ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF ECONOMICS

SECOND CYCLE DEGREE
IN
ECONOMICS AND ECONOMETRICS

Dynamic Matrix Factor Models and the EM Algorithm: A Nowcasting Framework for Mixed-Frequency Data in the Euro Area

Dissertation in Macroeconometrics

Defended by:
Davide Delfino

ID number:
1126595

Supervisor:
Prof. Matteo Barigozzi

Session 1st
Academic Year 2024-2025

Abstract

Recent literature in macroeconometrics has introduced Dynamic Matrix Factor Models (DMFMs) to address two major challenges in empirical analysis: managing the high dimensionality of modern datasets through low-rank latent structures, and enhancing forecast accuracy by explicitly modeling the dynamics of factors. The key innovation of these models, compared to traditional vector-based Dynamic Factor Models (DFMs), lies in their ability to preserve the original matrix structure of macroeconomic data.

This thesis presents a novel application of DMFMs for nowcasting matrix-variate time series, specifically focusing on the Gross Domestic Product (GDP) of major Euro Area (EA) countries. Nowcasts are generated through a pseudo real-time forecasting exercise, where the parameters of the DMFM are estimated via the Expectation-Maximization (EM) algorithm.

This empirical framework allows for handling (i) the matrix structure of the data, (ii) mixed-frequency datastes, and (iii) general patterns of missing data, such as those introduced by the COVID-19 disruptions.

Standard Dynamic Factor Models (DFMs) in vector form are used as a benchmark to evaluate the nowcasting performance of DMFMs, which incorporate cross-country information as an additional dimension. Empirical results show that the DMFM improves nowcasting accuracy both before and after the COVID-19 shock, and enhances responsiveness to monthly data updates, particularly during the COVID recession and recovery. Notably, Spain and France exhibit consistently better performance throughout the entire evaluation sample, while Germany and Italy show significant improvements particularly in the post-COVID period. These findings suggest that GDP nowcasts benefit most from the matrix-based approach in countries that are highly dependent for the Monetary Union economic status or during periods of intensified cross-country economic interdependence.

Keywords: Dynamic Matrix Factor Model; Expectation-Maximization Algorithm; Kalman Smoother; Missing Values; Mixed Frequency Data; Nowcasting; Euro Area;

Contents

1 Introduction

Recent literature in macroeconometrics is currently facing two major, interconnected, and demanding challenges: managing the high dimensionality of modern datasets and developing more sophisticated techniques to improve time series analysis and forecasting accuracy. This thesis contributes to this mission by providing an empirical application, based on Euro Area (EA) macroeconomic data, adopting recent methodologies developed in factor analysis in terms of the implementation of a new model, estimation methods, and forecasting strategies. Specifically, it addresses these challenges by extending traditional factor models to the lowest-order tensor data structure, namely matrix-variate time series, where observations are structured across three dimensions: units (rows), variables (columns), and time (slices). Moreover, the factors governing the data-generating process are allowed to evolve over time, resulting in a Dynamic Matrix Factor Model (DMFM). Model's parameters then estimated via Quasi Maximum Likelihood (QML) through the Expectation-Maximization (EM) algorithm, a flexible strategy that is also convenient for handling general patterns of missing data. Given the integration of Kalman filtering techniques, this strategy also enables a nowcasting application, which represents the true and novel contribution presented in this work. Indeed, the empirical exercise consists of nowcasting EA country-specific GDP by exploiting all the information that the model can retrieve from other countries and selected variables, preserving the matrix structure in which data are collected in the European Monetary Union. The nowcasting results obtained using the DMFM are systematically compared, for each EA country, with those generated by the standard vector-based Dynamic Factor Model (DFM), demonstrating the superior performance of the matrix-based approach in terms of both forecast accuracy and responsiveness to the information flow.

This chapter introduces the Dynamic Matrix Factor Model (DMFM), describing its theoretical foundations and briefly reviewing the key methodological contributions in the literature that the present work uses as reference. After discussing the main advantages that the matrix-based specification offers for economic analysis, particularly in capturing cross-sectional dynamics across EA countries, I turn to the estimation procedure via QML, highlighting the flexibility of the Expectation-Maximization (EM) algorithm in handling complex data structures and missing observations. The chapter concludes by presenting an original and practical application of the DMFM to nowcasting on EA data, an estension to the traditionally nowcasting procedure with vector-based Dynamic Factor Models (DFMs). I describe the implementation strategy, report comparative real-time forecasting results, and offer an interpretation of the findings.

Beginning with their historical development, matrix-variate time series have recently attracted increasing interest in applied macroeconomic and financial research. In these fields, it is common to collect data as a sequences of two-dimensional array evolving over time. Consider, for instance, macroeconomic indicators observed across multiple countries, or asset returns and volatilities across portfolios. Several methodologies have been proposed to analyze such data structures, including matrix autoregressive models (e.g.,R. Chen, Xiao, and Yang 2021, Billio et al. 2023), matrix panel regression models (e.g., Kapetanios, Serlenga, and Shin 2021), and matrix factor models (e.g., Wang, Liu, and R. Chen 2019, L. Yu et al. 2022, or Matteo Barigozzi and Trapin 2025). This thesis specifically explores the last class of the listed models.

Matrix Factor Models (MFMs) preserve the original structure of the data as collected, tracking the same variables over time across different economic entities, while simultaneously enabling substantial dimensionality reduction through factor analysis, a particularly desirable property in high-dimensional settings. Moreover, by introducing some dynamic in the latent factors, these models can also be adopted for forecasting purposes. Indeed, factors may evolve over time according to a matrix autoregressive (MAR) process. Indeed, during the last five years the literature extended from static model for matrix-valued time series to the so-called Dynamic Matrix Factor Model (DMFM). This model can outperform traditional vector-based Dynamic Factor Models (DFMs) as it account for a supplementary dimensionality. On one hand, DFMs typically applied to two-dimensional datasets within a multivariate time series framework, namely a set of variables observed over time, on the other DMFMs can jointly consider data collected in a matrix way repeatedly over time. This thesis provides a comparison between these two class of models in the context of nowcasting, resulting in one of the first application of this real-time forecasting on tensor data.

Before introducing the estimation methods and the nowcasting framework, it is worth stressing how preserving the matrix structure of the data, rather than vectorizing it, offers significant advantages. Indeed, even if it is a common practice to turn a matrix into a vector and apply standard vector time series analysis, it can be the case that the structural relationships embedded in the matrix get lost. In contrast, matrix factor models retain the original structure and exploit all the information that data collected in matrix form can convey. MFM can enhance the interpretability of latent factors, reduce the number of parameters to estimate, and more effectively capture dependencies and commonalities across rows (e.g. countries) and columns (e.g. variables). This is particularly important when dealing with datasets in which rows and columns represent distinct but structurally connected entities, as for instance in our empirical example where economic indicators and EA countries can represent different layers of commonalities. To get convinced on how matrix data are the most general case possible, consider the following Matrix-valued time series dataset for a specific time t:

Country	**GDP Growth**	**Unemployment Rate**	**...**	**Industrial Production**
Germany	$X_{t,11}$	$X_{t,12}$	...	$X_{t,1p}$
France	$X_{t,21}$	$X_{t,22}$	...	$X_{t,2p}$
Italy	$X_{t,31}$	$X_{t,32}$	...	$X_{t,3p}$
Spain	$X_{t,41}$	$X_{t,42}$	...	$X_{t,4p}$

Table 1: Matrix-Valued Time Series Dataset

Table 1 represents just a slice of a matrix-valued time series for a given time span $t = 1 \dots T$. This structure generalizes classical univariate and multivariate time series models. For instance, a univariate time series describes the temporal evolution of a single variable (i.e., a single entry of the table repeated over time), while multivariate or panel consider vectors of variables observed jointly for a single unit (e.g., a row observed over time). However, in many empirical contexts, as in the economic and financial examples mentioned above, the data are more appropriately structured as matrices evolving over time. Within this framework, it is reasonable to assume that raws, such as countries belonging to the same monetary union (for instance, in the European Monetary Union) share common features and can be

convenient implement matrix factor models to provide a more general, and informative framework for modeling such data jointly.

The DMFM considered in this thesis represents an evolution of the matrix-based factor models, namely the Matrix Factor Models (MFM), proposed by Wang, Liu, and R. Chen 2019. In the orginal framework data, observed at each time t, are treated in a matrix $Y_t \in \mathbb{R}^{p_1 \times p_2}$, represented as:

$$Y_t = RF_tC^\top + E_t, \tag{1}$$

This specification preserves the bilinear structure in both the common and idiosyncratic components. Specifically, $F_t \in \mathbb{R}^{k_1 \times k_2}$ is a matrix of latent factors; $R \in \mathbb{R}^{p_1 \times k_1}$ and $C \in \mathbb{R}^{p_2 \times k_2}$ are the row and column loading matrices, respectively. finally E_t is the idiosyncratic error term. In this thesis, E_t is allowed to exhibit both temporal and cross-sectional dependence across rows and columns. This class of models is referred to as approximate factor models, since it allows the idiosyncratic components to exhibit some weak cross-sectional dependence. In other words, this allows the idiosyncratic components, originally caracterized by white noise, and thus with a diagonal structure, to exhibit some correlation accounting for some limits. Indeed, having this component highly correlated would means to have some other factors that could be considered and this is the reason why the selection of the correct number of factors is a first relevant step in the empirical analysis.

Although Equation (1) defines a static factor structure, the model can be extended to a dynamic setting allowing latent factors to evolve over time. The problem can consequently be specified with a state-space representation where the measurement equation is given by Equation (1), and the transition equation for the latent factors, given by Equation (2), is assumed to follow a Matrix Autoregressive (*MAR*) process, as proposed by R. Chen, Xiao, and Yang 2021. *MAR* processes are a matrix-valued extension of the traditional *VAR* model and, for the latent variables, takes the form of:

$$F_t = \sum_{i=1}^{P} A_i F_{t-i} B_i^\top + U_t, \tag{2}$$

Equation (2) describes the temporal dynamics of the latent factors accounting for the bilinear structure for rows and columns in the autoregressive terms. For a generic order $i = \{1 \dots P\}$, A_i and B_i are coefficient matrices respectively for rows and columns. In this thesis, without loss of generality, the lag effect is assumed to last one period, namely $i = 1$.

Building on the original contribution of Wang, Liu, and R. Chen 2019, the literature also proposed the introduced of more general advancements, particularly regarding estimation methods. The first is the Projected Estimation method developed by L. Yu et al. 2022. This approach, based on Principal Component (PC) techniques, is employed for the initial estimation of the row and column loading matrices in the DMFM. As a PC-based method, it requires an imputation step in the presence of missing values. The imputation strategy adopted follows the methodology proposed by Cen and Lam 2025, which extends the "all-purpose estimator" of Xiong and Pelger 2023 to matrix-valued time series. The second and principal methodological contribution is the estimation framework by Matteo Barigozzi and Trapin 2025, which generalizes the EM algorithm, originally developed for vector factor models by Doz, Giannone, and Reichlin 2012, and later adapted to missing data settings by Bańbura and Modugno

2014, to the context of high-dimensional DMFMs. Within this framework, the projected estimator is used as the initialization step of the EM algorithm. It is also worth noting that the vector factor model, along with the estimation and forecasting strategy proposed by Bańbura and Modugno 2014, is used in this thesis as a benchmark to assess the performance of the DMFM in terms of nowcasting. The framework developed by Matteo Barigozzi and Trapin 2025 is in fact based on the same EM algorithmic structure as that of Bańbura and Modugno 2014, and it is consequently appealing to compare the performance of factor models in vector and matrix form by the results achieved using these two models for nowcasting.

Nowcasting, even if is a practical and increasingly popular method, adopted especially in Central Banks, at the best of the author's knowledge, was never adopted in the context of DMFM. Through this real-time forecasting strategy it is possible to exploit the different frequency of variables and the staggered release of economic indicators. It basically allows to manage jagged edge datasets. Through this approach one can produces real-time estimates of low-frequency targets, such as GDP, by incorporating more timely, high-frequency variables. The method is particularly effective in this setting, because by reproducing a real-time flow of information, within a mixed-frequency framework, where monthly variables are released at different times within a quarter, it is possible to track the monthly contribution of monthly indicators on nowcasts updates through the Kalman Filter. The nowcasting framework was originally proposed by Giannone, Reichlin, and Small 2008, who built a large factor model estimated via principal components and updated through a Kalman smoother. While nowcasting is now standard practice, its integration with tensor, or matrix-structured data, remains relatively recent and underexplored. The Rolling-Nowcast proposed in this thesis is a variant of the one in Giannone, Reichlin, and Small 2008 and extends the method by Bańbura and Modugno 2014 to the matrix case exploiting the EM algorithm.

Based on the methodological framework described above, this thesis proposes an empirical application using the "EA-MD-QD: Large Euro Area and Euro Member Countries Datasets for Macroeconomic Research" by Matteo Barigozzi, Lippi, and Luciani 2021, which includes both monthly and quarterly macroeconomic indicators from January 2000 to january 2025, for a total of $T = 300$ time periods. The empirical analysis focuses on Germany, France, Italy, and Spain, using 39 monthly variables and quarterly GDP. [1]. To implement the nowcasting exercise, the dataset is recursively truncated from the first quarter of 2017 (Q1:2017) to the first quarter of 2025 (Q1:2025), corresponding to the most recent available observation at the time of the implementation. At each iteration, the model is estimated by applying a masking procedure that replicates the frequency and the actual release schedule of each variable, simulating a pseudo-real-time flow of information. GDP is then nowcasted for each country using: (i) all available information across countries in the matrix-based approach, or (ii) only country-specific information in the vector-based model. The model is re-estimated at each step, and the final variable values,computed using the Kalman smoother, are retained as nowcasts, since they correspond to the filtered predictions of the Kalman filter. The estimation of the DMFM at each step is performed using the Expectation-Maximization (EM) algorithm, which is selected over alternative methods (discussed in Chapter 2) due to its strong suitability for handling missing data and

[1]The accompanying R replication code, available at https://github.com/dolpolo/Nowcasting_EA_DMFM, allows full reproducibility and flexible selection of countries and variables up to the limit of the reference dataset.

for enabling real-time nowcasting. In this way it is possible to work with mixed-frequency structures and the presence of outliers during the COVID-19 period, when real variables, the most affected, are masked. The EM algorithm is particularly appropriate in this context not just because it accommodates general missing data patterns, but also because, as an iterative approach, it allows to find a way out to a quasi-log likelihood with no closed-form solution.

To sum up, the application of nowcasting to GDP in Euro Area countries using DFMs in both vector and matrix form, addresses two central research questions:

- Does modeling data in matrix form improve forecasting accuracy at the country level?
- Does incorporating high-frequency information enhance GDP predictions?

Comparative results show that nowcasting GDP for Euro Area countries using a medium-sized set of monthly variables within a DMFM framework outperforms the vector-based approach, particularly in terms of accuracy and responsiveness during adverse phases of the business cycle. This advantage is especially evident for countries that are structurally more dependent on the common dynamics of the Euro Area, such as Spain and France. For Germany and Italy, by contrast, the DMFM provides significantly better forecasts particularly in the post-COVID period. This pattern suggests that the more a country is economically integrated within the Monetary Union, especially during times of crises and recoveries or coordinated monetary policy, the greater the benefits of adopting a matrix-based modeling strategy. Nevertheless, the vector-based DFM remains a valid alternative for computationally reasons.

The main contribution of this thesis lies in the empirical implementation of recent advanced models, estimation techniques, and forecasting strategies, as it:

- accommodates structured macroeconomic datasets collected across comparable observational units by means of matrix-variate time series models;
- handles complex patterns of missing data arising from mixed frequencies, irregular release schedules, or disruptions such as the COVID-19 pandemic, through the implementation of the EM algorithm for parameter estimation;
- supports nowcasting, representing one of the first applications of the nowcasting framework to matrix-variate time series data.

Structure of the Thesis: The remainder of this work is structured as follows. Chapter 2 reviews the origins of factor models and their application to matrix-valued time series data, discussing estimation strategies proposed in the literature and motivating the methodological choices adopted in this thesis for DMFMs. Chapter 3 presents the adopted methodological framework, introducing the dynamic matrix factor model and comparing it to its vector-based counterpart. The discussion focuses specifically on the QML estimation procedure via the EM algorithm, including initialization steps and its adaptation for nowcasting purposes. Chapter 4 introduces the empirical dataset, evaluates and compares the nowcasting performance of the models, and discusses the results suggesting also future works. The code is available in `MATLAB` for the vector-based Dynamic Factor Model and in `R` for the DMFM. Finally, Chapter 5 summarizes the key findings and outlines directions for future research.

2 Literature Review

This initial chapter reviews the relevant literature underlying this work, including alternative model specifications and estimation strategies. The focus is primarily on matrix factor models, potentially augmented with matrix autoregressive (MAR) dynamics for the latent factors.

Traditional Dynamic Factor Models (DFMs) have recently been extended to accommodate matrix-valued time series, leading to the development of Dynamic Matrix Factor Models (DMFMs). Estimation methods originally proposed for vector-valued factor models can often be adapted to the matrix setting. Among these, this work adopts Quasi-Maximum Likelihood Estimation (QMLE) via the Expectation-Maximization (EM) algorithm as the preferred approach. This estimation strategy effectively addresses the challenges posed by high dimensionality and the presence of mixed-frequency data (or in other words the presenence of missing observations), offering a high degree of modeling flexibility. Moreover, its integration with Kalman filtering techniques makes it particularly well-suited for nowcasting applications.

Dynamic Matrix Factor Models (DMFMs) represent a recent advancement in factor analysis, one of the most widely used techniques in unsupervised statistical learning. As a multivariate method, factor analysis is particularly effective in high-dimensional settings, where the number of time series n may be large relative to the number of time periods T. The core idea is to model the n observed time series as linear combinations of a small number $r \ll n$ of latent factors. Each factor influences the observed variables through a corresponding loading, while the idiosyncratic component captures the portion of variability not explained by the common structure. This leads to substantial dimensionality reduction.Factor models are powerful tools for analyzing co-movements in large datasets, identifying structural shocks, and forecasting macroeconomic aggregates—all central objectives of this work. By incorporating a dynamic structure into the evolution of the latent factors, these models can be successfully applied to forecasting tasks. DMFMs inherit all the strengths of traditional factor models—robustness in high-dimensional contexts and predictive capacity—but extend their applicability to matrix-valued time series. This additional layer of structure allows DMFMs to capture both cross-sectional and temporal interdependencies more effectively, thus offering a novel contribution to the existing literature.

Indeed, the most recent literature in econometrics has proposed a historical transition from vector to matrix factor models, particularly suited for observations that are naturally structured in matrix form, such as panels of macroeconomic indicators collected across countries over time. A pioneering contribution to this line of research is provided by Wang, Liu, and R. Chen 2019, who introduced a matrix-valued factor model in which each observation is linked to both row and column latent factors. The estimation of the Matrix Factor Models (MFM) was provided by the eigen-decomposition of the long-run covariance matrix. This method was then also applied in E. Y. Chen, Tsay, and R. Chen 2020, who further propose a general framework that incorporates prior knowledge or domain constraints into the model through linear restrictions.

Although the great intuition underlying this advancements in factor analysis, the original framework still poses two restrictive assumptions: the independence of the idiosyncratic component both in time and cross-sectionally, and a static structure of latent factors.

As a solution to the former, L. Yu et al. 2022 and E. Y. Chen and Fan 2023 proposed Projected Estimators that allow for autocorrelated idiosyncratic components. Their two-step methodology first estimates the number of factors and the loading matrices using principal components, and then projects the data accordingly to produce a significant dimensionality reduction. From this projection they recover through iterative algorithms the correct number of factors, based on the eigenvalue-ratio and a most effective estimates of loadings both for rows and columns. This approach is especially relevant here, as it is used to initialize the EM algorithm in our setup and provides a first level of generalization beyond the framework proposed by Wang, Liu, and R. Chen 2019. More recently Xu, Yuan, and Guo 2025 developed Q-MLE techniques specifically tailored for matrix factor models. These methods allow for heteroskedastic idiosyncratic components and avoid the need for a fully specified dynamic structure in the factors, thus enhancing robustness and feasibility in empirical applications.

Sequently, to introduce temporal dependence, a valuable feature for forecasting purposes, R. Chen, Xiao, and Yang 2021 proposed an adaptation of vector processes for time series to matrix data. They provided a generalization of traditional vector autoregressive techniques to the matrix setting, maintaining the data's structure and preserving bilinearity. As in the MFM proposed by Wang, Liu, and R. Chen 2019, this approach achieves substantial dimensionality reduction and enhanced interpretability, although without involving latent factors. It should be noted, however, that matrix autoregressive (MAR) models are not always the most appropriate specification. As shown by Tsay 2024, while MAR models are convenient for matrix-variate time series analysis, matrix moving average (MMA) models may perform better in some contexts, such as modeling matrix-valued seasonal time series. In that context, Tsay 2024 extends traditional VARMA models to the matrix setting and proposes maximum likelihood estimation based on recursive innovations.

More recently, the theory of factor analysis for matrix-valued data has drawn inspiration from these advances and from the literature on dynamic factor models in vector form, which suggest that temporal dynamics can be incorporated via a state-space representation. As a result, the matrix factor models (MFMs) discussed above have been extended by allowing the latent factors to follow a dynamic process, typically specified as a matrix autoregressive (MAR) model. The resulting model is referred to as a Dynamic Matrix Factor Model (DMFM), and it has been formally studied in recent contributions by R. Yu et al. 2024 and Matteo Barigozzi and Trapin 2025. The latter provides the main reference for this thesis, as the estimation strategy employed here is directly inspired by their framework. They estimate DMFMs using an adapted EM algorithm for the matrix case. The initialization relies on the static projected estimator from L. Yu et al. 2022, while the EM algorithm itself generalizes the frameworks proposed by Doz, Giannone, and Reichlin 2012 Bańbura and Modugno 2014. In particular, Doz, Giannone, and Reichlin 2012 demonstrate that the EM estimator for DFMs in vector form remains robust to mis-specification in the cross-sectional and serial correlation of idiosyncratic components using the Kalman smoother. Meanwhile, Bańbura and Modugno 2014 extend the EM algorithm—originally proposed by Watson and Engle 1983 for DFMs, to cases with arbitrary patterns of missing data and serially correlated idiosyncratic components. Their framework efficiently handles datasets with varying publication delays, mixed frequencies, and different sample lengths. It is well suited to tasks such as interpolation, backcasting (e.g., for emerging economies), and nowcasting. A further contribution of Bańbura and Modugno 2014 is the introduction of a model-based approach to

quantify the impact of data releases (referred to as "news") on forecast revisions. This mechanism allows for interpreting nowcasting updates in terms of sign and magnitude, and helps to decompose the contribution of individual releases to changes in the forecast—especially useful when multiple indicators are published simultaneously.

Summing up, the approached followed by Matteo Barigozzi and Trapin 2025, that inspires this work is actually a generalization to the matrix case of Bańbura and Modugno 2014. Moreover, a distinctive feature of the approach by Matteo Barigozzi and Trapin 2025, in comparison to R. Yu et al. 2024, regardless the treatment of missing data patterns, is that their estimation technique integrates Kalman filtering within the EM framework, thereby enabling the model to accommodate unbalanced panels and mixed-frequency data. Handling missingness in high-dimensional time series is a recurring challenge, and addressing it adds a significant layer of generality to empirical analysis. The EM algorithm, being iterative, requires parameter initialization. This is provided via projected estimates, which are based on Principal Component (PC) methods. Since PC approaches are not robust to missing data, an initial imputation step is required. This thesis follows the imputation strategy proposed by Cen and Lam 2025, which extends the "all-purpose estimator" by Xiong and Pelger 2023 to the tensor setting.

If not mis-specified, the log-likelihood estimation via Maximum-Likelihood of an approximate DMFMs, as specified in this work, would be computationally infeasible. Chapter 3 discusses this issue in detail, drawing from the approach of Matteo Barigozzi and Trapin 2025. The high dimensionality and complexity of DMFMs make the full maximum likelihood estimation unfeasible. To address this issue, they adopt a quasi-likelihood framework inspired by Tipping and Bishop 1999 in the context of probabilistic PCA. As in Doz, Giannone, and Reichlin 2012, the idiosyncratic components are treated as approximately i.i.d., allowing for a tractable and robust estimation process via the EM algorithm.

Finally, since the core of this dissertation lies in the empirical comparison of forecasting performance between vector and matrix formulations of Dynamic Factor Models (DFMs), I briefly review the nowcasting methodology. Nowcasting—the forecasting of the present or very near future—has become increasingly relevant for central banks and policymakers. The foundational work by Giannone, Reichlin, and Small 2008 introduced a two-step procedure that combines principal components with Kalman filtering to manage unbalanced data releases in real time.

In this thesis, I adopt the nowcasting strategy by Bańbura and Modugno 2014 adapted to matrix-valued data. In the vector case they reconstruct the full dataset using factors and extract the implied nowcasts directly from the smoothed estimates. This strategy enables a recursive update of GDP nowcasts month by month within each quarter, reflecting the real-time flow of incoming information.

While the majority of nowcasting applications to date have employed vector-valued models, the matrix formulation offers a novel advantage: it allows for the decomposition of nowcast revisions by both variable (columns) and cross-sectional unit (rows). This multidimensional structure opens the door to more granular analyses of informational content and shock transmission mechanisms across countries and indicators. This field remains underexplored and holds promise for future applications.

3 Methodology

This chapter outlines the methodology underlying the empirical analysis on Euro Area data presented in Chapter 4. It discusses how The data can be modeled using a Dynamic Matrix Factor Model (DMFM), the parameter estimation via Quasi Maximum Likelihood (QML) through the Expectation-Maximization (EM) algorithm, the extention to handle general patterns of missing values and the implementation of a recursive nowcasting procedure.

As a recent advancement in macroeconometrics, this chapter first provides a comparison between Dynamic Factor Models (DFMs) in the traditional vector and matrix in Subsection 3.1.1. Using as reference Matteo Barigozzi and Trapin 2025 I discuss how the matrix structure is capable of capturing the main dynamics of time series data by accounting for cross-sectional dependence across both rows and columns over time.

The estimation of the mis-specified log-likelihood for the DMFM is presented in Subsection 3.2, with particular focus on the nature of the mis-specification in Subsubsection 3.2.1, and on the EM algorithm implementation in Subsubsection 3.2.2. As an iterative procedure, the EM algorithm overcomes the lack of a closed-form solution for the quasi log-likelihood. However, it requires both an initialization strategy, discussed in Subsubsection 3.2.3, and a convergence rule, outlined in Subsubsection 3.2.4. The framework is then extended to handle missing data, as explained in Subsection 3.3. Finally, the integration of the EM algorithm within a recursive nowcasting setting is discussed in Subsection 3.4.

On these solid theoretical foundations, the pseudo real-time nowcasting of GDP for Euro Area countries is built and presented in the empirical chapter.

3.1 The Model

Nowadays, the high dimensionality of datasets represents a significant challenge for central banks and researchers. Factor models are one of the main techniques to address this issue as they summarize the large amount of information through a reduced set of unobserved latent components, namely the factors.

In this work, I consider two different specifications of factor models, accounting also for their dynamics over time: the traditional Dynamic Factor Model (DFM) in vector form, and its generalization to the matrix case, namely the Dynamic Matrix Factor Model (DMFM). For practitioners, the way data are collected can strongly influence the choice between these two classes of dynamic factor models. In particular, as addressed in this work, even if these models share common features and can be estimated using analogous techniques, the matrix formulation may offer specific advantages. Indeed, it is becoming increasingly appealing in macroeconomic and financial contexts, where data are often collected as a sequence of variables for multiple units observed over time.

In this initial methodological section, I describe the DFM by comparing its vector and matrix formulations, emphasizing first the strengths of the more general matrix case, the DMFM, and then providing a brief review of the standard vector-based approach.

3.1.1 Dynamic Matrix Factor Model

Consider a Matrix Factor Model (MFM) for the $p_1 \times p_2$ centered and standardized matrix-valued stationary process $\{Y_t\}$. Without loss of generality, assume that the latent factors follow a Matrix Autoregressive (MAR) process of order $P = 1$. Formally, for any $t \in \mathbb{Z}$, the model is given by:

$$Y_t = RF_tC^\top + E_t, \tag{3}$$

$$F_t = AF_{t-1}B^\top + U_t, \tag{4}$$

According to this equation system at each time t, data collected in matrix form, can be treated withing a factor model by preserving their bi-dimensionality over time, namely $Y_t \in \mathbb{R}^{p_1 \times p_2}$. The state space representation consists in a measurement equation, namely Equation (3), corresponing to the static Matrix Factor Model proposed by Wang, Liu, and R. Chen 2019, and in a transition equation, namely Equation (4), describing the Matrix Autoregressive (MAR) process introduced by R. Chen, Xiao, and Yang 2021. The key idea is to retrieve the observed variables as a combination of unobserved factors, which are also allowed to evolve over time through a dynamic process. In high-dimensional settings, this strategy can succesfull, as the latent structure enables the condensation of a large amount of information into a few common components, mitigating the risk of incurring in the so-called "curse of dimensionality". More formally, consider Equation (3) as a data-generating process consisting of a low-rank common component $\chi_t = RF_tC^\top$ and an idiosyncratic component $E_t \in \mathbb{R}^{p_1 \times p_2}$. On one hand, the common component χ_t is composed by $F_t \in \mathbb{R}^{k_1 \times k_2}$, the latent factor matrix, with $k_1, k_2 < \min(p_1, p_2)$, which drives the evolution of the matrix through row and column loading matrices $R \in \mathbb{R}^{p_1 \times k_1}$ and $C \in \mathbb{R}^{p_2 \times k_2}$, respectively. The loading matrices capture how much the factor impacts each row and column. This structure enhance the interpretability as loadings convey the identification of the main sources of variablity. In other words, through this procedure it is straightforward to assign a name to the unobserved latent variable by analyzing their impact on loading matrices. On the other hand, the idiosyncratic component $E_t \in \mathbb{R}^{p_1 \times p_2}$ is characterized by a row covariance matrix $H \in \mathbb{R}^{p_1 \times p_1}$ and a column covariance matrix $K \in \mathbb{R}^{p_2 \times p_2}$. Note that even if no explicit dynamic model is assumed for the evolution of E_t, such extensions are possible. For consistent parameter estimation, it is assumed that $\{F_t\}$ and $\{E_t\}$ are uncorrelated at all leads and lags. It is worth noting that the idiosyncratic component will play a central role in estimation, particularly during the Log-Likelihood specification. Considering the full structure of the covariance matrices the estimation is practically unfeasible, however the DMFM discussed so far is an *approximate* factor model, as it allows for full (non-diagonal) covariance matrices H and K.

A similar structure applies also to Equation (4), which models the temporal dynamics of the latent factors. Indeed, the transition equation introduces time dependence among the factors F_t through the Matrix Autoregressive (MAR) structure. Also in this equation the bilinear form is preserved, with $A \in \mathbb{R}^{k_1 \times k_1}$ and $B \in \mathbb{R}^{k_2 \times k_2}$ denoting the autoregressive coefficient matrices. $U_t \in \mathbb{R}^{k_1 \times k_2}$ represents the innovation matrix, which has row covariance matrix $P \in \mathbb{R}^{k_1 \times k_1}$ and column covariance matrix $Q \in \mathbb{R}^{k_2 \times k_2}$.

The bilinear structure, expressed through the loading matrices R and C, and the covariance matrices H

and K in the measurement equation, along with the autoregressive matrices A and B and the innovation covariance matrices P and Q in the transition equation, is the reason why the matrix representation of the Dynamic Factor Model can be seen as a generalization of the vector case. In fact, when the dimensionality is reduced to a single row or column, the matrix formulation simplifies to the vector version. Accounting for the matrix structure is especially appealing for macroeconomic and financial applications, as this methodology allows one to account for both time and cross-sectional dependence across rows and columns jointly, while also preserving the natural structure in which the data are collected.

The Vectorization: The DMFM presented above can be vectorized, as described in Section 7.2 in Durbin and Koopman 2012, to facilitate the estimation of the parameters via Quasi Maximum Likelihood (*QML*), by maximizing the prediction error decomposition of the Gaussian likelihood obtained from the Kalman filter. Unfortunately, the non trivial cost of vectorization of the DMFM is loosing the bilinear structure, that is exactly what this work aims to address and exploit. Moreover this yield a significant increase in the number of parameters to estimate, that would make the estimation as the data becomes larger. To show this let's consider the Vectorized form of the DMFM presented in Equations 3 and 4:

$$
\begin{aligned}
y_t &= (C \otimes R) f_t + e_t, \\
f_t &= (B_1 \otimes A) f_{t-1} + u_t,
\end{aligned}
$$

where $y_t = \text{vec}(Y_t)$ is the vectorized observed data at time t, resulting from f_t, the vectorized matrix factor at time t. $C \otimes R$ is the Kronecker product of the column and row loading matrices, and $B_1 \otimes A$ is the transition matrix. Seemingly, $K \otimes H$ is the measurement error covariance matrix in e_t, and $Q \otimes P$ is the state error covariance matrix in u_t.

While the DMFM described above generates a total number of row and column loading parameters equal to $p_1 k_1 + p_2 k_2$, the corresponding vectorized DMFM requires estimating a full loading matrix of dimension $(p_1 p_2) \times (k_1 k_2)$. This loss of the bi-dimensional structure reduces interpretability and makes estimation more challenging, especially in high-dimensional settings. However, this vectorized form of the DMFM, as I will explain more formally later on, is particularly useful during some steps of the estimation procedure. In particular, during the Kalman Filter procedure in the E-step (paragraph 3.2.2), the vectorized form of the MAR parameters is used. This adaptation does not compromise the results, since the focus is not on the autoregressive parameters and the related innovation covariances. Nevertheless, it is useful for running the Kalman smoother without resorting to matrix-based versions of the Kalman filter for matrix state-space models, which offer only negligible computational advantages. Therefore, vectorization will be used just to exploit the estimators obtained from the vectorized MAR, as in Matteo Barigozzi and Trapin 2025 and R. Yu et al. 2024.

DMFM Identification: A critical aspect of the Matrix Factor Model, as introduced by Wang, Liu, and R. Chen 2019, concerns its identifiability, that is, the impossibility of uniquely estimating the

single parameters diven the data without imposing additional restrictions. Identifiability is not an issue for forecasting purposes, because, as also noted by Bańbura and Modugno 2014, in such cases the primary interest is identifying the space spanned by the latent factors rather than uniquely estimating the individual factors themselves. In other words, the objective is estimating the product that, regardless the single parameters, produce the same combination. Nevertheless, addressing identifiability is essential to offer a rigorous theoretical understanding of the model and to provide the general overview of the estimation procedures proposed by R. Yu et al. 2024, used in this work to find the correct number of factors and initial parameters used for the EM-Algorithm.

The identifiability issue specifically affects the loading matrices R and C, and the factor matrix F_t, in the Matrix Factor Model (Equation (3)). These parameters are not separately identifiable because the product $RF_tC^\top$ remains unchanged under transformations involving invertible matrices $W_1 \in \mathbb{R}^{k_1 \times k_1}$ and $W_2 \in \mathbb{R}^{k_2 \times k_2}$, as shown below:

$$RF_tC^\top = (RW_1)(W_1^{-1}F_tW_2^{-1})(CW_2)^\top.$$

A complete ambiguity underlies the identifiability problem as both the row and column loading matrices can be arbitrarily transformed using invertible matrices. In particular, for any invertible matrix W_1 the transformation on R does not change the product since $RW_1 \cdot W_1^{-1} = R$, and if W_2 is orthogonal, i.e., $W_2^\top = W_2^{-1}$ also the transformation of C leaves the product unchanged. This implies that further constraints are needed to uniquely identify the expression $RF_tC^\top$ from the observed data, as multiple combinations of R, F_t, and C can produce the same observed matrix.

A common solution to this issue in the MFM framework is to enforce identifiability by imposing orthonormality constraints on the loading matrices, as proposed by Wang, Liu, and R. Chen 2019 and L. Yu et al. 2022:

$$\left\| \frac{1}{p_1} R^\top R - I_{k_1} \right\| \to 0, \quad \left\| \frac{1}{p_2} C^\top C - I_{k_2} \right\| \to 0. \tag{5}$$

By imposing these conditions, arbitrary invertible transformations are no longer permitted because, trough orthonormality constraints, the equivalence among multiple representations is broken, as any multiplication by a non-orthogonal matrix would violate the imposed conditions. This is why the model is said to be identifiable *up to orthogonal rotations*. This, in turn, allows for the estimation of a "unique" and a restricted set of parameters. It is worth noting that even under orthonormality constraints, the model remains identifiable also *up to the sign* of the factors and loadings. Inverting the sign of columns in R and C simultaneously leads to the same observed product $RF_tC^\top$.

To implement this constraint, the loading matrices can be decomposed as:

$$R = \sqrt{p_1}Q_1W_1, \quad C = \sqrt{p_2}Q_2W_2,$$

where $Q_1 \in \mathbb{R}^{p_1 \times k_1}$ and $Q_2 \in \mathbb{R}^{p_2 \times k_2}$ have orthonormal columns, meaning that each column has unit norm and the scalar product between different columns is zero [2]. The matrices W_1 and W_2 are full-rank,

[2] Unit norm means that in each column the weight is standardized across the observed variables, i.e., $\sum_i R_{i,j}^2 = 1$ allowing for column loadings comparison; orthogonality implies that each factor captures different information, that in other words means that is geometrically independent from the others.

i.e., invertible square matrices. Finally, the scaling factors $\sqrt{p_1}$ and $\sqrt{p_2}$ are included to satisfy the normalization condition. When multiplying by any matrix W, it is now necessary to respect the constraint structure. The decomposition above implies:

$$\frac{1}{p_1}R^\top R = W_1^\top W_1, \quad \frac{1}{p_2}C^\top C = W_2^\top W_2.$$

Assuming that W_1 and W_2 are orthogonal, one can impose $W_i^\top W_i = I$, thus enforcing orthonormality and addressing the identifiability issue. These assumptions not only allow for the identification, but also reduce the number of free parameters and reinforce the structural view of pervasive latent factors. Orthonormality ensures that loading columns have unit norm and are mutually orthogonal, facilitating both interpretation and estimation. Combined with the plausible assumption that loadings are spread across many variables, this structure implies that each factor affects a substantial portion of the data. This is especially useful in high-dimensional contexts, where a few common components must capture the dynamics of many variables.

3.1.2 Dynamic Factor Model

This thesis adopts an approximate vector DFM as a benchmark to evaluate the nowcasting performance of the DMFM, where the only structural difference lies in the bilinear factorization that characterizes the latter.Indeed, although this work builds upon the matrix formulation of Dynamic Factor Models (DFMs), such models can be regarded as a natural extension of the standard and widely used vector-form DFMs. The latter can be seen as a limiting case of the former: by reducing the number of rows to $p_1 = 1$ and the number of row factors to $k_1 = 1$, or alternatively the number of columns to $p_2 = 1$ and the number of column factors to $k_2 = 1$, the matrix structure collapses into a vector structure.

Let's consider a data-generating process analogous to the one in the DMFM, now applied to a vector context. The observed time series y_{nt} is an n-dimensional stationary vector process, standardized to have zero mean and unit variance. The latent factors are assumed to follow a Vector Autoregressive (*VAR*) process of order $P = 1$. The problem can be formulated through a state-space representation where the measurement equation links the observed data to the latent factors (Equation 6), and a transition equation drives the evolution of the factors through an autoregressive process (Equation 7):

$$y_{nt} = \Lambda_n F_t + \xi_{nt}, \qquad t = 1, \ldots, T \tag{6}$$

$$F_t = AF_{t-1} + v_t, \qquad v_t \sim \mathcal{N}(0, \Gamma_v) \tag{7}$$

In Equation 6, $y_{nt} \in \mathbb{R}^n$ is modeled as the sum of the common component composed of the factor loading matrix $\Lambda_n \in \mathbb{R}^{n \times r}$ and the latent factor vector $F_t \in \mathbb{R}^r$ and the idiosyncratic noise, that is ξ_{nt}, which may exhibit weak cross-sectional correlation with other series. Equation 7 introduces dynamics by specifying a *VAR* process for the latent factors, with autoregressive coefficients captured by matrix $A \in \mathbb{R}^{r \times r}$ and innovation $v_t \sim \mathcal{N}(0, \Gamma_v)$.

DFM vs. DMFM: Model Comparison

Dynamic Factor Model (DFM)

Observation equation:

$$y_t = \Lambda F_t + \xi_t$$

Transition equation:

$$F_t = AF_{t-1} + v_t$$

Key characteristics:

- Vector-valued data: $y_t \in \mathbb{R}^n$
- Factor loadings: $\Lambda \in \mathbb{R}^{n \times r}$

Dynamic Matrix Factor Model (DMFM)

Observation equation:

$$Y_t = RF_tC^\top + E_t$$

Transition equation:

$$F_t = AF_{t-1}B^\top + U_t$$

Key characteristics:

- Matrix-valued data: $Y_t \in \mathbb{R}^{p_1 \times p_2}$
- Bilinear structure of loadings: R, C

3.2 Estimation

In this subsection, I present the strategy followed for the estimation of the approximate Dynamic Matrix Factor Model (DMFM) discussed in Subsection 3.1.1. Borrowing the methodology from Matteo Barigozzi and Trapin 2025, the parameters of the DMFM are jointly estimated via Quasi Maximum Likelihood (QML), implementing the Expectation-Maximization (EM) algorithm, where the Kalman smoother updates the estimates of the latent factors and their covariances in a backward recursion. Overall, the EM algorithm allows for several layers of generality. As explained in Subsection 3.3, it is particularly well-suited for handling general patterns of missing data. Moreover, as discussed in Subsection 3.4, given the inclusion of Kalman filtering techniques, it can be effectively embedded within a nowcasting framework. It goes without saying that the flexibility to accommodate mixed-frequency datasets and more complex patterns of missing values in a recursive nowcasting exercise—when the data are naturally organized in matrix form—makes this approach particularly attractive for empirical forecasting applications, especially in central banking contexts.

The methodology presented here extends the EM algorithm to matrix-valued time series, as discussed in Matteo Barigozzi and Trapin 2025, whereas its validity in the vector case has already been successfully established by Doz, Giannone, and Reichlin 2012, Bańbura and Modugno 2014, or M. Barigozzi and Lissona 2024. In fact, the vector-form DFM can be regarded as a special case of the matrix formulation, both in terms of model structure and estimation procedure.

In order to estimate the parameters of the DMFM one needs to maximize the model's gaussian log likelihood function. However, the first obstacle to face for the model as originally specified is it's approximate nature. The fully specified idiosyncratic covariance matrix makes the standard Maximum Likelihood estimation computationally infeasible because the number of parameters to estimate growth rapidly at rate of $\mathcal{O}((p_1^2 + p_2^2)T)$. To reduces drastically the number of parameters the log-likelihood is intentionally mis-specified by imposing a diagonal structure on the idiosyncratic component, meaning restricting the cross-sectional and serial correlation. The second challenge is the no admitted closed-

form solution of the resulting quasi-log-likelihood. In this step the EM algorithm plays a vital role since it is used to iteratively compute closed-form updates for all model parameters. From the estimates of the latent factors obtained through the Kalman smoother, the E-step, described in Paragraph 3.2.2, reconstructs a "complete" dataset, while the M-step, detailed in Paragraph 3.2.2, provides updated estimates of the model parameters. The procedure iterates until the quasi-log-likelihood stabilizes around a local maximum, given the model specification and the observed data.

The following subsections are a rigorous discussion of this estimation strategy, with particular emphasis on the implementation of the EM algorithm, along with the initialization procedure and the convergence criteria adopted in the empirical analysis, as inspired by Matteo Barigozzi and Trapin 2025.

3.2.1 Pseudo Likelihood

The DMFM defined in Equations (3) and (4) implies the following covariance structure for the data:

$$\Omega_{Y_T} = (I_T \otimes C \otimes R)\,\Omega_{F_T}(A,B,P,Q)\,(I_T \otimes C \otimes R)^\top + \Omega_{E_T},$$

where the first term captures the common component, and the second term represents the idiosyncratic noise.

Estimating the full covariance matrix $\Omega_{E_T} = \mathrm{Cov}(E_1,\dots,E_T)$ without any restrictions is infeasible, since the total number of observations available in the data is Tp_1p_2, while the full estimation of Ω_{E_T}, accounting for both cross-sectional and serial dependence, would require the estimation of $\mathcal{O}(T^2(p_1p_2)^2)$ free parameters. To compute Ω_{E_T}, one must vectorize each matrix E_t into $\mathrm{vec}(E_t)$, resulting in a $T(p_1p_2)$-dimensional stacked vector. The covariance matrix Ω_{E_T} would then be of dimension $Tp_1p_2 \times Tp_1p_2$, implying $\frac{Tp_1p_2(Tp_1p_2+1)}{2} = \mathcal{O}(T^2p_1^2p_2^2)$ free parameters, assuming symmetry. This number quickly becomes infeasible in practice.

To overcome this issue, Matteo Barigozzi and Trapin 2025 impose both cross-sectional and serial uncorrelation of E_t, reducing the number of parameters to estimate from $\mathcal{O}((p_1p_2T)^2)$ to $\mathcal{O}(p_1p_2T)$. Through this assumption, the estimation becomes feasible since the idiosyncratic covariance structure reduces to:

$$\Omega_{E_T} \approx I_T \otimes \mathrm{diag}(K) \otimes \mathrm{diag}(H),$$

The gaussian quasi-log-likelihood function under this mis-specification of Ω_{E_T} becomes:

$$\begin{aligned}\ell(Y_T;\theta) = -\frac{p_1p_2T}{2}\log(2\pi) - \frac{1}{2}\log\Big|(I_T \otimes C \otimes R)\Omega_{F_T}(I_T \otimes C \otimes R)^\top + I_T \otimes \mathrm{diag}(K) \otimes \mathrm{diag}(H)\Big| \\ - \frac{1}{2}Y_T^\top\Big[(I_T \otimes C \otimes R)\Omega_{F_T}(I_T \otimes C \otimes R)^\top + I_T \otimes \mathrm{diag}(K) \otimes \mathrm{diag}(H)\Big]^{-1}Y_T. \qquad (8)\end{aligned}$$

Where Ω_{F_T} is the covariance matrix of the vectorized latent factor process that depends on the matrices $A \in \mathbb{R}^{k_1\times k_1}$, $B \in \mathbb{R}^{k_2\times k_2}$, and the innovation covariances $P \in \mathbb{R}^{k_1\times k_1}$ and $Q \in \mathbb{R}^{k_1\times k_1}$.

The maximizer of this quasi-log-likelihood is defined Quasi-Maximum Likelihood (QML) estimator. Through the mis-specification of Ω_{E_T} it became computationally feasible even in high-dimensional settings. indeed, by imposing a diagonal structure, ruling out any form of correlation, the number of

parameters to estimate decreases drastically.

3.2.2 EM Algorithm

After imposing a diagonal structure on the idiosyncratic covariance matrix Ω_{E_T} to ensure feasible estimation, the next step is to maximize the mis-specified Gaussian quasi-log-likelihood, as defined in Equation 3.2.4. However, this task presents another challenge since no closed-form solution is available.

In general, the log-likelihood of multivariate Gaussian data admits a closed-form expression only when the covariance structure can be explicitly expressed in terms of the model parameters. In this case, neither the factor covariance matrix Ω_{F_T} nor the observation covariance matrix Ω_{Y_T} is available in closed form. This prevents analytical maximization of the quasi-log-likelihood for two main reasons: first, Ω_{F_T} depends implicitly on the unknown parameters θ; second, the latent factors F_t are unobserved and thus cannot be directly substituted into the likelihood. Furthermore, even under the assumption that the factors were observed, the structure of Ω_{Y_T} still precludes closed-form expressions for both its determinant and inverse. In fact, Ω_{Y_T} is defined as a sum of Kronecker products, a form that does not allow for algebraic simplifications. As a result, computing $\log|\Omega_{Y_T}|$ and $\Omega_{Y_T}^{-1}$ analytically is not feasible.

A common solution to the absence of a closed-form expression is to implement an iterative estimation strategy, such as the Expectation-Maximization (EM) algorithm. The EM algorithm allows for the estimation of latent variables in a first step, followed by conditional optimization of the model parameters. The procedure consists of two steps:

- In the **E-step** (Expectation), given the current parameter estimates $\theta^{(n)}$, the Kalman filter followed by the Kalman smoother is used to compute the conditional expectation of the latent factors, $\hat{F}_t = \mathbb{E}[F_t \mid Y, \theta^{(n)}]$, and their conditional covariance, $\mathbb{V}[F_t \mid Y, \theta^{(n)}]$. These quantities effectively reconstruct the unobserved information, allowing the algorithm to operate on a "complete" dataset that includes the estimated factors.

- In the **M-step** (Maximization), the expected value of the complete-data log-likelihood—constructed using the outputs of the E-step—is maximized with respect to the parameters θ using Quasi Maximum Likelihood Estimation (QMLE). Although the factors are estimated conditionally on the observed data Y, this step enables optimization over all remaining parameters of the DMFM.

Each iteration produces a new set of parameter estimates, $\theta^{(n+1)}$, which serve as the starting point for the next iteration. Repeating this process until convergence—defined as the stabilization of the observed log-likelihood—leads to a local maximum of the quasi-log-likelihood function.

In the following paragraphs I describe rigorously these two steps along with a formal description of the Kalman Filtering techniques.

Kalman Filter and Smoother : Kalman filtering techniques, originally developed in physics and engineering, have been widely adopted in econometrics due to their effectiveness in forecasting and

state estimation. During the estimation procedure, for each EM iteration $n \geq 0$, and given the current parameter estimates $\hat{\Theta}^{(n)}$, the Kalman filter and smoother are employed to retrieve the key conditional expectations required in the E-step. Specifically, the procedure yields the expected value of the latent vectorized factor $f_t = \text{vec}(F_t)$, conditional on the full sample:

$$f_{t|T} = \mathbb{E}[f_t \mid Y_{1:T}; \hat{\Theta}^{(n)}],$$

along with the corresponding conditional covariances.

As described in Subsubsection 3.1.1, the DMFM is vectorized as $y_t = \text{vec}(Y_t) \in \mathbb{R}^{p_1 p_2}$, yielding the following state-space representation:

$$\begin{aligned} y_t &= Z f_t + e_t, \\ f_t &= T f_{t-1} + u_t, \end{aligned}$$

where $Z = C \otimes R$ and $T = A \otimes B$.

Starting from initial conditions $f_{0|0}$ and $P_{0|0}$, typically set to zero and the identity matrix respectively, the Kalman filter proceeds forward in time. When data are missing, a selection matrix can be introduced to handle the unobserved elements at each time step.

For each period $t = 1, \ldots, T$, the filter performs the following steps:

1. *Prediction step*:
$$f_{t|t-1} = T f_{t-1|t-1}, \quad P_{t|t-1} = T P_{t-1|t-1} T^\top + Q.$$

2. *Update step*:
$$S_t = Z_t P_{t|t-1} Z_t^\top + H_t, \quad K_t = P_{t|t-1} Z_t^\top S_t^{-1},$$
$$f_{t|t} = f_{t|t-1} + K_t (y_t - Z_t f_{t|t-1}), \quad P_{t|t} = P_{t|t-1} - K_t Z_t P_{t|t-1}.$$

The Kalman smoother then performs a backward recursion to refine the state estimates. The smoothed values are given by:

$$\begin{aligned} f_{t|T} &= f_{t|t} + P_{t|t} r_t, \\ P_{t|T} &= P_{t|t} - P_{t|t} N_t P_{t|t}, \end{aligned}$$

where the recursion terms are computed as:

$$\begin{aligned} r_t &= Z_t^\top S_t^{-1} (y_t - Z_t f_{t|t-1}) + L_t^\top r_{t+1}, \\ N_t &= Z_t^\top S_t^{-1} Z_t + L_t^\top N_{t+1} L_t. \end{aligned}$$

The smoothed estimates $(f_{t|T}, P_{t|T})$ are then used in the E-step to compute the conditional expectations of the data and latent factors, which are required for updating the parameters in the M-step of the EM algorithm.

E-step: Once the Kalman smoother has provided the smoothed estimates of the conditional first and second moments of the latent factors, the E-step of the EM algorithm computes the expected Gaussian quasi-log-likelihood of the approximate DMFM. Given the current parameter estimates $\hat{\theta}^{(n)}$, the observed-data log-likelihood can be decomposed, applying Bayes' rule, as follows:

$$\ell(Y_T;\theta) = \underbrace{\mathbb{E}_{\hat{\theta}^{(n)}}\left[\ell(Y_T \mid F_T;\theta) \mid Y_T\right]}_{\substack{\text{Expected log-likelihood} \\ \text{of the observed data}}} + \underbrace{\mathbb{E}_{\hat{\theta}^{(n)}}\left[\ell(F_T;\theta) \mid Y_T\right]}_{\substack{\text{Expected log-likelihood} \\ \text{of the latent states}}} - \underbrace{\mathbb{E}_{\hat{\theta}^{(n)}}\left[\ell(F_T \mid Y_T;\theta) \mid Y_T\right]}_{\text{Conditional entropy}} \tag{9}$$

The first term reflects how well the current parameters explain the observed data conditional on the latent states; the second measures the consistency of the model with the dynamics of the latent process; the third is a correction term related to the entropy of the conditional distribution.

As shown by Dempster, Laird, and Rubin 1977, and since the model belongs to the exponential family, the last term does not depend on the parameters being optimized. Therefore, it can be omitted when maximizing the observed-data log-likelihood $\ell(Y_T;\theta)$.

For this reason, the E-step focuses on computing the so-called *Q-function*, defined as:

$$Q(\theta,\hat{\theta}^{(n)}) = \mathbb{E}_{\hat{\theta}^{(n)}}\left[\ell(Y_T \mid F_T;\theta) \mid Y_T\right] + \mathbb{E}_{\hat{\theta}^{(n)}}\left[\ell(F_T;\theta) \mid Y_T\right] \tag{10}$$

Subsequently, the M-step of the EM algorithm maximizes this expected complete-data log-likelihood with respect to θ.

The expected log-likelihood of the observed data takes the following form:

$$\begin{aligned}\mathbb{E}_{\hat{\theta}^{(n)}}\left[\ell(Y_T \mid F_T;\theta) \mid Y_T\right] = &-\frac{T}{2}\left(p_1 \log|K| + p_2 \log|H|\right) \\ &-\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_{\hat{\theta}^{(n)}}\left[\operatorname{tr}\left(H^{-1}(Y_t - RF_tC^\top)K^{-1}(Y_t - RF_tC^\top)^\top\right)\Big|Y_T\right]\end{aligned} \tag{11}$$

while the expected log-likelihood of the latent states is given by:

$$\begin{aligned}\mathbb{E}_{\hat{\theta}^{(n)}}\left[\ell(F_T;\theta) \mid Y_T\right] = &-\frac{T-1}{2}\left(k_1 \log|Q| + k_2 \log|P|\right) \\ &-\frac{1}{2}\sum_{t=2}^{T}\mathbb{E}_{\hat{\theta}^{(n)}}\left[\operatorname{tr}\left(P^{-1}(F_t - AF_{t-1}B^\top)Q^{-1}(F_t - AF_{t-1}B^\top)^\top\right)\Big|Y_T\right]\end{aligned} \tag{12}$$

It is worth noting that both expressions above are directly formulated in terms of the original matrix-valued observations and factor processes.

M-step: In the M-step, Equations (11) and (12), derived using the current parameter estimates $\hat{\theta}^{(n)}$, are maximized to obtain an updated set of parameters $\hat{\theta}^{(n+1)}$. At each iteration $n \geq 0$, new estimates of the DMFM parameters are computed, beginning with the row and column loading matrices, denoted

by $\hat{R}^{(n+1)}$ and $\hat{C}^{(n+1)}$, respectively.

The row loadings are updated as follows:

$$\hat{R}^{(n+1)} = \left(\sum_{t=1}^{T} Y_t \hat{K}^{(n)-1} \hat{C}^{(n)} \hat{F}_{t|T}^{(n)\prime} \right) \left(\sum_{t=1}^{T} \left(\hat{C}^{(n)\prime} \hat{K}^{(n)-1} \hat{C}^{(n)} \right) \star \left(\hat{F}_{t|T}^{(n)} \hat{F}_{t|T}^{(n)\prime} + \Pi_{t|T}^{(n)} \right) \right)^{-1} \tag{13}$$

while the column loadings are updated via:

$$\hat{C}^{(n+1)} = \left(\sum_{t=1}^{T} Y_t^{\top} \hat{H}^{(n)-1} \hat{R}^{(n+1)} \hat{F}_{t|T}^{(n)} \right) \left(\sum_{t=1}^{T} \left(\hat{R}^{(n+1)\prime} \hat{H}^{(n)-1} \hat{R}^{(n+1)} \right) \star \left(K_{k_1 k_2} \left(\hat{F}_{t|T}^{(n)} \hat{F}_{t|T}^{(n)\prime} + \Pi_{t|T}^{(n)} \right) K_{k_1 k_2}' \right) \right)^{-1} \tag{14}$$

Since the estimation of R and C depends on each other, the empirical strategy adopted is to first compute $\hat{R}^{(n+1)}$ conditionally on $\hat{C}^{(n)}$, and then update $\hat{C}^{(n+1)}$ based on the newly estimated $\hat{R}^{(n+1)}$. However, the reverse order would also be valid.

Given $\hat{R}^{(n+1)}$ and $\hat{C}^{(n+1)}$, we can now estimate the idiosyncratic covariance matrices $\hat{H}^{(n+1)}$ and $\hat{K}^{(n+1)}$. As discussed in Section 3.2.1, these matrices are assumed to be diagonal, consistent with the quasi-log-likelihood mis-specification in Equation 3.2.4. For $i = 1, \ldots, p_1$, with $[\hat{H}^{(n+1)}]_{ij} = 0$ for $i \neq j$, $[\hat{H}^{(n+1)}]_{ij}$ is computed as:

$$\begin{aligned} [\hat{H}^{(n+1)}]_{ii} = \frac{1}{T p_2} \sum_{t=1}^{T} \Big[& Y_t \hat{K}^{(n)-1} Y_t^{\top} - Y_t \hat{K}^{(n)-1} \hat{C}^{(n+1)} \hat{F}_{t|T}^{(n)\prime} \hat{R}^{(n+1)\prime} - \hat{R}^{(n+1)} \hat{F}_{t|T}^{(n)} \hat{C}^{(n+1)\prime} \hat{K}^{(n)-1} Y_t^{\top} \\ & + \left(\hat{C}^{(n+1)\prime} \hat{K}^{(n)-1} \hat{C}^{(n+1)} \right) \star \left((I_{k_2} \otimes \hat{R}^{(n+1)}) (\hat{F}_{t|T}^{(n)} \hat{F}_{t|T}^{(n)\prime} + \Pi_{t|T}^{(n)}) (I_{k_2} \otimes \hat{R}^{(n+1)})' \right) \Big]_{ii} \end{aligned} \tag{15}$$

Similarly, for $i = 1, \ldots, p_2$, with $[\hat{K}^{(n+1)}]_{ij} = 0$ for $i \neq j$, $[\hat{K}^{(n+1)}]_{ij}$ takes the form of :

$$\begin{aligned} [\hat{K}^{(n+1)}]_{ii} = \frac{1}{T p_1} \sum_{t=1}^{T} \Big[& Y_t^{\top} \hat{H}^{(n+1)-1} Y_t - Y_t^{\top} \hat{H}^{(n+1)-1} \hat{R}^{(n+1)} \hat{F}_{t|T}^{(n)} \hat{C}^{(n+1)\prime} - \hat{C}^{(n+1)} \hat{F}_{t|T}^{(n)\prime} \hat{R}^{(n+1)\prime} \hat{H}^{(n+1)-1} Y_t \\ & + \left(\hat{R}^{(n+1)\prime} \hat{H}^{(n+1)-1} \hat{R}^{(n+1)} \right) \star \left((I_{k_1} \otimes \hat{C}^{(n+1)}) K_{k_1 k_2} (\hat{F}_{t|T}^{(n)} \hat{F}_{t|T}^{(n)\prime} + \Pi_{t|T}^{(n)}) K_{k_1 k_2}' (I_{k_1} \otimes \hat{C}^{(n+1)})' \right) \Big]_{ii} \end{aligned} \tag{16}$$

As discussed for the loadings, the estimation of $\hat{H}$ and $\hat{K}$ depend on each other. To maintain the bilinear structure of the model, $\hat{H}^{(n+1)}$ is computed first, conditional on $\hat{K}^{(n)}$.

Following Matteo Barigozzi and Trapin 2025, this work uses autoregressive matrices A and B and innovation covariance matrices P and Q solely for the implementation of the Kalman smoother on vectorized data. Therefore, the transition and innovation matrices are estimated via the vectorized

MAR model as:

$$\widehat{B \otimes A}^{(n+1)} = \left(\sum_{t=2}^{T} \hat{F}_{t|T}^{(n+1)} \hat{F}_{t-1|T}^{(n)\prime} + \Delta_{t|T}^{(n)} \right) \left(\sum_{t=2}^{T} \hat{F}_{t-1|T}^{(n)} \hat{F}_{t-1|T}^{(n)\prime} + \Pi_{t-1|T}^{(n)} \right)^{-1} \tag{17}$$

$$\widehat{Q \otimes P}^{(n+1)} = \frac{1}{T} \sum_{t=2}^{T} \left(\hat{F}_{t|T}^{(n)} \hat{F}_{t|T}^{(n)\prime} + \Pi_{t|T}^{(n)} - \left(\hat{F}_{t|T}^{(n)} \hat{F}_{t-1|T}^{(n)\prime} + \Delta_{t|T}^{(n)} \right) \left(\widehat{B \otimes A}^{(n+1)} \right)' \right) \tag{18}$$

In this case, the estimators do not enforce the bilinear structure of the MAR model directly. However, since singular estimates of the dynamic parameters are not required, and the estimated loadings and factors are asymptotically unaffected by this simplification, this approach is adopted without loss of generality.

3.2.3 EM Initialization

The EM algorithm, being an iterative procedure for likelihood maximization, requires a starting point, that is an initial set of parameter estimates. Since the quasi-log-likelihood function is generally non-concave, the algorithm may converge to a local rather than a global maximum. The quality of the initialization is crucial since the closer the initial values are to the region of the global maximum, the faster and more reliably the algorithm will converge. Moreover, because the E-step relies on these initial estimates to compute the expected complete-data log-likelihood, a poor initialization can lead to slow convergence or convergence to suboptimal solutions.

In order obtain initial estimates of DMFM parameters, I first discuss the MFM's parameters initialization and then the MAR once. For the former the methodology adopted are the projected estimates (PE) proposed by L. Yu et al. 2022 while, for the latter, the pre-estimators are given by the OLS estimates of the vectorized transition and innovation covariance matrices.

MFM Parameters: L. Yu et al. 2022 introduce the Projected Estimator for large-dimensional matrix factor models. This approach allows for the estimation of the correct number of row and column factors and the estimates of the corresponding loading matrices. The technique is based on the eigenvalue decomposition of specific covariance matrices constructed from projections of the data onto the row and column spaces, reducting the dimension drastically. Since this projection procedure is based on a PCA approach, it requires data imputation in presence of missing. This extension is discussed in Subsection 3.3.2 and it is inspired by the imputation method proposed by Cen and Lam 2025.

The discussion of the L. Yu et al. 2022's methodology presented in this section describes first the projected estimation method and its consistency. Then, focuses on the initial projections when both row and column spaces are unknown, and finally describes how to extract matrix factor loadings using principal components and how to determine the number of row and column factors via the eigenvalue ratio algorithm proposed by Lam and Yao 2012. For these two algorithms I provide a supplementary

concise boxed summary.

Let's start from a matrix factor model without explicit factor dynamics as proposed by L. Yu et al. 2022. This means the model can be considered static, but this does not limit the applicability of their method to dynamic matrix factor models. Given a column loading matrix C satisfying the orthogonality condition discussed in paragraph 5 it is possible to project the data matrix onto the column space. The result is this projected data matrix:

$$Y_t = \frac{1}{p_2} X_t C = \frac{1}{p_2} R F_t C^\top C + \frac{1}{p_2} E_t C = R F_t + \tilde{E}_t,$$

where $\tilde{E}_t = \frac{1}{p_2} E_t C$ is the transformed noise term. The shrinkage of the number of columns from p_2 to k_2 provides immediate dimensionality reduction. Moreover, since the variance of $\tilde{E}_t$ is of order $\mathcal{O}(1/p_2)$, the noise level decreases significantly when p_2 is large. Finally, the projected data Y_t behave like a near noise-free factor model:

$$Y_t = R F_t + \tilde{E}_t.$$

Thus, the projection turns the matrix factor model into a standard vector factor model, with the idiosyncratic component that asymptotically vanishes as $p_2 \to \infty$.

Given the projected data, the next step is to construct the sample covariance matrix and apply a Principal component analysis. From:

$$\tilde{M}_1 = \frac{1}{T p_1} \sum_{t=1}^{T} Y_t Y_t^\top,$$

It is possible to extract the eigenvectors corresponding to the leading k_1 eigenvalues of M_1 to obtain a consistent estimator of the row loading matrix R. It is worth noting that the number of leading eigenvalue relies on the Algorithm 2 discussed in paragraph 3.2.3. Finally, to ensure that the signal does not diminish as dimensionality increases, the matrix R must be rescaled by $\sqrt{p_1}$. The specular discussion is true for C as the box 3.2.3 shows. At this point it is possible to Update the loadings, namely $\tilde{R}$ and $\tilde{C}$ that can be used as new initial values for a sequent estimation. However, as shown by L. Yu et al. 2022, one iteration is often sufficient for accurate estimation.

Since column loading matrix C (or equally R) is unknown and must be estimated from the data a natural approach to obtain some initial estimators, namely $\hat{C}$ (or seemingly $\hat{R}$). To estimate $\hat{C}$, first transpose the data matrices $X_t^\top$ and treat the rows of X_t as observations from a vector factor model. The next step is to compute the sample covariance matrix:

$$\hat{M}_2 = \frac{1}{T p_1 p_2} \sum_{t=1}^{T} X_t^\top X_t.$$

and from this $p_2 \times p_2$ matrix $\hat{M}_2$, extract the leading k_2 eigenvectors to obtain $\hat{Q}_2$. The initial estimator of the column loading matrix is then given by:

$$\hat{C} = \sqrt{p_2} \hat{Q}_2.$$

Similarly, the row loading matrix $\hat{R}$ is estimated by applying the same procedure but specular.

This estimates are noisier then the ones obtained after projection but are sufficient to initialize the iterative procedure summarized in 3.2.3. Indeed, even the first projection improves the estimates significantly, reducing the estimation variance, and enhancing convergence properties.

Algorithm 1: Projected Estimation

1. **Initial Estimation:** Obtain preliminary estimates $\hat{R}$ and $\hat{C}$, for example via PCA.
2. **Projection:** Define projected data:
$$\hat{Y}_t = \frac{1}{p_2} X_t \hat{C}, \quad \hat{Z}_t = \frac{1}{p_1} X_t^\top \hat{R}.$$
3. **Covariance Matrices:** Extract the leading k_1 and k_2 eigenvectors $\tilde{Q}_1$ and $\tilde{Q}_2$ from the Covariance Matrices:
$$\tilde{M}_1 = \frac{1}{Tp_1} \sum_{t=1}^{T} \hat{Y}_t \hat{Y}_t^\top, \quad \tilde{M}_2 = \frac{1}{Tp_2} \sum_{t=1}^{T} \hat{Z}_t \hat{Z}_t^\top.$$
4. **Loadings:** Update the loadings as: $\tilde{R} = \sqrt{p_1}\tilde{Q}_1, \tilde{C} = \sqrt{p_2}\tilde{Q}_2$.

From the pre-estimators of $\tilde{R}$ and $\tilde{C}$ the factor matrix one is obtained via linear projection:

$$\hat{F}_t = \frac{\hat{R}^{(0)\top} Y_t \hat{C}^{(0)}}{p_1 p_2}.$$

Now it is possible to compute residuals as:

$$\hat{E}_t^{(0)} = Y_t - \hat{R}^{(0)} \hat{F}_t \hat{C}^{(0)\top},$$

end extract pre-estimators of the noise covariance matrices H and K are given by:

$$\hat{K}^{(0)} = \frac{1}{Tp_1} \sum_{t=1}^{T} \operatorname{tr}\left(\hat{E}_t^{(0)\top} \hat{E}_t^{(0)}\right), \quad \hat{H}^{(0)} = \frac{1}{Tp_1} \sum_{t=1}^{T} \operatorname{tr}\left(\hat{E}_t^{(0)} \hat{K}^{(0)-1} \hat{E}_t^{(0)\top}\right). \tag{19}$$

where only the the diagonal terms are required to run the EM algorithm.

Since initial estimates of the column or row loadings are unknown, another step to initialize this algorithm is needed, and the following paragraph provides the methodology.

Estimation of the Number of Row and Column Factors As mentioned during the discussion of Algorithm 1 in subsection 3.2.3 to be practically efficient one needs to determine the correct number of row and column factors. This step is crucial both for or projected estimation and for the factors' number considered for the EM algorithm.

The number of row and column factors, respectively denoted by k_1 and k_2, provide the dimensions of the latent factor matrix $F_t \in \mathbb{R}^{k_1 \times k_2}$, and consequently the number of columns in the row and column

loading matrices, respectively $R \in \mathbb{R}^{p_1 \times k_1}$ and $C \in \mathbb{R}^{p_2 \times k_2}$.

k_1 and k_2 are estimated using the eigenvalue ratio criterion, originally proposed by Lam and Yao 2012, which is based on detecting sudden jumps between variance explained by eigenvectors in the covariance matrix resulted from the projection process. Specifically, k_1 is selected as the index j that maximizes the ratio between the j-th and $(j+1)$-th largest eigenvalues of the projected covariance matrix $\tilde{M}_1$, formally:

$$\hat{k}_1 = \arg\max_{j \le k_{\max}} \frac{\lambda_j(\tilde{M}_1)}{\lambda_{j+1}(\tilde{M}_1)},$$

where $\lambda_j(\tilde{M}_1)$ is the j-th largest eigenvalue, and $k_{\max}$ is a pre-specified upper bound. Once the first k_1 eigenvalues significantly larger than the rest are detected, one can use them to explain the substantial part of the variance. The corresponding factor will be retained and the others discarded.

Since small eigenvalues could affect the process, the strategy to stabilize the ratio is adding a regularization term to the denominator and some small constant c as follows:

$$\lambda_{j+1} + c\delta, \quad \text{with } \delta = \max\left\{\frac{1}{\sqrt{Tp_2}}, \frac{1}{\sqrt{Tp_1}}, \frac{1}{p_1}\right\},$$

In practice, estimating $\tilde{M}_1$ requires a preliminary estimate of the column loading matrix $\hat{C}$, which in turn depends on k_2. Similarly, estimating $\tilde{M}_2$ requires knowledge of $\hat{R}$ and k_1. To address this interdependency, another iterative algorithm, Algorithm 2 is proposed by L. Yu et al. 2022 and it is breifly summarized in box 3.2.3.

Algorithm 2: Number of Factors Estimation

1. **Initialization:** Set initial guesses: $\hat{k}_1^{(0)} = k_{\max}, \hat{k}_2^{(0)} = k_{\max}$.
2. **Iterative Updates:** For iterations $t = 1, \ldots, m$, where m is arbirtrary setted:
   (a) Given $\hat{k}_2^{(t-1)}$, estimate the column loading matrix $\hat{C}^{(t)}$ via PCA.
   (b) Compute the projected covariance matrix:
   $$\tilde{M}_1^{(t)} = \frac{1}{Tp_1} \sum_{t=1}^{T} \hat{Y}_t \hat{Y}_t^\top.$$
   (c) Update $\hat{k}_1^{(t)}$ as:
   $$\hat{k}_1^{(t)} = \arg\max_j \frac{\lambda_j(\tilde{M}_1^{(t)})}{\lambda_{j+1}(\tilde{M}_1^{(t)})}.$$
   (d) Given $\hat{k}_1^{(t)}$, estimate the row loading matrix $\hat{R}^{(t)}$, compute $\tilde{M}_2^{(t)}$ and update $\hat{k}_2^{(t)}$
3. **Stopping Criterion:** Stop if m or the convergence is reached:
   $$\hat{k}_1^{(t)} = \hat{k}_1^{(t-1)} \quad \text{and} \quad \hat{k}_2^{(t)} = \hat{k}_2^{(t-1)},$$

MAR Parameters: Given the latent factor matrices the initial estimator of the MAR parameters are obtained through the OLS estimates of the vectorized transition and innovation covariance matrices. The first step is vectorizing factors:

$$\tilde{f}_t = \frac{(\hat{C}^{(0)} \otimes \hat{R}^{(0)})^\top y_t}{p_1 p_2},$$

where $y_t = \text{vec}(Y_t)$ is the vectorized projected data matrix at time t.

and then, the pre-estimators for the MAR transition parameters are given by:

$$\widehat{B \otimes A^{(0)}} = \left(\sum_{t=2}^{T} \tilde{f}_t \tilde{f}_{t-1}^\top \right) \left(\sum_{t=2}^{T} \tilde{f}_{t-1} \tilde{f}_{t-1}^\top \right)^{-1},$$

$$\widehat{Q \otimes P^{(0)}} = \sum_{t=2}^{T} \left(\tilde{f}_t - \widehat{B \otimes A^{(0)}} \, \tilde{f}_{t-1} \right) \left(\tilde{f}_t - \widehat{B \otimes A^{(0)}} \, \tilde{f}_{t-1} \right)^\top.$$

Initialization Process Borrowed from L. Yu et al. 2022

Step 1:Number of Factors Estimation

Use the eigenvalue ration criterion form Lam and Yao 2012 to estimate $(\hat{k}_1, \hat{k}_2)$:

$$\hat{k}_1 = \arg\max_{j \le k_{\max}} \frac{\lambda_j(\tilde{M}_1)}{\lambda_{j+1}(\tilde{M}_1)}, \quad \hat{k}_2 = \arg\max_{j \le k_{\max}} \frac{\lambda_j(\tilde{M}_2)}{\lambda_{j+1}(\tilde{M}_2)}$$

.

Step 2: Initial Estimation of $\hat{R}$ and $\hat{C}$

Given $\hat{k}_1, \hat{k}_2$, estimate initial loading matrices using PCA:

$$\hat{R}^{(0)} = \sqrt{p_1} \cdot \text{eigvecs}(\hat{M}_1), \quad \hat{C}^{(0)} = \sqrt{p_2} \cdot \text{eigvecs}(\hat{M}_2)$$

with

$$\hat{M}_1 = \frac{1}{T p_1 p_2} \sum_{t=1}^{T} X_t X_t^\top, \quad \hat{M}_2 = \frac{1}{T p_1 p_2} \sum_{t=1}^{T} X_t^\top X_t.$$

Step 3: Projected Estimators $\tilde{R}, \tilde{C}$

Using $\hat{R}^{(0)}$, $\hat{C}^{(0)}$, project the data:

$$\hat{Y}_t = \frac{1}{p_2} X_t \hat{C}^{(0)}, \quad \hat{Z}_t = \frac{1}{p_1} X_t^\top \hat{R}^{(0)},$$

and construct

$$\tilde{M}_1 = \frac{1}{T p_1} \sum_{t=1}^{T} \hat{Y}_t \hat{Y}_t^\top, \quad \tilde{M}_2 = \frac{1}{T p_2} \sum_{t=1}^{T} \hat{Z}_t \hat{Z}_t^\top.$$

extract the eigenvectors associated to the leading eigenvalues $\hat{k}_1, \hat{k}_2$:

$$\tilde{R} = \sqrt{p_1} \cdot \tilde{Q}_1, \quad \tilde{C} = \sqrt{p_2} \cdot \tilde{Q}_2.$$

3.2.4 EM Convergence

Any iterative procedure, especially within optimization frameworks, requires a well-defined convergence criterion to determine when the algorithm should stop. This also applies to the EM Algorithm. At each iteration n, the algorithm provides an updated set of estimated parameters, denoted by $\hat{\Theta}^{(n+1)}$. Based on these values, it is possible to compute the Prediction Error Log-Likelihood (PE-LLK).

The algorithm stops when the improvement in log-likelihood between two successive iterations becomes negligible. Specifically, convergence is reached when the relative difference in PE-LLK falls below an arbitrary and user-specified tolerance level ε. Formally this can be expressed as:

$$\Delta L_n = \frac{L(Y_T; \hat{\theta}^{(n+1)}) - L(Y_T; \hat{\theta}^{(n)})}{\frac{1}{2}\left(L(Y_T; \hat{\theta}^{(n+1)}) + L(Y_T; \hat{\theta}^{(n)})\right)}$$

for the seek of clarity, the log-likelihood $L(Y_T; \theta)$ is computed using the Kalman filter, which produces one-step-ahead predictions $f_{t|t-1}$ and associated prediction error covariances $\Pi_{t|t-1}$, as follows:

$$L(Y_T; \theta) = -\frac{1}{2}\sum_{t=1}^{T}\Bigg[\log\det\Big((C\otimes R)\,\Pi_{t|t-1}\,(C\otimes R)^\top + K\otimes H\Big) + (y_t - (C\otimes R) f_{t|t-1})^\top \Big((C\otimes R)\,\Pi_{t|t-1}\,(C\otimes R)^\top + K\otimes H\Big)^{-1}(y_t - (C\otimes R) f_{t|t-1})\Bigg] \quad (20)$$

The numerator in Equation 3.2.4 represents the absolute improvement in the PE-LLK between two consecutive iterations of the algorithm, while the denominator provides a normalization based on the mean value of the PE-LLK at these two iterations. This expression measures the relative percentage improvement of the log-likelihood with respect to the average log-likelihood value across the two iterations.

However, convergence may not occur within a finite number of steps, or the procedure may become computationally intensive in high or ultra-high dimensional settings such as in this case. A possible solution for practical implementation is setting an additional stopping rule based on a maximum number of iterations $n_{\max}$.

Then, the EM algorithm will stop the iterative process as soon as one of the two following conditions is met:

- The relative log-likelihood increment ΔL_n falls below ε,
- The number of iterations reaches $n_{\max}$.

When the stopping condition is satisfied at iteration n^*, the final EM estimate is defined as:

$$\hat{\Theta} = \hat{\Theta}^{(n^*+1)}.$$

and a last run of the Kalman smoother is performed to compute the smoothed estimates of the latent factor matrices.

3.3 Extension to missing data

The presence of missing observations is often addressed by aggregating the data at a higher level or lower frequency (in time series contexts) where all entries are available, or by applying simple imputation methods to fill in the gaps. However, especially in the context of factor models—where each observation may convey relevant information—aggregating data may lead to the exclusion of meaningful variation, while simple imputation methods risk introducing noise and bias into the dataset.

Fortunately, the EM algorithm is naturally suited to handle general patterns of missing values, and this Subsection presents its extension to such cases. The methodology described here is adapted from Matteo Barigozzi and Trapin 2025, who extend the EM algorithm for missing data originally proposed in the vector setting by Bańbura and Modugno 2014. This extension adds an important layer of flexibility to the present work, as the empirical application must account for COVID-19-related disruptions in real variables and for mixed-frequency data.

Before discussing the EM extension, however, it is necessary to address the issue of initialization. The strategy presented in Subsubsection 3.2.3 is based on principal components and does not accommodate missing values. To overcome this limitation, this work adopts a methodology originally developed for the vector case—namely, the "all-purpose estimator" proposed by Xiong and Pelger 2023—and applies its matrix-form extension as introduced in Cen and Lam 2025.

3.3.1 Imputation for the EM initialization

The initial parameters required by the EM algorithm for its first iteration can be obtained as described in Subsubsection 3.2.3. However, the presence of missing data introduces additional complexity. In particular, PCA techniques cannot be directly applied when observations are incomplete, as they rely on a fully observed dataset. A common workaround is to impute the missing entries prior to applying Projected Estimators.

Following the approach of Matteo Barigozzi and Trapin 2025, this work adopts the methodology proposed by Cen and Lam 2025, which generalizes the so-called "all-purpose estimator" originally developed by Xiong and Pelger 2023 for vector time series. This Subsection focuses on the matrix case.

In this setting, given a partially observed dataset $Y \in \mathbb{R}^{T \times N}$, the missing data problem is tackled by constructing an adjusted sample covariance matrix that uses only the available observations. For each pair of variables (i, j), the adjusted covariance is computed as:

$$\widetilde{\Sigma}_{ij} = \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} Y_{it} Y_{jt},$$

where Q_{ij} denotes the set of time periods in which both series i and j are simultaneously observed.

This estimator corrects for the bias introduced by missing values and does not require any distributional assumptions on the missingness pattern. Once $\widetilde{\Sigma}$ is computed, PCA is performed to extract the factor loading matrix $\widetilde{\Lambda}$, whose columns correspond to the eigenvectors associated with the largest eigenvalues.

The common factors at each time t are then estimated by solving a weighted least squares problem:

$$\widetilde{F}_t = \left(\sum_{i=1}^{N} W_{it} \widetilde{\Lambda}_i \widetilde{\Lambda}_i^\top \right)^{-1} \left(\sum_{i=1}^{N} W_{it} \widetilde{\Lambda}_i Y_{it} \right),$$

where W_{it} is an indicator equal to 1 if Y_{it} is observed and 0 otherwise.

This procedure is particularly attractive as it accommodates general missing data patterns, including block of missings and mixed-frequency structures.

3.3.2 Modification of the EM-algorithm

Once the initial estimates have been successfully computed—accounting also for the presence of missing values—the EM algorithm can be initialized. Even in this case, some refinements need to be discussed. Although this extension is appealing, it comes at the cost of slightly more demanding theoretical and computational considerations.

Recalling the discussion in Paragraph 3.2.2, the expected log-likelihood can be decomposed into a component that depends only on the latent factors, as shown in Equation 10, and another that depends on the observed data, namely Equation 11.

Since the first component involves only the latent factors and not the observed data directly, the presence of missing values does not alter its analytical form. As a result, the parameters A, B, P, and Q remain unaffected. On the other hand, the second component is influenced by missingness, as it depends explicitly on the available data. Consequently, the MFM parameters related to the measurement equation—such as the loading matrices R and C, and the idiosyncratic variance matrices H and K—must be adjusted. The presence of missing values is handled in practice through a selection matrix.

The M-step is thus modified following the approach of Bańbura and Modugno 2014 for the vector case and Matteo Barigozzi and Trapin 2025 for the matrix case. In the latter, within a matrix-variate time series framework, the selection matrix $W_t \in \mathbb{R}^{p_1 \times p_2}$, with the same dimensions as Y_t, is defined as:

$$(W_t)_{ij} = \begin{cases} 1 & \text{if } y_{t,ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the MFM parameters are updated at each iteration. In particular, the explicit update formulas for the loading matrices R and C, given the current parameter estimates $\hat{\theta}^{(n)}$, are computed as follows during the estimation process:

$$\begin{aligned} \operatorname{vec}\left(\hat{R}^{(n+1)}\right) = & \left(\sum_{t=1}^{T} \sum_{s=1}^{p_1} \sum_{q=1}^{p_1} \left[\left(\hat{C}^{(n)\top} D^{[s,q]} W_t \hat{K}^{(n)-1} \hat{C}^{(n)} \right) \star \left(\hat{f}_{t|T}^{(n)} \hat{f}_{t|T}^{(n)\top} + \Pi_{t|T}^{(n)} \right) \right] \otimes \left(E_{p_1,p_1}^{[s,q]} \hat{H}^{(n)-1} \right) \right)^{-1} \\ & \times \left(\sum_{t=1}^{T} \operatorname{vec}\left(\left[W_t \circ \hat{H}^{(n)-1} Y_t \hat{K}^{(n)-1} \right] \hat{C}^{(n)} \hat{f}_{t|T}^{(n)\top} \right) \right) \end{aligned}$$

$$\text{vec}\left(\hat{C}^{(n+1)}\right) = \left(\sum_{t=1}^{T}\sum_{k=1}^{p_2}\sum_{q=1}^{p_2}\left[\left(\hat{R}^{(n+1)\top}D^{[k,q]}W_t^{\top}\hat{H}^{(n)-1}\hat{R}^{(n+1)}\right)\star K_{k_1,k_2}\left(\hat{f}_{t|T}^{(n)}\hat{f}_{t|T}^{(n)\top}+\Pi_{t|T}^{(n)}\right)K_{k_1,k_2}^{\top}\right]\otimes\left(E_{p_2,p_2}^{[k,q]}\hat{K}^{(n)-1}\right)\right)^{-1}$$
$$\times\left(\sum_{t=1}^{T}\text{vec}\left(\left[W_t\circ\hat{H}^{(n)-1}Y_t\hat{K}^{(n)-1}\right]^{\top}\hat{R}^{(n+1)}\hat{f}_{t|T}^{(n)}\right)\right)$$

for the same iteration $n \geq 0$, and the current parameter estimates $\hat{\theta}^{(n)}$, the idiosyncratic variances $\hat{H}$ and $\hat{K}$ in the presence of missing data are computed are:

$$\begin{aligned}\left[\hat{H}^{(n+1)}\right]_{ii} = \frac{1}{Tp_2}\sum_{t=1}^{T}\Bigg\{&(W_t\circ Y_t)\hat{K}^{(n)-1}(W_t\circ Y_t)^{\top} - (W_t\circ Y_t)\hat{K}^{(n)-1}\left(W_t\circ\left[\hat{R}^{(n+1)}\hat{f}_{t|T}^{(n)}\hat{C}^{(n+1)\top}\right]\right)^{\top}\\ &-\left(W_t\circ\left[\hat{R}^{(n+1)}\hat{f}_{t|T}^{(n)}\hat{C}^{(n+1)\top}\right]\right)\hat{K}^{(n)-1}(W_t\circ Y_t)^{\top}\\ &+\hat{K}^{(n)-1}\star\left[D_{W_t}\left(\hat{C}^{(n+1)}\otimes\hat{R}^{(n+1)}\right)\left(\hat{f}_{t|T}^{(n)}\hat{f}_{t|T}^{(n)\top}+\Pi_{t|T}^{(n)}\right)\left(\hat{C}^{(n+1)}\otimes\hat{R}^{(n+1)}\right)^{\top}D_{W_t}^{\top}\right]\\ &+\left[\hat{H}^{(n)}\mathbf{1}_{p_1,p_2}\hat{K}^{(n)}\right]\hat{K}^{(n)-1}\left(\mathbf{1}_{p_1,p_2}-W_t\right)^{\top}\Bigg\}_{ii}\end{aligned}$$

$$\begin{aligned}\left[\hat{K}^{(n+1)}\right]_{ii} = \frac{1}{Tp_1}\sum_{t=1}^{T}\Bigg\{&(W_t\circ Y_t)^{\top}\hat{H}^{(n+1)-1}(W_t\circ Y_t) - (W_t\circ Y_t)^{\top}\hat{H}^{(n+1)-1}\left(W_t\circ\left[\hat{R}^{(n+1)}\hat{f}_{t|T}^{(n)}\hat{C}^{(n+1)\top}\right]\right)\\ &-\left(W_t\circ\left[\hat{R}^{(n+1)}\hat{f}_{t|T}^{(n)}\hat{C}^{(n+1)\top}\right]\right)^{\top}\hat{H}^{(n+1)-1}(W_t\circ Y_t)\\ &+\hat{H}^{(n+1)-1}\star\left[D_{W_t}^{\top}\left(\hat{R}^{(n+1)}\otimes\hat{C}^{(n+1)}\right)K_{k_1,k_2}\left(\hat{f}_{t|T}^{(n)}\hat{f}_{t|T}^{(n)\top}+\Pi_{t|T}^{(n)}\right)K_{k_1,k_2}^{\top}\left(\hat{R}^{(n+1)}\otimes\hat{C}^{(n+1)}\right)^{\top}D_{W_t}^{\top}\right]\\ &+(\mathbf{1}_{p_1,p_2}-W_t)^{\top}\hat{H}^{(n+1)-1}\left(\hat{H}^{(n+1)}\mathbf{1}_{p_1,p_2}\hat{K}^{(n)}\right)\Bigg\}_{ii}\end{aligned}$$

3.4 Nowcasting

Over the past decade, real-time forecasting—commonly referred to as *nowcasting*—has gained increasing relevance in empirical macroeconomics. The foundational framework was introduced by Giannone, Reichlin, and Small 2008, who proposed a two-step approach combining principal component analysis and Kalman filtering to manage unbalanced data releases in real time. Most applications of Dynamic Factor Models (DFMs) for nowcasting have been developed in the vector setting.

The main contribution of this work is to extend that framework to *matrix-variate* time series. By doing so, we evaluate the effectiveness of the Dynamic Matrix Factor Model (DMFM) described in previous sections, and assess whether it can outperform traditional vector-based DFMs in nowcasting performance.

Nowcasting involves predicting low-frequency target variables by leveraging the information contained in high-frequency indicators, which are released asynchronously over time. In this thesis, the original structure proposed by Giannone, Reichlin, and Small 2008 is adapted to the matrix setting, modifying the way nowcasts are extracted. Specifically, we build on the methodology of Bańbura and Modugno 2014, where nowcasts are implicitly derived from data reconstructed via the estimated latent factors. This approach is here extended to the matrix framework, explicitly accounting for the *jagged-edge* nature of macroeconomic datasets—where variables are released at different points in time—while maintaining the data in their natural matrix organization.

Consider, for instance, a quarterly variable embedded within a monthly dataset: whenever the quarterly observations are not yet available, they can be treated as missing and predicted using the latent factor dynamics. This takes advantage of both cross-sectional and temporal dependencies among monthly and quarterly indicators.

To implement this, we extend the procedure of Bańbura and Modugno 2014 to the matrix-variate setting. Vectorizing the DMFM defined in Equation 3, the observation equation becomes:

$$\text{vec}(Y_t) = (C \otimes R) \cdot \text{vec}(F_t) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, K \otimes H),$$

while the latent factor dynamics, assuming a matrix autoregressive process, are given by:

$$\text{vec}(F_t) = \Phi \cdot \text{vec}(F_{t-1}) + u_t, \quad \Phi = B \otimes A.$$

Once the model is estimated via the EM algorithm, nowcasts are obtained through Kalman filtering applied to the vectorized state-space representation. As described in Paragraph 3.2.2, the Kalman filter computes the expected values and second moments of the latent factors, and updates them as new data become available. The Kalman smoother then refines these estimates by applying a backward recursion across the sample.

From the smoothed factors, the predicted values of the original matrix-valued variable can be reconstructed as:

$$\hat{Y}_{t|T} = R \cdot \text{mat}(\hat{f}_{t|T}) \cdot C^\top.$$

However, for the most recent observations—where no future data are yet available—the Kalman smoother cannot operate, and the nowcasts coincide with the output of the Kalman filter. These constitute the real-time estimates we are interested in.

This procedure can be implemented recursively. At each data vintage v_j, the model is re-estimated and the Kalman filter is used to update the nowcast of the target variable. When the next vintage v_{j+1} becomes available, the information set expands with new releases—either new monthly observations, new quarterly values, or revisions of previously released data. Repeating this process over successive vintages simulates real-time forecasting and revision. The result is a time series of nowcasts that can be directly compared to the realized values of the variable of interest.

Recursive Nowcasting Procedure with DMFM

For each information set (vintage) at time t_v:

1. Construct the observed tensor $Y^{(v)} = \{Y_1, \ldots, Y_{t_v}\}$ and the corresponding availability mask $W^{(v)}$, according to the data release schedule;

2. Estimate the DMFM parameters using the Expectation-Maximization (EM) algorithm, accounting for missing values through the observation mask;

3. Reconstruct the completed dataset as:

$$\hat{Y}_t = R f_t C^\top$$

   where f_t are the filtered latent factors at time t_v, and R, C are the estimated loading matrices;

4. Generate the nowcast for the quarterly variable $y^q_{t_v+1}$:

   - If $y^q_{t_v+1}$ is not yet released, store the estimate $\hat{y}^q_{t_v+1|t_v}$ as the nowcast;
   - If $y^q_{t_v+1}$ has been released, compute the forecast error as:

$$\text{Error}_{t_v+1} = y^q_{t_v+1} - \hat{y}^q_{t_v+1|t_v}$$

4 Empirics

This chapter represents the core contribution of this work, as it empirically assesses the validity of the methodology provided in Chapter 3 by applying the estimation and nowcasting techniques to Euro Area data modeled through a DMFM. The central research question is whether the DMFM, using factors extracted from a set of monthly indicators along with quarterly GDP, can outperform its vector counterpart in terms of nowcasting performance.

In doing so, the approach offers several layers of flexibility, making the methodology well-suited for real-world applications, as it (*i*) preserves the original structure of the data organized as a matrix-variate time series; (*ii*) accommodates missing values arising from the mixed-frequency nature of the dataset and the treatment of real variables during the COVID-19 period; (*iii*) exploits the asynchronous release of macroeconomic indicators, enabling sequential updates of nowcasts as more recent information becomes available over time within the quarter.

The first step in the empirical analysis consists in constructing the matrix-variate time series, namely the lowest-rank tensor where two dimensions are observed over time. Hence, using the dataset by M. Barigozzi and Lissona 2024, I first extracted a subset of countries and variables according to selection criteria detailed in Subsections 4.1.1 and 4.1.2. Then I built a matrix-variate data where countries are arranged as rows and the macroeconomic variables specific to each country as columns. This bilinear structure is repeated over time, reflecting how data are naturally collected within a Monetary Union. The empirical analysis presented here focuses on Germany, France, Italy, and Spain. For each of these EA economic drivers, I considered the full set of monthly indicators along with GDP as the key national account variable, amounting to a total of 40 indicators[3].

The selected Euro Area (EA) economies, as members of a Monetary Union, collect comparable macroeconomic variables in a harmonized fashion. As key drivers of Euro Area activity, these countries are plausibly linked by common latent factors, whose interpretation is discussed in Subsection 4.2. Confirming this idea throughnsome anecdotal evidence presented in Subsection 4.1, and building on a formal theoretical framework, this chapter aims to evaluate the benefits of preserving the matrix structure for real-time forecasting. To this end, the forecasting performance of matrix-based DFMs is compared with their vector-based counterparts. The empirical results are analyzed in Subsection 4.3.1. In practice, the exercise involves recursively truncating the dataset month by month and applying a release mask that reflects the actual publication calendar and frequency of each variable. Once the DMFM is estimated, regardless of whether the specification is matrix-based or vector-based, model's parameters are obtained via Quasi-Maximum Likelihood (QML) using the Expectation-Maximization (EM) algorithm. Quarterly nowcasts for target variable, namely national GDP, are then extracted using the Kalman Filter. The EM algorithm plays a central role in this approach, as it enables consistent parameter estimation in the presence of missing values, which naturally arise from mixed data frequencies and from the masking of real variables during the COVID-19 period. Combined with the Kalman filter, this strategy offers a robust solution for handling the asynchronous release of information and producing reliable nowcasts.

[3] From the replication code in R available at: https://github.com/dolpolo/Nowcasting_EA_DMFM, this framework can be extended to include more variables and countries, up to the limits of the dataset.

While this recursive nowcasting approach has become standard in vector-based DFM implementations, to the best of the author's knowledge, this work represents one of the first applications of a nowcasting framework to tensor-structured data, even in its simplest three-dimensional form of matrix-variate time series. The empirical results suggest that the matrix-based approach outperforms traditional vector-based Dynamic Factor Models.

To summarize, the research questions that guide this empirical investigation are directed to understand whether co-movements among Euro Area countries can be exploited to improve the nowcast of an individual EA country's GDP, and if this can be achieved nowcasting high-frequency indicators, such as GDP, just extracting factors from a medium-sized set of monthly indicators. In other words what follows can be read keeping in mind the following question: can the inclusion of monthly indicators and cross-country information in a Matrix-varate setup improve the accuracy of GDP nowcasts for individual EA countries?

4.1 Data

The dataset used for this analysis is the result of a country and variable selection from the "EA-MD-QD" dataset by M. Barigozzi and Lissona 2024. In this subsection, I discuss the selection criteria applied to extract the final set of EA countries and macroeconomic indicators, along with the procedure used for data preparation.

The original dataset includes quarterly and monthly macroeconomic indicators from January 2000 to January 2025, recorded at monthly frequency, yielding a total of $T = 300$ observations for ten Euro Area member countries: Germany, France, Italy, Spain, Netherland, Belgium, Portugal, Austria, Ireland, and Greece. During this period, Europe experienced three major global shocks: the 2008 Global Financial Crisis, the 2012 Sovereign Debt Crisis, and the COVID-19 pandemic in 2020. Each of these shocks differs in origin and nature, with COVID-19 being a purely exogenous event that poses unique challenges for empirical modeling. To avoid the introduction of noise during this period, real variables have been masked and replaced with filtered forecasts obtained via the Kalman filter.

The core application presented in this work focuses on the first four countries, which are considered the primary drivers of the Euro Area business cycle. From the full set of available macroeconomic indicators, the analysis retains all monthly variables and includes GDP as the only quarterly variable, which also serves as the nowcasting target. The structure of the resulting dataset, in terms of variable selection, aims to replicate that used in Giannone, Reichlin, and Small 2008. Moreover, it is similar to the dataset adopted by Cascaldi-Garcia et al. 2024, although the model specification differs, as they construct a multi-country framework to nowcast economic conditions in the Euro Area and the same set of member countries, excluding Spain.

For the implementation of the methodology, the variables are required to be stationary, with zero mean and unit variance. Stationarity can be automatically imposed during the data construction phase using tools provided by M. Barigozzi and Lissona 2024. Centering and standardization are initially applied, and then restored after the estimation procedure.

To summarize, after the selection phase, the stationary time series are arranged in matrix format and

stored in a three-dimensional tensor. The first dimension corresponds to countries (rows), the second to variables (columns), and the third to time (slices). The resulting matrix-valued time series contains $T = 300$ monthly observations for $p_1 = 4$ countries and $p_2 = 40$ variables. Within this dataset, as expected, two main sources of missing data are observed: (*i*) blocks of missing values for real variables during the COVID-19 period, and (*ii*) a regular pattern of missing entries for lower-frequency variables (e.g., GDP), consistent with a mixed-frequency framework.

4.1.1 Country Selection

The countries considered in this analysis are Germany, France, Italy, and Spain. These economies were selected because they are the largest within the Euro Area and together account for the majority of its aggregate output, as shown in Figure 1. All ten countries included in the "EA-MD-QD" dataset are highlighted in red and colored accounting for their aggregate output contribution according to Eurostat data.

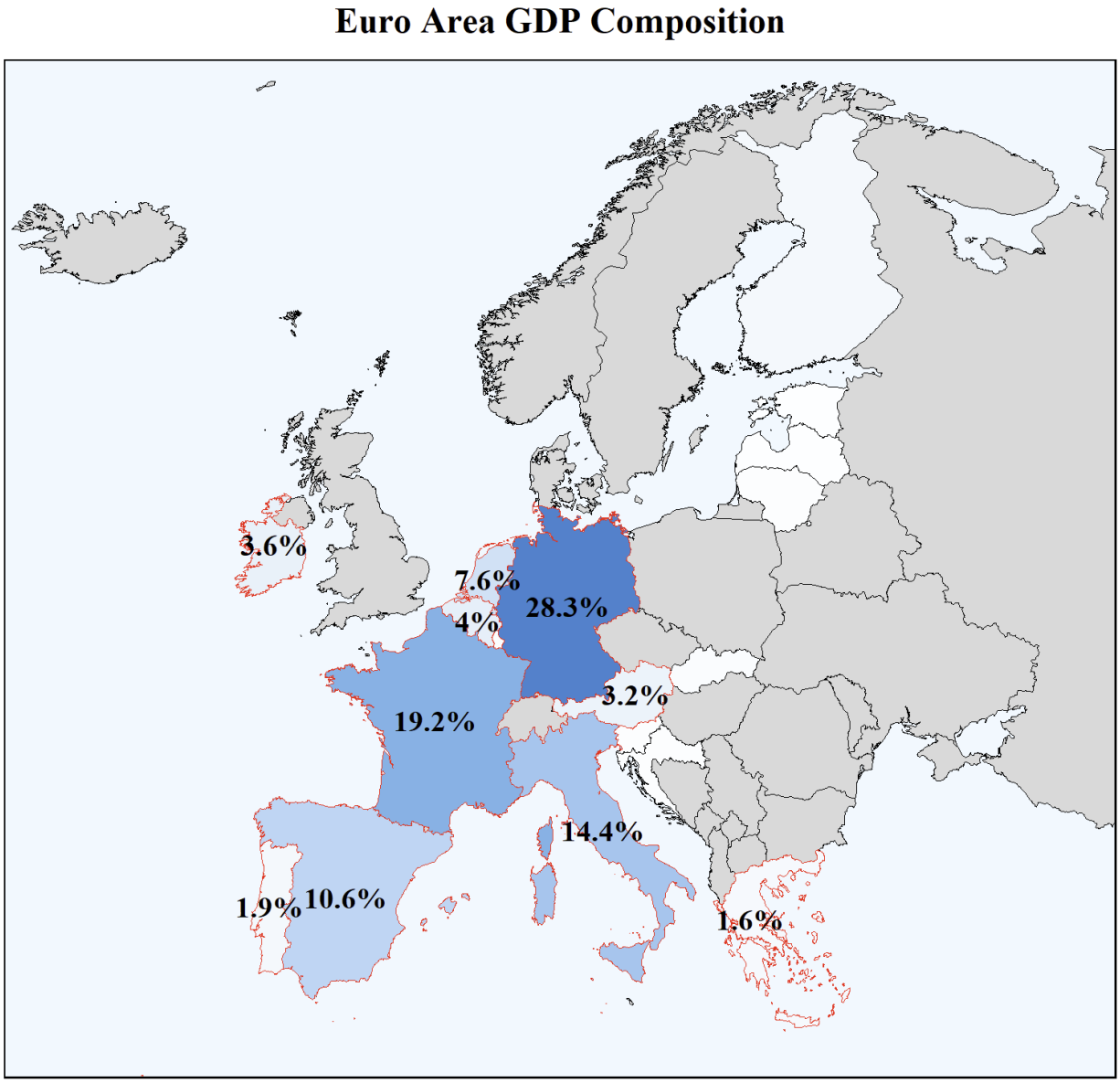
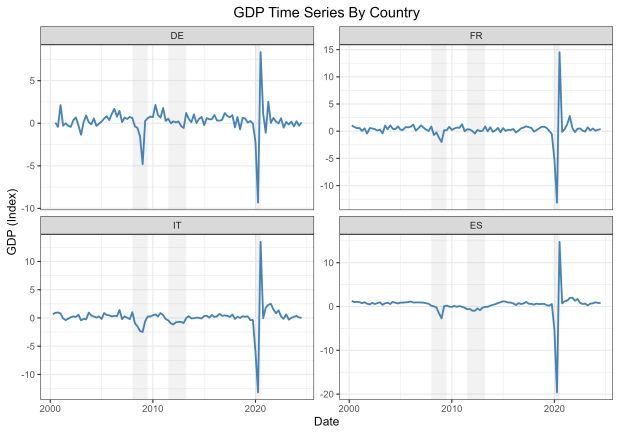


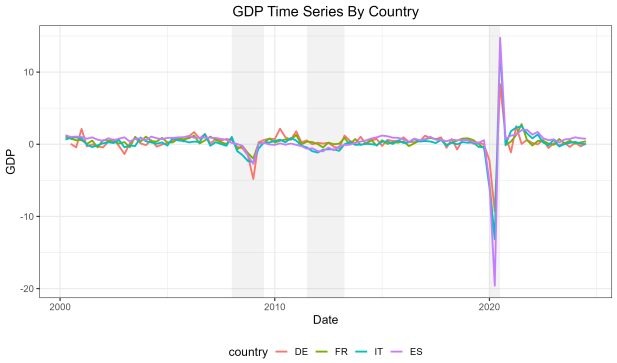
Figure 1: Share of Euro Area GDP by country, based on Eurostat data.

As shown in the figure, Germany is the largest contributor, followed by France, Italy, and Spain.

In addition to their economic relevance, these countries exhibit strong co-movements in their business cycles, providing anecdotal but strong evidence of shared latent structures. This is illustrated in Figures 2a and 2b, which display the individual and overlapping quarterly GDP series for the selected countries.



(a) Quarterly GDP series



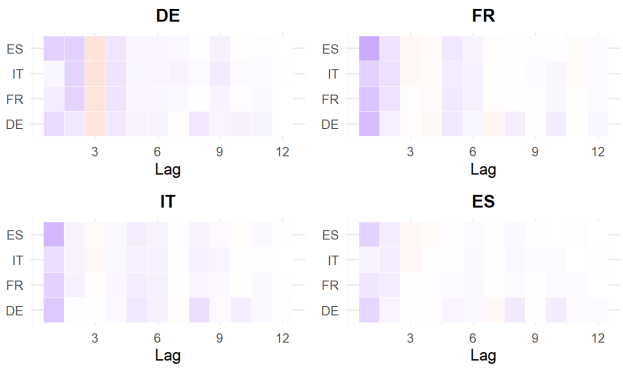
(b) Overlapping GDP series

Figure 2: Visual comparison of GDP dynamics across selected Euro Area countries.

This co-movement is further supported by high pairwise correlations among the GDP series, as reported in Table 3a. All selected countries exhibit strong contemporaneous linear dependence in their business cycles, with correlation coefficients consistently above 0.85. However, these relationships weaken substantially once lags are introduced. As shown in Figure 3b, correlations drop significantly with a one-quarter lag and become negligible at one-year lags. This empirical evidence highlights the importance of modeling contemporaneous cross-country dependencies, one of the key strengths of the DMFM framework adopted in this study. These results support the existence of a common latent factor structure that is autocorrelated up to few lags, justifying the modeling choices made in subsequent sections.

	DE	FR	IT	ES
DE	1.00	0.88	0.88	0.86
FR	0.88	1.00	0.97	0.95
IT	0.88	0.97	1.00	0.96
ES	0.86	0.95	0.96	1.00

(a) Contemporaneous GDP Correlation



(b) Lagged GDP Correlations (One-Year Lag)

Figure 3: Cross-country correlations in GDP series: contemporaneous (left) and lagged (right).

4.1.2 Variables Selection

From the extensive set of monthly and quarterly variables in the dataset developed by M. Barigozzi and Lissona 2024, only a subset has been selected for the present analysis.

The final selection includes 39 monthly indicators along with GDP, the only quarterly variable, which also serves as the target for the nowcasting exercise. This composition of $p_2 = 40$ variables mirrors the empirical setup adopted by Giannone, Reichlin, and Small 2008, where the authors rely solely on monthly indicators and GDP. This modeling choice directly addresses the first core research question, namely whether factors extracted from monthly indicators alone are sufficient to capture GDP fluctuations in Euro Area countries reguardless the vector or matrix-based DFM used. Additionally, this medium-scale design ensures that the nowcasting procedure remains computationally feasible within the DMFM framework. It is worth noting that opting for a medium-sized dataset does not imply a limitation: previous literature, including Bańbura and Modugno 2014 and Cen and Lam 2025, has demonstrated that small to medium models often outperform larger ones in terms of forecasting accuracy.

The selected variables span the period from February 2000 to January 2025, for a total of $T = 300$ months, thus covering also the COVID-19 shock. The main challenge posed by the pandemic is the disruption of the economic cycle, which affected real variables in particular. As a result, real variables reporting significant outliers during the pandemic have been masked in the dataset for the affected period. Accounting for missing data due to both the mixed-frequency structure of the dataset and the masking of COVID-distorted variables, the final dataset features a missing value rate of 4.7%.

Table 4.1.2 provides an overview of the selected indicators, highlighting their key characteristics, including class, frequency, transformation, and release delays. The column "Delay" reports the average number of days between the end of the reference period and the release date of each variable. For instance, GDP, which is reported quarterly, is typically released 45 days after the end of the reference period. Hence, the GDP for the first quarter is usually published around mid-April, i.e., approximately 45 days after March 31st.

The most timely indicators are survey-based confidence indicators, typically published within the same month. These are commonly referred to as *soft data*. While less precise than quantitative (*hard*) data such as GDP, CPI, or industrial production, soft data are available earlier and are particularly useful for nowcasting purposes. As noted by Bańbura and Modugno 2014 and Giannone, Lenza, and Primiceri 2021, central banks often rely on such indicators to detect turning points in the business cycle. In general, soft data capture expectations and sentiment, whereas hard data are objective but subject to publication delays and subsequent revisions.

Table 4.1.2 also provides additional information on the variables. The "Class" column indicates whether a variable is real (and thus masked during the COVID-19 period, specifically from Q1 2020 to Q2 2021). The "Category" distinguishes soft from hard data. The "Frequency" column identifies monthly indicators. Finally, the "Transformation" column specifies the transformation applied to each series to ensure stationarity. For example, GDP is transformed using the logarithmic first difference scaled by 100, $100 \times \Delta \log(x_t)$, which approximates the quarterly percentage change. Acronyms and transformations are explained in the Appendix.

Lastly, anecdotal evidence supports the existence of common latent structures among the EA variables. Figure 4 displays a heatmap of pairwise correlations among the full set of variables in the dataset of M. Barigozzi and Lissona 2024. The high level of correlation observed among many indicators reinforces the hypothesis of an underlying common factor structure, further motivating the use of factor models in the empirical analysis.

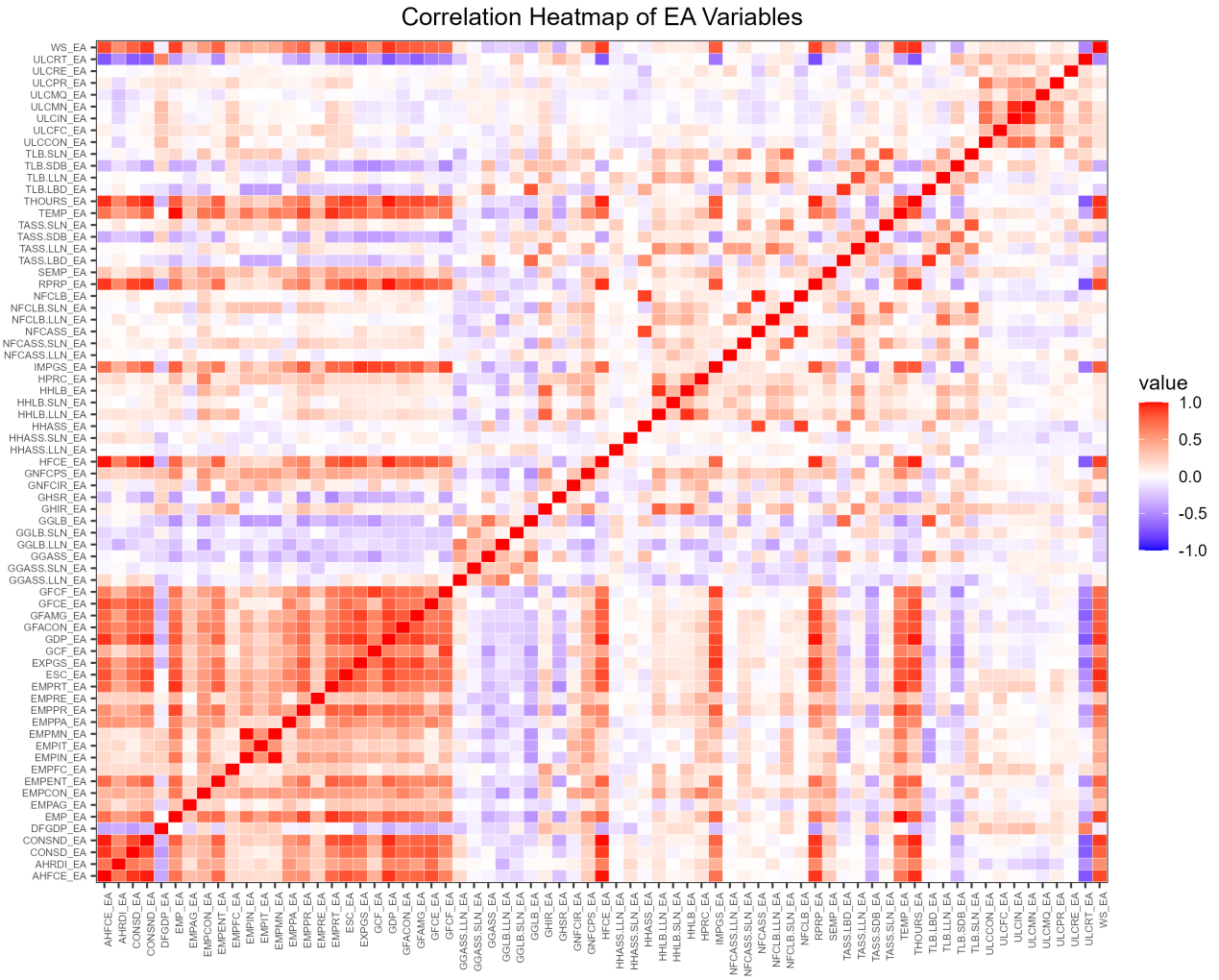


Figure 4: Heatmap of correlations among M. Barigozzi and Lissona 2024 variables for the Euro Area

As a Final remark, through the R code, a different set of variables can be selected based on their correlation with GDP. For each country, a set of variables is extracted according to their correlation strength. These are then aggregated with the variables most correlated with the GDP of other countries (e.g., selecting the top 5 variables per country and aggregating them across countries considered for each country also then 5 variables selected for the others if different). Through this replication code is also possible to select a different set of EA countries.

Table 2: Macroeconomic variables selected and release delays

N	ID	Series Name	Class	Category	Transformation	Freq.	Delay
(1) National Accounts / Real Economy							
1	GDP	Real Gross Domestic Product	R	H	1	Q	45
(2) Labor Market							
2	UNETOT	Unemployment: Total	R	H	0	M	45
3	UNEU25	Unemployment: Under 25 years	R	H	0	M	45
4	UNEO25	Unemployment: Over 25 years	R	H	0	M	45
(4) Exchange and Interest Rates							
5	REER42	Real Exchange Rate (42 countries)	F	H	1	M	35
6	LTIRT	Long-Term Interest Rates (EMU)	F	H	2	M	35
(5) Industrial Production and Turnover							
7	IPMN	Industrial Production Index: Manufacturing	R	H	1	M	45
8	IPING	Industrial Production Index: Energy	R	H	1	M	45
9	IPCAG	Industrial Production Index: Capital Goods	R	H	1	M	45
10	IPDCOG	Industrial Production Index: Durable C. Goods	R	H	1	M	45
11	IPNDCOG	Industrial Production Index: Non-Durable C.G.	R	H	1	M	45
12	IPCOG	Industrial Production Index: Consumer Goods	R	H	1	M	45
13	IPNRG	Industrial Production Index: Energy	R	H	1	M	45
14	TRNING	Turnover Index: Intermediate Goods	R	H	1	M	45
15	TRNDCOG	Turnover Index: Durable Consumer Goods	R	H	1	M	45
16	TRNCAG	Turnover Index: Capital Goods	R	H	1	M	45
17	TRNNRG	Turnover Index: Energy	R	H	1	M	45
18	TRNCOG	Turnover Index: Consumer Goods	R	H	1	M	45
19	TRNNDCOG	Turnover Index: Non-Durable C. Goods	R	H	1	M	45
(6) Prices							
20	PPIING	Producer Price Index: Intermediate Goods	N	H	1	M	40
21	PPINRG	Producer Price Index: Energy	N	H	1	M	40
22	PPICAG	Producer Price Index: Capital Goods	N	H	1	M	40
23	PPICOG	Producer Price Index: Consumer Goods	N	H	1	M	40
24	PPINDCOG	Producer Price Index: Non-Durable C. Goods	N	H	1	M	40
25	PPIDCOG	Producer Price Index: Durable C. Goods	N	H	1	M	40
26	HICPNG	HICP: Energy	N	H	1	M	40
27	HICPNEF	HICP: All Items excl. Energy and Food	N	H	1	M	40
28	HICPIN	HICP: Industrial Goods	N	H	1	M	40
29	HICPOV	HICP: Overall Index	N	H	1	M	40
30	HICPG	HICP: Goods	N	H	1	M	40
31	HICPSV	HICP: Services	N	H	1	M	40
(3) Confidence Indicators							
32	ICONFIX	Industrial Confidence Indicator	C	S	0	M	5
33	CCONFIX	Consumer Confidence Indicator	C	S	0	M	5
34	KCONFIX	Construction Confidence Indicator	C	S	0	M	5
35	SCONFIX	Services Confidence Indicator	C	S	0	M	5
36	ESENTIX	Economic Sentiment Indicator	C	S	0	M	5
37	RTCONFIX	Retail Confidence Indicator	C	S	0	M	5
38	BCI	Composite Business Confidence Index	C	S	1	M	5
39	CCI	Composite Consumer Confidence Index	C	S	1	M	5
(7) Others							
40	SHIX	Share Price Index	F	S	1	M	1

4.2 Selection and Interpretation of DMFM Factors

This subsection implements the methodology discussed in Subsubsection 3.2.3 to select the number of factors, using the algorithm presented in Box 3.2.3, and provides an interpretation of these factors based on the estimated row and column loadings retrieved through the projected estimation method proposed by L. Yu et al. 2022, as detailed in Box 3.2.3. These estimated loadings are then used to initialize the EM algorithm, offering a reliable approximation of a local maximum and thereby reducing the number of iterations required for the EM convergence.

As discussed in Subsections 4.1.1 and 4.1.2, anecdotal evidence supports the presence of strong co-movements across both countries and variables, suggesting a low-rank latent structure. To capture this structure, I apply the eigendecomposition, based method proposed by L. Yu et al. 2022. However, since standard principal component (PC) methods are not applicable in the presence of missing data, the first step involves imputing the unavailable values. This is done using the strategy developed by Cen and Lam 2025, an extension of the "all-purpose estimator" to matrix-variate time series proposed by Xiong and Pelger 2023.

The first step involves selecting the appropriate number of factors by initially specifying a relatively large number of row and column components. Using Algorithm 3.2.3, convergence was immediately reached at one row factor ($k_1 = 1$) and one column factor ($k_2 = 1$). Figure 5 displays the eigenvalue spectrum, clearly showing a pronounced drop between the first and second components for both dimensions. This behavior supports the chosen dimensionality and is consistent with heuristic methods such as the elbow criterion. Cumulative variance is often less informative and relying on the eigenvalue ratio provides a more robust basis for factor selection.

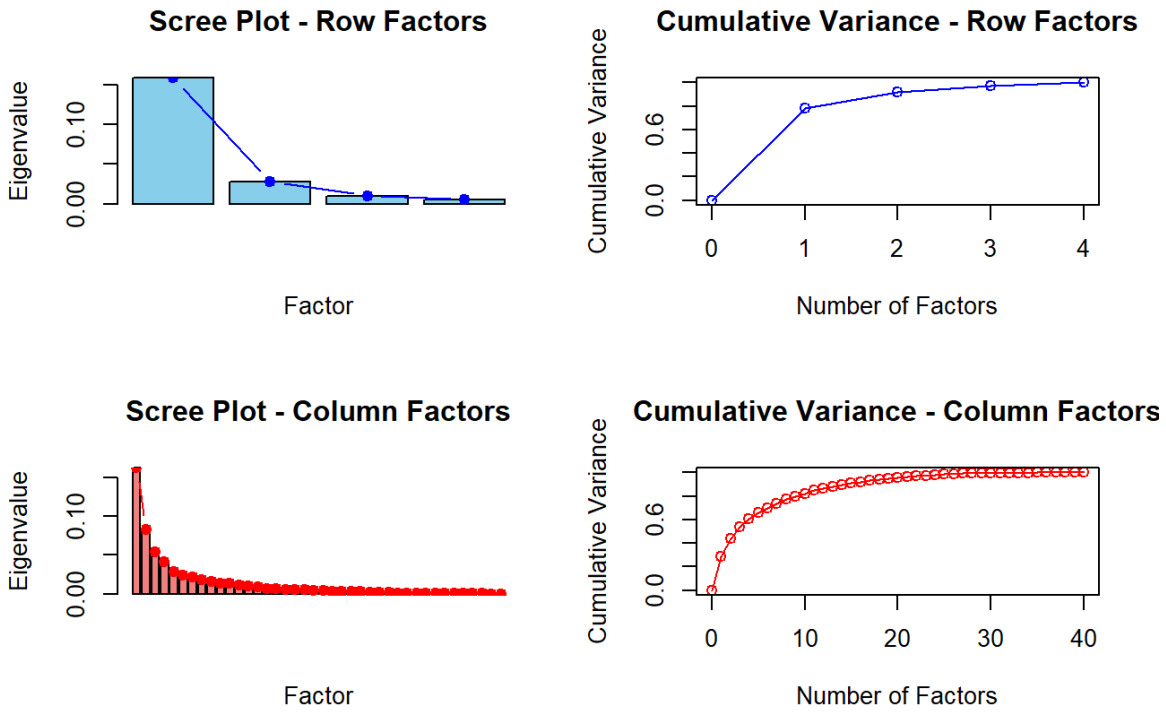


Figure 5: Scree plot and cumulative variance for row and column factor

After determining the number of factors, the next step is to estimate the row and column loadings. According to L. Yu et al. 2022, a single iteration typically suffices to yield reliable estimates. In this application, these first iteration estimates serve as an effective initialization for the EM algorithm.

Row loadings, in Table 3, have uniform values across Germany, France, Italy, and Spain, suggesting some mean effect of Euro Area on these countries. The row factor is indeed interpreted as a "Euro Area Membership Factor", representing general co-movement linked to their integration within the Monetary Union.

Table 3: Estimated Row Loadings

Country	DE	FR	IT	ES
Loading	0.899	1.078	1.024	0.990

In contrast, the column loadings reported in Table 4 rather than a spatial interpretation, convay an economic latent dynamic. Variables with high positive loadings include confidence indicators and GDP, while labor market indicators, such as unemployment, exhibit negative loadings. When the factor is high, confidence and output rise while unemployment falls. Accordingly, this latent component is interpretable as the "Business Cycle Factor". Figure 6 graphically convays the same idea by sorting the variables according to the impact of the factor.

Table 4: Estimated Column Loadings

Variable	**Loading**	**Variable**	**Loading**
TRNING	0.975	BCI	1.885
IPMN	0.654	PPIING	1.776
IPING	0.578	ESENTIX	1.708
TRNDCOG	0.504	TRNNDCOG	0.538
IPCAG	0.546	IPNDCOG	0.291
IPDCOG	0.415	CCONFIX	0.842
TRNCAG	0.408	CCI	0.840
TRNNRG	0.747	PPINRG	1.246
IPCOG	0.379	HICPNG	1.204
TRNCOG	0.597	UNEU25	-0.760
SHIX	0.305	HICPNEF	0.571
RTCONFIX	1.017	SCONFIX	1.476
IPNRG	0.104	REER42	-0.253
ICONFIX	1.866	LTIRT	0.664
PPINDCOG	1.468	UNETOT	-0.608
PPICOG	1.481	UNEO25	-0.514
HICPSV	0.473	PPICAG	0.976
PPIDCOG	0.825	HICPIN	1.182
HICPOV	1.265	HICPG	1.287
KCONFIX	0.692	GDP	1.322

It is important to note that the model is only identifiable up to orthogonal transformations. For the purposes of interpretation, the signs of the row and column loadings are normalized to be positive, which enhances their economic interpretation.

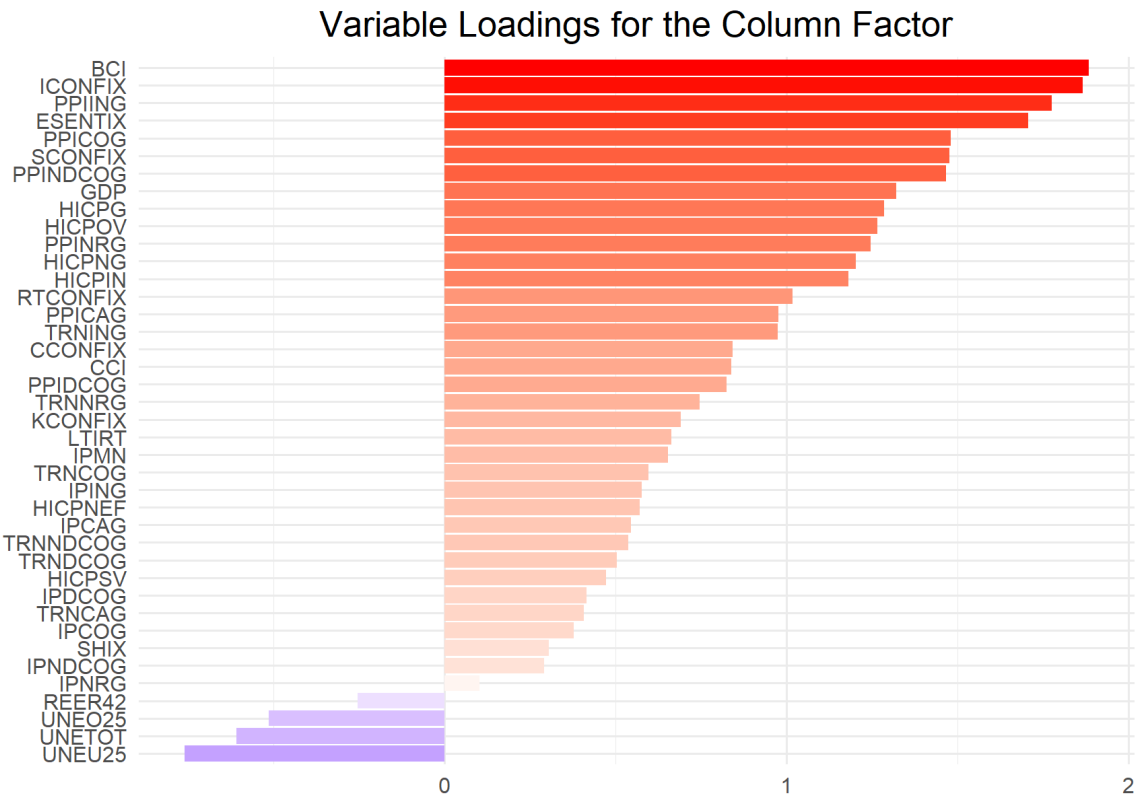


Figure 6: Column Loadings Ordered by Variable

The temporal dynamic of the extracted factor is shown in Figure 7. The plot clearly highlights its alignment with major recessionary periods in the Euro Area, including the crises of 2008, 2012, and 2020. This coherence with EA downturns reinforces its interpretation as the "Euro Area Business Cycle Factor".

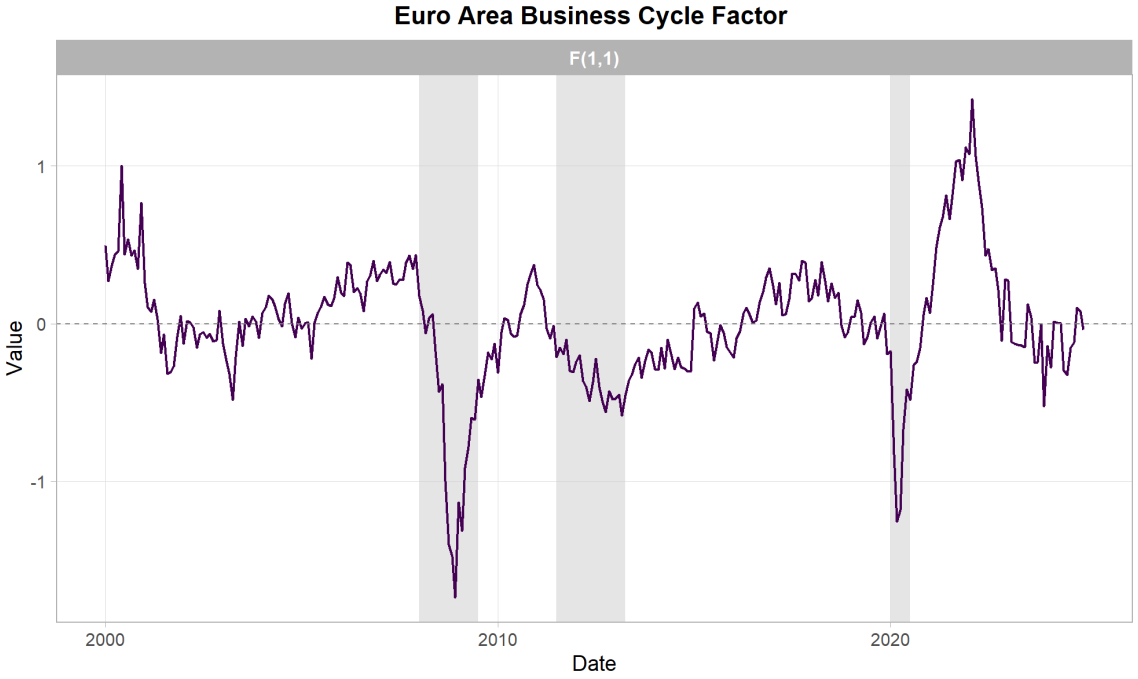


Figure 7: Estimated Euro Area Business Cycle Factor. Grey shaded areas indicate recession periods.

As this factor is derived jointly from row and column loadings, it represents a single latent component that spans both dimensions. Figure 8 illustrates its impact across countries by plotting the product of the row and column loadings. Given the similarity of row loadings, the factor's influence appears almost uniform across Germany, France, Italy, and Spain.

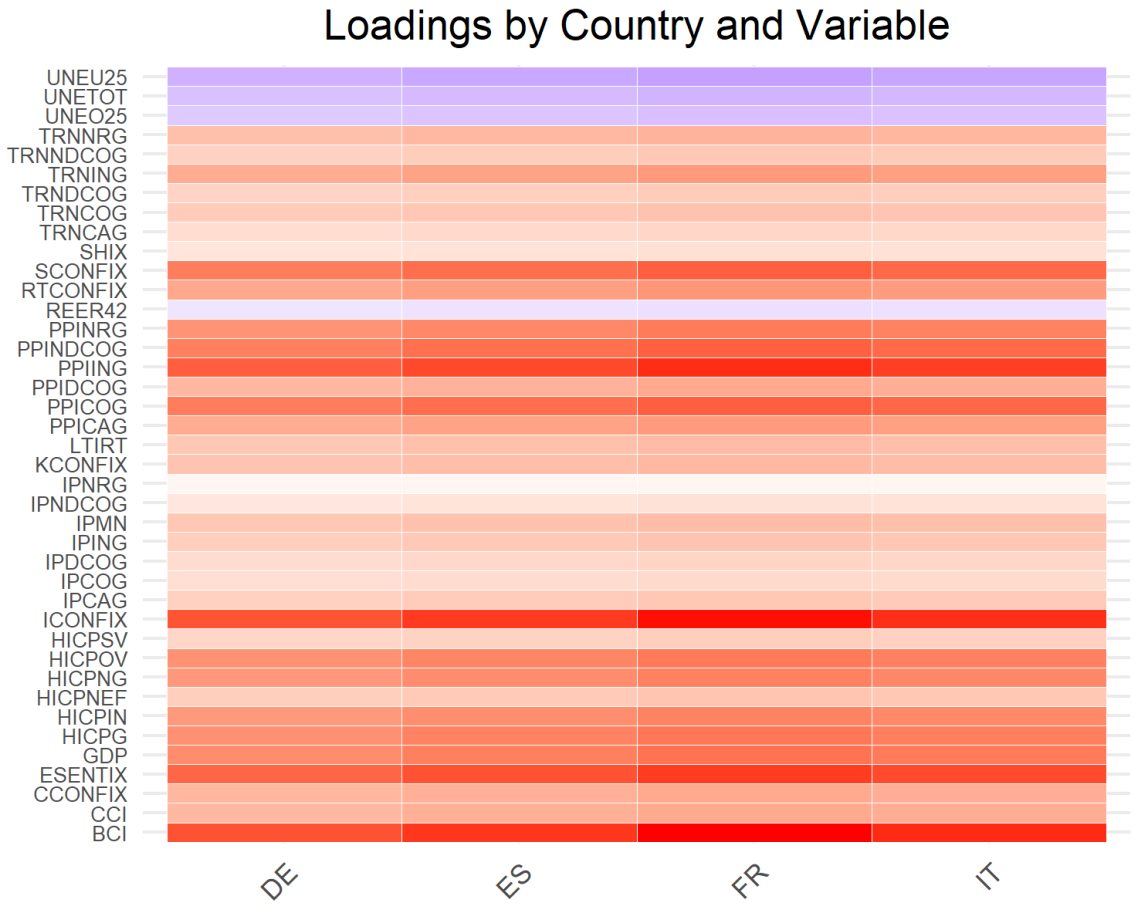


Figure 8: Combined Loadings (Row $\times$ Column)

4.3 Euro Area Nowcasting

This subsection represents the core contribution of the thesis and describes the empirical application, on Euro Area data, of the recursive nowcasting procedure discussed in subsection 3.4. This analysis compares the performance of the traditional vector-based Dynamic Factor Model (DFM) with its matrix-based extension, the Dynamic Matrix Factor Model (DMFM), to evaluate whether incorporating cross-country information can benefit real-time forecasting accuracy.

Unlike the vector DFM, which relies exclusively on indicators from the target country, the DMFM extracts common factors from a panel of indicators across multiple countries, allowing it to explicitly capture cross-country spillovers. To isolate the contribution of this additional dimension, both models are estimated under identical conditions: same dataset, frequency, release calendar, estimation method, and treatment of missing data.

Results show that both models provide decent nowcasts, confirming the sufficiency of medium-sized set of monthly indicators and GDP for real-time forecasting. However, the DMFM consistently delivers higher accuracy and greater responsiveness to new information released over the quarter, particularly during the COVID-19 crisis and recovery. These findings support the use of matrix-based models in times with strong economic interconnections, providing important implications for central banks aiming to implement more effective nowcasts for monetary policy interventions.

4.3.1 DMFM vs DFM

The comparison between the nowcasting performance of the DMFM and the DFM is organized into two parts. First, RMSFE values are reported before and after the COVID-19 shock as a measure of accuracy during stable economic times. Second, the analysis assesses how each model responds to the progressive release of information within each quarter, with a particular focus on the COVID-19 crisis and the subsequent recovery in the Euro Area.

Nowcast Accuracy The first key metric used to compare the models performances is the Root Mean Squared Forecast Error (RMSFE). Table 5 reports these values for each country and for each month within the quarter, namely $Q(0)M_1$, $Q(0)M_2$, and $Q(0)M_3$, both before and after the COVID-19 shock. Two main insights emerge: first, both models improve as more timely information becomes available within the quarter; second, the matrix formulation (DMFM) outperforms the vector model (DFM) during periods of economic stability for all EA countries considered.

Starting from the farmer result, Table 5 shows the evolution of forecast accuracy over the months of the quarter. For all countries and in both models, accuracy improves from M1 to M3, reflecting the progressive accumulation of information within the nowcasted quarter. In both models the improvement from M2 to M3 is often marginal, as M2 already achieves a level of accuracy very close to that of M3. This suggests that survey-based indicators and delayed hard data released in the second month effectively capture the bulk of relevant information for GDP forecasting. Hence, consistently with the literature, soft data emerge as reliable early signals, making also the information retrived from the first month of the quarter particularly informative.

More importantly, the table clearly highlights the superiority of the DMFM over the vector-based DFM in EA nowcasting applications. The matrix specification yields lower RMSFE values for all Euro Area countries, with the sole exception of Italy in the pre-COVID period, where the DFM performs slightly better. This minor deviation does not undermine the overall conclusion in favor of the DMFM but likely reflects a more idiosyncratic trend in Italy prior to the COVID-19 upheavals. Seemingly, for Germany, the benefits of the matrix approach are particularly evident in the post-COVID period, where the DMFM substantially reduces the RMSFE. This highlights the model's ability to capture stronger integration factors following the structural changes caused by COVID-19. On the other hand, France and Spain exhibit the most consistent improvements under the matrix model, both before and after the pandemic, suggesting that cross-country linkages are especially informative for these economies. While the vector model already performs well, the DMFM further enhances accuracy by capturing country-specific idiosyncrasies more precisely.

Overall, these results strongly support the DMFM as a superior nowcasting tool, particularly in the presence of structural breaks or increased interconnectedness, as observed after the COVID-19 crisis. This may reflect a greater integration of national business cycles, potentially driven by common shocks and the sequent coordinated policy responses, which strengthens the incentives for explicitly modeling cross-country dynamics through a matrix-based approach. This intuition is confirmed in the time series analysis presented in the following paragraph.

Table 5: RMSFE pre- and post-COVID for DMFM and DFM Across Euro Area Countries

Country	**Month**	**DMFM**		**DFM**	
		Pre-COVID	*Post-COVID*	*Pre-COVID*	*Post-COVID*
Germany	M1	**0.6469**	**0.7361**	0.6493	1.0490
	M2	**0.6373**	***0.6833***	0.6512	0.9078
	M3	***0.6289***	**0.6959**	*0.6294*	*0.8488*
France	M1	**0.3747**	**0.6481**	0.4505	0.6992
	M2	**0.3585**	**0.6098**	0.4468	*0.6861*
	M3	***0.3498***	***0.5916***	*0.4405*	0.6895
Italy	M1	0.3222	**0.8192**	**0.3031**	1.0025
	M2	*0.3084*	***0.8045***	**0.2964**	1.4274
	M3	0.3186	**0.8185**	***0.2953***	*1.3826*
Spain	M1	**0.2723**	***0.4434***	0.3882	0.6004
	M2	**0.2438**	**0.4753**	*0.3782*	0.5828
	M3	***0.2399***	**0.4775**	0.3796	*0.5761*

Note: Bold values indicate the lowest RMSFE for each month across models. Italics mark the best performance within each model (pre- and post-COVID).

Nowcast Reactivity A second result emerges when comparing the time series nowcasts produced by both formulations of DFM and true GDP for each Euro Area country. the DMFM appears significantly more responsive to the sequential release of monthly data within each quarter. This reactivity is particularly evident during the COVID-19 crisis and the subsequent recovery, when economic conditions evolved rapidly and Euro Area countries became increasingly interconnected due to the economic shock and the common monetary policies implemented.

With the DMFM and a medium-sized set of indicators, nowcasts adjust more promptly and accurately as new data become available. This is particularly valuable during downturns and recoveries, when economic signals in one country can provide useful insights for others. Indeed, a promising direction for future research would be to identify which countries, and at what times, contribute most to the update of nowcasts in the others.

Overall, while the vector model tends to produce flatter and less responsive nowcast paths, often missing key turning points within the quarter and capturing just the avarage level of GDP growth, the DMFM consistently delivers more timely and reliable updates. Combined with the superior accuracy observed under stable economic conditions, these results support the DMFM as a more effective strategy for nowcasting applications.

According to Table 5, Spain and France emerge as the countries where the matrix formulation performs best during normal times. Therefore, I begin this paragraph by presenting the time series of nowcasts from both models compared to actual GDP growth for these two countries. The figures span from 2017Q4 to 2025Q1, covering three distinct volatility regimes in the Euro Area: pre-COVID, during COVID, and post-COVID. A similar discussion is then provided for Germany and Italy, where the DMFM outperforms the vector model especially in the post-COVID period.

Spain and France Starting with `Spain`, Figure 9 shows the evolution of GDP growth across the three volatility regimes in the evaluation sample.

Before the pandemic, both models behave similarly, although the DMFM tracks actual GDP dynamics more closely. More pronounced differences emerge during the COVID-19 period, highlighted by the grey-shaded region. In this phase, the DFM responds with a notable delay and produces smooth nowcasts. It underestimates not only the severity of the contraction but also the strength of the rebound, particularly evident in the Spanish case. In contrast, the DMFM reacts more promptly and accurately, capturing both the timing and the magnitude of the economic shock, at least partially, and subsequent recovery. This enhanced responsiveness likely stems from the DMFM's ability to leverage cross-country information, where early signals from other Euro Area economies contribute to more timely and precise nowcasts. Notably, while the DFM fails to reflect the intensity of the recovery, the DMFM closely follows the sharp rebound. In the post-COVID phase, both models converge, reflecting greater stability in the economic environment.

Turning now to the second consideration, Figure 10 reports the evolution of quarterly nowcasts by month, showing how the staggered release of data refines real-time forecasts within each quarter. First-month nowcasts, that are mainly driven by high-frequency indicators such as the Share Price Index, are substantially updated in the second month, when confidence surveys and delayed hard data

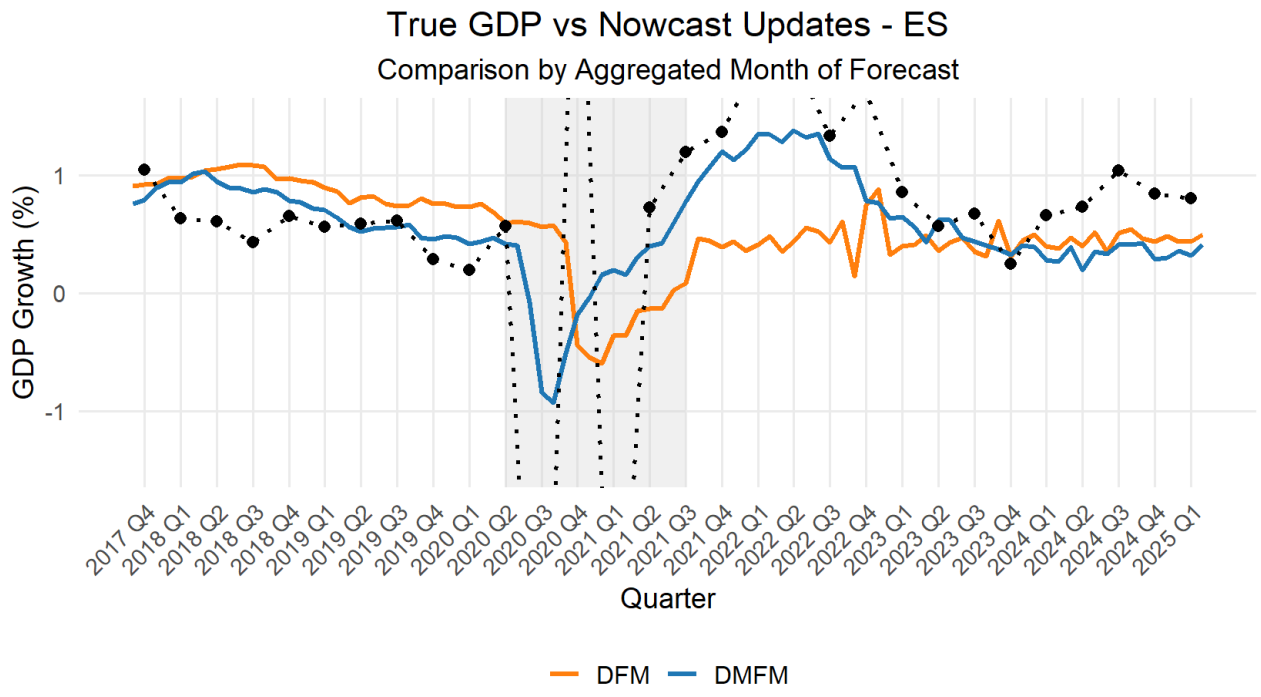


Figure 9: Nowcast comparison between DFM and DMFM (2017Q4–2025Q1) – Spain

referring to the first month become available. Third-month revisions offer marginal but still existent improvements, as most information is already incorporated in the second month data released. The graph clearly shows that at every stage, the DMFM better captures the dynamics of GDP growth in Spain, especially during the COVID downturn and the recovery phase, suggesting a strong reliance on cross-country information.

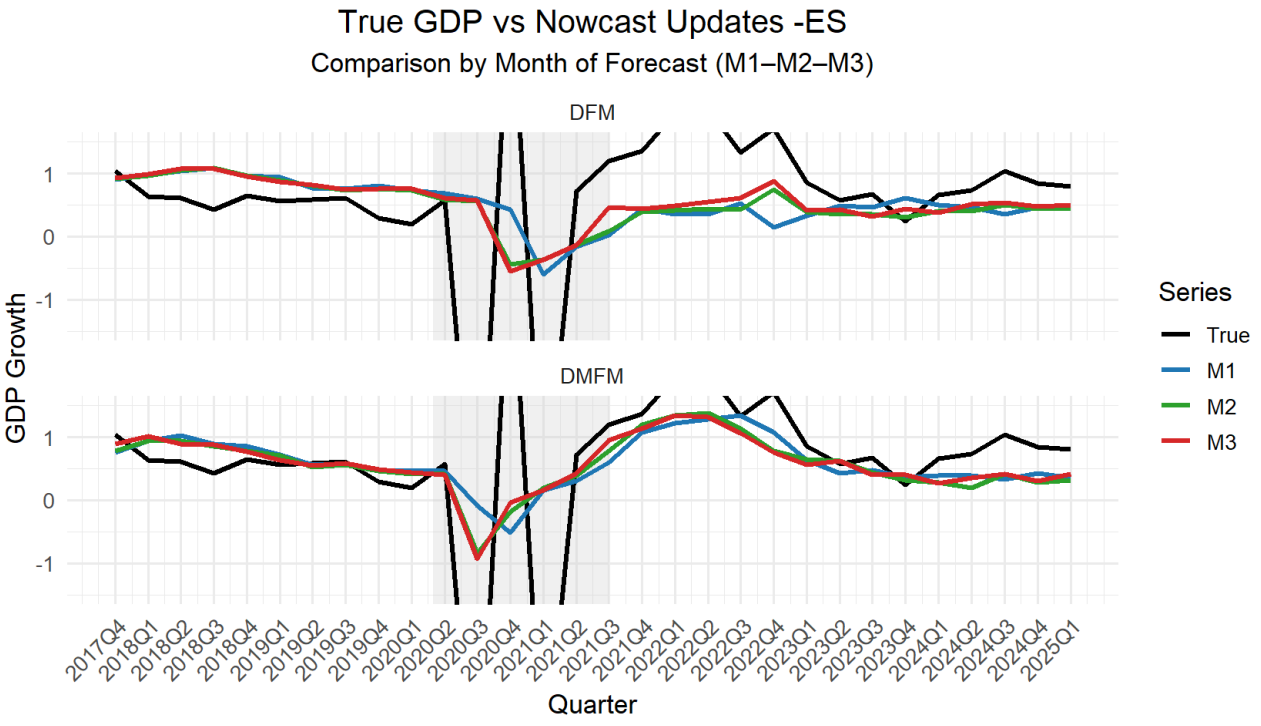


Figure 10: Monthly nowcast updates – Spain

As already discussed for Table 5, the DMFM not only better tracks the trajectory of GDP growth but also delivers a more stable forecasting performance. As shown in the boxplot below (Figures 11), the matrix model exhibits a significantly tighter distribution of forecast errors in the post-COVID period, with fewer outliers compared to the DFM.

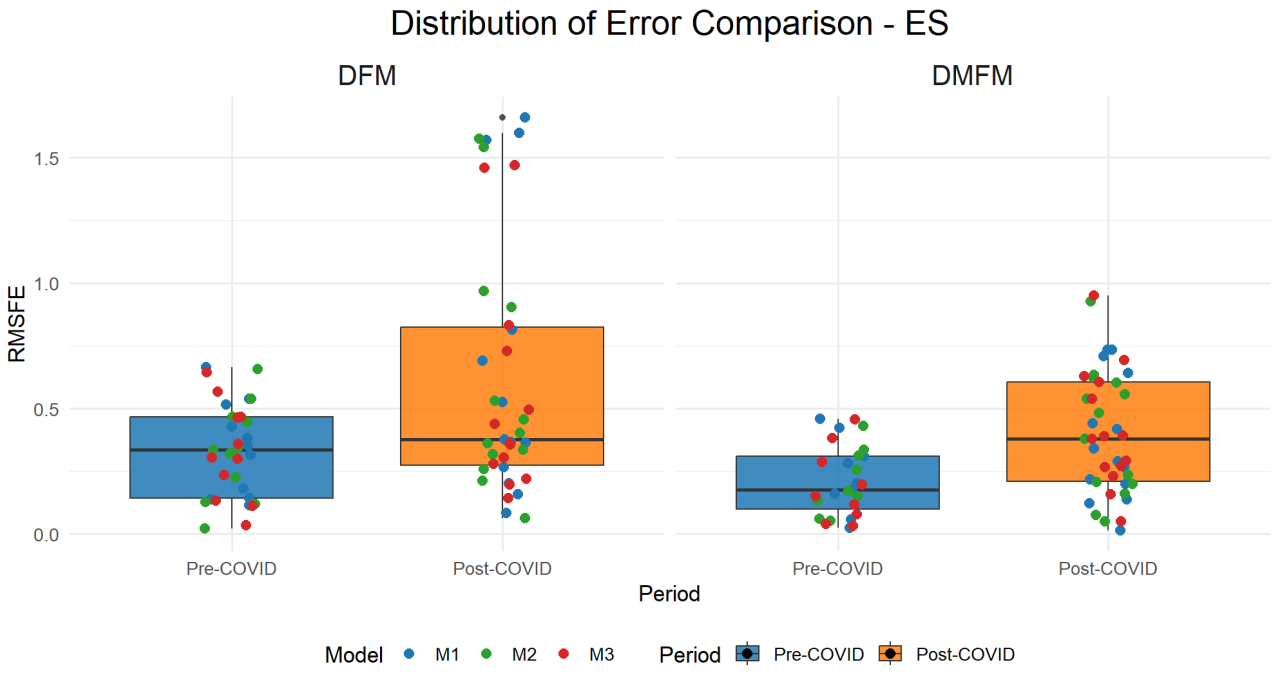


Figure 11: Error distribution comparison between DFM and DMFM Pre and Post Covid – Spain

This final evidence reinforces the conclusion that the DMFM offers significant advantages in capturing both sharp turning points and smoother trends, making it particularly well-suited for nowcasting Spanish GDP.

Seemingly for `France`, the GDP growth nowcast using the DFM provides limited informative value as the forecast series obtained through the vector-based approach are even flatter than those for Spain for all volatility regimes in the evaluation sample. However, the DMFM significantly enhances the sensitivity of nowcasts, both in terms of accuracy and reactivity to economic fluctuations.

As illustrated in Figure 12, integrating French data with information from other Euro Area countries is particularly effective. The largest improvements are observed during the COVID-19 crisis and the subsequent recovery phase. These results confirm that the superior RMSFE values achieved for France stem from the increased dynamic capabilities of the matrix-based model. On the other hand, the DFM, tends to capture only the average trend of GDP growth, whereas the DMFM can adapt more flexibly to evolving economic patterns.

Similar considerations apply to the responsiveness of monthly updates. As shown in Figure 19, in the Appendix, pre-Covid the DMFM and DFM are not able to exploits the sequential release of information, but the matrix-based formulation is more effectively in this during the pantemic and after the crises. Moreover, in both the pre- and post-COVID periods, the third month of the quarter consistently delivers higher accuracy.

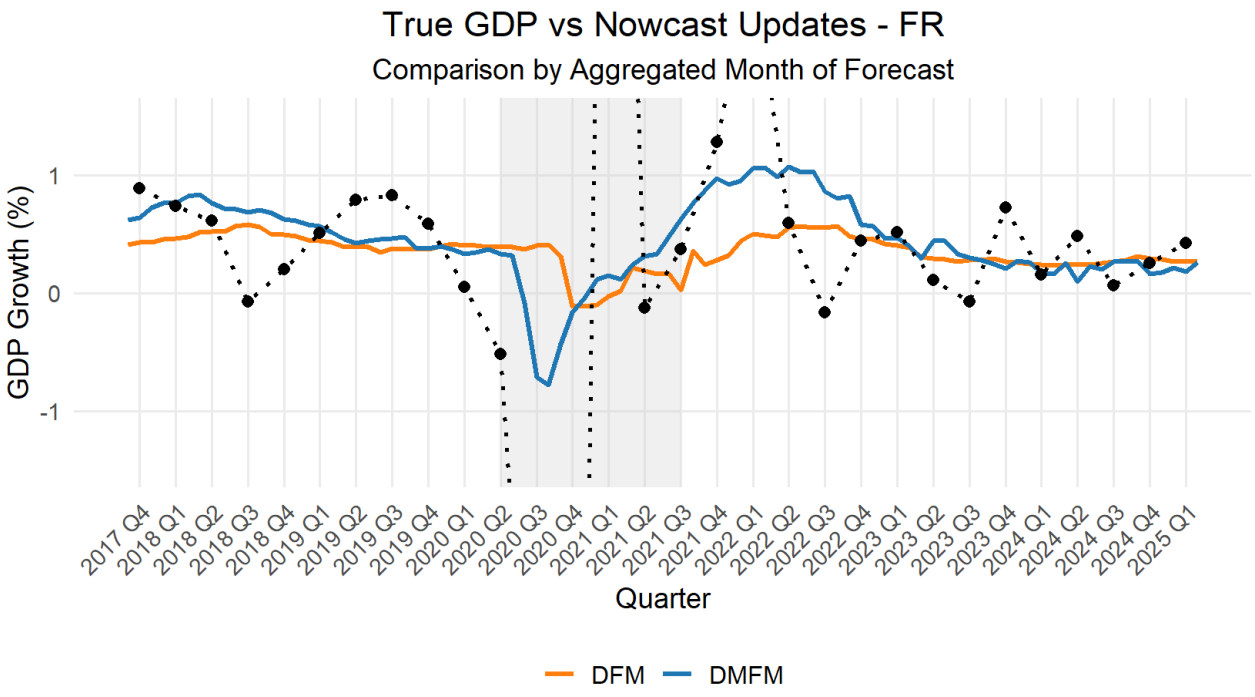


Figure 12: Nowcast comparison between DFM and DMFM (2017Q4–2025Q1) – France

Germany and Italy The DMFM consistently improves nowcast performance for France and Spain, while gains for Germany and Italy emerged mainly after the COVID-19 period.

For Germany, Figure 13 shows that DMFM nowcasts are particularly accurate during the COVID crisis, the recovery phase, and most recent months.

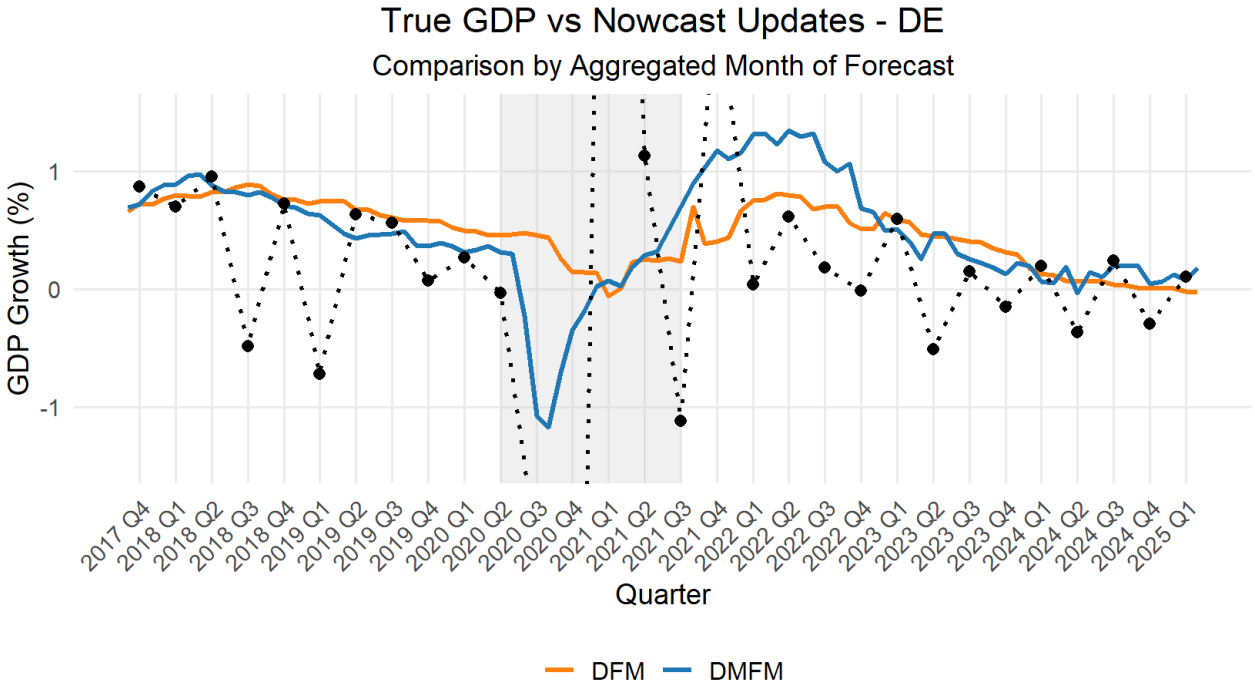


Figure 13: Nowcast time series for Germany: DMFM vs. DFM (2017Q4–2025Q1)

Similar to Spain and France, the DMFM promptly reacts to the sharp downturn during the COVID-19 crisis, whereas the vector-based DFM bearly responds with a smoothed and lagged trajectory. Unlike other Euro Area countries, Germany experiences a quicker recovery. By early 2022, GDP growth begins to stabilize, and the DMFM progressively aligns with this rebound. In contrast, the DFM remains largely unresponsive. By late 2022, the DMFM begins to closely follow the path of actual GDP and continues to deliver the most reliable forecasts throughout the last year of the sample.

Germany's strong integration within the European economy likely contributes to these improvements. Its industrial structure and macroeconomic trends are closely linked to those of neighboring countries. Thus, a joint modeling approach that incorporates cross-country information, such as the DMFM, yields more precise nowcasts, especially in a context of intensified cooperation and synchronized policy responses during the post-COVID recovery, when Germany acted as the main driver of the EA reborn.

What has been said for Germany can similarly be applied to `Italy`. Figure 14 shows that Italy also benefits from the matrix formulation, with noticeable gains in forecast accuracy emerging primarily after the pandemic. While the DFM performs adequately before COVID-19, its predictive path remains relatively flat and lacks responsiveness to new information, even during the recovery phase.

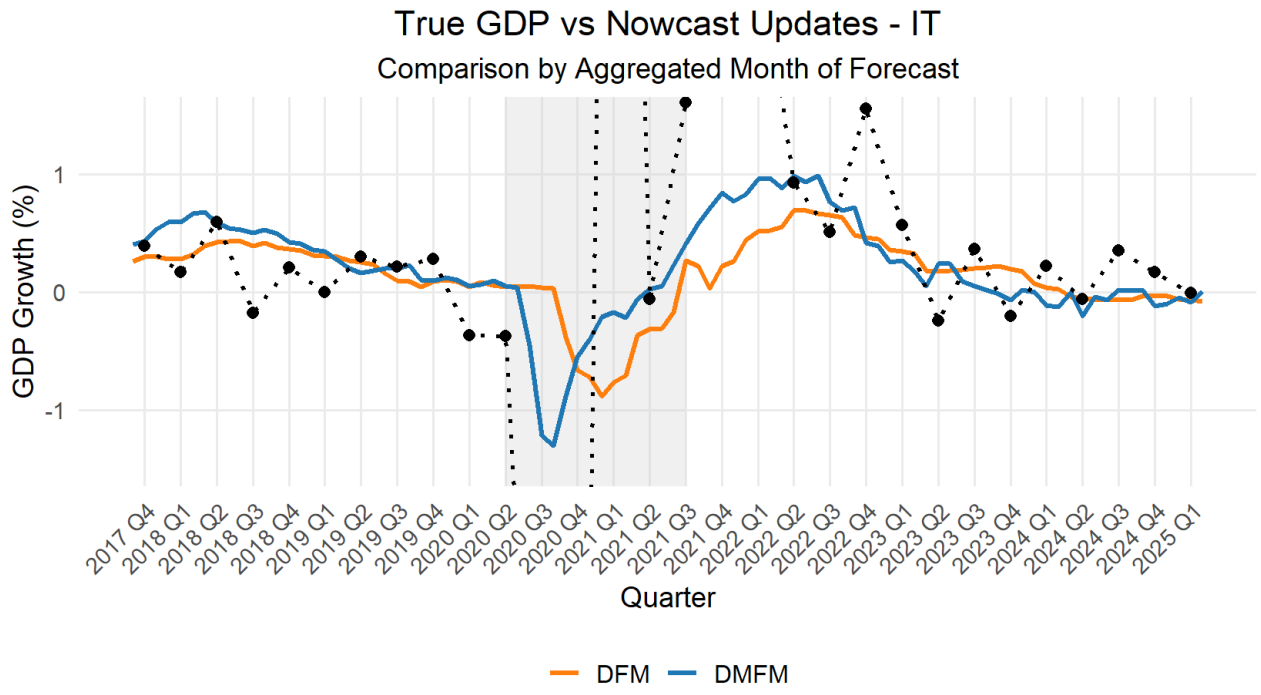


Figure 14: Nowcast comparison between DFM and DMFM (2017Q4–2025Q1) – Italy

For both Germany and Italy, during the crisis, the DFM exhibits a delayed reaction and fails to fully incorporate monthly updates. In contrast, the DMFM demonstrates a more dynamic adjustment process, particularly in the second and third months of the quarter when key data releases occur. These updates significantly refine the initial nowcasts generated in M1, underscoring the DMFM's superior capacity to integrate incoming information. This is clearly illustrated in Figures 20 and 21 in the Appendix, for Germany and Italy respectively, and as in the previous case particularly evident during the crises.

EM Algorithm discussion Some interesting insights can be drawn from the estimation results obtained in the final iteration of the recursive nowcasting procedure. As discussed in the methodological section, for each vintage in the evaluation sample—where each variable is considered according to its release schedule—the expected pseudo log-likelihood is computed using the Kalman smoother. Once this expectation is retrieved, the Kalman filter is applied to update and maximize the model parameters, and nowcasts are obtained by reconstructing the data based on the newly estimated factors and loadings. This recursive procedure, at its final iteration, incorporates the full dataset available up to the first quarter of 2025.

The smoothed estimates of GDP for each country, presented in Figure 15, closely track the observed series, reinforcing the reliability, consistency, and validity of the results.

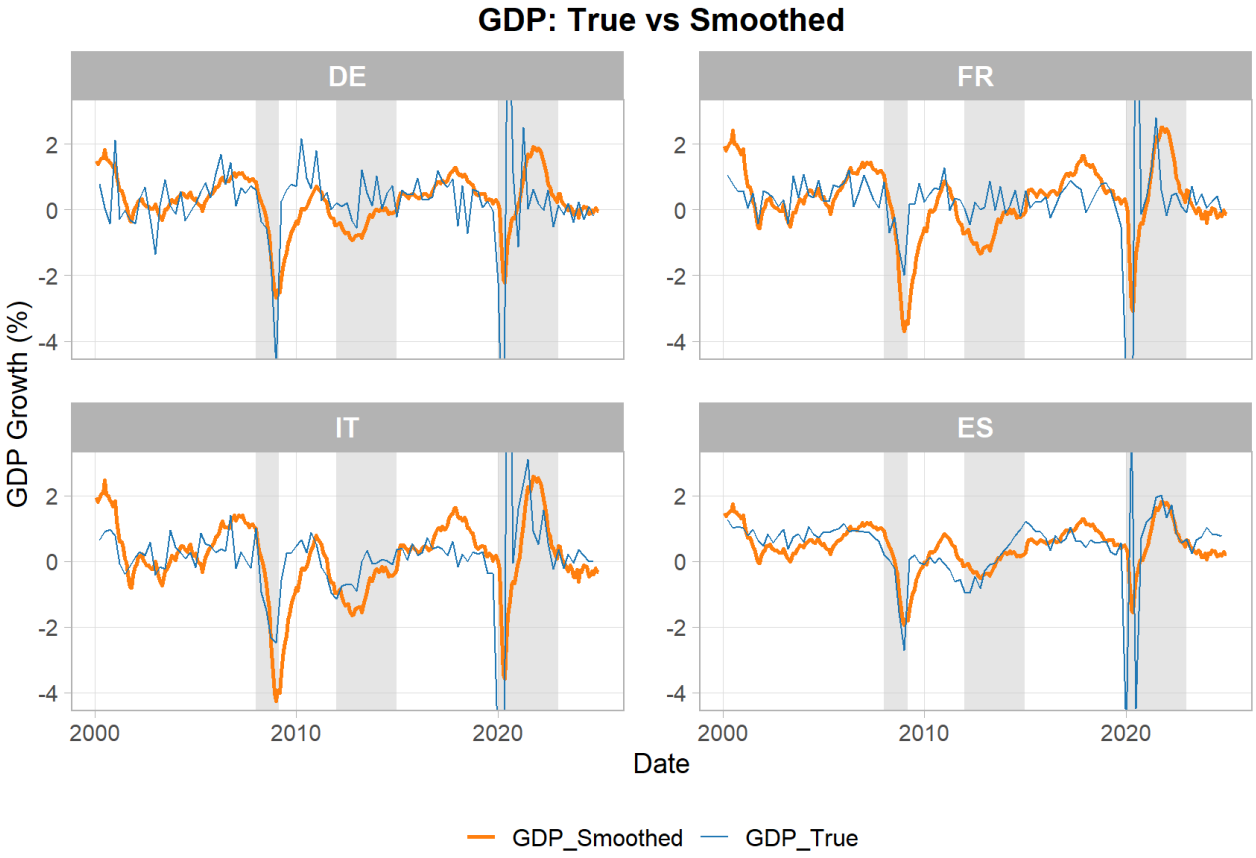


Figure 15: Smoothed estimates of GDP

Moreover, Figure 16 illustrates the convergence of the Expectation-Maximization (EM) algorithm during the final parameter estimation, i.e., for the full dataset up to January 2025. The log-likelihood values stabilize after six iterations, confirming that the algorithm has met the convergence criterion, which sets a tolerance level of $\varepsilon = 10 \times e^{-03}$.

Overall, this empirical analysis provides clear answers to the initial research questions, yielding two main findings.

First, the Dynamic Matrix Factor Model (DMFM) improves nowcasting performance both in terms of accuracy and responsiveness to the progressive release of timely information. Among the all Euro Area countries considered, its matrix-based formulation is particularly effective in the post-COVID period.

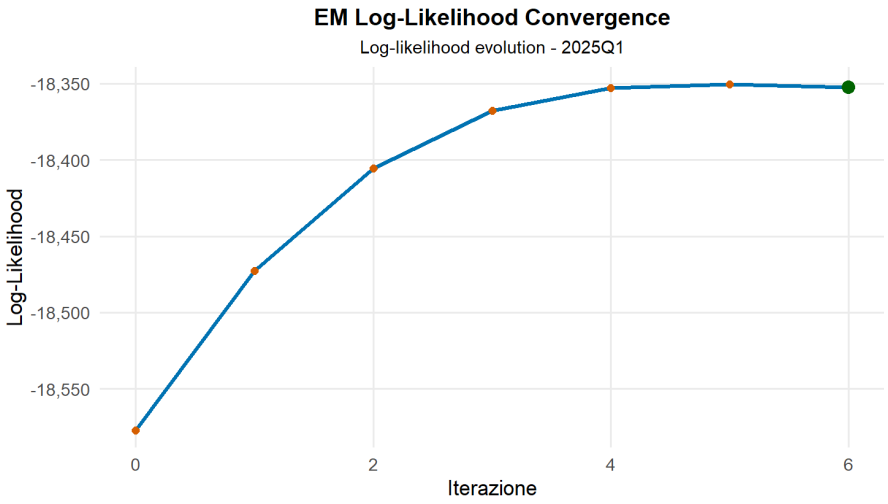


Figure 16: EM algorithm – Quasi log-likelihood convergence

This result likely reflects the increased interdependence among member states, driven by common monetary policies and more synchronized economic cycles, that are the features that the DMFM can better capture thanks to its cross-country structure. Notably, the model provides more timely nowcasts during periods of economic disruption, such as the COVID-19 crisis, and greater accuracy during the subsequent recovery. In contrast, the traditional vector-based DFM often misses key turning points and responds more slowly and smoothly to new data, failing to fully capture country-specific responses to common Euro Area shocks.

Second, regarding the sufficiency of a medium-scale dataset for nowcasting purposes, the results indicate that a set of monthly indicators, augmented by the quarterly GDP series, is adequate for generating reliable forecasts. Both the vector and matrix formulations benefit from this setup, but the DMFM demonstrates clear superiority, particularly by leveraging information from other countries to improve the timeliness and precision of national GDP estimates. While expanding the dataset to include additional quarterly variables may further improve performance, one must also consider the computational trade-offs. Indeed, the DFM in vector form, is generally preferable in terms of computationally efficiency. However, as emphasized by Giannone, Reichlin, and Small 2008 and Bańbura and Modugno 2014, small to medium-sized models are generally preferable over high-dimensional systems for operational nowcasting, thus this 40-variables system could be already a desirable setup.

In summary, the DMFM, applied to a medium-sized dataset comprising 39 monthly indicators and the target GDP series, enables the construction of significantly more accurate and responsive nowcasts for all the Euro Area countries considered, compared to its vector-based counterpart. The improvements are especially pronounced in Germany and Italy during the post-COVID period. For France and Spain, the matrix-based model also proves to be the most effective framework for current-quarter forecasting. These findings support the use of matrix factor models in a Monetary Union context, particularly for countries closely aligned with the aggregate Euro Area cycle and for periods of high intergation. Joint modeling of member states is a valuable strategy not only under normal conditions but, crucially, also during periods of heightened interdependence, such as during recoveries, underscoring the importance of cross-country dynamics in real-time macroeconomic forecasting.

4.3.2 Contributions and Future Works

This work offers several original contributions to the field of macroeconomic forecasting, particularly in the application of nowcasting to dynamic factor models applied to matrix-structured datasets.

First, it introduces a Dynamic Matrix Factor Model (DMFM) for nowcasting GDP. In this way it is possible to preserve the two-dimensional structure of macroeconomic data (variables over time and across countries) while addressing high dimensionality through latent factors.

Second, the estimation of the model using the Expectation-Maximization (EM) algorithm, provides an addition layer of flexibility since allows to handle missing data patterns typical of mixed-frequency settings and periods of economic disruption, such as the COVID-19 pandemic.

Lastly, the study implements a pseudo real-time forecasting exercise, simulating the actual release calendar of economic indicators. Through recursive estimation and nowcasting using the Kalman filter, the framework generates time-consistent forecasts for both the DMFM and a benchmark vector-based Dynamic Factor Model (DFM). To the best of our knowledge, this is among the first empirical applications of matrix-based nowcasting techniques to macroeconomic datasets.

The primary empirical contribution indeed lies in applying this nowcasting approach to a medium-scale dataset comprising 39 monthly indicators and quarterly GDP for key Euro Area economic drivers considered jointly. The comparison between the DMFM and the DFM reveals that the matrix-based model delivers markedly better performance. From the flatter nowcasts of the DFM, the DMFM provides for Spain and France an oustanding enhancment throughout the whole sample, and for Germany and Italy the same is true for the post-COVID period. Resuts achieved duirng the COVID period show that DMFM are not just more accurate in normal times, but are also more responsive to the adverse business cycle during recessions and recoveries. These results suggest that during periods in which countries exhibits a stronger economic dependency from the Monetary Union, nowcasts benefit more from cross-country information, providing evidences in favor of DMFM applications. Conversely, in periods of lower interconnection, such as pre-COVID Germany and Italy, the simpler vector-based DFM remains a valid and computationally efficient alternative.

These findings underscore the potential of matrix-based models to improve forecast accuracy in integrated economic regions, and highlight the importance of accounting for cross-sectional dependencies in real-time macroeconomic forecasting especially for the post-COVID period.

Nonetheless, several further researches could enhance and generalize the proposed approach.

First, the transition equation in Equation 4 could be extended to incorporate higher-order lags. In this thesis, the latent factors are assumed to follow a first-order matrix autoregressive process (MAR(1)). However, the BIC criterion reported in Figure 22 in the Appendix suggests that a two-lag specification may offer a better fit. Adopting a MAR(2) structure would require rewriting the DMFM in companion form.

A second extension could be the inclusion of additional countries and especially variables. The latter would allow testing the hypothesis that small- and medium-scale models often outperform larger models in terms of forecast accuracy, especially when the signal-to-noise ratio deteriorates with dimensionality.

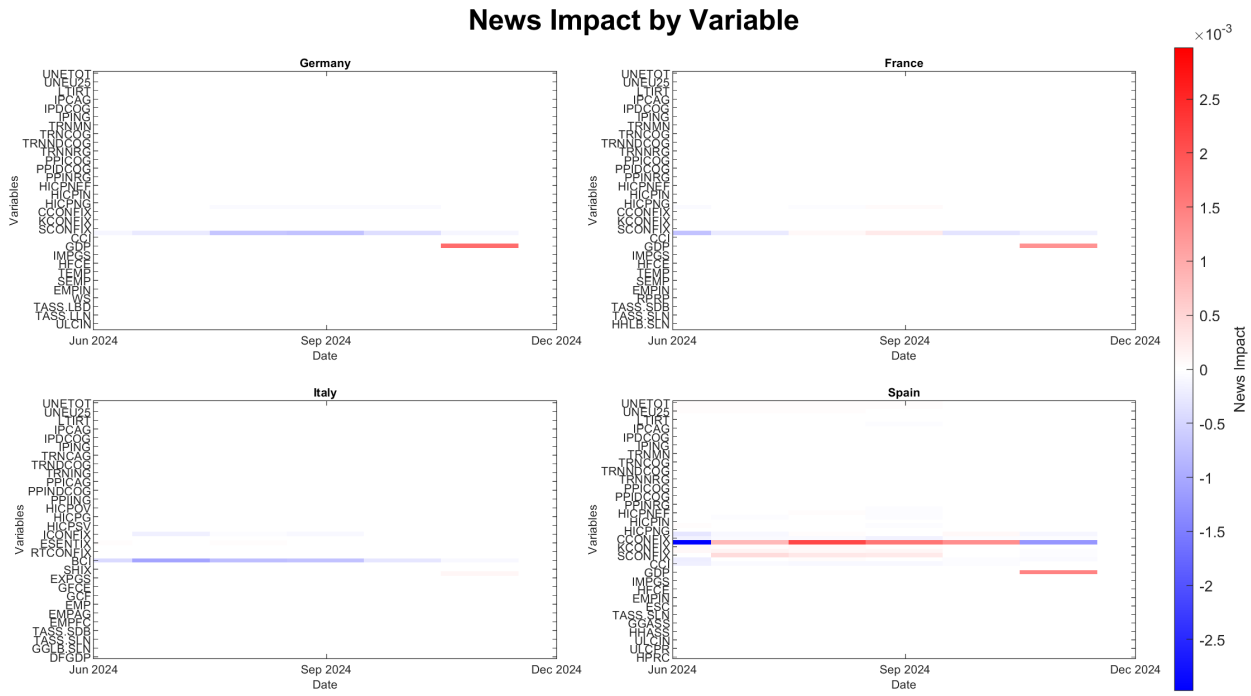


Figure 17: Decomposition of GDP Forecast Revisions by Variable Type

The most promising direction for future research, however, lies in refining the interpretation of forecast revisions. Specifically, decomposing nowcast updates not only by variable or group of variables, but also by country could help identify which member of the matrix is driving changes in the forecast for a particular target country.

To illustrate this idea, consider the forecast revision decomposition following the methodology of Bańbura and Modugno 2014. Between June and December 2024, it is possible, within the vector-form DFM framework, to monthly decompose GDP nowcast updates into contributions associated with the arrival of new data, either by variable or group of variables, for each EA country considered.

Assuming, that revisions are driven solely by the magnitude of the news (i.e., unexpected data releases), and not by re-estimation due to the smoothing step, the results are consistent with previous studies in literature (e.g., Giannone, Reichlin, and Small 2008, Bańbura and Modugno 2014), that underscore the importance of timely survey-based information, particularly confidence data, in shaping nowcasts during the early phases of a quarter.

Indeed, the decomposition by variable for each Euro Area country reveals that soft data, especially confidence indicators, contribute more significantly to nowcast revisions than other indicators, as shown in Figure 17. Their impact is generally associated with downward revisions across most countries, with the notable exception of Spain, where a positive economic trend in the past year has led to upward adjustments.

consistently, when grouping variables, confidence indicators (CIs) consistently emerge as the most influential drivers of forecast updates, particularly during the early part of the quarter. These effects are typically negative, reflecting early sentiment-based expectations. As the quarter progresses, hard

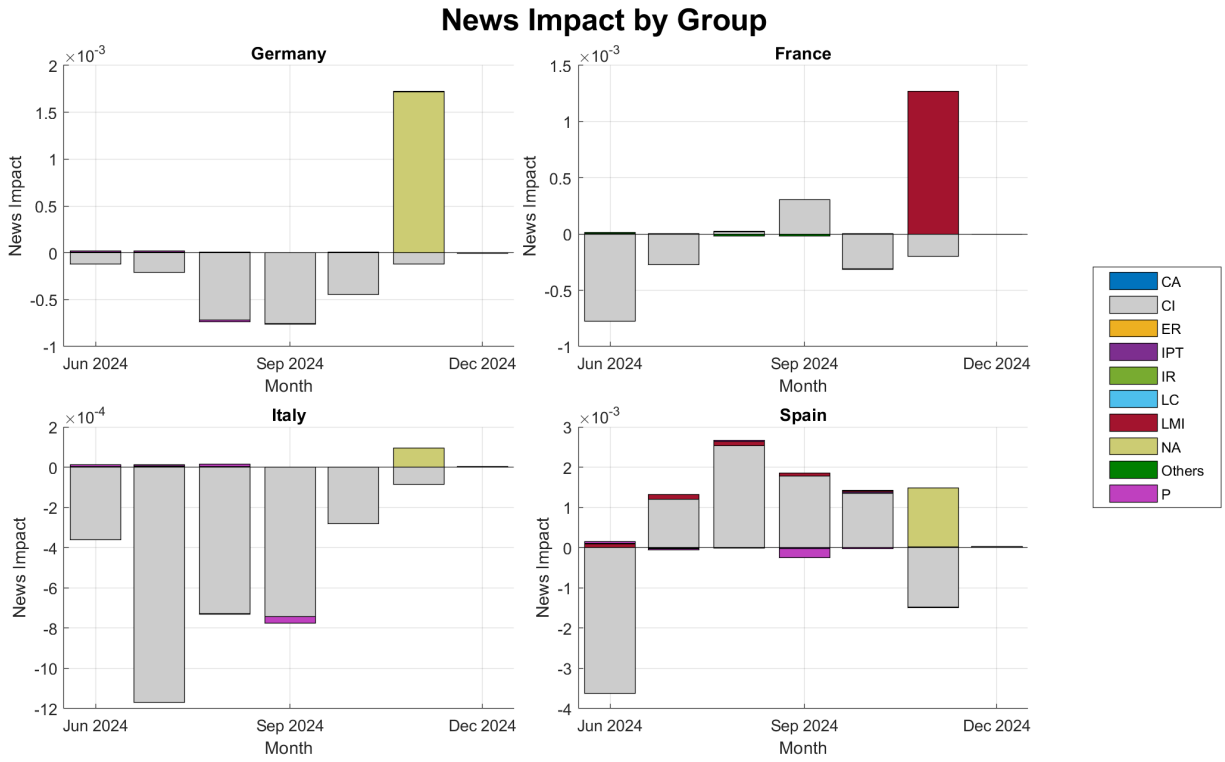


Figure 18: Decomposition of GDP Forecast Revisions by Variable Group

data, such as national accounts and labor market statistics, become increasingly important, as shown in Figure 18. These releases often offset or reverse the initial revisions prompted by soft data, underscoring the sequential dynamics of information flow in real-time forecasting.

Building on these results, a natural extension would be to attribute the unexpected news not only to specific variables or groups, but also to their country of origin. The goal would be to assess which countries contribute most to the revision of another country's GDP nowcast. More precisely, the innovative step would consist in decomposing forecast updates along both the variable and the country dimensions. This would make it possible, for example, to quantify whether Germany's data releases contribute more to improving Italy's forecast than those from France or Spain. Such an analysis would not only capture the relevance of domestic indicators, but also reveal how information expands across countries. This approach would provide a sort of of neural networks structure that could be used to model interdependencies in complex systems, where both the direction and intensity of information flows are learned from the data. Especially during periods of heightened interconnectedness, where the empirical analysis has demonstrated the superiority of the DMFM nowcasting procedure, this additional investigation could help identify which countries are most affected by negative economic cycles originating elsewhere, or which country is driving the recovery. Such insights would support more accurate, effective, and targeted monetary policy interventions especially in a Monetary Union such as the EMU.

5 Conclusion

In economics and finance, data are often collected in matrix form. Recent advances in econometric methodology have increasingly focused on preserving this structure for macroeconomic analysis. Starting from the pioneering work of Wang, Liu, and R. Chen 2019 on Matrix Factor Models (MFMs), subsequent contributions, such as Matteo Barigozzi and Trapin 2025, have introduced a Matrix Autoregressive (MAR) component, giving rise to the Dynamic Matrix Factor Model (DMFM): a state-space formulation particularly well-suited to high-dimensional time series and forecasting. Indeed, the latent factor structure allows the model to capture the main sources of variability in the data, while the dynamics among these factors provide insights into the evolution of underlying economic signals.

This thesis builds on the estimation framework proposed by Matteo Barigozzi and Trapin 2025, in which DMFMs are estimated via Quasi Maximum Likelihood (QML) using the Expectation-Maximization (EM) algorithm. The main contribution of this work lies in adapting this methodology to a nowcasting framework for Euro Area (EA) countries. This work addresses three main challenges: (i) it models matrix-valued data through the DMFM structure; (ii) it enables parameter estimation via a misspecified Gaussian log-likelihood solved iteratively through the EM algorithm; and (iii) it produces real-time nowcasts of quarterly GDP using the Kalman filter embedded in the EM algorithm, jointly exploiting cross-country information for DMFM.

The empirical implementation follows a pseudo real-time nowcasting exercise. Each month, the dataset is updated according to the actual release calendar and frequency of each variable. The model is re-estimated using the EM algorithm, and nowcasts are produced by reconstructing the dataset based on the estimated parameters and the factors extracted through the Kalman filter. To evaluate forecast performance, the Dynamic Matrix Factor Model (DMFM) is compared with a standard vector-based Dynamic Factor Model (DFM), focusing on the four largest Euro Area economies: Germany, France, Italy, and Spain. The analysis relies on a dataset of 39 monthly macroeconomic indicators per country, alongside the respective quarterly GDP series.Results indicate that this set of monthly indicators, combined with GDP, is sufficient to produce reliable nowcasts under both model specifications. However, the key finding is that the matrix-based DMFM consistently yields more accurate forecasts, as it allows the joint exploitation of cross-country information through a common latent factor structure. Spain and France benefit most from the matrix formulation, likely due to their greater dependency with the overall Euro Area economic cycle. Germany and Italy, on the other hand, show marked improvements especially in the post-COVID period. This latter result may reflect the heightened economic interdependence observed after the pandemic, driven by coordinated monetary policies and shared recovery strategies across the Euro Area. Such integration enhances the relevance of cross-country dynamics and explains the superior accuracy and responsiveness of the DMFM during the COVID-19 crisis and the sequent recovery period. In general, while the vector-based DFM remains a viable and efficient choice, especially in periods of economic stability and for countries with more domestically-driven dynamics, the matrix specification proves superior during turning points, such as recessions and recoveries. The main trade-off is computational: the DMFM is more demanding to estimate. Nevertheless, its ability to capture interconnected macroeconomic fluctuations makes it a valuable tool for real-time forecasting in highly integrated economic regions like the Euro Area.

These findings also suggest a promising direction for future research: decomposing nowcast revisions not only by variable or group of variables, as commonly done in vector-based DFMs, but also by country of origin. This would make it possible to identify which countries have the greatest impact on each national nowcast, thereby shedding light on cross-country linkages, as in a neural network framework. Such information is particularly valuable during periods when identifying the main driving country can meaningfully inform monetary policy interventions.

In conclusion, this thesis presents a flexible modeling framework suited to mixed-frequency datasets and capable of accommodating disruptions such as those caused by COVID-19. It constitutes a first step toward the application of nowcasting techniques to tensor-valued data, at least in the matrix case, and offers encouraging evidence supporting the adoption of DMFMs in real-time forecasting applications, particularly in Monetary Unions such as the Euro Area.

References

Bańbura, Marta and Michele Modugno (2014). "Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data". In: *Journal of applied econometrics* 29.1, pp. 133–160.

Barigozzi, M. and C. Lissona (2024). *EA-MD-QD: Large Euro Area and Euro Member Countries Datasets for Macroeconomic Research (Version 12.2023)*. Data set. URL: https://doi.org/10.5281/ZENODO.10514668.

Barigozzi, Matteo, Marco Lippi, and Matteo Luciani (2021). "Large-dimensional dynamic factor models: Estimation of impulse–response functions with I (1) cointegrated factors". In: *Journal of Econometrics* 221.2, pp. 455–482.

Barigozzi, Matteo and Luca Trapin (2025). "Quasi maximum likelihood estimation of high-dimensional approximate dynamic matrix factor models via the EM algorithm". In: *arXiv preprint arXiv:2502.04112*.

Billio, Monica et al. (2023). "Bayesian dynamic tensor regression". In: *Journal of Business & Economic Statistics* 41.2, pp. 429–439.

Cascaldi-Garcia, Danilo et al. (2024). "Back to the present: Learning about the euro area through a now-casting model". In: *International Journal of Forecasting* 40.2, pp. 661–686.

Cen, Zetai and Clifford Lam (2025). "Tensor time series imputation through tensor factor modelling". In: *Journal of Econometrics* 249, p. 105974.

Chen, Elynn Y and Jianqing Fan (2023). "Statistical inference for high-dimensional matrix-variate factor models". In: *Journal of the American Statistical Association* 118.542, pp. 1038–1055.

Chen, Elynn Y, Ruey S Tsay, and Rong Chen (2020). "Constrained factor models for high-dimensional matrix-variate time series". In: *Journal of the American Statistical Association*.

Chen, Rong, Han Xiao, and Dan Yang (2021). "Autoregressive models for matrix-valued time series". In: *Journal of Econometrics* 222.1, pp. 539–560.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1, pp. 1–22.

Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin (2012). "A quasi–maximum likelihood approach for large, approximate dynamic factor models". In: *Review of economics and statistics* 94.4, pp. 1014–1024.

Durbin, James and Siem Jan Koopman (2012). *Time series analysis by state space methods*. Oxford University Press (UK).

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri (2021). "Economic predictions with big data: The illusion of sparsity". In: *Econometrica* 89.5, pp. 2409–2437.

Giannone, Domenico, Lucrezia Reichlin, and David Small (2008). "Nowcasting: The real-time informational content of macroeconomic data". In: *Journal of monetary economics* 55.4, pp. 665–676.

Kapetanios, George, Laura Serlenga, and Yongcheol Shin (2021). "Estimation and inference for multi-dimensional heterogeneous panel datasets with hierarchical multi-factor error structure". In: *Journal of Econometrics* 220.2, pp. 504–531.

Lam, Clifford and Qiwei Yao (2012). "Factor modeling for high-dimensional time series: inference for the number of factors". In: *The Annals of Statistics*, pp. 694–726.

Tipping, Michael E and Christopher M Bishop (1999). "Probabilistic principal component analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3, pp. 611–622.

Tsay, Ruey S (2024). "Matrix-Variate Time Series Analysis: A Brief Review and Some New Developments". In: *International Statistical Review* 92.2, pp. 246–262.

Wang, Dong, Xialu Liu, and Rong Chen (2019). "Factor models for matrix-valued high-dimensional time series". In: *Journal of econometrics* 208.1, pp. 231–248.

Watson, Mark W and Robert F Engle (1983). "Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models". In: *Journal of Econometrics* 23.3, pp. 385–400.

Xiong, Ruoxuan and Markus Pelger (2023). "Large dimensional latent factor modeling with missing observations and applications to causal inference". In: *Journal of Econometrics* 233.1, pp. 271–301.

Xu, Sainan, Chaofeng Yuan, and Jianhua Guo (2025). "Quasi maximum likelihood estimation for large-dimensional matrix factor models". In: *Journal of Business & Economic Statistics* 43.2, pp. 439–453.

Yu, Long et al. (2022). "Projected estimation for large-dimensional matrix factor models". In: *Journal of Econometrics* 229.1, pp. 201–217.

Yu, Ruofan et al. (2024). "Dynamic matrix factor models for high dimensional time series". In: *arXiv preprint arXiv:2407.05624*.

Appendix A: Reference Coding Scheme

Table 6 provides the legend used in Table 4.1.2 to classify each variable by its economic nature (Class), data type (Category), frequency, and transformation method. This coding ensures concise yet informative labeling of variables across the study.

Table 6: Classification codes for variable type, frequency, and transformation

Class	Category	Frequency	Transformation
R = Real	S = Soft Data	Q = Quarterly	0 = No transformation
N = Nominal	H = Hard Data	M = Monthly	1 = $100 \times \Delta\log(x_t)$
F = Financial			2 = Δx_t
C = Confidence			

Appendix B: Supplementary Figures

B.1 Nowcasting Performance by Country

France, Germany, and Italy, like Spain, benefit from the matrix formulation due to its enhanced responsiveness to adverse business cycle conditions and the sequential inflow of monthly data, especially during the COVID-19 recession and the subsequent recovery period in the Euro Area. These advantages, particularly in terms of more timely nowcast reactivity, support the adoption of the matrix formulation of the DFM during periods of heightened interconnectedness.

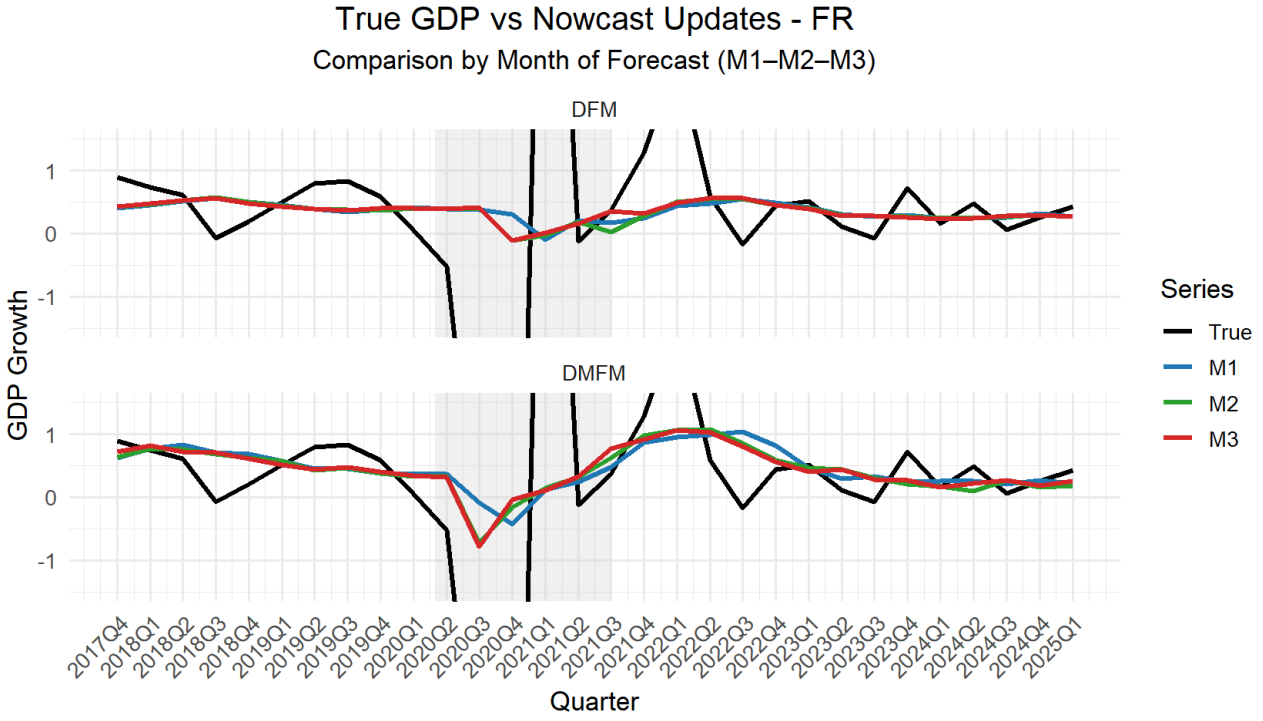


Figure 19: Monthly nowcast updates – France

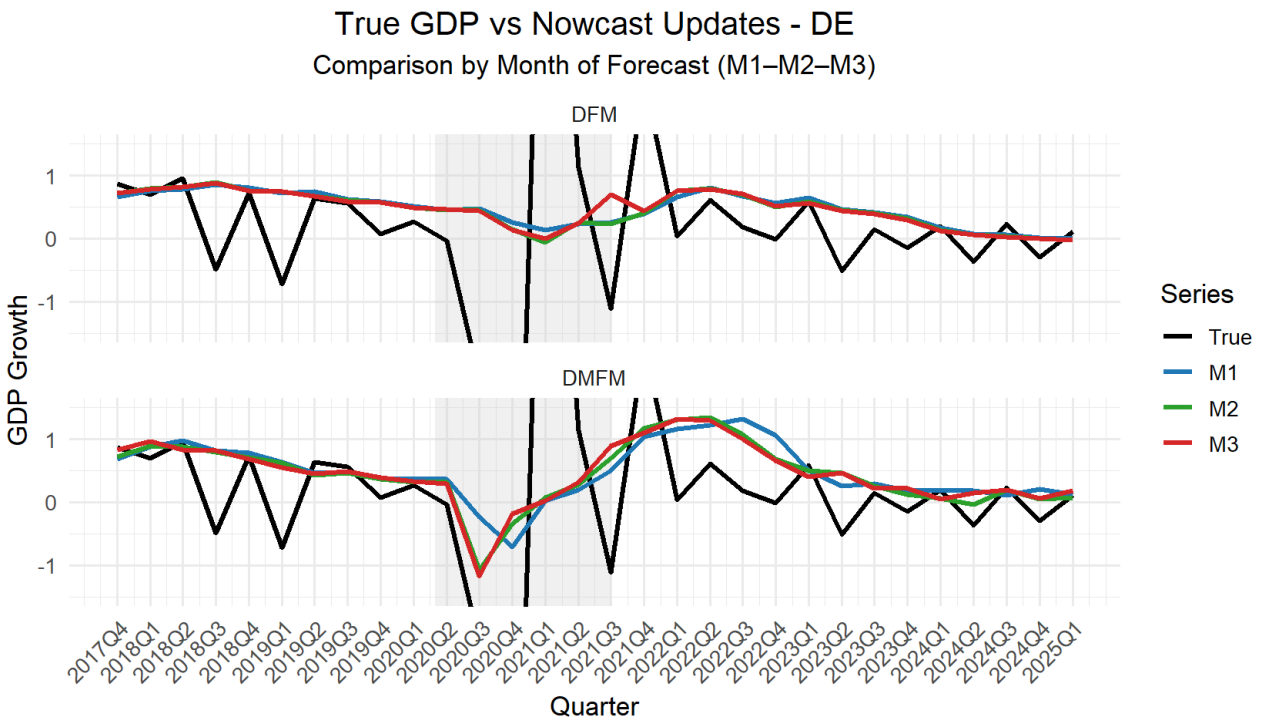


Figure 20: Monthly nowcast updates – Germany

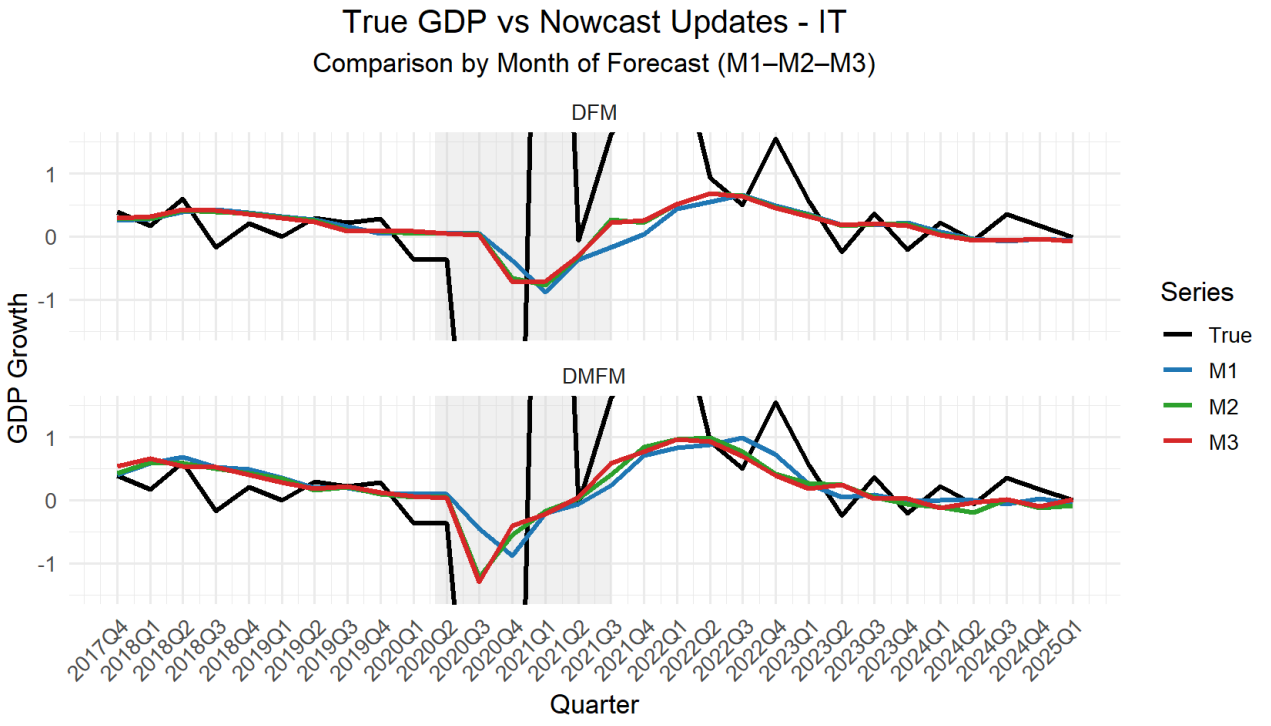


Figure 21: Monthly nowcast updates – Italy

B.2 Model Selection and Future Work

A potential refinement of this work involves extending the lag order of the transition equation in the DMFM (Equation 3). Figure 22 illustrates model selection via the BIC criteria, suggesting a MAR(2) structure. This is possible by extending the model to a companion form.

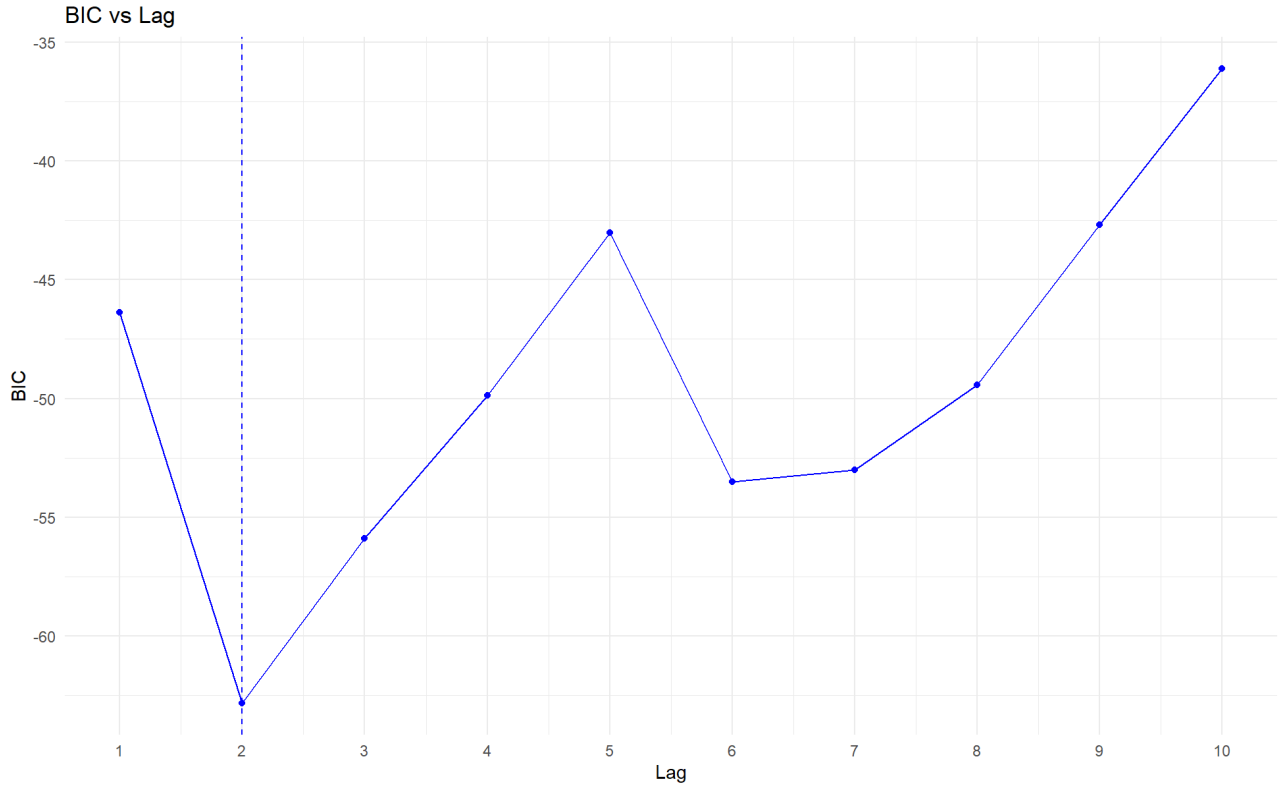


Figure 22: Model selection via BIC and AIC criteria