2011, have arisen new concerns about the optimality and sustainability of the EMU. These two crises led to recession phases in the business cycles of the country-members and the different ways of recovering of such periods gave rise to the phenomenon known as *The Two-Speed Europe*. This phenomenon represents the existence of two different groups of countries: the core-countries and the peripheral-countries. Although Euro Area country-members share a common monetary policy, they may face different phases (recession or expansion) of the business cycle. The business cycle (BC from now on) is defined in the literature (see Bai (2003)) as the comovement of economic variables, and the latent common factors in the DFM are interpreted as a proper illustration of such comovements.

This topic about the existence of two different groups within the Euro Area in terms of business cycle synchronization and convergence have been analyzed by Vymyatnina and Antonova (2014) amongothers, who found that the synchronization of GDP and its major components had increased since the creation of the EMU. On the contrary, Artis et al. (2004), Camacho et al. (2006) and Camacho et al. (2008), and Gogas and Kothroulas (2009) conclude that the level of comovements between the country members have not experimented a significant increased. Recently, Klaus and Ferroni (2015) analyzing the four largest European economies identify a considerable economic integration for France, Germany and Italy, but a disconnection for Spanish business cycle. Also, Borsi and Metiu (2015) analysis shows no global convergence in the EU based on per capital real income. Although the issue has been widely analyzed, there is no consensus about which countries would belong to each group, as well as, the possibility of there being a *Multi-Speed Europe*. Some of the reasons of such no consensus may be the time horizon and the macroeconomic indicators considered, together with the country-members included in the sample. Camacho et al. (2006) and Borsi and Metiu (2015) carry out their analysis considering just one macroeconomic indicator, the Industrial Production Index (IPI) and Per capita real income, respectively. Di Giorgio (2016) and Jiménez-Rodríguez et al. (2013) analyze comovements between Euro Area countries and Central and Eastern Europe Countries (CEECs), while Klaus and Ferroni (2015) studies the four largest EA economies. In addition, Breitung and Eickmeier (2006) consider Austria, Belgium, France, Germany, Italy, the Netherlands and Spain as core EA countries. Any of the previous studies have considered all the EMU countries at the same time, nor the existence of a *Multi-Speed Europe*. Nevertheless, the present study fill these gaps in the literature, since the data sample includes macroeconomic series of all EMU country-members and spans the period from the introduction of the single currency until the end of 2017.

As we mentioned in the introduction, one of the main advantages of the DFM is that we will summarize the comovements across the Euro Area country-members during the period of analysis in a reduce number of common factors. We estimate each country-member business cycle with the first principal component obtained from the eigen decomposition of the sample variance-covariance matrix $\widehat{M}_k$ in Equation 5. We have used this matrix because according to Section 3.3 for small number of time series and large number of observations, $\widehat{M}_k$ matrix provides more accurate estimations than $\widehat{\Gamma}_y(0)$. Once we estimate the business cycle of each country, we apply a Hierarchical clustering analysis in order to give answer to the existence of a *Two-* or *Multi-Speed Europe*. Results from this DFM application may provide relevant insights to economic policy makers, as well as, to the public in general.

4.1 Data

The dataset is composed by three macroeconomic indicators for each one of the 19 EA country-members. As suggested by Kose et al. (2003) , macroeconomic series of production, consumption and investment for each country are considered to be good proxies for the estimation of business cycles. In particular, seasonally and calendar adjusted series at a quarterly frequency of Gross Domestic Product (GDP), Household & NPISH Final Consumption Expenditure (CON), and Gross Fixed Capital Formation (INV) were obtained from the Eurostat database. Data availability differs for each one of the EA members, thus, in order to consider a balanced dataset, the sample spans the period between the first quarter of 2000 and the third quarter of 2017, covering the last 17 years since the introduction of the Euro as a single currency in 1999. Previous to the analysis all series were corrected from data anomalies, we use the program TRAMO from Gómez and Maravall (1996), where the outlier detection and correction procedure for each observation consist in computing the t-test for four types of outliers, as in Chen and Liu (1993). Furthermore, data were transformed to achieve stationarity by differencing or log-differencing. Finally, to avoid the problem of series with large-variance when stracting common factors all series were standardized to have zero mean and standard deviation equals to one. In general, quarterly series become stationary after a log-difference transformation.

4.2 Clustering time series by dependency

The algorithm we consider for the cluster analysis is the Hierarchical Agglomerative Clustering (HAC). This technique merges observations from bottom to top: it starts with each observation assigned to its own cluster, in each iteration the two most similar clusters are merged into one . Similarity between clusters is measure by a distance matrix. In this application the distance matrix is the Generalized Cross-Correlation matrix, recently proposed in Alonso and Peña (2017). This measure takes into account all the cross correlations of the observed variables until some lag k, where $k = 1$ in this application. We consider four clustering methodologies within the HAC: Average Linkage, Single Linkage, Complete Linkage and the Ward's method, and choose the one with the largest agglomerative coefficient, which represents the amount of clustering structure found. The dissimilarity measure is defined as:

$$\widehat{GCC}(Y_i, Y_j) = 1 - \frac{|\widehat{R}_{\chi(i,j)}|^{1/2(k+1)}}{|\widehat{R}_{\chi(i)}|^{1/2(k+1)}|\widehat{R}_{\chi(j)}|^{1/2(k+1)}} \tag{13}$$

where $\widehat{R}_\chi$ represents the sample correlation matrices of the χ's data matrices, see section 4 in Alonso and Peña (2017) for a detailed description of the measure.

4.3 Results

First, we take into account the whole dataset and estimate the Euro Area business cycle which give us an useful interpretations of the European Monetary Union' economic performance. As we mentioned above, the estimated common factor is considered as representative of the Euro Area business cycle given that is able to represent the main events that have occured during the period of analysis taking negative values during the two recession periods. Shaded areas in Figure 7 correspond to the Financial Crisis of 2008Q1 and to the European Sovereign Debt Crisis of 2011Q3 established by the Euro Area Business Cycle Dating Committee. Moreover, this estimated common factor is able to represent 42.05% of the variability present in the original data. Figure 8 displays the impact (sum of loading coefficients in absolute value) of each country explaining the Euro Area business cycle, and Figure 9 summarizes the effect in terms of Gross Domestic Product, consumption and investment indicators. In general, it represents the economic behavior of most countries in the Euro Area during the last 17 years, although we can see minor influences of countries such as Estonia, Greece, Luxembourg, Malta and Slovekia. In terms of macroeconomic indicators, GDP series are the most influential explaining comovements between countries.

The second application consists in estimating each country-member's business cycle and applying the clustering methodology to them in order to analize the groups' memberships. We consider the Ward's method since it presents the largest agglomerative coefficient. The objective of this method is to minimize the total within-cluster variance. Analyzing the dendogram of the Ward's method in Figure 11 we can distinguish two main clusters, which are supported by the results of the Elbow and the Average Silhoutte methods in Figure 10. These methods consist in optimizing a criterion: in the Elbow method, the optimal number of clusters is the one with minimum total within cluster Sum of Squares, and in the Average Silhoutte the one that maximizes the average silhouette. This result confirms the phenomenom of a Two-speed Europe, but the main finding is that groups do not have a geographic interpretation. One group is composed of Austria, Belgium, Cyprus, Finland, Germany, Greece, Irland, Italy, Latvie, Lithuania, Malta, The Netherlands, Portugal, Slovakia, Slovenia, and Spain; meanwhile just three countries formed group 2: Estonia, France, and Luxembourg. Even though the optimal number of clusters is two, it is worthy of our attention the likely existence of four clusters in Figure 12. Group 1 would be formed by Austria, Irland, Malta, Slovakia, and Slovenia, group 2 by Belgium, Cyprus, Finland, Greece, Italy, Latvia, and Lithuania, group 3 by Germany, the Netherlands, Portugal and Spain, and finally, group 4 by Estonia, France, and Luxembourg.

Results from the clustering of the business cycles of the Euro Area country-members are consistent with the literature about the existence of a *Two-speed Europe*, although such groups of countries do not represent geographical regions within the European Monetary Union as it has been assumed previously.

5 Concluding remarks

This article has evaluated the performance of the Pincipal Component estimator under two settings; when the sample covariance matrix of y_t only includes current information, and when it considers past information. Some simulation experiments have been conducted to analyze for which (N, T) combinations would be more advantageous to consider $\widehat{\Gamma}_0$ or $\widehat{M}_k$ under two scenarios depending on the idiosyncratic error structure. Simulation results show that when idiosyncratic errors are assumed to have a scalar structure both covariance matrix specifications provide similar

Canonical Correlation coefficients. Nevertheless, when errors are assumed to have a more realistic structure, as in case (2) with instantaneous cross-correlation, the $\widehat{M_k}$ matrix, which includes past information, performs better than $\widehat{\Gamma}_0$, which only considers current information. Just when the size of time series N grows more than proportionally with respect to the number of time observations T, we obtain slightly larger Canonical Correlation coeffients from $\widehat{\Gamma}_0$. The empirical application of the Euro Area dataset shows the usefulness of the Principal Component estimator with $\widehat{\Lambda}_k$, for analyzing and evaluating the economic behaviour of countries or regions of countries. Moreover, the estimation of the Euro Area business cycles together with the clustering analysis provide us relevant insights about the phenomenon of the Two-Speed Europe.

This work can be extended in many dimensions. The first, that is currently under investigation, is to consider different characterizations of the idiosyncratic comonent e_t as in cases (3), where errors present serial correlation and instantaneous cross-correlation, and case (4), where erros are serially and cross-sectionally correlated. A second theoretical extension may be to include common latent factors with lags in Equation 1, or to move beyond stationary settings.

References

Ahn, S. K. and Reinsel, G. C. (1990). Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association*, 85(411):813–823.

Alonso, A. and Peña, D. (2017). Clustering time series by dependency. *Working Paper of the Statistics Department at Carlos III University of Madrid*.

Artis, M., Krolzig, H.-M., and Toro, J. (2004). The european business cycle. *Oxford Economic Papers*, 56(1):1–44.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Bai, J., Ng, S., et al. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3(2):89–163.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Borsi, M. T. and Metiu, N. (2015). The evolution of economic convergence in the european union. *Empirical Economics*, 48(2):657–681.

Breitung, J. and Eickmeier, S. (2006). Dynamic factor models. *Allgemeines Statistisches Archiv*, 90(1):27–42.

Camacho, M., Perez-Quiros, G., and Saiz, L. (2006). Are european business cycles close enough to be just one? *Journal of Economic Dynamics and Control*, 30(9-10):1687–1706.

Camacho, M., Perez-Quiros, G., and Saiz, L. (2008). Do european business cycles look like one? *Journal of Economic Dynamics and Control*, 32(7):2165–2190.

Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421):284–297.

Connor, G. and Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of financial economics*, 15(3):373–394.

Di Giorgio, C. (2016). Business cycle synchronization of ceecs with the euro area: A regime switching approach. *JCMS: Journal of Common Market Studies*, 54(2):284–300.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4):540–554.

Forni, M. and Reichlin, L. (1998). Let's get real: a factor analytical approach to disaggregated business cycle dynamics. *The Review of Economic Studies*, 65(3):453–473.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models*.

Gogas, P. and Kothroulas, G. (2009). Two speed europe and business cycle synchronization in the european union: The effect of the common currency.

Gómez, V. and Maravall, A. (1996). Programs tramo (time series regression with arima noise, missing observations, and outliers) and seats (signal extraction in arima time series). instructions for the user. *Documento de Trabajo*, 9628.

Jiménez-Rodríguez, R., Morales-Zumaquero, A., and Égert, B. (2013). Business cycle synchronization between euro area and central and eastern european countries. *Review of Development Economics*, 17(2):379–395.

Klaus, B. and Ferroni, F. (2015). Euro area business cycles in turbulent times: convergence or decoupling? Technical report, ECB Working Paper.

Kose, M. A., Otrok, C., and Whiteman, C. H. (2003). International business cycles: World, region, and country-specific factors. *american economic review*, 93(4):1216–1239.

Ku, W., Storer, R. H., and Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and intelligent laboratory systems*, 30(1):179–196.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Motta, G., Hafner, C. M., and von Sachs, R. (2011). Locally stationary factor models: Identification and nonparametric estimation. *Econometric Theory*, 27(6):1279–1319.

Motta, G. and Ombao, H. (2012). Evolutionary factor analysis of replicated time series. *Biometrics*, 68(3):825–836.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, 95(2):365–379.

Peña, D. (2013). *Análisis de datos multivariantes*. McGraw-Hill España.

Peña, D. (2014). Big data and statistics: Trend or change? *BEIO, Boletín de Estadística e Investigación Operativa*, 30(3):313–324.

Peña, D. and Box, G. E. (1987). Identifying a simplifying structure in time series. *Journal of the American statistical Association*, 82(399):836–843.

Peña, D. and Poncela, P. (2006). Dimension reduction in multivariate time series. In *Advances in distribution theory, order statistics, and inference*, pages 433–458. Springer.

Peña, D. and Yohai, V. J. (2016). Generalized dynamic principal components. *Journal of the American Statistical Association*, 111(515):1121–1131.

Sargent, T. J., Sims, C. A., et al. (1977). Business cycle modeling without pretending to have too much a priori economic theory. *New methods in business cycle research*, 1:145–168.

Stock, J. H. and Watson, M. (2011). Dynamic factor models. *Oxford handbook on economic forecasting*.

Stock, J. H. and Watson, M. W. (1988). Testing for common trends. *Journal of the American statistical Association*, 83(404):1097–1107.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics*, volume 2, pages 415–525. Elsevier.

Tiao, G. C. and Tsay, R. S. (1989). Model specification in multivariate time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–213.

Vymyatnina, Y. and Antonova, D. (2014). *Creating a Eurasian Union: Economic integration of the former Soviet republics*. Springer.

Appendix

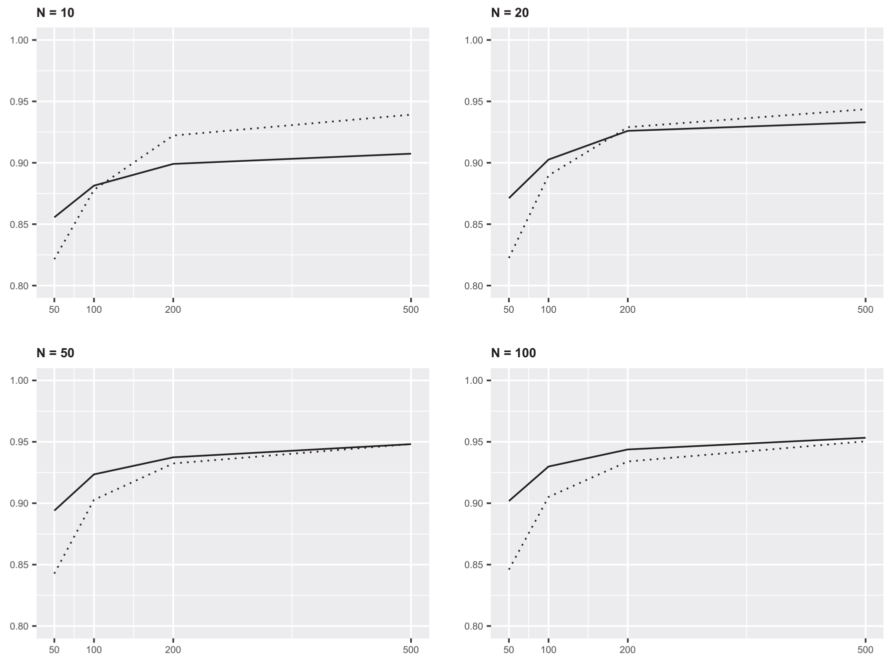

Figure 1: Similarity between theoretical and estimated subspaces measured by the maximum canonical correlation for the model with one common factor (DGP_1). CC_0 (solid line) and CC_k with $k = 1$ (dotted line). T in abscissa axis.

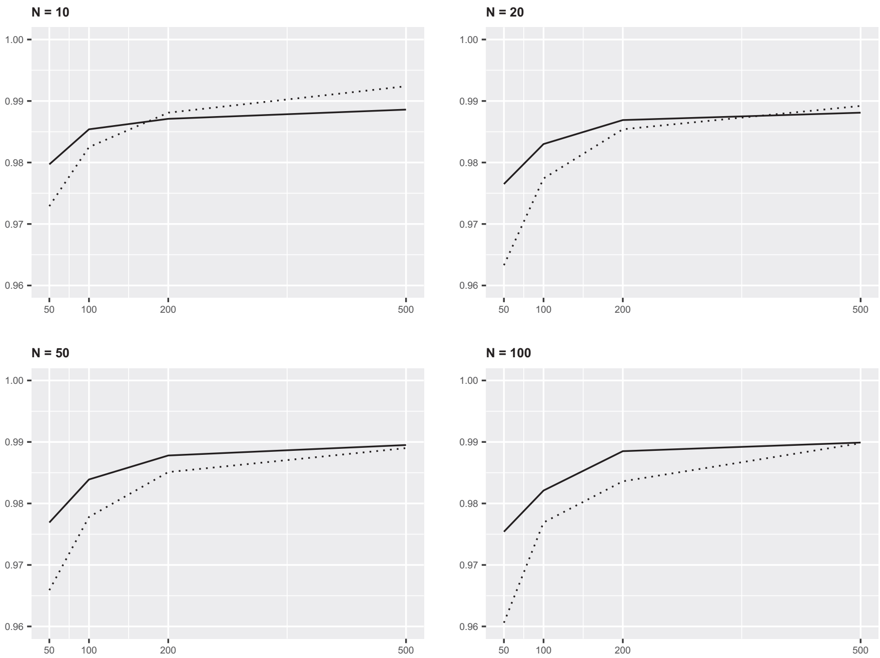

Figure 2: Similarity between theoretical and estimated subspaces measured by the maximum canonical correlation for the model with two common factors (DGP_2). CC_0 (solid line) and CC_k with $k = 1$ (dotted line). T in abscissa axis.

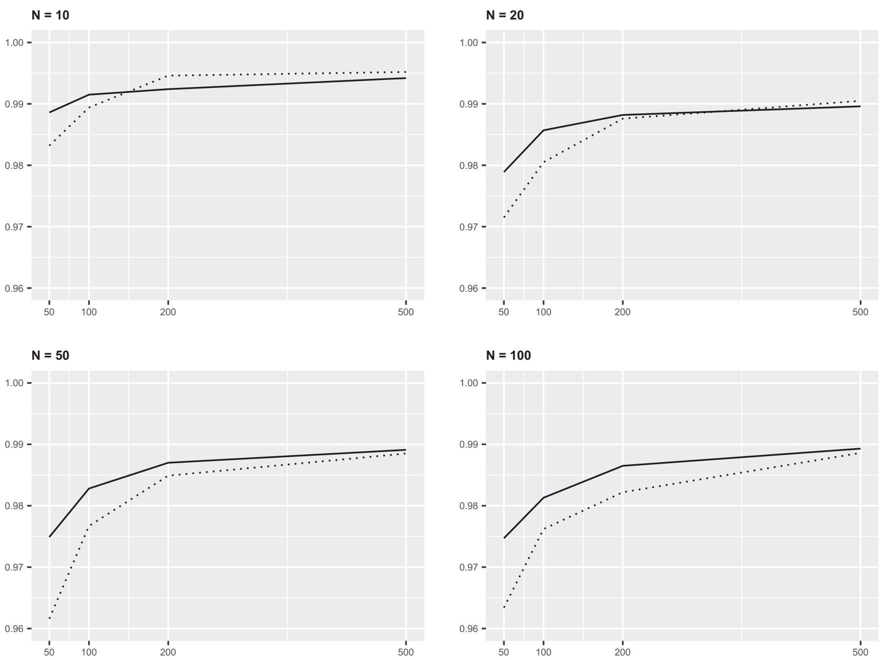

Figure 3: Similarity between theoretical and estimated subspaces measured by the maximum canonical correlation for the model with three common factors (DGP_3). CC_0 (solid line) and CC_k with $k = 1$ (dotted line). T in abscissa axis.

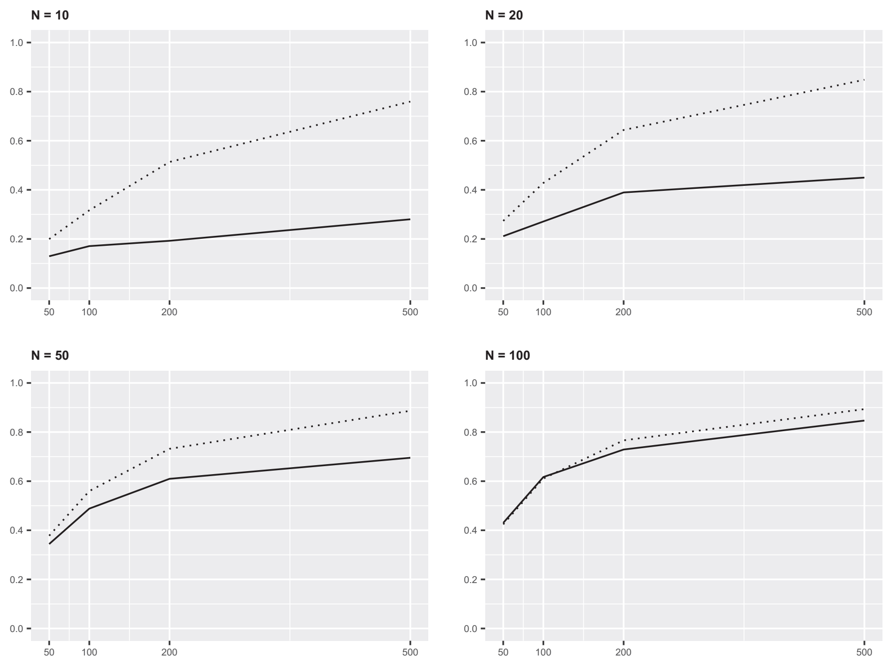

Figure 4: Similarity between theoretical and estimated subspaces measured by the maximum canonical correlation for the model with one common factor (DGP_1). CC_0 (solid line) and CC_k with $k = 2$ (dotted line). T in abscissa axis.

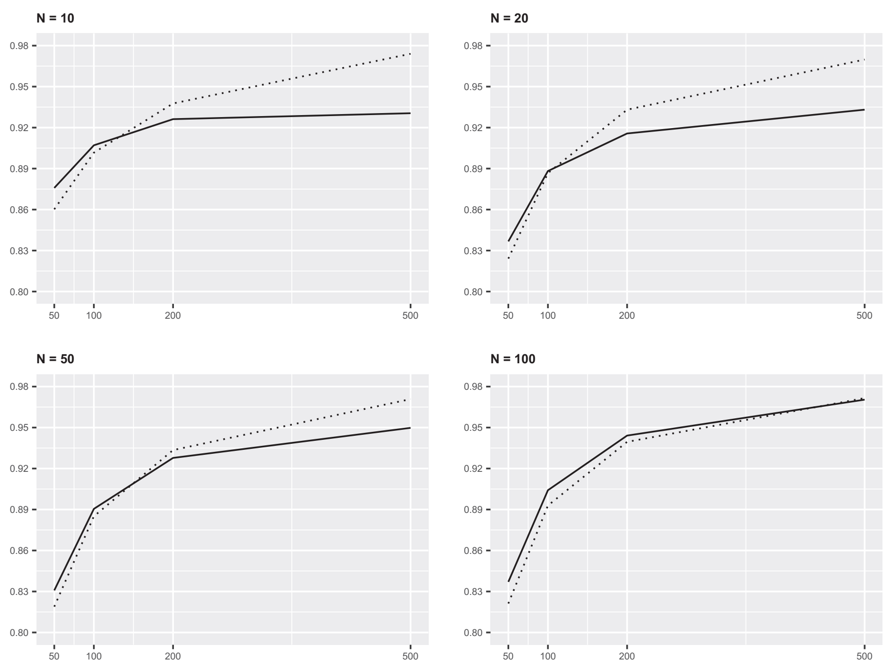

Figure 5: Similarity between theoretical and estimated subspaces measured by the maximum canonical correlation for the model with two common factors (DGP_2). CC_0 (solid line) and CC_k with $k = 5$ (dotted line). T in abscissa axis.

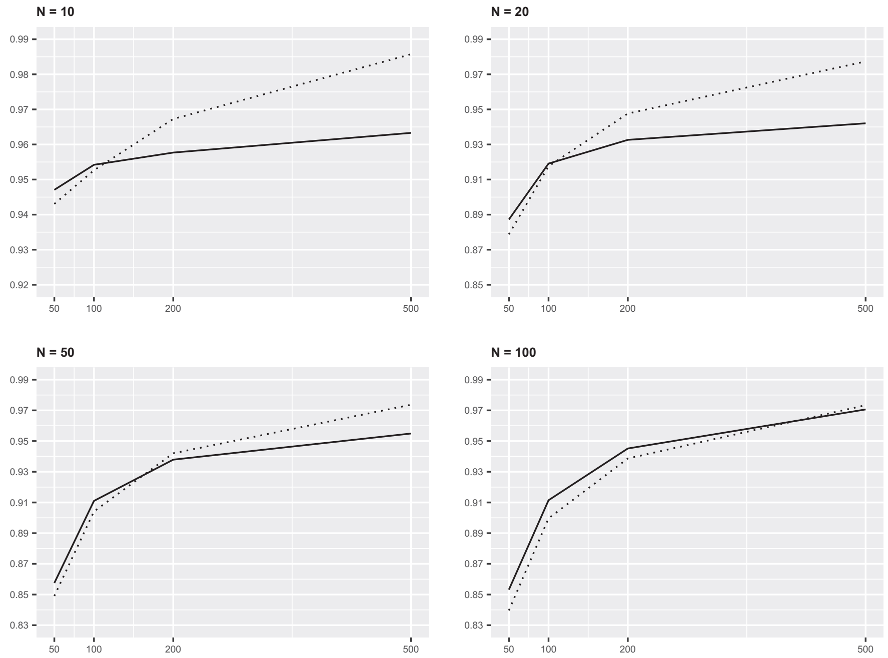

Figure 6: Similarity between theoretical and estimated subspaces measured by the maximum canonical correlation for the model with three common factors (DGP_3). CC_0 (solid line) and CC_k with $k = 4$ (dotted line). T in abscissa axis.

Figure 7: Estimated Euro Area Business Cycle.

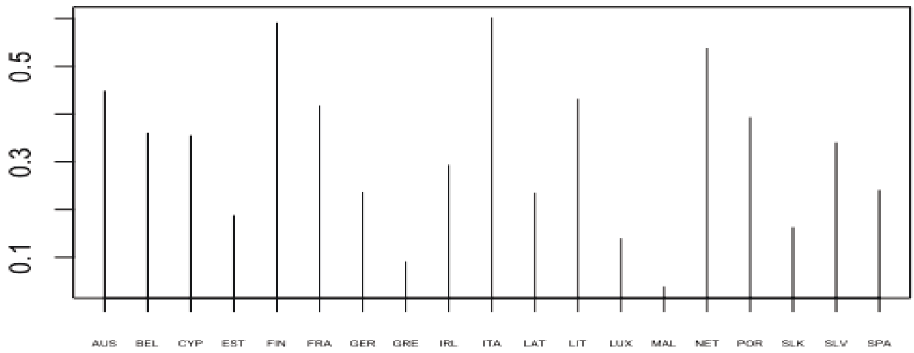

Figure 8: Loadings coefficients by country for the estimated Euro Area Business Cycle.

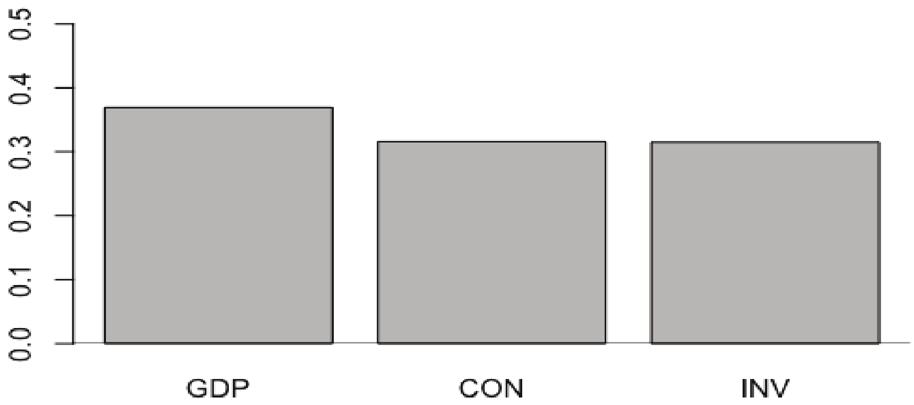

Figure 9: Loadings by macroeconomic series for the estimated Euro Area Business Cycle.

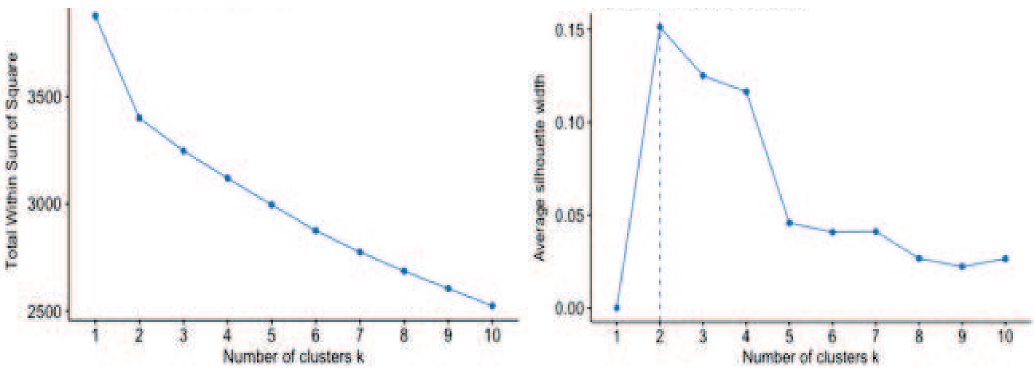

Figure 10: Elbow and Average Silhoutte methods for the optimal number of clusters of the Euro Area country-members.

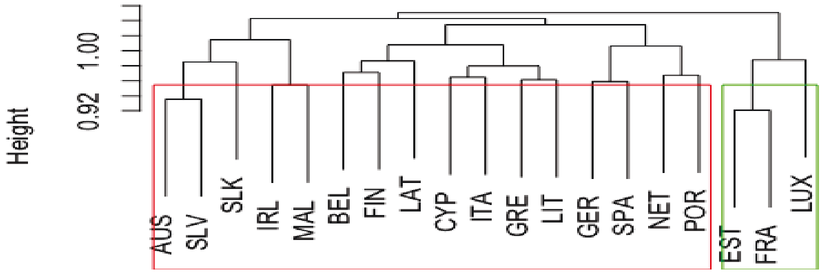

Figure 11: Dendogram from the Euro Area country-members clustering under Ward's method in two groups.

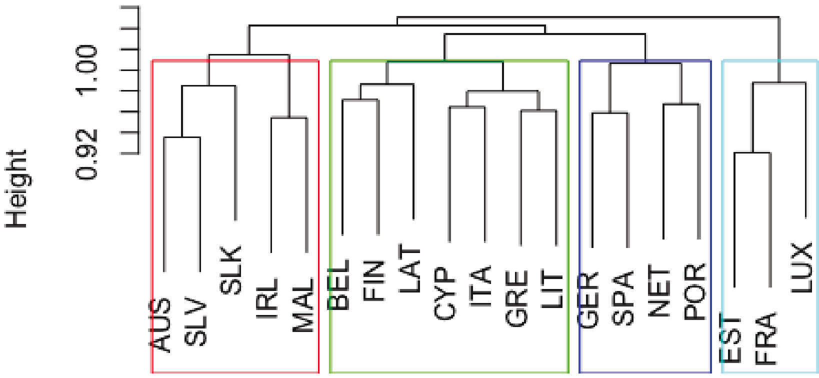

Figure 12: Dendogram from the Euro Area country-members clustering under Ward's method in four groups.