Dynamic Factor Models for Heterogeneous Data

by

Ángela Caro Navarro

A dissertation submitted by in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business and Quantitative Methods

Universidad Carlos III de Madrid

Advisors:
Máximo Camacho Alonso
Daniel Peña Sánchez de Rivera

October, 2020

A mis padres

Acknowledgements

Throughout the writing of this thesis I have received a great deal of support and assistance.

First, I would like to thank the Ministerio de Ciencia, Innovación y Universidades for the financial support under Ayuda para la Formación del Profesorado Universitario with reference FPU15/03983.

Second, I would like to thank my supervisors, Máximo Camacho and Daniel Peña, for all their dedication, for their invaluable experience, for their patience, for all the opportunities, and above all, for guiding me in my career as a researcher. I will always be grateful.

I am very grateful to Professor Koopman for its invitation to the Department of Econometrics and Data Science at the Vrije University of Amsterdam, and to Professor Qiwei for its invitation to the Department of Statistics at the London School of Economics and Political Sciences. I would like to thank them for their dedication and for their valuable advices and recommendations.

I would like to thank all the members of the Statistics Department at Carlos III University of Madrid, especially Helena, Juanmi, Andrés and Pedro. Thanks to Vanesa and Javier for all the good times we have spent together inside and outside the university and thanks to my big academic brother Germán, everything started with him.

Finally, I would like to thank my family, my friends and Antonio, for supporting me in the most difficult times and for celebrating the successes with me. Thanks to their trips to Amsterdam, London and Madrid I have never felt far from home.

Published and submitted contents

- Published contents:
  - Caro, A. and Peña, D. (2018). Estimation of the common component in Dynamic Factor Models. Universidad Carlos III de Madrid, Departamento de Estadística.
    * http://hdl.handle.net/10016/27047.
    * Co-author.
    * It is partially included in Chapter 3.
    * The material from this source included in this thesis is not indicated by typographical means or references.
  - Camacho, M., Caro, A., Lopez-Buenache, G. (2019). The two-speed Europe in business cycle synchronization. *Empirical Economics* 59, 1069–1084 (2020).
    * https://doi.org/10.1007/s00181-019-01730-4.
    * Co-author.
    * It is fully included in Chapter 4.
    * The material from this source included in this thesis is not indicated by typographical means or references.

- Contents submitted for publication:
  - Caro, A. and Peña, D. (2020). A Test for the Number of Factors in Dynamic Factor Models. Submitted.
    * Co-author.
    * It is fully included in Chapter 2.
    * The material from this source included in this thesis is not indicated by typographical means or references.

Abstract

The thesis *Dynamic Factor Models for Heterogeneous Data* has two major purposes: (1) to investigate the advantages and disadvantages of different Dynamic Factor Model (DFM) estimation methodologies and (2) to show DFM usefulness in real data applications. This thesis includes a literature review in the introduction. Chapter 2 presents a new approach for the estimation of the number of factors using an eigenvalues ratio test. Chapter 3 generalizes the proposed method in chapter 2 for the estimation of the factor space. Chapter 4 studies the business cycles synchronization between Euro Area countries by means of a DFM with known cluster structure and Chapter 5 analyses international energy prices interrelations using DFM with unknown cluster structure. Simulation results suggest that the new approach proposed in this thesis for finding the number of factors and estimating them, based on lagged correlation matrices, provides a good performance compared to methods already presented in the literature. Specially, when the data sample includes atypical series the proposed method outperforms its competitors. This is also corroborated by real data examples.

Resumen

La tesis *Dynamic Factor Models for Heterogeneous Data* tiene dos propósitos principales: (1) investigar las ventajas y desventajas de diferentes metodologías de estimación del Modelo de Factores Dinámicos (MFD) y (2) mostrar la utilidad del MFD en aplicaciones de datos reales. Esta tesis incluye una revisión bibliográfica en la introducción. El capítulo 2 presenta un nuevo enfoque para la estimación del número de factores utilizando un test basado en el uso de valores propios. El capítulo 3 generaliza el método propuesto en el capítulo 2 para la estimación del espacio factorial. El capítulo 4 estudia la sincronización de los ciclos económicos entre los países de la Zona Euro mediante un MFD con estructura de clúster conocida y el capítulo 5 analiza las interrelaciones existentes entre los precios internacionales de la energá utilizando un MFD con estructura de clúster desconocida. Los resultados de las simulaciones sugieren que el nuevo enfoque propuesto en esta tesis para determinar el número de factores y para la estimació de dichos factores, basado en matrices de correlación rezagadas, proporciona un buen desempeño en comparación con los métodos ya presentados en la literatura. Especialmente, cuando la muestra de datos incluye series atípicas, el método propuesto supera a sus competidores. Esto también se corrobora con ejemplos de datos reales.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: Dynamic Factor Models

One of the major limitations in classical econometric models and multivariate time series models is that the number of parameters to estimate increases with the square of the dimension of the vector of time series, with the consequently loss of degrees of freedom. When addressing empirical issues, it is crucial to find simplified structures which can be correctly estimated. As a solution to the problem of dimensionality, factor models have become one of the most useful tools between researchers and practitioners. This section reviews the main characteristics of this methodology and its different specifications along literature.

First applications of Dynamic Factor Models (DFM) to macroeconomic series were originally proposed by Geweke (1977) and Sargent et al. (1977), as an extension of the classical static factor models to the field of time series, and were initially known as *index* models. Since then, an extensive literature, both theoretical and empirical, about them has been developed. The main idea behind factor decomposition in time series analysis is that the co-movements of a high-dimensional vector of observed variables, $\mathbf{y}_t$, are driven by two mutually orthogonal components: a small number of latent dynamic factors, $\mathbf{F}_t$, and a vector of mean-zero idiosyncratic disturbances, $\mathbf{e}_t$, that are specific to an individual series. Let us consider the following factor model representation for observation y_{it}, with $i = 1, ..., N$ where N is the number of cross-section units and $t = 1, ..., T$ where T is the number of time series observations:

$$y_{it} = \mathbf{p}_i' \mathbf{F}_t + e_{it}, \tag{1.1}$$

each component of the $r \times 1)$ vector $\mathbf{p}_i$, given by $\mathbf{p}_{ij}$ for $i = 1, ..., N$ and $j = 1, ..., r$, is known as the *factor loading*, where r is the number of latent factors, and $\mathbf{p}_i' \mathbf{F}_t$ is con-

sidered as the *common component* for such observation. It is usually assume that the latent factors follow autoregressive dynamics of order p, such as

$$\mathbf{F}_t = \phi_1 \mathbf{F}_{t-1} + ... + \phi_p \mathbf{F}_{t-p} + \eta_t, \tag{1.2}$$

where $\phi_1, ..., \phi_p$ are the autoregressive coefficients and the factor innovation, η_t, is a Gaussian white noise vector with positive and finite covariance matrix Γ_η, independently distributed for all leads and lags.

In vector representation, the DFM is defined by the two equations:

$$\mathbf{y}_t = \mathbf{P}\mathbf{F}_t + \mathbf{e}_t, \tag{1.3}$$

$$\Phi(L)\mathbf{F}_t = \eta_t, \tag{1.4}$$

where $\mathbf{y}_t = (y_{1t}, y_{2t}, ..., y_{Nt})'$ and $\mathbf{e}_t = (e_{1t}, e_{2t}, ..., e_{Nt})'$ are $N \times 1$, and $\mathbf{F}_t$ and η_t are $r \times 1$. The factor loading matrix given by $\mathbf{P} = (\mathbf{p}_1, ..., \mathbf{p}_N)'$ is $N \times r$ and in equation (1.4), $\Phi(L) = (I - \phi_1 L - ... - \phi_p L^p)$ is a polynomial of the lag operator L, which can be of infinite order.

Given the advantages of these models for dimension reduction, the state-of-the-art about DFM has distinguished different versions and implementations. Attending to the literature, we consider here two possible classifications of the DFM: one depending on the amount of observable time series, N, used for the estimation of the latent factors; and the other depending on the assumptions made on $\mathbf{F}_t$ in order to be *common* and on $\mathbf{e}_t$ in order to be *idiosyncratic*.

Depending on the number of series, N, included in $\mathbf{y}_t$, DFMs can be considered as *small scale* or *large scale*. Methods applied in the estimation procedure will be different depending on the size of N. Due to the small number of series to be included in the estimation procedure in *small scale* DFMs, factors and parameters use to be estimated by means of the Maximum Likelihood (ML) via the Kalman filter and smoother (KFS), see e.g. Engle and Watson (1981, 1983), Stock and Watson (1989), Sargent (1989), and Quah and Sargent (1993). Differently, *large scale* DFMs allow to include a larger amount of information and Principal Component Analysis (PCA) is considered one of the most useful methodologies in order to estimate the common factors, see Stock and Watson (2002) and Forni et al. (2005) for a review of these methods. In past years, hybrid approaches attracted the attention of researchers. These approaches combined both methodologies, KFS and PCA, and were introduced in Doz et al. (2012) and applied for example in Giannone et al. (2008). Some studies have addressed the properties of the estimation methodologies applied in small scale and large scale DFMs, see e.g.

Alvarez et al. (2016) for DFMs that use aggregate and disaggregate data, and Poncela and Ruiz (2016) which compare PC, KFS, 2SKF (Two Steps Kalman Filter) and QML (Quasi-Maximum Likelihood) given the sample size of simulated and real data. Apart from these methodologies, some authors have considered the Bayesian estimation approach based on Markov Chain Monte Carlo (MCMC) methods, see for example Otrok and Whiteman (1998) for an estimation of the single factor model with application to economic activity in Iowa, and Kose et al. (2003, 2008) and Crucini et al. (2011) for an application of the multifactor model to illustrate global, regional and country business cycles and shocks. The latter studies, address the drawback about the pervasiveness of the factors in cross-country data sets, which is not a realistic assumption given that some latent factors could summarize the co-movements in a given country while not affecting others.

Depending on different characterizations of the idiosyncratic errors or disturbances, $\mathbf{e}_t$, there exist a variety of definitions of DFMs such as *exact*, *approximate* or *strict*. The strict and exact DFMs make the assumption that idiosyncratic errors are cross-sectionally strictly orthogonal to each other at all leads and lags. Nevertheless, while in the strict DFM, see Chamberlain and Rothschild (1983), errors present also serial independence, in the exact DFM of Sargent et al. (1977), this assumption is relaxed allowing serial correlation in the errors. In the approximate or *weak* DFMs previous assumptions are relaxed allowing serial correlation and 'limited' cross-correlation between the idiosyncratic errors, see Doz et al. (2012) for a detailed description of these models.

Along literature, two different representations of the DFM have been considered, *static* and *dynamic*, depending on the way in which the dynamic of the common component is introduced in the model. The word 'static' implies that all common dynamics features are introduced in (1.1) or (1.3) contemporaneously, although the static factors contain current and past values of the dynamic factors. The DFM in (1.3) and (1.4) could then be rewritten in dynamic form as:

$$\mathbf{y}_t = \mathbf{P}(L)\mathbf{f}_t + \mathbf{e}_t, \tag{1.5}$$

$$\mathbf{f}_t = C(L)\varepsilon_t, \tag{1.6}$$

where $\mathbf{P}(L) = (1 - p_1 L - ... - p_s L^s)$ is a vector of dynamic factor loadings of order s, and ε_t are iid errors.The vectors $\mathbf{f}_t$ and ε_t have dimension q, with q equal to the number of dynamic factors. The DFM previously described in (1.3) and (1.4) corresponds with the case in which the number of lags, s, is finite. Instead, when s is allowed to be infinite the model is the so called Generalized Dynamic Factor Model (GDFM) proposed in

Forni et al. (2000). If there are q dynamic factors, the model in (1.5) and (1.6) could be rewritten as a static model with r factors, where $r = q(s+1) \geq q$. It is worth noting that the dimension of $\mathbf{F}_t$ in (1.3) is expected to be different from the dimension of $\mathbf{f}_t$ given that F_t will include all leads and lags of $\mathbf{f}_t$. In what follows, DFM means that the latent common factors follow time series processes, and depending whether the factor loadings are constant or follow a dynamic process, these models are considered to have a 'static' or 'dynamic' representation, respectively.

Finally, the distinction existing between DFMs and GDFMs has to do also with the assumptions related with the underlying data-generating process. As mentioned in Hallin and Lippi (2013), most of the application of DFMs has the nature of 'statistics models' in the sense that they impose restrictions about the underlying data-generating process. Usually, these models have assumed that all the processes in (1.3) and (1.4) are stationary, as it is the case in Peña and Box (1987), Stock and Watson (1988), Bai and Ng (2002), and Lam and Yao (2012). Alternatively, Peña and Poncela (2006a) assumes nonstationarity for integrated process, Pan and Yao (2008) for general processes and Motta et al. (2011), Motta and Ombao (2012) for locally stationary processes. Such models which impose a structure in the underlying data-generating process have been considered suitable along literature when restrictions could be satisfied by the observed data or when they lead to good approximations. Nevertheless, when applied to real data those assumptions can be misleading and difficult to prove from observations. This drawback suggests to apply instead of the classical DFM the GDFM, which does not restric the number of lags in the factors and allows a low correlation between idiosyncratic components or noises, going further from the general structure assumptions (like stationarity). Studies in this topic are Forni et al. (2015) which proposed a model with possibly infinite-dimensional factor spaces and obtained a one-sided representation for the dynamic factor model, and Peña and Yohai (2016) which introduced Generalized Dynamic Principal Components (GDPC) as a generalization of the pioneer work of Brillinger (1981), showing how to reconstruct data set generated by GDPC.

Bai et al. (2008) provides a technical review dedicated mostly to the econometric theory and the restrictions presented in different specifications of DFMs. They focus on the use of estimated factors in subsequent estimation and inference, and differentiate between the static and dynamic representations of the model, introducing the assumptions taken into account in classical factor analysis, which are relaxed in the so-called new generation of 'large dimensional approximate factor models'. Stock and Watson (2011) offers a review focusing on the applications and empirical findings in terms of factor estimation, the determination of the number of factors and the uses of the estimated factors, as well as, some important extensions of DFMs. Finally, Hallin and

Lippi (2013) summarizes the methodological foundations of factor models, contrasting concepts as commonality and idiosyncrasy, factors and common shocks, dynamic and static factor models. The latter focuses on the GDFM, and defines the rest of factor models (static-dynamic, exact-approximate) as particular cases of the GDFM under the assumption of second-order stationarity.

The rest of the thesis is organized as follows: Chapter 2 presents a new eigenvalues ratio test for the number of factors with an empirical application to international business cycles. Chapter 3 introduces an improved method for the estimation of the common component with an empirical application to CO_2 emissions. Chapters 4 and 5 deal with the Dynamic Factor Model with Cluster Structure (DFMCS): Chapter 4 applies a DFM with known cluster structure to macroeconomic data in order to evaluate synchronization between business cycles in the Euro Area; Chapter 5 analyzes international energy prices by means of a DFM with unknown cluster structure. At the end of each chapter concluding remarks and future research extensions are given.

Chapter 2

A new eigenvalues ratio test for the number of factors

As we mention in the introduction, Dynamic Factor Models (DFM) are a useful approach to model and forecasts large sets of big dependent data. Nevertheless, an important problem in these models is to determine the number of common factors. First procedures considered heuristic inspection of eigenvalues of the lagged covariance matrices, see Peña and Box (1987), but three main approaches to solve this problem have been proposed. The first one is based on canonical correlation analysis, Tiao and Tsay (1989) used this technique to check the rank of some moment matrices with a chi-square statistic. A similar test for finding the number of factors in a static or exact dynamic factor model (EDFM) without lagged factor effects was proposed by Peña and Poncela (2006b). These authors showed that the number of common factors is equivalent to the number of non-zero canonical correlations between the vector of series and their lags. Jacobs and Otter (2008) related this test to entropy and used them for DFM with lagged factor effects. This type of test works well when the sample size, T, is much larger than the number of series, N, but its power deteriorates when N is large. The second alternative is to select the number of factors by an information criterion. As the BIC or AIC criteria are not appropriate when N is large and close to T, Bai and Ng (2002) have provided consistent model selection criteria that seem to work well in high dimensional cases. A third approach is to used the ratios of consecutive eigenvalues (in a proper order) of the covariance or spectral matrices. Hallin and Liška (2007) explored tests for the number of divergent eigenvalues in the spectral matrices to find the number of factors in a DFM with lagged factor effects, Onatski (2010) and Ahn and Horenstein (2013) used the covariance matrix, and Lam and Yao (2012) the cumulative sum of lagged covariance matrices for the number of static factors, or factors without lagged effects. The main

idea is that, under the DFM, the eigenvalues of these matrices can be separated into r spike eigenvalues that diverge to infinity and $N - r$ bounded eigenvalues. In addition to this three main approaches other procedures have been proposed, as estimating the number of factors in different subsamples and checking the stability of the results, see Hallin and Liška (2007) and Alessi et al. (2010). Recently, Fan et al. (2019) based on the random matrix theory have proposed an adjusted correlation threshold for determining the number of common factors in high dimensional static factor models.

In this chapter we propose a criterion for the estimation of the number of static factors, it can be seen like a test, as it is called by Ahn and Horenstein (2013), or like a criterion or rule, as it is called by Lam and Yao (2012). The proposed criterion is based on eigenvalue ratios and it combines the advantages of those proposed by Ahn and Horenstein (2013) and Lam and Yao (2012) (AH and LY from now on) and adds four others. First, it is based on the correlation instead of the covariance matrices and, therefore, the test is robust to a few atypical series with large variance that can dominate the results of a test based on the eigenvalues of the covariance matrices. Second, the new test uses all the information available about the dependency among the series as it incorporates both the information about the lag zero dependency (as the AH test) and the positive lagged dependency (as the LY criterion). Third, instead of adding the lagged covariance matrices they are combined with weights that depend on the precision estimation of each matrix. Fourth, when the series are heteroscedastic theoretical reasons are given to justify that the ratios of eigenvalues of correlation matrices are expected to be more powerful to detect the number of factors than those from the covariance matrices.

The rest of this chapter is organized as follows. In the next section we define the notation and the background of the dynamic factor model. In Section 2 tests for the number of factors based on the eigenvalues of covariance matrices are reviewed. In particular, we analyze the pros and cons of the AH and LY criteria based on eigenvalues ratios. Section 3 presents the new test and justify some possible advantages of the ratio of eigenvalues from correlation matrices to find the number of factor with heteroscedastic time series. Section 4 includes a Monte Carlo experiment to compare its performance to AH and LY criteria. Section 5 discusses an example of finding the factors with economic data where the proposed test leads to more interesting results than the AH and LY criteria. Finally, Section 6 presents some concluding remarks.

2.1 Theoretical framework

We consider here the static representation of the DFM. Let r represents the number of common factors, a model for the observed data $\mathbf{y}_t$ for $t = 1, ..., T$ is:

$$\mathbf{y}_t = \mathbf{P}\mathbf{f}_t + \mathbf{e}_t, \tag{2.1}$$

where $\mathbf{P}$ is a $(N \times r)$ matrix of factor loadings, $\mathbf{f}_t$ is a $(r \times 1)$ vector of common factors, and $\mathbf{e}_t$ is a $(N \times 1)$ vector of idiosyncratic disturbances or errors. We assume for model identification that $\mathbf{P}'\mathbf{P} = \mathbf{I}_r$, and also that $\mathbf{\Gamma}_f(k) = \text{cov}(\mathbf{f}_t, \mathbf{f}_{t-k}) \neq 0$ for some $k > 0$, meaning that the factors may present serial correlation. With the assumption that $\mathbf{e}_t$ is a white noise process the model is identified in finite samples and is called the Exact DFM or EDFM. However, this hypothesis is often unrealistic in practice. Nevertheless, we can allow weak autocorrelation and cross-section correlation in the noise or idiosyncratic term under assumptions that imply that this dynamic vanishes asymptotically, whereas the factor dynamics remains. For instance, Stock and Watson (2002) assume that $\frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t\mathbf{f}_t' \xrightarrow{p} \mathbf{\Gamma}_f(0) > 0$ for a $r \times r$ non-random diagonal matrix $\mathbf{\Gamma}_f(0)$ with diagonal elements $\gamma_i(0)$ for $i = 1, ...r$ and $N^{-1}\mathbf{P}'\mathbf{P} \xrightarrow{p} \boldsymbol{I}_r$. These two assumptions provide that the factors are pervasive, affecting almost all the series, and that the average signal provided for the factors does not disappear asymptotically. Note that, from (2.1), each series is given by:

$$y_{it} = \mathbf{p}_i'\mathbf{f}_t + e_{it}, \tag{2.2}$$

where $\mathbf{p}_i'$ is the ith row of $\mathbf{P}$, and the variance of the series is $\mathbf{p}_i'\mathbf{\Gamma}_f(0)\mathbf{p}_i + var(e_{it})$. The average variance of the signal, or common part, is asymptotically

$$\overline{var(y_t)} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{p}_i'\mathbf{\Gamma}_f(0)\mathbf{p}_i = tr(\mathbf{\Gamma}_f(0)\frac{1}{N}\sum_{i=1}^{N} \mathbf{p}_i\mathbf{p}_i') \xrightarrow{p} \sum_{i=1}^{r} \gamma_i(0).$$

The approximate DFM (ADFM) allows the errors, $\mathbf{e}_t$, to be autocorrelated, heteroscedastic and with some weak cross-section correlation, but all this dynamics disappears asymptotically, see Bai and Ng (2002). Given these assumptions, the estimation of the factors by principal components provides consistent estimators of the common part, see Bai and Ng (2013). Other estimators can be obtained by the eigenvectors of the cumulative sum of lagged covariance matrices, see Lam et al. (2011) where they prove that $P(\widehat{r} \geq r) \to 1$, and also by maximun likelihood, see Bai and Li (2016).

The eigenstructure of the covariance matrices can be used to find the number of factors. Note that from (2.1)

$$\mathbf{\Gamma}_y(k) = \mathbf{P}\mathbf{\Gamma}_f(k)\mathbf{P}' + \mathbf{\Gamma}_e(k), \tag{2.3}$$

and the properties of $\mathbf{\Gamma}_y(k)$ depend on the hypothesis about the covariance matrices of the noise $\mathbf{\Gamma}_e(k)$. Under the EDFM with white noise $\mathbf{\Gamma}_e(k) = 0$ for $k \neq 0$, and if $\mathbf{\Gamma}_e(0) = \sigma^2\mathbf{I}$, the homoscedastic case, the matrix $\mathbf{\Gamma}_y(0)$ has r large eigenvalues corresponding to the variance of the factors, and $N-r$ small eigenvalues σ^2. The matrices $\mathbf{\Gamma}_y(k)$, for $k > 0$, will have rank at most r with eigenvalues equal to the covariance of lag k of the factors. If $\mathbf{\Gamma}_e(0) = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix, then the number of large eigenvalues in $\mathbf{\Gamma}_y(0)$ depends on the relative size of the minimum variance of the factors and the maximum variance of the noises. If there is autocorrelation and $\mathbf{\Gamma}_e(k) \neq 0$ for $k \neq 0$, this will not affect the eigenvalues of $\mathbf{\Gamma}_y(0)$ but will affect those of $\mathbf{\Gamma}_y(k)$. Finally, in the more general case, with heteroscedasticity and cross-sectional and auto correlation in the errors, the eigenvalues of all the covariance matrices depend on the assumed structure.

In this chapter we will concentrate on finding the number of factors in model (2.1) using eigenvalues ratio tests. These tests are discussed in the next section.

2.2 Testing the number of factors with eigenvalues

A test on the number of factors in a DFM based on the properties of the eigenvalues of some covariance matrices was proposed by Peña and Poncela (2006b), based on previous results by Tiao and Tsay (1989). Note that if in model (2.1) $\mathbf{e}_t$ is white noise for $k > 0$, $\mathbf{\Gamma}_y(k) = \mathbf{P}\mathbf{\Gamma}_f(k)\mathbf{P}'$ and there exists a $N \times (N-r)$ matrix $\mathbf{P}_\perp$, such that $\mathbf{\Gamma}_y(k)\mathbf{P}_\perp = \mathbf{0}$. Thus, the $N-r$ independent linear combinations of the observed series given by $\mathbf{P}'_\perp\mathbf{y}_t$ are cross-sectionally and serially uncorrelated for all lags, and also uncorrelated with $\mathbf{P}'_\perp\mathbf{y}_{t-k}$. The $N \times N$ canonical correlation matrix between $\mathbf{y}_t$ and $\mathbf{y}_{t-k}$ assuming an EDFM satisfying $\mathbf{\Gamma}_y(k) = \mathbf{\Gamma}_y(-k)$, is

$$\mathbf{C}(k) = \left[\mathbf{\Gamma}_y(0)^{-1}\mathbf{\Gamma}_y(k)\right]^2.$$

As for $k > 0$ rank($\mathbf{\Gamma}_y(k)$) = rank($\mathbf{\Gamma}_f(k)$) = r, then rank[$\mathbf{C}(k)$] = r, and the number of zero canonical correlations between $\mathbf{y}_{t-k}$ and $\mathbf{y}_t$ is given by the number of zero eigenvalues of the $\mathbf{C}(k)$ matrix, that is $N-r$. Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_N$ be the ordered eigenvalues of the estimated matrix $\widehat{\mathbf{C}}(k)$. If we have r factors, the eigenvalues $\widehat{\lambda}_{r+1}, ..., \widehat{\lambda}_N$ are estimates of squared correlations equal to zero and have asymptotic variance $1/(T-k)$. Therefore, for $j > r$ the statistics $-(T-k)\log(1-\widehat{\lambda}_j) \simeq (T-k)\widehat{\lambda}_j$ follow a Chi-square

distribution and the statistics

$$S_{N-r} = -(T-k)\sum_{j=r+1}^{N}\log(1-\widehat{\lambda}_j), \tag{2.4}$$

can be shown to be asymptotically a $\chi^2_{(N-r)^2}$, see Peña and Poncela (2006b). The test is applied sequentially, it starts with $r=0$ and if the hypothesis is rejected the value $r=1$ is tried and so on. The testing procedure stops when the hypothesis of $r+1$ eigenvalues equal to zero cannot be rejected.

This test only works for the EDFM and in the frequent case in which idiosyncratic terms have autocorrelation it cannot be applied. It has been generalized by Bolivar et al. (2020) by finding a consistent estimate of the matrices $\mathbf{\Gamma}_e(k)$ and applying the test to the corrected matrices $\widehat{\mathbf{\Gamma}}^*_y(k) = \widehat{\mathbf{\Gamma}}_y(k) - \widehat{\mathbf{\Gamma}}_e(k)$. However, although this test works well when T is much large than N, it deteriorates when N is large.

An useful alternative is, instead of checking the eigenvalues themselves, looking at the ratios of consecutive eigenvalues (in a proper order) of the covariance matrix, or a cumulative sum of lagged covariance matrices. Lam and Yao (2012) proposed to compute the ordered estimated eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_N$ of the pooled covariance matrix

$$\mathbf{M}_{1,k_0} = \sum_{k=1}^{k_0}\widehat{\mathbf{\Gamma}}_y(k)\widehat{\mathbf{\Gamma}}_y(k)', \tag{2.5}$$

where k_0 is a pre-specified positive integer, and select r as

$$\widehat{r} = \arg\max_{1\leq i\leq r^*}\frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}},$$

for some $r^* = \alpha N$, where N is the number of series and $0 < \alpha < 1$ such as $\alpha = 0.2$. Suppose that the first r eigenvalues are large and the remaining eigenvalues are small. Then, the ratios $\lambda_{i+1}/\lambda_i \leq 1$ would have a big decrease for $i = r$.

A similar test has been proposed by Ahn and Horenstein (2013) by using the ordered estimated eigenvalues $\widehat{\nu}_1 \geq \widehat{\nu}_2 \geq ... \geq \widehat{\nu}_N$ of the covariance matrix $\widehat{\mathbf{\Gamma}}_y(0)$. The criterion is

$$\widehat{r} = \arg\max_{1\leq i\leq r^*}\frac{\widehat{\nu}_i}{\widehat{\nu}_{i+1}}.$$

The LY criterion cannot be applied if some factors are white noise because then there is no information about those factors in the lagged covariance matrices. However, the AH test continues to work in these cases. Both criteria are consistent when both T and N go to infinity under appropriate hypothesis so that the long run effect of the idiosyncratic terms disappears. However, their properties in finite samples are very different

depending on the structure of the idiosyncratic components. The advantage of the LY criterion is to pool the information about the factors that is present in the lagged covariance matrices and with strong factors and white noise errors is expected to be more powerful than the AH test. However, as the lagged covariance matrices also include the information about the dynamics of the idiosyncratic components, and if this dynamic exits and is different among the components, it will increase the noise in the estimation of the factor effects and with weak factors using the lagged information may not be useful. On the other hand, the covariance matrix is not affected by these dynamics as it only contains contemporaneous information of the variances of the factors and the idiosyncratic components. The situation is the opposite with heteroscedasticity: the eigenvalues of the covariance matrix will be different and the AH test is expected to be affected whereas the LY criterion will not. Thus the relative advantage of each test depends on the dynamics of the idiosyncratic term.

An additional problem of these two tests is their lack of robustness to a few atypical series. Suppose that one of the series is affected by some large measurement errors, as often happen in time series automatically collected by sensor devices. Then, the variance of this series will be much larger than the others. Also, the outliers due to the measurement errors may destroy the cross-section correlation between this series and the others. In the limit, the largest eigenvalue of the covariance matrix will be equal to the variance of the atypical series and the corresponding eigenvector will have close to the zero components in the rest of the uncontaminated series and a value close to one in the outlying series. This problem will not appear if we work with autocorrelation matrices. Of course, these matrices are not robust to other outliers and a more powerful procedure with many contaminated series will be to compute robust correlation matrices and this alternative will be explored in future works.

2.3 An eigenvalues test on the Pooled Correlation Matrices

We propose an eigenvalue test based on the weighted combination of the correlation matrices of the observed data. We define the *combined correlation matrix* as

$$\mathbf{R}_{k_0} = \sum_{k=0}^{k_0} w_k \mathbf{R}_y(k) \mathbf{R}_y(k)', \tag{2.6}$$

where k_0 is a pre-specified positive integer, the coefficients $w_k > 0$ are weights which verify $\sum_{k=0}^{k_0} w_k = 1$, and $\mathbf{R}_y(k)$ is the lag k correlation matrix of the series. Different weights can be considered but a simple solution is to use the asymptotic variance of the autocorrelation and cross correlation coefficients, r_{ij} for $i, j = 1, ..., N$, for white noise

stationary process, $var(r_{ij}(k)) \approx (T-k)^{-1}$. Then, as in the Box-Ljung pormanteau test of goodness of fit, we can standardize the squared correlations by their variance and define the weights as $(T-k)/c$, where c is chosen so that the weights add up to one by

$$c = \sum_{k=0}^{k_0} (T-k),$$

which implies $c = (k_0+1)(T-k_0/2)$ and $w_k = (T-k)/\left((k_0+1)(T-k_0/2)\right)$. Let $\widehat{\alpha}_1 \geq \widehat{\alpha}_2 \geq ... \geq \widehat{\alpha}_N$ be the ordered estimated eigenvalues of the matrix $\mathbf{R}_{k_0}$. The test selects the number of factors as

$$\widehat{r} = \arg \max_{1 \leq i \leq r^*} \frac{\widehat{\alpha}_i}{\widehat{\alpha}_{i+1}}.$$

To compare this test with the two previous ones also based on ratios of eigenvalues let $\mathbf{S} = diag(\mathbf{\Gamma}_y(0))$ be the diagonal matrix of the variances of the series, $\mathbf{M}_0 = \mathbf{\Gamma}_y(0)\mathbf{\Gamma}_y(0)'$ and $\mathbf{M}_{1,k_0}$ as given in (2.5). Then

$$\mathbf{R}_{k_0} = w_0 \mathbf{S}^{-1/2}\mathbf{M}_0\mathbf{S}^{-1/2} + (1-w_0)\mathbf{S}^{-1/2}\mathbf{M}_{1,k_0}\mathbf{S}^{-1/2}.$$

This equation shows that the pooled correlation matrix defined in (2.6) is a weighted combination of the matrices used in the AH and LY criteria. The advantage of using the correlation matrices instead of covariance matrices to check the ratio of the eigenvalues can be important when the series are heteroscedastic, as shown in the following theorem.

Theorem 1. Let $\mathbf{C} = \{c_{ij}\}$ be a $N \times N$ symmetric and positive definite random matrix with eigenvalues $\lambda_{C1} \geq \lambda_{C2} \geq ... \geq \lambda_{CN} \geq 0$, that are assumed to follow a non-negative distribution with finite moments. The corresponding eigenvectors, $u_{Cj} = (u_{Cj1}, ..., u_{CjN})/q_{Cj}$, where the u_{Cji}, are random variables with $E(u_{Cji}) = 0$ and $E(u_{Cji}^2) = 1/N$ and $q_{Cj} = \sqrt{\sum_{i=1}^N u_{Cji}^2}$. Let $\mathbf{V} = \mathbf{C} + \mathbf{S}$ where $\mathbf{S} = \{s_{ij}\}$ is, given $\mathbf{C}$, a non random matrix with $s_{11} = \sigma^2 - 1, s_{1j} = s_{j1} = (\sigma - 1)c_{1j}$ and $s_{ij} = 0$ otherwise, where $\sigma > 1$ and $m = \sum_{j=1}^N c_{1j} > 0$. The eigenvalues of the matrix $\mathbf{V}$, $\lambda_{V1} \geq \lambda_{V2} \geq ... \geq \lambda_{VN} \geq 0$, are related to those of matrix $\mathbf{C}$, up to a first order approximation, by

$$E(\lambda_{Vj}) \simeq E(\lambda_{Cj}) + (\sigma - 1)((\sigma + E(\lambda_{Cj}))/N + m).$$

Proof of **Theorem 1**

The relationship between the eigenvalues of the matrices C and V with eigenvalues λ_{Cj} and λ_{Vj} and eigenvectors $\mathbf{u}_{Cj}$ and $\mathbf{u}_{Vj}$ can be approximated by Stewart and Sun (1990)

$$\lambda_{Vj} = \lambda_{Cj} + \mathbf{u}'_{Cj}\mathbf{S}\mathbf{u}_{Cj} + o(\|\mathbf{S}\|^2),$$

where $\|\mathbf{S}\|$ is a norm of the matrix $\mathbf{S}$. As, calling $\mathbf{u}'_{Cj} = (u_{j1}, ..., u_{jN})$,

$$\begin{bmatrix} \sigma^2 - 1 & c_{12}(\sigma-1) & ... & c_{1N}(\sigma-1) \\ c_{12}(\sigma-1) & 0 & ... & 0 \\ ... & ... & ... & ... \\ c_{1N}(\sigma-1) & 0 & ... & 0 \end{bmatrix} \mathbf{u}_{Cj} = \begin{bmatrix} (\sigma^2-1)u_{Cj1} + (\sigma-1)\sum_{h=2}^{N} c_{1h}u_{Cjh} \\ (\sigma-1)c_{12}u_{Cj2} \\ ... \\ (\sigma-1)c_{1N}u_{CjN} \end{bmatrix},$$

and as $\sum_{h=1}^{N} c_{1j}u_{Cjh} = \lambda_{Cj}u_{Cj1}$, we have

$$\mathbf{u}'_{Cj}\mathbf{S}\mathbf{u}_{Cj} = \mathbf{u}'_{Cj}(\sigma-1)\begin{bmatrix} (\sigma+\lambda_{Cj})u_{Cj1} \\ c_{12}u_{Cj2} \\ ... \\ c_{1N}u_{CjN} \end{bmatrix} = (\sigma-1)[u^2_{Cj1}(\sigma+\lambda_{Cj}) + \sum_{h=1}^{N} c_{1j}u^2_{Cjh}],$$

and assuming independence between the random variables u_{Cj1} and λ_{Cj} and taking expected values with $E(u^2_{Cj1}) = 1/N$

$$E(\mathbf{u}'_{Cj}\mathbf{S}\mathbf{u}_{Cj}) = (\sigma-1)((\sigma + E(\lambda_{Cj}))/N + m),$$

where $m = E(\sum_{h=1}^{N} c_{1j}u^2_{Cjh}) = \sum_{h=1}^{N} c_{1j}/N > 0$, as the term c_{1j} are constant numbers. Thus, with a the first order approximation

$$E(\lambda_{Vj}) \simeq E(\lambda_{Cj}) + (\sigma-1)((\sigma + E(\lambda_{Cj}))/N + m).$$

Suppose that we have a set of N variables. The variance of the first one is $\sigma^2 > 1$ and all the others have variance equal to one. Calling $\mathbf{V}$ to the covariance matrix of these variables and $\mathbf{C}$ to their correlation matrix, they are related by $\mathbf{V} = \mathbf{C} + \mathbf{S}$, where $s_{11} = \sigma^2 - 1$, $s_{1j} = s_{j1} = (\sigma-1)c_{1j}$ and $s_{ij} = 0$ otherwise. Then, according to Theorem 1, the ratio between the expected values of the eigenvalues of these matrices is

$$r_V(j) = \frac{E(\lambda_{Vj})}{E(\lambda_{Vj+1})} \simeq \frac{E(\lambda_{Cj}) + (\sigma-1)((\sigma + E(\lambda_{Cj}))/N + m)}{E(\lambda_{Cj+1}) + (\sigma-1)((\sigma + E(\lambda_{Cj+1}))/N + m)} = \frac{E(\lambda_{Cj})(1+a) + b}{E(\lambda_{Cj+1})(1+a) + b},$$

where $a = (\sigma - 1)/N$ and $b = (\sigma - 1)(\sigma/N + m)$. Thus

$$r_V(j) \simeq \frac{E(\lambda_{Cj})(1+a)+b}{E(\lambda_{Cj+1})(1+a)+b} = \frac{E(\lambda_{Cj})(1+b/(1+a)E(\lambda_{Cj}))}{E(\lambda_{Cj+1})(1+b/(1+a)E(\lambda_{Cj+1}))} < \frac{E(\lambda_{Cj})}{E(\lambda_{Cj+1})} = r_C(j).$$

Therefore, when one of the series have a variance larger than the others the ratio of eigenvalues in the correlation matrix is expected to be larger than in the covariance matrix, and the difference between these ratios will increase with the heteroscedasticity (the value of σ) and in the larger ratios in the covariance matrix. Thus, standardizing the variables will increase the expected ratio in the correlation matrices when this ratio is already large in the covariance matrices. On the other hand, when these ratios are small in the covariance matrices the expected change will be small in the correlation matrices. This result implies that the standardization of the variables when the series are heteroscedastic is expected to increase the ratio of eigenvalues at the exact number of factors in the correlation matrices with respect to the covariance matrices, increasing, therefore, the power of the ratio of eigenvalues test. This result will be confirmed in the Monte Carlo experiment presented in the next section.

2.4 Monte Carlo experiment of tests performance

We run a simulation exercise to compare three criteria for the number of factors: AH, LY, and the one proposed in this chapter, CP. For this comparison we use different: (1) Data generating processes (DGP) for the idiosyncratic component in model (1); (2) Numbers of latent factors: two and three; (3) Signal to noise ratios, strong, medium and weak. The data is generated by equation (2.2) with factor loading coefficients $\mathbf{p}_i$ generated from the U(-0.5, 0.5) distribution. Each common factor, f_t, is generated as an AR(1) process, by the following equation:

$$f_t = \phi f_{t-1} + \eta_t, \tag{2.7}$$

We consider six DGP for the idiosyncratic component. The first three have serially uncorrelated noises with different idiosyncratic covariance matrices, $\Gamma_e(0)$, and the next three add serially correlated noises to the basic covariance structure. Thus, DGP_1 has homoscedastic errors and $\Gamma_e(0) = \sigma^2\mathbf{I}$, with $\sigma^2_{e_i} = 1$ for $i = 1, ..., N$, DGP_2 has heteroscedastic uncorrelated noises with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and DGP_3 has heteroscedastic and cross-section correlated errors, with $\Gamma_e(0)$ having diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. The next three DGP add to the three previous scenarios serially correlated errors $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, $u_{it} \sim N(0, 1)$ white

noise, and uncorrelated noises with $\sigma^2_{e_i} = 1$ for $i = N/2 + 1, ..., N$. Thus, this modification of the first DGP, that will be called $DGPC_1$, has $\Gamma_e(0)$ with variances around $1/(1-.5^2) = 1.33$ in half of the matrix and variances equal to one in the other half. Also $\Gamma_e(k)$ for $k > 0$ is a diagonal matrix different from zero. $DGPC_2$ has heteroscedastic variances around $1/(1-.5^2) = 1.33$ in half of $\Gamma_e(0)$ and variances equal to two in the other half, and non null diagonal lag covariance matrices. Finally $DGPC_3$ has full rank $\Gamma_e(0)$ with different variances in half of the matrix and equal to two in the other half, non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$, and non null $\Gamma_e(k)$ lagged covariance matrices.

Three signal to noise ratios (SN) are considered: Strong, with autoregressive coefficients $\phi_{f_1} = 0.9$, $\phi_{f_2} = 0.8$ and η_t white noise $N(0,1)$; Medium with $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ and, as before, η_t independent $N(0,1)$; Weak with $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ but errors η_t independent $N(0, 0.5)$ random variables. The errors η_t are independent of the idiosyncratic errors e_t, such that $E(\eta_t e_t') = 0$, for all data generating processes.
Finally, scenarios DGP_4, DGP_5, and DGP_6 and $DGPC_4$, $DGPC_5$, and $DGPC_6$ follow the same idiosyncratic covariance structures as DGP_1, DGP_2, and DGP_3, respectively, but with three common factors ($r = 3$). Each common factor follows an AR(1) process with $\phi_{f_1} = 0.9$, $\phi_{f_2} = 0.8$ and $\phi_{f_3} = 0.7$ in scenarios with strong SN ratio, and $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ and $\phi_{f_3} = 0.4$ in scenarios with medium and weak SN ratio. For the first two scenarios errors η_t are generated as independent $N(0,1)$ random variables and for the last scenario they are generated as independent $N(0, 0.5)$ random variables.
The number of cross-section variables considered are $N = 10, 50, 100, 200$, and the number of time observations are $T = 125, 250, 500, 1250$. For each one of the different (N, T) combinations we run 200 iterations and calculate the relative frequency estimates of the true number of common factors, $\widehat{r}$. The number of lags considered in the cumulative sum of lagged covariances matrices is $k_0 = 2$. We just consider this value given that the criterion is not sensitive to the choice of k_0, as stated in Lam et al. (2011).

Tables 2.1 - 2.7 report the relative frequency estimates of the true number of common factors $\widehat{r}$. We can see in all tables that the estimation of r demonstrates "the blessing of dimensionality", for fixed sample size T, the relative frequency estimates for $\widehat{r} = r$ increase with N, and the differences depend on the number of series N, and the signal to noise ratio. The last three columns in each table report the estimations when the idiosyncratic terms are autocorrelated. Under this scenario, all of the estimators have a decrease of power to detect the true number of factors. In broad terms autocorrelated errors have a similar effect to decrease the SN ratio.

Tables 2.1 - 2.4 show the results for two factors and Tables 2.5 - 2.7 for three factors. We have selected these 7 tables out of the 18 that have all the tested scenarios because the message obtained is quite consistent and these seven table summarize it well, the rest

of the tables are available in A.2 Tables section in the Appendix to Chapter 2. When the idiosyncratic terms have approximately the same variance, as in Table 2.1, the three tests are similar, but AH and CP are slightly better than LY. Only for small number of series $N = 10$ and very large $T = 1250$ LY has the best performance. Table 2.2 presents results for the two factor model with heteroscedastic errors and strong factors. In this more realistic scenario the CP test clearly outperforms the other two tests both when the errors are white noise and when they are correlated. For example, in Table 2.2 when $N = 100$ the advantage of CP test over LY and AH is approximately of 30%. This difference could be even larger than 100% when the SN ratio is weak, as shown in Table 2.3 for $N = 100$. For small number of series, AH has less power to detect the factors, even with strong SN ratio, see Table 2.4. The performance of the CP test is less sensitive than AH and LY to autocorrelated errors, see last three columns. Finally, if in addition to heteroscedasticity we consider lag zero cross-section correlation, as in Table 2.4, with medium SN ratio, LY and CP always provide better estimations than AH. For example, for $N = 200$ and $T = 125$ the performance of the CP test without autocorrelated noise is 60% more powerful than LY and larger than 100% than AH, and with autocorrelated error the differences in power are even greater: for $N = 200$ and $T = 500$ CP gives the right number of factors 78% more often than AH and 273% than LY. As expected, the performance of the AH test is much more sensitive to cross-section correlation than CP and LY criteria.

Tables 2.5 - 2.7 give some results for DFM with three factors. Table 2.5 is similar to Table 2.1 but now with three factors and weak signal to noise ratio. The best performance correspond to AH test follows closely by CP and both are more powerful than LY. Table 2.6 presents the results for heteroscedastic time series and the advantages of the CP test with respect to the other two are clear. Finally, Table 2.7 gives the performance with heteroscedasticiy and cross-section correlation. When the noises are not autocorrelated, see columns $1-3$, the best performance correspond to LY criterion, but when the noises are autocorrelated, see columns $4-6$, CP is more powerful than the others. In summary, we conclude that the CP test provides overall the more powerful performance for the number of common factors when errors present heteroscedasticity and autocorrelation. When the noises are homoscedastic white noises it has a performance close to the best test in this case.

Table 2.1: Relative frequency estimates of the true number of common factors $r = 2$. Homoscedastic errors and medium signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.42	0.22	0.36	0.33	0.22	0.28	0.31
T=250	0.55	0.34	0.47	0.37	0.16	0.33	0.37
T=500	0.52	0.44	0.48	0.43	0.1	0.38	0.39
T=1250	0.6	0.62	0.51	0.52	0.02	0.5	0.46
N=50							
T=125	0.96	0.56	0.8	0.86	0.4	0.74	0.72
T=250	0.99	0.84	0.98	0.96	0.64	0.93	0.89
T=500	1	0.97	1	1	0.75	0.99	0.95
T=1250	1	1	1	1	0.96	1	0.99
N=100							
T=125	1	0.68	0.97	0.96	0.62	0.92	0.86
T=250	1	0.9	1	1	0.84	1	0.96
T=500	1	1	1	1	0.94	1	0.99
T=1250	1	1	1	1	1	1	1
N=200							
T=125	1	0.82	0.98	1	0.76	0.96	0.92
T=250	1	0.94	1	1	0.98	1	0.99
T=500	1	1	1	1	1	1	1
T=1250	1	1	1	1	1	1	1
Mean	**0.88**	0.77	0.85	**0.84**	0.65	0.81	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$ and $\phi_{f_2} = 0.5$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix $\Gamma_e = \sigma_e \mathbf{I}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table 2.2: Relative frequency estimates of the true number of common factors $r = 2$. Heteroscedastic errors and strong signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.07	0.16	0.24	0.09	0.3	0.23	0.18
T=250	0.06	0.24	0.26	0.08	0.22	0.23	0.18
T=500	0.08	0.43	0.38	0.06	0.08	0.28	0.22
T=1250	0.06	0.63	0.32	0.08	0.08	0.27	0.24
N=50							
T=125	0.43	0.46	0.66	0.32	0.3	0.55	0.45
T=250	0.59	0.7	0.86	0.43	0.4	0.76	0.63
T=500	0.74	0.9	0.96	0.57	0.55	0.92	0.77
T=1250	0.86	0.99	0.99	0.7	0.72	0.98	0.87
N=100							
T=125	0.58	0.57	0.75	0.54	0.44	0.72	0.6
T=250	0.9	0.9	0.96	0.78	0.68	0.92	0.86
T=500	0.98	1	1	0.9	0.82	1	0.95
T=1250	1	1	1	0.98	0.96	1	0.99
N=200							
T=125	0.86	0.76	0.92	0.68	0.5	0.83	0.76
T=250	0.96	0.92	1	0.92	0.76	0.96	0.92
T=500	1	1	1	0.98	0.94	1	0.99
T=1250	1	1	1	1	1	1	1
Mean	0.64	0.73	**0.77**	0.57	0.55	**0.73**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$ and $\phi_{f_2} = 0.8$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table 2.3: Relative frequency estimates of the true number of common factors $r = 2$. Heteroscedastic errors and weak signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0	0.17	0.18	0.01	0.51	0.26	0.19
T=250	0	0.18	0.21	0	0.8	0.2	0.23
T=500	0	0.2	0.24	0	0.95	0.26	0.27
T=1250	0	0.16	0.22	0	0.99	0.24	0.27
N=50							
T=125	0.08	0.12	0.2	0.14	0.22	0.18	0.16
T=250	0.01	0.14	0.23	0.05	0.19	0.22	0.14
T=500	0	0.08	0.4	0	0.1	0.24	0.14
T=1250	0	0.14	0.7	0	0	0.44	0.21
N=100							
T=125	0.12	0.13	0.22	0.2	0.25	0.2	0.19
T=250	0.1	0.12	0.46	0.16	0.22	0.24	0.22
T=500	0.04	0.17	0.8	0.1	0.09	0.46	0.28
T=1250	0	0.26	0.98	0	0.06	0.78	0.35
N=200							
T=125	0.15	0.14	0.35	0.16	0.18	0.21	0.2
T=250	0.26	0.16	0.76	0.24	0.2	0.46	0.35
T=500	0.54	0.36	0.96	0.25	0.21	0.74	0.51
T=1250	0.88	0.65	1	0.47	0.24	0.99	0.7
Mean	0.14	0.2	**0.49**	0.11	0.33	**0.38**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$ and $\phi_{f_2} = 0.5$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1 - 3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4 - 6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table 2.4: Relative frequency estimates of the true number of common factors $r = 2$. Heteroscedastic and cross correlated errors, and medium signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.2	0.16	0.28	0.34	0.14	0.51	0.27
T=250	0.24	0.18	0.28	0.3	0.1	0.42	0.25
T=500	0.22	0.22	0.34	0.4	0.17	0.6	0.32
T=1250	0.26	0.16	0.25	0.45	0.16	0.63	0.32
N=50							
T=125	0.08	0.14	0.13	0.08	0.31	0.1	0.14
T=250	0.05	0.18	0.02	0.08	0.27	0.03	0.11
T=500	0.03	0.26	0.01	0.1	0.29	0.04	0.12
T=1250	0	0.53	0.02	0.02	0.23	0	0.14
N=100							
T=125	0.12	0.2	0.25	0.14	0.24	0.24	0.2
T=250	0.08	0.3	0.32	0.15	0.23	0.26	0.22
T=500	0.08	0.61	0.46	0.14	0.17	0.23	0.28
T=1250	0.14	0.86	0.53	0.1	0.07	0.39	0.35
N=200							
T=125	0.32	0.24	0.51	0.23	0.22	0.44	0.33
T=250	0.56	0.52	0.87	0.3	0.2	0.66	0.52
T=500	0.76	0.83	0.94	0.46	0.22	0.82	0.68
T=1250	0.94	1	1	0.56	0.18	0.98	0.78
Mean	0.26	**0.4**	0.39	0.24	0.2	**0.4**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$ and $\phi_{f_2} = 0.5$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Table 2.5: Relative frequency estimates of the true number of common factors $r = 3$. Homoscedastic errors and weak signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.12	0.18	0.11	0.12	0.18	0.12	0.14
T=250	0.13	0.14	0.1	0.16	0.11	0.16	0.13
T=500	0.08	0.1	0.08	0.14	0	0.13	0.09
T=1250	0.12	0.04	0.1	0.14	0	0.14	0.09
N=50							
T=125	0.12	0.04	0.12	0.14	0.14	0.12	0.12
T=250	0.23	0.04	0.17	0.16	0.1	0.09	0.13
T=500	0.57	0.07	0.5	0.36	0.06	0.3	0.31
T=1250	0.84	0.32	0.8	0.64	0.04	0.63	0.55
N=100							
T=125	0.31	0.06	0.16	0.13	0.08	0.09	0.14
T=250	0.82	0.22	0.68	0.32	0.08	0.21	0.39
T=500	0.98	0.38	0.96	0.74	0.1	0.59	0.63
T=1250	1	0.78	1	0.96	0.14	0.94	0.8
N=200							
T=125	0.7	0.18	0.46	0.22	0.11	0.14	0.3
T=250	0.98	0.36	0.89	0.76	0.1	0.48	0.6
T=500	1	0.78	1	0.99	0.2	0.96	0.82
T=1250	1	0.98	1	1	0.57	1	0.93
Mean	**0.56**	0.29	0.51	**0.43**	0.13	0.38	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ and $\phi_{f_3} = 0.4$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1 - 3$ idiosyncratic covariance matrix $\Gamma_e = \sigma_e \mathbf{I}$. Columns $4 - 6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table 2.6: Relative frequency estimates of the true number of common factors $r = 3$. Heteroscedastic errors and medium signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.03	0.16	0.08	0.04	0.08	0.09	0.08
T=250	0	0.12	0.1	0.02	0.12	0.1	0.08
T=500	0	0.1	0.05	0.02	0.14	0.08	0.06
T=1250	0	0.06	0.08	0	0.07	0.09	0.05
N=50							
T=125	0.1	0.08	0.43	0.16	0.14	0.21	0.19
T=250	0.23	0.06	0.72	0.13	0.1	0.4	0.27
T=500	0.36	0.16	0.9	0.18	0.06	0.73	0.4
T=1250	0.66	0.52	0.98	0.22	0.02	0.97	0.56
N=100							
T=125	0.4	0.08	0.68	0.2	0.13	0.36	0.31
T=250	0.74	0.2	0.96	0.35	0.06	0.82	0.52
T=500	0.95	0.55	1	0.72	0.1	0.99	0.72
T=1250	0.99	0.91	1	0.95	0.16	1	0.83
N=200							
T=125	0.8	0.17	0.85	0.37	0.1	0.64	0.49
T=250	1	0.52	1	0.84	0.16	0.96	0.75
T=500	1	0.86	1	1	0.28	1	0.86
T=1250	1	0.99	1	1	0.64	1	0.94
Mean	0.52	0.35	**0.68**	0.39	0.15	**0.59**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ and $\phi_{f_3} = 0.4$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table 2.7: Relative frequency estimates of the true number of common factors $r = 3$. Heteroscedastic and cross correlated errors, and strong signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.18	0.1	0.22	0.16	0.1	0.28	0.17
T=250	0.16	0.06	0.2	0.14	0.16	0.24	0.16
T=500	0.22	0.1	0.28	0.19	0.2	0.3	0.22
T=1250	0.14	0.16	0.21	0.14	0.12	0.26	0.17
N=50							
T=125	0.05	0.04	0.03	0.06	0.12	0.05	0.06
T=250	0.02	0.07	0.02	0.06	0.06	0.01	0.04
T=500	0.01	0.28	0.02	0.06	0.04	0	0.07
T=1250	0	0.77	0.02	0.02	0.01	0	0.14
N=100							
T=125	0.02	0.07	0.08	0.04	0.04	0.07	0.05
T=250	0.03	0.22	0.15	0.02	0.03	0.09	0.09
T=500	0.02	0.6	0.22	0.01	0	0.1	0.16
T=1250	0.02	1	0.28	0.02	0.02	0.09	0.24
N=200							
T=125	0.15	0.22	0.3	0.1	0.06	0.2	0.18
T=250	0.29	0.48	0.6	0.1	0.07	0.41	0.32
T=500	0.54	0.92	0.87	0.24	0.2	0.63	0.56
T=1250	0.76	1	0.98	0.32	0.33	0.92	0.72
Mean	0.16	**0.38**	0.28	0.11	0.1	**0.23**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$, $\phi_{f_2} = 0.8$ and $\phi_{f_3} = 0.7$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

2.5 An application to real data: Business Cycles

Our interest is to provide an estimate of the global business cycle from 1998 to 2019 using a DFM. Following literature in business cycles, see Kose et al. (2003) and Crucini et al. (2011), we analyze the total GDP, the private consumption expenditure (CON) and the gross fixed capital formation (INV) of 35 OECD countries, see Table 2.8. The data set is available at OECD Statistics (https://stats.oecd.org). The complete data sample has N=105 quarterly series (three per country) and T=88 time observations, from 1998:Q1 to 2019:Q4, and previous to the analysis we first difference the data that was already in logs. We implement the three criteria considered in previous sections, AH, LY and CP, in order to estimate the number of common factors $\widehat{r}$. Then, the factors are estimated using the $\widehat{r}$ eigenvectors linked to the $\widehat{r}$ largest eigenvalues, in descending order, of the matrices $\widehat{\mathbf{\Gamma}}_y(0)$, $\mathbf{M}_{1,k_0}$ and $\mathbf{R}_{k_0}$. Chapter 3 presents a comparison of the estimated eigenvectors of these three matrices. When using LY and CP criteria, the maximum number of lags considered in the cumulative sum of covariance and correlation matrices is $k_0 = 2$, see Lam et al. (2011) where it is shown that the estimated factor model is not sensitive to the choice of k_0. The AH and LY criteria provide similar results and we just show here those of the first of these tests. The AH ratio $\widehat{\nu}_i/\widehat{\nu}_{i+1}$ is plotted against i in the left panel of Figure 2.1. It can be seen that the estimator leads to $\widehat{r} = 1$.

Table 2.8: OECD countries included in the real data sample.

Australia	Austria	Belgium	Canada	Chile
Czech Republic	Denmark	Estonia	Finland	France
Germany	Hungary	Iceland	Ireland	Israel
Italy	Japan	Korea	Latvia	Lithuania
Luxembourg	Mexico	The Netherlands	New Zealand	Norway
Poland	Portugal	Slovak Republic	Slovenia	Spain
Sweden	Switzerland	Turkey	United Kingdom	United States

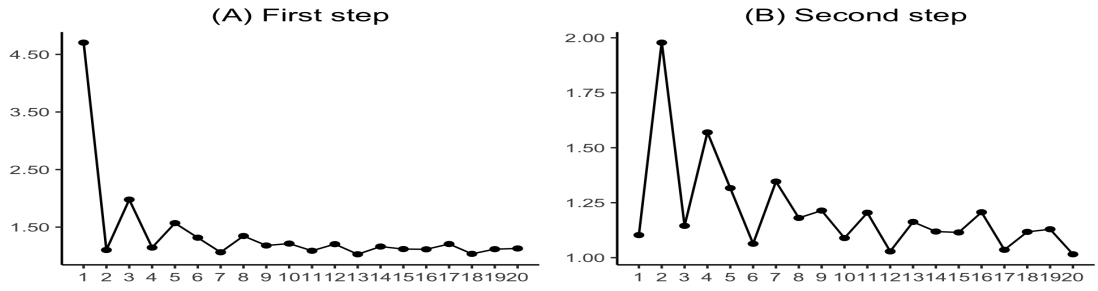

Figure 2.1: AH ratio of eigenvalues of $\widehat{\mathbf{\Gamma}}_y(0)$ for the first and the second steps.

The time series plot of AH estimated factor $\widehat{f}_t = y_t \widehat{P}'$ and its corresponding loading vector $\widehat{P}$ are plotted in Figures 2.2 and 2.3, respectively. The hidden x-axis labels after GDP series correspond with consumption and investment series for each country. The column vector of loadings plotted in Figure 2.3 is the normalized eigenvector of the covariance matrix $\widehat{\mathbf{\Gamma}}_y(0)$ corresponding to its largest eigenvalue. Both the AH and LY criteria estimate one common factor which is mainly the series of Ireland INV. In fact, the correlation between the factor and this series is 0.99. This situation appears because the large variance of the Ireland INV series makes that the first principal component is mostly generated by this series. Figure 2.4 shows a barplot of the variances of the series and it can be seen that Ireland INV is outlying in terms of variance followed by The Netherlands and Iceland investment series.

Figure 2.2: AH first estimated factor.

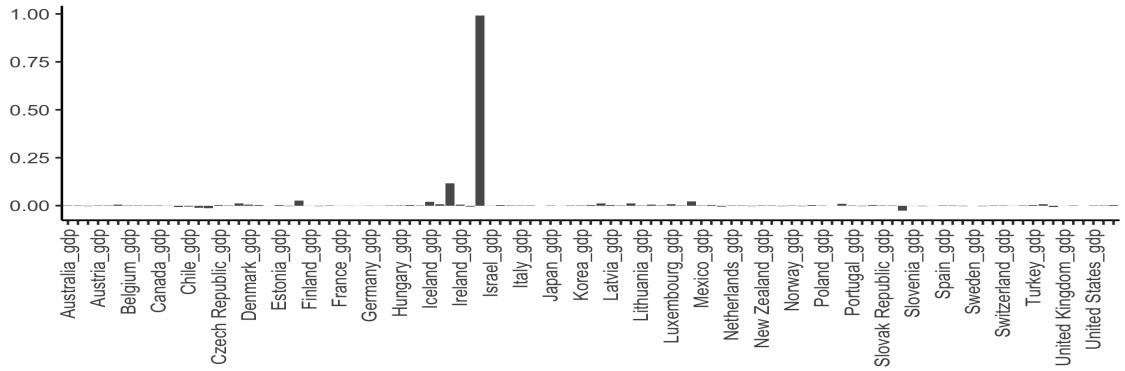

Figure 2.3: Estimated loadings corresponding to AH first common factor. In the x-axis only the labels for gdp series are shown. The hidden labels after country_gdp are country_con and country_inv.

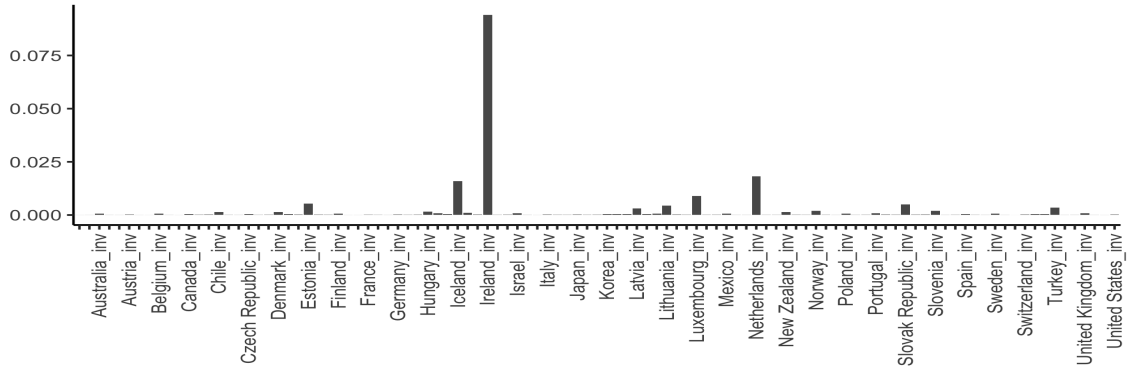

Figure 2.4: Variances of each time series in the sample. In the x-axis only the labels for investment series are shown. The hidden labels before country_inv are country_gdp and country_con.

The left panel of Figure 2.1 suggests that, in addition to the first strong factor, two additional weaker factors may exist, see the second largest ratio $\widehat{\nu}_3/\widehat{\nu}_4$. In a second step we remove the first estimated factor from the real data y_t and apply the AH ratio test to the resulting residuals. AH test estimates two additional factors, $\widehat{r} = 2$, see the right panel in Figure 2.1. We present the time series plot and the barplots of loadings in Figures 2.5, 2.6 and 2.7, respectively. Both factors are basically affected by the dynamics of The Netherlands INV and Iceland INV, which are the second and third series with the largest variances in the sample, see the barplot in Figure 2.4. The LY criterion estimates the same number of factors in the second step, $\widehat{r} = 2$: the second factor affects mainly INV series of The Netherlands, Latvia, Estonia and Iceland, and the third factor influences the dynamics in The Netherlands INV, followed by Luxembourg INV, Lithuania INV and Estonia INV to a lesser extent.

Figure 2.5: AH second (solid) and third (dashed) estimated factors.

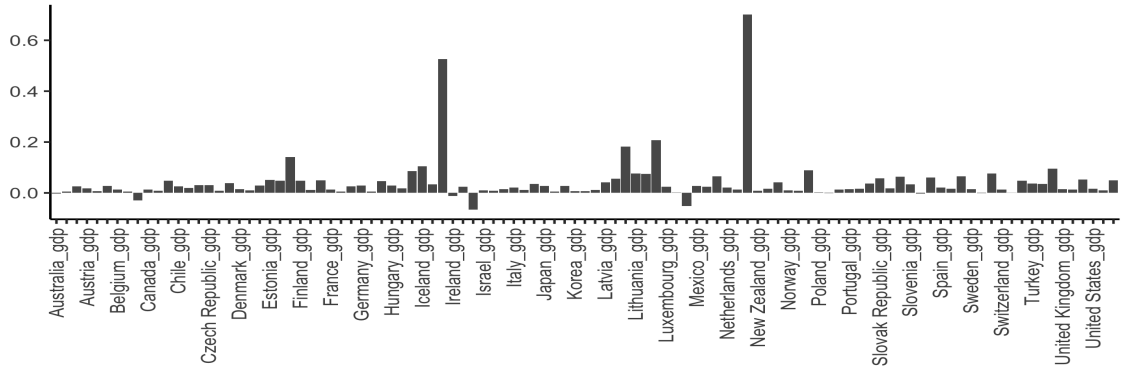

Figure 2.6: Estimated loadings corresponding to AH second common factor. In the x-axis only the labels for gdp series are shown. The hidden labels after country_gdp are country_con and country_inv.

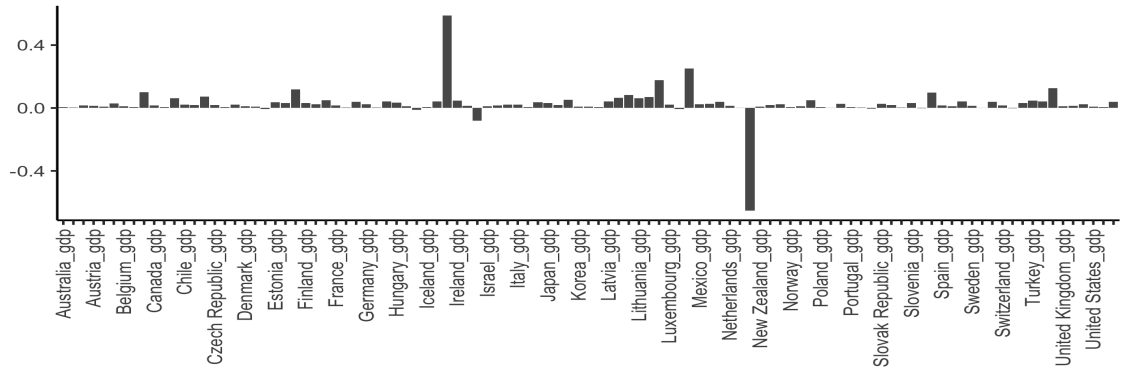

Figure 2.7: Estimated loadings corresponding to AH third common factor. In the x-axis only the labels for gdp series are shown. The hidden labels after country_gdp are country_con and country_inv.

The CP ratio test estimates one common factor $\widehat{r} = 1$ in the first step, see the left panel in Figure 2.8. The column vector of loadings $\widehat{P}$ is the normalized eigenvector of the matrix $\widehat{\mathbf{R}}_{k_0}$ corresponding to its largest eigenvalue. Contrary to AH and LY criteria, we can see in the barplot of Figure 2.10 how CP is able to capture the overall dynamics of all the countries in the sample. The first latent factor plotted in Figure 2.9 represents a global business cycle taking negative values during the worldwide financial crisis of 2008 and whose recovery did not reach pre-crisis levels. The largest factor loadings correspond to France GDP and INV, Spain GDP and Czech Republic GDP.

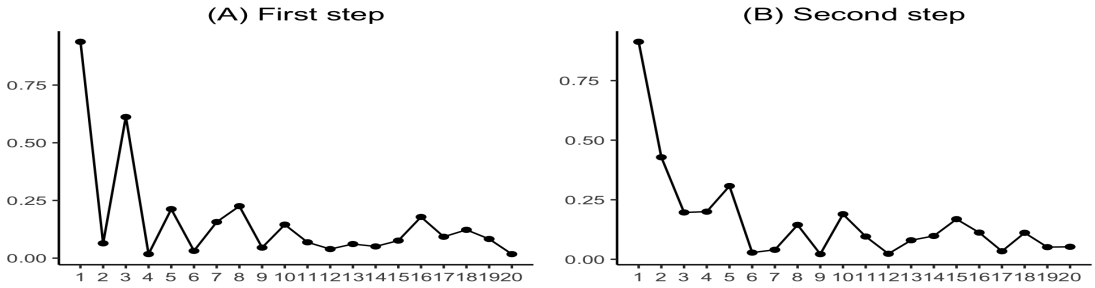

Figure 2.8: CP ratios of eigenvalues of $\mathbf{R}_{k_0}$ for the first and the second steps.

Figure 2.9: CP first estimated factor.

The ratio of eigenvalues plot suggests the existence of weaker factors, see the second largest ratio $\widehat{\alpha}_3/\widehat{\alpha}_4$, in Figure 2.8 left panel. In a second step, we subtract the first estimated component from the real data y_t and apply the CP ratio test to the resulting residuals. The estimate is one additional common factor $\widehat{r} = 1$. This second factor plotted in Figure 2.11 presents two drops: taking negative values at the beginning of the sample period and during the financial crisis of 2008. Differently from the estimated factor in the first step, this factor impacts mainly the dynamics of countries whose recovery reached pre-crisis levels or even superior. In terms of the percentage of the total variance of y_t explained by the estimated common factors, the first factor computed from the covariance matrix, called to simplify AH factor, accounts for the 50% of the total variability. The factor computed from the LY matrix accounts for 90% of the variability of the squared matrices and the factor computed from the correlation matrices accounts for 74% of the variability of the standardized data.

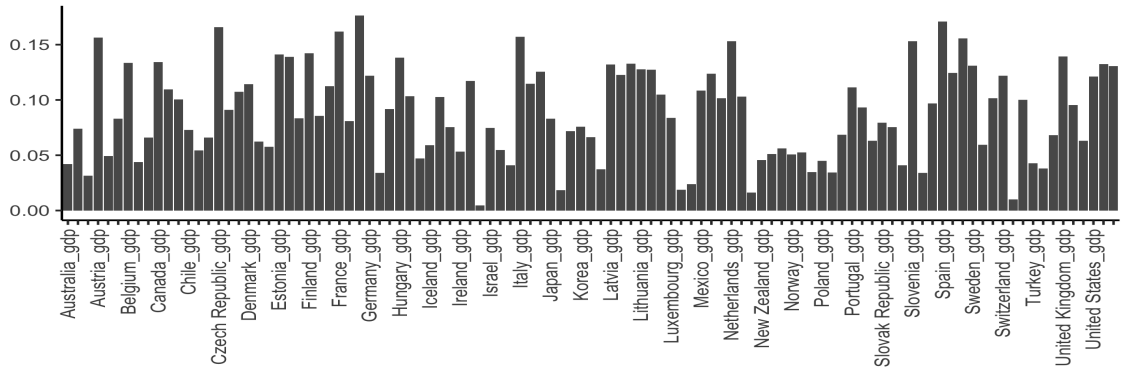

Figure 2.10: Estimated loadings corresponding to CP first common factor. In the x-axis only the labels for gdp series are shown. The hidden labels after country_gdp are country_con and country_inv.

Figure 2.11: CP second estimated factor.

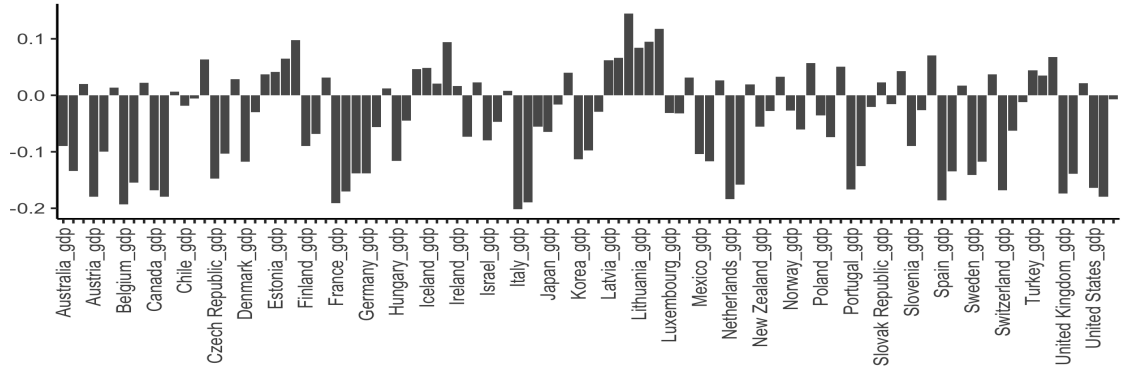

Figure 2.12: Estimated loadings corresponding to CP second common factor. In the x-axis only the labels for gdp series are shown. The hidden labels after country_gdp are country_con and country_inv.

2.6 Concluding remarks

The test proposed in this chapter has been evaluated in different scenarios depending on the idiosyncratic error structure. It has shown a better overall performance than the ones proposed by Ahn and Horenstein (2013) and Lam et al. (2011). The advantages of the test appear mostly under a realistic error structure that includes heteroscedasticity in the series and allows the errors to present cross-sectional and serial correlations. Also, it has been illustrated in a real example that this test is less affected by atypical series with large variability and, therefore, has clear advantages in empirical applications.

Chapter 3

An improved estimation method using correlation matrices

This chapter extends the proposed method introduced in Chapter 2, which is based on the use of correlation matrices, for the estimation of the factor space. We focus on the estimation of the DFM by means of non-parametric statistical tools. The most famous technique in this topic is Principal Component Analysis (PCA) which takes into account contemporaneous information about the data.

Up to our knowledge, little attention has been given to the estimation of the *common* component having into account past information, previous work in this topic are Peña and Box (1987), and Lam et al. (2011). We want to analyze how and in what degree different idiosyncratic error structures, which are more realistic than the classical scalar error structure, may affect to the estimation of the DFM. We compare the effect of different error structures on the PC estimator considered in Stock and Watson (2002), Bai and Ng (2002) and Bai and Ng (2006) between others; the pooling lagged estimator proposed in Lam et al. (2011), called LY estimator in what follows; and the one proposed in this chapter based on lagged correlation matrices, called CP in what follows. The main contribution is a Monte Carlo analysis of different data error structures in finite samples, where we analyze in deep the *exact* and the *approximate* DFM. Previous studies have considered contemporaneous covariance across the errors but not serial correlation, see Forni et al. (2005), or serial correlation and heteroscedasticity but not cross-correlation, see Stock and Watson (2005) and Breitung and Tenhofen (2011). As a novelty, we include the scenario where errors may present serial correlation and cross-section correlation for different sample sizes. The key idea of the PC estimator is to consider the covariance matrix of the observed data as a weighting average in order to estimate the factor space, whereas the pooling lagged estimators (LY and CP) take into account the accumulated

sum of lagged autocovariance/autocorrelation matrices of the observed data for the estimation. Therefore,we analyze the finite sample performances for the three estimators under different scenarios, filling the gap in nonparametric averaging methods.
The chapter is organized as follows; Section 1 introduces the DFM notation and the estimation methodologies. Section 2 illustrates the simulation exercise with the data-generating processes, scenarios and results. Section 3 provides an empirical application about CO_2 emissions. Finally, some concluding remarks and potential extensions are given in Section 4.

3.1 Theoretical framework: Nonparametric averaging methods

We consider the same DFM framework as in Chapter 2. The time series vector of observed data, $\mathbf{y}_t$ for $t = 1, ..., T$, is defined as:

$$\mathbf{y}_t = \mathbf{P}\mathbf{f}_t + \mathbf{e}_t \tag{3.1}$$

where $\mathbf{P}$ is the $(N \times r)$ matrix of factor loadings, with r being the number of common factors, $\mathbf{f}_t$ is the $(r \times 1)$ vector of common factors and $\mathbf{e}_t$ is a $(N \times 1)$ vector. Given that $\mathbf{P}\mathbf{f}_t = \mathbf{P}\mathbf{A}\mathbf{A}^{-1}\mathbf{f}_t$, with $\mathbf{A}$ being any nonsingular matrix, we need to assume that $\mathbf{P}'\mathbf{P} = \mathbf{I}_r$ and that the covariance matrix of the factors $\mathbf{\Gamma}_f(0) = E(\mathbf{f}_t\mathbf{f}_t')$ is diagonal, in order to uniquely define the factors. We also assume that the lag k covariance matrix of the factors $\mathbf{\Gamma}_f(k) = E(\mathbf{f}_t\mathbf{f}_{t-k}') \neq 0$, for some $k > 0$ to allow serial correlation in the common factors.

Assuming that the number of factors r is given, we analyze two non-parametric statistical tools for the estimation of model (3.1) in finite samples, and propose a new approach which generalizes the idea behind the LY estimator. Let $\mathbf{\Gamma}_y(k) = cov(\mathbf{y}_t, \mathbf{y}_{t-k})$, $\mathbf{\Gamma}_f(k) = cov(\mathbf{f}_t, \mathbf{f}_{t-k})$ and $\mathbf{\Gamma}_e(k) = cov(\mathbf{e}_t, \mathbf{e}_{t-k})$ be the lag k covariance matrices of the observed data $\mathbf{y}_t$, the common factors, $\mathbf{f}_t$, and the errors, $\mathbf{e}_t$, respectively.

Given the DFM and the independence between factors and noises we have that

$$\mathbf{\Gamma}_y(k) = \mathbf{P}\mathbf{\Gamma}_f(k)\mathbf{P}' + \mathbf{\Gamma}_e(k). \tag{3.2}$$

Suppose the simplest EDFM, where $\mathbf{\Gamma}_e(0) = \sigma^2\mathbf{I}$, then as

$$\mathbf{\Gamma}_y(0)\mathbf{P} = \mathbf{P}(\mathbf{\Gamma}_f(0) + \sigma^2\mathbf{I}) \tag{3.3}$$

the columns of $\mathbf{P}$ are the r eigenvectors of $\mathbf{\Gamma}_y(0)$ corresponding to their r largest eigenvalues $\gamma_{f_i}(0)+\sigma^2$ and the columns of the null space of $\mathbf{P}$ given by the $(N\times(N-r))$ matrix $\mathbf{P}^\perp$ such that $\mathbf{P}'\mathbf{P}^\perp = 0$, are the eigenvectors of the common eigenvalue σ^2. Based

on these results the PC estimator computes the first r leading eigenvectors, corresponding to the r largest eigenvalues, of the lag zero covariance matrix as estimates of the loadings $\mathbf{P}$. Also, from (3.1) we have $\mathbf{f}_t = \mathbf{P}'\mathbf{y}_t - \mathbf{P}'\mathbf{e}_t$ and for the central limit theorem the linear combination of the noises goes to zero and a natural estimate of the factors is $\mathbf{P}'\mathbf{y}_t$. Similar results can be obtained asymptotically when $N, T \to \infty$ if $\mathbf{\Gamma}_e(0) = \sigma^2\mathbf{D}$, being $\mathbf{D}$ a diagonal matrix, assuming that the signal to noise $s_0 = \gamma_f(0)/\sigma^2$ is large, where $\gamma_f(0) = \min_{i\in r}\gamma_{f_i}(0)$, and $\sigma^2 = \max_{i\in N}\sigma_i^2$. Again, in the ADFM, where $\mathbf{\Gamma}_e(0)$ is general, assuming weak cross correlation structure, as shown in Stock and Watson (2012) and Bai (2012) consistency can be obtained.

The second approach, proposed by Lam et al (2011) finds the r dominant eigenvectors of a combination of lagged covariance matrices given by

$$\mathbf{M}_{1,k_0} = \sum_{k=1}^{k_0} \mathbf{\Gamma}_y(k)\mathbf{\Gamma}_y(k)', \tag{3.4}$$

where $k_0 \geq 1$ is an arbitrary integer. Note that under the EDFM, as $\mathbf{\Gamma}_e(k) = 0$ for $k > 0$, we have

$$\mathbf{M}_{1,k_0} = \mathbf{P}\left(\sum_{k=1}^{k_0} \mathbf{\Gamma}_f^2(k)\right)\mathbf{P}' \tag{3.5}$$

and, therefore, the matrix $\mathbf{M}_{1,k_0}$ has rank r with the r dominant eigenvectors given by the columns of $\mathbf{P}$ and eigenvalues equal to $\sum_{k=1}^{k_0}\gamma_{f_i}^2(k)$. Calling $\gamma_f = \min_{i\in r}\sum_{k=1}^{k_0}\gamma_{f_i}^2(k)$ the signal to noise ratio is $s_M = \gamma_f/\sigma^4$. The larger this ratio is with respect to s_0 the better the performance of this method with respect to the covariance matrix when we have an homocedastic EDFM. However, if the dynamics of the factor is weak and some of them are white noise the advantages of this method may be worse than using the covariance matrix. On the other hand, in the heteroscedastic case, when $\mathbf{\Gamma}_e(0) = \sigma^2\mathbf{D}$, with $\mathbf{D}$ being a diagonal matrix with different diagonal elements, and in the ADFM when $\mathbf{\Gamma}_e(0)$ is general, the eigenvectors of $\mathbf{M}$ estimate the loading matrix in finite samples whereas those of $\mathbf{\Gamma}_y(0)$ does not. Thus, in this case if also the factors have strong dynamics this procedure is expected to work better than using the covariance matrix.

As we introduced in Chapter 2, these two estimation methodologies present a lack of robustness to the existence of a few atypical series. In order to overcome this limitation, we consider the new approach which generalize the LY estimator. The *combined dynamic correlation matrix* is

$$\mathbf{R}_{k_0} = \sum_{k=0}^{k_0} w_k \boldsymbol{R}_y(k)\boldsymbol{R}_y(k)' \tag{3.6}$$

where $\boldsymbol{R}_y(k) = cor(\mathbf{y}_t, \mathbf{y}_{t-k})$ is the lag k correlation matrix of the series in model (3.1), and the coefficients $w_k > 0$ are weights which verify $\sum_{k=0}^{k_0} w_k = 1$. The weights, which

were introduced in Chapter 2, are defined as $w_k = (T-k)/\left((k_0+1)(T-k_0/2)\right)$ using the standardized squared correlations, r_{ij} for $i,j = 1,...,N$, with asymptotic variance $var(r_{ij}(k)) \approx (T-k)^{-1}$.

Note that we now incorporate in the $\mathbf{R}_{k_0}$ matrix the information in the lag zero correlation matrix $\mathbf{R}_y(0)$, and the weights define the relative importance given by this matrix and the $\mathbf{M}_{1,k_0}$ matrix with correlation matrices $\mathbf{R}_y(k)$ instead of the variance matrices $\mathbf{\Gamma}_y(k)$.

The factor space in model (3.1) is estimated using the eigen decomposition of the combined correlation matrix $\mathbf{R}_{k_0}$ with standardized data. The r columns of $\mathbf{P}$ are the eigenvectors associated to the r largest eigenvalues of $\mathbf{R}_{k_0}$, in descending order. Finally, the estimates of the common factors are obtained as $\mathbf{f}_t = \mathbf{P}'\mathbf{y}_t$.

3.2 Monte Carlo experiment of estimation performance

In this section we compare the estimation performance of the three procedures considered in this chapter: PC, LY, and the one proposed in previous section, CP, for the estimation of the factor space. For this comparison we simulate data samples in a similar manner to the ones generated in Chapter 2. We consider six different data generating processes (DGP) depending on the error structure. The signal to noise ratios are strong, medium and weak.

Each data point is generated following the equation

$$y_{it} = \mathbf{p}_i'\mathbf{F}_t + e_{it}, \tag{3.7}$$

where the factor loading coefficients $\mathbf{p}_i$ are generated from the U(-0.5, 0.5) distribution, and the common factor, f_t, follows an autoregressive process of order 1, given by:

$$f_t = \phi f_{t-1} + \eta_t. \tag{3.8}$$

The first three DGP have serially uncorrelated noises with different idiosyncratic covariance matrices, $\Gamma_e(0)$, and the next three add serially correlated noises to the basic covariance structure. Thus, DGP_1 has homoscedastic errors and $\Gamma_e(0) = \sigma^2\mathbf{I}$, with $\sigma^2_{e_i} = 1$ for $i = 1,...,N$, DGP_2 has heteroscedastic uncorrelated noises with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and DGP_3 has heteroscedastic and cross-section correlated errors, with $\Gamma_e(0)$ having diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. The next three DGP add to the three previous scenarios serially correlated errors $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1,...,N/2$ with

$\theta \sim N(0.5, 0.05)$, $u_{it} \sim N(0,1)$ white noise, and uncorrelated errors with $\sigma^2_{e_i} = 1$ for $i = N/2 + 1, ..., N$. Thus, this modification of the first DGP, that will be called $DGPC_1$, has $\Gamma_e(0)$ with variances around $1/(1 - .5^2) = 1.33$ in half of the matrix and variances equal to one in the other half. Also $\Gamma_e(k)$ for $k > 0$ is a diagonal matrix different from zero. $DGPC_2$ has heteroscedastic variances around $1/(1 - .5^2) = 1.33$ in half of $\Gamma_e(0)$ and variances equal to two in the other half, and non null diagonal lagged covariance matrices. Finally $DGPC_3$ has full rank $\Gamma_e(0)$ with different variances in half of the matrix and equal to two in the other half, non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$, and non null $\Gamma_e(k)$ lagged covariance matrices.

Three signal to noise ratios (SN) are considered: Strong, with autoregressive coefficient $\phi_{f_1} = 0.9$; Medium with $\phi_{f_1} = 0.6$; Weak with $\phi_{f_1} = 0.3$,. Errors η_t are independent $N(0,1)$ random variables. The errors η_t are independent of the idiosyncratic errors e_t, such that $E(\eta_t e_t') = 0$, for all data generating processes.
Finally, scenarios DGP_4, DGP_5, and DGP_6 and $DGPC_4$, $DGPC_5$, and $DGPC_6$ follow the same idiosyncratic covariance structures as DGP_1, DGP_2, and DGP_3, respectively, but with two common factors ($r = 2$). Each common factor follows an AR(1) process with $\phi_{f_1} = 0.9$, and $\phi_{f_2} = 0.6$ in scenarios with strong SN ratio, and $\phi_{f_1} = 0.6$ and $\phi_{f_2} = 0.3$ in scenarios with medium and weak SN ratio. For the first two scenarios errors η_t are generated as independent $N(0,1)$ random variables and for the last scenario they are generated as independent $N(0, 0.5)$ random variables.
The number of cross-section variables considered are $N = 10, 50, 100, 200$, and the number of time observations are $T = 125, 250, 500, 1250$. For each one of the different (N, T) combinations we run 200 iterations and calculate the mean of a similarity measure that compares the linear space spanned by the columns of the theoretical loadings, $\mathcal{M}(\mathbf{P})$, with the linear space spanned by the columns of the estimated loadings, $\mathcal{M}(\widehat{\mathbf{P}})$, using the following expression

$$S(\mathcal{M}(\mathbf{P}), \mathcal{M}(\widehat{\mathbf{P}})) = \frac{tr(H_{\mathbf{P}} H_{\widehat{\mathbf{P}}})}{r} \tag{3.9}$$

where $\mathbf{P}$ is the theoretical loading matrix having rank r, $\widehat{\mathbf{P}}$ is the estimated loading matrix having rank r, and $H_{\mathbf{P}} = \mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'$ and $H_{\widehat{\mathbf{P}}} = \widehat{\mathbf{P}}(\widehat{\mathbf{P}}'\widehat{\mathbf{P}})^{-1}\widehat{\mathbf{P}}'$. This measure is equal to 1 if and only if $\mathcal{M}(\widehat{\mathbf{P}}) \subset \mathcal{M}(\mathbf{P})$ or $\mathcal{M}(\mathbf{P}) \subset \mathcal{M}(\widehat{\mathbf{P}})$, and is equal to 0 if and only if $\mathcal{M}(\mathbf{P}) \perp \mathcal{M}(\widehat{\mathbf{P}})$.
The number of lags considered in the sum of lagged covariances matrices is $k_0 = 2$, see Lam et al. (2011) where they shown that the method is not sensitive to the choice of k_0.

We present the results for seven of the 18 scenarios considered in the Monte Carlo exercise. These tables provide a general picture of the results and summarize the main highlights from each method. The rest of the tables are available in the B.1 Tables section

in the Appendix to Chapter 3. Tables 3.1 - 3.4 show the results for the scenarios with one factor and Tables 3.5 - 3.7 for the scenarios with two factors. Results by columns are the mean of the similarity measure between the original loading matrix $\mathbf{P}$ and the estimated ones $\widehat{\mathbf{P}}$ using $\mathbf{\Gamma}(0)$, $\mathbf{M}$ with $k_0 = 2$, and the combined dynamic correlation matrix $\mathbf{R}$, respectively. In general, although the three methodologies provide similar results under DGP_1 when the errors are homocedastic, we identify big differences when we assume a more realistic idiosyncratic error structure under DGP_2 and DGP_3. The relative precision growth rate (RPG) of using lags over $\mathbf{\Gamma}(0)$ in the estimation is

$$RPG = \frac{(S - S_0)}{S_0}. \tag{3.10}$$

.

where S is the mean of the similarity measure using the LY method with the $\mathbf{M}$ matrix or the CP method with the $\mathbf{R}$ matrix, and S_0 is the mean of the similarity measure using PC with the variance-covariance matrix $\mathbf{\Gamma}(0)$.

We can see in all tables that the estimation of $\mathbf{P}$ demonstrates "the blessing of dimensionality", for fixed sample size T, the mean of the similarity measure for $\widehat{P} = P$ increase for $N = 10, 50, 100, 200$. In general, differences depend on the number of series N, and the signal to noise ratio. The last three columns in each table report the estimations when the idiosyncratic terms are autocorrelated. Under this scenario, all of the estimators have a decrease of power to estimate the true factors. In broad terms autocorrelated errors have a similar effect to decrease the SN ratio. When the idiosyncratic terms have approximately the same variance, as in Table 3.1, the three methods are similar, but PC and CP are slightly better than LY. Only for data samples with very large number of observations $T = 1250$ LY has similar performance. Table 3.2 presents results for the one factor model with heteroscedastic errors and medium signal to noise ratio. In this more realistic scenario the CP method clearly outperforms the other two especially when the errors are correlated. When the errors are white noise we find large differences for small number of series $N = 10, 50$. Similar results are presented in Table 3.3 for weak SN ratio. For example, when $N = 50$ the advantage of using CP over PC is approximately of 57%, this difference could be even larger than 100% with respect to LY. The performance of the CP method is less sensitive than PC and LY to autocorrelated errors, see last three columns in Table 3.3. Finally, if in addition to heteroscedasticity we consider lag zero cross-section correlation, as in Table 3.4, with medium SN ratio, LY and CP always provide better estimations than PC. For example, for $N = 100$ and $T = 125$ the performance of the CP method without autocorrelated noise is 95% more powerful than PC, and with autocorrelated errors the differences in power are even greater: for $N = 100$ and $T = 125$ CP gives a RPG of 139% with respect to PC and

165% with respect to LY. As expected, the performance of PC is much more sensitive to cross-section correlation than CP and LY methodologies.

Table 3.1: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Homoscedastic errors and medium signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.866	0.725	0.839	0.772	0.486	0.753	0.74
T=250	0.937	0.863	0.925	0.864	0.605	0.86	0.84
T=500	0.971	0.933	0.962	0.913	0.694	0.921	0.9
T=1250	0.989	0.971	0.977	0.961	0.808	0.97	0.95
N=50							
T=125	0.928	0.842	0.919	0.88	0.714	0.857	0.86
T=250	0.966	0.912	0.96	0.938	0.827	0.924	0.92
T=500	0.983	0.951	0.978	0.968	0.899	0.958	0.96
T=1250	0.993	0.98	0.989	0.987	0.955	0.981	0.98
N=100							
T=125	0.934	0.845	0.924	0.889	0.727	0.861	0.86
T=250	0.967	0.913	0.961	0.942	0.841	0.926	0.93
T=500	0.983	0.953	0.978	0.971	0.912	0.96	0.96
T=1250	0.993	0.981	0.99	0.988	0.962	0.981	0.98
N=200							
T=125	0.935	0.85	0.925	0.891	0.74	0.864	0.87
T=250	0.967	0.914	0.959	0.945	0.849	0.927	0.93
T=500	0.984	0.954	0.979	0.973	0.917	0.961	0.96
T=1250	0.994	0.981	0.99	0.989	0.964	0.981	0.98
Mean	**0.96**	0.91	0.95	**0.93**	0.81	0.92	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$ and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix $\Gamma_e = \sigma_e \mathbf{I}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table 3.2: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Heteroscedastic errors and medium signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.276	0.287	0.648	0.245	0.191	0.541	0.36
T=250	0.364	0.494	0.809	0.259	0.208	0.692	0.47
T=500	0.385	0.63	0.857	0.3	0.21	0.799	0.53
T=1250	0.468	0.852	0.915	0.276	0.2	0.868	0.6
N=50							
T=125	0.672	0.634	0.801	0.442	0.261	0.682	0.58
T=250	0.807	0.772	0.859	0.666	0.432	0.808	0.72
T=500	0.873	0.871	0.891	0.777	0.569	0.861	0.81
T=1250	0.917	0.944	0.91	0.859	0.74	0.897	0.88
N=100							
T=125	0.786	0.685	0.805	0.626	0.366	0.711	0.66
T=250	0.88	0.814	0.861	0.8	0.599	0.81	0.79
T=500	0.932	0.889	0.89	0.885	0.748	0.862	0.87
T=1250	0.965	0.952	0.908	0.939	0.871	0.895	0.92
N=200							
T=125	0.829	0.719	0.81	0.723	0.512	0.727	0.72
T=250	0.908	0.823	0.862	0.844	0.675	0.811	0.82
T=500	0.951	0.895	0.889	0.917	0.804	0.861	0.89
T=1250	0.978	0.953	0.906	0.963	0.904	0.893	0.93
Mean	0.75	0.76	**0.85**	0.66	0.52	**0.79**	

NOTES: Factor autoregressive coefficient $\phi_{f_1} = 0.6$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table 3.3: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Heteroscedastic errors and weak signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.214	0.171	0.493	0.186	0.13	0.37	0.26
T=250	0.258	0.2	0.672	0.182	0.125	0.527	0.33
T=500	0.296	0.19	0.807	0.226	0.146	0.663	0.39
T=1250	0.361	0.334	0.892	0.241	0.136	0.823	0.46
N=50							
T=125	0.472	0.262	0.74	0.227	0.066	0.585	0.39
T=250	0.656	0.378	0.829	0.422	0.101	0.772	0.53
T=500	0.781	0.556	0.875	0.56	0.1	0.844	0.62
T=1250	0.844	0.732	0.901	0.723	0.123	0.887	0.7
N=100							
T=125	0.678	0.466	0.771	0.436	0.091	0.659	0.52
T=250	0.815	0.604	0.842	0.673	0.153	0.792	0.65
T=500	0.885	0.715	0.876	0.813	0.193	0.853	0.72
T=1250	0.937	0.816	0.9	0.898	0.306	0.887	0.79
N=200							
T=125	0.764	0.603	0.785	0.627	0.182	0.705	0.61
T=250	0.87	0.698	0.847	0.793	0.284	0.805	0.72
T=500	0.927	0.762	0.879	0.889	0.451	0.857	0.79
T=1250	0.966	0.848	0.901	0.949	0.649	0.888	0.87
Mean	0.67	0.52	**0.81**	0.55	0.2	**0.74**	

NOTES: Factor autoregressive coefficient $\phi_{f_1} = 0.3$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table 3.4: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Heteroscedastic and cross correlated errors, and medium signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.126	0.213	0.134	0.11	0.141	0.114	0.14
T=250	0.121	0.323	0.125	0.123	0.117	0.126	0.16
T=500	0.131	0.408	0.13	0.12	0.133	0.13	0.18
T=1250	0.133	0.737	0.132	0.119	0.11	0.121	0.23
N=50							
T=125	0.093	0.314	0.226	0.098	0.113	0.207	0.18
T=250	0.106	0.588	0.299	0.069	0.091	0.189	0.22
T=500	0.117	0.822	0.343	0.071	0.095	0.229	0.28
T=1250	0.126	0.932	0.36	0.065	0.091	0.251	0.3
N=100							
T=125	0.315	0.544	0.614	0.184	0.166	0.44	0.38
T=250	0.425	0.762	0.749	0.202	0.206	0.607	0.49
T=500	0.529	0.872	0.807	0.227	0.25	0.711	0.57
T=1250	0.646	0.946	0.848	0.264	0.341	0.782	0.64
N=200							
T=125	0.685	0.667	0.767	0.469	0.313	0.638	0.59
T=250	0.83	0.815	0.84	0.674	0.494	0.772	0.74
T=500	0.895	0.893	0.874	0.791	0.665	0.831	0.82
T=1250	0.934	0.953	0.896	0.88	0.839	0.874	0.9
Mean	0.39	**0.67**	0.51	0.28	0.26	**0.44**	

NOTES: Factor autoregressive coefficient $\phi_{f_1} = 0.6$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Tables 3.5 - 3.7 give some results for scenarios with two factors. Table 3.5 give the results under weak strength factors and homoscedastic series. This table is similar to Table 3.1 but now with two factors. The best performance correspond to PC followed closely by CP and both being more powerful than LY. Table 3.6 presents the result for heteroscedastic time series and the advantages of CP with respect to the other two are clear especially when the errors are autocorrelated. Finally Table 3.7 gives the performance with heteroscedasticiy and cross-section correlation. When the noises are not autocorrelated the best performance correspond to LY, but when the noises are autocorrelated CP is more powerful than the others. In summary, we conclude that the proposed CP method provides overall the more powerful test for the number of common factors when errors

present heteroscedasticity and autocorrelation and when the noises are homoscedastic white noises have a performance close to the best test in this case.

Table 3.5: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Homocedastic errors and weak signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.265	0.188	0.258	0.201	0.137	0.214	0.21
T=250	0.33	0.246	0.33	0.24	0.153	0.27	0.26
T=500	0.395	0.299	0.394	0.279	0.173	0.33	0.31
T=1250	0.457	0.379	0.45	0.352	0.203	0.405	0.37
N=50							
T=125	0.344	0.272	0.334	0.243	0.126	0.229	0.26
T=250	0.412	0.351	0.41	0.337	0.182	0.327	0.34
T=500	0.453	0.404	0.451	0.411	0.268	0.41	0.4
T=1250	0.477	0.448	0.477	0.458	0.365	0.46	0.45
N=100							
T=125	0.371	0.312	0.363	0.285	0.159	0.257	0.29
T=250	0.428	0.373	0.423	0.379	0.249	0.362	0.37
T=500	0.46	0.417	0.457	0.432	0.334	0.423	0.42
T=1250	0.482	0.456	0.48	0.469	0.41	0.465	0.46
N=200							
T=125	0.384	0.33	0.376	0.312	0.2	0.281	0.31
T=250	0.436	0.384	0.43	0.393	0.29	0.372	0.38
T=500	0.466	0.425	0.462	0.441	0.364	0.429	0.43
T=1250	0.485	0.463	0.483	0.474	0.429	0.468	0.47
Mean	**0.42**	0.36	0.41	**0.36**	0.25	**0.36**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, and $\phi_{f_2} = 0.3$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ for $i = 1, ..., N$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table 3.6: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Heteroscedastic errors and medium signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.21	0.203	0.376	0.174	0.139	0.32	0.24
T=250	0.242	0.271	0.426	0.181	0.134	0.381	0.27
T=500	0.264	0.331	0.446	0.214	0.147	0.428	0.3
T=1250	0.287	0.421	0.46	0.215	0.156	0.452	0.33
N=50							
T=125	0.345	0.308	0.405	0.258	0.154	0.353	0.3
T=250	0.408	0.386	0.431	0.347	0.222	0.407	0.37
T=500	0.438	0.432	0.445	0.395	0.288	0.432	0.4
T=1250	0.457	0.466	0.455	0.433	0.375	0.451	0.44
N=100							
T=125	0.399	0.356	0.408	0.335	0.222	0.363	0.35
T=250	0.44	0.402	0.433	0.401	0.3	0.407	0.4
T=500	0.466	0.443	0.446	0.442	0.377	0.433	0.43
T=1250	0.481	0.472	0.454	0.468	0.433	0.447	0.46
N=200							
T=125	0.418	0.366	0.408	0.364	0.259	0.365	0.36
T=250	0.454	0.411	0.433	0.423	0.338	0.407	0.41
T=500	0.475	0.448	0.445	0.458	0.4	0.432	0.44
T=1250	0.488	0.475	0.455	0.48	0.451	0.447	0.47
Mean	0.39	0.39	**0.43**	0.35	0.27	**0.41**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, and $\phi_{f_2} = 0.3$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table 3.7: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Heteroscedastic and cross correlated errors, and medium signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.118	0.159	0.124	0.118	0.116	0.124	0.13
T=250	0.112	0.184	0.117	0.105	0.108	0.112	0.12
T=500	0.112	0.266	0.115	0.112	0.105	0.112	0.14
T=1250	0.106	0.399	0.11	0.1	0.108	0.099	0.15
N=50							
T=125	0.089	0.19	0.171	0.063	0.064	0.114	0.12
T=250	0.082	0.296	0.167	0.059	0.059	0.123	0.13
T=500	0.092	0.408	0.198	0.061	0.065	0.148	0.16
T=1250	0.095	0.461	0.211	0.051	0.054	0.143	0.17
N=100							
T=125	0.184	0.292	0.335	0.108	0.101	0.246	0.21
T=250	0.238	0.383	0.381	0.128	0.116	0.312	0.26
T=500	0.299	0.437	0.406	0.143	0.141	0.353	0.3
T=1250	0.324	0.47	0.425	0.156	0.171	0.393	0.32
N=200							
T=125	0.357	0.349	0.39	0.259	0.175	0.328	0.31
T=250	0.414	0.404	0.42	0.343	0.26	0.388	0.37
T=500	0.447	0.446	0.439	0.4	0.342	0.418	0.42
T=1250	0.468	0.475	0.45	0.439	0.416	0.438	0.45
Mean	0.22	**0.35**	0.28	0.17	0.15	**0.24**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, and $\phi_{f_2} = 0.3$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

3.3 An application to real data: CO_2 emissions

The primary driver of global climate change are the carbon dioxide emissions. We are interested in evaluating the global behavior of CO_2 emissions around the world given its direct impact to global warming, threatening human and natural habitats. The dataset which is available at www.worldbank.org includes information about emissions in 124 countries from 1960 to 2016. Each time series is in logs and first differenced previous to the analysis. Then, the data matrix is (56×124). We implement the three estimation methodologies, PC, LY and the one proposed in this chapter, CP, in order to estimate the common factors. First, we estimate the number of common factors, $\widehat{r}$, using eigenvalues ratio tests, such as Ahn and Horenstein (2013) test for PC, Lam et al. (2011) criterion for LY, and Caro and Peña (2020) test for CP. Each criterion estimates one common factor, $\widehat{r} = 1$. Given that PC and LY methodologies provide pretty similar results, we just consider here the ones from PC. Figure 3.1 represents the estimated common factor, $\widehat{f}_t = y_t \widehat{P}'$, using PC (black line) together with the time series of Cameroon (red line). We can see how this common factor which is the normalized eigenvector of the covariance matrix $\widehat{\mathbf{\Gamma}}(0)$, associated to its largest eigenvalue, is mainly the series of Cameroon. This happens because the large variability presented in Cameroon time series makes that the first principal component is mostly generated by this series. The corresponding factor loadings are plotted in Figure 3.2 top panel. It is clear that the CO_2 emissions time series from Cameroon is the one with largest effect in the estimated factor. Same result is obtained using LY methodology given that the $\widehat{\mathbf{M}}$ matrix also takes into account lagged covariance matrices $\widehat{\mathbf{\Gamma}}(k)$.

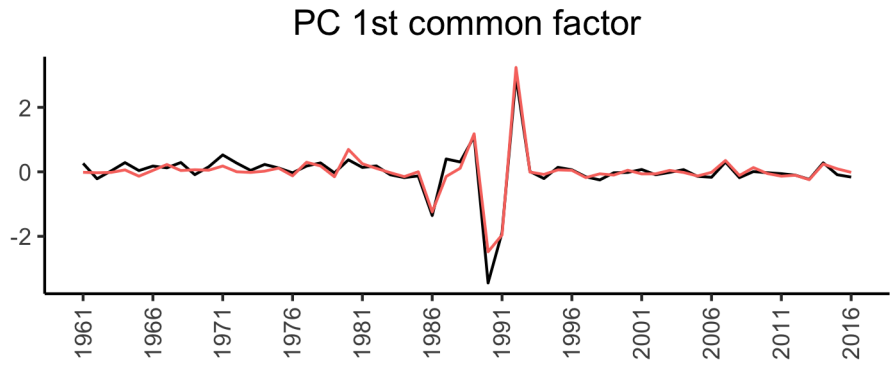

Figure 3.1: PC first estimated factor (black line) and CO_2 emissions in Cameroon (red line).

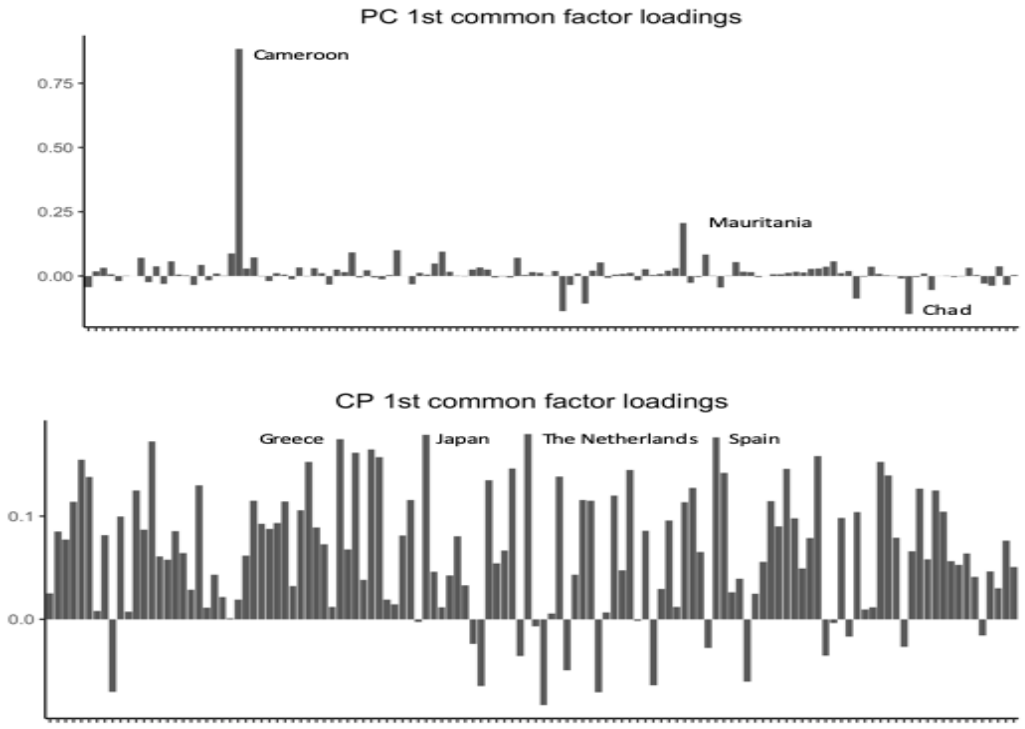

Figure 3.2: PC and CP estimated factor loadings.

We can see in the top panel of Figure 3.3 the first common factor using the proposed approach, CP, which combine lagged information by means of the combined dynamic correlation matrix $\widehat{\mathbf{R}}$. Contrary to PC and LY methods, this common factor is able to capture the overall dynamics of CO_2 emissions for all the countries included in the analysis representing a decreasing trend at the end of the sample period. We plot together to CP first common factor the series of CO_2 emissions for Japan, Greece, The Netherlands and Spain, which correspond with the four largest loadings coefficients, see the bottom panel in Figure 3.2. In terms of the percentage of the total variance of y_t explained by the estimated common factors, the first factor computed from the covariance matrix, called to simplify PC factor, accounts for the 14% of the total variability. The factor computed from the LY matrix accounts for 23% of the variability of the sum of squared covariance matrices and the factor computed from the correlation matrices accounts also for 23% of the variability of the standardized data.

The empirical application has shown the lack of interpretability of PC and LY when atypical series are presented, whereas the CP method overcomes this limitation providing interpretable results for researchers and practitioners.

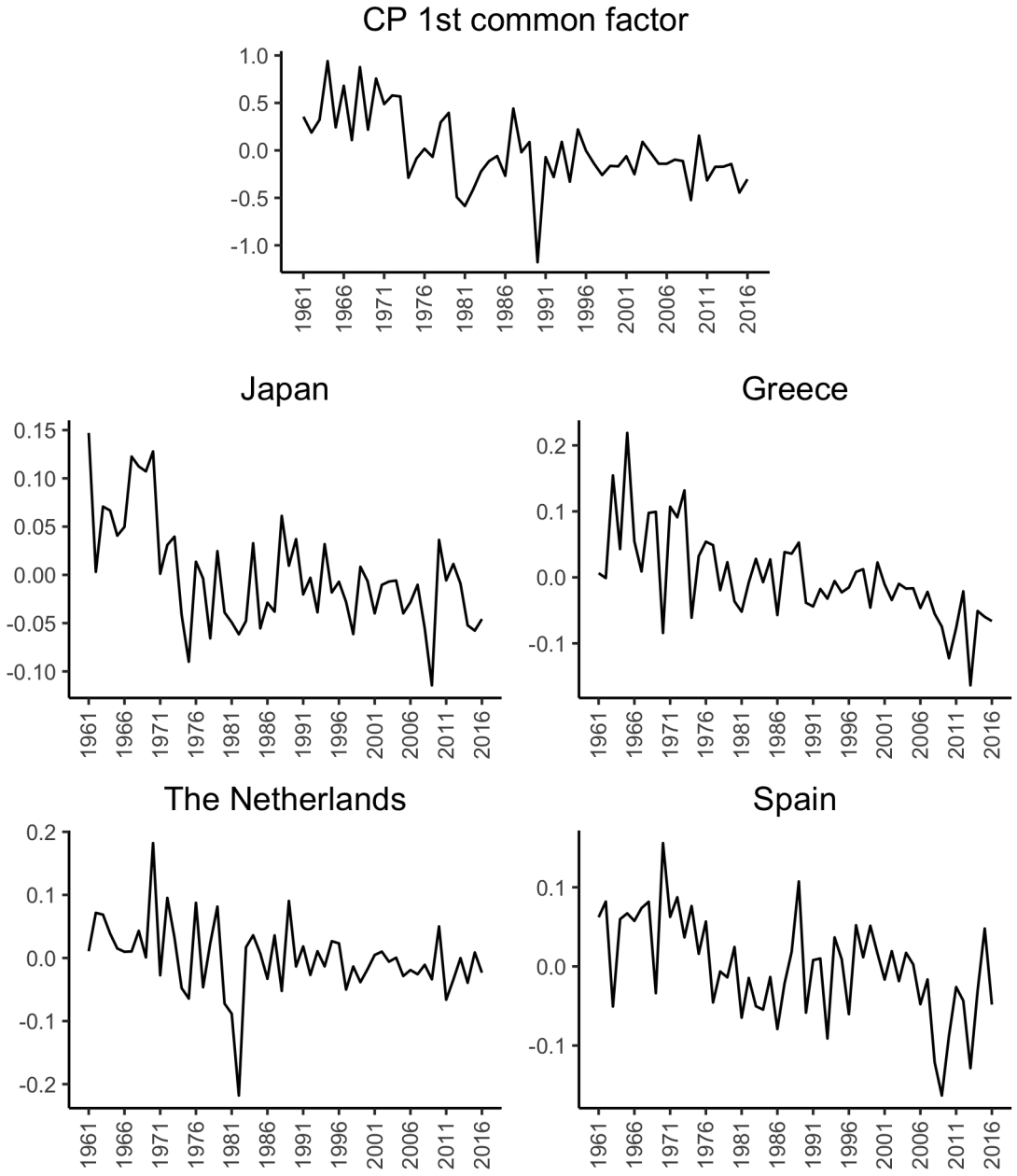

Figure 3.3: CP first estimated factor and CO_2 emissions in Japan, Greece, The Netherlands and Spain.

3.4 Concluding remarks

This chapter has evaluated the finite sample performances of the principal component estimator, the estimator based on the eigenvectors of the pooling lagged matrices and the proposed estimator based on the eigenvectors of the combined dynamic correlation matrix. Some simulation experiments have been conducted to analyze for which sample size, T, and dimension of time series, N, would be more advantageous to consider the classical principal component estimator or the lagged estimators. Simulation results comparing the three methodologies, under different idiosyncratic error structures, have shown that the Relative Precision Growth rate of using CP procedure with the $\mathbf{R}$ matrix can be up to 140%, whereas the disadvantages would be as maximum of 3%. Furthermore, these gains would be obtain under a more realistic error structure than the classical one with homocedastic errors, given that errors may present some degree of serial and cross-sectional dependence.

Chapter 4

DFM with known cluster structure: An application to business cycles

The European Union (EU) was considered a successful process of integration, which could contribute to the economic development and the creation of wealth among its members. However, after the financial and the sovereign debt crises, the initial Euro-enthusiasm was followed by doubts about the advantages of the union. The differences in the economic performance of the EU members after these crises led to a phenomenon known as the *Two-Speed Europe* with two groups of countries: *Core countries*, formed by states with similar fiscal restraint and solid economic growth, and a second group of *Peripheral countries*, with weaker economic performance and higher fiscal deficits and public debt.

The collapse of the financial system after 2008, the sluggish economic recovery, and the deflationary pressures forced the European Central Bank (ECB) to employ a set of unconventional monetary policies. The disparity in the economic performance of the members, which shared the common expansionary monetary policy, also raised new concerns about the benefits of the existence of a monetary union. For these reasons, analyzing the characteristics of the European countries' business cycles has been a source of research in the literature. Among others, Camacho et al. (2006) and Borsi and Metiu (2015) used a single country-specific indicator of aggregate economic activity as the Industrial Production Index or the Gross Domestic Product (GDP) to evaluate the economic convergence and business cycle synchronization across the members. Similarly, Di Giorgio (2016) implemented Markov-switching (MS) models to estimate the changes in the business cycle phases of the Euro Area (EA) and some Central and Eastern European countries (CEECs) by using one series of GDP per each CEEC and a series for the aggregate GDP of the EA. Nevertheless, as the author acknowledges, a univari-

ate analysis may fail to capture some recessionary phases and, under the event of imperfect cyclical synchronization among macroeconomic categories, several variables per country should be considered for the estimation of the business cycle.

With the aim of filling this gap, this chapter examines the disparities in the evolution of the business cycle synchronization across the members of the EA by proposing the following two-step procedure. In the first step, we obtain EA and country specific measures of aggregate economic activity by constructing a large dataset of cross-country series from several macroeconomic categories, whose co-movements are captured by a Dynamic Factor Model with Cluster Structure (DFMCS). In the case of Europe, previous studies have relied in DFMs to describe macroeconomic interactions between CEECs and some EA members, see Breitung and Eickmeier (2006) and Jiménez-Rodríguez et al. (2013). Recently, Klaus and Ferroni (2015) estimate a DFM to analyze the business cycles characteristics of the four largest EA countries. In our proposal, we use the DFM of Kose et al. (2003) and Crucini et al. (2011) because it allows us to distinguish between common sources of variation in the Union and nation-specific factors.

In the second step, the measures of aggregate economic activities obtained in the factor analysis are used in the Markov-switching framework developed by Leiva-Leon (2017) to draw inferences about the synchronization of business cycles across the EA members. In contrast to other standard approaches, which summarize the overall level of synchronization in a single number for the entire sample period, this multivariate Markov-switching approach allows us to compute a measure of pairwise synchronization at each time observation along the sample. Therefore, we can examine the evolution of the time-varying dynamic interactions across the business cycles of the EA members. See Égert and Kočenda (2011) which examine the time-varying synchronization across European stock markets.)

Using a recent dataset, which encompasses the financial and the sovereign debt crises, we find that, overall, the degree of synchronization of the EA members remained stable until the financial crisis, which implied a dramatic reduction in the degree of synchronization due to the different effects of this shock on each country. Thereafter, all the countries showed a progressive recovery in the synchronization to pre-crisis levels. Notably, we find significant discrepancies in the recovery paths. Some countries have been able to catch up their pre-crisis level of synchronization very fast, letting even some countries to improve their initial levels. However, some EA members are still far from recovering their pre-crisis degrees of business cycle synchronization. In an independent proposal, based on GDP data at the regional level of the NUTS2 classification, Gadea-Rivas et al. (2019) have also documented the different patterns in the synchronization in Europe since the introduction of the euro. These findings support the existence of a

Two-Speed Europe in terms of synchronization. The rest of the chapter is organized as follows. Section 1 describes the DFM with cluster structure and Section 2 the measure of synchronization between the common factor and each of the country-specific factors. Section 3 examines the empirical results and Section 4 states the conclusions.

4.1 A model to examine synchronization

This section describes the procedure applied for the estimation of the latent factors that summarize the common behavior of the economic indicators.

The estimation of the factors is based on the DFM proposed in Crucini et al. (2011). This is a suitable framework to deal with the large dimension of the dataset and the specific characteristic of the cross-country data. In particular, each macroeconomic indicator in a given country is assumed to be explained by three components: a common latent factor affecting all the series in the dataset; a second latent factor, which only affects the group of series in a particular country; and an idiosyncratic term, which is specific to each series. Hence, every data observation is decomposed according to the following equation:

$$y_{i,t} = \alpha_i + \beta_{EA,i} f_{EA,t} + \beta_{n,i} f_{n,t} + \varepsilon_{i,t}, \tag{4.1}$$

where $f_{EA,t}$ is the EA factor; $f_{n,t}$ is the country factor; $n = 1, \ldots, N$, where N is the number of countries; and $\varepsilon_{i,t}$ is the idiosyncratic component. Observable series at period t are denoted as $y_{i,t}$ for $i = 1, ..., M \times N$, where M is the number of macroeconomic series per country. The factor loadings, $\beta_{EA,i}$ and $\beta_{n,i}$, measure the amount of variation in $y_{i,t}$ that is explained by each factor. The dynamic of the factors is assumed to follow an autoregressive process of order p_k:

$$f_{k,t} = \phi_{f_{k,1}} f_{k,t-1} + \phi_{f_{k,2}} f_{k,t-2} + \cdots + \phi_{f_{k,p}} f_{k,t-p_k} + u_{f_k,t}, \tag{4.2}$$

where $E\left[u_{f_k,t} u_{f_k,t}\right] = \sigma^2_{f_k}$, $k = 1, ..., K$, and K is the number of latent factors. In addition, the idiosyncratic terms, $\varepsilon_{i,t}$, are assumed to follow autoregressive processes of order q_i:

$$\varepsilon_{i,t} = \phi_{i,1} \varepsilon_{i,t-1} + \phi_{i,2} \varepsilon_{i,t-2} + \cdots + \phi_{i,q_i} \varepsilon_{i,t-q_i} + u_{i,t}, \tag{4.3}$$

where $E\left[u_{i,t} u_{j,t-s}\right] = \sigma^2_i$ for $i = j, s = 0$, and 0 otherwise; $E\left[u_{f_k,t} u_{i,t-s}\right] = 0$ for all k, i, and s. Following Kose et al. (2003) and Crucini et al. (2011), we set the lag length of the autoregressive processes to three. The error terms $u_{f_k,t}$ and $u_{i,t}$ are assumed to be normally distributed variables with zero mean. In the empirical application, the dataset

is formed by eighteen countries $(N = 18)$, and there are nineteen dynamic unobserved factors $(K = N + 1)$ that represent the common interrelations that take place in the cross-country dataset.

The estimation of the multifactor model (1) - (3) for a large set of countries requires the estimation of the latent factors and a sizable number of parameters relating them with the observable series. Moreover, given the short life of the EA, the temporal dimension of the dataset is relatively small with respect to cross section dimension. For these reasons, we apply the Bayesian estimation procedure proposed by Kose et al. (2003) and Crucini et al. (2011), which has been shown to work efficiently in this context. PP4(Kose et al. (2003) estimate world, region and country factors for 60 countries with 30 years of annual data.)

Let θ be the set of parameters to be estimated, F the vector of dynamic latent factors $(KT \times 1)$ with Gaussian probability density $p_f(F)$, and Y the vector of observable data $(MNT \times 1)$. The Gaussian probability density for the observable series conditional on the factors and the parameters is $p_y(Y \mid \theta, F)$. According to the Bayes' theorem, for a given prior distribution of θ, $p(\theta)$, the joint posterior distribution for the factors and parameters is the product of the likelihood and prior:

$$p(\theta, F \mid Y) \subset p_y(Y \mid \theta, F)p_f(F)p(\theta). \tag{4.4}$$

However, while the joint posterior is difficult to handle, a sample of θ and F can be generated by means of Markov Chain Monte Carlo methods sampling from the conditional density of the parameters given factors and data and the conditional density of the factors given parameters and data. Specifically, the parameters and the factors are generated by sampling both iteratively from the next two steps:

1. Sampling θ^1 from the conditional density $p\left(\theta \mid F^0, Y\right)$ where F^0 is a starting value in the support of the posterior distribution of F.

2. Sampling F^1 from the conditional density $p\left(F \mid \theta^1, Y\right)$. In a first step, we sample from the distribution of the EA factor conditional on the parameters and the specific country factors. In a second step, we sample from the distribution of each country factor conditional on the EA factor and the parameters.

These two steps generate in each stage of the Markov Chain drawings $\left\{\theta^j, F^j\right\}$ for $j = 1, \ldots, J$ where $\theta^j \sim p\left(\theta \mid F^{j-1}, Y\right)$ and $F^j \sim p\left(F \mid \theta^{j-1}, Y\right)$. Given the proper priors, this iterative process produces a realization of a Markov chain whose invariant distribution is the joint posterior of the model parameters and the unobservable factors (Otrok and Whiteman 1998). The posterior distribution for the parameters is

built by computing the likelihood for the first p_i observations, sequentially conditioning to compute the rest of the likelihood and using the usual prior densities which are considered sufficiently uninformative. To be precise, the prior for the factor loadings is $(\beta_{EA,i}\beta_{n,i})' \sim N\left(0, (0.001 * I_2)^{-1}\right)$, where I_2 is the 2×2 identity matrix.

Autoregressive parameters for both, the factors and the idiosyncratic components, $\phi_i = (\phi_{i,1}, \phi_{i,2}, \phi_{i,3})'$ and $\phi_{f_k} = (\phi_{f_k,1}, \phi_{f_k,2}, \phi_{f_k,3})'$, follow a prior distribution $N(0, \Sigma)$ with

$$\Sigma = \begin{bmatrix} 0.85 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.25 \end{bmatrix}. \tag{4.5}$$

This represents the belief that data in growth rates is not serially correlated and that the impacts of the lags mitigate as the lag length increases. As in the previous literature, to identify the scale of the latent factors, $\sigma^2_{f_k}$ are set equal to a constant. In addition, the prior distribution for σ^2_i is $IG(6,\ 0.001)$.

Finally, the conditional distribution of the factors given the data and the parameters is derived as in Otrok and Whiteman (1998). In particular, we compute the joint density for the observable data and the latent factors given the parameters as the product of NMK independent Gaussian densities. Then, we use this joint distribution to obtain the conditional distribution of the factors given the data and the parameters. Next section describes the process to measure the synchronization among the factors.

4.2 Synchronization between factors

This section describes the procedure followed to evaluate the potential variations in the cyclical interdependencies between the EA factor and each of the country-specific factors. Using the bivariate Markov-switching model proposed by Leiva-Leon (2017), we obtain a full characterization of the regime inferences and inferences on the type of synchronicity that the Euro Area factor and the specific-country factors bear at each period of time. Following the previous notation, let $f_{k,t}$ be the unobservable dynamic factors that describe the macroeconomic co-movements among the countries included in the panel of cross-country data.

When index $k = EA$, the factor describes the evolution of the EA aggregate economic activity, while it represents the country-specific economic activity when $k = n$, where $n = 1, ..., 18$. Therefore, $f_{k,t}$ can be modeled using a MS model as proposed by Hamilton (1989), where the dynamics of the factors depends on an unobserved state variable ($S_{k,t}$) that controls the regime changes, an idiosyncratic component, $\epsilon_{k,t}$, and

a set of model parameters, Θ_k, where $k = EA, n$. Each of the state variables, $S_{EA,t}$ and $S_{n,t}$ evolve according to an irreducible two-state Markov chain, whose transition probabilities are given by:

$$\Pr\left(S_{k,t} = j \mid S_{k,t-1} = i\right) = p_{k,ij}. \tag{4.6}$$

To compute inference on the interactions between the two state variables, we adopt the following bivariate two-state Markov-switching specification:

$$\begin{bmatrix} f_{EA,t} \\ f_{n,t} \end{bmatrix} = \begin{bmatrix} \mu_{EA,0} + \mu_{EA,1} S_{EA,t} \\ \mu_{n,0} + \mu_{n,1} S_{n,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{EA,t} \\ \epsilon_{n,t} \end{bmatrix}, \tag{4.7}$$

$$\begin{bmatrix} \epsilon_{EA,t} \\ \epsilon_{n,t} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{EA}^2 & \sigma_{EA,n} \\ \sigma_{EA,n} & \sigma_n^2 \end{bmatrix} \right), \tag{4.8}$$

for $i, j = 0, 1$ and $k = EA, n$. If the state variable $S_{k,t} = 0, f_{k,t}$ is in regime 0 with mean equals to $\mu_{k,0}$, while if $S_{k,t} = 1$, $f_{k,t}$ is in regime 1 with mean equals to $\mu_{k,0} + \mu_{k,1}$. If we assume $\mu_{k,1} > 0$, the latent variable $S_{k,t}$ identifies periods of low and high economic performance, which are interpreted as recessions and expansions, respectively. Phillips (1991) was pioneering in evaluating the transmission of business cycles between two countries in the context of bivariate Markov-switching processes. Although there are two possible states for each separate country, modeling the interactions would require a new unobservable state variable $S_{EA,n,t}$that encompasses the four different combinations: $S_{EA,n,t} = 1$ when $\left(S_{EA,t} = 0, S_{n,t} = 0\right), S_{EA,n,t} = 2$ when $\left(S_{EA,t} = 1, S_{n,t} = 0\right), S_{EA,n,t} = 3$ when $\left(S_{EA,t} = 0, S_{n,t} = 1\right)$, and $S_{EA,n,t} = 4$ when $(S_{EA,t} = 1, S_{n,t} = 1)$.

Regarding the case of business cycle interdependence between the business cycles of two countries, Phillips (1991) describes two extreme cases. The first case characterizes pairs of countries for which their individual business cycle fluctuations are completely independent. In this case, the separate regime-shifting processes, $S_{EA,t}$ and $S_{n,t}$ are independent and

$$\Pr\left(S_{EA,t} = j_{EA}, S_{n,t} = j_n\right) = \Pr\left(S_{EA,t} = j_{EA}\right) \Pr\left(S_{n,t} = j_n\right), \tag{4.9}$$

where $j_{EA} = 0, 1$ and $j_n = 0, 1$. In the opposite case of perfect synchronization (or dependence), both countries share the state of the business cycle and the probabilities of $S_{EA,n,t}$ are in fact those of one of the countries, implying that $S_{EA,t} = S_{n,t} = S_t$ and that

$$\Pr\left(S_{EA,t} = j_{EA}, S_{n,t} = j_n\right) = \Pr\left(S_t = j\right), \tag{4.10}$$

where $j = 0, 1$. Bengoechea et al. (2006) proposed a new framework to measure the degree of business cycle correlation between EA and country n. These authors realized that independence and perfect synchronization are two extreme possibilities that never occur in practice. For two countries, the actual probabilities will be a linear combination of these two extremes:

$$\Pr\left(S_{EA,t} = j_{EA}, S_{n,t} = j_n\right) = \delta \Pr\left(S_t = j\right) + (1 - \delta) \Pr\left(S_{EA,t} = j_{EA}\right) \Pr\left(S_{n,t} = j_n\right). \tag{4.11}$$

Then, δ which is estimated by the data, measures the degree of overall pairwise business cycle synchronization.

Leiva-Leon (2017) went one step further, although at the cost of complicating the approach a bit. The contribution of this author was summarizing the information about the relationship of dependency between the two separate latent variables, by defining another latent variable $V_{EA,n,t}$ that governs the transition between the two extreme cases, independent cycles and perfect synchronization. This latent variable $V_{EA,n,t}$ is equal to 1 if the business cycle phases are in a fully synchronized regime at time t, and 0 otherwise. To complete the statistical characterization of the model, $V_{EA,n,t}$ is also assumed to evolve according to a two-state Markov chain with transition probabilities given by

$$\Pr\left(V_{EA,n,t} = j_v \mid V_{EA,n,t-1} = i_v\right) = p_{v,ij}, \quad \text{for } i_v, j_v = 0, 1 \tag{4.12}$$

The potential regimes of the model implies that the four cases of the regime-switching variable $S_{EA,n,t}$ could appear either when $V_{EA,n,t} = 1$ or when $V_{EA,n,t} = 0$. The resulting eight different states are summarized by the latent variable $S^*_{EA,nt}$ for each period of time t. In particular, the eight different regimes are

$$S^*_{EA,n,t} = \left\{ \begin{array}{llll} 1, & \text{if } S_{EA,t} = 0, & S_{n,t} = 0, & V_{EA,n,t} = 0 \\ 2, & \text{if } S_{EA,t} = 0, & S_{n,t} = 1, & V_{EA,n,t} = 0 \\ 3, & \text{if } S_{EA,t} = 1, & S_{n,t} = 0, & V_{EA,n,t} = 0 \\ 4, & \text{if } S_{EA,t} = 1, & S_{n,t} = 1, & V_{EA,n,t} = 0 \\ 5, & \text{if } S_{EA,t} = 0, & S_{n,t} = 0, & V_{EA,n,t} = 1 \\ 6, & \text{if } S_{EA,t} = 0, & S_{n,t} = 1, & V_{EA,n,t} = 1 \\ 7, & \text{if } S_{EA,t} = 1, & S_{n,t} = 0, & V_{EA,n,t} = 1 \\ 8, & \text{if } S_{EA,t} = 1, S_{n,t} = 1, V_{EA,n,t} = 1 & & \end{array} \right\}. \tag{4.13}$$

Finally, the joint dynamic of $S_{EA,t}$ and $S_{n,t}$ is described by a weighted average between the fully synchronized and fully independent scenarios as follows:

$$\begin{aligned}\Pr\left(S_{EA,t}=j_{EA},S_{n,t}=j_n\right)&=\Pr\left(V_{EA,n,t}=1\right)\Pr\left(S_t=j\right)\\&+\left(1-\Pr\left(V_{EA,n,t}=1\right)\right)\Pr\left(S_{EA,t}=j_{EA}\right)\Pr\left(S_{n,t}=j_n\right),\end{aligned}\tag{4.14}$$

where $\Pr\left(V_{EA,n,t}=1\right)=\delta_t$ measures the dynamic synchronicity between $S_{EA,t}$ and $S_{n,t}$ and determines the weights attributed to each scenario. Leiva-Leon (2017) describes a Bayesian method to estimate the model parameters and to compute inferences on the state variables. Therefore, in our empirical application, δ_t quantifies the time-varying degree of synchronization between the business cycle of the EA and the particular macroeconomic fluctuations in each country included in the dataset along the sample period.

4.3 Empirical results

Next sections describe the selected data, the results regarding the estimation of the factors as a summary of the aggregate economic activity, and the characterization of the evolution of the business cycle synchronization among the EA members.

The dataset is composed by several macroeconomic indicators for all the EA members. As suggested by Kose et al. (2003) and Crucini et al. (2011), we select macroeconomic series of production, consumption and investment for each country. In particular, we use the demeaned growth rates of GDP, Household & NPISH Final Consumption Expenditure, and Gross Fixed Capital Formation. The seasonally adjusted series were downloaded from the Eurostat database at a quarterly frequency. Data availability differs for each of the EA members. Thus, to consider a balanced dataset, our effective sample spans the period between the first quarter of 2000 to the last quarter of 2015 for all the nineteen countries of the EA but Cyprus. The database from Eurostat has only a few observations available for this country. Hence, our dataset is composed by three economic indicators from 18 different countries and 64 quarterly observations, which cover the last 16 years since the introduction of the euro as a single currency in 1999.

4.3.1 Aggregate economic activity

Macroeconomic co-movements along the eighteen countries in the dataset are estimated through a common EA factor and the particular co-movements within each country through the country-specific factors. Figure 4.1 presents the results of the estimation of the latent factors. For the sake of space, and due to the large amount of countries composing the EA, the figure only depicts some illustrative examples. In particular, the

figure represents the factors for the EA, Spain and Italy, along with 33 and 66-percent quantile bands (doted lines); these tight bands show that the factors are accurately estimated. The upper graph describes the evolution of the EA factor together with the periods defined by the Euro Area Business Cycle Dating Committee of the Center for Economic Policy Research (CEPR) as recessions (shaded areas).

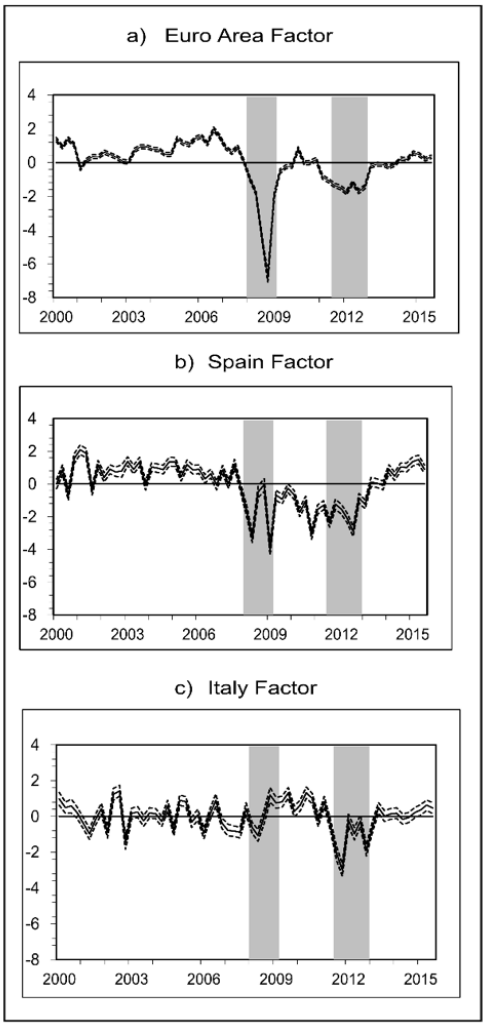

Figure 4.1: The estimated Euro Area (a), Spain (b) and Italy (c) factors.

The figure shows that the EA factor is able to track the economic evolution of the EA according to the dating of the CEPR. Overall, the factor takes negative values during the two recession periods between 2008.Q1-2009.Q2 and 2011.Q3-2013.Q1 and positive values elsewhere. The estimates of the EA factor suggest that the downturn at the beginning of 2008 was much severe than the recession that followed the European sovereign debt crisis of 2011 and measures the economic recovery between the two recessions.

However, the factor in Spain (middle graph) shows how the Spanish economy kept a low economic performance between the two recessions and starts an expansion after the second quarter of 2013. On the other hand, the factor in Italy (lower graph) shows that this country was more affected by the sovereign debt crisis during the second recession. These results provide an illustrative example of the different reactions of the EA members to the economic events that affected Europe after 2008. Regarding others country-specific factors not included in Figure 4.1, some countries recover earlier than the Spanish economy from the first recession and show a stable improvement during the following years, not being affected by the second recession severely. These results are omitted to save space. This is the case of Belgium, Estonia, France, Germany, Latvia, Lithuania, Slovenia and Slovakia factors. In the case of Greece, the estimated factor shows that the Greek economy had a lower economic performance between the two recessions as in the case of Spain. Finland and Portugal show a pattern similar to Italy, being more affected by the second crisis. Finally, the factors corresponding to Luxembourg, Ireland, Malta, Austria and the Netherlands remained relatively stable during that period.

4.3.2 Business cycle synchronization

Once the latent factors are estimated, they are included in the Markov-switching specification described in Section 2. Figure 4.2 depicts the comparison of the regime switches of the EA factor with those corresponding to the Spain factor. The upper and middle graphs represent the MS filtered probabilities of regime switches in the EA and Spain factors respectively. These estimated probabilities take values close to one during the CEPR recession periods. Hence, they are interpreted as an accurate characterization of the business cycle phases estimated with the information included in a large panel of cross-country data. The filtered probabilities for the EA factor show a decrease between the two recession periods. However, given that the Spanish economy showed a worse economic performance at that dates (see Figure 4.1), its recession probabilities are higher during that period and remain closer to one.

The lower graph depicts the probability of synchronization in the business cycle phase changes of the EA and Spain factors. This probability shows high values and a slight positive trend during the first part of the sample since the introduction of the euro. In the period between the two recessions, synchronization decreases drastically due to the poorer economic performance of Spain with respect to the EA. Then, the degree of synchronization rises again as the EA economy enters into the second recession. After that, synchronization shows a slight decrease since the middle of the second recession given that Spain needed more time to begin the economic recovery. Finally,

synchronization increases again once Spain starts to recover in the second quarter of 2013 and reaches values like those observed at the beginning of the sample.

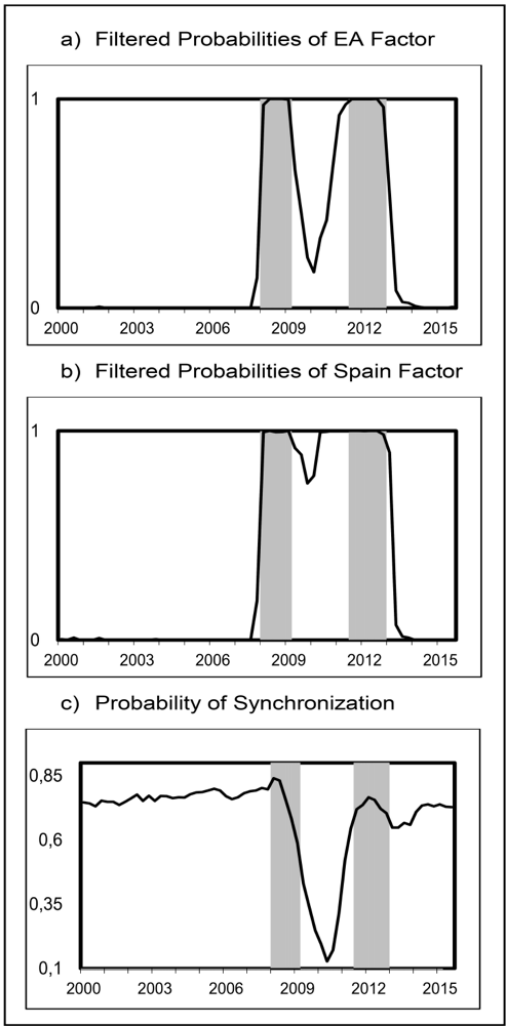

Figure 4.2: Filtered probabilities of regime switches for the EA factor (a) and the Spain factor (b). Probability of synchronization between the regime changes of EA and Spain factors (c).

Figure 4.3 shows the estimated pairwise synchronization of the EA economic activity and Germany, Italy, France and Portugal. Germany shows a high level of synchronization along the sample period. According to this figure, the German business cycles become less synchronized after the financial shock and during the second CEPR recession. However, those drops take place with a delay with respect to the beginning of the CEPR recessions. These facts highlight the robust connection of the German economy to the EA fluctuations, especially during the first stages of the financial and the debt crisis. Italy, France and Portugal present a lower level of synchronization during the pre-recessions period at the beginning of the sample. In these three countries, synchronization drops after the financial crisis and shows a significant increase after 2009. In the case of France and Portugal, the high synchronization is only slightly reduced during the second recession between 2011 and 2013. Notably, the level of synchronization of these countries shows an improvement at the end of the sample with respect to the period between the introduction of the euro and the first recession.

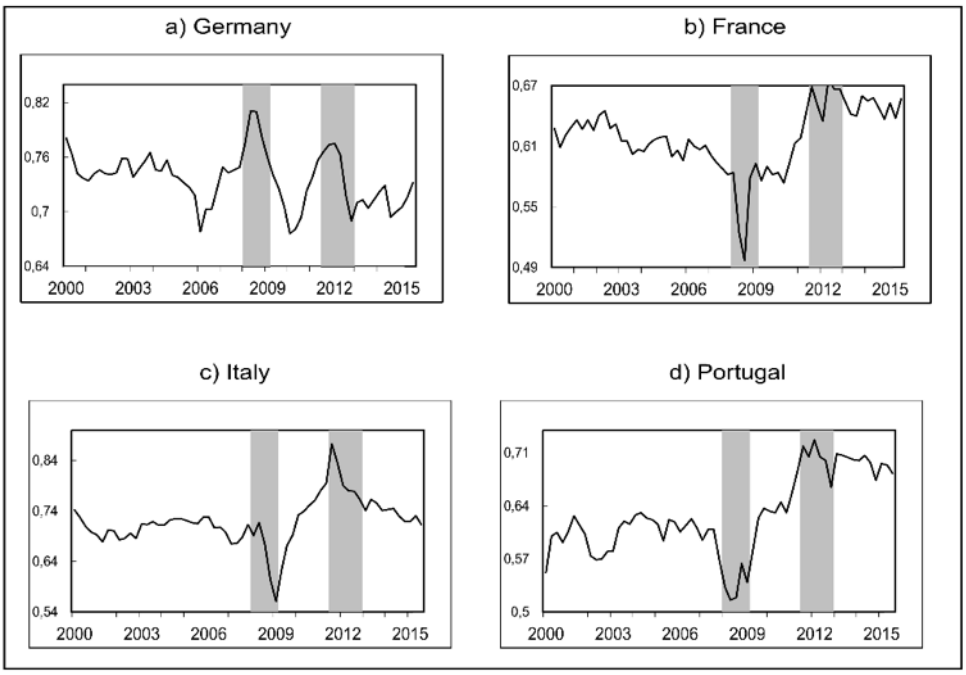

Figure 4.3: Probabilities of synchronization between the regime switches of the Euro Area and the Germany (a), the France (b), the Italy (c), and the Portugal (d) factors.

Figure 4.4 describes the evolution of the synchronization of EA and Greece, Austria, Finland, Slovakia, and Slovenia. As in Figure 4.3, there is a fall in synchronization that took place close to the financial crisis. The lack of synchronization in Greece, Finland and Austria started earlier while macroeconomic fluctuations in Slovakia and Slovenia decoupled once the recession had started. However, the effects of the recessions were more persistent in the second set of countries. Synchronization remains at low values for a longer time after 2008 and does not improve before the second recession (Finland and Slovenia) or falls again after 2011 (Greece, Austria and Slovakia). Furthermore, these countries show higher levels of synchronization before the recessions than at the end of the sample, suggesting that they will require more time to reconnect their macroeconomic behavior with the EA fluctuations.

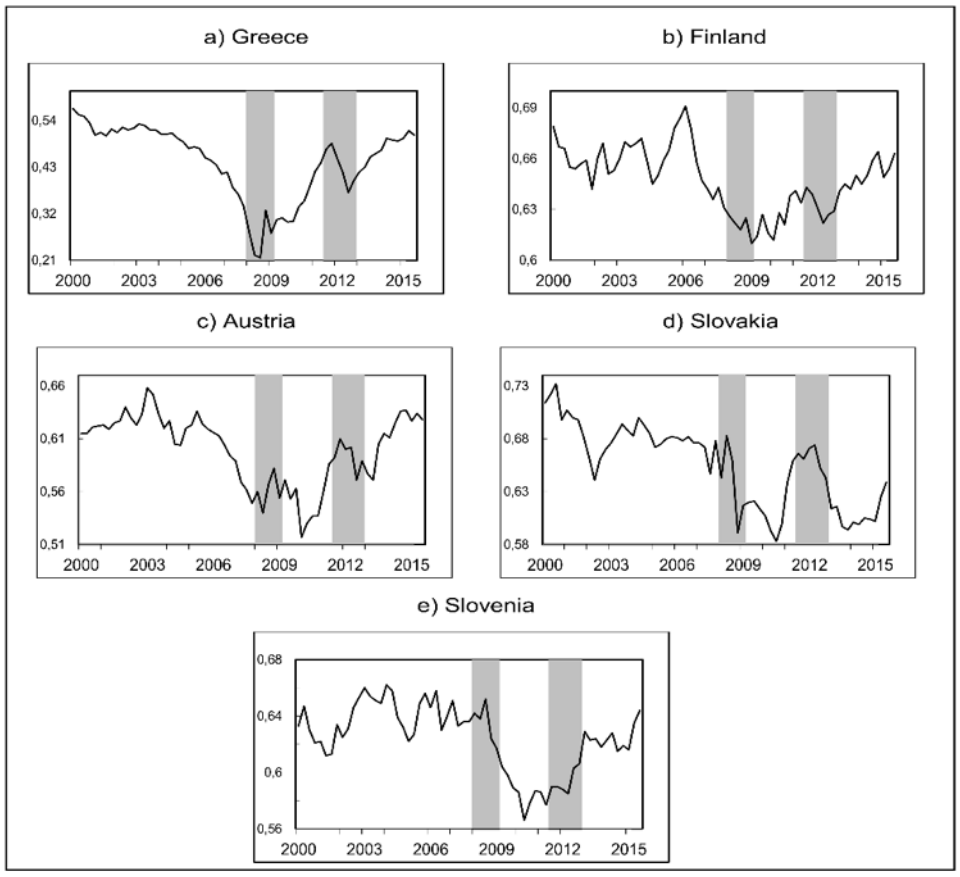

Figure 4.4: Probabilities of synchronization between the regime switches of the Euro Area and the Greece (a), the Finland (b), the Austria (c), the Slovakia (d), and the Slovenia (e) factors.

The results for Ireland, the Netherlands, Latvia and Luxembourg are presented in Figure 4.5. In contrast to the analysis of the EA members showed by Figures 4.3 and 4.4, this set of countries shows a progressive decrease of synchronization and reaches the maximum level of desynchronization during the 2011-2013 recession. This decline starts earlier in the case of Ireland and the Netherlands while it is more abrupt in Latvia and Luxembourg. Although Ireland, the Netherlands and Latvia show an increasing trend in their estimates at the end of the sample, only Luxembourg reaches a degree of synchronization as high as the one observed before the recessions.

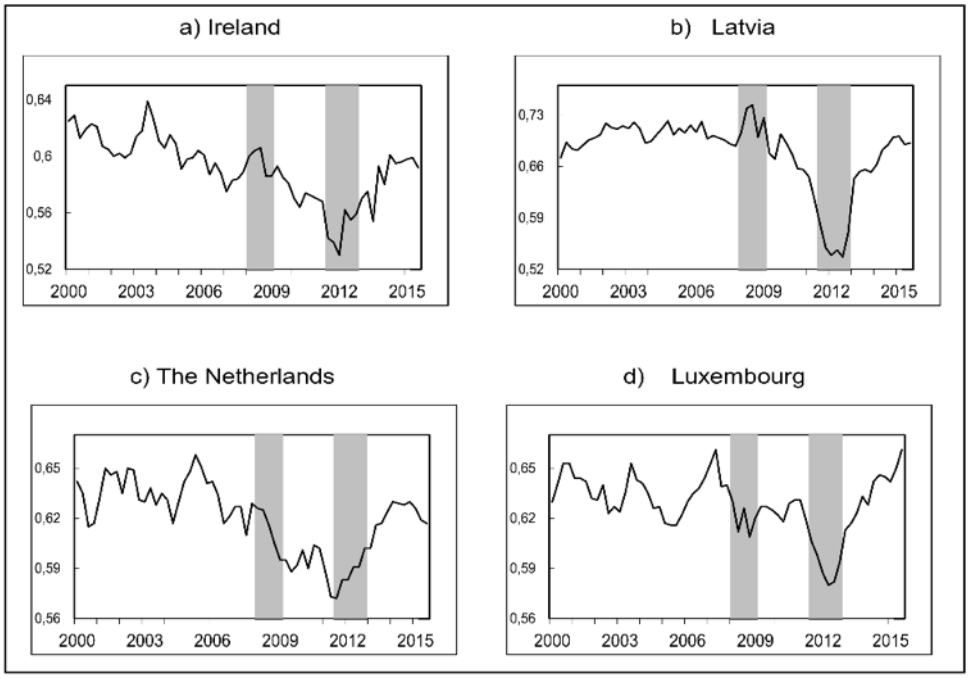

Figure 4.5: Probabilities of synchronization between the regime switches of the Euro Area and the Ireland (a), the Latvia (b), the Netherlands (c), and the Luxembourg (d) factors.

Finally, Figure 4.6 incorporates the results corresponding to Belgium, Estonia, Malta and Lithuania. In the case of Malta, the figure depicts an increasing trend in its synchronization since the beginning of the sample until the start of the first recession. After that date, it decreases constantly and reaches a minimum in 2013. Once the second recession ended, synchronization improves until the end of the sample. As in the analysis of the countries included in Figure 4.3, Belgium, Estonia and Lithuania suffer from a drastic decrease in their synchronization during the first recession, while it recovers quickly after this period. In particular, Belgium and Estonia react very fast in terms of synchronization after the financial crisis and required a short period to recover from that shock. Instead, Lithuania desynchronizes some quarters later after the start of the

first recession although, thereafter, its synchronization rises sharply before the second recession. Nevertheless, the main difference between these countries and those depicted in Figure 4.3 is that none of them shows an increase in its degree of synchronization after the recessions with respect to the beginning of the sample.

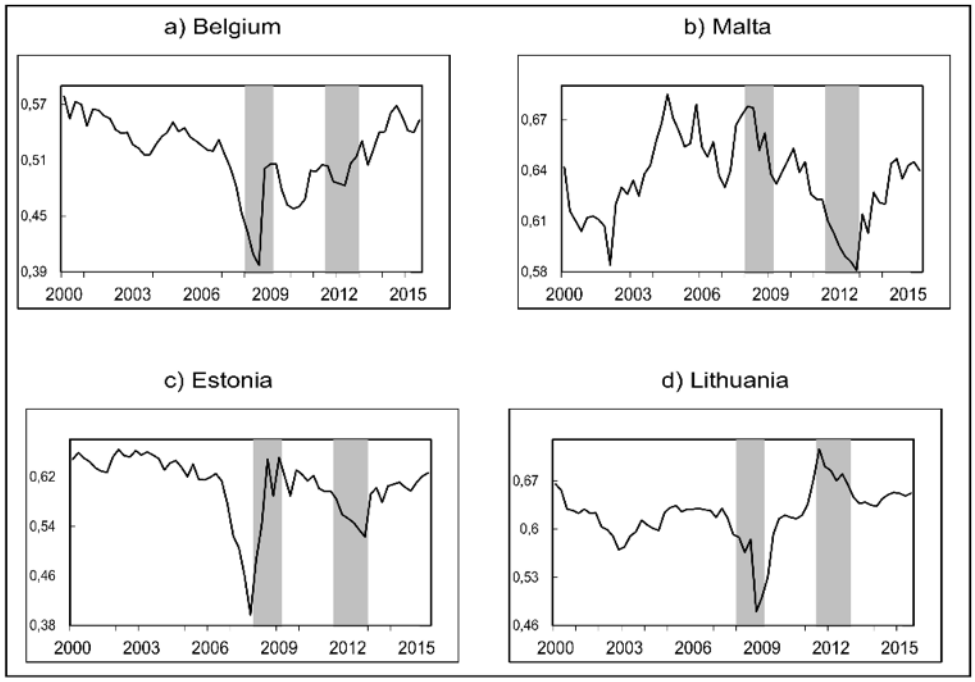

Figure 4.6: Probabilities of synchronization between the regime switches of the Euro Area and the Belgium (a), the Malta (b), the Estonia (c), and the Lithuania (d) factors.

4.4 Concluding remarks

The global effects of the financial crisis and its resulting consequences on government debt levels forced central bankers to implement unconventional monetary stimuli to avoid an economic collapse. In the case of EA, this was particularly challenging because a single and highly expansionary monetary policy was applied to a large set of countries in different phases of their respective business cycles. This chapter evaluates the consequences of these historic events on the evolution of the business cycle synchronization among all the members of the EA using a large panel of cross-country data. The analysis focuses on combining the dimension reduction properties of DFMCSs with a MS specification that provides a time-varying measure of synchronization for each observation along the sample. Our results show that, although the countries exhibit an overall decline in the synchronization in the financial crisis, they recover the levels of

synchronization that characterized the pre-recessions period. However, we also find that there are notable differences in the magnitude of the fall in synchronization and in the period of time required to recover the pre-recessions synchronization levels. Hence, the findings provided here support the presence of a *Two-Speed Europe* after the financial crisis in terms of economic synchronization. Countries as Germany, France, Italy, Spain, Portugal, Belgium, Estonia and Lithuania recover quickly their level of synchronization after the first recession and some of them even improve it with respect to the period before 2008. On the contrary, Ireland, the Netherlands, Luxembourg, Latvia and Malta suffered a larger desynchronization after 2011 and show a slower recovery. Greece, Austria, Slovakia, Slovenia and Finland keep a decrease on their levels of synchronization during a larger period between the two recessions. Therefore, we fail to find evidence suggesting that the recent implementation of the unconventional monetary stimuli applied by the ECB amplified the desynchronization of the EA members. By contrast, these countries show an increasing degree of synchronization after 2013, some of which reached an improvement with respect to their pre-recession synchronization levels.

Chapter 5

DFM with unknown cluster structure: An application to energy prices

Energy economics and energy policy are two important concerns for the development of countries. Governments and investors have to pay close attention to international energy prices given that they affect not only the competitiveness of industry yet households. Recently, the intervention of governments in the regulation of energy is gaining even more importance, since the environment is suffering the consequences of the improper use of energy fuels.

We already know that there exist differences in the energy market structures across countries due to production and transportion costs, trade restrictions, and contractual terms, between others factors. Nevertheless, after taking into account these differences, in this chapter we are interested in the existence of common factors that describe a global behavior in the international energy market, together with group-specific factors explaining energy prices related to regions, countries or industrial sectors.

After three decades since the pioneering work of Griffin (1980) about energy economics and policy, research in this topic has increased considerably according to the economic events. The lack of data about international energy prices has been an obstacle for researchers, nevertheless, the establishment of statistical agencies and developments in data storage and the internet have provided new tools which are capable to analyze big data sets about energy. Previous works have analyzed energy prices paying attention to individual fuel types, for example Brown et al. (2008) studies what affects natural gas and crude oil prices in USA, and Nick and Thoenes (2014) pays attention to natural gas prices in Germany. International oil prices have been considered by Van Benthem

and Romani (2009) and Aastveit et al. (2015) together with domestic end-use energy and demand, respectively. How energy prices relate to energy comsumption and investment was examined in Mahadevan and Asafu-Adjaye (2007) and Bretschger (2015) using the consumer price index and energy use as proxies for energy price. Only a few of these articles consider different fuel types together in their analysis of energy prices, and in such cases they are used as control variables.

Up to our knowledge, we are pioneers in analyzing the co-movements of international energy prices in a bigdata scenario of 30 countries and 12 industrial sectors. The data set is from Sato et al. (2019). We consider for the analysis the Dynamic Factor Model with Cluster Structure (DFMCS) which allows us to investigate if there exists a group structure between international energy prices, to characterize the heterogeneity of the global energy market based on industry, country or region, to quantify the extent to which "crisis" affected the global energy prices, and to identify the sources that explain the cross-section variations in energy prices through control variables which are country-specific.

We extend the methodology proposed in Alonso et al. (2020) in order to study the effect of control variables, which are country specific, over energy prices. We also run a Monte Carlo simulation to evaluate the performance, in finite samples, of Alonso et al. (2020) clustering procedure when we take into account control variables. Results provide useful interpretations about the existence of trading groups of countries in the global energy market.

The rest of the chapter is organized as follows. Section 1 introduces the model. The estimation method and the Monte Carlo experiment are described in Section 2. Section 3 presents the data and Section 4 contains the empirical results. Finally, some concluding remarks and future extensions are given in Section 5.

5.1 Theoretical framework

In this section we introduce an extension of the DFMCS framework to analyze what drives energy prices for 12 industrial sectors and over 30 OECD and non-OECD countries. It consists in including in the DFMCS an 'observed' component of macroeconomic variables which are country-dependent.

Let $t = 1, ..., T$ represents the time index and $i = 1, ..., N$ the cross-section index, the unknown and fixed number of groups is S and $G = g_1, ..., g_N$ is the group membership with $g_i \in 1, ..., S$. The number of countries is Q and $C = c_1, ..., c_N$ represent the country membership with $c_i \in 1, ..., Q$. There are N_j units within group j ($j = 1, ..., S$) such that $N = \sum_{j=1}^{S} N_j$. The response variable of the ith unit, observed at time t, y_{it}, is defined as

$$y_{it} = x'_{it}\beta_{c_i} + f'_{0,t}\lambda_{0,i} + f'_{g_i,t}\lambda_{g_i,i} + \varepsilon_{it}. \tag{5.1}$$

where x_{it} is a $p \times 1$ vector of observable variables, $f'_{0,t}$ is a $r_0 \times 1$ vector of unobserved global factors affecting all the N series in the sample, $f'_{g_i,t}$ is a $r_j \times 1$ vector of unobserved group-specific factors that affect only the units in group g_i, β_{c_i} is a $p \times 1$ vector of unknown regression coefficients for country q $(q = 1, ..., Q)$, $\lambda_{0,i}$ and $\lambda_{g_i,i}$ are the corresponding factor loadings, and $\varepsilon_{i,t}$ is the unit specific error. Here β_{c_i} is common for all i units belonging to country q.

In line with Wang (2008) identifying conditions for large dimensional factor models, we consider the following assumptions to identify the model and make possible the estimation. Let Λ_0 and Λ_j be the corresponding matrix of factor loadings, then we assume that $\Lambda'_0\Lambda_0 = \mathbf{I}_{r_0}$, $\Lambda'_j\Lambda_j = \mathbf{I}_{r_j}$ for $j = 1, ..., S$, the $r = \sum_{j=0}^{S} r_j$ covariance matrices of the factors are diagonal, $\Lambda'_0\Lambda_j = \mathbf{0}_{r_0 \times r_j}$ and $\Lambda'_j\Lambda_i = \mathbf{0}_{r_j \times r_i}$ for $j \neq i$.

5.2 Estimation method

The model is estimated using the procedure introduced in Alonso et al. (2020). This method works well and seems to be better than the one in (Ando and Bai, 2017) in terms of estimation of factors and loadings, as shown by their Monte Carlo simulation results.

Given the DFMCS in (5.1), we just have information about the left hand side and the control variables included in matrix X. The unknown parameters to be estimated are the number of groups, S, the number of common and group-specific factors, r_0 and r_j, the corresponding factors and its loadings, $\widehat{F}_0, \widehat{F}_1, ..., \widehat{F}_S, \widehat{\Lambda}_0, \widehat{\Lambda}_1, ..., \widehat{\Lambda}_S$, the membership of each variable to a given group, g_i, and the sensitivity to observable factors, β_{c_i}. We need to modify Alonso et al. (2020) procedure, AGP in what follows, because they do not consider exogenous variables and we want to take into account the effect of macroeconomic variables, which are country-specific, over energy prices.

AGP includes a prior step in which the observed time series, y_{it}, are cleaned from additive outliers and level shifts, see Alonso et al. (2020) for more details about the cleaning procedure. We extend AGP to consider the estimation of the regression coefficients, β_{c_i} and divide the estimation of model (5.1) in five steps:

1. Estimate the country-dependent regression coefficients, β_{c_i}, by minimizing

$$L(\beta_{c_i}) = \sum_{i=1}^{N} \parallel y_i - X_i\beta_{c_i} \parallel^2 + T\sum_{i=1}^{N} p_i(\beta_{c_i}). \tag{5.2}$$

where

$$p_i(\beta_{c_i}) \equiv p_{\kappa_{c_i},\gamma}(\beta_{c_i}) = \sum_{c_i=1}^{p_{c_i}} p_{\kappa_{c_i},\gamma}(\mid \beta_{c_i} \mid),$$

is the SCAD (*smooth clipped absolute deviation*) penalty of (Fan and Li, 2001) as suggested in Ando and Bai (2017), with $\kappa_{c_i} > 0$ and $\gamma = 3.7$, which minimize a Bayesian risk criteria for the regression coefficients. The size of the penalty which is control by κ_{c_i}, for $c_i \in 1, ..., Q$, is the same for those cross-sectional units related to country q with $q = 1, ..., Q$ in the sample. Given $\widehat{\beta}_{c_i}$, we subtract the corresponding 'observed' component from each time series, y_i, obtaining

$$y_i^* = y_i - X_i\widehat{\beta}_{c_i}.$$

2. Estimate an initial set of global factors F_0 and their corresponding loadings Λ_0. The number of global factors r_0 is obtained using the test proposed in Caro and Peña (2020) and introduced in Chapter 2. The factors are estimated by $\widehat{f}_{0t} = \widehat{\lambda}_0 y_t^*$, and the common component by $c_t = \widehat{\Lambda}_0\widehat{\Lambda}_0' y_t^*$, where $\widehat{\Lambda}_0$ is the estimated matrix of factor loadings which columns are the eigenvectors of the *combined dynamic correlation matrix* of the observed data, $\mathbf{R}_{k_0}$, introduced in Chapter 2, associated to the r_0 largest eigenvalues.

3. Apply the clustering algorithm proposed in Alonso and Peña (2019), based on similar linear dependency measures between time series, to the estimated common component $\widehat{\Lambda}_0\widehat{f}_{0,t}$. Once the optimal number of groups S is calculated using a modification of the Silhouette algorithm proposed by Rousseeuw (1987), the memberships g_i for $i = 1, ..., N$ are obtained.

4. Obtain the number of group-specific factors r_j for $j = 1, .., S$ using the CP test taking into acount the time series $y_{i,t}$ belonging to each group. The corresponding factors $F_1, ..., F_S$ and their loadings $\Lambda_1, ..., \Lambda_S$ are estimated as describe in step (2) for the global factors. In this step, all the global and group-specific factors from steps 2 and 4 are compared and classified according to the decision rules in Alonso et al. (2020) based on empirical canonical correlations.

5. As suggested in Alonso et al. (2020), it must be verified that the groups are a consequence of specific factors and not due to different loadings corresponding to global factors. Therefore, group-specific residuals $v_t = y_t^* - \widehat{\Lambda}_0\widehat{f}_{0,t}$ are obtained and used to re-estimate the group-specific factors. Finally, each group must be analyzed to check the existence of at least one specific factor.

Our objective in the rest of the chapter is, by means of a Monte Carlo simulation, to evaluate the clustering performance of AGP under different data generation processes, and to applied the proposed extention to the analysis about international energy prices.

5.2.1 Monte Carlo experiment

In this section we simulate a data structure similar to the one in our data set about energy prices and evaluate the clustering performance of AGP procedure. We consider three data-generating processes (DGP) and we set the number of countries in each DGP to be three, such that $c_i \in 1, 2, 3$. Each variable y_i for $i = 1, ..., N$ is generated as

$$y_i = X_i\beta_{c_i} + F_0\lambda_{0,i} + F_{g_i}\lambda_{g_i,i} + \varepsilon_i, \tag{5.3}$$

where the r-dimensional global common factor $f_{0,t}$ is a vector of $U(0, 1)$ variables and the corresponding elements of the loading matrix Λ_0 follow $U(-2, 2)$ distribution, the r_j-dimensional group-specific factor $f_{g_i,t}$ $(j = 1, ..., S)$, is a vector of $N(0, 1)$ variables, and each element of the factor loading matrix Λ_j is generated from the $N(0.5j, 1)$ distribution.

The number of columns of X_i is set to $p = 30$, while the true number of predictors is $q = 3$. Each of the elements of X_i is generated from $N(0, 1)$ distribution. The non-zero true parameter values of β_{c_i} are set to be $(4, 3.5, 3)$ for country 1, $(-2.5, -2, -2.5)$ for country 2, and $(1, 0.5, 1.5)$ for country 3. These non-zero elements are put into the first three elements of β_{c_i}, for example, the true parameter vector is $\beta_1 = (4, 3.5, 3, 0, 0, ..., 0)'$ for country 1. We set the number of groups $S = 3$. We assume that each country has 99 series, the first 33 series of each country belong to group 1, the second 33 series, from 34 to 66, of each country belong to group 2, and the last 33 series, from 67 to 99, of each country belong to group 3. We consider the sample sizes $N = \{297, 594\}$ and the number of time observations $T = \{100, 200\}$.

First DGP assumes that the N-dimensional vector ε_t has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma_e^2\mathbf{I}_N$. Second DGP has non-homoscedastic errors with cross-sectional dependence such that $\varepsilon_{it} = 0.2e_{it}^1 + \delta_t e_{it}^2$, where $\delta_t = 1$ if t is odd and zero if t is even, and the N-dimensional vectors $e_t^1 = (e_{1t}^1, ..., e_{Nt}^1)'$ and $e_t^2 = (e_{1t}^2, ..., e_{Nt}^2)'$ follow multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$ with $s_{ij} = 0.3^{|i-j|}\sigma_e^2$ and e_t^1 and e_t^2 are independent. The third DGP presents errors which are serial and cross-sectional correlated such that, $\varepsilon_{it} = 0.2\varepsilon_{i,t-1} + e_{it}$, where $t = 1, ..., T$, the N-dimensional vector $e_t = (e_{1t}, ..., e_{NT})'$ follows multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$ with $s_{ij} = 0.3^{|i-j|}\sigma_e^2$. We consider the noise variances $\sigma_e^2 = 1, 2$.

We generate 100 replications using each of the three data-generating process. In each replication the proposed procedure is applied to the simulated data in order to select, simultaneously, the number of groups, the number of global common factors, the number of group-specific pervasive factors and the size of the regularization parameter. We set the possible numbers of group-specific and global factors to range from zero to eight. The number of groups ranges from two to twelve. Possible candidates for the regularization parameter κ_i are $\kappa_i = 10^{-3+0.25k}$ for $k = 0, ..., 12$.

Table 5.1 shows the mean of the selected number of clusters for the 100 iterations, and below each mean the total number of iterations out of 100 where the true number was selected.

Table 5.1: Mean of the selected number of clusters (first row) and number of iterations out of 100 where the true number of clusters was selected (second row).

	DGP_1		DGP_2		DGP_3	
(T, N)	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 1$	$\sigma^2 = 2$
(100, 297)	2.96	2.08	2.98	2.89	2.92	2.51
	97	29	98	89	93	62
(100, 594)	3.00	2.17	3.00	2.90	3.00	2.72
	100	25	100	91	100	72
(200, 297)	3.00	2.58	3.01	2.99	3.00	2.95
	100	60	99	99	100	95
(200, 594)	3.00	2.88	3.00	3.00	3.00	3.00
	100	88	100	100	100	100

We see how AGP tends to underestimate the number of groups, S. The method suffers with the increase of variance under DGP_1, over all when $T = 100$. Under DGP_2 and DGP_3, where errors are allowed to present serial and limited cross-section correlation, AGP provides very accurate results.

The similarity between the original data clustering and the estimated one is measured by the Adjusted Rand Index in Hubert and Arabie (1985) using the Permutation Model,

$$ARI(C, C') = \frac{\sum_{i=1}^{S}\sum_{j=1}^{S'}\binom{\#(C_i \cap C'_j)}{2} - \sum_{i=1}^{S}\binom{\#(C_i)}{2}\sum_{j=1}^{S'}\binom{\#(C'_j)}{2} / \binom{n}{2}}{\left(\sum_{i=1}^{S}\binom{\#(C_i)}{2} + \sum_{j=1}^{S'}\binom{\#(C'_j)}{2}\right)/2 - \sum_{i=1}^{S}\binom{\#(C_i)}{2}\sum_{j=1}^{S'}\binom{\#(C'_j)}{2} / \binom{n}{2}},$$

where C is the original partition with S groups, and C' the estimated one with S' groups, and it represents the probability that C and C' will agree on a randomly chosen pair. Table 5.2 reports the mean of the Adjusted Rand Index (ARI) for the 100 iterations. The closer the index is to 1 the better the agreement between the real partition and the estimated one.

Table 5.2: Clustering performance evaluation using the Adjusted Rand Index.

	DGP_1		DGP_2		DGP_3	
(T, N)	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 1$	$\sigma^2 = 2$
(100, 297)	0.8392	0.3195	0.8690	0.7511	0.8269	0.6399
(100, 594)	0.9248	0.3140	0.9172	0.8898	0.9372	0.7387
(200, 297)	0.9251	0.5508	0.9144	0.9027	0.9074	0.8658
(200, 594)	0.9792	0.8887	0.9738	0.9204	0.9393	0.9727

In general, AGP procedure shows pretty good allocation performance. Nevertheless, as we mention before for Table 5.1, AGP is very sensitive to the increase of variance under homoscedastic errors and time dimension $T = 100$.

5.3 Data

We analyze the Energy Price Index with fixed weights (FEPI) constructed in Sato et al. (2019). The data set includes 30 OECD and non-OECD countries and 12 sectors for the time period 1995-2015. Countries and sectors included in the sample are listed in Table 5.3. The cross-section dimension is $N = 30 \times 12 = 360$ and the time-series dimension is $T = 21$ years. Previous to the analysis the data are first differenced in order to achieve stationarity. Using subscripts $i = 1, ..., 30$ for the country, $s = 1, ..., 12$ for the sector and $t = 1, ..., 21$ for time, the $FEPI_{ist}$ is defined as

$$FEPI_{ist} = \sum_j w_{is}^j \cdot log(P_{it}^j) \tag{5.4}$$

where $w_{is}^j = \sum_j \frac{F_{is}^j}{\sum_j F_{is}^j}$ is the time invariant weight with F_{is}^j being the input quantity of fuel type j in tons of oil equivalent for sector s in country i, and P_{it}^j denotes the real price of fuel type j per toe of aggregate industry in country i at time t in constant 2010 USD. In our case, we choose the weights to be the average of the weights from 1995-2015. The FEPI captures only energy price changes that come from changes in fuel prices, and not through changes in the mix fuel inputs. The fuel types considered are electricity, natural gas, coal and oil.

From a general point of view, there are many factors that influence directly or indirectly energy prices, for example Dahl (2015) enumerates the following: population growth, demographic shifts and elongation of human life, income growth, environmental concerns, technology (investment capital available), renewable energies, waste storage and proliferation, government intervention, transportation/travel (moving freight, commuting, recreation and tourism, socializing, shopping, other services, industry travel),

Table 5.3: Countries and sector coverage

Countries		**Sectors**
Australia	Japan	Chemical & petrochemical
Austria	Korea, Republic of	Construction
Belgium	Mexico	Food & tobacco
Brazil	Netherlands	Iron & steel
Canada	New Zealand	Machinery
Croatia	Norway	Mining & quarrying
Cyprus	Poland	Non-Ferrous metals
Czech Republic	Portugal	Non-metallic minerals
Denmark	Romania	Paper, pulp & print
Finland	Slovakia	Textile & leather
France	Sweden	Transport equipment
Germany	Switzerland	Wood & wood products
Greece	Turkey	
Hungary	United Kingdom	
Italy	United States of America	

household heating, cooling, transport, and nuclear energy. We consider the effect of some of these factors over energy prices in our analysis. Having into account the data scarcity, we just include those control variables which are available for all the countries in the sample. Table 5.4 describes the control variables together with their transformations.

Table 5.4: Control variables

	Transformation
Energy imports	1
Energy intensity level of primary energy	1
GDP per unit of energy use	1
Renewable energy consumption	1
Inflation rate	0
Population growth	0
Electricity production from renewable sources	1
Combustible renewables and waste	1

NOTES: Transformation = 1 denotes that the series is in growth rates.
Transformation = 0 denotes no transformation is needed.

Figure 5.1 plots 12 time series, each one corresponding to the FEP of an industrial sector of Australia, the first country in the sample. We observe that Textile & leather, Transport equipment and Wood & wood products sectors are the ones with less variability, whereas Construction, Iron & stell and Mining & quarrying are the ones with largest

variability. It is clear that FEP for Construction sector was the most affected by the financial crisis in 2008, followed by FEP in Mining and Quarrying sector, both of them have not yet recover since then. This analysis gives the intuition that it may exists a cluster structure in the data given by group-specific factors related to industrial sectors. Figure 5.2 represents the construction sector for each of the 30 countries in the sample. The most different country is Turkey, which FEP experiment large decreases around 1998 and 2004, and the one having positive values at the end of the sample period is Brazil. We observe different patterns and magnitudes between countries which could be interpreted as the existence of group-specific factors related to countries.

We implement the DFMCS to the FEP data set in order to analyze whether there exits a cluster structure describing FEP based on countries/industrial sectors, or just global factors. Empirical results are given in next section.

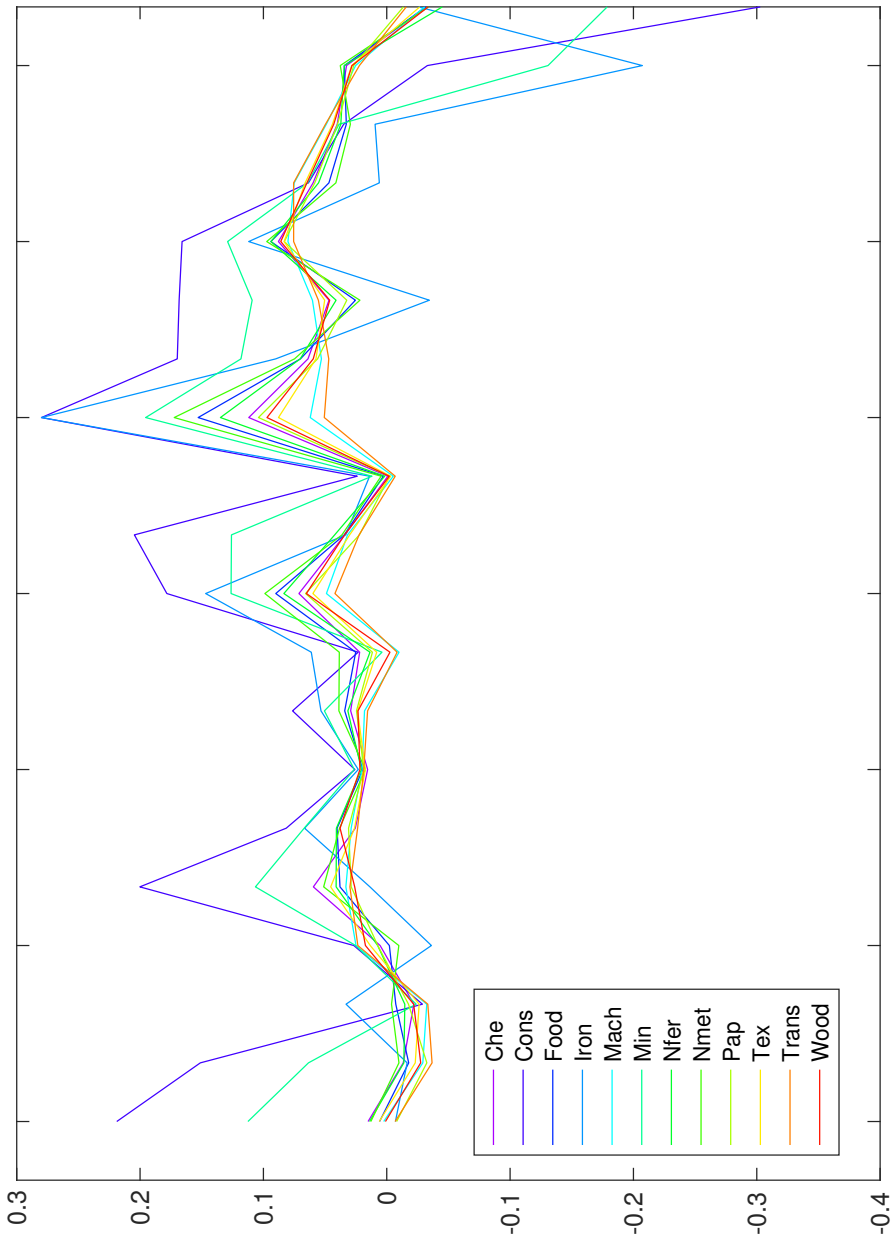

Figure 5.1: Fixed Energy Price for 12 industrial sectors in Australia from 1995 to 2015

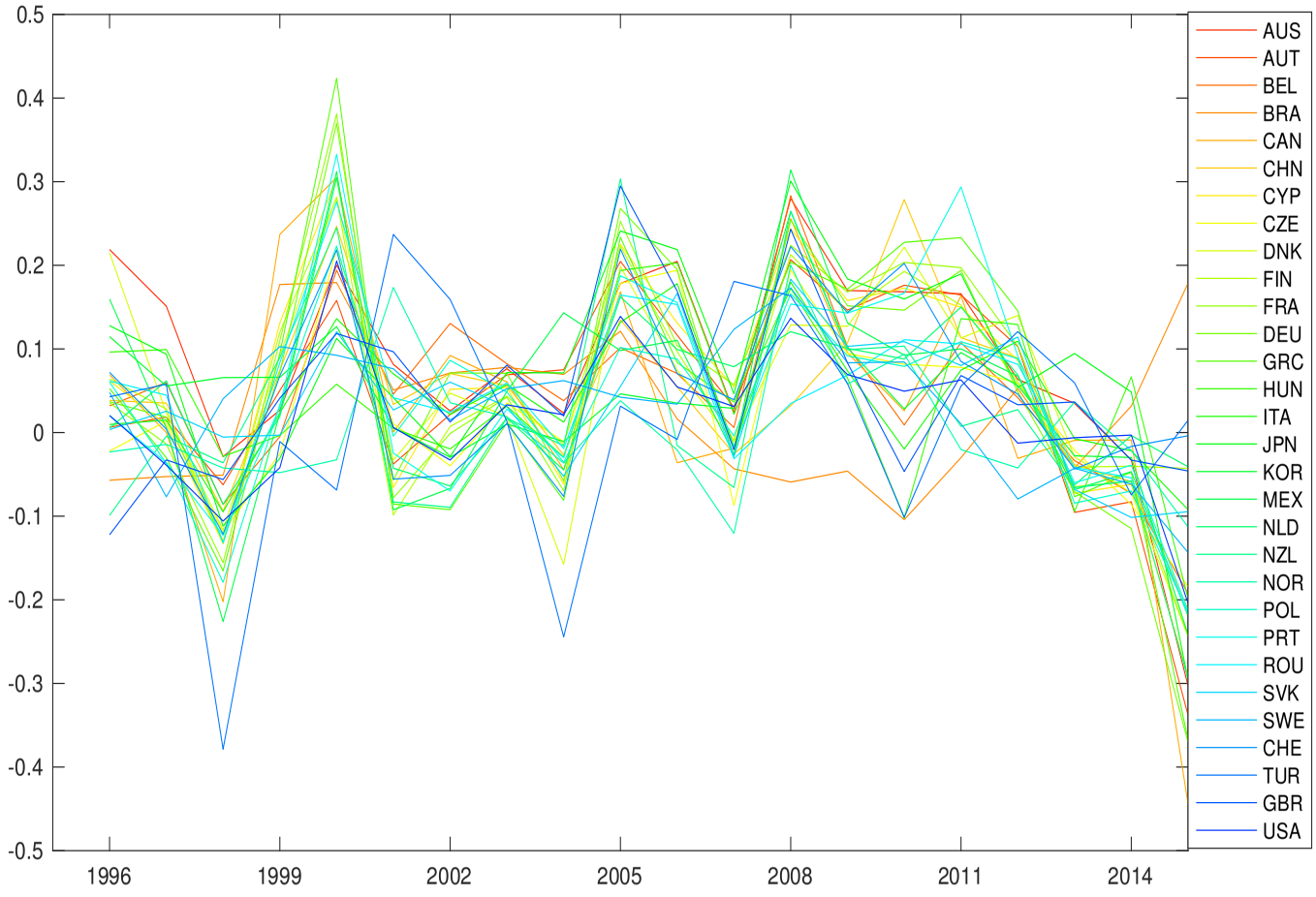

Figure 5.2: Construction sector Fixed Energy Prices for each country in the sample.

5.4 Empirical results

The Monte Carlo simulations clearly show that AGP procedure performs well when estimating the number of groups S and the allocations within groups. Previous to the analysis, the observed time series y_i, are cleaned from additive outliers and level shifts as proposed in Alonso et al. (2020).

First, we run the SCAD-penalized regression, in order to evaluate the effect of control variables in international FEP. The penalization parameter κ_{c_i} is optimized using a BIC criterion and the function `penalized`, in the MATLAB toolbox: 'penalized', see McIlhagga (2016). The estimation results show one relevant coefficient for each country. Given the close relation between prices and inflation, it is the control variable which affects the FEPs in a larger number of countries: Japan, Switzerland, USA, Australia, Austria, Belgium, Cyprus, Czech Republic and Denmark. Energy imports affect energy prices in Greece, Sweden and United Kingdom; Energy intensity level of primary energy is significant in Korea and Norway; GDP per unit of energy use in Romania; Renewable energy consumption in Italy, the Netherlands and Brazil; Population growth in Finland and Germany; Electricity production from renewable sources in France, Mexico, New Zealand and Slovakia; and Combustible renewables and waste in Croatia, Hungary, Poland and Turkey. Once we estimate $\widehat{\beta}_{c_i}$ we subtract the observed component from the observed data, y_i, obtaining $y_i^* = y_i - X_i\widehat{\beta}_{c_i}$.

Second, we estimate the number of common factors in y_i^* using the CP eigenvalue ratio test based on correlation matrices. As suggested by Lam and Yao (2012) we applied the test twice for the possible existence of weak factors. The estimated number of global factors is $r_0 = 2$, see the CP ratio of eigenvalues in Figure 5.3. Each factor explains 72.7% and 7.3% of the total variability, respectively. The time series plots of CP estimated factors and its corresponding loadings are plotted in Figures 5.4 and 5.5, respectively. Loadings corresponding to the first global factor take negative values for all the series in the sample with the exception of Brazil FEP. This factor is able to capture the global dynamic of the differenced series, for example taking negative values during the 2000s energy crisis and the global recession in 2008. The loadings of the second factor takes positive and negative values across countries with differences in magnitude. The largest positive ones correspond to Slovakia, New Zealand and Poland FEP, and the largest negative ones to Greece and Cyprus.

Third, we apply the algorithm proposed in Alonso and Peña (2019), based on the Generalized Cross Correlation (GCC) measure of linear dependency between time series, to the common component. After applying a hierarchical clustering with average linkage to the dissimilarity matrix obtained from $\widehat{GCC}$ measure, we consider a modification of the Silhouette algorithm proposed by Rousseeuw (1987), which give us the num-

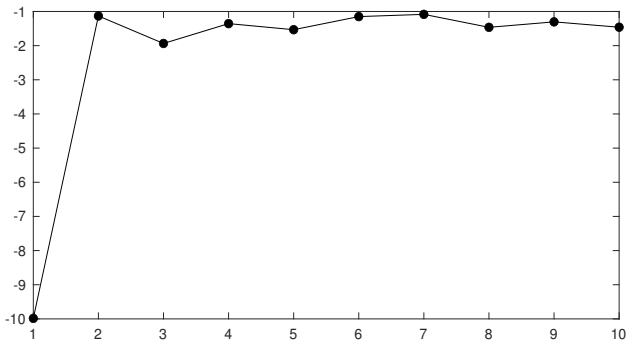

Figure 5.3: CP ratio of eigenvalues for the estimation of the initial factors.

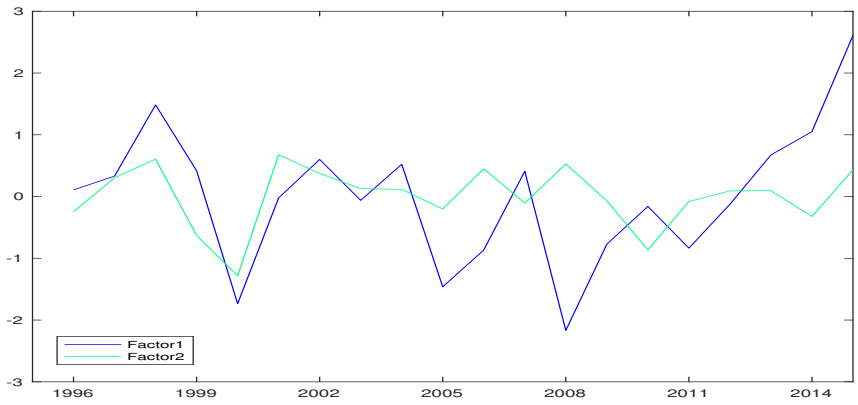

Figure 5.4: Estimated initial factors.

ber of groups $S = 6$ together with the allocation of each series to one of the groups, $G = g_1, ..., g_N$.

Results from the clustering are available in C.1 Empirical Section in the Appendix to Chapter 5. First of all, we observe that groups are classified according to countries instead to industrial sectors. For example, group 1 includes over all FEP of Cyprus, Mexico, USA, Switzerland, Greece, Romania and the Netherlands. Given that USA is the largest trading partner of Mexico it was expected that both of them were in the same cluster. We also observe that series of Construction from different countries are clustered in this group. Group 2 contains principally FEP of United Kingdom, Czech Republic, Poland, Austria, Germany, Slovakia and Denmark. This cluster is representative of countries with high industrial electricity prices according to the study of International industrial energy prices from the Department for Business, Energy & Industrial Strategy at the government of the UK (https://www.gov.uk). From this study we also conclude that group 3 represents mainly FEP of countries with high indus-

trial gas prices, for example Finland, Japan, Korea, France, Italy, Sweden, and Portugal between others. Group 4 includes over all FEP from Hungary, Norway, Belgium and New Zealand. Finally, group 5 represents FEP from Austria and Brazil, and group 6 contains three series which may be consider atypical.

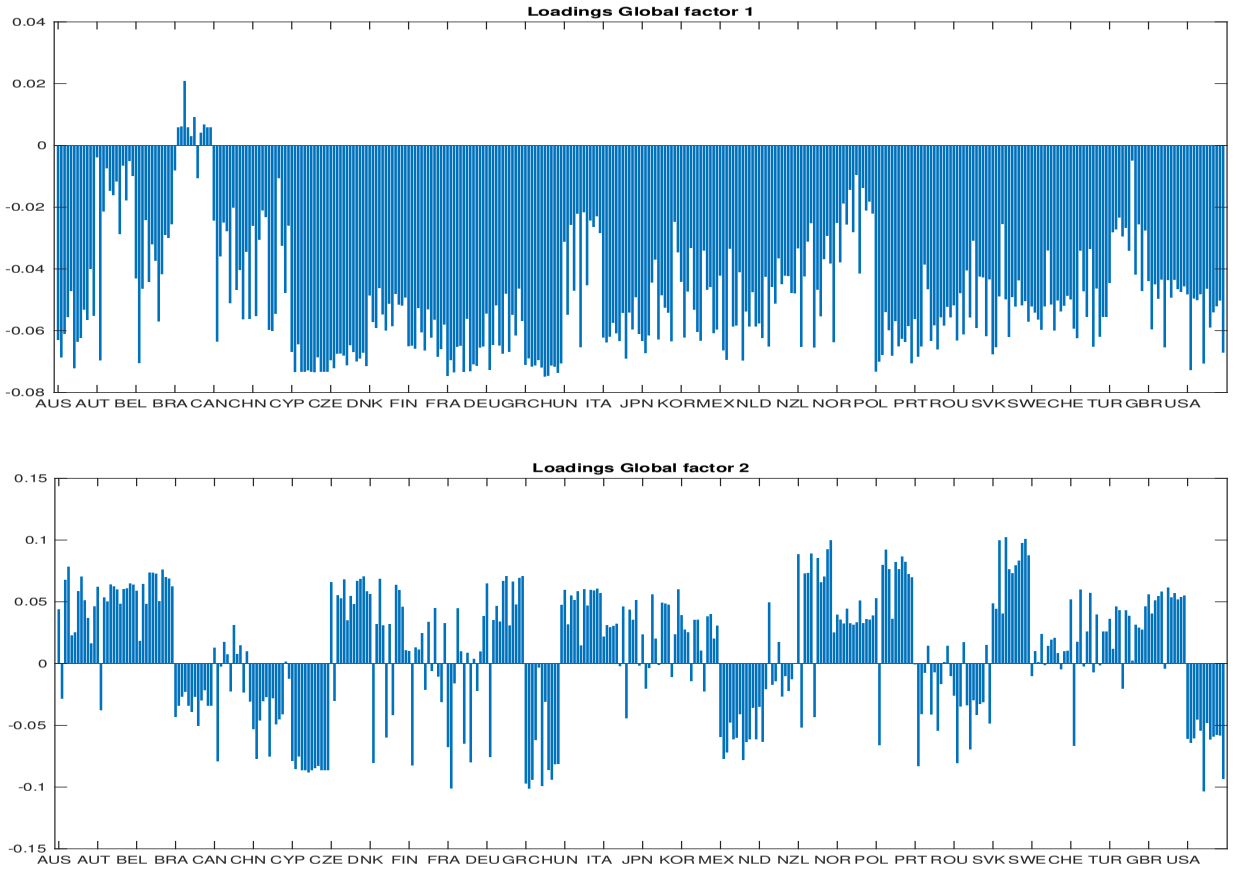

Figure 5.5: Estimated loadings of the two initial factors.

Fourth, once we have the membership of each series to the corresponding group, we estimate the group-specific factors and their corresponding loadings using the time series in each group. The test estimates 3, 2, 3, and 10 factors in groups 1, 2, 3 and 4 respectively. Groups 5 and 6 do not have common factors. These specific-group factors may contain some of the global factors, for these reason all the 'global' and 'group-specific' factors are classified following the rules proposed by Alonso et al. (2020) based on the empirical canonical correlation between each global factor from step (2) and the group-specific factors from step (4). The first factor from the common component is classified as global, given that it is highly correlated with two or more groups, and the second factor is classified as specific because it is highly correlated with group 1.

Finally, we subtract the global common component from y_i^* and re-estimate the group-specific factors. The new estimates are 1, 2, 3, and 5 factors in groups 1, 2, 3 and 4, respectively. Group 4 is the group with largest number of factors, this is because each factor is representative of the FEPs in a specific country. For example, see upper plot in 5.6, the first group-specific factor represents the dynamics in Hungary, whereas the second group-specific factor, bottom plot, is representative of FPE in Norway. The rest of the bar plots representing the group-specific factor loadings are available in C.1 Empirical section in the Appendix to Chapter 5.

In summary, our DFM with cluster structure has 5 groups, 4 of them with group-specific factors and one group only only affected by the global factor.

5.5 Concluding remarks

This chapter has three main contributions: first, we have presented an extension of the methodology proposed by Alonso et al. (2020) for DFMCS. The goodness of fit from this extension has been evaluated in a Monte Carlo experiment and has allowed us to evaluate the effect of macroeconomic variables which are country-specific over a large sample of international energy prices; second, the number of global and group-specific factors are estimated using the test proposed in Caro and Peña (2020) and introduced in Chapter 2; third, the factors and their corresponding loadings are estimated following the approach presented in Chapter 3 based on correlation matrices.

Results from the application of international energy prices have provided useful interpretations about the existence of co-movements between energy prices related to group of countries instead of groups related to industrial sectors. Country connections within groups may be also explained by the high price of a specific fuel type. This analysis gives new insights for public policy decision making, to formulate and implement environmental policies, and for energy market investors to diversify their portfolios.

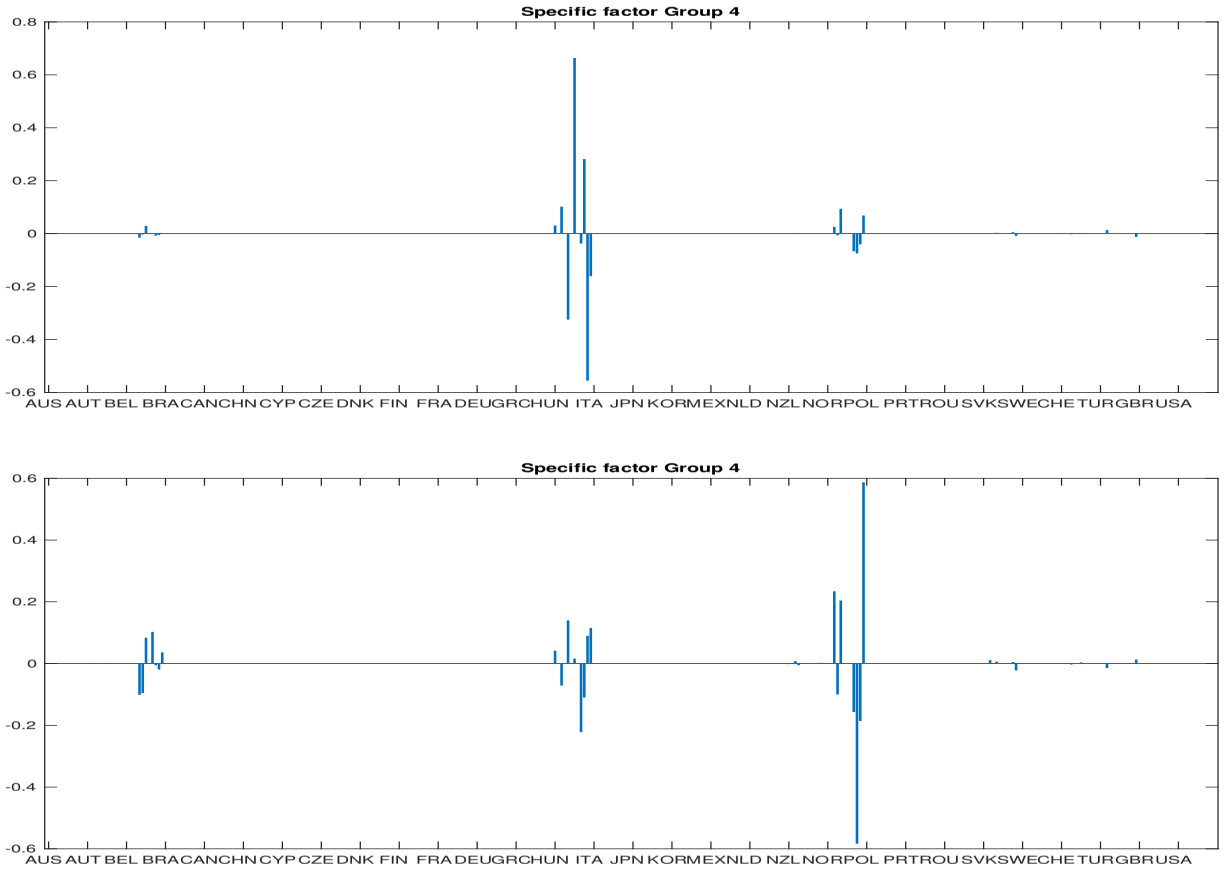

Figure 5.6: Estimated loadings of the first and second specific-factors in Cluster 4.

Chapter 6

Conclusions

This research aimed to improve the way dynamic factor models (DFMs) are built and shows its applications to large databases. Results from this thesis may be useful for economic policy decisions and for analyzing high dimensional heterogeneous dynamic data.

The main contributions of the thesis are as follows: First, this thesis presents a new approach for finding the number of factors in a DFM and estimating them. Second, it extends the methodology proposed by (Alonso et al., 2020) to build DFM with cluster structure by introducing the effect of macroeconomic variables. Third, it shows how DFM can contribute to the analysis of macroeconomic variables which are representative of the business cycles, to the study of CO_2 emission, to the evaluation of the synchronization of Euro Area business cycles and to investigate co-movements between international energy prices.

The first drawback that researchers encounter when estimating the DFM is to select the number of common factors. In this thesis a new eigenvalue ratio test is proposed and theoretical reasons for the advantages of the proposal are presented, especially when the error structures include heteroscedasticity and serial and cross-sectional dependencies. These properties have been confirmed by Monte Carlo simulations and by an application to real macroeconomic data.

Furthermore, we extend the proposed approach to the estimation of the common component of the model. Using Monte Carlo simulations improvements in the estimation of common factors with respect to other alternative methods are observed. This happens especially when the errors are heteroscedastic and present serial and cross-sectional correlation. We show in an application with real data, on CO_2 emissions, that our proposal provides interpretable results which are more meaningful than alternative methods.

Next, we analyze the usefulness of DFMs with cluster structure (DFMCS) in two empirical applications: one on the synchronization of Euro Area business cycles and the other on international energy prices.

The first application, on the synchronization of Euro Area business cycles, shows the advantages of considering DFMCS when analyzing economic relations between country members, evaluating the effect of expansionary monetary policies and studying the effect of the financial crisis in 2008 and the European sovereign debt crisis in 2011. Results conclude that, although the countries experience a generalized fall in synchronization in the financial crisis, they recover the levels of synchronization that characterized the pre-recessive period. Furthermore, results support the presence of a *Two-speed Europe* after the financial crisis in terms of economic synchronization.

Finally, we include an extension in the methodology proposed by (Alonso et al., 2020). In the thesis a penalized regression is proposed to estimate the coefficients associated to explanatory variables. The effect of this estimation method on the group factor structure has been evaluated with Monte Carlo simulations under different data generating processes. This new proposal has been applied to a large data set of international energy prices with country-specific explanatory variables. Results from the analysis identify the existence of co-movements between energy prices related to groups of countries and highlight the effect of some macroeconomic variables.

Future extensions of this research are: first, to develop the theoretical framework behind the estimation of the factor space for the proposed new approach. Second, build up the theoretical assumptions associated to the effect of including exogenous variables that are country-specific in the DFMCS. Third, to evaluate the performance of the new approach under different models specifications for the common latent factors. Fourth, to extend the new approach to the analysis of nonstationary data and time-varying parameters. Fifth, to evaluate the potential of the proposed estimation method for forecasting large data bases.

Bibliography

Aastveit, K. A., Bjørnland, H. C., and Thorsrud, L. A. (2015). What drives oil prices? Emerging versus developed economies. *Journal of Applied Econometrics*, 30(7):1013–1028.

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23-24):1806–1813.

Alonso, A. M., Galeano, P., and Peña, D. (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics*.

Alonso, A. M. and Peña, D. (2019). Clustering time series by linear dependency. *Statistics and Computing*, 29(4):655–676.

Alvarez, R., Camacho, M., and Perez-Quiros, G. (2016). Aggregate versus disaggregate information in dynamic factor models. *International Journal of Forecasting*, 32(3):680–694.

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(519):1182–1198.

Bai, J. and Li, K. (2016). Maximum likelihood estimation and inference for approximate factor models of high dimension. *Review of Economics and Statistics*, 98(2):298–309.

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. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29.

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

Bengoechea, P., Camacho, M., and Perez-Quiros, G. (2006). A useful tool for forecasting the euro-area business cycle phases. *International Journal of Forecasting*, 22(4):735–749.

Bolivar, S., Nieto, F., and Peña, D. (2020). On a new procedure for identifying a dynamic factor model. Working Paper of the Statistics Department at National University of Colombia.

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.

Breitung, J. and Tenhofen, J. (2011). Gls estimation of dynamic factor models. *Journal of the American Statistical Association*, 106(495):1150–1166.

Bretschger, L. (2015). Energy prices, growth, and the channels in between: Theory and evidence. *Resource and Energy Economics*, 39:29–52.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. Classics in Applied Mathematics, San Francisco, CA: Society for Industrial and Applied Mathematics.

Brown, S. P., Yucel, M. K., et al. (2008). What drives natural gas prices? *Energy Journal*, 29(2):45.

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.

Caro, A. and Peña, D. (2020). A test for the number of factors in dynamic factor models. *Submitted*.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets.

Crucini, M. J., Kose, M. A., and Otrok, C. (2011). What are the driving forces of international business cycles? *Review of Economic Dynamics*, 14(1):156–175.

Dahl, C. (2015). *International energy markets: understanding pricing, policies, & profits*. PennWell Books.

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.

Doz, C., Giannone, D., and Reichlin, L. (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, 94(4):1014–1024.

Égert, B. and Kočenda, E. (2011). Time-varying synchronization of european stock markets. *Empirical Economics*, 40(2):393–407.

Engle, R. and Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, 76(376):774–781.

Engle, R. F. and Watson, M. W. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23(3):385–400.

Fan, J., Guo, J., and Zheng, S. (2019). Estimating number of factors by adjusted eigenvalues thresholding. *arXiv preprint arXiv:1909.10710*.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

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., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471):830–840.

Forni, M., Hallin, M., Lippi, M., and Zaffaroni, P. (2015). Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. *Journal of Econometrics*, 185(2):359–371.

Gadea-Rivas, M. D., Gómez-Loscos, A., and Leiva-Leon, D. (2019). Increasing linkages among european regions. the role of sectoral composition. *Economic Modelling*, 80:222–243.

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

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Griffin, J. M. (1980). Henry b. steele. *Energy Economics and Policy (New York: Acadmic Press, 1980)*.

Hallin, M. and Lippi, M. (2013). Factor models in high-dimensional time series—a time-domain approach. *Stochastic Processes and their Applications*, 123(7):2678–2695.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Jacobs, J. P. and Otter, P. W. (2008). Determining the number of factors and lag order in dynamic factor models: A minimum entropy approach. *Econometric Reviews*, 27(4-6):385–397.

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.

Kose, M. A., Otrok, C., and Whiteman, C. H. (2008). Understanding the evolution of world business cycles. *Journal of international Economics*, 75(1):110–130.

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.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Leiva-Leon, D. (2017). Measuring business cycles intra-synchronization in us: A regime-switching interdependence framework. *Oxford Bulletin of Economics and Statistics*, 79(4):513–545.

Mahadevan, R. and Asafu-Adjaye, J. (2007). Energy consumption, economic growth and prices: A reassessment using panel vecm for developed and developing countries. *Energy policy*, 35(4):2481–2490.

McIlhagga, W. H. (2016). penalized: A matlab toolbox for fitting generalized linear models with penalties.

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.

Nick, S. and Thoenes, S. (2014). What drives natural gas prices? A structural var approach. *Energy Economics*, 45:517–527.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Otrok, C. and Whiteman, C. H. (1998). Bayesian leading indicators: measuring and predicting economic conditions in iowa. *International Economic Review*, pages 997–1014.

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

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

Peña, D. and Poncela, P. (2006b). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, 136(4):1237–1257.

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

Phillips, K. L. (1991). A two-country model of stochastic output with changes in regime. *Journal of international economics*, 31(1-2):121–142.

Poncela, P. and Ruiz, E. (2016). Small-versus big-data factor extraction in dynamic factor models: An empirical assessment. In *Dynamic Factor Models*. Emerald Group Publishing Limited.

Quah, D. and Sargent, T. J. (1993). A dynamic index model for large cross sections. In *Business cycles, indicators, and forecasting*, pages 285–310. University of Chicago Press.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Sargent, T. J. (1989). Two models of measurements and the investment accelerator. *Journal of Political Economy*, 97(2):251–287.

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.

Sato, M., Singer, G., Dussaux, D., and Lovo, S. (2019). International and sectoral variation in industrial energy prices 1995–2015. *Energy Economics*, 78:235–258.

Stewart, G. and Sun, J. (1990). *Computer science and scientific computing. Matrix perturbation theory*. Academic Press.

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. (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*, 4:351–394.

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. (2005). Implications of dynamic factor models for var analysis. Technical report, National Bureau of Economic Research.

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.

Van Benthem, A. and Romani, M. (2009). Fuelling growth: what drives energy demand in developing countries? *The Energy Journal*, pages 91–114.

Wang, P. (2008). Large dimensional factor models with a multi-level factor structure: identification, estimation and inference. *Unpublished manuscript, New York University*.

Appendix A

Chapter 2

A.1 Tables

Table A.1: Relative frequency estimates of the true number of common factors $r = 2$. Homoscedastic errors and strong signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.375	0.38	0.35	0.395	0.345	0.355	0.37
T=250	0.545	0.675	0.485	0.385	0.355	0.395	0.47
T=500	0.61	0.795	0.555	0.54	0.5	0.51	0.58
T=1250	0.715	0.95	0.605	0.605	0.57	0.52	0.66
N=50							
T=125	0.94	0.85	0.89	0.865	0.685	0.795	0.84
T=250	0.98	0.97	0.97	0.97	0.875	0.975	0.96
T=500	0.995	1	0.995	0.99	0.965	0.98	0.99
T=1250	1	1	1	1	0.995	0.995	1
N=100							
T=125	0.96	0.885	0.92	0.93	0.805	0.9	0.9
T=250	1	0.99	0.99	0.995	0.915	0.985	0.98
T=500	1	1	1	1	1	1	1
T=1250	1	1	1	1	1	1	1
N=200							
T=125	0.995	0.92	0.96	0.995	0.865	0.945	0.95
T=250	1	0.995	0.995	1	0.985	0.995	1
T=500	1	1	1	1	1	1	1
T=1250	1	1	1	1	1	1	1
Mean	0.88	**0.9**	0.86	**0.85**	0.8	0.83	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$ and $\phi_{f_2} = 0.8$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix $\Gamma_e = \sigma_e \mathbf{I}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table A.2: Relative frequency estimates of the true number of common factors $r = 2$. Homoscedastic errors and weak signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.225	0.135	0.205	0.285	0.23	0.295	0.23
T=250	0.22	0.15	0.255	0.215	0.09	0.22	0.19
T=500	0.2	0.125	0.2	0.2	0	0.27	0.17
T=1250	0.365	0.16	0.32	0.135	0	0.205	0.2
N=50							
T=125	0.37	0.185	0.29	0.26	0.195	0.235	0.26
T=250	0.57	0.275	0.495	0.36	0.2	0.33	0.37
T=500	0.83	0.505	0.79	0.525	0.23	0.505	0.56
T=1250	0.95	0.815	0.955	0.835	0.22	0.84	0.77
N=100							
T=125	0.535	0.235	0.425	0.29	0.215	0.255	0.33
T=250	0.87	0.61	0.835	0.615	0.27	0.485	0.61
T=500	0.995	0.765	0.99	0.905	0.315	0.825	0.8
T=1250	1	0.955	1	0.99	0.565	0.995	0.92
N=200							
T=125	0.785	0.41	0.575	0.475	0.17	0.32	0.46
T=250	1	0.765	0.95	0.835	0.415	0.765	0.79
T=500	1	0.935	1	0.995	0.615	0.965	0.92
T=1250	1	1	1	1	0.945	1	0.99
Mean	**0.68**	0.5	0.64	**0.56**	0.29	0.53	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$ and $\phi_{f_2} = 0.5$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1 - 3$ idiosyncratic covariance matrix $\Gamma_e = \sigma_e \mathbf{I}$. Columns $4 - 6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table A.3: Relative frequency estimates of the true number of common factors $r = 2$. Heteroscedastic errors and medium signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.015	0.155	0.215	0.035	0.43	0.25	0.18
T=250	0	0.14	0.265	0.005	0.675	0.265	0.22
T=500	0	0.155	0.26	0.01	0.78	0.295	0.25
T=1250	0	0.25	0.365	0.01	0.865	0.295	0.3
N=50							
T=125	0.365	0.2	0.695	0.215	0.235	0.47	0.36
T=250	0.445	0.345	0.865	0.305	0.26	0.69	0.48
T=500	0.665	0.57	0.95	0.41	0.145	0.935	0.61
T=1250	0.825	0.925	1	0.495	0.08	0.965	0.72
N=100							
T=125	0.675	0.435	0.79	0.39	0.205	0.615	0.52
T=250	0.865	0.655	0.98	0.71	0.28	0.94	0.74
T=500	0.965	0.87	1	0.86	0.36	0.985	0.84
T=1250	0.99	0.99	1	0.97	0.555	1	0.92
N=200							
T=125	0.885	0.565	0.93	0.68	0.31	0.81	0.7
T=250	1	0.78	1	0.955	0.475	0.985	0.87
T=500	1	0.975	1	0.995	0.695	1	0.94
T=1250	1	1	1	1	0.955	1	0.99
Mean	0.61	0.56	**0.77**	0.5	0.46	**0.72**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$ and $\phi_{f_2} = 0.5$, and errors η_t are independent $N(0, 1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even..

Table A.4: Relative frequency estimates of the true number of common factors $r = 2$. Heteroscedastic and cross correlated errors and strong signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.26	0.135	0.215	0.26	0.355	0.25	0.25
T=250	0.265	0.21	0.19	0.27	0.405	0.2	0.26
T=500	0.21	0.345	0.195	0.28	0.33	0.18	0.26
T=1250	0.24	0.455	0.17	0.23	0.265	0.14	0.25
N=50							
T=125	0.065	0.27	0.17	0.18	0.185	0.17	0.17
T=250	0.07	0.495	0.27	0.11	0.155	0.205	0.22
T=500	0.07	0.805	0.275	0.1	0.17	0.215	0.27
T=1250	0.085	0.965	0.24	0.075	0.08	0.18	0.27
N=100							
T=125	0.245	0.385	0.41	0.14	0.23	0.39	0.3
T=250	0.31	0.66	0.58	0.19	0.3	0.525	0.43
T=500	0.415	0.945	0.765	0.215	0.32	0.56	0.54
T=1250	0.415	1	0.845	0.255	0.545	0.745	0.63
N=200							
T=125	0.485	0.58	0.695	0.42	0.44	0.62	0.54
T=250	0.65	0.825	0.855	0.525	0.51	0.805	0.7
T=500	0.86	0.98	0.975	0.72	0.755	0.91	0.87
T=1250	0.965	1	1	0.86	0.955	0.995	0.96
Mean	0.35	**0.63**	0.49	0.3	0.38	**0.44**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$ and $\phi_{f_2} = 0.8$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-4$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $5-8$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Table A.5: Relative frequency estimates of the true number of common factors $r = 2$. Heteroscedastic and cross correlated errors and weak signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.105	0.245	0.43	0.195	0.11	0.54	0.27
T=250	0.055	0.22	0.485	0.115	0.175	0.66	0.28
T=500	0.035	0.25	0.425	0.145	0.215	0.86	0.32
T=1250	0.005	0.17	0.53	0.05	0.555	0.925	0.37
N=50							
T=125	0.07	0.155	0.09	0.085	0.3	0.055	0.13
T=250	0.04	0.13	0.04	0.03	0.315	0.07	0.1
T=500	0.025	0.135	0.02	0.025	0.29	0.035	0.09
T=1250	0.005	0.165	0.01	0.03	0.235	0.01	0.08
N=100							
T=125	0.11	0.155	0.09	0.18	0.285	0.2	0.17
T=250	0.075	0.15	0.05	0.165	0.255	0.105	0.13
T=500	0.025	0.155	0.03	0.12	0.2	0.07	0.1
T=1250	0.01	0.145	0	0.17	0.125	0.02	0.08
N=200							
T=125	0.14	0.175	0.15	0.205	0.235	0.22	0.19
T=250	0.09	0.135	0.11	0.21	0.205	0.18	0.16
T=500	0.045	0.175	0.09	0.215	0.235	0.175	0.16
T=1250	0.005	0.14	0.04	0.135	0.13	0.08	0.09
Mean	0.05	**0.17**	0.16	0.13	0.24	**0.26**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$ and $\phi_{f_2} = 0.5$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1 - 4$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $5 - 8$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Table A.6: Relative frequency estimates of the true number of common factors $r = 3$. Homoscedastic errors and strong signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.18	0.09	0.08	0.14	0.11	0.12	0.12
T=250	0.315	0.34	0.21	0.16	0.11	0.09	0.2
T=500	0.315	0.45	0.195	0.18	0.13	0.155	0.24
T=1250	0.325	0.71	0.205	0.25	0.19	0.195	0.31
N=50							
T=125	0.805	0.505	0.65	0.625	0.315	0.485	0.56
T=250	0.97	0.91	0.905	0.87	0.58	0.8	0.84
T=500	1	1	1	0.98	0.86	0.97	0.97
T=1250	1	1	0.995	1	0.995	0.99	1
N=100							
T=125	0.975	0.715	0.855	0.87	0.52	0.735	0.78
T=250	0.995	0.96	0.99	0.985	0.835	0.965	0.96
T=500	1	1	1	1	0.985	1	1
T=1250	1	1	1	1	1	1	1
N=200							
T=125	1	0.735	0.895	0.96	0.585	0.77	0.82
T=250	1	0.995	0.995	1	0.95	1	0.99
T=500	1	1	1	1	1	1	1
T=1250	1	1	1	1	1	1	1
Mean	**0.8**	0.78	0.75	**0.75**	0.64	0.7	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$, $\phi_{f_2} = 0.8$ and $\phi_{f_3} = 0.7$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ for $i = 1, ..., N$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table A.7: Relative frequency estimates of the true number of common factors $r = 3$. Homoscedastic errors and medium signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.145	0.045	0.11	0.11	0.195	0.115	0.12
T=250	0.225	0.05	0.145	0.195	0.1	0.125	0.14
T=500	0.32	0.15	0.27	0.145	0.015	0.145	0.17
T=1250	0.295	0.23	0.21	0.185	0.02	0.205	0.19
N=50							
T=125	0.875	0.18	0.77	0.69	0.16	0.49	0.53
T=250	0.995	0.565	0.965	0.955	0.23	0.86	0.76
T=500	1	0.81	1	0.99	0.435	0.98	0.87
T=1250	1	0.985	1	1	0.735	0.995	0.95
N=100							
T=125	0.995	0.365	0.895	0.935	0.285	0.78	0.71
T=250	1	0.735	0.995	1	0.485	0.985	0.87
T=500	1	0.96	1	1	0.855	1	0.97
T=1250	1	0.995	1	1	0.995	1	1
N=200							
T=125	1	0.51	0.965	0.99	0.435	0.885	0.8
T=250	1	0.86	1	1	0.805	1	0.94
T=500	1	0.99	1	1	0.985	1	1
T=1250	1	1	1	1	1	1	1
Mean	**0.8**	0.59	0.77	**0.76**	0.48	0.72	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ and $\phi_{f_3} = 0.4$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ for $i = 1, ..., N$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table A.8: Relative frequency estimates of the true number of common factors $r = 3$. Heteroscedastic errors and strong signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.055	0.04	0.035	0.045	0.18	0.03	0.06
T=250	0.025	0.045	0.04	0.03	0.23	0.03	0.07
T=500	0.015	0.045	0.045	0.045	0.22	0.04	0.07
T=1250	0.01	0.21	0.05	0.035	0.18	0.055	0.09
N=50							
T=125	0.095	0.07	0.29	0.06	0.03	0.135	0.11
T=250	0.26	0.29	0.675	0.095	0.045	0.395	0.29
T=500	0.375	0.615	0.85	0.165	0.08	0.72	0.47
T=1250	0.58	0.98	0.975	0.26	0.16	0.93	0.65
N=100							
T=125	0.345	0.21	0.605	0.21	0.075	0.37	0.3
T=250	0.61	0.54	0.885	0.395	0.145	0.76	0.56
T=500	0.91	0.93	0.995	0.685	0.46	0.955	0.82
T=1250	0.99	1	1	0.905	0.715	1	0.94
N=200							
T=125	0.67	0.41	0.715	0.35	0.15	0.54	0.47
T=250	0.94	0.8	0.97	0.745	0.395	0.925	0.8
T=500	1	1	1	0.975	0.8	0.995	0.96
T=1250	1	1	1	1	0.985	1	1
Mean	0.49	0.51	**0.63**	0.38	0.3	**0.56**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$, $\phi_{f_2} = 0.8$ and $\phi_{f_3} = 0.7$, and errors η_t are independent $N(0, 1)$ random variables. Columns $1 - 3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4 - 6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even..

Table A.9: Relative frequency estimates of the true number of common factors $r = 3$. Heteroscedastic errors and weak signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.005	0.155	0.18	0.01	0.025	0.1	0.08
T=250	0	0.2	0.175	0.005	0.02	0.12	0.09
T=500	0	0.16	0.115	0	0.005	0.135	0.07
T=1250	0	0.205	0.085	0	0	0.1	0.06
N=50							
T=125	0.05	0.06	0.06	0.065	0.225	0.12	0.1
T=250	0	0.08	0.06	0.02	0.105	0.135	0.07
T=500	0	0.09	0.18	0	0.07	0.06	0.07
T=1250	0	0.02	0.455	0	0	0.16	0.11
N=100							
T=125	0.04	0.095	0.055	0.105	0.11	0.1	0.08
T=250	0.08	0.055	0.15	0.13	0.135	0.085	0.11
T=500	0.015	0.015	0.525	0.025	0.1	0.1	0.13
T=1250	0.005	0.04	0.935	0	0.075	0.585	0.27
N=200							
T=125	0.035	0.03	0.155	0.1	0.14	0.1	0.09
T=250	0.1	0.02	0.51	0.085	0.095	0.145	0.16
T=500	0.255	0.025	0.94	0.17	0.105	0.465	0.33
T=1250	0.76	0.185	1	0.15	0.055	0.98	0.52
Mean	0.08	0.09	**0.35**	0.05	0.08	**0.22**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ and $\phi_{f_3} = 0.4$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table A.10: Relative frequency estimates of the true number of common factors $r = 3$. Heteroscedastic and cross correlated errors and medium signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.125	0.23	0.115	0.06	0.045	0.15	0.12
T=250	0.075	0.145	0.09	0.02	0.035	0.085	0.08
T=500	0.085	0.13	0.095	0.035	0.03	0.06	0.07
T=1250	0.085	0.185	0.075	0.045	0.015	0.075	0.08
N=50							
T=125	0.095	0.065	0.095	0.045	0.145	0.095	0.09
T=250	0.05	0.08	0.05	0.065	0.2	0.07	0.09
T=500	0.07	0.055	0.05	0.085	0.17	0.075	0.08
T=1250	0.055	0.115	0.03	0.06	0.2	0.04	0.08
N=100							
T=125	0.065	0.05	0.095	0.09	0.15	0.12	0.1
T=250	0.045	0.025	0.105	0.055	0.13	0.115	0.08
T=500	0.05	0.075	0.135	0.065	0.14	0.07	0.09
T=1250	0.045	0.435	0.215	0.085	0.12	0.06	0.16
N=200							
T=125	0.145	0.055	0.32	0.12	0.165	0.175	0.16
T=250	0.33	0.17	0.73	0.145	0.09	0.325	0.3
T=500	0.585	0.42	0.915	0.185	0.11	0.62	0.47
T=1250	0.79	0.88	0.995	0.3	0.01	0.89	0.64
Mean	0.17	0.19	**0.26**	0.09	0.11	**0.19**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ and $\phi_{f_3} = 0.4$, and errors η_t are independent $N(0, 1)$ random variables. Columns $1 - 3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4 - 6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Table A.11: Relative frequency estimates of the true number of common factors $r = 3$. Heteroscedastic and cross correlated errors and weak signal to noise ratio.

	AH	LY	CP	AH	LY	CP	Mean
N=10							
T=125	0.035	0.185	0.145	0.025	0.035	0.06	0.08
T=250	0.02	0.175	0.075	0.005	0.045	0.08	0.07
T=500	0.005	0.175	0.085	0.005	0.005	0.045	0.05
T=1250	0	0.15	0.03	0	0	0.005	0.03
N=50							
T=125	0.06	0.12	0.11	0.065	0.18	0.11	0.11
T=250	0.05	0.11	0.11	0.045	0.18	0.055	0.09
T=500	0.045	0.11	0.07	0.02	0.23	0.06	0.09
T=1250	0.015	0.12	0.035	0.04	0.235	0.035	0.08
N=100							
T=125	0.095	0.095	0.13	0.1	0.15	0.085	0.11
T=250	0.03	0.115	0.055	0.06	0.145	0.04	0.07
T=500	0.01	0.11	0.02	0.055	0.145	0.02	0.06
T=1250	0	0.105	0	0.055	0.125	0.005	0.05
N=200							
T=125	0.05	0.105	0.09	0.12	0.135	0.145	0.11
T=250	0.05	0.08	0.085	0.1	0.12	0.1	0.09
T=500	0.02	0.095	0.04	0.14	0.11	0.09	0.08
T=1250	0.005	0.005	0.01	0.12	0.07	0.045	0.04
Mean	0.03	**0.12**	0.07	0.06	**0.12**	0.06	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, $\phi_{f_2} = 0.5$ and $\phi_{f_3} = 0.4$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Appendix B

Chapter 3

B.1 Tables

Table B.1: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Homoscedastic errors and strong signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.965	0.96	0.952	0.931	0.9	0.917	0.94
T=250	0.984	0.981	0.968	0.966	0.949	0.951	0.97
T=500	0.994	0.993	0.978	0.985	0.978	0.972	0.98
T=1250	0.998	0.997	0.98	0.993	0.988	0.979	0.99
N=50							
T=125	0.975	0.966	0.962	0.95	0.929	0.935	0.95
T=250	0.989	0.985	0.977	0.976	0.966	0.964	0.98
T=500	0.995	0.993	0.982	0.988	0.984	0.977	0.99
T=1250	0.998	0.997	0.986	0.996	0.994	0.983	0.99
N=100							
T=125	0.976	0.967	0.964	0.949	0.925	0.932	0.95
T=250	0.99	0.986	0.977	0.977	0.968	0.964	0.98
T=500	0.995	0.993	0.984	0.989	0.985	0.977	0.99
T=1250	0.998	0.998	0.986	0.996	0.994	0.984	0.99
N=200							
T=125	0.975	0.964	0.962	0.951	0.929	0.935	0.95
T=250	0.99	0.986	0.978	0.977	0.968	0.964	0.98
T=500	0.995	0.994	0.984	0.989	0.985	0.977	0.99
T=1250	0.998	0.997	0.987	0.996	0.994	0.984	0.99
Mean	**0.99**	0.98	0.98	**0.98**	0.96	0.96	

NOTES: Factor autoregressive coefficient $\phi_{f_1} = 0.9$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix $\Gamma_e = \sigma_e \mathbf{I}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table B.2: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Homoscedastic errors and weak signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.796	0.426	0.781	0.668	0.254	0.662	0.6
T=250	0.879	0.556	0.861	0.793	0.289	0.795	0.7
T=500	0.952	0.694	0.937	0.858	0.295	0.876	0.77
T=1250	0.979	0.827	0.968	0.915	0.429	0.942	0.84
N=50							
T=125	0.899	0.719	0.896	0.854	0.375	0.834	0.76
T=250	0.949	0.787	0.947	0.923	0.523	0.915	0.84
T=500	0.974	0.845	0.972	0.961	0.663	0.956	0.9
T=1250	0.989	0.911	0.988	0.983	0.802	0.98	0.94
N=100							
T=125	0.908	0.753	0.905	0.869	0.476	0.851	0.79
T=250	0.952	0.8	0.95	0.932	0.641	0.923	0.87
T=500	0.976	0.854	0.975	0.965	0.746	0.96	0.91
T=1250	0.99	0.919	0.989	0.986	0.857	0.981	0.95
N=200							
T=125	0.912	0.765	0.909	0.872	0.549	0.856	0.81
T=250	0.956	0.815	0.955	0.937	0.688	0.929	0.88
T=500	0.978	0.863	0.976	0.968	0.784	0.963	0.92
T=1250	0.991	0.921	0.99	0.987	0.876	0.982	0.96
Mean	**0.94**	0.78	**0.94**	**0.9**	0.58	**0.9**	

NOTES: Factor autoregressive coefficient $\phi_{f_1} = 0.3$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix $\Gamma_e = \sigma_e \mathbf{I}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table B.3: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Heteroscedastic errors and strong signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.691	0.841	0.87	0.589	0.64	0.822	0.74
T=250	0.762	0.947	0.917	0.653	0.76	0.89	0.82
T=500	0.822	0.978	0.931	0.746	0.837	0.92	0.87
T=1250	0.875	0.992	0.939	0.786	0.878	0.931	0.9
N=50							
T=125	0.912	0.911	0.883	0.849	0.828	0.848	0.87
T=250	0.961	0.964	0.915	0.921	0.912	0.89	0.93
T=500	0.978	0.983	0.926	0.958	0.956	0.912	0.95
T=1250	0.989	0.994	0.932	0.979	0.982	0.925	0.97
N=100							
T=125	0.935	0.922	0.889	0.868	0.833	0.84	0.88
T=250	0.97	0.964	0.913	0.933	0.917	0.886	0.93
T=500	0.985	0.983	0.925	0.97	0.963	0.912	0.96
T=1250	0.993	0.993	0.93	0.986	0.985	0.923	0.97
N=200							
T=125	0.939	0.921	0.887	0.876	0.837	0.84	0.88
T=250	0.971	0.964	0.912	0.939	0.919	0.886	0.93
T=500	0.987	0.983	0.924	0.97	0.961	0.91	0.96
T=1250	0.995	0.994	0.931	0.988	0.985	0.922	0.97
Mean	0.92	**0.96**	0.91	0.88	0.89	0.89	

NOTES: Factor autoregressive coefficient $\phi_{f_1} = 0.9$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table B.4: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Heteroscedastic and cross correlated errors and strong signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.234	0.815	0.358	0.214	0.464	0.331	0.4
T=250	0.252	0.942	0.366	0.189	0.471	0.33	0.42
T=500	0.251	0.976	0.437	0.222	0.537	0.371	0.47
T=1250	0.244	0.992	0.427	0.222	0.594	0.377	0.48
N=50							
T=125	0.686	0.904	0.847	0.541	0.699	0.765	0.74
T=250	0.825	0.962	0.897	0.734	0.874	0.862	0.86
T=500	0.883	0.982	0.914	0.778	0.933	0.893	0.9
T=1250	0.931	0.994	0.927	0.856	0.97	0.91	0.93
N=100							
T=125	0.881	0.917	0.878	0.784	0.805	0.826	0.85
T=250	0.943	0.963	0.909	0.886	0.902	0.88	0.91
T=500	0.97	0.982	0.922	0.941	0.955	0.907	0.95
T=1250	0.983	0.994	0.931	0.97	0.982	0.921	0.96
N=200							
T=125	0.923	0.917	0.881	0.848	0.824	0.834	0.87
T=250	0.967	0.965	0.912	0.926	0.916	0.884	0.93
T=500	0.983	0.983	0.924	0.964	0.961	0.91	0.95
T=1250	0.992	0.994	0.931	0.984	0.984	0.922	0.97
Mean	0.75	**0.96**	0.78	0.69	**0.8**	0.75	

NOTES: Factor autoregressive coefficient $\phi_{f_1} = 0.9$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Table B.5: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 1$. Heteroscedastic and cross correlated errors and weak signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.118	0.135	0.119	0.112	0.113	0.11	0.12
T=250	0.127	0.141	0.126	0.129	0.113	0.124	0.13
T=500	0.1	0.136	0.102	0.114	0.118	0.114	0.11
T=1250	0.102	0.194	0.112	0.11	0.091	0.118	0.12
N=50							
T=125	0.056	0.072	0.103	0.053	0.035	0.083	0.07
T=250	0.06	0.095	0.108	0.047	0.034	0.086	0.07
T=500	0.07	0.151	0.123	0.044	0.033	0.089	0.09
T=1250	0.066	0.341	0.123	0.043	0.03	0.078	0.11
N=100							
T=125	0.127	0.125	0.311	0.064	0.036	0.177	0.14
T=250	0.15	0.202	0.432	0.062	0.03	0.228	0.18
T=500	0.19	0.37	0.55	0.065	0.03	0.317	0.25
T=1250	0.244	0.674	0.627	0.088	0.03	0.428	0.35
N=200							
T=125	0.46	0.328	0.675	0.186	0.058	0.438	0.36
T=250	0.624	0.498	0.777	0.322	0.061	0.642	0.49
T=500	0.765	0.657	0.834	0.436	0.055	0.77	0.59
T=1250	0.846	0.806	0.866	0.639	0.05	0.832	0.67
Mean	0.26	0.31	**0.37**	0.16	0.06	**0.29**	

NOTES: Factor autoregressive coefficient $\phi_{f_1} = 0.3$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Table B.6: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Homoscedastic errors and strong signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.468	0.458	0.458	0.454	0.436	0.447	0.45
T=250	0.476	0.47	0.467	0.469	0.457	0.461	0.47
T=500	0.479	0.475	0.469	0.474	0.465	0.466	0.47
T=1250	0.477	0.473	0.467	0.479	0.473	0.47	0.47
N=50							
T=125	0.481	0.476	0.474	0.469	0.457	0.461	0.47
T=250	0.487	0.484	0.48	0.481	0.476	0.474	0.48
T=500	0.49	0.489	0.484	0.488	0.485	0.482	0.49
T=1250	0.491	0.49	0.485	0.49	0.489	0.484	0.49
N=100							
T=125	0.484	0.479	0.477	0.472	0.461	0.464	0.47
T=250	0.49	0.488	0.485	0.484	0.48	0.478	0.48
T=500	0.493	0.492	0.488	0.491	0.489	0.485	0.49
T=1250	0.496	0.495	0.49	0.493	0.492	0.487	0.49
N=200							
T=125	0.486	0.48	0.479	0.473	0.461	0.464	0.47
T=250	0.493	0.491	0.487	0.487	0.482	0.48	0.49
T=500	0.495	0.494	0.49	0.492	0.49	0.486	0.49
T=1250	0.497	0.497	0.491	0.496	0.495	0.49	0.49
Mean	0.49	0.48	0.48	0.48	0.47	0.47	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$, and $\phi_{f_2} = 0.6$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ for $i = 1, ..., N$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table B.7: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Homoscedastic errors and medium signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.452	0.39	0.445	0.415	0.282	0.41	0.4
T=250	0.473	0.429	0.466	0.452	0.324	0.454	0.43
T=500	0.484	0.45	0.477	0.467	0.374	0.468	0.45
T=1250	0.492	0.472	0.486	0.48	0.404	0.479	0.47
N=50							
T=125	0.463	0.418	0.458	0.443	0.357	0.431	0.43
T=250	0.48	0.45	0.476	0.47	0.415	0.462	0.46
T=500	0.489	0.47	0.485	0.483	0.448	0.477	0.48
T=1250	0.493	0.483	0.49	0.49	0.472	0.486	0.49
N=100							
T=125	0.467	0.425	0.461	0.446	0.365	0.433	0.43
T=250	0.482	0.456	0.478	0.471	0.419	0.462	0.46
T=500	0.489	0.473	0.486	0.484	0.455	0.478	0.48
T=1250	0.494	0.486	0.492	0.491	0.477	0.487	0.49
N=200							
T=125	0.469	0.429	0.463	0.449	0.377	0.435	0.44
T=250	0.483	0.457	0.479	0.473	0.428	0.464	0.46
T=500	0.491	0.476	0.487	0.485	0.457	0.479	0.48
T=1250	0.495	0.488	0.493	0.493	0.48	0.488	0.49
Mean	**0.48**	0.45	**0.48**	**0.47**	0.41	0.46	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, and $\phi_{f_2} = 0.3$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ for $i = 1, ..., N$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ $i = 1, ..., N$.

Table B.8: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Heteroscedastic errors and strong signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.355	0.42	0.427	0.302	0.311	0.4	0.37
T=250	0.408	0.454	0.445	0.352	0.381	0.436	0.41
T=500	0.414	0.468	0.453	0.386	0.415	0.444	0.43
T=1250	0.423	0.472	0.454	0.399	0.43	0.451	0.44
N=50							
T=125	0.454	0.452	0.441	0.415	0.401	0.417	0.43
T=250	0.472	0.473	0.451	0.455	0.448	0.441	0.46
T=500	0.482	0.484	0.459	0.473	0.471	0.452	0.47
T=1250	0.487	0.489	0.462	0.482	0.482	0.458	0.48
N=100							
T=125	0.463	0.457	0.443	0.429	0.41	0.418	0.44
T=250	0.481	0.478	0.456	0.464	0.455	0.442	0.46
T=500	0.488	0.487	0.461	0.48	0.476	0.454	0.47
T=1250	0.492	0.492	0.464	0.489	0.488	0.461	0.48
N=200							
T=125	0.468	0.46	0.445	0.439	0.42	0.421	0.44
T=250	0.484	0.481	0.457	0.469	0.46	0.444	0.47
T=500	0.491	0.489	0.462	0.483	0.479	0.455	0.48
T=1250	0.495	0.495	0.466	0.492	0.491	0.462	0.48
Mean	0.46	**0.47**	0.45	0.44	0.44	0.44	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$, and $\phi_{f_2} = 0.6$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table B.9: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Heteroscedastic errors and weak signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.134	0.106	0.189	0.114	0.101	0.151	0.13
T=250	0.116	0.11	0.226	0.11	0.107	0.179	0.14
T=500	0.13	0.134	0.294	0.115	0.102	0.222	0.17
T=1250	0.135	0.148	0.37	0.109	0.095	0.291	0.19
N=50							
T=125	0.069	0.067	0.2	0.047	0.044	0.126	0.09
T=250	0.093	0.087	0.308	0.056	0.043	0.187	0.13
T=500	0.111	0.145	0.378	0.058	0.04	0.298	0.17
T=1250	0.162	0.288	0.423	0.077	0.048	0.393	0.23
N=100							
T=125	0.087	0.089	0.251	0.045	0.032	0.139	0.11
T=250	0.142	0.153	0.336	0.064	0.036	0.236	0.16
T=500	0.221	0.255	0.392	0.107	0.055	0.335	0.23
T=1250	0.33	0.382	0.429	0.173	0.089	0.406	0.3
N=200							
T=125	0.148	0.145	0.279	0.071	0.044	0.165	0.14
T=250	0.258	0.246	0.353	0.116	0.063	0.269	0.22
T=500	0.353	0.335	0.399	0.232	0.111	0.352	0.3
T=1250	0.421	0.41	0.431	0.358	0.231	0.411	0.38
Mean	0.18	0.19	**0.33**	0.12	0.08	**0.26**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, and $\phi_{f_2} = 0.3$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e is diagonal with $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u is diagonal with $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even.

Table B.10: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Heteroscedastic and cross correlated errors and strong signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.162	0.391	0.204	0.139	0.267	0.201	0.23
T=250	0.148	0.455	0.224	0.149	0.275	0.214	0.24
T=500	0.153	0.468	0.233	0.147	0.291	0.227	0.25
T=1250	0.167	0.473	0.242	0.152	0.311	0.219	0.26
N=50							
T=125	0.355	0.45	0.421	0.286	0.362	0.389	0.38
T=250	0.42	0.475	0.446	0.351	0.432	0.429	0.43
T=500	0.442	0.484	0.454	0.397	0.462	0.444	0.45
T=1250	0.46	0.49	0.459	0.429	0.477	0.453	0.46
N=100							
T=125	0.444	0.458	0.44	0.392	0.4	0.411	0.42
T=250	0.471	0.479	0.455	0.445	0.452	0.44	0.46
T=500	0.481	0.487	0.46	0.467	0.473	0.452	0.47
T=1250	0.488	0.492	0.464	0.48	0.486	0.459	0.48
N=200							
T=125	0.461	0.459	0.443	0.427	0.416	0.419	0.44
T=250	0.481	0.48	0.457	0.463	0.458	0.444	0.46
T=500	0.49	0.49	0.463	0.48	0.478	0.455	0.48
T=1250	0.494	0.495	0.466	0.49	0.49	0.462	0.48
Mean	0.38	**0.47**	0.4	0.36	**0.41**	0.38	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.9$, and $\phi_{f_2} = 0.6$, and errors η_t are independent $N(0,1)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Table B.11: Mean of the similarity measures between $\mathbf{P}$ and $\widehat{\mathbf{P}}$ when $r = 2$. Heteroscedastic and cross correlated errors and weak signal to noise ratio.

	PC	LY	CP	PC	LY	CP	Mean
N=10							
T=125	0.089	0.096	0.087	0.1	0.104	0.102	0.1
T=250	0.095	0.095	0.097	0.094	0.1	0.096	0.1
T=500	0.099	0.112	0.098	0.105	0.109	0.103	0.1
T=1250	0.091	0.112	0.095	0.103	0.098	0.101	0.1
N=50							
T=125	0.029	0.03	0.033	0.026	0.028	0.033	0.03
T=250	0.03	0.036	0.034	0.024	0.026	0.028	0.03
T=500	0.025	0.038	0.032	0.026	0.027	0.029	0.03
T=1250	0.028	0.076	0.034	0.027	0.027	0.032	0.04
N=100							
T=125	0.021	0.027	0.028	0.018	0.02	0.028	0.02
T=250	0.019	0.032	0.029	0.019	0.019	0.027	0.02
T=500	0.022	0.047	0.032	0.016	0.016	0.027	0.03
T=1250	0.022	0.177	0.036	0.018	0.018	0.028	0.05
N=200							
T=125	0.021	0.041	0.056	0.017	0.018	0.038	0.03
T=250	0.023	0.075	0.081	0.018	0.019	0.048	0.04
T=500	0.025	0.158	0.111	0.017	0.018	0.061	0.07
T=1250	0.033	0.358	0.162	0.016	0.018	0.081	0.11
Mean	0.04	**0.09**	0.07	0.04	0.04	**0.05**	

NOTES: Factor autoregressive coefficients $\phi_{f_1} = 0.6$, and $\phi_{f_2} = 0.3$, and errors η_t are independent $N(0, 0.5)$ random variables. Columns $1-3$ idiosyncratic covariance matrix Γ_e has diagonal elements $\sigma_{e_i} = 1$ if i is odd and $\sigma_{e_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{e_i}\sigma_{e_j}$. Columns $4-6$ idiosyncratic errors, e_{it}, present serial dependence such that $e_{it} = \theta e_{i,t-1} + u_{it}$ for $i = 1, ..., N/2$ with $\theta \sim N(0.5, 0.05)$, and Γ_u has diagonal elements $\sigma_{u_i} = 1$ if i is odd and $\sigma_{u_i} = 2$ if i is even, and non-diagonal elements $0.7^{|i-j|}\sigma_{u_i}\sigma_{u_j}$.

Appendix C

Chapter 5

C.1 Empirical

Table C.1: Fixed energy prices series included in Cluster 1.

Cluster 1
″CYP.Che″ ″CYP.Cons″ ″CYP.Food″ ″CYP.Iron″ ″CYP.Mach″ ″CYP.Min″ ″CYP.Nfer″
″CYP.Nmet″ ″CYP.Pap″ ″CYP.Tex″ ″CYP.Trans″ ″CYP.Wood″
″MEX.Che″ ″MEX.Cons″ ″MEX.Food″ ″MEX.Iron″ ″MEX.Mach″ ″MEX.Min″ ″MEX.Nfer″
″MEX.Nmet″ ″MEX.Pap″ ″MEX.Tex″ ″MEX.Trans″ ″MEX.Wood″
″USA.Che″ ″USA.Cons″ ″USA.Food″ ″USA.Iron″ ″USA.Mach″ ″USA.Min″ ″USA.Nfer″
″USA.Nmet″ ″USA.Pap″ ″USA.Tex″ ″USA.Trans″ ″USA.Wood″
″CHN.Che″ ″CHN.Cons″ ″CHN.Food″ ″CHN.Iron″ ″CHN.Mach″ ″CHN.Min″ ″CHN.Nfer″
″CHN.Nmet″ ″CHN.Tex″ ″CHN.Wood″
″GRC.Che″ ″GRC.Cons″ ″GRC.Food″ ″GRC.Iron″ ″GRC.Min″ ″GRC.Nfer″ ″GRC.Nmet″
″GRC.Pap″ ″GRC.Tex″ ″GRC.Trans″
″ROU.Che″ ″ROU.Cons″ ″ROU.Food″ ″ROU.Mach″ ″ROU.Min″ ″ROU.Nfer″ ″ROU.Nmet″
″ROU.Pap″ ″ROU.Tex″ ″ROU.Wood″
″NLD.Che″ ″NLD.Cons″ ″NLD.Food″ ″NLD.Mach″ ″NLD.Nmet″ ″NLD.Tex″
″FRA.Che″ ″FRA.Cons″ ″FRA.Min″ ″FRA.Nmet″
″CAN.Cons″ ″CAN.Min″ ″CAN.Tex″ ″CAN.Wood″
″PRT.Cons″ ″PRT.Food″ ″PRT.Min″ ″PRT.Nmet″
″DNK.Cons″ ″DNK.Min″ ″DNK.Nmet″
″FIN.Cons″ ″FIN.Trans″
″KOR.Iron″ ″KOR.Nmet″
″NZL.Cons″ ″NZL.Min″
″AUS.Cons″
″AUT.Cons″
″CZE.Cons″
″DEU.Cons″
″ITA.Nmet″
″POL.Cons″
″CHE.Cons″
″TUR.Mach″

Table C.2: Fixed energy prices series included in Cluster 2.

Cluster 2
”UK.Che” ”UK.Cons” ”UK.Food” ”UK.Iron” ”UK.Mach” ”UK.Nfer” ”UK.Nmet” ”UK.Pap” ”UK.Tex” ”UK.Trans” ”UK.Wood”
”CZE.Che” ”CZE.Food” ”CZE.Iron” ”CZE.Mach” ”CZE.Nfer” ”CZE.Nmet” ”CZE.Pap” ”CZE.Tex” ”CZE.Trans” ”CZE.Wood”
”POL.Che” ”POL.Food” ”POL.Mach” ”POL.Nfer” ”POL.Nmet” ”POL.Pap” ”POL.Tex” ”POL.Trans” ”POL.Wood”
”AUS.Che” ”AUS.Food” ”AUS.Iron” ”AUS.Nfer” ”AUS.Nmet” ”AUS.Pap” ”AUS.Tex” ”AUS.Wood”
”DEU.Che” ”DEU.Iron” ”DEU.Min” ”DEU.Nfer” ”DEU.Pap” ”DEU.Tex” ”DEU.Trans” ”DEU.Wood”
”SVK.Che” ”SVK.Cons” ”SVK.Iron” ”SVK.Min” ”SVK.Nfer” ”SVK.Nmet” ”SVK.Wood”
”DNK.Che” ”DNK.Iron” ”DNK.Nfer” ”DNK.Pap” ”DNK.Tex” ”DNK.Trans”
”JPN.Iron” ”JPN.Nfer” ”JPN.Nmet” ”JPN.Pap” ”JPN.Trans”
”TUR.Che” ”TUR.Min” ”TUR.Nfer” ”TUR.Pap” ”TUR.Tex”
”KOR.Che” ”KOR.Mach” ”KOR.Pap” ”KOR.Tex”
”NOR.Che” ”NOR.Cons” ”NOR.Min” ”NOR.Nmet”
”BEL.Che” ”BEL.Food” ”BEL.Nmet”
”ITA.Nfer” ”ITA.Pap” ”ITA.Trans”
”CAN.Iron” ”CAN.Nfer”
”FIN.Nfer” ”FIN.Pap”
”HUN.Iron” ”HUN.Nmet”
”CHE.Che” ”CHE.Pap”
”BRA.Iron”
”FRA.Iron”
”GRC.Wood”
”NLD.Iron”
”NZL.Nmet”

Table C.3: Fixed energy prices series included in Cluster 3.

Cluster 3
″SWE.Che″ ″SWE.Cons″ ″SWE.Food″ ″SWE.Iron″ ″SWE.Mach″ ″SWE.Min″ ″SWE.Nfer″ ″SWE.Nmet″ ″SWE.Pap″ ″SWE.Tex″ ″SWE.Trans″ ″SWE.Wood″
″FIN.Che″ ″FIN.Food″ ″FIN.Iron″ ″FIN.Mach″ ″FIN.Min″ ″FIN.Nmet″ ″FIN.Tex″ ″FIN.Wood″
″ITA.Che″ ″ITA.Cons″ ″ITA.Food″ ″ITA.Iron″ ″ITA.Mach″ ″ITA.Min″ ″ITA.Tex″ ″ITA.Wood″
″PRT.Che″ ″PRT.Iron″ ″PRT.Mach″ ″PRT.Nfer″ ″PRT.Pap″ ″PRT.Tex″ ″PRT.Trans″ ″PRT.Wood″
″FRA.Food″ ″FRA.Mach″ ″FRA.Nfer″ ″FRA.Pap″ ″FRA.Tex″ ″FRA.Trans″ ″FRA.Wood″
″CHE.Food″ ″CHE.Mach″ ″CHE.Min″ ″CHE.Nmet″ ″CHE.Tex″ ″CHE.Trans″ ″CHE.Wood″
″CAN.Che″ ″CAN.Food″ ″CAN.Mach″ ″CAN.Nmet″ ″CAN.Pap″ ″CAN.Trans″
″KOR.Cons″ ″KOR.Food″ ″KOR.Min″ ″KOR.Nfer″ ″KOR.Trans″ ″KOR.Wood″
″JPN.Che″ ″JPN.Cons″ ″JPN.Food″ ″JPN.Mach″ ″JPN.Min″ ″JPN.Tex″
″NLD.Min″ ″NLD.Nfer″ ″NLD.Pap″ ″NLD.Trans″ ″NLD.Wood″
″AUS.Mach″ ″AUS.Min″ ″AUS.Trans″
″DNK.Food″ ″DNK.Mach″ ″DNK.Wood″
″DEU.Food″ ″DEU.Mach″ ″DEU.Nmet″
″TUR.Cons″ ″TUR.Nmet″ ″TUR.Trans″
″HUN.Cons″ ″HUN.Min″
″ROU.Iron″ ″ROU.Trans″
″BEL.Cons″
″CHN.Trans″
″CZE.Min″
″GRC.Mach″
″NZL.Wood″
″POL.Min″
″UK.Min″

Table C.4: Fixed energy prices series included in Clusters 4, 5 and 6.

Cluster 4
”HUN.Che” ”HUN.Food” ”HUN.Mach” ”HUN.Nfer” ”HUN.Pap””HUN.Tex” ”HUN.Trans” ”HUN.Wood”
”NOR.Food” ”NOR.Iron” ”NOR.Mach” ”NOR.Pap” ”NOR.Tex” ”NOR.Trans” ”NOR.Wood”
”BEL.Iron” ”BEL.Mach” ”BEL.Min” ”BEL.Nfer” ”BEL.Pap” ”BEL.Tex”
”NZL.Che” ”NZL.Food” ”NZL.Iron” ”NZL.Nfer” ”NZL.Pap” ”NZL.Trans”
”SVK.Food” ”SVK.Mach” ”SVK.Pap” ”SVK.Tex” ”SVK.Trans”
”TUR.Food” ”TUR.Iron” ”TUR.Wood”
”CHE.Iron” ”CHE.Nfer”
”AUT.Food” ”AUT.Nmet”
”BEL.Trans” ”BEL.Wood”
”BRA.Nfer”
”JPN.Wood”
”POL.Iron”
Cluster 5
”AUT.Che” ”AUT.Iron” ”AUT.Mach” ”AUT.Min” ”AUT.Nfer” ”AUT.Pap” ”AUT.Tex” ”AUT.Trans” ”AUT.Wood”
”BRA.Cons” ”BRA.Food” ”BRA.Mach” ”BRA.Min” ”BRA.Pap” ”BRA.Tex” ”BRA.Trans” ”BRA.Wood”
”NZL.Mach” ”NZL.Tex”
”NOR.Nfer”
Cluster 6
”BRA.Che” ”BRA.Nmet”
”CHN.Pap”

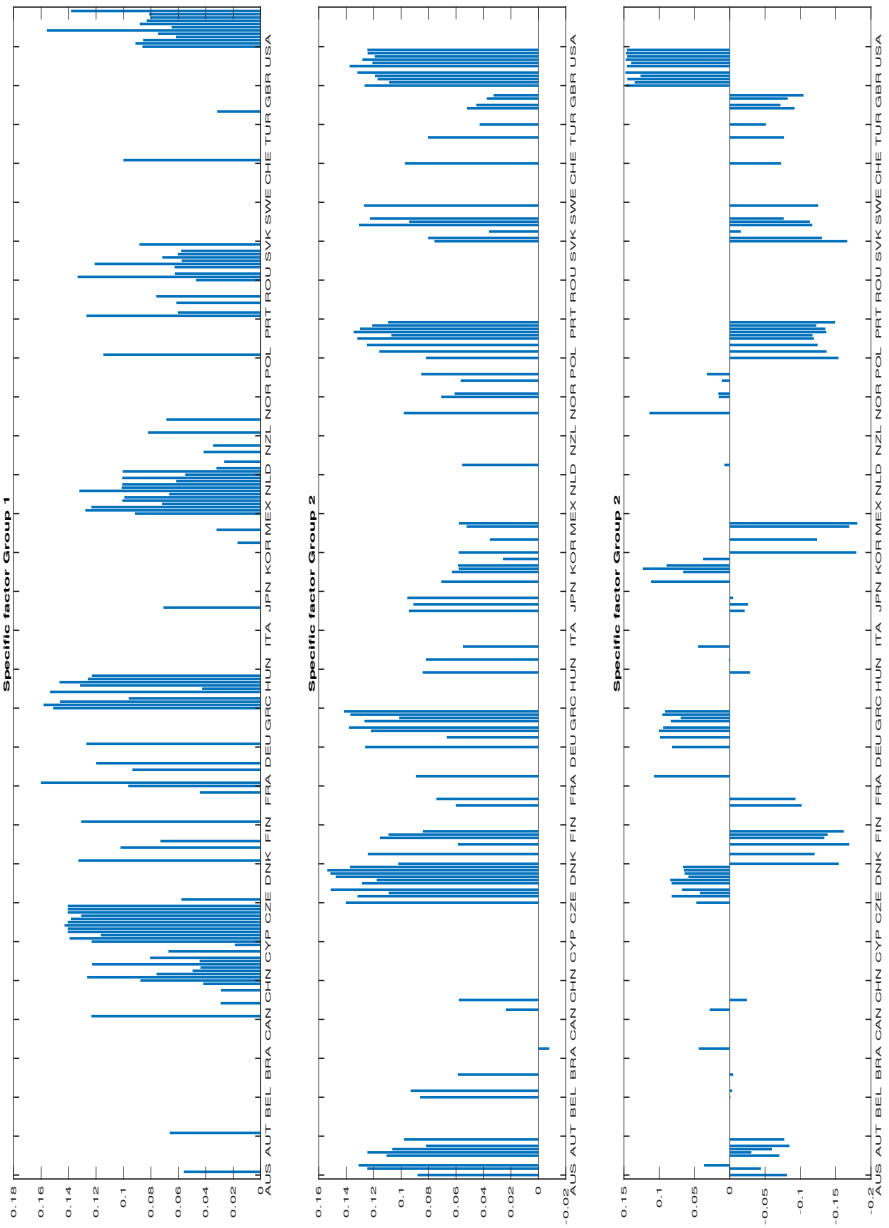

Figure C.1: Estimated loadings of the specific-factor in Cluster 1, first row, and the first and second specific-factors in Cluster 2, second and third row, respectively.

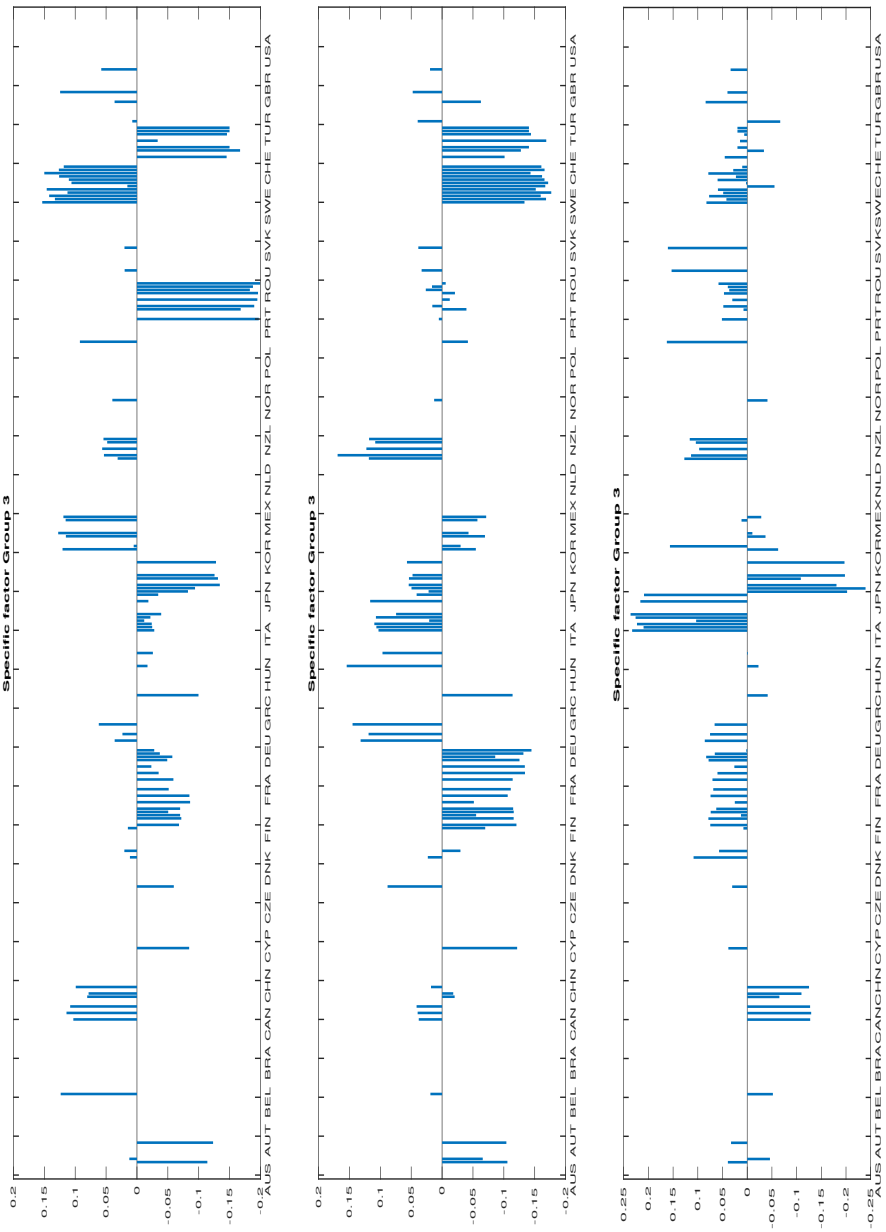

Figure C.2: Estimated loadings of the first, second and third specific-factors in Cluster 3.

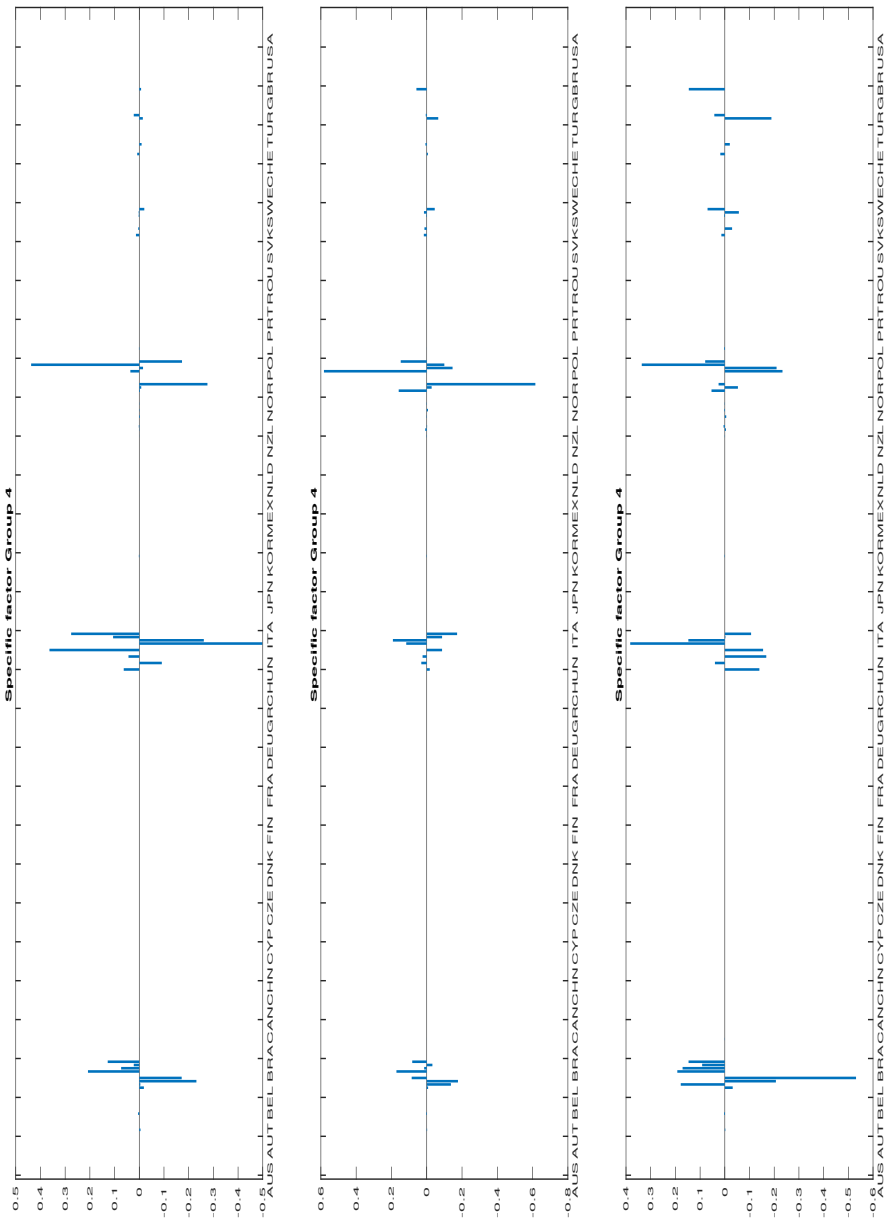

Figure C.3: Estimated loadings of the first, second and third specific-factors in Cluster 3.