Nonlinear Dynamic Factor Models

Gianluca Giudice

A thesis presented for the degree of
Doctor of Philosophy

London School of Economics and Political Science
London, United Kingdom
Department of Statistics
June 2022

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.
I declare that my thesis consists of less than 100,000 words.

Statement of co-authored work:
I confirm that Chapters 1 and 2 are co-authored with Prof. Matteo Barigozzi while Chapter 3 is jointly co-authored with Prof. Kostas Kalogeropoulos and Dr. Sara Geneletti.

Acknowledgements

Firstly, I would like to express my gratitude to my initial supervisor, Matteo Barigozzi. It has been a very long journey and I am very thankful for his guidance and encouragement since day one. His valuable insights and rigour gave me the platform to achieve this result. A no less important thanks to my current supervisor, Kostas Kalogeropoulos, for taking me under his wing on this already initiated Ph.D. odyssey. It has been an honour to work with such a talented and versatile researcher and I am grateful for the invaluable support, even during the inconceivable hours of the night. Furthermore, I want to express my recognition to Sara Geneletti for all the critical advice, but, more importantly, her infectious laugh that makes you forget a major proof gone wrong.

A huge shout-out to the Department of Statistics for providing a highly stimulating environment and a special thanks to Penny for the constant help throughout the years. Thanks to my friends and colleagues: Davide and Filippo, be it a ridge regression or a drink, Italians really do it better; and Anica, the most skilled statistician of the modern age, I could not have asked for a better desk mate.

Of course, a huge 'grazie' to my parents, Giuseppe and Claudia, for allowing me to take part in this quest in the first place. Dad for teaching me how to avoid dull mistakes and mom for teaching me how to make them.

Finally to the people that shared these years with me, by my side. To Valerio, for his unconditional support and motivation; Elvira for her non-stop talking and for new 'fatti' every day; Frensis for always having my back; Chiara for the constant source of laughter mixed with tears.

Thanks.

Abstract

This thesis discusses latent variable models with the aim of uncovering hidden structure in multi-dimensional data. In rich data settings, dimension reduction has made latent variable methods, such as Dynamic Factor Models, extremely popular. Nonetheless, the dynamics of the factors have in most cases been modelled as independent and identically distributed (*i.i.d.*) white noise even though many financial and economic variables exhibit conditional heteroskedasticity, i.e., their variances conditional on the past evolve with time. This feature, modelled in the literature by GARCH models, has been applied to factor analysis either in the finite n case or by use of two-step estimators, producing inefficient results. In the first chapter, we show that when $n \to \infty$, estimators for the latent factors and their conditional variance are, indeed, consistent. First, we convert the model in state-space form explicitly taking into account heteroskedasticity. Subsequently, we apply the Kalman filter to jointly estimate the parameter via the Expectation Conditional Maximization Either algorithm (ECME). This version of the EM replaces some of the steps which conditionally maximize the expected complete-data log-likelihood, with steps that maximize the real prediction-error likelihood, thus dealing with the lack of closed-form solution for the GARCH parameters. We then propose further modifications to the original model, introducing potential Dynamic Conditional Correlation (DCC) dynamics in the factors and a time-varying volatility for the observation equation disturbances. These extensions are subsequently assessed empirically, in the context of portfolio allocation and the economically relevant Growth at Risk (GaR). Finally, when the data dimension is limited, a Multi-Output Gaussian Process with Semiparametric Latent Factor structure can provide an extremely valuable opportunity to explore unobserved states in a multivariate setting. These non-linear models offer a novel and efficient approach to estimate the causal effect of interventions in time. As such, we analyse whether the early and intense vaccination campaign introduced in the UK affected the number of deaths and level of contagiousness of Covid-19 in the first semester of 2021.

Contents

List of Figures

List of Tables

Chapter 1

Conditionally Heteroskedastic Dynamic Factor Models

1.1 Introduction

High-dimensional data are undeniably one of the most significant challenges in current statistics, and they have grown in prevalence in almost all disciplines linked to data sciences. The study of high-dimensional time series, huge cross-sections of univariate time series or panels, has not avoided this trend, and it is now one of the most active subjects in theoretical and applied econometrics. The so-called factor models are the foundation of one of the most successful frameworks in the analysis and prediction of high-dimensional time series thus far. This framework, in its various structures, is based on the sum of two mutually orthogonal components: the *common component*, driven by a small number of factors or common shocks, and an *idiosyncratic component*, which is specific to each series. In particular, the General or Generalized Dynamic Factor Model (GDFM) proposed by Forni et al. (2000) encompasses most other models, such as the static factor approaches proposed by Bai (2003), Stock and Watson (2002), and Fan et al. (2013), by taking into account all leading and lagging linear dependencies among the data. Furthermore, as pointed out by Forni and Lippi (2001) and Hallin and Lippi (2013), the GDFM decomposition into a common and idiosyncratic component requires the usual second-order stationarity and the existence of spectral densities with no further structural restriction on the data generating process.

Prediction is an apparent and natural goal in traditional analysis of univariate and multivariate time series; it is no less important in high-dimensional data. Due to the numerous and complex cross-dependencies among the many cross-sectional components, an efficient forecast should exploit the amount of information available in the present and lagged values of the entire cross-section; the larger the cross-section (i.e., the higher the dimension n), the more crucial the role of that information, and the more delicate its recovery. Stock and Watson (2002), Bai and Ng (2008), and Forni et al. (2018), to name just a few, have all employed factor models in the design of point-predictors, and have done so effectively.

However, those writers are largely working with macroeconomic data, while factor model approaches in the study and prediction of financial returns have received less attention: see, for example, Chamberlain and Rothschild (1983), Connor and Korajczyk (1993), or Ait-Sahalia and Xiu (2017). Conditional volatility phenomena are particularly important when dealing with returns, due to the presence of conditional distribution heterogeneity (of which conditional heteroskedasticity is only one example) and should be considered when constructing conditional prediction limits or conditional prediction intervals. Most multivariate approaches for the analysis of conditional heterogeneity in the literature are limited to the study of conditional heteroskedasticity and rely on Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) or Stochastic Volatility type parameterizations: see, for example, the reviews by Laurent et al. (2006) for the former and Asai et al. (2006) or Aguilar (2009) for the latter. However, because of the *curse of dimensionality*, only the most basic models may be evaluated in high-dimensional panels, potentially resulting in a significant loss of efficiency. Engle (1987) was the first to propose the concept of a conditional heteroskedastic factor model, in which the observed series' conditional covariance follows a one-factor process. The factor GARCH technique is the most often used of these; see, for example, Diebold and Nerlove (1989), Ng et al. (1992), Harvey et al. (1992), and Sentana et al. (2008). The former authors, for example, develop an heteroskedastic one-factor model for exchange rate series. In these papers, the Kalman filter is used for estimation rather than Engle et al. (1990)'s two-step technique, in which static elements are extracted from the unconditional covariance matrix before being treated as univariate GARCH processes. Static factor models based only on volatilities have also been examined by Fan et al. (2015) and Connor et al. (2006), but this approach fails to take advantage of the information contained in the idiosyncratic components of returns. For these reasons, Barigozzi and Hallin (2017a) propose a two-step GDFM approach in which factor models' nonparametric and model-free properties are combined in a joint study of returns and volatilities. That two-step GDFM is combined with a GARCH strategy to produce point-forecasts for volatilities. Then, Barigozzi and Hallin (2017b) and Barigozzi et al. (2019) study the dynamic interdependencies of the US and international financial markets and Chicheportiche and Bouchaud (2015) present a comparable two-stage factor technique, but in a static factor model context. Finally, Trucíos et al. (2021) use a similar approach to construct the minimum variance portfolio for a high-dimensional panel of assets on the basis of the one-step-ahead variance forecast. Most of the above literature, however, employs a two-step estimation approach, which presents a strong limitation as it is not clear how the inefficiency generated by the two-step procedure is incorporated into the volatility panel.

Separately, there is strand of literature that studies conditionally heteroschedastic (non-dynamic) factor models, but in the context of indirect or simulation-based estimation. In order to bypass the inconsistencies associated with the Kalman filter approximations to the log-likelihood function provided by Diebold and Nerlove (1989) and Harvey et al. (1992), Fiorentini et al. (2004) develop computationally efficient Markov chain Monte Carlo (MCMC) simulation methods that provide more precise likelihood-based estimators of factor models with GARCH structures. Subsequently, Sentana et al. (2008) derive an indirect

estimators to deal with the lack of a closed form solution to the log-likelihood function, without resorting to simulation. Even so, the procedure requires some approximation to accommodate large-scale multivariate processes thus the authors resort to sequential estimators. Although both works mention consistency proprieties of the latent factor as n increases, this framework is never properly adopted and no formal proof is presented, especially in regard to the errors and, consequently, conditional variances.
In this chapter we propose a Conditionally Heteroskedastic Dynamic Factor Model (CHDFM) to explain and forecast the conditional covariances of a large number of series with a minimal number of parameters. Each series is assumed to have a common portion that carries a dynamic structure and an idiosyncratic part, independent of time. The common component volatilities evolve according to the GARCH rule so that we can describe the development of the conditional covariances of observable series by simply modelling the evolution of the conditional covariances of a few components. This work extends the results of Harvey et al. (1992) for the ARCH(1) to the more general AR(1)-GARCH(1,1) in the context of asymptotic n. In the paper, the authors developed a modified Kalman filter for models with unobservable heteroskedastic components, a structural ARCH, and applied it to a dynamic factor model. The derivation of the Kalman filter estimator is performed in the finite n case, and the authors argue against the good capacity of the estimator to correctly extract the unobserved state. In this first chapter we prove that, in the framework of the Generalized Dynamic Factor Model, as n grows, the factors, including the conditional variances, can be consistently extracted and parameters efficiently estimated by the use of a variation of the Expectation Maximization (EM) algorithm of Dempster et al. (1977). As no closed form solution exists for GARCH parameters, we employ the Expectation Conditional Maximization Either (ECME) algorithm (Liu and Rubin, 1994), which is obtained by replacing some (all or none) CM-steps of the ECM, with steps that maximize the correspondingly constrained actual likelihood function, directly available from the Kalman filter. Upon investigating the consistency features of our estimation procedure for the cross-sectional (n) and sample (T) dimensions going to infinity, we evaluate the goodness of our estimation approach and the strengths of our model using Monte Carlo simulations and two empirical applications.

The main contribution of this chapter is to set the theoretical background of the CHDFM. We address the limitations of the unobserved component model with ARCH disturbances of Harvey et al. (1992) and show that in the context of asymptotic n, both unobserved factors and model parameters can be estimated consistently. Furthermore, we generalize the approach to the more exhaustive and flexible AR-GARCH model. Finally, we propose and derive an efficient estimation procedure based on the ECME algorithm.

The chapter is structured as follows. Section 1.2 describes the general structure of the model and the main assumptions. Section 1.3 deals with estimation, introducing the Kalman filter to extract the unobserved factors and the Expectation Conditional Maximization Either algorithm to determine the optimal parameters. Section 1.4 extends the theoretical framework by discussing topics such as identification and prediction distribu-

tion. Section 1.5 illustrates some specific cases and examines the numerical aspects in more detail. Section 1.6 presents a simulation exercise in which estimation results are discussed. The final section contains the Appendix, which encompasses the main proofs of the statements discussed in this chapter.

1.2 The Model

1.2.1 A Conditionally Heteroskedastic DFM framework

For an n-dimensional panel of time series of size T: $\{x_{i,t} : i = 1, ..., n, t = 1, ..., T\}$ consider the approximate dynamic factor model given by

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}\mathbf{G}_t + \boldsymbol{\xi}_t, \tag{1.1}$$
$$\mathbf{G}_t = \boldsymbol{\Phi}\mathbf{G}_{t-1} + \boldsymbol{\eta}_t \tag{1.2}$$

where $\boldsymbol{x}_t = (x_{1,t} \cdots x_{n,t})'$, $\boldsymbol{\xi}_t = (\xi_{1,t} \cdots \xi_{n,t})'$ are n-dimensional vectors, $\mathbf{G}_t = (G_{1,t} \cdots G_{r,t})'$ is the $r \times 1$ zero-mean unobserved state vector, with r finite and $r \ll n$, and $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1 \cdots \boldsymbol{\lambda}_n)'$ is the matrix of factor loadings with dimension $n \times r$. We have that $\mathbb{E}[\mathbf{G}_t] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{x}_t] = \mathbf{0}$ and unconditional covariance for $\boldsymbol{x}_t$ and $\mathbf{G}_t$ are denoted as $\mathrm{Var}[\boldsymbol{x}_t] = \boldsymbol{\Sigma}$ and $\mathrm{Var}[\mathbf{G}_t] = \boldsymbol{\Omega}$. The $r \times r$ matrix $\boldsymbol{\Phi}$ describes the dynamic relationship between the factors and it is time independent.[1] The observation equation disturbance vector $\boldsymbol{\xi}_t$ is homoskedastic and normally distributed with mean zero, that is $\boldsymbol{\xi}_t \sim \mathrm{NID}(\mathbf{0}, \boldsymbol{\Gamma})$, with $\boldsymbol{\Gamma}$ positive definite.[2] Furthermore, we assume that $\boldsymbol{\xi}_t$ is uncorrelated at all lags and leads with $\mathbf{G}_t$. The conditional heteroskedasticity is introduced through the elements of $\boldsymbol{\eta}_t$, the factor disturbances $\eta_{i,t}$ with $i = 1, ..., r$, which evolves according to:

$$\boldsymbol{\eta}_t = \mathbf{Q}_t^{1/2}\widetilde{\boldsymbol{\eta}}_t \tag{1.3}$$
$$q_{i,t} = \omega_i + \alpha_i \eta_{i,t-1}^2 + \beta_i q_{i,t-1} \tag{1.4}$$

with $\widetilde{\boldsymbol{\eta}}_t \sim \mathrm{NID}(\mathbf{0}, \mathbf{I}_r)$ being an r-dimensional vector and $\mathbf{Q}_t$ being the $r \times r$ diagonal conditional covariance matrix with diagonal entries $q_{i,t}$.
Denote by $\mathcal{I}_{t-1}$ the σ-field generated by $\boldsymbol{x}_t$ and $\mathbf{G}_t$ up to, and including, time $t-1$. It is of crucial importance to distinguish between $\mathcal{I}_{t-1}$ and the econometrician's information set $\mathcal{X}_{t-1} = \{\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, ...\}$, defined such that $\mathcal{I}_{t-1} = \mathcal{X}_{t-1} \cup \mathcal{F}_{t-1}$, given $\mathcal{F}_{t-1} = \{\mathbf{G}_{t-1}, \mathbf{G}_{t-2}, ...\}$. Although we work with both $\boldsymbol{x}_t$ and $\mathbf{G}_\mathbf{t}$, let us remark that the only process we observe is $\boldsymbol{x}_t$. As it will be explained later, working with a smaller subset complicates the analysis as we define the variance process as a latent process itself.
From the above equations, it is straightforward to see that $\mathbf{G}_t$ has conditional moments $\mathbb{E}[\mathbf{G}_t|\mathcal{I}_{t-1}] = \boldsymbol{\Phi}\mathbf{G}_{t-1}$ and $\mathrm{Var}[\mathbf{G}_t|\mathcal{I}_{t-1}] = \mathbf{Q}_t$, while for $\boldsymbol{x}_t$ it holds that $\mathbb{E}[\boldsymbol{x}_t|\mathcal{I}_{t-1}] = \boldsymbol{\Lambda}\mathbf{G}_t$

[1] Although a generalization of it depending on t or the information available at time $t-1$ is possible (Harvey et al., 1992).

[2] Generalization to $\boldsymbol{\xi}_t$ being heteroskedastic is also possible making $\mathrm{Var}(\boldsymbol{\xi}_t|\mathcal{X}_{t-1})$ diagonal and time-dependent. We will show how the model can be modified accordingly in the following chapter.

and $\text{Var}[\boldsymbol{x}_t|\mathcal{I}_{t-1}] = \boldsymbol{\Sigma}_t$, where $\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda}\mathbf{Q}_t\boldsymbol{\Lambda}' + \boldsymbol{\Gamma}$.

Let $\|\mathbf{A}\|_F$ be the Frobineus norm of a $p \times p$ matrix $\mathbf{A}$ such that $\|\mathbf{A}\|_F = \text{tr}(\mathbf{A}'\mathbf{A})^{1/2}$ with $\text{tr}(\mathbf{A})$ being the trace of $\mathbf{A}$. Also, $\|\mathbf{A}\|_1 = \max_{i=1,...,p} \sum_{j=1}^{p} |[\mathbf{A}]_{ij}|$ and $\|\mathbf{A}\| = \lambda_{max}(\mathbf{A}\mathbf{A}')^{1/2} = \lambda_{max}(\mathbf{A})$ if $\mathbf{A}$ is symmetric. They represent the maximum absolute column sum and largest eigenvalue of $\mathbf{A}$, respectively. Denote n and T as the dataset dimension and sample size, respectively. We focus on asymptotic analysis where both $n, T \to \infty$.
The following assumptions about the model are made:

Assumption 1 (Dynamics) $\{\mathbf{G}_t\}$ *is a stationary process with* $\mathbb{E}[\mathbf{G}_t] = \mathbf{0}$ *and* $\text{Var}[\mathbf{G}_t] < \infty$. *More specifically:* $\det(\mathbf{I}_r - \boldsymbol{\Phi}z) \neq 0$ *for all* $z \in \mathbb{C}$ *such that* $|z| \leq 1$. *Furthermore each element of* $\boldsymbol{\eta}_t = \mathbf{Q}_t^{1/2}\widetilde{\boldsymbol{\eta}}_t$ *follows a GARCH(1,1) dynamic with* $\widetilde{\boldsymbol{\eta}}_t \sim NID(\mathbf{0}, \mathbf{I}_r)$ *and* $\mathbf{Q}_t$ *diagonal with entries* $q_{i,t}$. $\omega_i, \alpha_i, \beta_i > 0$ *and* $\alpha_i + \beta_i < 1$, *for all* $i = 1, ..., r$.

Assumption 2 (Errors independence) *The processes* $\{\eta_{j,t}, j = 1, \cdots, r, t \in \mathbb{Z}\}$ *and* $\{\xi_{i,t}, i \in \mathbb{N}, t \in \mathbb{Z}\}$ *are mutually independent.*

Assumption 3 (Measurement errors covariance) *For all* $n \in \mathbb{N}$ *and all* $t \in \mathbb{Z}$, $\mathbb{E}[\boldsymbol{\xi}_t] = \mathbf{0}$ *and* $\boldsymbol{\Gamma} = \mathbb{E}[\boldsymbol{\xi}_t\boldsymbol{\xi}_t']$ *is positive definite and for all* $i \in \mathbb{N}$, $C_\xi^{-1} \leq [\boldsymbol{\Gamma}]_{ii} \leq C_\xi$ *for some finite positive real* C_ξ *independent of* i. *Furthermore, for all* $i, j \in \mathbb{N}$, *all* $t \in \mathbb{Z}$, *and all* $h \in \mathbb{Z}$, $|\mathbb{E}[\xi_{i,t}\xi_{j,t-h}]| \leq \rho^{|h|}M_{i,j}$ *where* ρ *and* $M_{i,j}$ *are finite positive reals, independents of* t *and such that* $0 \leq \rho < 1$, $\sum_{j=1}^{n} M_{i,j} \leq M_\xi$, *and* $\sum_{i=1}^{n} M_{i,j} \leq M_\xi$ *for some real* $M_\xi < \infty$ *and independent of* n.

Assumption 4 (Factors covariance) $\mathbb{E}[\mathbf{G}_t\mathbf{G}_t'] = \boldsymbol{\Omega}$, *with* $\boldsymbol{\Omega}$ *positive definite with distinct eigenvalues.*

Assumption 5 (Loadings) *There exists an integer* n_0 *such that for all* $n > n_0$, $\|n^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Lambda} - \boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}\| = 0$, *where* $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}$ *is positive definite with distinct eigenvalues. Furthermore, for all* $n \in \mathbb{N}$, $m_\lambda < \max_{i=1,...,n}\|\boldsymbol{\lambda}_i\| < M_\lambda$, *for some finite positive reals* m_λ *and* M_λ *and independent of* n.

Assumption (A1) expresses the conditions such that the factor process is weakly stationary and guarantees that the matrix $\mathbf{Q}_t$ is positive definite for each t.
Assumption (A2) specifies no correlation between the factors and the idiosyncratic component while Assumption (A3) states the conditions under which the zero-mean idiosyncratic error $\xi_{i,t}$ can be weakly cross-sectionally correlated. In this last case, factor structure is said to be *approximate* in comparison to the *strict* case where the idiosyncratic terms are uncorrelated. Nevertheless, to address serial dependence issues, it is possible to treat each serially correlated idiosyncratic component as a latent state. Given the set $\mathcal{C} = \{i \in \mathbb{N} : \text{Cov}[\xi_{i,t}, \xi_{i,t-k}] \neq 0\}$ we can add an additional state to (1.2) such that $\xi_{i,t} = \rho_i\mathbb{I}_{i\in\mathcal{C}}\xi_{i,t-1} + e_{i,t}$ for $i = 1, ..., n$ as long as $e_{it} \sim NID(0, \sigma_e^2)$.
Assumptions (A4) and (A5) are standard in factor analysis and they will be revised in the

next sections to identify the model. Indeed, given this specification, there exists an observationally equivalent model such that $\mathbf{\Lambda G}_t = \mathbf{\Lambda U U}^{-1}\mathbf{G}_t$ for any $r \times r$ invertible matrix $\mathbf{U}$. This leaves an r^2 restriction to be imposed to uniquely identify the individual columns of $\mathbf{G}_t$ and $\mathbf{\Lambda}$. Restrictions are imposed depending on the factor dependencies and they will be modified in the next chapter to take into account potential correlation among factors. Furthermore, (A5) specifies Chamberlain and Rothschild (1983)'s conditions for common component pervasiveness, that is $\liminf_{n\to\infty} n^{-1}\lambda_{\min}(\mathbf{\Lambda}'\mathbf{\Lambda}) > 0$, and that all the eigenvalues of $\mathbf{\Lambda}'\mathbf{\Lambda}$ diverge at the same rate, i.e. $\limsup_{n\to\infty} n^{-1}\lambda_{\max}(\mathbf{\Lambda}'\mathbf{\Lambda})$ is *finite*.
Nonetheless, the conditionally heteroskedastic structure of (1.3) requires further restriction to be addressed. In particular, the diagonality of $\mathbf{Q}_t$ is of major relevance in the correct parameter estimation and consistent specification of the covariance matrix. Thus, we will explore identification conditions in more details in the next sections.

1.2.2 The Augmented Model

The estimation of the model relies on the augmented state space form of (1.1) - (1.2), in which we introduce a misspecification error $\boldsymbol{\eta}^*$. Following Diebold and Nerlove (1989), the disturbance $\boldsymbol{\eta}_t$ is treated as both a state variable and error so that the measurement and transition equations become:

$$\boldsymbol{x}_t = \underbrace{\begin{bmatrix}\mathbf{\Lambda} & \mathbf{0}_{n\times r}\end{bmatrix}}_{\mathbf{\Lambda}^\dagger}\mathbf{F}_t^\dagger + \boldsymbol{\xi}_t, \tag{1.5}$$

$$\mathbf{F}_t^\dagger = \begin{bmatrix}\mathbf{F}_t \\ \boldsymbol{\eta}_t\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{\Phi} & \mathbf{0}_r \\ \mathbf{0}_r & \mathbf{0}_r\end{bmatrix}}_{\mathbf{\Phi}^\dagger}\underbrace{\begin{bmatrix}\mathbf{F}_{t-1} \\ \boldsymbol{\eta}_{t-1}\end{bmatrix}}_{\mathbf{F}_{t-1}^\dagger} + \underbrace{\begin{bmatrix}\mathbf{I}_r & \mathbf{I}_r \\ \mathbf{0}_r & \mathbf{I}_r\end{bmatrix}}_{\mathbf{\Psi}^\dagger}\underbrace{\begin{bmatrix}\boldsymbol{\eta}_t^* \\ \boldsymbol{\eta}_t\end{bmatrix}}_{\boldsymbol{\eta}_{t-1}^\dagger}. \tag{1.6}$$

Matrices $\mathbf{\Phi}^\dagger$ and $\mathbf{\Psi}^\dagger$ are now both of dimension $2r \times 2r$ and $\mathbf{\Lambda}^\dagger$ is $n \times 2r$. $\mathbf{F}_t^\dagger$ is the $2r \times 1$ augmented unobserved state vector and $\boldsymbol{\eta}_t^\dagger$ is the $2r \times 1$ disturbance component consisting of $\boldsymbol{\eta}_t^*$ and $\boldsymbol{\eta}_t$, whose first two conditional moments are given by

$$\boldsymbol{\eta}_t^\dagger|\mathcal{I}_{t-1} \sim \mathcal{N}\left[\begin{pmatrix}\mathbf{0}_{r\times 1} \\ \mathbf{0}_{r\times 1}\end{pmatrix}, \begin{pmatrix}\mathbf{Q}^* & \mathbf{0}_r \\ \mathbf{0}_r & \mathbf{Q}_t\end{pmatrix}\right],$$

where $\mathcal{N}$ indicates the Normal probability density function. In this way, the dynamic of the conditional variance is given by the lower-right $r \times r$ block matrix $\mathbf{Q}_t$. On the other hand, the *iid* disturbance vector $\boldsymbol{\eta}_t^*$ is constrained to have a homoskedastic covariance matrix $\mathbf{Q}^* = \epsilon\mathbf{I}_r$ with $\epsilon \to 0$ on the diagonal and 0 elsewhere, and it is independent of $\boldsymbol{\eta}_t$ and $\boldsymbol{\xi}_t$. This means that $\mathbf{Q}^* = O(\epsilon)$. If $\boldsymbol{\eta}_t^* = \mathbf{0}$, then the model reverts back to the original case of (1.1) - (1.2). More precisely, we have that $\mathbf{F}_t = \mathbf{G}_t + \sum_{i=1}^{\infty}\mathbf{\Phi}^i\boldsymbol{\eta}_{t-i}^*$. When calculating the variance we have $\mathrm{Var}[\mathbf{F}_t] = \mathrm{Var}[\mathbf{G}_t] + \sum_{i=1}^{\infty}\mathbf{\Phi}^i\mathbf{Q}^*\mathbf{\Phi}^{i'}$. Let us call the last term $\mathbf{Q}^\star$, then this is a Lyapunov equation with $\|\mathbf{\Phi}\| \leq 1$ and whose solution is given by $\mathrm{vec}(\mathbf{Q}^\star) = (\mathbf{I}_{r^2} - \mathbf{\Phi}\otimes\mathbf{\Phi})^{-1}\mathrm{vec}(\mathbf{Q}^*)$. Thus, $\mathbf{Q}^\star$ is a rescaled version of $\mathbf{Q}^*$ which is defined by the user and $O(\epsilon)$ so the two are equivalent. For this purpose we will use the relation

$\text{Var}[\mathbf{F}_t] = \text{Var}[\mathbf{G}_t] + \mathbf{Q}^\star = \text{Var}[\mathbf{G}_t] + O(\epsilon)$ throughout.
An important remark must be made about the modification we introduced with $\boldsymbol{\eta}_t^*$. In practice, when the variance of the state equation, i.e. $\mathbf{Q}_t^\dagger$ has a zero row (or column), the algorithm, defined as in Shumway and Stoffer (2006) and presented in the Appendix, still works out and provides the appropriate state output and maximized likelihood.[3] The issue, however, is purely theoretical as the likelihood form is not defined because it involves the inversion of $\mathbf{Q}_t^\dagger$. Furthermore, if $\boldsymbol{\eta}_t^* = \mathbf{0}$ the unconditional variance of $\mathbf{F}_t^\dagger$ is not invertible as well, given the matrix extension of Cauchy-Schwarz inequality. To invert the block matrix $\text{Var}[\mathbf{F}_t^\dagger]$, we need the Schur component $\text{Var}[\mathbf{F}_t] - Cov[\mathbf{F}_t, \boldsymbol{\eta}_t]\text{Var}[\boldsymbol{\eta}_t]^{-1}Cov[\mathbf{F}_t, \boldsymbol{\eta}_t]'$ to be invertible as well. But when $\boldsymbol{\eta}_t^* = \mathbf{0}$, i.e. $\mathbf{F}_t = \mathbf{G}_t$ by Cauchy-Schwarz inequality, we have that $\text{Var}[\mathbf{F}_t] = Cov[\mathbf{F}_t, \boldsymbol{\eta}_t]\text{Var}[\boldsymbol{\eta}_t]^{-1}Cov[\mathbf{F}_t, \boldsymbol{\eta}_t]'$. As a result, a constrained variable $\boldsymbol{\eta}_t^* \neq \mathbf{0}$ is introduced in the model. It can be shown, however, that this misspecification is negligible and mainly dependent on the selected value of $\text{Var}[\eta_{i,t}^*] = \epsilon$, which is set to a very small value, such as $\epsilon = 10^{-8}$.
From here on we will employ the Augmented Model as the main reference. All the assumptions listed in Section 1.2 hold replacing $\mathbf{G}_t$ with $\mathbf{F}_t$.

1.3 Estimation

1.3.1 The Kalman Filter and the Kalman Smoother

Let's consider first the case of known parameters. Here, we apply the same framework of Harvey et al. (1992) but in the context of $n \to \infty$, where we exploit the *blessing of dimensionality*. The Dynamic Factor Model as indicated in (1.5) - (1.6) is only conditionally Gaussian. As a consequence, the Kalman filter and smoothing algorithm yield the *minimum mean square estimate* (MMSE) of the unobserved states. Hence, given the parameters the Kalman filter provides $\mathbf{F}_{t|t}^\dagger = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{F}_t^\dagger|\mathcal{X}_t]$ and $\mathbf{P}_{t|t}^\dagger = \text{Var}_{\boldsymbol{\theta}}[\mathbf{F}_t^\dagger|\mathcal{X}_t] = \mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{F}_t^\dagger - \mathbf{F}_{t|t}^\dagger)(\mathbf{F}_t^\dagger - \mathbf{F}_{t|t}^\dagger)'|\mathcal{X}_t]$ for $t \leq T$. The Kalman smoother, the recursion following the Kalman filter and starting from $t = T$ and going backwards to $t = 0$, provides the solution, delivering the smoothed process $\mathbf{F}_{t|T}^\dagger = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{F}_t^\dagger|\mathcal{X}_T]$ and its covariance $\mathbf{P}_{t|T}^\dagger = \text{Var}_{\boldsymbol{\theta}}[\mathbf{F}_t^\dagger|\mathcal{X}_T] = \mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{F}_t^\dagger - \mathbf{F}_{t|T}^\dagger)(\mathbf{F}_t^\dagger - \mathbf{F}_{t|T}^\dagger)'|\mathcal{X}_T]$.

Proposition 1 *Given the true value of the parameters* $\boldsymbol{\theta}$ *and initial condition* $\mathbf{F}_{0|0}^\dagger$ *and* $\mathbf{P}_{0|0}^\dagger$*, under Assumptions (A1) through (A5) one can prove that, as* $n \to \infty$*,*

$$\|\mathbf{F}_t - \mathbf{F}_{t|t}\| = O_p\left(\frac{1}{\sqrt{n}}\right) \qquad \textit{and} \qquad \|\mathbf{F}_t - \mathbf{F}_{t|T}\| = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{1.7}$$

[3] Computationally speaking, this doesn't affect parameter estimation as $(\mathbf{Q}_t^\dagger)^{-1}$ does not appear in any first-order condition (thus, after derivatives are computed) for Maximum Likelihood Estimators.

for any given t. Furthermore, the errors

$$\|\boldsymbol{\eta}_t - \boldsymbol{\eta}_{t|t}\| = O_p\left(\frac{1}{\sqrt{n}}\right) \qquad \textit{and} \qquad \|\boldsymbol{\eta}_t - \boldsymbol{\eta}_{t|T}\| = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{1.8}$$

for any given t.

Proof. See Appendix 1.7.4 and 1.7.3.

Thus, if the cross-sectional dimension tends to infinity both the Kalman filter and the Kalman smoother yield consistent estimates of the underlying factors (Doz et al., 2011). This result relies on the fact that as n increases the filter stochastic uncertainty decreases since we are averaging the cross-sectional errors. These proprieties are essential for estimation purposes as the model in (1.1) - (1.2) poses some non-trivial issues to deal with. In this regard, in standard conditionally heteroskedastic models the distribution of $\eta_t|\mathcal{I}_{t-1}$ is assumed to be Gaussian. Thus, if the state errors were directly observable, the model would be conditionally normal. However, $\mathbf{F}_t$ is a latent process and the filter manages the distribution of $\boldsymbol{\eta}_t$ conditional on past observations, $\boldsymbol{\eta}_t|\mathcal{X}_{t-1}$ not $\boldsymbol{\eta}_t|\mathcal{I}_{t-1}$. This leads the Kalman filter estimates to be *quasi-optimal* and the model can be treated as if it were conditionally Gaussian. Nevertheless, given the Kalman filter it is possible to evaluate both the conditional mean and the variance of $\eta_{i,t}$. While the former is simply zero the latter is given by:

$$\mathrm{Var}_{\boldsymbol{\theta}}[\eta_{i,t}|\mathcal{X}_{t-1}] = \mathbb{E}_{\boldsymbol{\theta}}[\eta_{i,t}^2|\mathcal{X}_{t-1}] = \omega_i + \alpha_i \mathbb{E}_{\boldsymbol{\theta}}[\eta_{i,t-1}^2|\mathcal{X}_{t-1}] + \beta_i \mathbb{E}_{\boldsymbol{\theta}}[q_{i,t-1}|\mathcal{X}_{t-1}]. \tag{1.9}$$

Furthermore, we can replace the expectation terms with their Kalman Filter estimates. Denoting $\eta_{i,t|t} = \mathbb{E}_{\boldsymbol{\theta}}[\eta_{i,t}|\mathcal{X}_t]$ and $P^{\eta}_{i,t|t} = \mathbb{E}_{\boldsymbol{\theta}}[(\eta_{i,t} - \eta_{i,t|t})^2|\mathcal{X}_t]$, we can use the the fact that $\mathbb{E}_{\boldsymbol{\theta}}[\eta_{i,t}^2|\mathcal{X}_t] = \eta_{i,t|t}^2 + P^{\eta}_{i,t|t}$. We then have

$$q_{i,t|t-1} = \omega + \alpha(\eta^2_{i,t-1|t-1} + P^{\eta}_{i,t-1|t-1}) + \beta q_{i,t-1|t-2} + \delta_{i,t}, \tag{1.10}$$

where δ_t is the correction term and it is equal to:

$$\delta_{i,t} \;=\; \sum_{j=1}^{\infty} \alpha\beta^j[(\eta^2_{i,t-j-1|t-1} - \eta^2_{i,t-j-1|t-2}) + (P^{\eta}_{i,t-j-1|t-1} - P^{\eta}_{i,t-j-1|t-2})]. \tag{1.11}$$

This filter requires past disturbances conditional on future time points. One solution is to use the fixed-point interval smoothing, adding to the state vector of (1.2) j lags of $\eta_{i,t-j}$. Nevertheless, this procedure dramatically increases the dimension of the matrices to be handled by the filter. It is important to note, however, that the correction terms are moderately small. First of all, the $\alpha\beta^j$ term is less then 1 and approach zero as j increases. The term within the square brackets in (1.11) is, instead, the difference between the filter and the smoother ($j = 1$) or, more generally ($j > 1$), between two smoothers calculated backwardly with $T = t - 1$ and $T = t - 2$, plus their variance difference.

Since we are working in an $n \to \infty$ environment, we can show (see Appendix) that those expressions go to 0 asymptotically, therefore having $\delta_t \to 0$. These results represent the key of a consistent estimate of the disturbance term $\boldsymbol{\eta}_t$, and the subsequent more convenient specification of the variance $q_{i,t}$, given that $\delta_{i,t}$ can be neglected.

Proposition 2 *Given the true value of the parameters $\boldsymbol{\theta}$, under assumptions (A1) through (A5) and Proposition 1 we have that, as $n \to \infty$*

$$\left|q_{i,t} - q_{i,t|t-1}\right| = O_p\left(\frac{1}{n}\right) \tag{1.12}$$

for any t and any i.

Proof See Appendix 1.7.6.

The Kalman filter variance term becomes the main proxy for the variance prediction. From the Kalman filter equations (see Appendix 1.7.1) we have that for a factor i

$$P^{\eta}_{i,t|t-1} = q_{i,t|t-1}. \tag{1.13}$$

Hence, the variance process is assumed to be equal to the estimate from the Kalman filter prediction. Thus, estimating the factors and their prediction errors also makes possible the correct estimation of the conditional variance $q_{t|t-1}$, which follows the GARCH(1,1) dynamic

$$q_{i,t|t-1} = \omega + \alpha\eta^2_{i,t-1|t-1} + \beta q_{i,t-1|t-2}. \tag{1.14}$$

Vector $\boldsymbol{\theta}$, however, is not known. Let us consider now the case in which parameters are not given.

1.3.2 Expectation Conditional Maximization Either

As pointed out in Calzolari et al. (2004), estimation issues arise from the fact that we observe only $\boldsymbol{x}_t$ so that our information set is actually $\mathcal{X}_t$ and not $\mathcal{I}_t$. Consequently, we cannot use the normality assumption to derive the log-likelihood. The diagonal elements of $\mathbf{Q}_t$ are not, indeed, measurable functions of $\mathcal{X}_t$ but are a function of lagged values of $\mathbf{F_t}$, which makes the exact form of the conditional density of $\boldsymbol{x}_t$ given $\mathcal{X}_t$ altogether unknown. As a result, the set of the parameters of interest $\boldsymbol{\theta}$

$$\boldsymbol{\theta} = [\underbrace{\text{vec}(\boldsymbol{\Lambda}), \text{vech}(\boldsymbol{\Gamma}), \text{vec}(\boldsymbol{\Phi})}_{\boldsymbol{\theta}_{(\mathcal{Q})}}, \underbrace{\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}}_{\boldsymbol{\theta}_{(\ell)}}], \tag{1.15}$$

cannot be estimated simultaneously on the basis of the log-likelihood function obtained from the observables $\boldsymbol{x}_t$. For this reason, we propose a methodology that efficiently exploits the available information set $\mathcal{X}_t$, which relies on the Kalman filter and, more generally, the ECME algorithm(Liu and Rubin, 1994).

Denote as $\boldsymbol{X}_T = (\boldsymbol{x}_1', \cdots, \boldsymbol{x}_T')'$ the nT-dimensional vectors of observations and $\boldsymbol{F}_T^\dagger = (\mathbf{F}_1^{\dagger\prime}, \cdots, \mathbf{F}_T^{\dagger\prime})'$ the $2rT$-dimensional vectors of factors. The Expectation Maximization algorithm is based on the idea that if we could observe the factors along with the observation, for generic values of the parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the joint log-likelihood of the complete data would be:

$$\begin{aligned} \ell(\boldsymbol{X}_T, \boldsymbol{F}_T^\dagger; \boldsymbol{\theta}) &= \ell(\boldsymbol{X}_T | \boldsymbol{F}_T^\dagger; \boldsymbol{\theta}) + \ell(\boldsymbol{F}_T^\dagger; \boldsymbol{\theta}) \\ &= \sum_{t=1}^{T} \ell(\boldsymbol{x}_t | \mathbf{F}_t^\dagger; \boldsymbol{\theta}) + \sum_{t=1}^{T} \ell(\mathbf{F}_t^\dagger | \mathbf{F}_{t-1}^\dagger; \boldsymbol{\theta}) + \ell(\mathbf{F}_0^\dagger; \boldsymbol{\theta}), \end{aligned} \tag{1.16}$$

where the last equality is given by the proprieties of State Space Models, specifically: (i) the state process $\mathbf{F}_t^\dagger$ is assumed to be a Markov process; (ii) conditionally on $\mathbf{F}_t^\dagger$, the $\boldsymbol{x}_t$s are independent and $\boldsymbol{x}_t$ depend on $\mathbf{F}_t^\dagger$ only (Petris et al., 2009). Under the Gaussian assumption and ignoring constants, this can be rewritten as

$$\begin{aligned} \ell(\boldsymbol{X}_T, \boldsymbol{F}_T^\dagger; \boldsymbol{\theta}) = & - \frac{T}{2} \log |\boldsymbol{\Gamma}| - \frac{1}{2} \sum_{t=1}^{T} (\boldsymbol{x}_t - \boldsymbol{\Lambda}^\dagger \mathbf{F}_t^\dagger)' \boldsymbol{\Gamma}^{-1} (\boldsymbol{x}_t - \boldsymbol{\Lambda}^\dagger \mathbf{F}_t^\dagger) \\ & - \frac{1}{2} \sum_{t=1}^{T} \log |\mathbf{Q}_t^\dagger| - \frac{1}{2} \sum_{t=1}^{T} (\mathbf{F}_t^\dagger - \boldsymbol{\Phi}^\dagger \mathbf{F}_{t-1}^\dagger)' \mathbf{Q}_t^{\dagger -1} (\mathbf{F}_t^\dagger - \boldsymbol{\Phi}^\dagger \mathbf{F}_{t-1}^\dagger) \\ & - \frac{1}{2} \log |\boldsymbol{\Omega}_0^\dagger| - \frac{1}{2} (\mathbf{F}_0^\dagger)' \boldsymbol{\Omega}_0^{\dagger -1} (\mathbf{F}_0^\dagger). \end{aligned} \tag{1.17}$$

Although we don't have the complete data, the algorithm provides an iterative method to find Maximum Likelihood (ML) estimates of the parameters, maximizing the likelihood only based on the observed data. As a matter of fact, further decomposing (1.16) using Bayes' rule one obtains

$$\ell(\boldsymbol{X}_T; \boldsymbol{\theta}) = \ell(\boldsymbol{X}_T | \boldsymbol{F}_T^\dagger; \boldsymbol{\theta}) + \ell(\boldsymbol{F}_T^\dagger; \boldsymbol{\theta}) - \ell(\boldsymbol{F}_T^\dagger | \boldsymbol{X}_T; \boldsymbol{\theta}). \tag{1.18}$$

Now, let us assume we have an estimate of the parameters for an iteration $j \geq 0$, say, $\widehat{\boldsymbol{\theta}}^{(j-1)}$. Taking expectation with respect to the distribution of $\boldsymbol{F}_t^\dagger$ conditionally on $\boldsymbol{X}_t$ and $\boldsymbol{\theta}$ on both sides of (1.18) we obtain

$$\ell(\boldsymbol{X}_T; \boldsymbol{\theta}) = \mathbb{E}_{\hat{\theta}^{(j-1)}} \left[\ell(\boldsymbol{X}_T | \boldsymbol{F}_T^\dagger; \boldsymbol{\theta}) + \ell(\boldsymbol{F}_T^\dagger; \boldsymbol{\theta}) | \boldsymbol{X}_T \right] - \mathbb{E}_{\hat{\theta}^{(j-1)}} \left[\ell(\boldsymbol{F}_T^\dagger | \boldsymbol{X}_T; \boldsymbol{\theta}) | \boldsymbol{X}_T \right] \tag{1.19}$$

$$= \mathcal{Q}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)}) - \mathcal{H}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)}). \tag{1.20}$$

Hence, maximizing $\mathcal{Q}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)})$ with respect to $\boldsymbol{\theta}$ in order to find $\hat{\boldsymbol{\theta}}^{(j-1)}$ is the same as maximizing the actual likelihood $\ell(\boldsymbol{X}_T; \boldsymbol{\theta})$. This result relies on the fact that for each j, $\mathcal{Q}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)}) \geq \mathcal{Q}(\widehat{\boldsymbol{\theta}}^{(j-1)}; \widehat{\boldsymbol{\theta}}^{(j-1)})$ by optimization and $\mathcal{H}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)}) \leq \mathcal{H}(\widehat{\boldsymbol{\theta}}^{(j-1)}; \widehat{\boldsymbol{\theta}}^{(j-1)})$ by Jensen's inequality, implying that $\ell(\boldsymbol{X}_T; \widehat{\boldsymbol{\theta}}^{(j)}) > \ell(\boldsymbol{X}_T; \widehat{\boldsymbol{\theta}}^{(j-1)})$ (Dempster et al., 1977). Calculation of $\mathcal{Q}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)})$, and thus expectation, is carried out through the Kalman smoother

which gives estimates for the factors at iteration j: $\mathbf{F}_{t|T}^{\dagger(j)} = \mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j)}}[\mathbf{F}_t^\dagger|\mathcal{X}_T]$. Intuitively, given the ML estimator of the parameters (M-step), the Kalman smoother gives an estimate of the factors (E-step), and vice versa, until convergence is reached. At this point, the last estimate of $\boldsymbol{\theta}$, under standard regularity conditions and up to a numerical error, is equivalent to the ML estimator of the parameters (Wu, 1983).
Nonetheless, not all parameters have closed-form solutions. This calls for the necessity to perform numerical optimization, possibly on a subset $\widetilde{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta}$, given all the other parameters that have an analytical solution. For this purpose, Meng and Rubin (1993) introduced the Expectation Conditional Maximization (ECM) which replaces the M-step of each iteration with a sequence of conditional or constrained maximization, or CM, steps. Each step $s \in \mathcal{S} = \{1, ..., S\}$, which may admit a closed-form solution or may require numerical optimization routines such as Newton-Raphson, maximizes the expected complete-data log-likelihood determined in the previous E-step. Since any individual step takes place in a reduced dimensional space, the optimization routine is faster and more reliable than maximizing the likelihood function on the whole parameter space.
In this framework, within the CM-step, instead of maximizing the expected complete-data log-likelihood function, we can perform the optimization on the actual log-likelihood. This approach leads to the adoption of the ECME algorithm, which replaces the constrained expected complete-data likelihood with the constrained actual function, subject to the same constraints on $\boldsymbol{\Theta}$. Using the same notation as Liu and Rubin (1994), the steps s are divided into two subspaces depending on whether the actual ($s \in \mathcal{S}_\ell$) or the expected likelihood ($s \in \mathcal{S}_\mathcal{Q}$) is maximized and such that $\mathcal{S}_\ell \cup \mathcal{S}_\mathcal{Q} = \mathcal{S}$. The authors demonstrated that the procedure shares the same proprieties and simplicity as the original one, but the convergence rate is considerably faster. Indeed, the likelihood to be maximized is the actual one rather than an approximation of it. It is important to note that, being the likelihood constructed from the filter, it has to be regarded as a Quasi Maximum Likelihood Estimation (QMLE), given that the disturbance $\boldsymbol{\eta}_t|\mathcal{X}_{t-1}$ has non-normal, although symmetric, distribution (Harvey et al., 1992).

In detail, the ECME algorithm consists of:

Expectation step (E-step)
The Kalman smoother is used to estimate the factors and compute the expected likelihood given the parameters of the model. Let us consider (1.18) with $\mathcal{Q}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)}) = \mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j-1)}}\big[\ell(\boldsymbol{X}_T, \boldsymbol{F}_T^\dagger; \boldsymbol{\theta})|\boldsymbol{X}_T\big]$; then, up to an initial condition, which is negligible for large T, we have:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)}) = & - \frac{T}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\mathrm{tr}\Big\{\boldsymbol{\Gamma}^{-1}\sum_{t=1}^{T}\mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j-1)}}\big[(\boldsymbol{x}_t - \boldsymbol{\Lambda}^\dagger\mathbf{F}_t^\dagger)(\boldsymbol{x}_t - \boldsymbol{\Lambda}^\dagger\mathbf{F}_t^\dagger)'|\boldsymbol{X}_T\big]\Big\} \qquad (1.21)\\ & - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{Q}_t^\dagger| - \frac{1}{2}\mathrm{tr}\Big\{\sum_{t=1}^{T}\mathbf{Q}_t^{\dagger-1}\mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j-1)}}\big[(\mathbf{F}_t^\dagger - \boldsymbol{\Phi}^\dagger\mathbf{F}_{t-1}^\dagger)(\mathbf{F}_t^\dagger - \boldsymbol{\Phi}^\dagger\mathbf{F}_{t-1}^\dagger)'|\boldsymbol{X}_T\big]\Big\}.\end{aligned}$$

Expectations are then replaced in the CM steps by the sufficient statistics given by the smoother:

$$\mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j)}}[\boldsymbol{x}_t\mathbf{F}_t^{\dagger\prime}|\mathcal{X}_T] = \boldsymbol{x}_t\mathbf{F}_{t|T}^{\dagger(j)\prime}, \tag{1.22}$$

$$\mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j)}}[\mathbf{F}_t^{\dagger}\mathbf{F}_t^{\dagger\prime}|\mathcal{X}_T] = \mathbf{F}_{t|T}^{\dagger(j)}\mathbf{F}_{t|T}^{\dagger(j)\prime} + \mathbf{P}_{t|T}^{\dagger(j)}, \tag{1.23}$$

$$\mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j)}}[\mathbf{F}_t^{\dagger}\mathbf{F}_{t-1}^{\dagger\prime}|\mathcal{X}_T] = \mathbf{F}_{t|T}^{\dagger(j)}\mathbf{F}_{t-1|T}^{\dagger(j)\prime} + \mathbf{P}_{t,t-1|T}^{\dagger(j)}, \tag{1.24}$$

where $\mathbf{P}_{t,t-1|T}^{\dagger(j)} = \mathrm{Cov}_{\widehat{\boldsymbol{\theta}}^{(j)}}[\mathbf{F}_{t|T}^{\dagger}, \mathbf{F}_{t-1|T}^{\dagger}|\mathcal{X}_T]$. and $\mathbf{P}_{t|T}^{\dagger(j)} = \mathrm{Var}_{\widehat{\boldsymbol{\theta}}^{(j)}}[\mathbf{F}_t^{\dagger}|\mathcal{X}_T]$.

Conditional Maximization steps (CM-steps)
These involve finding the solution to the maximization problem:

$$\widehat{\boldsymbol{\theta}}^{(j)} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\ \mathcal{Q}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(j-1)}) \qquad \text{or} \qquad \widehat{\boldsymbol{\theta}}^{(j)} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\ \ell(\boldsymbol{X}_T; \boldsymbol{\theta}), \tag{1.25}$$

depending on whether we want to maximize the complete-data or actual likelihood, respectively. The unknown $\boldsymbol{\theta}$ is partitioned into $\boldsymbol{\theta}_{(\mathcal{Q})}$ for $s \in \mathcal{S}_{\mathcal{Q}}$ and $\boldsymbol{\theta}_{(\ell)}$ for $s \in \mathcal{S}_{\ell}$, with $\mathcal{S}_{\mathcal{Q}} = \{1\}$ and $\mathcal{S}_{\ell} = \{2, ...r+1\}$.

(i) **CM-Step 1**. Analytical maximization of the expected likelihood given the factors for the parameters that have a closed form, i.e. $\boldsymbol{\theta}_{(\mathcal{Q})} = [\mathrm{vec}(\boldsymbol{\Lambda}), \mathrm{vech}(\boldsymbol{\Gamma}), \mathrm{vec}(\boldsymbol{\Phi})]$. Maximizing the expression above results in the estimators

$$\boldsymbol{\Lambda}^{\dagger(j)} = \left(\sum_{t=2}^{T} \mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j-1)}}\left[\boldsymbol{x}_t\mathbf{F}_t^{\dagger\prime}|\mathcal{X}_T\right]\right)\left(\sum_{t=2}^{T} \mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j-1)}}\left[\mathbf{F}_t^{\dagger}\mathbf{F}_t^{\dagger\prime}|\mathcal{X}_T\right]\right)^{-1}, \tag{1.26}$$

$$\boldsymbol{\Phi}^{\dagger(j)} = \left(\sum_{t=2}^{T} \mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j-1)}}\left[\mathbf{F}_t^{\dagger}\mathbf{F}_{t-1}^{\dagger\prime}|\mathcal{X}_T\right]\right)\left(\sum_{t=2}^{T} \mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j-1)}}\left[\mathbf{F}_{t-1}^{\dagger}\mathbf{F}_{t-1}^{\dagger\prime}|\mathcal{X}_T\right]\right)^{-1}, \tag{1.27}$$

$$\boldsymbol{\Gamma}^{(j)} = \mathrm{diag}\left\{\frac{1}{T}\sum_{t=2}^{T} \mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(j-1)}}\left[(\boldsymbol{x}_t - \boldsymbol{\Lambda}^{\dagger(j)}\mathbf{F}_t^{\dagger})(\boldsymbol{x}_t - \boldsymbol{\Lambda}^{\dagger(j)}\mathbf{F}_t^{\dagger})'|\mathcal{X}_T\right]\right\}. \tag{1.28}$$

All the equations above conveniently also apply to block-diagonal matrices. Therefore, the derivation generalizes to a subset of the matrices of interest, allowing some freedom when the model, such as (1.5) and (1.6), requires restriction. That is why, in the next section, we will consider the sub-matrices $\boldsymbol{\Lambda}$, $\mathbf{F}_t$ or $\boldsymbol{\Phi}$ instead of $\boldsymbol{\Lambda}^{\dagger}$, $\mathbf{F}_t^{\dagger}$ or $\boldsymbol{\Phi}^{\dagger}$. Thus, if we are interested in row a to b and column c to d of a matrix, all matrices on the right-hand side of the equations should be taken with respect of those indices.

(ii) **CM-Step 2,...,r+1**. Numerical maximization of the actual likelihood given the factors for parameters $\boldsymbol{\theta}_{(\ell)} = [\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}]$. Following the usual prediction error decomposition, with $\boldsymbol{e}_t = \boldsymbol{x}_t - \boldsymbol{\Lambda}^{\dagger}\mathbf{F}_t^{\dagger} = \boldsymbol{x}_t - \boldsymbol{\Lambda}\mathbf{F}_t$, the likelihood of $\boldsymbol{X}_T$ is the product of all

the conditional distributions of $\boldsymbol{x}_t$. The log-likelihood becomes:

$$\ell(\boldsymbol{X}_T;\boldsymbol{\theta}) = -\frac{1}{2}\sum_{t=1}^{T}\log|\boldsymbol{\Sigma}_t| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{e_t'}\boldsymbol{\Sigma_t^{-1}}\boldsymbol{e_t} \tag{1.29}$$

and parameters $\boldsymbol{\theta}_{(\ell)}$ enter into the maximization through $\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda}\mathbf{P}_{t|t-1}\boldsymbol{\Lambda}' + \boldsymbol{\Gamma}$ since $\mathbf{P}_{t|t-1} = \boldsymbol{\Phi}\mathbf{P}_{t-1|t-2}\boldsymbol{\Phi}' + \boldsymbol{\Psi}\mathbf{Q}_t(\boldsymbol{\theta}_{(\ell)})\boldsymbol{\Psi}' - \mathbf{K}_t\boldsymbol{\Sigma}_{t-1}\mathbf{K}_t'$. where $\mathbf{K}_t$ is the Kalman gain and it is defined as $\boldsymbol{\Phi}\mathbf{P}_{t|t-1}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_t^{-1}$.
For each factor $i = 1, ..., r$ it is possible to carry a separate and subsequent numerical optimization step s for the parameters ω_i, α_i and β_i. To reduce the computational complexity of the optimization and to satisfy $\text{Var}[F_{i,t}] = 1$, we employ variance targeting estimation (Francq and Zakoïan, 2010; Francq et al., 2011). This relies on reparameterization of the volatility equation in which the intercept ω_i is replaced by the unconditional variance, thus obtaining $\omega_i = (1-\phi_i^2)(1-\alpha_i-\beta_i)$.[4]

The algorithm terminates when the stopping criterion is achieved : $|\ell(\boldsymbol{X}_T;\widehat{\boldsymbol{\theta}}^{(j)}) - \ell(\boldsymbol{X}_T;\widehat{\boldsymbol{\theta}}^{(j-1)})| / |\ell(\boldsymbol{X}_T;\widehat{\boldsymbol{\theta}}^{(j)})| < \tau$, where $\ell(\boldsymbol{\theta}^{(j)})$ represent the actual log-likelihood at the j^{th} iteration and φ is the tolerance parameter.

1.3.3 Two-step PCA estimator

A straightforward way to estimate this model is in two steps, using Principal Components Analysis (PCA) to estimate factors and loadings and then using Quasi Maximum Likelihood to obtain GARCH parameters.
More specifically, define the sample covariance matrix of $\boldsymbol{x}_t$ as

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'. \tag{1.30}$$

Using spectral decomposition we have that $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{W}}\widehat{\mathbf{L}}\widehat{\mathbf{W}}'$, with $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{L}}$ being the matrix of normalized eigenvectors and eigenvalues of the sample covariance matrix of $\boldsymbol{x}_t$, respectively. Let us assume that the number of factors r is given. If both Assumptions (A4) and (A5) hold, then, following Fan et al. (2013), $\mathbf{F}_t$ are the the principal components of $\boldsymbol{\Lambda}\mathbf{F}_t$ rescaled by the diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}^{1/2}$. PCA estimates are given by:

$$\widehat{\boldsymbol{\Lambda}}^{PCA} = \widehat{\mathbf{W}}_r\widehat{\mathbf{L}}_r^{1/2}, \qquad \widehat{\mathbf{F}}_t^{PCA} = \widehat{\mathbf{L}}_r^{-1/2}\widehat{\mathbf{W}}_r'\boldsymbol{x}_t, \tag{1.31}$$

where the columns in $\widehat{\mathbf{W}}_r$ are the r eigenvectors corresponding to the largest eigenvalues collected in the diagonal matrix $\widehat{\mathbf{L}}_r$ of dimension $r \times r$. If the non-zero eigenvalues diverge linearly in n and that they are asymptotically distinct, then it implies that $\widehat{\boldsymbol{\Lambda}}^{PCA}$ is consistent.

[4]For an AR(1)-GARCH(1,1) process the unconditional variance is given by $\text{Var}[F_{i,t}] = \frac{\omega_i}{(1-\phi_i^2)(1-\alpha_i-\beta_i)}$.

In the second step, GARCH parameters are estimated using Quasi Maximum Likelihood on the obtained $\widehat{\mathbf{F}}_t$:

$$\widehat{\boldsymbol{\theta}}^{PCA} = \underset{\boldsymbol{\theta}}{\text{argmax}} \ \ell(\widehat{\mathbf{F}}_t; \widehat{\boldsymbol{\theta}}). \tag{1.32}$$

This procedure is not entirely appropriate as the log-likelihood on $\widehat{\mathbf{F}}_t$ is not defined since $\widehat{\mathbf{F}}_t \neq \mathbf{F}_t$. However, PCA still provides efficient estimates of the factors $\mathbf{F}_t$, the loadings $\boldsymbol{\Lambda}$, and the idiosyncratic variances $\boldsymbol{\Gamma}$ (Doz et al., 2012), so we are going to use these pre-estimators to initialize the ECME. These parameters are independent of $\boldsymbol{\alpha}, \boldsymbol{\beta}, and \boldsymbol{\omega}$ since the latter maximizes the actual likelihood, instead of the complete one (Barigozzi and Luciani, 2019).
Denote as $\boldsymbol{\theta}^{(0)}$ the pre-estimator of the parameters. Let $\widehat{\mathbf{L}}_r$ be the diagonal matrix whose entries are the r-largest eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{W}}_r$ be the matrix of corresponding eigenvectors. Then,

$$\widehat{\boldsymbol{\Lambda}}^{(0)} = \widehat{\mathbf{W}}_r \widehat{\mathbf{L}}_r^{1/2}, \qquad \widehat{\mathbf{F}}_t^{(0)} = \widehat{\mathbf{L}}_r^{-1/2} \widehat{\mathbf{W}}_r' \boldsymbol{x}_t, \tag{1.33}$$

$$[\widehat{\boldsymbol{\Gamma}}^{(0)}]_{ii} = \frac{1}{T} \sum_{t=1}^{T} \left(x_{i,t} - \widehat{\boldsymbol{\lambda}}_i^{(0)\prime} \widehat{\mathbf{F}}_t^{(0)} \right)^2, \qquad i = 1, ..., n. \tag{1.34}$$

with $\widehat{\boldsymbol{\lambda}}_i^{(0)}$ being the i-th row of $\widehat{\boldsymbol{\Lambda}}^{(0)}$ and $[\widehat{\boldsymbol{\Gamma}}^{(0)}]_{ii} = 0$ if $i \neq j$.
AR-GARCH parameters are subsequently estimated by Quasi Maximum Likelihood (QML) on $\widehat{\mathbf{F}}_t^{(0)}$. Denote by $q_{t|t-1}^{PCA}$ the conditional variance estimated using this approach.

1.4 Further Topics

1.4.1 On Identification

Let us rewrite the augmented model of section (1.5) - (1.6)

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\Lambda} \mathbf{F}_t + \boldsymbol{\xi}_t, & (1.35) \\ \mathbf{F}_t &= \boldsymbol{\Phi} \mathbf{F}_{t-1} + \boldsymbol{\eta}_t + \boldsymbol{\eta}_t^* & (1.36) \\ \boldsymbol{\eta}_t &= \mathbf{Q}_t^{1/2} \widetilde{\boldsymbol{\eta}}_t. & (1.37) \end{aligned}$$

As mentioned in Section 1.2.2, one can obtain all the observationally equivalent structures through two $r \times r$ invertible matrices $\mathbf{U}$ and $\mathbf{U}^\star$ as follows (Burmeister et al., 1986):

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\Lambda} \mathbf{U} \mathbf{U}^{-1} \mathbf{F}_t + \boldsymbol{\xi}_t, & (1.38) \\ \mathbf{U}^{-1} \mathbf{F}_t &= \mathbf{U}^{-1} \boldsymbol{\Phi} \mathbf{U} \mathbf{U}^{-1} \mathbf{F}_{t-1} + \mathbf{U}^{-1} \boldsymbol{\eta}_t + \mathbf{U}^{-1} \boldsymbol{\eta}_t^* & (1.39) \\ \mathbf{U}^{-1} \boldsymbol{\eta}_t &= \mathbf{U}^{-1} \mathbf{Q}_t^{1/2} \mathbf{U}^\star \mathbf{U}^{\star -1} \widetilde{\boldsymbol{\eta}}_t. & (1.40) \end{aligned}$$

In order to fully identify the model we need some structure to preclude any other transformation different than $\mathbf{U} = \mathbf{U}^\star = \mathbf{I}_r$. This requires us to specify r^2 restriction on the model.

Identification Condition 1 (IC1) $\mathbb{E}[\mathbf{F}_t\mathbf{F}_t'] = \mathbf{I}_r$ *and the stochastic processes* $q_{i,t}$*'s for each* $i = 1, \cdots, r$ *are linearly independent, i.e.* $\nexists \boldsymbol{\delta} \in \mathbb{R}^r, \boldsymbol{\delta} \neq \mathbf{0} : \boldsymbol{\delta}' \boldsymbol{q}_t = 0 \ \forall t$*. This implies* $\mathbf{Q}_t$*,* $\mathbf{Q}$ *and* $\boldsymbol{\Phi}$ *are diagonal matrices.*

This identification scheme builds on the work of Sentana (1992) and Sentana and Fiorentini (2001), where the authors show that conditional heteroskedasticy actually alleviate identification problem. As a matter of fact, a *uniquely identified unconditional covariance* coupled with *linearly independent conditional variances* of the factor disturbances, ensures that the system (1.38) - (1.40) is statistically identified. Practically, this condition is satisfied if the conditional variances of at least $r > 1$ structural shocks are time-varying, given that these variances are empirically parametrized by the GARCH(1,1) processes (King et al., 1994). The framework of Sentana and Fiorentini (2001) is a specific case of (1.5) - (1.6) with $\boldsymbol{\Phi} = \mathbf{0}_r$ and $\epsilon = 0$, or $\mathbf{F}_t = \boldsymbol{\eta}_t$. They prove that if $\mathbf{Q}_t = \text{Var}[\boldsymbol{\eta}_t|\mathcal{I}_{t-1}]$ is diagonal (but not scalar) and $\mathbb{E}[\mathbf{Q}_t] = \mathbf{I}_r$, then these constrain ensure that $\mathbf{F}_t$ can be identified up to a sign. Furthermore, this result does not rely on any particular parameterisation of the dynamic conditional heteroskedasticity, only on the conditional orthogonality of the factors, the time-variation of their variances and the constancy of $\boldsymbol{\Lambda}$. This identification scheme has been employed in many conditionally heteroskedastic factor model applications such as King et al. (1994), Normandin and Phaneuf (2004) and Normandin (2004).
In our case, factors are autoregressive processes and their dynamic structure is specified by $\boldsymbol{\Phi}$. (A1) guarantees that $q_{i,t}$'s are linearly independent. This is implemented in the ECME simply by updating the recursive equation for $q_{i,t|t-1}$ and leaving 0 on the off-diagonal elements. Thus, we only need make sure that the unconditional variance $\boldsymbol{\Omega} = \text{Var}[\mathbf{F}_t]$ is identified. Given that that the data generating process of the factors is given by a stationary vector autoregression (VAR), we can calculate the unconditional variance $\boldsymbol{\Omega}$ as

$$\begin{aligned} \text{vec}(\boldsymbol{\Omega}) &= (\mathbf{I}_{r^2} - \boldsymbol{\Phi} \otimes \boldsymbol{\Phi})^{-1}\text{vec}(\mathbf{Q} + \mathbf{Q}^*) && (1.41) \\ &= (\mathbf{I}_{r^2} - \boldsymbol{\Phi} \otimes \boldsymbol{\Phi})^{-1}\text{vec}(\mathbf{Q}) + O(\epsilon) && (1.42) \end{aligned}$$

given that $\|\boldsymbol{\Phi}\| < 1$ and $\mathbf{Q}^* = O(\epsilon)$. The first $r(r-1)/2$ conditions come from the diagonality of $\mathbf{Q}$. This is a natural consequence of the model, as each $\eta_{i,t}$ follows its own GARCH(1,1) evolution, with no interaction with the the errors $\eta_{j,t}$, for $i \neq j$. Then, we make some restriction on $\boldsymbol{\Phi}$. At first let us assume that the matrix is diagonal. In particular, we assume that the coefficient matrix is diagonal, $\boldsymbol{\Phi} = \text{diag}(\phi_1, \cdots, \phi_r)$. Such types of constriction can be naturally handled by the ECME, as all the equations in the analytical CM steps (1.26) - (1.28) apply to block-diagonal matrices (Holmes, 2013). This implies further $r(r-1)/2$ restriction, since $\boldsymbol{\Omega}$ is now diagonal. Given the current specification, $r^2 - r$ restrictions, factors are identified up to a scale normalization. To achieve identification up to a sign we need further r constraints. While Bai and Wang (2015) assume that $\mathbf{Q}$ is an identity matrix, we will restrict $\boldsymbol{\Omega}$ instead. The reason becomes naturally apparent by employing the variance targeting estimator (VTE) in the conditional maximization step of the ECME. Indeed, if we impose that $\mathbf{Q} = \mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}'$, the right-hand side of (1.42) cancels out and we obtain that $\boldsymbol{\Omega} = \mathbf{I}_r + O(\epsilon)$. Then for $\epsilon \to 0$, $\boldsymbol{\Omega} = \mathbf{I}_r$. This resolve the remaining r restrictions needed.

This setup is useful when we are using mid-to-high frequency data and are mostly interested in the covariance matrix dynamics as opposed to factors relationships. It is also very natural, as we impose restrictions directly on the population variance and not on sample moments. Given that the ECME produces consistent estimators $\widehat{\mathbf{\Phi}}$, (Barigozzi and Luciani, 2019) then as $n, T \to \infty$, the identifying constraints implied by Assumption 4 hold asymptotically. We leave the factor sign indeterminacy untouched, as our main interest lies in the covariance matrix. Indeed, $q_{i,t}$, being the variance process, it is uniquely identified.

1.4.2 Prediction Distribution

Further to optimal prediction for the values of $\{\mathbf{F}_{t|T}\}$, for each t, the models in 1.5 and 1.6 deliver a straightforward 1-step-ahead prediction of $T+1$. The model assumes that $\mathbf{\Lambda}^\dagger, \mathbf{\Phi}^\dagger, \mathbf{\Psi}^\dagger$ are time-invariant.
Using Kalman filter formulas in 1.7.1, we have that the distribution of $\mathbf{F}_{T+1}$ given $\mathcal{X}_T$ is Gaussian with mean and variance

$$\begin{aligned}
\mathbf{F}^\dagger_{T+1|T} &= \mathbf{\Phi}^\dagger \mathbf{F}^\dagger_{T|T-1} + \mathbf{K}^\dagger_T(\boldsymbol{x}_t - \mathbf{\Lambda}^\dagger \mathbf{F}^\dagger_{T|T+1}) && (1.43)\\
\mathbf{P}^\dagger_{T+1|T} &= \mathbf{\Phi}^\dagger \mathbf{P}^\dagger_{T|T-1} \mathbf{\Phi}^{\dagger\prime} + \mathbf{\Psi}^\dagger \mathbf{Q}^\dagger_{T+1|T} \mathbf{\Psi}^{\dagger\prime} + \mathbf{K}^\dagger_T \mathbf{\Sigma}_{T|T-1} \mathbf{K}^{\dagger\prime}_T && (1.44)\\
q_{i,T+1|T} &= \omega_i + \alpha_i(\eta^2_{i,T|T} + P^\eta_{i,T|T}) + \beta_i q_{i,T|T-1} \qquad i = 1, \cdots r && (1.45)
\end{aligned}$$

where $q_{i,T+1|T}$ is the $2i^{th}$ element of the block matrix $\mathbf{Q}^\dagger_{T+1|T} = \text{diag}(\mathbf{Q}^*, \mathbf{Q}_{T+1|T})$ and $\mathbf{Q}^*$ is time-independent. [5]
Then, the distribution of $\boldsymbol{x}_{T+1}$ given $\mathcal{X}_T$ is also Gaussian with mean and variance

$$\begin{aligned}
\boldsymbol{x}_{T+1|T} &= \mathbf{\Lambda}^\dagger \mathbf{F}^\dagger_{T+1|T} && (1.46)\\
\mathbf{\Sigma}_{T+1|T} &= \mathbf{\Lambda}^\dagger \mathbf{P}^\dagger_{T+1|T} \mathbf{\Lambda}^{\dagger\prime} + \mathbf{\Gamma}. && (1.47)
\end{aligned}$$

Now consider the task of an h-step-ahead forecast for $h > 1$, i.e. making a prediction of future observation at times $T+2, \cdots, T+h$. Repeatedly substituting in the *transition equation* of 1.6 we obtain

$$\mathbf{F}^\dagger_{T+h} = (\mathbf{\Phi}^\dagger)^h \mathbf{F}^\dagger_T + \sum_{j=0}^{h-1} (\mathbf{\Phi}^\dagger)^j \mathbf{\Psi}^\dagger \boldsymbol{\eta}^\dagger_{T+j+1}. \tag{1.48}$$

Taking conditional expectation and variance at time T in (1.48), respectively, we have

$$\begin{aligned}
\mathbf{F}^\dagger_{T+h|T} &= (\mathbf{\Phi}^\dagger)^h \mathbf{F}^\dagger_{T|T} && (1.49)\\
\mathbf{P}^\dagger_{T+h|T} &= (\mathbf{\Phi}^\dagger)^h \mathbf{P}^\dagger_{T|T} (\mathbf{\Phi}^{\dagger\prime})^h + \sum_{j=0}^{h-1} (\mathbf{\Phi}^\dagger)^j \mathbf{\Psi}^\dagger \mathbf{Q}^\dagger_{T+j+1|T} \mathbf{\Psi}^{\dagger\prime} (\mathbf{\Phi}^{\dagger\prime})^j. && (1.50)
\end{aligned}$$

[5] An equivalent formulation that involves the filter instead of the prediction is $\mathbf{F}^\dagger_{T+1|T} = \mathbf{\Phi}^\dagger \mathbf{F}^\dagger_{T|T}$ and $\mathbf{P}^\dagger_{T+1|T} = \mathbf{\Phi}^\dagger \mathbf{P}^\dagger_{T|T} \mathbf{\Phi}^{\dagger\prime} + \mathbf{\Psi}^\dagger \mathbf{Q}^\dagger_{T+1|T} \mathbf{\Psi}^{\dagger\prime}$.

The only unknown in (1.50) is $\mathbf{Q}^{\dagger}_{T+h|T} = \mathbb{E}[\boldsymbol{\eta}^{\dagger}_{T+h}|\mathcal{X}_{\mathcal{T}}]$. The only time-dependent elements are the $q_{i,T+h|T}$ for $i = 1, \cdots, r$. One can prove (1.7.7):

$$\mathbb{E}[q_{i,T+h}|\mathcal{X}_T] = q_{i,T+h|T} = 1 - \phi_i^2 + (\alpha_i + \beta_i)^{h-1}(q_{i,T+1|T} - (1 - \phi_i^2)). \tag{1.51}$$

This result shows that as $h \to \infty$, $\mathbb{E}[q_{t+h}|\mathcal{X}_t]$ converges to its unconditional variance $1-\phi^2$. A high persistence, $(\alpha_i + \beta_i)$ close to 1, implies that shocks that deviate the conditional variance from its unconditional value will persist for a long time, but eventually the long-horizon prediction will be the long-run variance, $1 - \phi_i^2$.
Finally, we can derive the distribution of $\boldsymbol{x}_{T+h}$ taking the conditional expectation and variance with respect to $\mathcal{X}_T$ of the *measurement equation* (1.5)

$$\boldsymbol{x}_{T+h|T} = \boldsymbol{\Lambda}^{\dagger}\mathbf{F}^{\dagger}_{T+h|T} \tag{1.52}$$

$$\boldsymbol{\Sigma}_{T+h|T} = \boldsymbol{\Lambda}^{\dagger}\mathbf{P}^{\dagger}_{T+h|T}\boldsymbol{\Lambda}^{\dagger'} + \boldsymbol{\Gamma}. \tag{1.53}$$

It's also possible to evaluate multi-step prediction by iteratively applying the Kalman filter prediction equation in (1.43) - (1.45).

1.4.3 Factor consistency with unknown parameters

Convergence rates hold when using the true value of the parameters $\boldsymbol{\theta}$, as $n \to \infty$. The conditions also remain valid when using the QML estimators of the parameters as long as both $T, n \to \infty$. Following Barigozzi and Luciani (2019), given the QML estimator of the parameters $\boldsymbol{\theta}^*$ obtained from the ECME algorithm, let $\mathbf{F}^*_{t|t}$ and $\mathbf{F}^*_{t|T}$ be the factor estimates obtained by the Kalman filter and smoother, respectively, then we have that

$$\min(\sqrt{n}, \sqrt{T})\|\mathbf{F}^*_{t|t} - \mathbf{F}_t\| = O_p(1), \qquad \min(\sqrt{n}, \sqrt{T})\|\mathbf{F}^*_{t|T} - \mathbf{F}_t\| = O_p(1) \tag{1.54}$$

with $\mathbf{F}^*_{t|t} = \mathbb{E}_{\boldsymbol{\theta}^*}[\mathbf{F}_t|\mathcal{X}_t]$ and $\mathbf{F}^*_{t|T} = \mathbb{E}_{\boldsymbol{\theta}^*}[\mathbf{F}_t|\mathcal{X}_T]$.

1.4.4 Standard Errors

The ECME algorithm does not directly generate standard errors. However, the Hessian matrix at the time of convergence can be used as an estimate of

$$\mathcal{I}(\boldsymbol{\theta}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[-\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] \tag{1.55}$$

to subsequently obtain standard errors estimates. In this case, no analytical derivatives are calculated, but following Shumway and Stoffer (2006) we can include a numerical evaluation of the Hessian matrix at the time of convergence. The main peculiarity of the ECME is that we can calculate the value of the actual likelihood, in contrast to the complete data likelihood, at any moment. Thus, we replace the numerical Hessian of $\ell(\boldsymbol{\theta})$ with the one calculated using $\ell(\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is the QML estimator of the parameters at convergence.

1.5 Numerical Aspects

Let us examine separately some simple cases.

1.5.1 ARCH(1) with one factor

With one factor, the Model of (1.5) - (1.6) can be written as:

$$\begin{aligned} \boldsymbol{x}_t &= \underbrace{\begin{bmatrix}\boldsymbol{\lambda} & \mathbf{0}\end{bmatrix}}_{\boldsymbol{\Lambda}^\dagger} \mathbf{F}_t^\dagger + \boldsymbol{\xi}_t \\ \mathbf{F}_t^\dagger &= \begin{bmatrix} F_t \\ \eta_t \end{bmatrix} = \underbrace{\begin{bmatrix}\phi & 0 \\ 0 & 0\end{bmatrix}}_{\boldsymbol{\Phi}^\dagger} \underbrace{\begin{bmatrix} F_{t-1} \\ \eta_{t-1} \end{bmatrix}}_{\boldsymbol{F}_{t-1}^\dagger} + \underbrace{\begin{bmatrix}1 & 1 \\ 0 & 1\end{bmatrix}}_{\boldsymbol{\Psi}^\dagger} \underbrace{\begin{bmatrix} \eta_t^* \\ \eta_t \end{bmatrix}}_{\boldsymbol{\eta}_t^\dagger} \end{aligned}$$

with $\boldsymbol{x}_t$ and $\boldsymbol{\xi}$ being $n \times 1$ vectors, $\boldsymbol{\Lambda}^\dagger$ an $n \times 2$ matrix, and $\boldsymbol{\Phi}^\dagger$ and $\boldsymbol{\Psi}^\dagger$ square matrices of dimension 2×2. As for the variance, using the results in (1.10) we have:

$$\text{Var}[\eta_t|\mathcal{X}_{t-1}] = q_{t|t-1} = \omega + \alpha(\eta^2_{t-1|t-1} + P^\eta_{t-1|t-1}).$$

The first term in the parentheses is the estimate from the Kalman filter and the last one is its variance.

In order to initialize the algorithm, PCA decomposition is performed in order to extract the factor. Given $\mathbf{W}$, the $n \times n$ matrix of eigenvectors, with $\mathbf{w}_1$ the first eigenvector and λ_1 its largest eigenvalue, then

$$\widehat{F}_t^{(0)} = \frac{\mathbf{w}_1^{'}\boldsymbol{x}_t}{\sqrt{\lambda_1}}, \qquad \widehat{\boldsymbol{\lambda}}^{(0)} = \mathbf{w}_1\sqrt{\lambda_1}. \tag{1.56}$$

Then, we obtain the diagonal entries of the matrix $\boldsymbol{\Gamma}^{(0)} = \text{diag}(\sigma^2_{\xi_0,1}, \ldots \sigma^2_{\xi_0,n})$, by calculating the sample variance of each observation residuals such that

$$\sigma^2_{\xi_0,i} = \frac{1}{T-1}\sum(x_{i,t} - \hat{\lambda}_i^{(0)}\widehat{F}_t^{(0)})^2. \tag{1.57}$$

ARCH parameters ω_0, α_0, and ϕ_0 are estimated on $\widehat{F}_t^{(0)}$ using usual AR-ARCH(1) Maximum Likelihood optimization. This procedure is called the two-step estimation.
Knowing that $\text{Var}[\eta_t] = Cov[F_t, \eta_t] = (1-\phi^2)$ we set q_0 equal to the unconditional variance of η_t , $q_0 = (1-\phi^2)$. The initial state F_0 is fixed at 0 and its variance $\sigma^2_{F_0} = \text{Var}[F_t] = 1$.

Given the initial values, the ECME algorithm is used to estimate parameters.

CM-Step 1. In regards to the maximization procedure, given the formulas in (1.26) - (1.28) the parameters $\boldsymbol{\theta}_{(\mathcal{Q})}$ are obtained by:

$$\begin{aligned}
\boldsymbol{\lambda}^{(j)} &= \left(\sum_{t=1}^{T} \boldsymbol{x}_t F_{t|T}\right)\left(\sum_{t=1}^{T} F_{t|T}^2 + P_{t|T}^F\right)^{-1}, \\
\phi^{(j)} &= \left(\sum_{t=1}^{T} F_{t-1|T} F_{t|T} + P_{t,t-1|T}^F\right)\left(\sum_{t=1}^{T} F_{t-1|T}^2 + P_{t-1|T}^F\right)^{-1}, \\
\boldsymbol{\Gamma}^{(j)} &= \operatorname{diag}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{u}_t \boldsymbol{u}_t^{'} + P_{t|T}^F \boldsymbol{\lambda}^{(j)} \boldsymbol{\lambda}^{(j)'}\right)\right\},
\end{aligned}$$

with $\boldsymbol{u}_t = \boldsymbol{x}_t - \boldsymbol{\lambda} F_t$. The term $P_{t|T}^F$ represents the variance of the state F_t given by the Kalman smoother. Also, the matrix containing ϕ is restricted to be 0 on all parameters except the upper-left corner so the ϕ calculation is treated as univariate instead of using the full $\boldsymbol{\Phi}^{\dagger(j)}$ matrix. The same reasoning applies for the vector $\boldsymbol{\lambda}^{(j)}$.
This procedure also provides the starting values of F_0 and Γ_0 given by the KS estimates $F_{0|T}$ and $P_{0|T}$.

CM-Step 2. As for the variance persistence parameters, $\boldsymbol{\theta}_{(\ell)} = (\omega, \alpha)'$, they are estimated via numerical constrained optimization of (1.16) through the Broyden–Fletcher–Goldfarb–Shanno (BFGS) procedure of Byrd et al. (1995). It is a quasi-Newton method in which each variable is given a lower and upper bound. Furthermore, variance targeting is employed for the parameter ω given that the long-run (unconditional) variance of F_t is equal to one. The constraints are the following:

$$\begin{aligned}
&1 < \alpha < 0, \\
&\omega = (1-\alpha)(1-\phi^2).
\end{aligned}$$

The optimization is now carried out only on the parameter α. As a matter of fact, once ϕ and α are estimated they are substituted in the last equality to obtain ω. This procedure, other than reducing the likelihood parameter space, assures the expected unitary unconditional variance of the factor. The Hessian matrix, used to estimate standard errors, is calculated numerically and the tolerance φ for the ECME algorithm is set to 10^{-6}.

1.5.2 GARCH(1,1) with one factor

With one factor, the model for the GARCH assumes the same structure as the previous case

$$\begin{aligned}\boldsymbol{x}_t &= \underbrace{\begin{bmatrix}\boldsymbol{\lambda} & \mathbf{0}\end{bmatrix}}_{\boldsymbol{\Lambda}^\dagger}\mathbf{F}_t^\dagger + \boldsymbol{\xi}_t \\ \mathbf{F}_t^\dagger &= \begin{bmatrix}F_t \\ \eta_t\end{bmatrix} = \underbrace{\begin{bmatrix}\phi & 0 \\ 0 & 0\end{bmatrix}}_{\boldsymbol{\Phi}^\dagger}\underbrace{\begin{bmatrix}F_{t-1} \\ \eta_{t-1}\end{bmatrix}}_{\mathbf{F}_{t-1}^\dagger} + \underbrace{\begin{bmatrix}1 & 1 \\ 0 & 1\end{bmatrix}}_{\boldsymbol{\Psi}^\dagger}\underbrace{\begin{bmatrix}\eta_t^* \\ \eta_t\end{bmatrix}}_{\boldsymbol{\eta}_t^\dagger}\end{aligned}$$

but in this case, following Harvey et al. (1992) plus the results in 2, the variance of η_t is given by

$$\text{Var}[\eta_t|\mathcal{X}_{t-1}] = q_{t|t-1} = \omega + \alpha(\eta_{t-1|t-1}^2 + P_{t-1|t-1}^{\eta}) + \beta q_{t-1|t-2},$$

where $\eta_{t-1|t-1}$ is the estimates from the Kalman Filter and $P_{t-1|t-1}^{\eta}$ its variance.

The model is estimated by means of the ECME algorithm, employing the same first conditional step as in the ARCH(1) and incorporating the variance persistence parameter β in the numerical optimization.

CM-Step 1. Same as ARCH(1).

CM-Step 2. Parameters $\boldsymbol{\theta}_{(\ell)} = (\omega, \alpha, \beta)'$ are obtained via numerical optimization of (1.16) employing the BFGS algorithm subject to the following restrictions:

$$\begin{aligned}&\alpha,\ \beta > 0, \\ &\alpha + \beta < 1, \\ &\omega = (1 - \alpha - \beta)(1 - \phi^2).\end{aligned}$$

To deal with the stricter inequality constraints the model is re-parametrized to take into account both non-negativity and stationarity. The sine transformation is applied to α and β so that $\alpha^\star = 0.99\text{sin}(\alpha)^2$ and $\beta^\star = (0.99 - \alpha)\text{sin}(\beta)^2$ maintain the domain within (0,1).

Again, the Hessian matrix is calculated numerically and the tolerance φ for the ECME algorithm is set to 10^{-6}.

1.5.3 GARCH(1,1) with two uncorrelated factors

With two uncorrelated factors following univariate GARCH(1,1) the model becomes:

$$\begin{aligned}
\boldsymbol{x}_t &= \underbrace{\begin{bmatrix}\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2 & \mathbf{0} & \mathbf{0}\end{bmatrix}}_{\boldsymbol{\Lambda}^\dagger} \mathbf{F}_t^\dagger + \boldsymbol{\xi}_t \\
\mathbf{F}_t^\dagger &= \begin{bmatrix}F_{1,t}\\ F_{2,t}\\ \eta_{1,t}\\ \eta_{2,t}\end{bmatrix} = \underbrace{\begin{bmatrix}\phi_1 & 0 & 0 & 0\\ 0 & \phi_2 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}}_{\boldsymbol{\Phi}^\dagger} \underbrace{\begin{bmatrix}F_{1,t-1}\\ F_{2,t-1}\\ \eta_{1,t-1}\\ \eta_{2,t-1}\end{bmatrix}}_{\mathbf{F}_{t-1}^\dagger} + \underbrace{\begin{bmatrix}1 & 0 & 1 & 0\\ 0 & 1 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}}_{\boldsymbol{\Psi}^\dagger} \underbrace{\begin{bmatrix}\eta^*_{1,t}\\ \eta^*_{2,t}\\ \eta_{1,t}\\ \eta_{2,t}\end{bmatrix}}_{\boldsymbol{\eta}_t^\dagger}
\end{aligned}$$

with $\boldsymbol{x}_t$ and $\boldsymbol{\xi}$ being $n \times 1$ vectors, $\boldsymbol{\Lambda}^\dagger$ an $n \times 4$ matrix, and $\boldsymbol{\Phi}^\dagger$ and $\boldsymbol{\Psi}^\dagger$ square matrices of dimension 4×4. The variance of the states is given by

$$\boldsymbol{\eta}_t^\dagger = \begin{bmatrix}\eta^*_{1t}\\ \eta^*_{2t}\\ \eta_{1t}\\ \eta_{2t}\end{bmatrix} = \begin{bmatrix}q^*_1 & 0 & 0 & 0\\ 0 & q^*_2 & 0 & 0\\ 0 & 0 & q_{1,t} & 0\\ 0 & 0 & 0 & q_{2,t}\end{bmatrix}^{1/2} \tilde{\boldsymbol{\eta}}_{\boldsymbol{t}},$$

with $\tilde{\boldsymbol{\eta}}_t \sim NID(\mathbf{0}, \mathbf{I}_4)$ and q^*_1, q^*_2 equal to 10^{-6}. Univariate variances in the lower-right block matrix follow the usual GARCH(1,1) dynamic

$$q_{1,t|t-1} = \mathrm{Var}[\eta_1|\mathcal{X}_{t-1}] = \omega_1 + \alpha_1(\eta^2_{1,t-1|t-1} + P^{\eta_1}_{t-1|t-1}) + \beta_1 q_{1,t-1|t-2}, \tag{1.58}$$

$$q_{2,t|t-1} = \mathrm{Var}[\eta_2|\mathcal{X}_{t-1}] = \omega_2 + \alpha_2(\eta^2_{2,t-1|t-1} + P^{\eta_2}_{t-1|t-1}) + \beta_2 q_{2,t-1|t-2}, \tag{1.59}$$

where the first term in the parentheses is the estimate from the Kalman filter of the 3^{rd} and 4^{th} states and the last one is their variances.
As before, to start the algorithm, PCA is performed to extract the factors. Given $\widehat{\mathbf{W}}$, the $n \times 2$ matrix of the first two eigenvectors, and $\widehat{\mathbf{L}}$, the diagonal 2×2 matrix of corresponding eigenvalues, we have

$$\widehat{\boldsymbol{\Lambda}}^{(0)} = \widehat{\mathbf{L}}^{1/2}\widehat{\mathbf{W}}', \qquad \widehat{\mathbf{F}}_t^{(0)} = \widehat{\mathbf{L}}^{-1/2}\widehat{\mathbf{W}}'\boldsymbol{x}_t.$$

The diagonal entries of the matrix $\boldsymbol{\Gamma}_0$ are obtained by the sample variance of the residual of the observation equation using $\widehat{\mathbf{F}}_t^{(0)}$, as in (1.57). Variance persistence parameters $\boldsymbol{\omega}_0$, $\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0$, and $\boldsymbol{\phi}_0$ are estimated separately on the two factors $\widehat{F}^{(0)}_{1,t}$ and $\widehat{F}^{(0)}_{2,t}$ using usual AR-GARCH(1,1) QMLE. Starting variances $q_{1,0}$ and $q_{2,0}$ are initiated at their unconditional values.
The initial state $\mathbf{F}_0$ is fixed at $\mathbf{0}$, while its initial state variance is given by:

$$\boldsymbol{\Omega}_0 = \begin{bmatrix}1 & 0 & 1-\phi_1^2 & 0\\ 0 & 1 & 0 & 1-\phi_2^2\\ 1-\phi_1^2 & 0 & 1-\phi_1^2 & 0\\ 0 & 1-\phi_2^2 & 0 & 1-\phi_2^2\end{bmatrix}. \tag{1.60}$$

In this case, the ECME Algorithm is made up of 3 steps, such that $\mathcal{S}_{\mathcal{Q}} = \{1\}$ and $\mathcal{S}_{\ell} = \{2, 3\}$.

CM-Step 1. With two factors the analytical solutions to the expected likelihood optimization problem become:

$$\begin{aligned}
\boldsymbol{\Lambda}^{(j)} &= \left(\sum_{t=1}^{T} \boldsymbol{x}_t \mathbf{F}'_{t|T}\right)\left(\sum_{t=1}^{T} \mathbf{F}_{t|T}\mathbf{F}'_{t|T} + \mathbf{P}^{F}_{t|T}\right)^{-1}, \\
\phi_i^{(j)} &= \left(\sum_{t=1}^{T} F_{i,t-1|T}F_{i,t|T} + P^{F_i}_{t,t-1|T}\right)\left(\sum_{t=1}^{T} F^2_{i,t-1|T} + P^{F_i}_{t-1|T}\right)^{-1}, \qquad i = 1, 2 \\
\boldsymbol{\Gamma}^{(j)} &= \operatorname{diag}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{u}_t\boldsymbol{u}_t' + \boldsymbol{\Lambda}^{(j)}\mathbf{P}^{F}_{t|T}\boldsymbol{\Lambda}^{(j)'}\right)\right\}.
\end{aligned}$$

Having the off-diagonal elements restricted to be zero, the matrix $\boldsymbol{\Phi}$ is treated as containing two diagonal block matrices ϕ_1 and ϕ_2, to be estimated separately. The terms involved in the calculation are only the ones referring to the first or second elements of the matrices of interest.

CM-Step 2. The first numerical step involves the optimization of (1.16) with reference to the GARCH parameter of the first factor. Using the same *sine* transformation as in the univariate case, to take into account the usual restrictions:

$$\begin{aligned}
&\alpha_1,\ \beta_1 > 0, \\
&\alpha_1 + \beta_1 < 1, \\
&\omega_1 = (1 - \alpha_1 - \beta_1)(1 - \phi_1^2)
\end{aligned}$$

the optimization is performed to obtain α_1, β_1 and ω_1.

CM-Step 3. Same as the previous step, but the focus is on the second factor. The parameters of interest are α_2, β_2, and ω_2.

Tolerance φ is set to a slightly higher level equal to 10^{-5}. This is due to the fact that near the optimum the likelihood with two factors becomes very flat and for any new iterations the parameter value difference becomes negligible.

1.6 Simulations

1.6.1 Monte Carlo Simulation

We now describe a simulation study to explore the properties of our proposed approach for a correctly specified model and a misspecified one. Throughout the Monte Carlo study we

let $n \in \{75, 150, 300\}$, $T \in \{750, 1, 250\}$, $r = 2$ and we simulate according to (1.5), (1.6) and (1.3), (1.4), such that

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\Lambda}^\dagger \mathbf{F}_t^\dagger + \kappa^{1/2} \boldsymbol{\xi}_t, & (1.61) \\ \mathbf{F}_t^\dagger &= \boldsymbol{\Phi}^\dagger \mathbf{F}_{t-1}^\dagger + \boldsymbol{\eta}_t + \boldsymbol{\eta}_t^*, & (1.62) \end{aligned}$$

with $\boldsymbol{\Lambda}^\dagger = [\boldsymbol{\Lambda}', \mathbf{0}']'$ and $\mathbf{F}_t^\dagger = [\mathbf{F}_t', \boldsymbol{\eta}_t']'$. κ is a scalar which describes the inverse of the signal-to-noise ratio (SNR) of each factor since we have that $\text{Var}[f_{i,t}] = 1$ and $1/\kappa = \text{Var}[f_{i,t}]/\text{Var}[\xi_{i,t}]$. The conditional heteroskedasticity is given by the scalar factor disturbance $\eta_{i,t}$ with $i = 1, 2$

$$\begin{aligned} \boldsymbol{\eta}_t &= \mathbf{Q}_t^{1/2} \widetilde{\boldsymbol{\eta}}_t & (1.63) \\ q_{i,t} &= \omega_i + \alpha_i \eta_{i,t-1}^2 + \beta_i q_{i,t-1}. & (1.64) \end{aligned}$$

The factor loadings are *iid* and such that $[\boldsymbol{\Lambda}]_{1,j} \sim \mathcal{N}(0, 1)$ and $[\boldsymbol{\Lambda}]_{2,j} \sim \mathcal{N}(0, 0.5)$. The common factors evolve according to a VAR(1) with $\text{diag}(\boldsymbol{\Phi}) = [0.7, 0.2]$ and 0 elsewhere, so that the first factor has a stronger autoregressive component. The idiosyncratic innovations are such that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. As for the common innovations, they evolve according to a GARCH(1,1) with parameters $\boldsymbol{\alpha} = [0.3, 0.1]'$ and $\boldsymbol{\beta} = [0.6, 0.8]'$; we apply *variance targeting* to guarantee that the unconditional variance of the factor i is $\text{Var}[f_{i,t}] = 1$, restricting $\omega_i = (1 - \phi_i^2)(1 - \alpha_i - \beta_i)$. η_t^* is the misspecification term and it is distributed as $\eta_t^* \sim \mathcal{N}(\mathbf{0}, 10^{-8}\mathbf{I}_2)$. Given this scheme, we set $\kappa = 0.5$.
We consider B = 5,000 replications, and at each replication b, we simulate data and we estimate $\boldsymbol{\phi}, \boldsymbol{\alpha}, and \boldsymbol{\beta}$ with both the two-stage approach as in (1.32) and through the Kalman filter and ECME algorithm (1.25).
In order to assess the robustness of the model we also test:

1) Different values of SNR. In particular, we simulate with different values of $\kappa \in \{1, 2, 4\}$ which translates into SNR $\in \{1, 0.5, 0.25\}$. We fix $n, T = (100, 1, 000)$.

2) The response to a misspecified model with regards to the idiosyncratic components. Following Ahn and Horenstein (2013) we define $\xi_{i,t}$ as

$$\begin{aligned} \xi_{i,t} &= \sqrt{\frac{1 - \rho^2}{1 + 2J\tau^2}} e_{i,t} & (1.65) \\ e_{i,t} &= \rho e_{i,t-1} + \nu_{i,t} + \sum_{h=j^+}^{j-1} \tau \nu_{h,t} + \sum_{j+1}^{h=j^-} \tau \nu_{h,t} & (1.66) \end{aligned}$$

with $j^+ = \max(j - J, 1)$ and $j^- = \min(j + J, n)$. In practice, ρ specifies the serial correlation of the time series, while the cross-sectional correlation is defined by the two parameters τ and J. The former controls the magnitude of the correlation and the latter controls the number of cross-section units that are correlated. Figure 1.1 indicates the cross correlation structure among 100 components, varying the value of τ. We set up $\rho = 0.5, \tau = 0.3, J = \max(10, n/20)$.

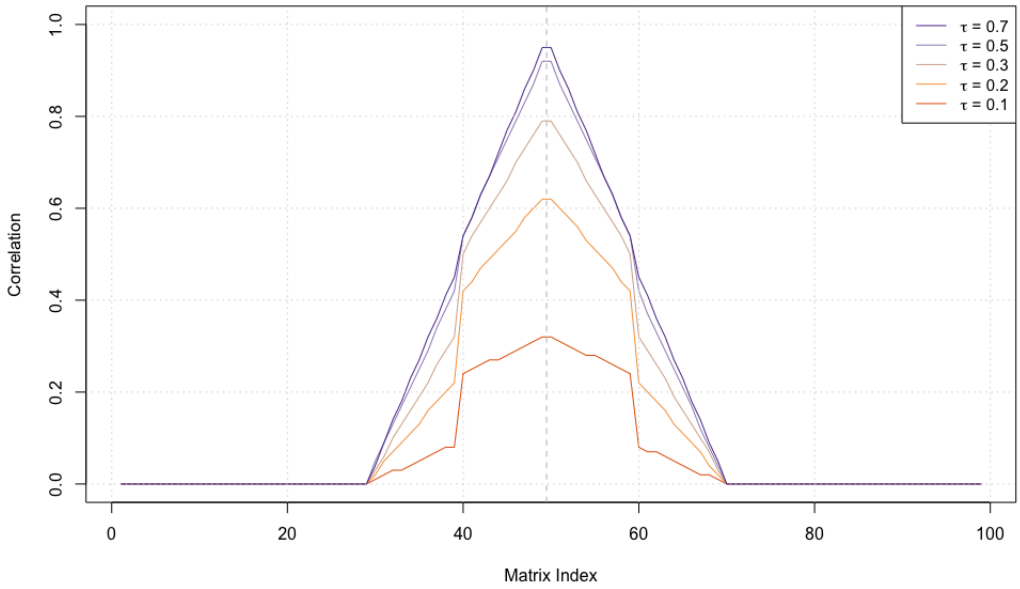

Figure 1.1: *Values of cross correlation for the element $\xi_{50,t}$ simulated according to* (1.66) *for different values of τ and $\rho = 0$. Given $n = 100$, we find $J = 10$, which shows that closer units exhibit higher degrees of correlation compared to distant counterparts.*

3) Small sample proprieties with $n, T = (50, 250)$, with SNR $\in \{2, 1\}$ and modest serial and cross correlation of the idiosyncratic component, i.e $\tau \in \{0, 0.2\}$ and $\rho \in \{0, 0.4\}$.

For each simulation exercise we also report the distribution of the Mean Squared Error (MSE) of the volatility process calculated as

$$\text{MSE} = \frac{1}{BT} \sum_{t=1}^{T} (q_t - \hat{q}_t)^2 \tag{1.67}$$

where $\hat{q}_t = q_{t|t-1}$ for the CHDFM model and $\hat{q}_t = q^{PCA}_{t|t-1}$ for the two-stage approach.

1.6.2 Results

As we can see from Figure 1.2 and Figure 1.11, the results in the graphs numerically validate the results outlined in Proposition 1 and Proposition 2. The CHDFM estimates are consistent even when n is not too large. For the two-stage approach (2SPCA) the higher the ϕ the higher the inconsistency in the estimates. This is shown particularly in Figure 1.2, since the first factor has a $\phi_1 = 0.7$ which produces inconsistent parameter estimates. As mentioned in Francq and Zakoïan (2010), when the innovation distribution

$\tilde{\eta}$ is symmetric then the asymptotic variance of the ARMA coefficient and GARCH coefficient is block diagonal, implying that those estimators are independent. Conversely, when $\tilde{\eta}$ is not symmetric the asymptotic distribution of the ARMA parameters depends on the GARCH coefficients. This makes the two-step approach inappropriate as the parameters are treated as independent. When ϕ approaches zero, estimates display a less pronounced bias, as shown by the parameter estimates of the second factor, whose $\phi_1 = 0.2$. It is also important to remark that when one or more components of $\boldsymbol{\theta}$ are null, asymptotic normality of QMLE is not satisfied. In this specific case, the asymptotic distribution of $\boldsymbol{\theta}$ cannot be Gaussian because the estimator is constrained. For instance, if $\alpha_i = 0$, the distribution of α_i is concentrated in $[0, \infty)$, for all n, and thus cannot be asymptotically normal. This 'boundary' problem is treated in Francq and Zakoïan (2010) and is the object of a specific study in Chapter 8 of the book. As indicated in Table 1.1, increasing the value of n from $n = 75$ to $n = 300$ reduces the MSE of the CHDFM model by about 20% for the first factor and 13% for the second one. 2SPCA achieves higher decrease in the MSE, 23% and 20% for the first and second factors respectively, but the absolute values are still two to three times higher compared to the CHDFM method. Parameter distributions are approximately normal for the CHDFM method. 2SPCA presents heavier tailed distributions whose means are biased when ϕ is significantly different from 0. In particular, one can see how the average ϕ and α are regularly underestimated, while β estimation is above the true values. Similar patterns are observed for $T = 1{,}250$. In this case, for both CHDFM and 2SPCA, the marginal reduction in MSE is more pronounced as n increases.
The successive simulation experiments are set to assess the robustness of the model. In particular, the model is first stressed introducing some degree of serial correlation whose magnitude depends on ρ. This specification has little impact on the CHDFM model while affecting 2SPCA parameter distributions to a greater extent. Conversely, cross correlation induced by τ and J seems to have a bigger influence on CHFDM parameters estimates compared to 2SPCA: MSE increases by 15% and 11% for the former model and 1% and 2% for the latter, for the first and second factor, respectively. The simulation that includes both serial and cross correlation displays similar estimates and MSE values to the one with $\tau = 0.3$ and $\rho = 0$. The second set of simulations tests the responsiveness of the model to different value of the signal-to-noise ratio. Higher values indicate that the variance of the common component (the signal) is lower in proportion to the idiosyncratic variance κ (the noise). Results indicate that parameter distribution of CHDFM changes moderately, with a slight increase of the standard deviations of the parameters' distributions. As a consequence, MSE increases by 21%, 47%, and 89% for $\kappa = 1, 2, 4$ for the first factor and by 13%, 34% and 56% for the second one. Higher SNR translates into higher MSE for the 2SPCA model as well, but, contrarily to its counterpart, a larger level of SNR deteriorates parameters' distributions significantly. Once again, the effect is greater for the first factor. For example, for $\alpha_1 = 0.3$ the two-step estimation process produces average estimates of 0.24, 0.22, and 0.2 for SNR equal to 1, 0.5, and 0.25, respectively. This is in contrast to the CHDFM estimates, which average around 0.3, 0.31, and 0.32 for the same values of SNR. Finally, we focus on smaller sample proprieties of the estimators. Surprisingly, the CHDFM estimation procedure guarantees parameter distributions to be approximately normal, al-

though with bigger standard errors with respect to previous estimation procedures. These are also mostly unaltered from error misspecification and higher SNRs. Two-step estimators are, on the contrary, not normally distributed and have average MSEs around 140% higher for the first factor and 40% higher for the second one.

					$q_{1,t}$		$q_{2,t}$	
n	T	κ	ρ	τ	CHDFM	2SPCA	CHDFM	2SPCA
75	750	0.5			2.9581	7.6852	2.6592	3.9931
150	750	0.5			2.5537	6.3707	2.4245	3.4683
300	750	0.5			2.3663	5.9385	2.3157	3.1979
75	1,250	0.5			3.4963	9.0125	2.8722	4.3064
150	1,250	0.5			2.8673	7.1958	2.5499	3.5268
300	1,250	0.5			2.6688	6.5079	2.3811	3.2004
100	1,000	0.5			2.9893	7.4371	2.7246	3.8052
100	1,000	1.0			3.6202	8.3306	2.9715	4.0739
100	1,000	2.0			4.3853	8.8643	3.5257	4.5584
100	1,000	4.0			5.6594	10.4406	4.4218	5.3338
100	1,000	0.5	0.0	0.0	2.9893	7.4371	2.7246	3.8052
100	1,000	0.5	0.5	0.0	2.9930	7.4108	2.5592	3.7562
100	1,000	0.5	0.0	0.3	3.4381	7.5054	3.0138	3.8949
100	1,000	0.5	0.4	0.2	3.3205	7.5007	3.0453	3.8930
50	250	0.5	0.0	0.0	2.3351	5.9086	2.4060	3.6628
50	250	1.0	0.0	0.0	2.7344	6.4347	2.6695	3.7599
50	250	1.0	0.4	0.2	2.6883	5.9911	2.8243	3.8298

Table 1.1: MSE values calculated as in (1.67) for different values of n, T, κ, ρ, τ.

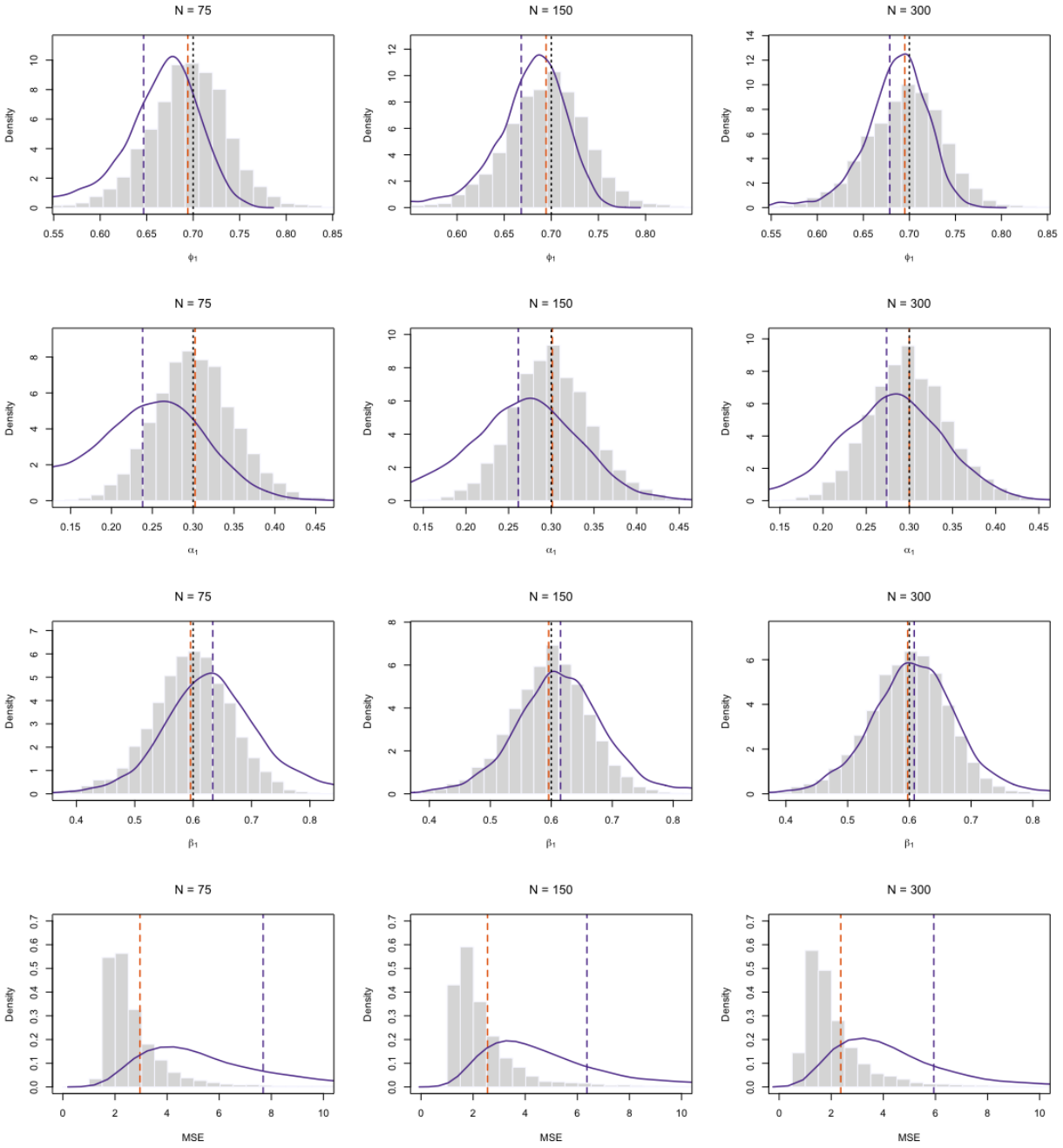

Figure 1.2: *Distributions of AR(1)-GARCH(1,1) parameters of the first factor (large ϕ), according to the two different models for $T = 750$. Last row indicates MSE. CHDFM parameters distributions are the grey histograms, while 2SPCA parameters densities are indicated by the purple lines. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

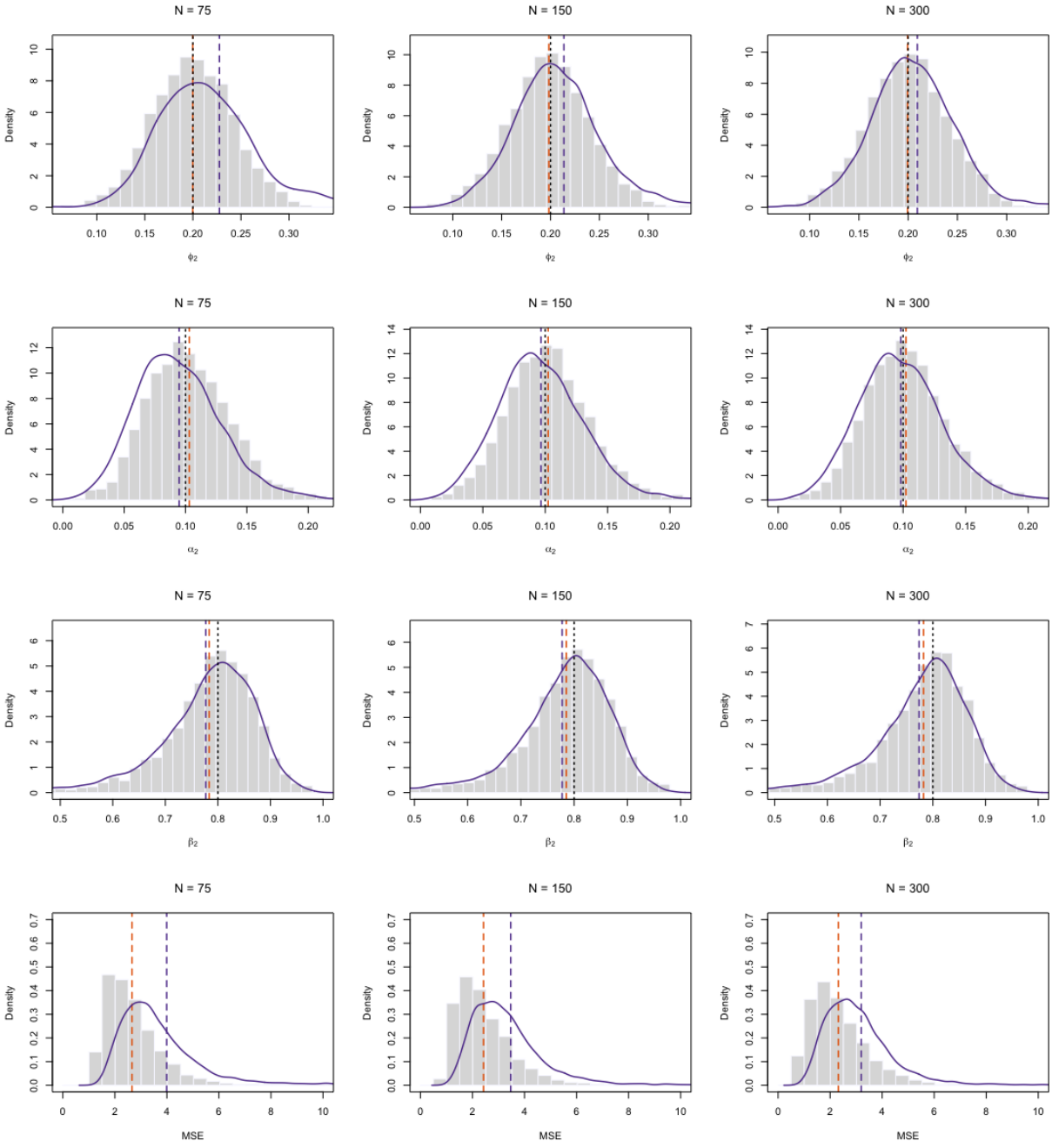

Figure 1.3: *Distributions of AR(1)-GARCH(1,1) parameters of the second factor (small ϕ), according to the two different models for $T = 750$. Last row indicates MSE. CHDFM parameters distributions are the grey histogram, while 2SPCA kernel density is indicated by the purple line. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

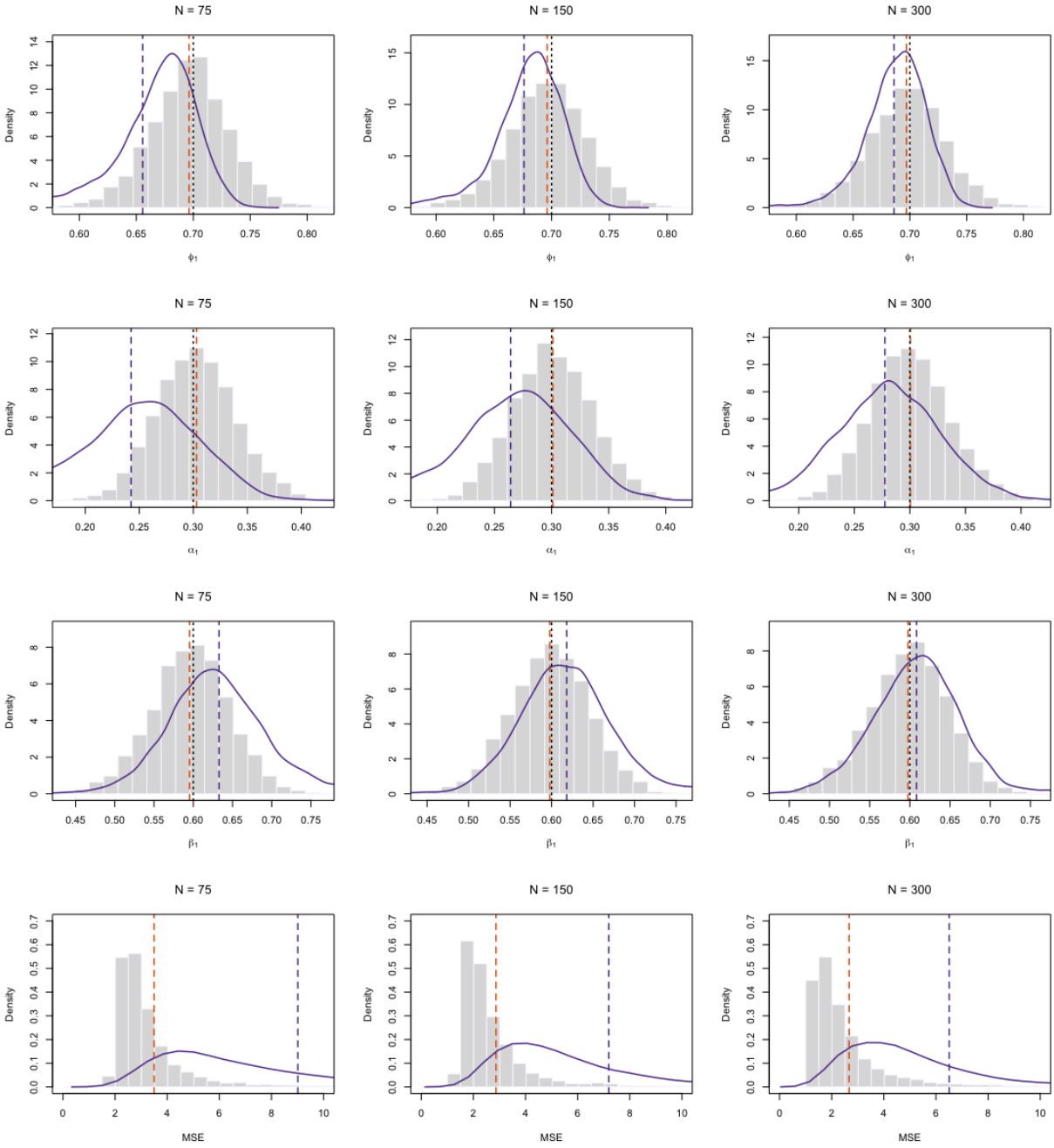

Figure 1.4: *Distributions of AR(1)-GARCH(1,1) parameters of the first factor (large ϕ), according to the two different models for $T = 1,250$. Last row indicates MSE. CHDFM parameters distributions are the grey histograms, while 2SPCA parameters densities are indicated by the purple lines. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

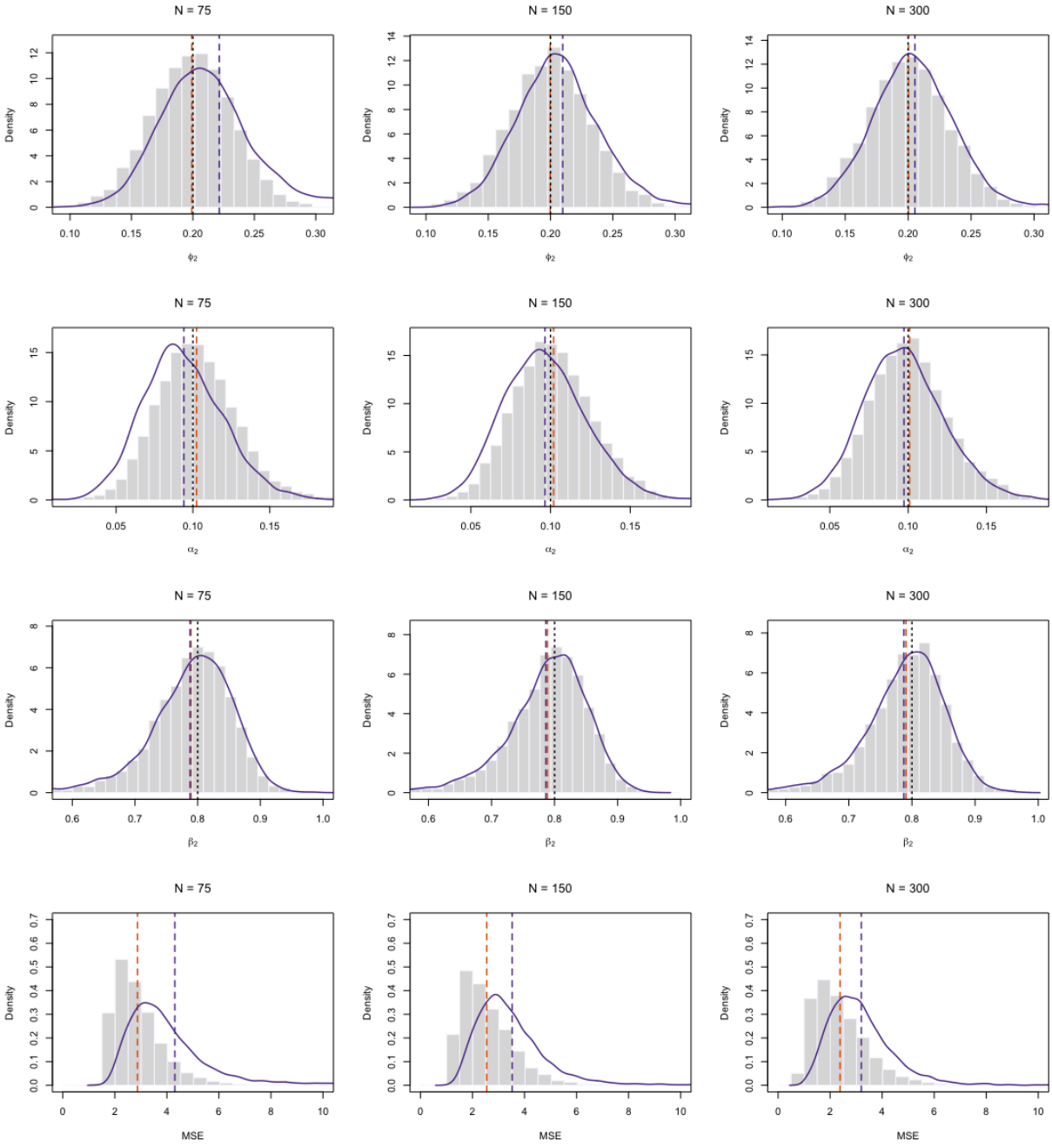

Figure 1.5: *Distributions of AR(1)-GARCH(1,1) parameters of the second factor (small ϕ), according to the two different models for $T = 1,250$. Last row indicates MSE. CHDFM parameters distributions are the grey histogram, while 2SPCA kernel density is indicated by the purple line. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

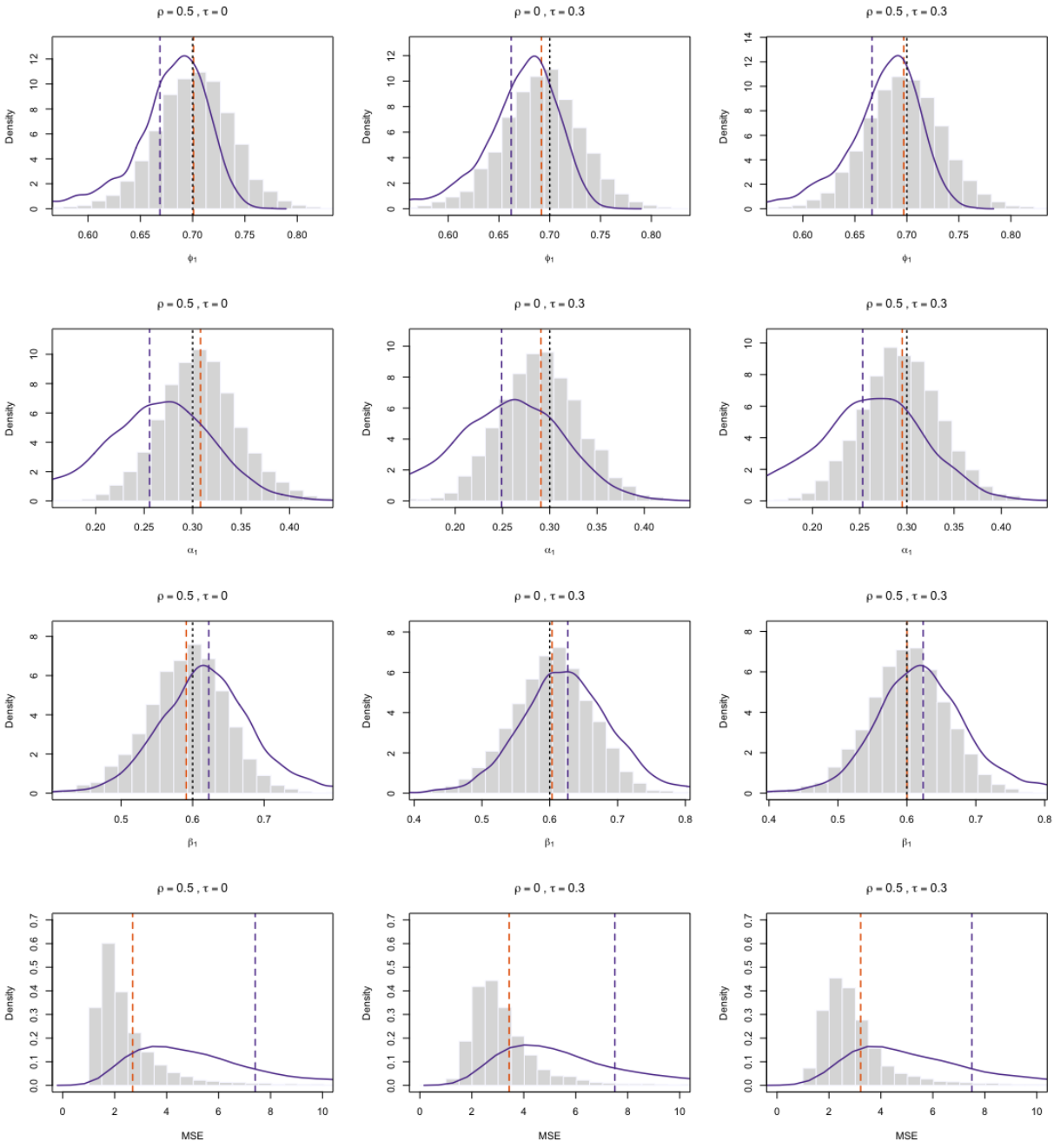

Figure 1.6: *Distributions of AR(1)-GARCH(1,1) parameters of the first factor (large ϕ), according to the two different models for $T = 1000$ and $n = 100$. Last row indicates MSE. CHDFM parameters distributions are the grey histograms, while 2SPCA parameters densities are indicated by the purple lines. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

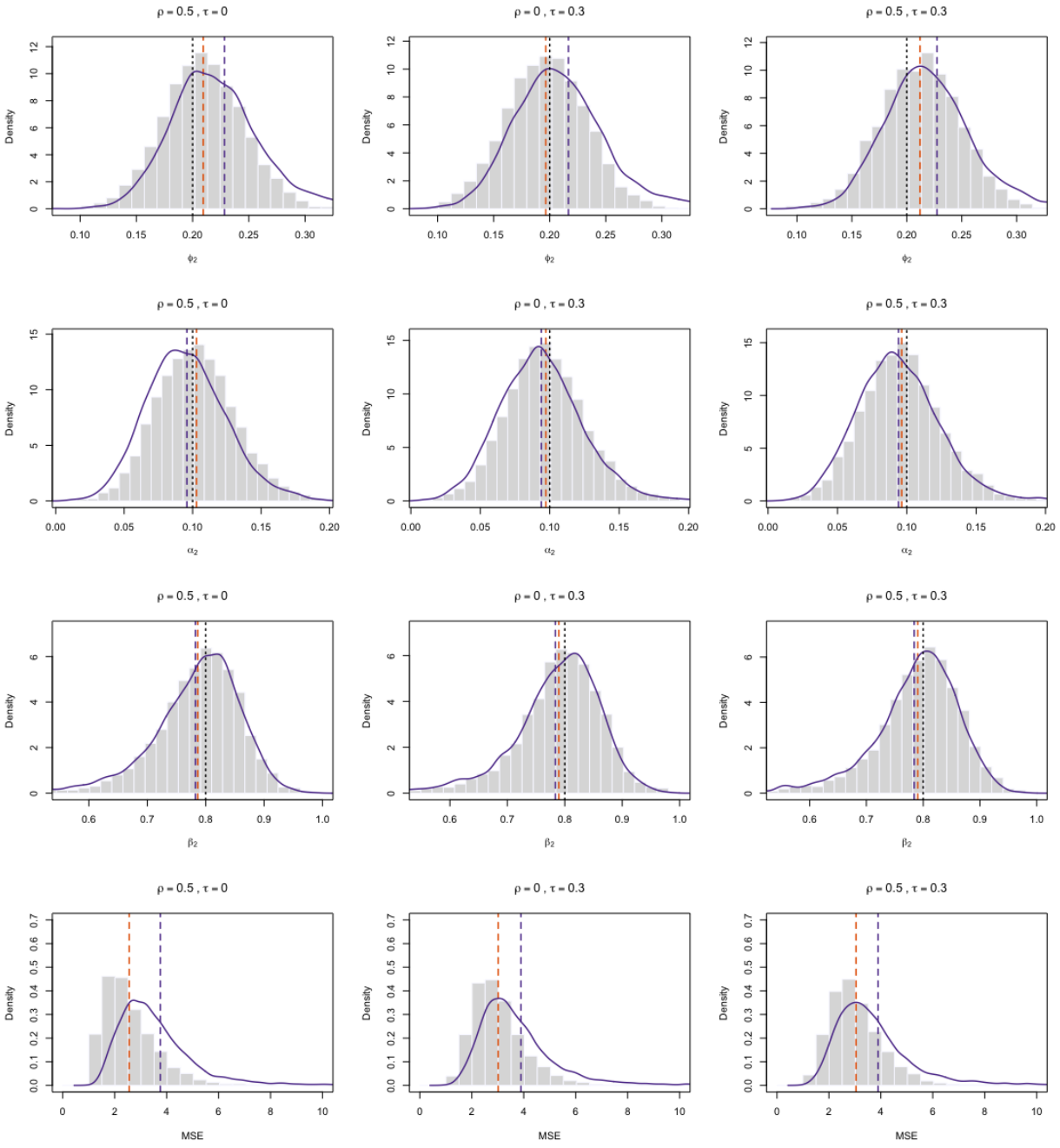

Figure 1.7: *Distribution of AR(1)-GARCH(1,1) parameters of the second factor (small ϕ), according to the two different models for $T = 1000$ and $n = 100$. Last row indicates MSE. CHDFM parameters distributions are the grey histogram, while 2SPCA kernel density is indicated by the purple line. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

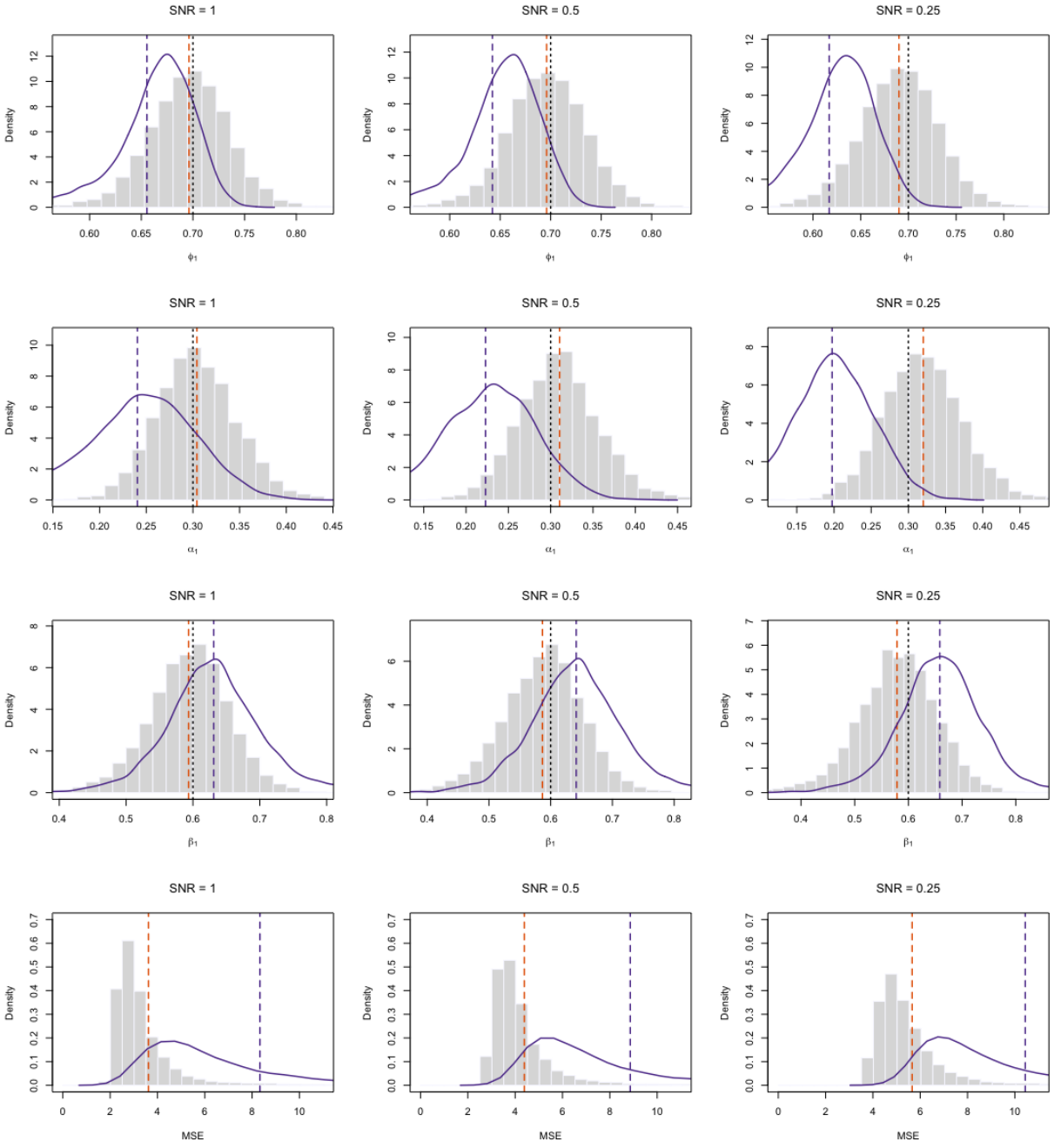

Figure 1.8: *Distributions of AR(1)-GARCH(1,1) parameters of the first factor (large ϕ), according to the two different models for $T = 1000$ and $n = 100$. Last row indicates MSE. CHDFM parameters distributions are the grey histograms, while 2SPCA parameters densities are indicated by the purple lines. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

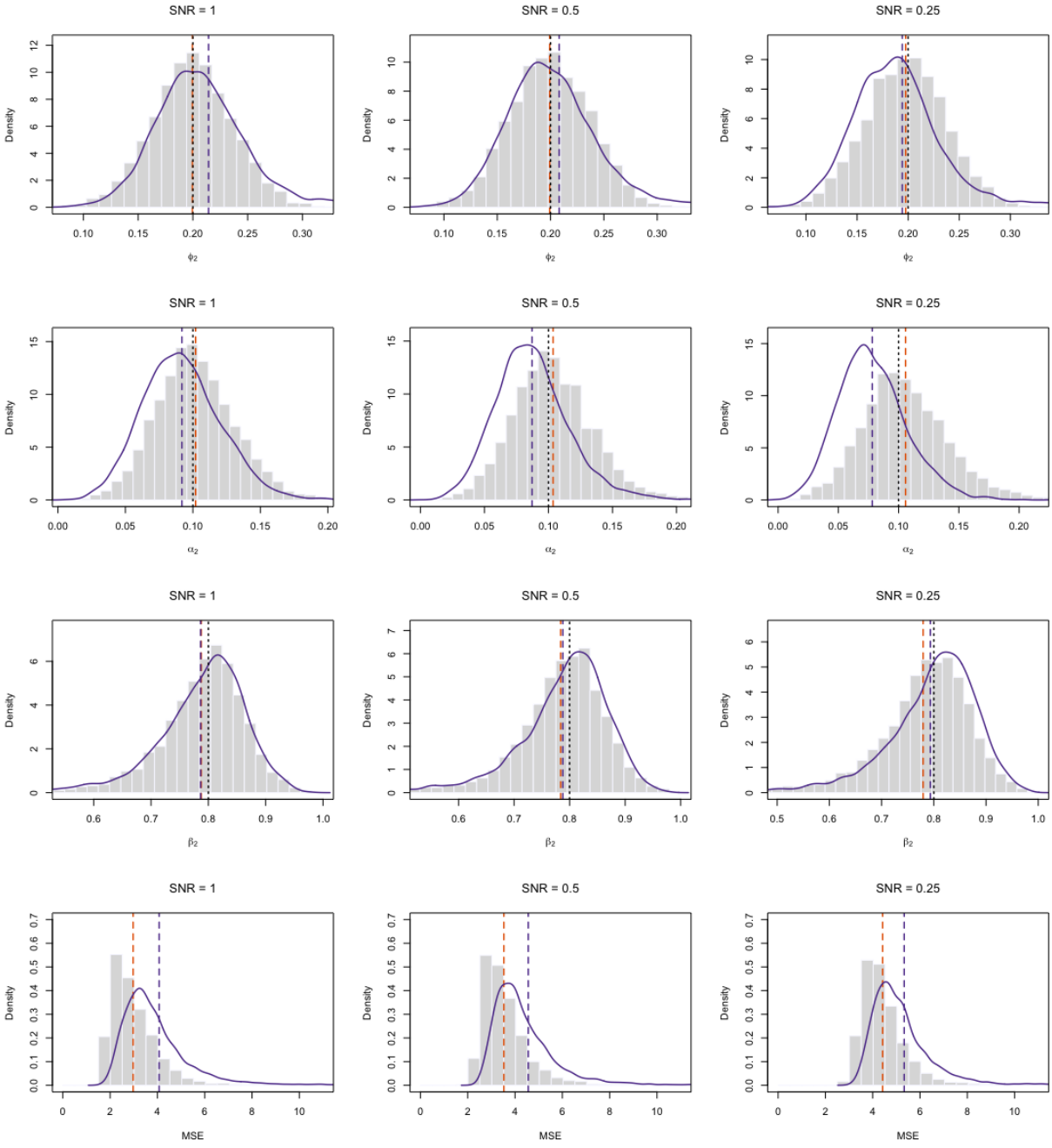

Figure 1.9: *Distributions of AR(1)-GARCH(1,1) parameters of the second factor (small ϕ), according to the two different models for $T = 1000$ and $n = 100$. Last row indicates MSE. CHDFM parameters distributions are the grey histogram, while 2SPCA kernel density is indicated by the purple line. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

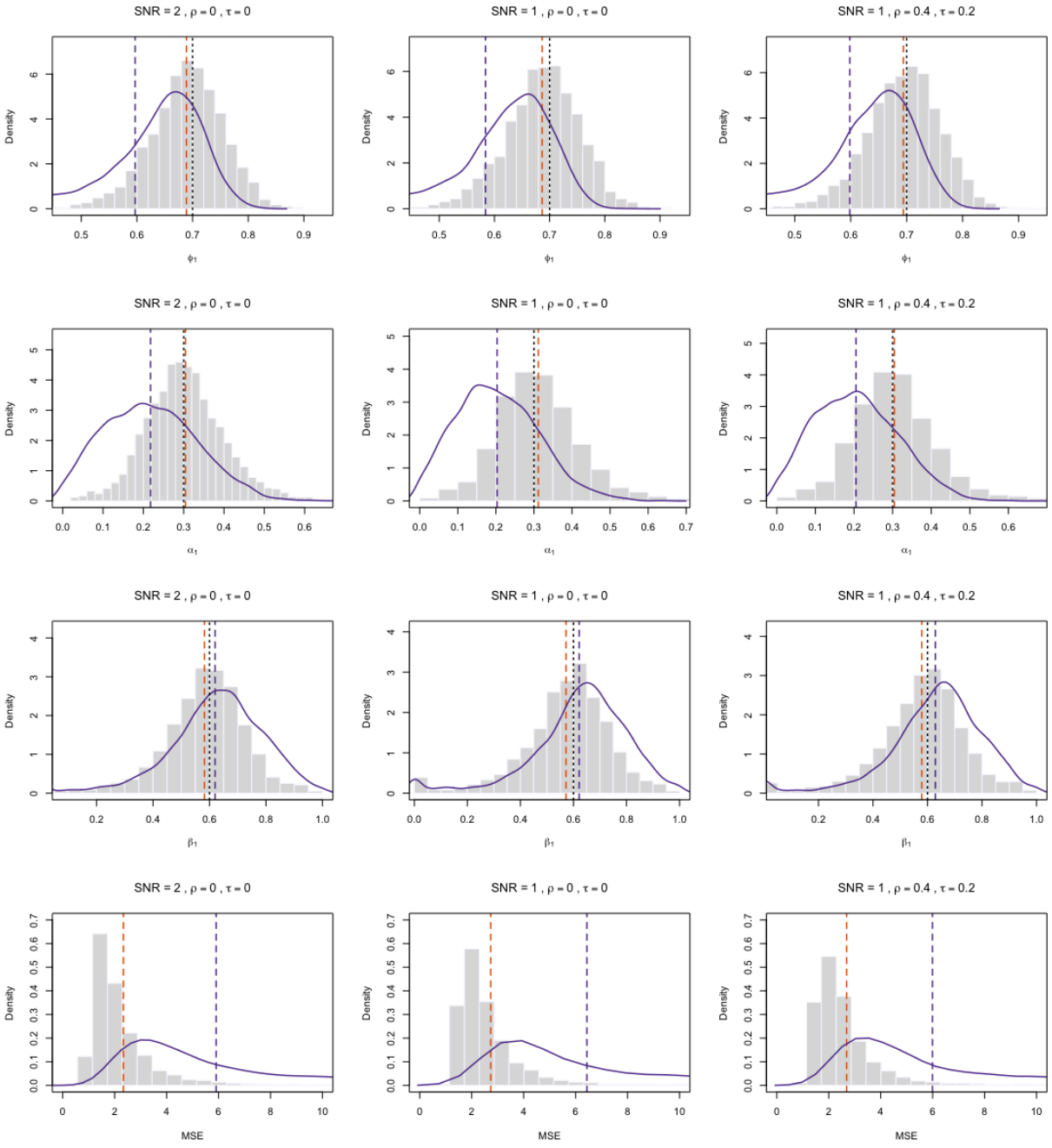

Figure 1.10: *Distributions of AR(1)-GARCH(1,1) parameters of the first factor (large ϕ), according to the two different models for $T = 250$ and $n = 50$. Last row indicates MSE. CHDFM parameters distributions are the grey histograms, while 2SPCA parameters densities are indicated by the purple lines. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

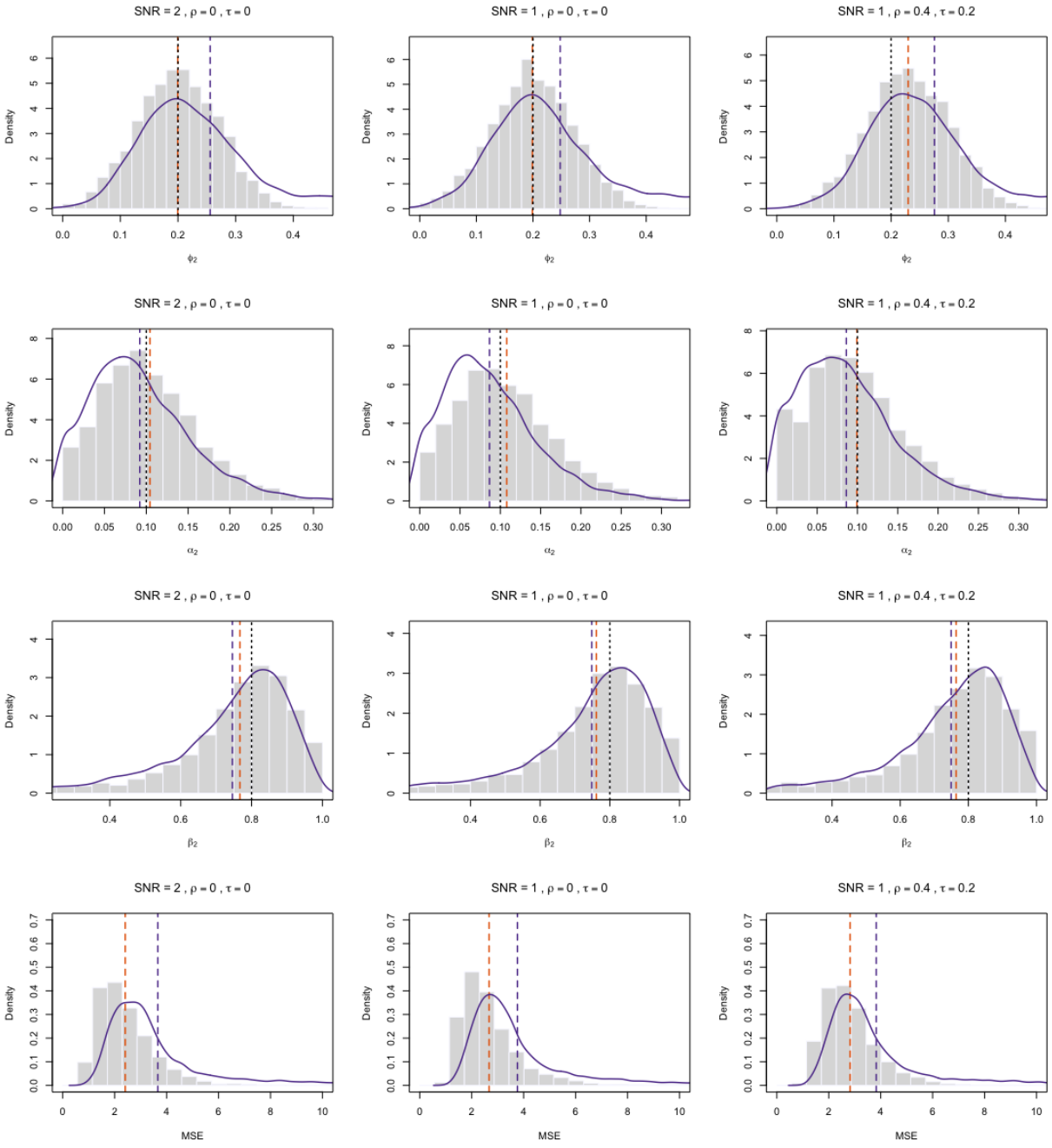

Figure 1.11: *Distributions of AR(1)-GARCH(1,1) parameters of the second factor (small ϕ), according to the two different models for $T = 250$ and $n = 50$. Last row indicates MSE. CHDFM parameters distributions are the grey histogram, while 2SPCA kernel density is indicated by the purple line. Orange (Purple) vertical bars indicate average parameter estimates for the CHDFM (2SPCA). Ground truth in dotted black.*

1.7 Appendix

1.7.1 Kalman Filter and Kalman Smoother

Let us assume that the true value of the parameter $\boldsymbol{\theta}$ is known and initial conditions $F_{0|0}$ and $P_{0|0}$ are given. For $s < T$ we have

$$\begin{aligned} \mathbf{F}_{t|s} &= \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{F}_t|\mathcal{X}_s] && (1.68) \\ \mathbf{P}_{t|s} &= \mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{F}_t - \mathbf{F}_{t|s})(\mathbf{F}_t - \mathbf{F}_{t|s})^{'}|\mathcal{X}_s] && (1.69) \end{aligned}$$

where $\mathcal{X}_s$ is the set of information that consists of $\boldsymbol{x}_s$ up to time s.

Kalman Filter - Forward Iterations

The Kalman filter is based on two sets of forward iterations. For $t = 1, \ldots, T$, the *predictions equations* are:

$$\begin{aligned} \mathbf{F}_{t|t-1} &= \boldsymbol{\Phi}\mathbf{F}_{t-1|t-1} && (1.70) \\ \mathbf{P}_{t|t-1} &= \boldsymbol{\Phi}\mathbf{P}_{t-1|t-1}\boldsymbol{\Phi}^{'} + \boldsymbol{\Psi}\mathbf{Q}_{t|t-1}\boldsymbol{\Psi}^{'}. && (1.71) \end{aligned}$$

When a new observation $\boldsymbol{x}_t$ become available the estimators for $\mathbf{F}_t$ are updated. The *updating equations* are:

$$\begin{aligned} \mathbf{F}_{t|t} &= \mathbf{F}_{t|t-1} + \mathbf{P}_{t|t-1}\boldsymbol{\Lambda}^{'}(\boldsymbol{\Lambda}\mathbf{P}_{t|t-1}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Gamma})^{-1}(\boldsymbol{x}_t - \boldsymbol{\Lambda}\mathbf{F}_{t|t-1}) && (1.72) \\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\boldsymbol{\Lambda}^{'}(\boldsymbol{\Lambda}\mathbf{P}_{t|t-1}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}\mathbf{P}_{t|t-1}. && (1.73) \end{aligned}$$

Moreover, by combining the two we obtain the recursion for the error covariance matrix

$$\mathbf{P}_{t+1|t} = \boldsymbol{\Phi}\mathbf{P}_{t|t-1}\boldsymbol{\Phi}^{'} - \boldsymbol{\Phi}\mathbf{P}_{t|t-1}\boldsymbol{\Lambda}^{'}(\boldsymbol{\Lambda}\mathbf{P}_{t|t-1}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}\mathbf{P}_{t|t-1}\boldsymbol{\Phi}^{'} + \boldsymbol{\Psi}\mathbf{Q}_{t+1|t}\boldsymbol{\Psi}^{'} \quad (1.74)$$

also known as *Riccati difference equation.*

Kalman Smoother - Backward Iterations

The Kalman smoother is based on the backward iterations for $t = T, \ldots, 1$:

$$\begin{aligned} \mathbf{F}_{t|T} &= \mathbf{F}_{t|t} + \mathbf{P}_{t|t}\boldsymbol{\Phi}^{'}\mathbf{P}_{t+1|t}^{-1}(\mathbf{F}_{t+1|T} - \mathbf{F}_{t+1|t}) && (1.75) \\ \mathbf{P}_{t|T} &= \mathbf{P}_{t|t} + \mathbf{P}_{t|t}\boldsymbol{\Phi}^{'}\mathbf{P}_{t+1|t}^{-1}(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{P}_{t+1|t}^{-1}\boldsymbol{\Phi}\mathbf{P}_{t|t} && (1.76) \end{aligned}$$

with starting point $\mathbf{F}_{T|T} = \mathbf{F}_{T|t=T}$ and $\mathbf{P}_{T|T} = \mathbf{P}_{T|t=T}$.

1.7.2 Proprieties of the Factor Covariance Matrix

Lemma 1 . *Given the model as defined in* (1.5) - (1.6)*, then:*

(i) $Var[\mathbf{F}_t] = \mathbf{I}_r + O(\epsilon)$

(ii) $Var[\boldsymbol{\eta}_t] = (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'})$

(iii) $Cov[\mathbf{F}_t, \mathbf{F}_{t+k}] = \boldsymbol{\Phi}^{|k|}$*, for any* $k \in \mathbb{Z}$

(iv) $Cov[\mathbf{F}_t, \boldsymbol{\eta}_{t+k}] = \boldsymbol{\Phi}^{|k|}(\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'})$*, for any* $k \in \mathbb{Z}^{+}$

(v) $Cov[\mathbf{F}_t, \boldsymbol{\eta}_{t+k}] = \mathbf{0}$*, for any* $k \in \mathbb{Z}^{-}$.

Proof. Straightforwardly from Section 1.5 and VAR(1) proprieties.

Lemma 2 . *Define* $\boldsymbol{F}_T = (\mathbf{F}_1^{'} \cdots \mathbf{F}_T^{'})^{'}$ *the* rT*-dimensional vector of unobserved factors with variance* $\boldsymbol{\Sigma}_F = \mathbb{E}[\boldsymbol{F}_T \boldsymbol{F}_T^{'}]$*. Under assumptions (A1) - (A5), the following proprieties hold:*

(i) $\|\boldsymbol{\Sigma}_F\| = O_p(1)$

(ii) $\|\boldsymbol{\Sigma}_F^{-1}\| = O_p(1)$.

Proof. The proof is based on Doz et al. (2012). Define by $\boldsymbol{S}_F(\omega)$ the spectral density matrix of $\boldsymbol{F}_t$, having autocovariance matrix $\boldsymbol{\Omega}(t-\tau) = \mathbb{E}[\mathbf{F}_t \mathbf{F}_\tau^{'}]$. Then,

$$\boldsymbol{S}_F(\omega) = \frac{1}{2\pi} \sum_{1 \leq t,\tau \leq T} \boldsymbol{\Gamma}(t-\tau) e^{-i\omega(t-\tau)}, \tag{1.77}$$

$$\boldsymbol{\Omega}(t-\tau) = \int_{-\pi}^{+\pi} \boldsymbol{S}_F(\omega) e^{-i\omega(t-\tau)} d\omega. \tag{1.78}$$

Let us denote $\underline{\mathbf{w}} = (\mathbf{w}_1^{'}, \mathbf{w}_2^{'}, ..., \mathbf{w}_T^{'}) \in \mathbb{R}^{rT}$ any eigenvector of $\boldsymbol{\Sigma}_F$ such that $\|\underline{\mathbf{w}}\|^2 = \sum_{t=1}^{T} \|\mathbf{w}_t\| = 1$ and λ^* a generic eigenvalues of $\boldsymbol{\Sigma}_F$. Thus, we can write

$$\lambda^* = \underline{\mathbf{w}}^{'} \boldsymbol{\Sigma}_F \underline{\mathbf{w}} = \sum_{t=1}^{T} \sum_{\tau=1}^{T} \mathbf{w}_t^{'} \boldsymbol{\Gamma}(t-\tau) \mathbf{w}_\tau. \tag{1.79}$$

Using the inverse transform of the spectral density we have that

$$\underline{\mathbf{w}}^{'} \boldsymbol{\Sigma}_F \underline{\mathbf{w}} = \sum_{1 \leq t,\tau \leq T} \mathbf{w}_t^{'} \left(\int_{-\pi}^{+\pi} \boldsymbol{S}_F(\omega) e^{-i\omega(t-\tau)} d\omega \right) \mathbf{w}_\tau \tag{1.80}$$

$$= \int_{-\pi}^{+\pi} \left(\sum_{1 \leq t,\tau \leq T} \mathbf{w}_t^{'} \boldsymbol{S}_F(\omega) \mathbf{w}_\tau e^{-i\omega(t-\tau)} \right) d\omega \tag{1.81}$$

$$= \int_{-\pi}^{+\pi} \left(\sum_{1 \leq t \leq T} \mathbf{w}_t^{'} e^{-i\omega t} \right) \boldsymbol{S}_F(\omega) \left(\sum_{1 \leq \tau \leq T} \mathbf{w}_\tau e^{i\omega\tau} \right) d\omega. \tag{1.82}$$

Now, indicate m and M as the minimum and maximum eigenvalue of $\boldsymbol{S}_F(\omega)$, that is

$$m = \min_{\omega\in[-\pi,+\pi]} \lambda_{\min}(\boldsymbol{S}_F(\omega)), \tag{1.83}$$

$$M = \max_{\omega\in[-\pi,+\pi]} \lambda_{\max}(\boldsymbol{S}_F(\omega)). \tag{1.84}$$

Hence,

$$\underline{\mathbf{w}}'\boldsymbol{\Sigma}_F\underline{\mathbf{w}} \in \left[m\int_{-\pi}^{+\pi} \left\| \sum_{1\le t\le T} \mathbf{w}_t' e^{-i\omega t} \right\|^2 d\omega \ , M\int_{-\pi}^{+\pi} \left\| \sum_{1\le t\le T} \mathbf{w}_t' e^{-i\omega t} \right\|^2 d\omega \right]. \tag{1.85}$$

But noticing that

$$\int_{-\pi}^{+\pi} \left\| \sum_{1\le t\le T} \mathbf{w}_t' e^{-i\omega t} \right\|^2 d\omega = \int_{-\pi}^{+\pi} \left(\sum_{1\le t,\tau\le T} \mathbf{w}_t' e^{-i\omega t}\mathbf{w}_\tau e^{i\omega\tau} \right) d\omega \tag{1.86}$$

$$= \sum_{1\le t,\tau\le T} \int_{-\pi}^{+\pi} \mathbf{w}_t'\mathbf{w}_\tau e^{-i\omega(t-\tau)} d\omega \tag{1.87}$$

$$= 2\pi \sum_{1\le t\le T} \mathbf{w}_t'\mathbf{w}_t = 2\pi, \tag{1.88}$$

we have shown that any eigenvalue of $\boldsymbol{\Sigma}_F$, $\lambda^* \in [2\pi m, 2\pi M]$. Finally, we have to prove that m and M are bounded. Starting from M, under assumption (A1) - (A1') we have that the process $\{\mathbf{F}_t\}$ is second-order stationary. Then $\mathbf{F}_t$ admits a Wold representation of the form: $\mathbf{F}_t = \mathcal{B}(L)\boldsymbol{\epsilon}_t = \sum_{j=1}^{\infty} \mathbf{B}^{(j)}\boldsymbol{\epsilon}_{t-j}$, with $\sum_{j=1}^{\infty}\|\mathbf{B}^{(j)}\| < \infty$ and $\boldsymbol{\epsilon}_t$ stationary at order four. Thus, for any $\omega \in [-\pi, +\pi]$ we have that

$$\|\boldsymbol{S}_F(\omega)\| = \frac{1}{2\pi} \left\| \sum_{h=-\infty}^{\infty} \boldsymbol{\Gamma}(h) e^{ih\omega} \right\| \le \sum_{h=-\infty}^{\infty} \|\boldsymbol{\Gamma}(h)\| < \infty. \tag{1.89}$$

Since $\lambda_{\max}(\boldsymbol{S}_F(\omega)) = \|\boldsymbol{S}_F(\omega)\|_2 \le \|\boldsymbol{S}_F(\omega)\| < \infty$ we have proven that $M < \infty$ and, hence, $\|\boldsymbol{\Sigma}_F\| = O_p(1)$.

For the (ii) part of the Lemma we need to show that $m > 0$. We start rewriting the process in (1.2) in the general VAR form $\mathcal{A}(L)\mathbf{F}_t = \boldsymbol{\eta}_t + \boldsymbol{\eta}_t^*$ where $\mathcal{A}(0) = \mathbf{I}_r$ and with $\mathcal{A}(z) \neq 0$ for $|z| \le 1$. As indicated in (1.3) and (1.4), the process $\{\boldsymbol{\eta}_t^*\}$ is *iid* while $\{\boldsymbol{\eta}_t\}$ is a GARCH(1,1) and, under assumptions (A1) and (A1'), $\boldsymbol{\eta}_t$ is a *weak white noise*.[6] Specifically, one has that

$$Cov[\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t+h}] = \mathbb{E}[\boldsymbol{\eta}_t\boldsymbol{\eta}_{t+h}'] = 0, \qquad \forall h \in \mathbb{Z}, h \neq 0. \tag{1.90}$$

Furthermore, $\{\boldsymbol{\eta}_t + \boldsymbol{\eta}_t^*\}$ is a *white noise* itself, meaning that $\boldsymbol{S}_{\eta+\eta^*}(\omega) = \boldsymbol{\Upsilon}/2\pi$. Now, taking a look at the spectral density of $\mathbf{F}_t$, using VAR representation,

$$\boldsymbol{S}_F(\omega) = \big(\mathcal{A}(e^{i\omega})\big)^{-1} \boldsymbol{S}_{\eta+\eta^*}(\omega) \big(\mathcal{A}'(e^{-i\omega})\big)^{-1}. \tag{1.91}$$

[6] A process $\{w_t\}$ is called white noise if, for some positive constant σ^2: (i) $\mathbb{E}[w_t] = 0\ \forall t \in \mathbb{Z}$; (ii) $\mathbb{E}[w_t] = \sigma^2 < \infty\ \forall t \in \mathbb{Z}$; (iii) $Cov(w_t, w_{t+h}) = 0\ \forall t, h \in \mathbb{Z}, h \neq 0$.

Then, for any eigenvector $\mathbf{w} \in \mathbb{C}^n$, such that $\|\mathbf{w}\| = 1$ and indicating with $\mathbf{w}^*$ the complex conjugate of $\mathbf{w}$, we can show that

$$\begin{aligned}
\mathbf{w}^{'} \boldsymbol{S}_F(\omega)\mathbf{w}^* &= \frac{1}{2\pi}\mathbf{w}^{'}\big(\mathcal{A}(e^{i\omega})\big)^{-1}\boldsymbol{\Upsilon}\big(\mathcal{A}^{'}(e^{-i\omega})\big)^{-1}\mathbf{w}^*. && (1.92)\\
&\geq \frac{1}{2\pi}\lambda_{\min}(\boldsymbol{\Upsilon})\mathbf{w}^{'}\|\big(\mathcal{A}(e^{i\omega})\big)^{-1}\big(\mathcal{A}^{'}(e^{-i\omega})\big)^{-1}\|\mathbf{w}^* && (1.93)\\
&\geq \frac{1}{2\pi}\lambda_{\min}(\boldsymbol{\Upsilon})\ \lambda_{\min}\big([\mathcal{A}^{'}(e^{-i\omega})\mathcal{A}(e^{i\omega})]^{-1}\big) && (1.94)\\
&= \frac{1}{2\pi}\frac{\lambda_{\min}(\boldsymbol{\Upsilon})}{\lambda_{\max}(\mathcal{A}^{'}(e^{-i\omega})\mathcal{A}(e^{i\omega}))} && (1.95)\\
&= \frac{1}{2\pi}\frac{\lambda_{\min}(\boldsymbol{\Upsilon})}{\|\mathcal{A}(e^{i\omega})\|_2^2}. && (1.96)
\end{aligned}$$

Denote $a = \min_{\omega\in[-\pi,+\pi]}\|\mathcal{A}(e^{i\omega})\|_2^2$. Knowing that a is *finite*, and $\boldsymbol{\Upsilon}$ is positive definite, we finally get

$$\lambda_{\min}(\boldsymbol{S}_F(\omega)) \geq \frac{1}{2\pi}\frac{\lambda_{\min}(\boldsymbol{\Upsilon})}{a}, \tag{1.97}$$

proving that $m > 0$ and, consequently, $\|\boldsymbol{\Sigma}_F^{-1}\| = O_p(1)$.

Lemma 3 . *Define* $\boldsymbol{F}_T^{\dagger} = (\mathbf{F}_1^{'}, \cdots, \mathbf{F}_T^{'}, \boldsymbol{\eta}_1^{'}, \cdots, \boldsymbol{\eta}_T^{'})^{'}$ *the* $2rT$*-dimensional vector of unobserved factors with variance* $\boldsymbol{\Sigma}_{F^\dagger} = \mathbb{E}[\boldsymbol{F}_T^{\dagger}\boldsymbol{F}_T^{\dagger'}]$.[7] *Under Assumptions (A1) - (A5) and Lemma 2 the following proprieties hold:*

(i) $\|\boldsymbol{\Sigma}_{F^\dagger}\| = O_p(1)$

(ii) $\|\boldsymbol{\Sigma}_{F^\dagger}^{-1}\| = O_p(1)$.

Proof. Let us start partitioning the $2rT \times 2rT$ matrix $\boldsymbol{\Sigma}_{F^\dagger}$ such that

$$\boldsymbol{\Sigma}_{F^\dagger} = \begin{bmatrix} \boldsymbol{\Sigma}_F & \boldsymbol{\Sigma}_{F,\eta} \\ \boldsymbol{\Sigma}_{F,\eta}^{'} & \boldsymbol{\Sigma}_\eta \end{bmatrix} \tag{1.98}$$

where $\boldsymbol{\Sigma}_F = \mathbb{E}[\boldsymbol{F}_T\boldsymbol{F}_T^{'}]$ is the matrix containing all autocovariances of $\mathbf{F}_t$ and such that $\boldsymbol{\Sigma}_F = \boldsymbol{\Sigma}_G + \mathbf{I}_T \otimes \mathbf{Q}^\star$, where $\mathbf{Q}^\star = \text{vec}^{-1}[(\mathbf{I}_{r^2} - \boldsymbol{\Phi}\otimes\boldsymbol{\Phi})^{-1}\text{vec}(\mathbf{Q}^\star)] = O(\epsilon)$. Then, $\boldsymbol{\Sigma}_\eta = \mathbf{I}_T \otimes \mathbf{Q}$, with $\mathbf{Q}$ being the unconditional variance of $\boldsymbol{\eta}_t$, given by $\mathbf{Q} = (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'})$. Finally, the matrix $\boldsymbol{\Sigma}_{F,\eta}$ is an upper triangular matrix containing all covariances of $\mathbf{F}_{t+k}$ and $\boldsymbol{\eta}_t$ for all positive

[7] For the proofs we are going to use this notation instead of $\boldsymbol{F}_T^{\dagger} = (\mathbf{F}_1^{\dagger'} \cdots \mathbf{F}_T^{\dagger'})^{'}$ because with the former specification it is possible to exploit block matrix proprieties.

values of the lag $k = 1, \ldots, T$.[8] By contruction, $\boldsymbol{\Sigma}_{\eta,F} = \boldsymbol{\Sigma}_{F,\eta}^{'}$ is lower triangular.
Taking a closer look to the submatrices we have

$$\boldsymbol{\Sigma}_F = \begin{bmatrix} \mathbf{I}_r & \boldsymbol{\Phi} & \ldots & \boldsymbol{\Phi}^{T-1} \\ \boldsymbol{\Phi} & \mathbf{I}_r & & \\ \vdots & & \ddots & \\ \boldsymbol{\Phi}^{T-1} & & & \mathbf{I}_r \end{bmatrix} + \begin{bmatrix} \mathbf{Q}^\star & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^\star & & \\ \vdots & & \ddots & \\ \mathbf{0} & & & \mathbf{Q}^\star \end{bmatrix} \tag{1.99}$$

$$\tag{1.100}$$

$$\boldsymbol{\Sigma}_\eta = \begin{bmatrix} (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}) & & \\ \vdots & & \ddots & \\ \mathbf{0} & & & (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}) \end{bmatrix}, \tag{1.101}$$

$$\tag{1.102}$$

$$\boldsymbol{\Sigma}_{F,\eta} = \begin{bmatrix} (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}) & \boldsymbol{\Phi}(\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}) & \ldots & \boldsymbol{\Phi}^{T-1}(\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}) \\ \mathbf{0} & (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}) & & \\ \vdots & & \ddots & \\ \mathbf{0} & & & (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}) \end{bmatrix}. \tag{1.103}$$

$$\tag{1.104}$$

To prove (i) we can use proprieties of psd block matrices (Mhanna, 2015) so that we have

$$\|\boldsymbol{\Sigma}_{F^\dagger}\| \leq \|\boldsymbol{\Sigma}_F\| + \|\boldsymbol{\Sigma}_\eta\|. \tag{1.105}$$

The first part is $O(1)$ as shown in Lemma 1 (i) while for the second it suffices to show that $\|\boldsymbol{\Sigma}_\eta\| = \|\mathbf{I}_T \otimes (\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'})\| = \|\mathbf{I}_r - \boldsymbol{\Phi}\boldsymbol{\Phi}^{'}\| = O(1)$ by Assumption (A1).
For the second part we can use block matrix inversion.[9] $\boldsymbol{\Sigma}_\eta$ is invertible as it is a diagonal matrix. Let us check, then, that the Schur complement $\mathbf{S} = \boldsymbol{\Sigma}_F - \boldsymbol{\Sigma}_{F,\eta}\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\Sigma}_{F,\eta}^{'}$ is invertible. One way of proving this relation is by means of the matrix extension of the Cauchy Schwarz inequality (Tripathi, 1999). As in Radhakrishna Rao (2000), let $\boldsymbol{F}_T \in \mathbb{R}^{rT}$ and $\boldsymbol{\eta}_T \in \mathbb{R}^{rT}$ be random vectors such that $\|\boldsymbol{\Sigma}_F\| = O(1)$ and $\|\boldsymbol{\Sigma}_\eta\| = O(1)$, then,

$$\boldsymbol{\Sigma}_F - \boldsymbol{\Sigma}_{F,\eta}\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\Sigma}_{F,\eta}^{'} \geq 0 \tag{1.107}$$

i.e. the difference is positive semi-definite. The inequality is sharp if all the mass of the distribution of $(\boldsymbol{F}_T, \boldsymbol{\eta}_T)$ lies on a proper linear subspace of $\mathbb{R}^{rT}$, as in the case when $\boldsymbol{F}_T, \boldsymbol{\eta}_T$

[8] It is upper triangular since $\text{Cov}(\boldsymbol{F}_t, \boldsymbol{\eta}_{t+k}) = 0$ for any value of $k = 1, \ldots, T$.

[9] Consider a matrix $M \in \mathbb{R}^{2n \times 2n}$ partitioned in four blocks $A, B, C, D \in \mathbb{R}^{n \times n}$. If D and the Schur complement $E = (A - BD^{-1}C)$ are invertible, then M can be inverted block-wise as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} E^{-1} & -E^{-1}BD^{-1} \\ -D^{-1}CE^{-1} & D^{-1} + D^{-1}CE^{-1}BD^{-1} \end{bmatrix} \tag{1.106}$$

.

are perfectly correlated. If one considers the model $\mathbf{G}_t = \mathbf{\Phi}\mathbf{G}_{t-1} + \boldsymbol{\eta}_t$ as in (1.2), then the relation of (1.107) becomes strict and the invertibility conditions of $\mathbf{S}$ are not met. On the other hand, $\mathbf{F}_t = \mathbf{\Phi}\mathbf{F}_{t-1} + \boldsymbol{\eta}_t + \boldsymbol{\eta}_t^*$ ensures that the matrix is invertible and furthermore we have:

$$\mathbf{\Sigma}_G - \mathbf{\Sigma}_{G,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{G,\eta}^{'} = \mathbf{0} \tag{1.108}$$

$$\mathbf{\Sigma}_G + \mathbf{I}_T \otimes \mathbf{Q}^\star - \mathbf{\Sigma}_{G,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{G,\eta}^{'} = \mathbf{I}_T \otimes \mathbf{Q}^\star \tag{1.109}$$

$$\mathbf{\Sigma}_F - \mathbf{\Sigma}_{F,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{F,\eta}^{'} = \mathbf{I}_T \otimes \mathbf{Q}^\star, \tag{1.110}$$

where we used the fact that $\mathbf{\Sigma}_{G,\eta} = \mathbf{\Sigma}_{F,\eta}$ by construction, since $\boldsymbol{\eta}_t^*$ is uncorrelated across time and with $\boldsymbol{\eta}_t$ for each value of t. The relation also holds when interchanging $\mathbf{F}_t$ and $\boldsymbol{\eta}_t$. In this case we have:

$$\mathbf{\Sigma}_\eta - \mathbf{\Sigma}_{F,\eta}^{'}\mathbf{\Sigma}_F^{-1}\mathbf{\Sigma}_{F,\eta} \geq 0. \tag{1.111}$$

This will play a central role in Kalman smoother consistency. Moreover, using the relation $(A+B)^{-1} = A^{-1} - (A+B)^{-1}BA^{-1}$ this can be seen to be

$$\mathbf{\Sigma}_\eta - \mathbf{\Sigma}_{F,\eta}^{'}(\mathbf{\Sigma}_G + \mathbf{I}_T \otimes \mathbf{Q}^\star)^{-1}\mathbf{\Sigma}_{F,\eta} \tag{1.112}$$

$$= \mathbf{\Sigma}_\eta - \mathbf{\Sigma}_{F,\eta}^{'}\mathbf{\Sigma}_G^{-1}\mathbf{\Sigma}_{F,\eta} + \mathbf{\Sigma}_{F,\eta}^{'}(\mathbf{\Sigma}_G + \mathbf{I}_T \otimes \mathbf{Q}^\star)^{-1}\mathbf{I}_T \otimes \mathbf{Q}^\star\mathbf{\Sigma}_G^{-1}\mathbf{\Sigma}_{F,\eta} \tag{1.113}$$

$$= \mathbf{\Sigma}_{F,\eta}^{'}\mathbf{\Sigma}_F^{-1}\mathbf{I}_T \otimes \mathbf{Q}^\star\mathbf{\Sigma}_G^{-1}\mathbf{\Sigma}_{F,\eta}. \tag{1.114}$$

If we were to solve it analytically, without loss of generality, let us assume that $\mathbf{\Phi}$ is symmetric so that $(\mathbf{I}_r - \mathbf{\Phi}\mathbf{\Phi}^{'}) = (\mathbf{I}_r - \mathbf{\Phi}^2)$ and indicate by $\boldsymbol{A}_{[i,j]}$ the $r \times r$ dimensional block of a matrix $\boldsymbol{A}$.[10] Then,

$$(\mathbf{\Sigma}_{F,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{F,\eta}^{'})_{[1,1]} = (\mathbf{I}_r - \mathbf{\Phi}^2) + \mathbf{\Phi}(\mathbf{I}_r - \mathbf{\Phi}^2)\mathbf{\Phi} + \cdots + \mathbf{\Phi}^{T-1}(\mathbf{I}_r - \mathbf{\Phi}^2)\mathbf{\Phi}^{T-1} \tag{1.115}$$

$$= (\mathbf{I}_r - \mathbf{\Phi}^2) + (\mathbf{I}_r - \mathbf{\Phi}^2)\mathbf{\Phi}^2 + \cdots + (\mathbf{I}_r - \mathbf{\Phi}^2)\mathbf{\Phi}^{2(T-1)} \tag{1.116}$$

$$= (\mathbf{I}_r - \mathbf{\Phi}^2)(\mathbf{I}_r + \mathbf{\Phi}^2 + \cdots + \mathbf{\Phi}^{2(T-1)}) \tag{1.117}$$

$$= (\mathbf{I}_r - \mathbf{\Phi}^{2T}) \tag{1.118}$$

$$(\mathbf{\Sigma}_{F,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{F,\eta}^{'})_{[1,2]} = \mathbf{\Phi}(\mathbf{I}_r - \mathbf{\Phi}^2) + \mathbf{\Phi}^3(\mathbf{I}_r - \mathbf{\Phi}^2)\cdots + \mathbf{\Phi}^{2T-1}(\mathbf{I}_r - \mathbf{\Phi}^2) \tag{1.119}$$

$$= \mathbf{\Phi}(\mathbf{I}_r - \mathbf{\Phi}^2)(\mathbf{I}_r + \mathbf{\Phi}^2 + \cdots + \mathbf{\Phi}^{2(T-2)}) \tag{1.120}$$

$$= \mathbf{\Phi}(\mathbf{I}_r - \mathbf{\Phi}^{2(T-1)}) \tag{1.121}$$

$$\ldots \tag{1.122}$$

$$(\mathbf{\Sigma}_{F,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{F,\eta}^{'})_{[1,T]} = \mathbf{\Phi}^{T-1}(\mathbf{I}_r - \mathbf{\Phi}^2) \tag{1.123}$$

$$\ldots \tag{1.124}$$

$$(\mathbf{\Sigma}_{F,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{F,\eta}^{'})_{[T,T]} = (\mathbf{I}_r - \mathbf{\Phi}^2). \tag{1.125}$$

where we used the partial sum of the Neumann series $(I-A)(I+A+A^2+\cdots A^n) = (I-A^{n+1})$. This can be generalized to:

$$(\mathbf{\Sigma}_{F,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{F,\eta}^{'})_{[i,j]} = \mathbf{\Phi}^{|i-j|}(\mathbf{I}_r - \mathbf{\Phi}^{2(T+1-\max(i,j))}), \tag{1.126}$$

[10] Identifying restrictions (IC1) imposes that $\mathbf{\Phi}$ is diagonal.

and then we obtain that the blocks of $\mathbf{S} = \mathbf{\Sigma}_F - (\mathbf{\Sigma}_{F,\eta}\mathbf{\Sigma}_\eta^{-1}\mathbf{\Sigma}_{F,\eta}^{'})$ are given by

$$\begin{aligned}
\mathbf{S}_{[i,j]} &= \mathbf{\Phi}^{|i-j|} - \mathbf{\Phi}^{|i-j|}(\mathbf{I}_r - \mathbf{\Phi}^{2(T+1-\max(i,j))}) + \mathbb{I}_{i=j}\mathbf{Q}^* && (1.127)\\
&= \mathbf{\Phi}^{|i-j|}\mathbf{\Phi}^{2(T+1-\max(i,j))} + \mathbb{I}_{i=j}\mathbf{Q}^* && (1.128)
\end{aligned}$$

where $\mathbb{I}_{i=j}$ is an indicator matrix which is equal to $\mathbf{I}_r$ when $i = j$ and $\mathbf{0}_{r\times r}$ otherwise. When $T \to \infty$, we have that $\mathbf{S} \to \mathbf{I}_T \otimes \mathbf{Q}^\star$, this guarantees the matrix is invertible even for large T.
Finally, we can rewrite the block-wise inverse as:

$$\begin{bmatrix} \mathbf{\Sigma}_F & \mathbf{\Sigma}_{F,\eta} \\ \mathbf{\Sigma}_{F,\eta}^{'} & \mathbf{\Sigma}_\eta \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}^{-1} & -\mathbf{S}^{-1}\mathbf{\Sigma}_{F,\eta}\mathbf{I}_T \otimes \mathbf{Q}^{-1} \\ \mathbf{I}_T \otimes \mathbf{Q}^{-1}\mathbf{\Sigma}_{F,\eta}^{'}\mathbf{S}^{-1} & \mathbf{I}_T \otimes \mathbf{Q}^{-1} + \mathbf{I}_T \otimes \mathbf{Q}^{-1}\mathbf{\Sigma}_{F,\eta}^{'}\mathbf{S}^{-1}\mathbf{\Sigma}_{F,\eta}\mathbf{I}_T \otimes \mathbf{Q}^{-1} \end{bmatrix} \tag{1.129}$$

Then we can use norm inequalities to find an upper bound.[11] In particular, we have

$$\begin{aligned}
\|\mathbf{\Sigma}_{F^\dagger}^{-1}\| &\leq \|\mathbf{S}^{-1}\| + \|\mathbf{I}_T \otimes \mathbf{Q}^{-1} + \mathbf{I}_T \otimes \mathbf{Q}^{-1}\mathbf{\Sigma}_{F,\eta}^{'}\mathbf{S}^{-1}\mathbf{\Sigma}_{F,\eta}\mathbf{I}_T \otimes \mathbf{Q}^{-1}\| && (1.130)\\
&\leq \|\mathbf{I}_T \otimes \mathbf{Q}^{\star-1}\| + \|\mathbf{I}_T \otimes \mathbf{Q}^{-1}\| + \|\mathbf{I}_T \otimes \mathbf{Q}^{\star-1}\|\|\mathbf{I}_T \otimes \mathbf{Q}^{-1}\|^2\|\mathbf{\Sigma}_{F,\eta}\|^2 && (1.131)\\
&= \|\mathbf{Q}^{\star-1}\| + \|\mathbf{Q}^{-1}\| + \|\mathbf{Q}^{\star-1}\|\|\mathbf{Q}^{-1}\|^2\|\mathbf{\Sigma}_{F,\eta}\|^2 && (1.132)\\
&= O(1) && (1.133)
\end{aligned}$$

since $\|\mathbf{Q}^{\star-1}\| = \lambda_{min}(\mathbf{Q}^\star)^{-1}$, which is *finite* by construction, and the other matrices depend on $\mathbf{\Phi}$ and we know that $\|\mathbf{\Phi}\| \leq 1$ by Assumption (A1).

1.7.3 Kalman Smoother Consistency

Denote $\boldsymbol{X}_T = (\boldsymbol{x}_1^{'} \cdots \boldsymbol{x}_T^{'})^{'}$ and $\boldsymbol{Z}_T = (\boldsymbol{\xi}_1^{'} \cdots \boldsymbol{\xi}_T^{'})^{'}$ as the nT-dimensional vectors of observed values and disturbances with corresponding variances $\mathbf{\Sigma}_X = \mathbb{E}[\boldsymbol{X}_T\boldsymbol{X}_T^{'}]$ and $\mathbf{\Sigma}_Z = \mathbb{E}[\boldsymbol{Z}_T\boldsymbol{Z}_T^{'}] = \mathbf{I}_T \otimes \mathbf{\Gamma}$. Let $\boldsymbol{L}_T^\dagger = (\mathbf{I}_T \otimes \mathbf{\Lambda}, \mathbf{I}_T \otimes \mathbf{0}_{n\times r})$, the $nT \times 2rT$ matrix of factor loadings, and $\boldsymbol{F}_T^\dagger = (\mathbf{F}_1^{'}, \cdots, \mathbf{F}_T^{'}, \boldsymbol{\eta}_1^{'}, \cdots, \boldsymbol{\eta}_T^{'})^{'}$, the $2rT$-dimensional vector of unobserved factors with variance $\mathbf{\Sigma}_{F^\dagger} = \mathbb{E}[\boldsymbol{F}_t^\dagger \boldsymbol{F}_T^{\dagger'}]$. Let us rewrite the model in (1.1) as:

$$\boldsymbol{X}_T = \boldsymbol{L}_T^\dagger \boldsymbol{F}_T^\dagger + \boldsymbol{Z}_T, \qquad \mathbf{\Sigma}_X = \boldsymbol{L}_T^\dagger \mathbf{\Sigma}_{F^\dagger} \boldsymbol{L}_T^{\dagger'} + \mathbf{\Sigma}_Z. \tag{1.134}$$

Given the true values of the parameters, the optimal predictor of the factors as a linear combination of the observables $(\boldsymbol{x}_1^{'} \cdots \boldsymbol{x}_T^{'})^{'}$ is its projection $\mathbf{F}_{t|T}^\dagger = Proj[\mathbf{F}_t^\dagger | \boldsymbol{X}_T]$. Under Gaussianity, the best linear projection is given by the expected value. Nonetheless, we will allow some degrees of flexibility in the covariance matrix of $\boldsymbol{\xi}_t$, showing that we can achieve consistency of the factors even though the true matrix $\mathbf{\Gamma}$ is non-diagonal and the

[11] Using the result from (Hayashi, 2018), consider a matrix $M \in \mathbb{R}^{2n\times 2n}$ partitioned in four blocks $A, B, C, D \in \mathbb{R}^{n\times n}$. By triangle inequality, $\|M\| \leq 2\|A + D\|$. If M is positive semi-definite then $\|M\| \leq \|A\| + \|D\|$. If M is upper triangular then $\|M\|^2 \leq \|A\|^2 + \|B\|^2 + \|D\|^2$. If M is lower triangular then $\|M\|^2 \leq \|A\|^2 + \|C\|^2 + \|D\|^2$.

the idiosyncratic components are autocorrelated. Denote $\mathbf{\Gamma}_0 = \text{diag}(\sigma_{\xi,11} \cdots \sigma_{\xi,nn})$ and $\mathbf{\Sigma}_{0Z} = \mathbf{I}_T \otimes \mathbf{\Gamma}_0$, thus we have $Proj[\mathbf{F}_t^\dagger | \boldsymbol{X}_T] = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{F}_t^\dagger | \boldsymbol{X}_T]$ and this is given by:

$$\begin{aligned}
\mathbf{F}_{t|T}^\dagger &= \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{F}_t^\dagger \boldsymbol{X}_T^{'}]\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{X}_T \boldsymbol{X}_T^{'}]^{-1}\boldsymbol{X}_T && (1.135)\\
&= \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{F}_t^\dagger(\boldsymbol{F}_T^{\dagger'}\boldsymbol{L}_T^{\dagger'} + \boldsymbol{Z}_T^{'})]\mathbf{\Sigma}_X^{-1}\boldsymbol{X}_T && (1.136)\\
&= \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{F}_t^\dagger \boldsymbol{F}_T^{\dagger'}]\boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_X^{-1}\boldsymbol{X}_T && (1.137)\\
&= (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{\Sigma}_{F^\dagger}\boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_X^{-1}\boldsymbol{X}_T, && (1.138)
\end{aligned}$$

where $(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)$ is the $2r \times 2rT$matrix with identity $\mathbf{I}_r$ at time t. It selects the appropriate time block t of a matrix of dimension $2rT$. Specifically, $\boldsymbol{\iota}_{2t}^{'}$ is given by

$$\boldsymbol{\iota}_{2t}^{'} = \begin{bmatrix} \boldsymbol{\iota}_t^{'} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota}_t^{'} \end{bmatrix} \tag{1.139}$$

with $\boldsymbol{\iota}_t$ being the t-th column of the identity matrix $\mathbf{I}_T$. Using Woodbury identity[12] we can rewrite the inverse of the variance of $\boldsymbol{X}_T$ as

$$\mathbf{\Sigma}_X^{-1} = \mathbf{\Sigma}_{0Z}^{-1} - \mathbf{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^\dagger(\mathbf{\Sigma}_{F^\dagger}^{-1} + \boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^\dagger)^{-1}\boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}, \tag{1.141}$$

and pre-multiplying by $\boldsymbol{L}_T^{\dagger'}$ we obtain

$$\begin{aligned}
\boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_X^{-1} &= \boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1} - \boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^\dagger(\mathbf{\Sigma}_{F^\dagger}^{-1} + \boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^\dagger)^{-1}\boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1} && (1.142)\\
&= \mathbf{\Sigma}_{F^\dagger}^{-1}(\mathbf{\Sigma}_{F^\dagger}^{-1} + \boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^\dagger)^{-1}\boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}, && (1.143)
\end{aligned}$$

where we used the relation $A - B(C+B)^{-1}A = C(C+B)^{-1}A$. Let us call the matrix $\mathbf{M}^\dagger = \boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^\dagger$ and replace what we obtained in Expression (1.135)

$$\mathbf{F}_{t|T}^\dagger = (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)(\mathbf{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}\boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}\boldsymbol{X}_T. \tag{1.144}$$

Now, we can use the relation in (1.134) to decompose the above equation

$$\begin{aligned}
\mathbf{F}_{t|T}^\dagger &= (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)(\mathbf{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}\mathbf{M}^\dagger\boldsymbol{F}_T^\dagger + \\
&+ (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)(\mathbf{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}\boldsymbol{L}_T^{\dagger'}\mathbf{\Sigma}_{0Z}^{-1}\boldsymbol{Z}_T. && (1.145)
\end{aligned}$$

Denote by $\mathbf{F}_{1,t|T}^\dagger$ the first term of the previous expression. We have

$$\begin{aligned}
\mathbf{F}_{1,t|T}^\dagger &= (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)(\mathbf{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}\mathbf{M}^\dagger\boldsymbol{F}_T^\dagger && (1.146)\\
&= (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)(\mathbf{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}(\mathbf{M}^\dagger + \mathbf{\Sigma}_{F^\dagger}^{-1} - \mathbf{\Sigma}_{F^\dagger}^{-1})\boldsymbol{F}_T^\dagger && (1.147)\\
&= (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\boldsymbol{F}_T^\dagger - (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)(\mathbf{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}\mathbf{\Sigma}_{F^\dagger}^{-1}\boldsymbol{F}_T^\dagger && (1.148)\\
&= \mathbf{F}_t^\dagger - (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)(\mathbf{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}\mathbf{\Sigma}_{F^\dagger}^{-1}\boldsymbol{F}_T^\dagger. && (1.149)
\end{aligned}$$

[12]For any invertible square A, C $\in \mathbb{R}^{n\times n}$ and rectangular matrices B and $D^{'} \in \mathbb{R}^{m\times n}$, we have

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{1.140}$$

.

To prove consistency, the second term in the addition should converge to 0. Let us recall that for the two norms $\|A\|_F = \mathrm{tr}(A'A)^{1/2} = \mathrm{tr}(AA')^{1/2}$ and $\|A\| = \lambda_{\max}(A'A)^{1/2}$ the following relation holds: $\|A\| \leq \|A\|_F \leq \sqrt{r}\|A\|$, with $r = rank(A)$. Equality holds trivially for vectors or,if and only if, the matrix A is a rank-one matrix or a zero matrix since the trace of a matrix is equal to the sum of its eigenvalues.[13] Denote $\mathbf{K} = (\mathbf{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}$, then, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}\big[\|\mathbf{F}_t^\dagger - \mathbf{F}_{1,t|T}^\dagger\|^2\big] &= \mathbb{E}_{\boldsymbol{\theta}}\big[\mathrm{tr}\big\{(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{K}\mathbf{\Sigma}_{F^\dagger}^{-1}\boldsymbol{F}_T^\dagger\boldsymbol{F}_T^{\dagger'}\mathbf{\Sigma}_{F^\dagger}^{-1}\mathbf{K}^{'}(\boldsymbol{\iota}_{2t} \otimes \mathbf{I}_r)\big\}\big] && (1.150)\\
&= \mathrm{tr}\big\{(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{K}\mathbf{\Sigma}_{F^\dagger}^{-1}\mathbf{\Sigma}_{F^\dagger}\mathbf{\Sigma}_{F^\dagger}^{-1}\mathbf{K}^{'}(\boldsymbol{\iota}_{2t} \otimes \mathbf{I}_r)\big\} && (1.151)\\
&= \|(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{K}\mathbf{\Sigma}_{F^\dagger}^{-1/2}\|^2. && (1.152)\\
&\leq \|(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\|^2\|\mathbf{K}\|^2\|\mathbf{\Sigma}_{F^\dagger}\| && (1.153)\\
&= \|\mathbf{K}\|^2\|\mathbf{\Sigma}_{F^\dagger}\| && (1.154)
\end{aligned}$$

since $\|(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\|^2 = 1$. Furthermore we know from Lemma 2 that $\|\mathbf{\Sigma}_{F^\dagger}^{-1}\| = O(1)$, so let us temporarily focus on the first term. Specifically, $\mathbf{K}$ can be decomposed into block matrices such that

$$\mathbf{K} = \left(\begin{bmatrix}\mathbf{\Sigma}_F & \mathbf{\Sigma}_{F,\eta}\\ \mathbf{\Sigma}_{F,\eta}^{'} & \mathbf{\Sigma}_\eta\end{bmatrix}^{-1} + \begin{bmatrix}\mathbf{M} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)^{-1} \quad (1.155)$$

where $\mathbf{M} = \mathbf{I}_T \otimes \mathbf{\Lambda}^{'}\mathbf{\Gamma}_0^{-1}\mathbf{\Lambda}$. Then, we can use the relation $(A+B)^{-1} = (I + A^{-1}B)^{-1}A^{-1}$ which holds for any invertible A to get

$$\begin{aligned}
\mathbf{K} &= \left(\begin{bmatrix}\mathbf{I}_{rT} & \mathbf{0}_{rT}\\ \mathbf{0}_{rT} & \mathbf{I}_{rT}\end{bmatrix} + \begin{bmatrix}\mathbf{\Sigma}_F & \mathbf{\Sigma}_{F,\eta}\\ \mathbf{\Sigma}_{F,\eta}^{'} & \mathbf{\Sigma}_\eta\end{bmatrix}\begin{bmatrix}\mathbf{M} & \mathbf{0}_{rT}\\ \mathbf{0}_{rT} & \mathbf{0}_{rT}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{\Sigma}_F & \mathbf{\Sigma}_{F,\eta}\\ \mathbf{\Sigma}_{F,\eta}^{'} & \mathbf{\Sigma}_\eta\end{bmatrix} && (1.156)\\
&= \left(\begin{bmatrix}\mathbf{I}_{rT} & \mathbf{0}_{rT}\\ \mathbf{0}_{rT} & \mathbf{I}_{rT}\end{bmatrix} + \begin{bmatrix}\mathbf{\Sigma}_F\mathbf{M} & \mathbf{0}_{rT}\\ \mathbf{\Sigma}_{F,\eta}^{'}\mathbf{M} & \mathbf{0}_{rT}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{\Sigma}_F & \mathbf{\Sigma}_{F,\eta}\\ \mathbf{\Sigma}_{F,\eta}^{'} & \mathbf{\Sigma}_\eta\end{bmatrix} && (1.157)\\
&= \begin{bmatrix}\mathbf{I}_{rT} + \mathbf{\Sigma}_F\mathbf{M} & \mathbf{0}_{rT}\\ \mathbf{\Sigma}_{F,\eta}^{'}\mathbf{M} & \mathbf{I}_{rT}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{\Sigma}_F & \mathbf{\Sigma}_{F,\eta}\\ \mathbf{\Sigma}_{F,\eta}^{'} & \mathbf{\Sigma}_\eta\end{bmatrix} && (1.158)\\
&= \begin{bmatrix}(\mathbf{I}_{rT} + \mathbf{\Sigma}_F\mathbf{M})^{-1} & \mathbf{0}_{rT}\\ -\mathbf{\Sigma}_{F,\eta}^{'}\mathbf{M}(\mathbf{I}_{rT} + \mathbf{\Sigma}_F\mathbf{M})^{-1} & \mathbf{I}_{rT}\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_F & \mathbf{\Sigma}_{F,\eta}\\ \mathbf{\Sigma}_{F,\eta}^{'} & \mathbf{\Sigma}_\eta\end{bmatrix} && (1.159)\\
&= \begin{bmatrix}(\mathbf{\Sigma}_F^{-1} + \mathbf{M})^{-1}\mathbf{\Sigma}_F^{-1} & \mathbf{0}_{rT}\\ -\mathbf{\Sigma}_{F,\eta}^{'}\mathbf{M}(\mathbf{\Sigma}_F^{-1} + \mathbf{M})^{-1}\mathbf{\Sigma}_F^{-1} & \mathbf{I}_{rT}\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_F & \mathbf{\Sigma}_{F,\eta}\\ \mathbf{\Sigma}_{F,\eta}^{'} & \mathbf{\Sigma}_\eta\end{bmatrix} && (1.160)\\
&= \begin{bmatrix}(\mathbf{\Sigma}_F^{-1} + \mathbf{M})^{-1} & (\mathbf{\Sigma}_F^{-1} + \mathbf{M})^{-1}\mathbf{\Sigma}_F^{-1}\mathbf{\Sigma}_{F,\eta}^{'}\\ \mathbf{\Sigma}_{F,\eta}^{'} - \mathbf{\Sigma}_{F,\eta}^{'}\mathbf{M}(\mathbf{\Sigma}_F^{-1} + \mathbf{M})^{-1} & \mathbf{\Sigma}_\eta - \mathbf{\Sigma}_{F,\eta}^{'}\mathbf{M}(\mathbf{\Sigma}_F^{-1} + \mathbf{M})^{-1}\mathbf{\Sigma}_F^{-1}\mathbf{\Sigma}_{F,\eta}\end{bmatrix}. && (1.161)
\end{aligned}$$

We can use block matrix norm inequalities to find that

$$\begin{aligned}
\|\mathbf{K}\|^2 &\leq \|(\mathbf{\Sigma}_F^{-1} + \mathbf{M})^{-1}\|^2 + && (1.162)\\
&\quad \|(\mathbf{\Sigma}_\eta - \mathbf{\Sigma}_{F,\eta}^{'}\mathbf{M}(\mathbf{\Sigma}_F^{-1} + \mathbf{M})^{-1}\mathbf{\Sigma}_F^{-1}\mathbf{\Sigma}_{F,\eta})\|^2. && (1.163)
\end{aligned}$$

[13] A matrix in $\mathbb{R}^{m\times n}$ has rank one if, and only if, it can be written as the outer product of two nonzero vectors in $\mathbb{R}^{m\times n}$. Given a vector $\boldsymbol{u} \in \mathbb{R}^n$, then $\boldsymbol{u}\boldsymbol{u}^{'}$ is a rank-one matrix such that its 2-norm and F-norm are equivalent.

Now, given that $\boldsymbol{\Sigma}_F^{-1}$ is positive definite, we have $\boldsymbol{\Sigma}_F^{-1}+\mathbf{M} \geq \mathbf{M}$ and $(\boldsymbol{\Sigma}_F^{-1}+\mathbf{M})^{-1} \leq \mathbf{M}^{-1}$. Then we can use the fact that $\|A \otimes B\| = \|A\|\|B\|$ and $\|\mathbf{I}_T\| = 1$ to obtain

$$\begin{aligned} \|(\boldsymbol{\Sigma}_F^{-1}+\mathbf{M})^{-1}\|^2 &\leq \|\mathbf{M}^{-1}\|^2 && (1.164)\\ &\leq \|(\mathbf{I}_T \otimes \boldsymbol{\Lambda}^{'}\boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Lambda})^{-1}\|^2 && (1.165)\\ &\leq \|(\boldsymbol{\Lambda}^{'}\boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Lambda})^{-1}\|^2, && (1.166)\end{aligned}$$

which converge to 0 with rate n^{-2} as $n \to \infty$ by Assumptions (A3) and (A5). For the second part, instead, we first use the relation $(A+B)^{-1} = A^{-1} - (A+B)^{-1}BA^{-1}$ to obtain

$$\begin{aligned} &\|(\boldsymbol{\Sigma}_\eta - \boldsymbol{\Sigma}_{F,\eta}^{'}\mathbf{M}(\boldsymbol{\Sigma}_F^{-1}+\mathbf{M})^{-1}\boldsymbol{\Sigma}_F^{-1}\boldsymbol{\Sigma}_{F,\eta})\|^2 \\ &\leq \|\boldsymbol{\Sigma}_\eta - \boldsymbol{\Sigma}_{F,\eta}^{'}\boldsymbol{\Sigma}_F^{-1}\boldsymbol{\Sigma}_{F,\eta}\|^2 + \|\boldsymbol{\Sigma}_{F,\eta}^{'}\mathbf{M}(\boldsymbol{\Sigma}_F^{-1}+\mathbf{M})^{-1}\boldsymbol{\Sigma}_F^{-1}\mathbf{M}^{-1}\boldsymbol{\Sigma}_F^{-1}\boldsymbol{\Sigma}_{F,\eta}\|^2 && (1.167)\\ &\leq \|\boldsymbol{\Sigma}_{F,\eta}^{'}\boldsymbol{\Sigma}_F^{-1}\mathbf{I}_T \otimes \mathbf{Q}^\star\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\Sigma}_{F,\eta}\|^2 + \|\boldsymbol{\Sigma}_{F,\eta}^{'}\boldsymbol{\Sigma}_F^{-1}\mathbf{M}^{-1}\boldsymbol{\Sigma}_F^{-1}\boldsymbol{\Sigma}_{F,\eta}\|^2 && (1.168)\\ &\leq \|\mathbf{I}_T \otimes \mathbf{Q}^\star\|^2\|\boldsymbol{\Sigma}_{F,\eta}^{'}\|^4\|\boldsymbol{\Sigma}_F^{-1}\|^2\|\boldsymbol{\Sigma}_G^{-1}\|^2 + \|\boldsymbol{\Sigma}_{F,\eta}^{'}\boldsymbol{\Sigma}_F^{-1}\mathbf{M}^{-1}\boldsymbol{\Sigma}_F^{-1}\boldsymbol{\Sigma}_{F,\eta}\|^2 && (1.169)\\ &\leq \|\mathbf{Q}^\star\|^2\|\boldsymbol{\Sigma}_{F,\eta}^{'}\|^4\|\boldsymbol{\Sigma}_F^{-1}\|^2\|\boldsymbol{\Sigma}_G^{-1}\|^2 + \|\boldsymbol{\Sigma}_{F,\eta}^{'}\boldsymbol{\Sigma}_F^{-1}\|^4\|\mathbf{I}_T \otimes (\boldsymbol{\Lambda}^{'}\boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Lambda})^{-1}\|^2 && (1.170)\\ &\leq \|\mathbf{Q}^\star\|^2\|\boldsymbol{\Sigma}_{F,\eta}^{'}\|^4\|\boldsymbol{\Sigma}_F^{-1}\|^2\|\boldsymbol{\Sigma}_G^{-1}\|^2 + \|(\boldsymbol{\Lambda}^{'}\boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Lambda})^{-1}\|^2\|\boldsymbol{\Sigma}_{F,\eta}^{'}\|^4\|\boldsymbol{\Sigma}_F^{-1}\|^4, && (1.171)\end{aligned}$$

where we used (1.114) in the third passage. The first term $\|\mathbf{Q}^\star\|^2 = O(\epsilon)$ with $\epsilon \to 0$. $\|\boldsymbol{\Sigma}_F^{-1}\|$, $\|\boldsymbol{\Sigma}_G^{-1}\|$, and $\|\boldsymbol{\Sigma}_{F,\eta}^{'}\|$ are *finite* and $\|(\boldsymbol{\Lambda}^{'}\boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Lambda})^{-1}\|^2 = O(n^{-2})$. So we have proven that:

$$\mathbf{F}_{1,t|T}^{\dagger} \to \mathbf{F}_t^{\dagger} \qquad \text{and} \qquad \mathbf{F}_{1,t|T}^{\dagger} = \mathbf{F}_t^{\dagger} + O\left(\frac{1}{n}\right). \tag{1.172}$$

Finally, let us focus on the second term of Equation (1.145), which we will denote as $\mathbf{F}_{2,t|T}^{\dagger}$

$$\mathbf{F}_{2,t|T}^{\dagger} = (\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)(\boldsymbol{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}\boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{Z}_T. \tag{1.173}$$

As above, denote $\mathbf{K} = (\boldsymbol{\Sigma}_{F^\dagger}^{-1} + \mathbf{M}^\dagger)^{-1}$ with $\mathbf{M}^\dagger = \boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^{\dagger}$ and take the expected value of the norm

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}\big[\|\mathbf{F}_{2,t|T}^{\dagger}\|^2\big] &= \mathbb{E}_{\boldsymbol{\theta}}\big[\|(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{K}\boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{Z}_T\|^2\big] && (1.174)\\ &= \mathbb{E}_{\boldsymbol{\theta}}\big[tr\big\{(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{K}\boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{Z}_T\boldsymbol{Z}_T^{'}\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^{\dagger}\mathbf{K}^{'}(\boldsymbol{\iota}_{2t} \otimes \mathbf{I}_r)\big\}\big] && (1.175)\\ &= tr\big\{(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{K}\boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{\Sigma}_Z\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{L}_T^{\dagger}\mathbf{K}^{'}(\boldsymbol{\iota}_{2t} \otimes \mathbf{I}_r)\big\} && (1.176)\\ &= \|(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{K}\boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1/2}\boldsymbol{\Sigma}_{0Z}^{-1/2}\boldsymbol{\Sigma}_Z^{1/2}\|^2 && (1.177)\\ &\leq \|(\boldsymbol{\iota}_{2t}^{'} \otimes \mathbf{I}_r)\mathbf{K}\boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1/2}\|^2\|\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{\Sigma}_Z\| && (1.178)\\ &\leq \|\mathbf{K}\|^2\|\boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1/2}\|^2\|\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{\Sigma}_Z\|. && (1.179)\end{aligned}$$

We already proved that $\|\mathbf{K}\|^2 = O(n^{-2})$. The second term is

$$\begin{aligned} \|\boldsymbol{L}_T^{\dagger'}\boldsymbol{\Sigma}_{0Z}^{-1/2}\|^2 &= \|\mathbf{I}_T \otimes \boldsymbol{\Lambda}\boldsymbol{\Gamma}_{0Z}^{-1/2}\|^2 && (1.180)\\ &= \|\boldsymbol{\Lambda}\boldsymbol{\Gamma}_{0Z}^{-1/2}\|^2 && (1.181)\\ &= \|\boldsymbol{\Lambda}^{'}\boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Lambda}\|, && (1.182)\end{aligned}$$

which is $O(n)$. Finally, we know from Assumption (A3) that the last norm is *finite*. Indeed,

$$\begin{aligned} \|\boldsymbol{\Sigma}_{0Z}^{-1}\boldsymbol{\Sigma}_Z\| &\leq \frac{\lambda_{max}(\boldsymbol{\Sigma}_Z)}{\lambda_{min}(\boldsymbol{\Sigma}_{0Z})} && (1.183)\\ &\leq \frac{\lambda_{max}(\boldsymbol{\Gamma})}{\lambda_{min}(\boldsymbol{\Gamma}_0)}. && (1.184) \end{aligned}$$

It follows that:

$$\mathbf{F}_{2,t|T}^{\dagger} \to \mathbf{0} \qquad \text{and} \qquad \mathbf{F}_{2,t|T}^{\dagger} = O\left(\frac{1}{\sqrt{n}}\right). \tag{1.185}$$

We can finally put this result together with the one obtained for $\mathbf{F}_{1,t|T}^{\dagger}$. This implies the mean square convergence of the Kalman smoother for all values of t and when parameters are known, i.e.

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{F}_t^{\dagger} - \mathbf{F}_{t|T}^{\dagger}\|^2\right] = O\left(\frac{1}{n}\right). \tag{1.186}$$

1.7.4 Kalman Filter Consistency

Let us assume that the true value of the parameter $\boldsymbol{\theta}$ is known and initial conditions $F_{0|0}^{\dagger}$ and $P_{0|0}^{\dagger}$ are given. For $t < T$ we have

$$\begin{aligned} \mathbf{F}_{t|t}^{\dagger} &= \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{F}_t^{\dagger}|\mathcal{X}_t] && (1.187)\\ \mathbf{P}_{t|t}^{\dagger} &= \mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{F}_t^{\dagger} - \mathbf{F}_{t|t}^{\dagger})(\mathbf{F}_t^{\dagger} - \mathbf{F}_{t|t}^{\dagger})^{'}|\mathcal{X}_t] && (1.188) \end{aligned}$$

which are of dimension $2r \times 1$ and $2r \times 2r$, respectively. As $\mathbf{F}_t^{\dagger} = [\mathbf{F}_t^{'} \quad \boldsymbol{\eta}_t^{'}]^{'}$ we can denote $\mathbf{P}_{t|t}^{F} = \mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{F}_t - \mathbf{F}_{t|t})(\mathbf{F}_t - \mathbf{F}_{t|t})^{'}|\mathcal{X}_t]$ and $\mathbf{P}_{t|t}^{\eta} = \mathbb{E}_{\boldsymbol{\theta}}[(\boldsymbol{\eta}_t - \boldsymbol{\eta}_{t|t})(\boldsymbol{\eta}_t - \boldsymbol{\eta}_{t|t})^{'}|\mathcal{X}_t]$. First of all, let's examine the Riccati difference equation as in (1.74):

$$\mathbf{P}_{t+1|t}^{\dagger} = \boldsymbol{\Phi}^{\dagger}\mathbf{P}_{t|t-1}^{\dagger}\boldsymbol{\Phi}^{\dagger'} - \boldsymbol{\Phi}^{\dagger}\mathbf{P}_{t|t-1}^{\dagger}\boldsymbol{\Lambda}^{\dagger'}(\boldsymbol{\Lambda}^{\dagger}\mathbf{P}_{t|t-1}^{\dagger}\boldsymbol{\Lambda}^{\dagger'} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}^{\dagger}\mathbf{P}_{t|t-1}^{\dagger}\boldsymbol{\Phi}^{\dagger'} + \boldsymbol{\Psi}^{\dagger}\mathbf{Q}_{t+1|t}^{\dagger}\boldsymbol{\Psi}^{\dagger'}. \tag{1.189}$$

Given that

$$\boldsymbol{\Psi}^{\dagger}\mathbf{Q}_{t+1|t}^{\dagger}\boldsymbol{\Psi}^{\dagger'} = \begin{bmatrix}\mathbf{I}_r & \mathbf{I}_r\\ \mathbf{0} & \mathbf{I}_r\end{bmatrix}\begin{bmatrix}\mathbf{Q}^* & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}_{t+1|t}\end{bmatrix}\begin{bmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{I}_r & \mathbf{I}_r\end{bmatrix} = \begin{bmatrix}\mathbf{Q}^* + \mathbf{Q}_{t+1|t} & \mathbf{Q}_{t+1|t}\\ \mathbf{Q}_{t+1|t} & \mathbf{Q}_{t+1|t}\end{bmatrix}, \tag{1.190}$$

we can write the upper-left $r \times r$ block matrix that refers to $\mathbf{F}_t$ as:

$$\mathbf{P}_{t+1|t}^{F} = \boldsymbol{\Phi}\mathbf{P}_{t|t-1}^{F}\boldsymbol{\Phi}^{'} - \boldsymbol{\Phi}\mathbf{P}_{t|t-1}^{F}\boldsymbol{\Lambda}^{'}(\boldsymbol{\Lambda}\mathbf{P}_{t|t-1}^{F}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}\mathbf{P}_{t|t-1}^{F}\boldsymbol{\Phi}^{'} + \mathbf{Q}_{t+1|t} + \mathbf{Q}^*. \tag{1.191}$$

Using a modified version of the Woodbury identity, such that for any matrix $A, B, C \in \mathbb{R}^{n\times n}$ with A and C invertible, $AC(B + C'AC)C'A = A - (A^{-1} + CB^{-1}C')^{-1}$ we obtain

$$\begin{aligned} \mathbf{P}_{t+1|t}^{F} &= \boldsymbol{\Phi}\mathbf{P}_{t|t-1}^{F}\boldsymbol{\Phi}^{'} - \boldsymbol{\Phi}\mathbf{P}_{t|t-1}^{F}\boldsymbol{\Phi}^{'} + \boldsymbol{\Phi}(\mathbf{P}_{t|t-1}^{F-1} + \boldsymbol{\Lambda}^{'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Phi}^{'} + \mathbf{Q}_{t+1|t} + \mathbf{Q}^* && (1.192)\\ &= \boldsymbol{\Phi}(\mathbf{P}_{t|t-1}^{F-1} + \boldsymbol{\Lambda}^{'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Phi}^{'} + \mathbf{Q}_{t+1|t} + \mathbf{Q}^* && (1.193) \end{aligned}$$

as $n \to \infty$ and for a negligible ϵ, $\mathbf{P}^F_{t+1|t} \to \mathbf{Q}_{t+1|t}$. As a matter of fact,

$$\begin{aligned}
\|\mathbf{P}^F_{t+1|t} - \mathbf{Q}_{t+1|t}\| &= \|\mathbf{\Phi}(\mathbf{P}^{F-1}_{t|t-1} + \mathbf{\Lambda}'\mathbf{\Gamma}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Phi}' + \mathbf{Q}^*\| && (1.194)\\
&\leq \|\mathbf{\Phi}(\mathbf{P}^{F-1}_{t|t-1} + \mathbf{\Lambda}'\mathbf{\Gamma}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Phi}'\| + \|\mathbf{Q}^*\| && (1.195)\\
&\leq \|\mathbf{\Phi}\|^2\|(\mathbf{\Lambda}'\mathbf{\Gamma}^{-1}\mathbf{\Lambda})^{-1}\| + \|\mathbf{Q}^*\| && (1.196)\\
&= O(n^{-1}) + O(\epsilon) && (1.197)
\end{aligned}$$

since $\|\mathbf{\Phi}\|$ is *finite* and $\|(\mathbf{\Lambda}'\mathbf{\Gamma}^{-1}\mathbf{\Lambda})^{-1}\| = O(n^{-1})$. As for the other blocks, we can notice that

$$\mathbf{P}^{\eta}_{t+1|t} = \mathbf{P}^{F,\eta}_{t+1|t} = \mathbf{P}^{\eta,F}_{t+1|t} = \mathbf{Q}_{t+1|t} \tag{1.198}$$

because all blocks but the upper-left in the first and second addends of (1.189) are $\mathbf{0}$. Given these premises, Kalman filter consistency can be proved based on recursion formula (1.73). Using Woodbury identity we can rewrite it as

$$\begin{aligned}
\mathbf{P}^{\dagger}_{t|t} &= \mathbf{P}^{\dagger}_{t|t-1} - \mathbf{P}^{\dagger}_{t|t-1}\mathbf{\Lambda}^{\dagger'}(\mathbf{\Lambda}^{\dagger}\mathbf{P}^{\dagger}_{t|t-1}\mathbf{\Lambda}^{\dagger'} + \mathbf{\Gamma})^{-1}\mathbf{\Lambda}^{\dagger}\mathbf{P}^{\dagger}_{t|t-1} && (1.199)\\
&= (\mathbf{P}^{\dagger-1}_{t|t-1} + \mathbf{\Lambda}^{\dagger'}\mathbf{\Gamma}^{-1}\mathbf{\Lambda}^{\dagger})^{-1} && (1.200)\\
&= \left(\begin{bmatrix} \mathbf{P}^F_{t|t-1} & \mathbf{P}^{F,\eta}_{t|t-1} \\ \mathbf{P}^{\eta,F}_{t|t-1} & \mathbf{P}^{\eta}_{t|t-1} \end{bmatrix}^{-1} + \begin{bmatrix} \mathbf{V} & \mathbf{0}_r \\ \mathbf{0}_r & \mathbf{0}_r \end{bmatrix}\right)^{-1} && (1.201)
\end{aligned}$$

with $\mathbf{V} = \mathbf{\Lambda}'\mathbf{\Gamma}^{-1}\mathbf{\Lambda}$. We can now use a similar step to that used in the Kalman smoother consistency to obtain

$$\begin{aligned}
\mathbf{P}^{\dagger}_{t|t} &= \left(\begin{bmatrix} \mathbf{I}_r & \mathbf{0}_r \\ \mathbf{0}_r & \mathbf{I}_r \end{bmatrix} + \begin{bmatrix} \mathbf{P}^F_{t|t-1} & \mathbf{P}^{F,\eta}_{t|t-1} \\ \mathbf{P}^{\eta,F}_{t|t-1} & \mathbf{P}^{\eta}_{t|t-1} \end{bmatrix}\begin{bmatrix} \mathbf{V} & \mathbf{0}_r \\ \mathbf{0}_r & \mathbf{0}_r \end{bmatrix}\right)^{-1}\begin{bmatrix} \mathbf{P}^F_{t|t-1} & \mathbf{P}^{F,\eta}_{t|t-1} \\ \mathbf{P}^{\eta,F}_{t|t-1} & \mathbf{P}^{\eta}_{t|t-1} \end{bmatrix} && (1.202)\\
&= \left(\begin{bmatrix} \mathbf{I}_r & \mathbf{0}_r \\ \mathbf{0}_r & \mathbf{I}_r \end{bmatrix} + \begin{bmatrix} \mathbf{P}^F_{t|t-1}\mathbf{V} & \mathbf{0}_r \\ \mathbf{P}^{\eta,F}_{t|t-1}\mathbf{V} & \mathbf{0}_r \end{bmatrix}\right)^{-1}\begin{bmatrix} \mathbf{P}^F_{t|t-1} & \mathbf{P}^{F,\eta}_{t|t-1} \\ \mathbf{P}^{\eta,F}_{t|t-1} & \mathbf{P}^{\eta}_{t|t-1} \end{bmatrix} && (1.203)\\
&= \begin{bmatrix} \mathbf{I}_r + \mathbf{P}^F_{t|t-1}\mathbf{V} & \mathbf{0}_r \\ \mathbf{P}^{\eta,F}_{t|t-1}\mathbf{V} & \mathbf{I}_r \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{P}^F_{t|t-1} & \mathbf{P}^{F,\eta}_{t|t-1} \\ \mathbf{P}^{\eta,F}_{t|t-1} & \mathbf{P}^{\eta}_{t|t-1} \end{bmatrix} && (1.204)\\
&= \begin{bmatrix} (\mathbf{I}_r + \mathbf{P}^F_{t|t-1}\mathbf{V})^{-1} & \mathbf{0}_r \\ -\mathbf{P}^{\eta,F}_{t|t-1}\mathbf{V}(\mathbf{I}_r + \mathbf{P}^F_{t|t-1}\mathbf{V})^{-1} & \mathbf{I}_r \end{bmatrix}\begin{bmatrix} \mathbf{P}^F_{t|t-1} & \mathbf{P}^{F,\eta}_{t|t-1} \\ \mathbf{P}^{\eta,F}_{t|t-1} & \mathbf{P}^{\eta}_{t|t-1} \end{bmatrix} && (1.205)\\
&= \begin{bmatrix} (\mathbf{P}^{F-1}_{t|t-1} + \mathbf{V})^{-1}\mathbf{P}^{F-1}_{t|t-1} & \mathbf{0}_r \\ -\mathbf{P}^{\eta,F}_{t|t-1}\mathbf{V}(\mathbf{P}^{F-1}_{t|t-1} + \mathbf{V})^{-1}\mathbf{P}^{F-1}_{t|t-1} & \mathbf{I}_r \end{bmatrix}\begin{bmatrix} \mathbf{P}^F_{t|t-1} & \mathbf{P}^{F,\eta}_{t|t-1} \\ \mathbf{P}^{\eta,F}_{t|t-1} & \mathbf{P}^{\eta}_{t|t-1} \end{bmatrix} && (1.206)\\
&= \begin{bmatrix} (\mathbf{P}^{F-1}_{t|t-1} + \mathbf{V})^{-1} & (\mathbf{P}^{F-1}_{t|t-1} + \mathbf{V})^{-1}\mathbf{P}^{F-1}_{t|t-1}\mathbf{P}^{\eta,F}_{t|t-1} \\ \mathbf{P}^{\eta,F}_{t|t-1} - \mathbf{P}^{\eta,F}_{t|t-1}\mathbf{V}(\mathbf{P}^{F-1}_{t|t-1} + \mathbf{V})^{-1} & \mathbf{P}^{\eta}_{t|t-1} - \mathbf{P}^{\eta,F}_{t|t-1}\mathbf{V}(\mathbf{P}^{F-1}_{t|t-1} + \mathbf{V})^{-1}\mathbf{P}^{F-1}_{t|t-1}\mathbf{P}^{F,\eta}_{t|t-1} \end{bmatrix}.
\end{aligned}$$

Now, passing to the norm, since $\mathbf{P}_{t|t}^{\dagger}$ is positive definite, then we can use block matrix norm inequalities to find that

$$\begin{aligned}
\|\mathbf{P}_{t|t}^{\dagger}\| &\leq \|(\mathbf{P}_{t|t-1}^{F-1}+\mathbf{V})^{-1}\| + \qquad (1.207)\\
&\quad \|\mathbf{P}_{t|t-1}^{\eta}-\mathbf{P}_{t|t-1}^{\eta,F}\mathbf{V}(\mathbf{P}_{t|t-1}^{F-1}+\mathbf{V})^{-1}\mathbf{P}_{t|t-1}^{F-1}\mathbf{P}_{t|t-1}^{F,\eta}\| \\
&\leq \|(\mathbf{P}_{t|t-1}^{F-1}+\mathbf{V})^{-1}\| + \qquad (1.208)\\
&\quad \|\mathbf{P}_{t|t-1}^{\eta}-\mathbf{P}_{t|t-1}^{\eta,F}\mathbf{P}_{t|t-1}^{F-1}\mathbf{P}_{t|t-1}^{F,\eta}\| + \\
&\quad \|\mathbf{P}_{t|t-1}^{\eta,F}\mathbf{V}(\mathbf{P}_{t|t-1}^{F-1}+\mathbf{V})^{-1}\mathbf{P}_{t|t-1}^{F-1}\mathbf{V}^{-1}\mathbf{P}_{t|t-1}^{F-1}\mathbf{P}_{t|t-1}^{F,\eta}\|,
\end{aligned}$$

where we used the relation $(A+B)^{-1}=A^{-1}-(A+B)^{-1}BA^{-1}$ in the second step. As for the first term, we know that

$$\begin{aligned}
\|(\mathbf{P}_{t|t-1}^{F-1}+\mathbf{V})^{-1}\| &= \|(\mathbf{P}_{t|t-1}^{F-1}+\mathbf{\Lambda}'\mathbf{\Gamma}^{-1}\mathbf{\Lambda})^{-1}\| \leq \|(\mathbf{\Lambda}'\mathbf{\Gamma}^{-1}\mathbf{\Lambda})^{-1}\| \qquad (1.209)\\
&= O(n^{-1}) \qquad (1.210)
\end{aligned}$$

for each value of $\mathbf{P}_{t|t-1}^{F-1}$, since the matrix is positive definite by construction and the term on the right-hand side converges to $O(n^{-1})$ For the second part, we can instead use relation (1.198) such that

$$\begin{aligned}
\|\mathbf{P}_{t|t-1}^{\eta}-\mathbf{P}_{t|t-1}^{\eta,F}\mathbf{P}_{t|t-1}^{F-1}\mathbf{P}_{t|t-1}^{F,\eta}\| &= \|\mathbf{Q}_{t|t-1}-\mathbf{Q}_{t|t-1}(\mathbf{P}_{t|t-1}^{F})^{-1}\mathbf{Q}_{t|t-1}\| \qquad (1.211)\\
&= O(n^{-1})+O(\epsilon) \qquad (1.212)
\end{aligned}$$

as we proved that $\mathbf{P}_{t+1|t}^{F}\to\mathbf{Q}_{t+1|t}$ as $n\to\infty$ and $\epsilon\to 0$. Finally, the last term

$$\begin{aligned}
&\|\mathbf{P}_{t|t-1}^{\eta,F}\mathbf{V}(\mathbf{P}_{t|t-1}^{F-1}+\mathbf{V})^{-1}\mathbf{P}_{t|t-1}^{F-1}\mathbf{V}^{-1}\mathbf{P}_{t|t-1}^{F-1}\mathbf{P}_{t|t-1}^{F,\eta}\| \qquad (1.213)\\
&\leq \|\mathbf{P}_{t|t-1}^{\eta,F}\mathbf{P}_{t|t-1}^{F-1}\mathbf{V}^{-1}\mathbf{P}_{t|t-1}^{F-1}\mathbf{P}_{t|t-1}^{F,\eta}\| \\
&\leq \|\mathbf{Q}_{t|t-1}\|^2\|\mathbf{P}_{t|t-1}^{F-1}\|^2\|(\mathbf{\Lambda}'\mathbf{\Gamma}^{-1}\mathbf{\Lambda})^{-1}\| \qquad (1.214)\\
&= O(n^{-1}), \qquad (1.215)
\end{aligned}$$

given that the fist term is bounded by the unconditional variance, being a stationary GARCH(1,1), and the second one is *finite* since the matrix $\mathbf{P}_{t|t-1}^{F}$ is positive definite by construction. So that for each $t=1,\ldots,T$ we have

$$\mathbf{P}_{t|t}^{\dagger}=\mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{F}_{t}^{\dagger}-\mathbf{F}_{t|t}^{\dagger})(\mathbf{F}_{t}^{\dagger}-\mathbf{F}_{t|t}^{\dagger})'|\mathcal{X}_t]=O\left(\frac{1}{n}\right). \qquad (1.216)$$

Then, by using the law of iterated expectation we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}\big[\|\mathbf{F}_{t}^{\dagger}-\mathbf{F}_{t|t}^{\dagger}\|^2\big] &= \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{E}_{\boldsymbol{\theta}}[tr\big\{(\mathbf{F}_{t}^{\dagger}-\mathbf{F}_{t|t}^{\dagger})(\mathbf{F}_{t}^{\dagger}-\mathbf{F}_{t|t}^{\dagger})'|\mathcal{X}_t\big\}] \qquad (1.217)\\
&= \mathbb{E}_{\boldsymbol{\theta}}[tr\big\{\mathbf{P}_{t|t}^{\dagger}\big\}]]=\sum_{i=1}^{2r}\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{P}_{t|t}^{\dagger}]_{ii} \qquad (1.218)\\
&\leq 2r\ \max\mathbb{E}_{\boldsymbol{\theta}}|\mathbf{P}_{t|t}^{\dagger}|_{ij}=r\ \mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{P}_{t|t}^{\dagger}\|_{\max} \qquad (1.219)\\
&\leq 2r\ \mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{P}^{\dagger}{}_{t|t}\|. \qquad (1.220)
\end{aligned}$$

Given that $\|\mathbf{P}_{t|t}^{\dagger}\|$ is smaller than the term in (1.207), by the *Dominated Covngergence Theorem* we can interchange expectation and limit so the proof completes and we finally have that

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{F}_{t}^{\dagger}-\mathbf{F}_{t|t}^{\dagger}\|^{2}\right]=O_{p}\left(\frac{1}{n}\right). \tag{1.221}$$

1.7.5 Further results on the Kalman Filter and Smoother

Lemma 4 *Denote with $t \in \mathbb{N}$ the generic time index and with $S, T \in \mathbb{N}$ the initial values for the Kalman smoother backward iterations, so that $t < S < T$. Given the Kalman filter and smoother consistency proprieties obtained in (A4.3) and (A4.4) the following hold:*

(i) $\|\mathbf{F}_{t|T}^{\dagger}-\mathbf{F}_{t|t}^{\dagger}\|=O(n^{-1})$;

(ii) $\|\mathbf{F}_{t|T}^{\dagger}-\mathbf{F}_{t|S}^{\dagger}\|=O(n^{-1})$;

(iii) $\|\mathbf{P}_{t|T}^{\dagger}-\mathbf{P}_{t|t}^{\dagger}\|=O(n^{-2})$;

(iv) $\|\mathbf{P}_{t|T}^{\dagger}-\mathbf{P}_{t|S}^{\dagger}\|=O(n^{-2})$.

Proof. We know already from (1.186) and (1.221) that $\mathbf{P}_{t|T}^{\dagger}=O(n^{-1})$ and $\mathbf{P}_{t|t}^{\dagger}=O(n^{-1})$. Now, if we use Kalman smoother backward iteration as in (1.75), we have that

$$\mathbf{F}_{t|T}^{\dagger}=\mathbf{F}_{t|t}^{\dagger}+\mathbf{P}_{t|t}^{\dagger}\boldsymbol{\Phi}^{\dagger\prime}\mathbf{P}_{t+1|t}^{\dagger-1}(\mathbf{F}_{t+1|T}^{\dagger}-\mathbf{F}_{t+1|t}^{\dagger}) \tag{1.222}$$

$$\mathbf{F}_{t|T}^{\dagger}-\mathbf{F}_{t|t}^{\dagger}=\mathbf{P}_{t|t}^{\dagger}\boldsymbol{\Phi}^{\dagger\prime}\mathbf{P}_{t+1|t}^{\dagger-1}(\mathbf{F}_{t+1|T}^{\dagger}-\mathbf{F}_{t+1|t}^{\dagger}) \tag{1.223}$$

$$\|\mathbf{F}_{t|T}^{\dagger}-\mathbf{F}_{t|t}^{\dagger}\|=O(n^{-1}), \tag{1.224}$$

because $\mathbf{P}_{t|t}^{\dagger}=O(n^{-1})$, and all the other terms can be shown to be finite, proving (i). Another important asymptotic result concerns the starting point from which the algorithm begins to iterate backwards. Generally, one sets up the most forward estimate of the Kalman filter as the initial condition for the smoother, $\mathbf{F}_{t|T}^{\dagger}=\mathbf{F}_{t|t}^{\dagger}$ when $t=T$. To prove (ii), let us consider a different starting point $S<T$. In this case

$$\mathbf{F}_{t|S}^{\dagger}=\mathbf{F}_{t|t}^{\dagger}+\mathbf{P}_{t|t}^{\dagger}\boldsymbol{\Phi}^{\dagger\prime}\mathbf{P}_{t+1|t}^{\dagger-1}(\mathbf{F}_{t+1|S}^{\dagger}-\mathbf{F}_{t+1|t}^{\dagger}). \tag{1.225}$$

Subtracting (1.225) from (1.222) we obtain

$$\mathbf{F}_{t|T}^{\dagger}-\mathbf{F}_{t|S}^{\dagger}=\mathbf{P}_{t|t}^{\dagger}\boldsymbol{\Phi}^{\dagger\prime}\mathbf{P}_{t+1|t}^{\dagger-1}(\mathbf{F}_{t+1|T}^{\dagger}-\mathbf{F}_{t+1|S}^{\dagger}) \tag{1.226}$$

$$\|\mathbf{F}_{t|T}-\mathbf{F}_{t|S}\|=O(n^{-1}), \tag{1.227}$$

as $n\to\infty$, since again, $\mathbf{P}_{t|t}^{\dagger}=O(n^{-1})$, and all the other terms are finite. Thus, as long as we increase the dimension of the cross-section, the starting point for the recursion is not

relevant.
For (iii) we can start using the Kalman smoother backward iteration for $\mathbf{P}^{\dagger}_{t|T}$:

$$\mathbf{P}^{\dagger}_{t|T} = \mathbf{P}^{\dagger}_{t|t} + \mathbf{P}^{\dagger}_{t|t}\mathbf{\Phi}^{\dagger\prime}\mathbf{P}^{\dagger-1}_{t+1|t}(\mathbf{P}^{\dagger}_{t+1|T} - \mathbf{P}^{\dagger}_{t+1|t})\mathbf{P}^{\dagger-1}_{t+1|t}\mathbf{\Phi}^{\dagger}\mathbf{P}^{\dagger}_{t|t} \tag{1.228}$$

$$\mathbf{P}^{\dagger}_{t|T} - \mathbf{P}^{\dagger}_{t|t} = \mathbf{P}^{\dagger}_{t|t}\mathbf{\Phi}^{\dagger\prime}\mathbf{P}^{\dagger-1}_{t+1|t}(\mathbf{P}^{\dagger}_{t+1|T} - \mathbf{P}^{\dagger}_{t+1|t})\mathbf{P}^{\dagger-1}_{t+1|t}\mathbf{\Phi}^{\dagger}\mathbf{P}^{\dagger}_{t|t} \tag{1.229}$$

$$\|\mathbf{P}_{t|T} - \mathbf{P}_{t|t}\| = O(n^{-2}), \tag{1.230}$$

as $\mathbf{P}^{\dagger}_{t|t} = O(n^{-1})$, the term in parentheses is positive semi-definite and the others are *finite*. Finally, let us consider a different starting point $S < T$,

$$\mathbf{P}^{\dagger}_{t|S} = \mathbf{P}^{\dagger}_{t|t} + \mathbf{P}^{\dagger}_{t|t}\mathbf{\Phi}^{\dagger\prime}\mathbf{P}^{\dagger-1}_{t+1|t}(\mathbf{P}^{\dagger}_{t+1|S} - \mathbf{P}^{\dagger}_{t+1|t})\mathbf{P}^{\dagger-1}_{t+1|t}\mathbf{\Phi}^{\dagger}\mathbf{P}^{\dagger}_{t|t}. \tag{1.231}$$

Then, subtracting (1.231) from (1.228) we have

$$\mathbf{P}^{\dagger}_{t|T} - \mathbf{P}^{\dagger}_{t|S} = \mathbf{P}^{\dagger}_{t|t}\mathbf{\Phi}^{\dagger\prime}\mathbf{P}^{\dagger-1}_{t+1|t}(\mathbf{P}^{\dagger}_{t+1|T} - \mathbf{P}^{\dagger}_{t+1|S})\mathbf{P}^{\dagger-1}_{t+1|t}\mathbf{\Phi}^{\dagger}\mathbf{P}^{\dagger}_{t|t} \tag{1.232}$$

$$\|\mathbf{P}^{\dagger}_{t|T} - \mathbf{P}^{\dagger}_{t|S}\| = O(n^{-2}) \tag{1.233}$$

since, again $\mathbf{P}^{\dagger}_{t|t} = O(n^{-1})$, and the other terms are *finite*.

1.7.6 Conditional Variance Consistency

Define by $\mathcal{X}_{t-1}$, the σ-field generated by $\boldsymbol{x}_t$ up to and including time $t-1$. To avoid unnecessary notational burden, let us consider a single factor $\mathbf{F}^{\dagger}_t = [F_t \ \eta_t]^{\prime}$ with from the Augmented Model as in (1.5) and (1.6)[14]. We want to calculate:

$$q_{t|t-1} = \mathrm{Var}_{\boldsymbol{\theta}}[\eta_t|\mathcal{X}_{t-1}] = \mathbb{E}_{\boldsymbol{\theta}}[\eta^2_t|\mathcal{X}_{t-1}] = \omega + \alpha\mathbb{E}_{\boldsymbol{\theta}}[\eta^2_{t-1}|\mathcal{X}_{t-1}] + \beta\mathbb{E}_{\boldsymbol{\theta}}[q_{t-1}|\mathcal{X}_{t-1}]. \tag{1.234}$$

Now, for the first term, denote by $\eta_{t-1|t-1}$ the estimate from the Kalman filter, i.e. $\eta_{t-1|t-1} = \mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-1}|\mathcal{X}_{t-1}]$, then we can write

$$\eta_{t-1} = \eta_{t-1|t-1} + (\eta_{t-1} - \eta_{t-1|t-1}). \tag{1.235}$$

Squaring both sides and taking expectation conditionally on $\mathcal{X}_{t-1}$ we obtain

$$\mathbb{E}_{\boldsymbol{\theta}}[\eta^2_{t-1}|\mathcal{X}_{t-1}] = \eta^2_{t-1|t-1} + P^{\eta}_{t-1|t-1}, \tag{1.236}$$

with $P^{\eta}_{t-1|t-1} = \mathbb{E}_{\boldsymbol{\theta}}[(\eta_{t-1} - \eta_{t-1|t-1})^2|\mathcal{X}_{t-1}]$ being the conditional variance of η_{t-1} given the information at time $t-1$. Here, we used the fact that the estimate from the KF is known at time $t-1$ so that we don't need expectation for the first term. Likewise, the cross product is zero.
The second term is more difficult to deal with since we don't know how to calculate

[14]Generalization to multiple factors is straightforward as $\mathbf{Q}_{t|t-1}$ is diagonal and each $q_{i,t|t}$ depends on the factor i only.

$\mathbb{E}_{\boldsymbol{\theta}}[q_{t-1}|\mathcal{X}_{t-1}]$. However, we could handle the expression $q_{t-1|t-2} = \mathbb{E}_{\boldsymbol{\theta}}[q_{t-1}|\mathcal{X}_{t-2}]$ more easily by recursion. As in Harvey et al. (1992), we repeatedly substitute q_{t-1} from (1.4). For $J \geq 1$ we get

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\theta}}[\eta_t|\mathcal{X}_{t-1}] &= \omega(1+\beta+\cdots+\beta^{J-1}) && (1.237)\\
&+ \alpha\sum_{j=1}^{J}\beta^{j-1}\mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-j}^2|\mathcal{X}_{t-1}] + \beta^J\mathbb{E}_{\boldsymbol{\theta}}[q_{t-J}|\mathcal{X}_{t-1}]. && (1.238)
\end{aligned}$$

We can simplify this expression by setting $J = \infty$. Subsequently, we multiply both sides of the equation for $t-1$ by β and subtract from the same expression for t. We then obtain

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{\theta}}[\eta_t|\mathcal{X}_{t-1}] &= \omega + \alpha\mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-1}^2|\mathcal{X}_{t-1}] && (1.239)\\
&+ \alpha\sum_{j=1}^{\infty}\beta^j(\mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-j-1}^2|\mathcal{X}_{t-1}] - \mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-j-1}^2|\mathcal{X}_{t-2}]) && (1.240)\\
&+ \beta\mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-1}^2|\mathcal{X}_{t-2}]. && (1.241)
\end{aligned}$$

Finally, we substitute the expectation with the values we got in (1.236)

$$q_{t|t-1} = \mathrm{Var}_{\boldsymbol{\theta}}[\eta_t|\mathcal{X}_{t-1}] = \omega + \alpha(\eta_{t-1|t-1}^2 + P_{t-1|t-1}^{\eta}) + \beta q_{t-1|t-2} + \delta_t, \quad (1.242)$$

where

$$\delta_t = \sum_{j=1}^{\infty}\alpha\beta^j(\mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-j-1}^2|\mathcal{X}_{t-1}] - \mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-j-1}^2|\mathcal{X}_{t-2}]). \quad (1.243)$$

Now, looking at the first term in parentheses, with $\eta_{t-j-1|t-1} = \mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-j-1}|\mathcal{X}_{t-1}]$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-j-1}^2|\mathcal{X}_{t-1}] &= \mathbb{E}_{\boldsymbol{\theta}}[(\eta_{t-j-1} - \eta_{t-j-1|t-1} + \eta_{t-j-1|t-1})^2|\mathcal{X}_{t-1}] && (1.244)\\
&= \mathbb{E}_{\boldsymbol{\theta}}[(\eta_{t-j-1} - \eta_{t-j-1|t-1})^2|\mathcal{X}_{t-1}] + \eta_{t-j-1|t-1}^2 && (1.245)\\
&= P_{t-j-1|t-1}^{\eta} + \eta_{t-j-1|t-1}^2, && (1.246)
\end{aligned}$$

where we used the fact that the estimates from the smoother $\eta_{t-j-1|t-1}$ is fixed and known at time $t-1$ so that the cross product is also 0. In the same way,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[\eta_{t-j-1}^2|\mathcal{X}_{t-2}] &= \mathbb{E}_{\boldsymbol{\theta}}[(\eta_{t-j-1} - \eta_{t-j-1|t-2} + \eta_{t-j-1|t-2})^2|\mathcal{X}_{t-2}] && (1.247)\\
&= \mathbb{E}_{\boldsymbol{\theta}}[(\eta_{t-j-1} - \eta_{t-j-1|t-2})^2|\mathcal{X}_{t-2}] + \eta_{t-j-1|t-2}^2 && (1.248)\\
&= P_{t-j-1|t-2}^{\eta} + \eta_{t-j-1|t-2}^2. && (1.249)
\end{aligned}$$

Then, we can rewrite (1.243) as

$$\begin{aligned}
\delta_t &= \sum_{j=1}^{\infty}\alpha\beta^j[(P_{t-j-1|t-1}^{\eta} - P_{t-j-1|t-2}^{\eta}) + (\eta_{t-j-1|t-1}^2 - \eta_{t-j-1|t-2}^2)] && (1.250)\\
&= \alpha\beta[(P_{t-2|t-1}^{\eta} - P_{t-2|t-2}^{\eta}) + (\eta_{t-2|t-1}^2 - \eta_{t-2|t-2}^2)] + && (1.251)\\
&+ \sum_{j>1}^{\infty}\alpha\beta^j[(P_{t-j-1|t-1}^{\eta} - P_{t-j-1|t-2}^{\eta}) + (\eta_{t-j-1|t-1}^2 - \eta_{t-j-1|t-2}^2)]. && (1.252)
\end{aligned}$$

We have previously proved that

$$\|\mathbf{F}^{\dagger}_{t|T} - \mathbf{F}^{\dagger}_{t|t}\| = O(n^{-1}) \qquad \text{and} \qquad \|\mathbf{F}^{\dagger}_{t|T} - \mathbf{F}^{\dagger}_{t|S}\| = O(n^{-1}), \tag{1.253}$$

which means that consistency also holds for $\boldsymbol{\eta}_t$, as this is part of $\mathbf{F}^{\dagger}_t$. Additionally, all the proprieties in Lemma 4 are valid.
In particular, by Lemma 4 (iii),

$$P^{\eta}_{t-2|t-1} - P^{\eta}_{t-2|t-2} = O(n^{-2}), \tag{1.254}$$

while

$$\eta^2_{t-2|t-1} - \eta^2_{t-2|t-2} = (\eta_{t-2|t-1} - \eta_{t-2|t-2})(\eta_{t-2|t-1} + \eta_{t-2|t-2}) = O(n^{-1}), \tag{1.255}$$

since the first term is $O(n^{-1})$ by Lemma 4 (i) and the second is *finite*. We can then use Lemma 4 (iv) to prove that

$$P^{\eta}_{t-j-1|t-1} - P^{\eta}_{t-j-1|t-2} = O(n^{-2}), \tag{1.256}$$

and by Lemma 4 (ii),

$$\eta^2_{t-j-1|t-1} - \eta^2_{t-j-1|t-2} = (\eta_{t-j-1|t-1} - \eta_{t-j-1|t-2})(\eta_{t-j-1|t-1} + \eta_{t-j-1|t-2}) = O(n^{-1}). \tag{1.257}$$

Finally, we have that

$$\delta_t = \sum_{j=1}^{\infty} \alpha\beta^j[(P^{\eta}_{t-j-1|t-1} - P^{\eta}_{t-j-1|t-2}) + (\eta^2_{t-j-1|t-1} - \eta^2_{t-j-1|t-2})] = O(n^{-1}), \tag{1.258}$$

since the O-terms do not depend on the summation index. This implies that as $n \to \infty$ the correction term vanishes to 0, i.e. $\delta_t \to 0$, at rate n^{-1}. Now, to prove conditional variance consistency, let us first transform the GARCH(1,1) into an ARCH(∞). In the same way as Bollerslev (1986), we have that

$$q_t = \frac{\omega}{1-\beta} + \alpha \sum_{j=1}^{\infty} \beta^{j-1}\eta^2_{t-j}, \tag{1.259}$$

and for the Kalman filter estimator, repeatedly substituting in (1.242) we obtain

$$q_{t|t} = \frac{\omega}{1-\beta} + \alpha \sum_{j=1}^{\infty} \beta^{j-1}(\eta^2_{t-j|t-j} + P^{\eta}_{t-j|t-j}) + \sum_{j=1}^{\infty} \beta^{j-1}\delta_{t+1-j}. \tag{1.260}$$

Then,

$$|q_t - q_{t|t}| \;=\; \left|\alpha \sum_{j=1}^{\infty} \beta^{j-1}(\eta^2_{t-j} - \eta^2_{t-j|t-j}) - \alpha \sum_{j=1}^{\infty} \beta^{j-1}P^{\eta}_{t-j|t-j} - \sum_{j=1}^{\infty} \beta^{j-1}\delta_{t+1-j}\right|. \tag{1.261}$$

We know that $\alpha, \beta, > 0$ and $\alpha + \beta < 1$ from Assumption (A1) and each of the terms in the summation is $O(n^{-1})$ and these O-terms are independent of j. Thus,

$$|q_t - q_{t|t}| = O(n^{-1}), \tag{1.262}$$

suggesting that as $n \to \infty$, the estimate from the Kalman filter converges to the GARCH(1,1) conditional variance, i.e. $q_{t|t} \to q_t$, with rate n^{-1}.

1.7.7 Multistep Forecast for $q_{t+h|t}$

Consider the problem of predicting $\text{Var}[\eta_{t+h}|\mathcal{X}_t]$, the conditional variance of η_{t+h} with information up to time t. Using $r = 1$ for simplicity, from (1.3) and (1.4)

$$\eta_t = \sqrt{q_t}\,\widetilde{\eta}_t \tag{1.263}$$
$$q_t = \omega + \alpha\eta_{t-1}^2 + \beta q_{t-1} \tag{1.264}$$

with $\widetilde{\eta}_t \sim \text{NID}(0,1)$. Define $q = 1 - \phi^2 = \text{Var}[\eta_t]$, the unconditional variance of η_t. We also know, from *variance targeting*, that $\omega = (1-\alpha-\beta)(1-\phi^2) = (1-\alpha-\beta)q$. Then,

$$q_{t+h} = (1-\alpha-\beta)q + \alpha\eta_{t+h-1}^2 + \beta q_{t+h-1} \tag{1.265}$$
$$q_{t+h} - q = \alpha(\eta_{t+h-1}^2 - q) + \beta(q_{t+h-1} - q). \tag{1.266}$$

Taking expectation with respect to $\mathcal{X}_t$ we obtain

$$\mathbb{E}[q_{t+h}|\mathcal{X}_t] - q = \alpha(\mathbb{E}[\eta_{t+h-1}^2|\mathcal{X}_t] - q) + \beta(\mathbb{E}[q_{t+h-1}|\mathcal{X}_t] - q). \tag{1.267}$$

But we know that $\mathbb{E}[\eta_{t+h-1}^2|\mathcal{X}_t] = \mathbb{E}[q_{t+h-1}|\mathcal{X}_t]$, since

$$\mathbb{E}[\eta_{t+h-1}^2|\mathcal{X}_t] = \mathbb{E}[q_{t+h-1}\widetilde{\eta}_{t+h-1}^2|\mathcal{X}_t] \tag{1.268}$$
$$= \mathbb{E}\left[\mathbb{E}[q_{t+h-1}\widetilde{\eta}_{t+h-1}^2|\mathcal{X}_{t+h-2}]|\mathcal{X}_t\right] \tag{1.269}$$
$$= \mathbb{E}[q_{t+h-1}|\mathcal{X}_t] \tag{1.270}$$

by the law of iterated expectation and using the fact that $\mathbb{E}[\widetilde{\eta}_t^2] = 1$ and $\mathbb{E}[q_{t+1}|\mathcal{X}_t] = q_{t+1}$. Substituting this result in (1.266)

$$\mathbb{E}[q_{t+h}|\mathcal{X}_t] - q = (\alpha+\beta)(\mathbb{E}[q_{t+h-1}|\mathcal{X}_t] - q) \tag{1.271}$$
$$= (\alpha+\beta)^{h-1}(\mathbb{E}[q_{t+1}|\mathcal{X}_t] - q) \tag{1.272}$$
$$= (\alpha+\beta)^{h-1}(q_{t+1|t} - q) \tag{1.273}$$
$$\mathbb{E}[q_{t+h}|\mathcal{X}_t] = 1 - \phi^2 + (\alpha+\beta)^{h-1}(q_{t+1|t} + \phi^2 - 1) \tag{1.274}$$

As $h \to \infty$, $\mathbb{E}[q_{t+h}|\mathcal{X}_t]$ converges to its unconditional variance $1 - \phi^2$.

1.7.8 Identification Condition on Factor Loadings

Denote by $\widehat{\mathbf{\Lambda}}$ the $n \times r$ matrix estimated using the ECME, that is:

$$\widehat{\mathbf{\Lambda}} = \left(\sum_{t=1}^{T} \boldsymbol{x}_t \widehat{\mathbf{F}}_{t|T}'\right)\left(\sum_{t=1}^{T} \widehat{\mathbf{F}}_{t|T}\widehat{\mathbf{F}}_{t|T}' + \mathbf{P}_{t|T}^F\right)^{-1}. \tag{1.275}$$

As $\widehat{\mathbf{F}}_{t|T}$ is a consistent estimator of $\mathbf{F}_t$, from (1.186) and (1.221) we know that

$$\mathbf{D}_1^{-1} = \left(\sum_{t=1}^{T} \widehat{\mathbf{F}}_{t|T}\widehat{\mathbf{F}}_{t|T}' + \mathbf{P}_{t|T}^F\right) \tag{1.276}$$
$$= n\mathbf{I}_r + O(n^{-1}) \tag{1.277}$$

is asymptotically diagonal, and so it will be its inverse. Now, we want to calculate

$$\widehat{\mathbf{\Lambda}}^{'}\widehat{\mathbf{\Lambda}} = \mathbf{D}_1 \left(\sum_{t=1}^{T} \widehat{\mathbf{F}}_{t|T} \boldsymbol{x}_t^{'} \right) \left(\sum_{t=1}^{T} \boldsymbol{x}_t \widehat{\mathbf{F}}_{t|T}^{'} \right) \mathbf{D}_1. \tag{1.278}$$

Now, replacing the sum $\boldsymbol{x}_t \widehat{\mathbf{F}}_{t|T}^{'}$ with their matrix multiplication $\widehat{\mathbf{F}}^{'}\mathbf{X}$, where $\widehat{\mathbf{F}} = (\widehat{\mathbf{F}}_{1T}, \cdots, \widehat{\mathbf{F}}_{T|T})$ is the $r \times T$ matrix containing the unobserved factors and $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T)$, then,

$$\widehat{\mathbf{\Lambda}}^{'}\widehat{\mathbf{\Lambda}} = \mathbf{D}_1 \widehat{\mathbf{F}} \mathbf{X}^{'} \mathbf{X} \widehat{\mathbf{F}}^{'} \mathbf{D}_1. \tag{1.279}$$

Using the fact that $\mathbf{X} = \mathbf{\Lambda F} + \mathbf{\Xi}$, with $\mathbf{\Xi} = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_T)$ we have that

$$\begin{aligned} \mathbf{X}^{'}\mathbf{X} &= (\mathbf{\Lambda F} + \mathbf{\Xi})^{'}(\mathbf{\Lambda F} + \mathbf{\Xi}) && (1.280) \\ &= (\mathbf{F}^{'}\mathbf{\Lambda}^{'} + \mathbf{\Xi}^{'})(\mathbf{\Lambda F} + \mathbf{\Xi}) && (1.281) \\ &= \mathbf{F}^{'}\mathbf{\Lambda}^{'}\mathbf{\Lambda F} + \mathbf{F}^{'}\mathbf{\Lambda}^{'}\mathbf{\Xi} + \mathbf{\Lambda F}\mathbf{\Xi}^{'} + \mathbf{\Xi}^{'}\mathbf{\Xi}, && (1.282) \end{aligned}$$

and then,

$$\widehat{\mathbf{F}}\mathbf{X}^{'}\mathbf{X}\widehat{\mathbf{F}}^{'} = \widehat{\mathbf{F}}\mathbf{F}^{'}\mathbf{\Lambda}^{'}\mathbf{\Lambda F}\widehat{\mathbf{F}}^{'} + \widehat{\mathbf{F}}\mathbf{F}^{'}\mathbf{\Lambda}^{'}\mathbf{\Xi}\widehat{\mathbf{F}}^{'} + \widehat{\mathbf{F}}\mathbf{\Lambda F}\mathbf{\Xi}^{'}\widehat{\mathbf{F}}^{'} + \widehat{\mathbf{F}}\mathbf{\Xi}^{'}\mathbf{\Xi}\widehat{\mathbf{F}}^{'}. \tag{1.283}$$

The last three terms on the right-hand side of the equation go to $\mathbf{0}$ asymptotically because of the product $\mathbf{\Xi}\widehat{\mathbf{F}}^{'}$. Furthermore, in a similar fashion to (1.277), the product $\mathbf{D}_2 = \widehat{\mathbf{F}}\mathbf{F}^{'}$ is diagonal when $n \to \infty$. Finally, substituting back in (1.279) and multiplying by n^{-1} we obtain

$$\frac{\widehat{\mathbf{\Lambda}}^{'}\widehat{\mathbf{\Lambda}}}{n} = \mathbf{D}_1\mathbf{D}_2\frac{\mathbf{\Lambda}^{'}\mathbf{\Lambda}}{n}\mathbf{D}_2\mathbf{D}_1. \tag{1.284}$$

Thus, $n^{-1}\widehat{\mathbf{\Lambda}}^{'}\widehat{\mathbf{\Lambda}}$ being diagonal depends on the main assumption of $n^{-1}\mathbf{\Lambda}^{'}\mathbf{\Lambda}$ and how the ECME is initialized. PCA identifying restrictions (Bai and Li, 2012), for example, require the product to be diagonal.

Chapter 2

Applications and Model Extensions

2.1 Growth at Risk

2.1.1 Introduction

Policymakers' attention to downside risk has shifted dramatically in recent years (Sánchez and Röhn, 2016; Prasad et al., 2019), prompting the creation of techniques to quantify the possibility and magnitude of severe occurrences in important economic variables (Ghysels et al., 2018). The International Monetary Fund (IMF) has lately popularized a risk metric known as Growth-at-Risk (GaR), which is the worst-case scenario for GDP growth at a certain coverage level and is the risk management equivalent of Value-at-Risk (VaR). Several institutions now report GaR for major international economies on a regular basis. This measure has been introduced by Adrian et al. (2019) to study the downside risks in periods of tight financial conditions. The authors used a quantile regression to model the 5% tail distribution of the GDP using a collection of quarterly financial variables provided by the IMF. Despite GaR's rapid success, less research has been done on its out-of-sample prediction performance. Interestingly, Plagborg-Møller et al. (2020) showed that financial variables contribute very little to GDP forecast distributions and none of the predictors they consider provide a robust signal of future tail risks. Along the same lines, Brownlees and Souza (2021) demonstrated that, when compared to the quantile regression method, fitting a simple AR(1)-GARCH(1,1) model using 24 OECD nations yields better in-sample and out-of-sample results. Thus, confirming that business cycle and real indicators are the main driver to forecast potential tail risks. Finally, Carriero et al. (2020) model the GDP conditional predictive distribution using a Bayesian VAR which features a generalized factor structure in the stochastic volatility to capture macroeconomic uncertainty.
By using a GARCH(1,1)-CHDFM we retain the strengths of the AR(1)-GARCH(1,1) model proposed by Brownlees and Souza (2021), but we expand the framework, enabling a factor structure in the variance. In particular, we acknowledge the existence of one unobserved common factor among the OECD countries that drives the mean process and we complement it with additional idiosyncratic errors that may better explain the variance. This is done by extending the model in Chapter 1 to account for a factor structure enclosed in the

observation errors. Then, we forecast the the GDP distribution at time $t+1$. A similar approach, the use of a common time-varying variance component in the observation disturbances using a GARCH(1,1) specification, has been pursued by Koopman et al. (2010) in an attempt to model the term structure of interest rates. However, the work does not rely on the infinite n framework proper of DFM and uses approximation to evaluate the variance process. Lastly, we backtest the the GaR predictions, using a variety of methods employed in the risk management literature. To determine if the GaR predictions are effective with regard to various information sets, we use the dynamic quantile test developed by Engle and Manganelli (2004) and other statistical tests elaborated by Christoffersen (1998). Additionally, the tick loss, a loss function frequently used to gauge the accuracy of VaR predictions, is employed to examine the marginal GaR projections (Giacomini and Komunjer, 2005). We also compare CHDFM with the historical unconditional distribution of GDP growth rates.

The contribution of this section is to extend the theoretical framework from the previous chapter to better align the model to the empirical task at hand, i.e. correctly backtesting the GDP growth worst-case scenario. Specifically, we allow for unobserved heteroskedasticity in the measurement error, too. The key learning from the exercise is that the CHDFM better fits the data and demonstrates superior performance when compared to the benchmark methodology.

The first part of this chapter is organized as follows. Subsection 2.1.2 introduces the main modification to the CHDFM to account for heteroskedasticity in the observation equation. Subsections 2.1.3 and 2.1.4 present a mathematical formulation of the GaR and related backtesting methodologies. Subsection 2.1.5 describes the data for the empirical application, which consists of GDP growth rates for 20 OECD countries. Finally, Subsections 2.1.6 and 2.1.7 evaluate the in-sample efficacy and out-of-sample performance of the model.

2.1.2 CHDFM with idiosyncratic GARCH(1,1)

Let us extend the Conditionally Heteroskedastic model of 1.5 and 1.6 to take into account potential volatility dynamics of the observational error:

$$\boldsymbol{x}_t = \underbrace{\begin{bmatrix}\boldsymbol{\Lambda}_F & \mathbf{0} & \boldsymbol{\Lambda}_\xi\end{bmatrix}}_{\boldsymbol{\Lambda}^\dagger} \mathbf{F}_t^\dagger + \boldsymbol{\xi}_t^*, \tag{2.1}$$

$$\mathbf{F}_t^\dagger = \begin{bmatrix}\mathbf{F}_t \\ \boldsymbol{\eta}_t \\ \boldsymbol{\xi}_t\end{bmatrix} = \underbrace{\begin{bmatrix}\boldsymbol{\Phi} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix}}_{\boldsymbol{\Phi}^\dagger} \underbrace{\begin{bmatrix}\mathbf{F}_{t-1} \\ \boldsymbol{\eta}_{t-1} \\ \boldsymbol{\xi}_{t-1}\end{bmatrix}}_{\mathbf{F}_{t-1}^\dagger} + \underbrace{\begin{bmatrix}\mathbf{I} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\end{bmatrix}}_{\boldsymbol{\Psi}^\dagger} \underbrace{\begin{bmatrix}\boldsymbol{\eta}_t^* \\ \boldsymbol{\eta}_t \\ \boldsymbol{\xi}_t\end{bmatrix}}_{\boldsymbol{\eta}_{t-1}^\dagger}. \tag{2.2}$$

with $\boldsymbol{\xi}_t^* \sim \mathcal{N}(\mathbf{0}, \mathbf{H}^*)$. Matrices $\boldsymbol{\Phi}^\dagger$ and $\boldsymbol{\Psi}^\dagger$ are now both of dimension $(2r+m)\times(2r+m)$ where m is the dimension of the vector $\boldsymbol{\xi}_t$. $\mathbf{F}_t^\dagger$ is the $(2r+m)\times 1$ augmented unobserved

state vector and $\boldsymbol{\eta}_t^\dagger$ is the $(2r+m)\times 1$ disturbance component consisting of $\boldsymbol{\eta}_t^\star$, $\boldsymbol{\eta}_t$ and $\boldsymbol{\xi}_t$. The first two moments are given by

$$\boldsymbol{\eta}_t^\dagger|\mathcal{I}_{t-1} \sim \mathcal{N}\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{Q}^\star & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}_t & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{H}_t\end{pmatrix}\right].$$

In this way we can model the dynamic of the conditional variance of both $\boldsymbol{\eta}_t$ and $\boldsymbol{\xi}_t$. For $i = 1, ..., r$ and for $j = r+1, ..., m$ we have:

$$\boldsymbol{\eta}_t = \mathbf{Q}_t^{1/2}\widetilde{\boldsymbol{\eta}}_t \qquad q_{i,t} = \omega_i + \alpha_i\eta_{i,t-1}^2 + \beta_i q_{i,t-1} \tag{2.3}$$

$$\boldsymbol{\xi}_t = \mathbf{H}_t^{1/2}\widetilde{\boldsymbol{\xi}}_t \qquad h_{j,t} = \omega_j + \alpha_j\eta_{i,t-1}^2 + \beta_j h_{i,t-1} \tag{2.4}$$

where $h_{jt} = \mathbf{H}_t^{[j,j]}$, $q_{i,t} = \mathbf{Q}_t^{[i,i]}$, and with $\widetilde{\boldsymbol{\eta}}_t, \widetilde{\boldsymbol{\xi}}_t, \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and independent. Noting that this specification relies on the same general model defined in 1.5 - 1.6, then 1 still holds and the Kalman filter delivers consistent factors (and errors). In particular, denoting $\eta_{t-1|t-1} = \mathbb{E}[\eta_{t-1}|\mathcal{X}_{t-1}]$ and $P^\eta_{t-1|t-1} = \mathbb{E}[(\eta_{t-1} - \eta_{t-1|t-1})^2|\mathcal{X}_{t-1}]$ and $\xi_{t-1|t-1} = \mathbb{E}[\xi_{t-1}|\mathcal{X}_{t-1}]$ and $P^\xi_{t-1|t-1} = \mathbb{E}[(\xi_{t-1} - \xi_{t-1|t-1})^2|\mathcal{X}_{t-1}]$ we have that

$$q_{t|t-1} = \omega_i + \alpha_i(\eta^2_{t-1|t-1} + P^\eta_{t-1|t-1}) + \beta_i q_{t-1|t-2} \tag{2.5}$$

$$h_{t|t-1} = \omega_j + \alpha_j(\xi^2_{t-1|t-1} + P^\xi_{t-1|t-1}) + \beta_j h_{t-1|t-2}. \tag{2.6}$$

The estimation is performed in the same way as described in Section 1.25, with the CM-step having $r+m$ optimization routines.

2.1.3 Evaluating the GaR

Recently, economists and policymakers have concentrated their efforts on modelling and predicting the marginal and joint distributions of GDP growth rates in order to quantify the negative risk associated with extreme events in the 5% conditional quantile, named Growth-at-Risk. Define as $x_{i,t}$ the i^{th} country's GDP growth rate, then the h-step ahead GaR is defined as the maximum loss that can occurr with a given degree of certainty p and such that:

$$\mathbb{P}_t(x_{i,t+h} \leq GaR^p_{i,t+h|t}) = p \tag{2.7}$$

with $p = 0.05$ and $\mathbb{P}_t$ the probability measure conditional on the information available up to time t. This can be rewritten as the p-quantile of the GDP conditional distribution:

$$GaR^p_{i,t+h|t} = \mathbb{F}^{-1}_{x_{i,t+h}}(p) \tag{2.8}$$

$\mathbb{F}^{-1}_{x_{i,t+h}}(p)$ being the cumulative density function of $x_{i,t+h}$. The $(1-p)\%$ marginal GaR is the prediction distribution area that should contain the GDP growth of each country with that probability. As Brownlees and Souza (2021) demonstrated, modelling the variance of an

autoregressive process could generate superior results both in and out of sample compared to a model which included external financial regressors.
As we will deal with dynamic factors that model all the variables jointly we will use the "joint marginal" approach, which employs the prediction region obtained by setting the joint GaR equal to the marginal GaR.
This section will introduce the historical approach plus two models that incorporate similar (but not exactly equal) assumptions about the data generation process and the determinants of GDP conditional distributions but different estimation procedures.

1) The Historical Benchmark is a non-parametric estimation method for quantiles, and GaR is calculated as the sample quantile estimate based on historical GDP growth rates. Given n ordered data points for the country i, $x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}$, ordered from the smallest to the largest, the Empirical Cumulative Density Function (ECDF) is defined as:

$$\widehat{\mathbb{F}}_n(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{x_i^{(j)} < x_i\right\}} \tag{2.9}$$

where $\mathbb{1}$ is the indicator function of the event $\{x_i^{(j)} < x_i\}$. Assuming each observation in time is *i.i.d* we can remove any dependency of t from $x_{i,t}$. Then the GaR can be calculated by the inverse function of $\widehat{\mathbb{F}}_n(x_i)$, for a given confidence level p:

$$GaR^p_{i,t|t+h} = GaR^p_i = \widehat{\mathbb{F}}^{-1}_{n,x_i}(p). \tag{2.10}$$

2) The second model employs the CHDFM with idiosyncratic error as defined in Section 2.1.2. Using that specification it is possible to construct the GaR forecast at time $t+1$ using the $p\%$ conditional quantile

$$GaR^p_{i,t+1|t} = y_{i,t+1|t} + \sqrt{\sigma^2_{i,t+1|t}}\,\widehat{\mathbb{F}}^{-1}_{z_i}(p). \tag{2.11}$$

Let us focus on the first two elements. Forecast distribution formulas are defined in 1.4.2. In particular, for the model defined in (2.1) and (2.2) and $h = 1$ we have

$$y_{i,t+1|t} = \boldsymbol{\lambda}_i^F \boldsymbol{\Phi} \mathbf{F}_{t|t} \tag{2.12}$$

$$\sigma_{i,t+1|t} = \boldsymbol{\lambda}_i^F (\boldsymbol{\Phi}\mathbf{P}_{t|t}\boldsymbol{\Phi}' + \mathbf{Q}_{t+1|t} + \mathbf{Q}^*)\boldsymbol{\lambda}^{F\prime}_i + \boldsymbol{\lambda}^{\xi}_i \mathbf{H}_{t+1|t}\boldsymbol{\lambda}^{\xi\prime}_i + \mathbf{H}^*, \tag{2.13}$$

where the diagonal elements of $\mathbf{Q}_{t+h|t}$ and $\mathbf{H}_{t+h|t}$ are given by (2.5) and (2.6).[1] In this way one can decompose the country conditional variance into two elements: a

[1] Although the analysis focuses on one-step-ahead prediction, multi-step forecasting is also possible either by recursively updating the *prediction equations* or using the formulas in 1.4.2. An h-step forecast of the conditional variances would be

$$q_{i,t+h|t} = 1 - \phi_i^2 + (\alpha_i + \beta_i)^{h-1}(q_{i,t+1|t} - (1-\phi_i^2)) \tag{2.14}$$

$$h_{j,t+h|t} = 1 + (\alpha_j + \beta_j)^{h-1}(h_{i,t+1|t} - 1) \tag{2.15}$$

for $i = 1, ..., r$ and for $j = r+1, ..., m$.

driving factor common component plus an observational error component.
Lastly, $\widehat{\mathbb{F}}_{Z_i}^{-1}(p)$ is the inverse cumulative distribution function of z_i, that is the standardized innovations $z_{i,t} = e_{i,t}/\sigma_{i,t}$ distributed according to $z_{i,t} \sim NID(0,1)$.[2]
For the probability density function (p.d.f.) of z_i we will compare both the quantiles from the Normal $\mathcal{N}(0,1)$ and the p%-quantiles obtained from the ECDF of z_i, $\widehat{\mathbb{F}}_{n,z_i}^{-1}(p)$.

3) The third approach is a general two-stage principal component in which one estimates the factor first (1.31) and then estimates the GARCH parameters afterwards (1.32). Subsequently, one can iterate variance prediction using standard GARCH(1,1) theory to obtain forecasts.
There is a small difference in the data generating process between 2) and 3), as the latter does not distinguish between r, the number of driving factors, and m, the number of shared idiosyncratic errors.

The models are subsequently compared using backtesting.

2.1.4 Backtesting the GaR

This section will discuss the tests that we are using to rate the accuracy of the GaR models presented previously. In terms of the in-sample backtesting exercise, we will rank all of the Growth-at-Risk models for each country using four separate criteria.
The first is based on the individual absolute number of violations. Define the function $I_{i,t}$ as the 'hit' indicator

$$I_{i,t} = \begin{cases} 1 & \text{if } x_{i,t} < GaR^{p}_{i,t|t-1} \\ 0 & \text{otherwise.} \end{cases} \tag{2.16}$$

Then, the empirical coverage metric (Kupiec, 1995) is the average proportion of Growth-at-Risk violations that occurred across the entire sample:

$$\widehat{\pi}_i = \frac{1}{T}\sum_{t=1}^{T} I_{i,t}. \tag{2.17}$$

The cross-sectional average can be easily derived and equals $\widehat{\pi} = n^{-1}\sum \widehat{\pi}_i$. If the distribution of the GDP growth rate is correctly specified, then the number of violations $\widehat{\pi}_i$ of a country i should be equal to the confidence level p. Since each 'hit' can be considered an i.i.d. Bernoulli(π) sequence, it is possible to perform the unconditional coverage tests, whose null hypothesis states that the empirical coverage is not statistically different from p. In a sample T, indicate with T_1 the number of violations and $T_0 = T - T_1$. Then, the test statistics

$$\mathrm{LR}_{uc} = -2\log\left(\frac{p^{T_1}(1-p)^{T_0}}{\widehat{\pi}_i^{T_1}(1-\widehat{\pi}_i)^{T_0}}\right) \sim \chi^2_1 \tag{2.18}$$

[2] $e_{i,t}$ are the innovations from the Kalman filter, $e_{i,t} = x_{i,t} - \boldsymbol{\lambda}\mathbf{F}_{t|t-1}$.

are distributed as χ^2_1 with one degree of freedom. The higher the values of LR, the more unlikely is the null hypothesis.
The independence of exceedances should also be a feature of a successful VaR model. The Christoffersen (1998) test also verifies the independence of those exceedances. When the model appropriately predicts the VaR, an exception today should not be affected by whether or not one happened the day before. Dropping the subscript i to avoid notational burden, the sequence of exceedances is represented using a first-order Markov chain with a matrix of transition probabilities

$$\widehat{\Pi}_1 = \begin{bmatrix} \widehat{\pi}_{00} & \widehat{\pi}_{01} \\ \widehat{\pi}_{10} & \widehat{\pi}_{11} \end{bmatrix} \tag{2.19}$$

with $\widehat{\pi}_{jk} = T_{jk}/(T_{j0} + T_{j1}),\ j,k = 0,1$, where T_{jk} is the number of days when condition j occurred assuming that k occurred the previous day. If the hits are independent across time, the probability of a violation tomorrow is independent of whether or not there was a violation today, thus $\widehat{\pi}_{01} = \widehat{\pi}_{11} = \widehat{\pi}$. The transition matrix becomes

$$\widehat{\Pi} = \begin{bmatrix} 1-\widehat{\pi} & \widehat{\pi} \\ 1-\widehat{\pi} & \widehat{\pi} \end{bmatrix}. \tag{2.20}$$

Assuming a sample T, one can test the independence hypothesis using the likelihood ratio test

$$\mathrm{LR}_{ind} = -2\log\left(\frac{\widehat{\pi}_{01}^{T_{01}}(1-\widehat{\pi}_{01})^{T_{00}}\widehat{\pi}_{11}^{T_{11}}(1-\widehat{\pi}_{11})^{T_{10}}}{\widehat{\pi}^{T_1}(1-\widehat{\pi})^{T_0}}\right) \sim \chi^1_2. \tag{2.21}$$

Finally, one can combine the two tests, i.e., the unconditional coverage test and the independence test, into the conditional coverage test to assess altogether $\widehat{\pi}_{01} = \widehat{\pi}_{11} = p$. The test statistics has the form

$$\mathrm{LR}_{cc} = \mathrm{LR}_{uc} + \mathrm{LR}_{ind} \sim \chi^2_2 \tag{2.22}$$

and has the asymptotic chi-square distribution with two degrees of freedom.
One of the limits of Christoffersen's test is that it only controls for the independence of the first exceedance. For this reason, we employ the the dynamic quantile test of Engle and Manganelli (2004) as a supplementary test to assess the absence of a serial correlation in the hit sequence. Given the sequence of violation I_t, the authors define a new variable $\tilde{I}_t = I_t - p$ and test the linear coefficients of

$$\tilde{I}_t = \gamma_0 + \sum_{c=1}^{C} \gamma_i I_{t-i} + \varepsilon_t. \tag{2.23}$$

The GaR is correctly estimated if the coefficients of the regression in 2.23 are zero. Thus, the dynamic quantile test is based on testing the null $H_0 : \gamma_0 = ... = \gamma_C = 0$ versus the alternative $H_1 : \gamma_c \neq 0$ for some $c = 0, ..., C$. We will use a number of lags equal to $C = 4$.

Using matrix notation, define $\widehat{\boldsymbol{\gamma}} = [\widehat{\gamma}_0, \cdots, \widehat{\gamma}_C]'$, the $(C+1) \times 1$ vector of linear regression coefficients, and $W = [\mathbf{1}, I_t, \cdots I_{t-C}]$, the $(C+1) \times (T-C)$ matrix of lagged hits, then the test statistics is given by

$$DQ_i = \frac{\widehat{\boldsymbol{\gamma}}' W_i' W_i \widehat{\boldsymbol{\gamma}}}{p(1-p)} \sim \chi^2_{C+1} \tag{2.24}$$

and has an asymptotic chi-square distribution with $C+1$ degrees of freedom.
Finally, GaR forecasts are evaluated using the tick loss function (González-Rivera et al., 2004; Giacomini and Komunjer, 2005) defined as

$$\mathrm{TL}_i = \frac{1}{T} \sum_{t=1}^{T} (p - I_{i,t})(x_{i,t} - GaR^p_{i,t|t-1}), \tag{2.25}$$

which penalises with weight $-(1-p)$ the violations of GaR while weighting the other cases with magnitude p. Lower absolute values of the metric identify correctly specified models.

2.1.5 Data Exploration

We define our analysis on the dataset used by Brownlees and Souza (2021), focusing on a total of 20, out of the 24, OECD countries. Most of those countries are located in Europe plus the United States and Canada.[3] The data consist of GDP growth rates for each country and spans from 1961Q2 to 2019Q3. The indicator is defined as the percentage changes from the previous quarter of seasonally adjusted real GPD, also known as GDP at constant prices or GDP in volume.[4] A main difference from the work of Brownlees is that we won't incorporate any exogenous predictor of downside risk.[5]
The dataset is pre-processed by winsorising extreme outliers, i.e. values that are less (greater) than the value at the 0.5th (99.5th) percentile of the whole dataset are replaced by the value at 0.5th (99.5th) percentile. The operation caps 48 observations out of 4,700. Then, the GDP vectors are individually demeaned.

2.1.6 In-sample Analysis

We begin the analysis by taking a look at the spectral decomposition of the covariance matrix. At first, we try to determine the number of factors in approximate factor models using a data-driven approach. In particular, we minimize the second information criteria

[3] Austria (AUT), Belgium (BEL), Canada (CAN), Denmark (DNK), Finland (FIN), France (FRA), Germany (DEU), Greece (GRC), Iceland (ISL), Ireland (IRL), Italy (ITA), Luxembourg (LUX), the Netherlands (NLD), Norway (NOR), Portugal (PRT), Spain (ESP), Sweden (SWE), Switzerland (CHE), the U.K. (GBR) and the U.S.A. (USA).

[4] Data is downloaded from the OECD website https://data.oecd.org/gdp/quarterly-gdp. By convention the base year is set to 2005.

[5] The model is still able to incorporate exogenous regressors, but here we rely on the endogenous factors' power to explain the data.

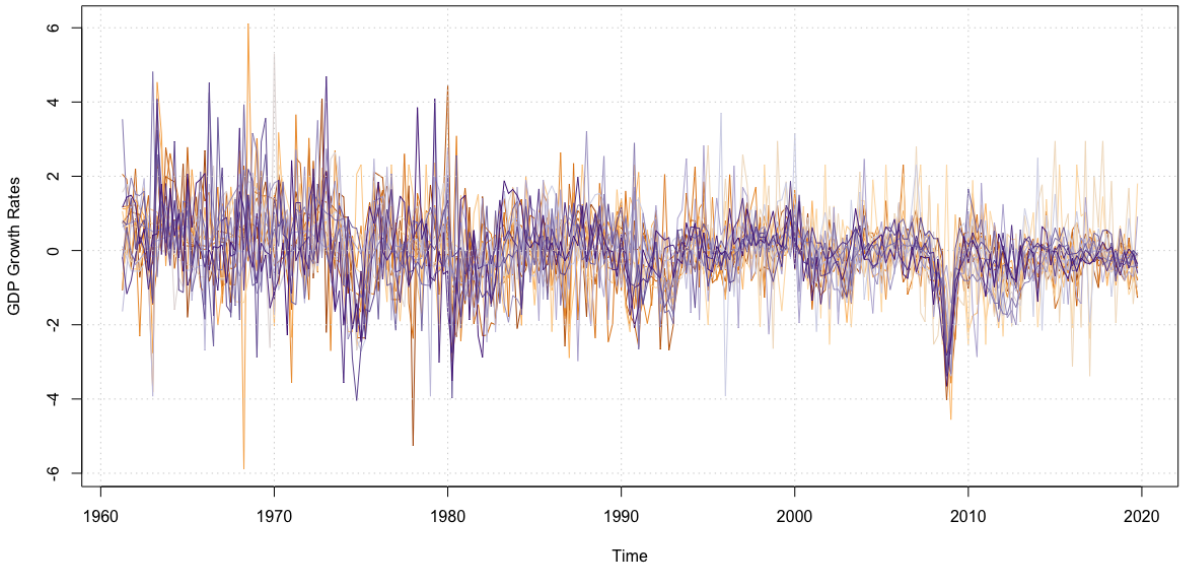

Figure 2.1: *Standardized time series of GDP growth rates for all the twenty countries.*

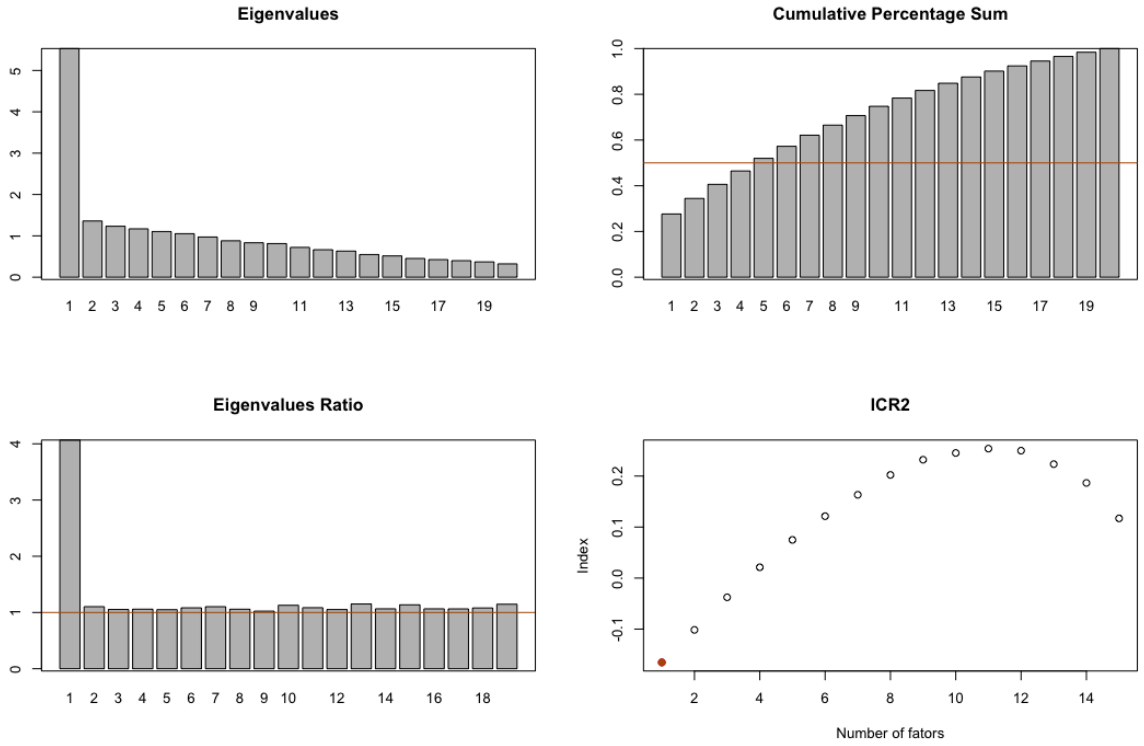

Figure 2.2: *Top left: Eigenvalues. Top right: Cumulative percentage sum of eigenvalues; horizontal line indicate 50%. Bottom left: Eigenvalue Ratio. Bottom right: Information criteria for determining the number of factors in a factors model.*

	2SPCA				CHDFM			
	q_1	q_2	q_3	q_4	q_1	h_2	h_3	h_4
ϕ	0.4440	-0.0716	-0.0022	0.0925	0.5308	-	-	-
	(0.0714)	(0.0801)	(0.0764)	(0.0813)	(0.0782)	( )	( )	( )
ω	0.0619	0.0687	0.1044	0.0381	0.0243	0.0572	0.0584	0.0241
	(0.0405)	(0.0703)	(0.0540)	(0.0260	(0.0216)	(0.0489)	(0.0349)	(0.0143)
α	0.1772	0.1553	0.2190	0.1268	0.2509	0.2126	0.3055	0.3167
	(0.0657)	(0.0894)	(0.0893)	(0.0557)	(0.0767)	(0.1064)	(0.0959)	(0.0961)
β	0.7533	0.7791	0.6933	0.8394	0.7153	0.7302	0.6361	0.6592
	(0.0794)	(0.1431)	(0.0944)	(0.0751)	(0.0717)	(0.1395)	(0.0870)	(0.0944)

Table 2.1: QML estimates for the CHDFM and 2SPCA models. Whole sample period. Standard errors are indicated in parenthesis.

(IC2) in Bai and Ng (2002) since we work with $n \ll T$. Although these criteria consistently estimate r, one should not rely only on these indicators for small values of n as they generally tend to underestimate the number of factors in small samples. Results of the test are showcased in the bottom-right plot of Figure 2.2: IC2 indicates the presence of one leading factor. This result is supported by a visual inspection of the eigendecomposition. The left graphs of 2.2, which display the absolute values of the eigenvalues (top) and the ratios of two consecutive eigenvalues (bottom), indicate a clear spike on the first component, while the others are capped. For this reason we will select $r = 1$ for the rest of the exercise. As for the choice of m, we will adopt a more pragmatic approach: we define m such that the relative sum of explained variability, i.e. $r + m$, is around, but not over, 50%. The top right graph of Figure 2.2 indicates that four factors explain around half of the variability, without surpassing the threshold. This brings us to the choice of $r = 1$ and $m = 3$.
We then fit the model for both the 2SPCA and CHDFM. The first model is estimated without restricting $\phi_i = 0$, as we don't want to limit the parametrization form of the latent factors, and without *variance targeting*.[6] For the latter one we initialize with estimates from the 2SPCA and run the ECME algorithm with tolerance $\varphi = 10^{-3}$. Table 2.1 indicates parameter estimates for the two models. Numbers are comparable, but CHDFM estimates the mean persistence ϕ around 20% higher compared to the 2SPCA counterpart. As for the conditional variance of the driving factor, both models identify the main source of dynamic volatility to be past shocks as seen by the $\alpha = 0.25$, $\beta = 0.72$ and $\alpha = 0.18$, $\beta = 0.75$ for the CHDFM and 2SPCA, respectively. In general the two-stage approach puts more emphasis on the persistence of volatility clusters, as seen from over-

[6]Numeric issues arise when optimizing with variance targeting. The BFGS algorithm does not converge using this specification.

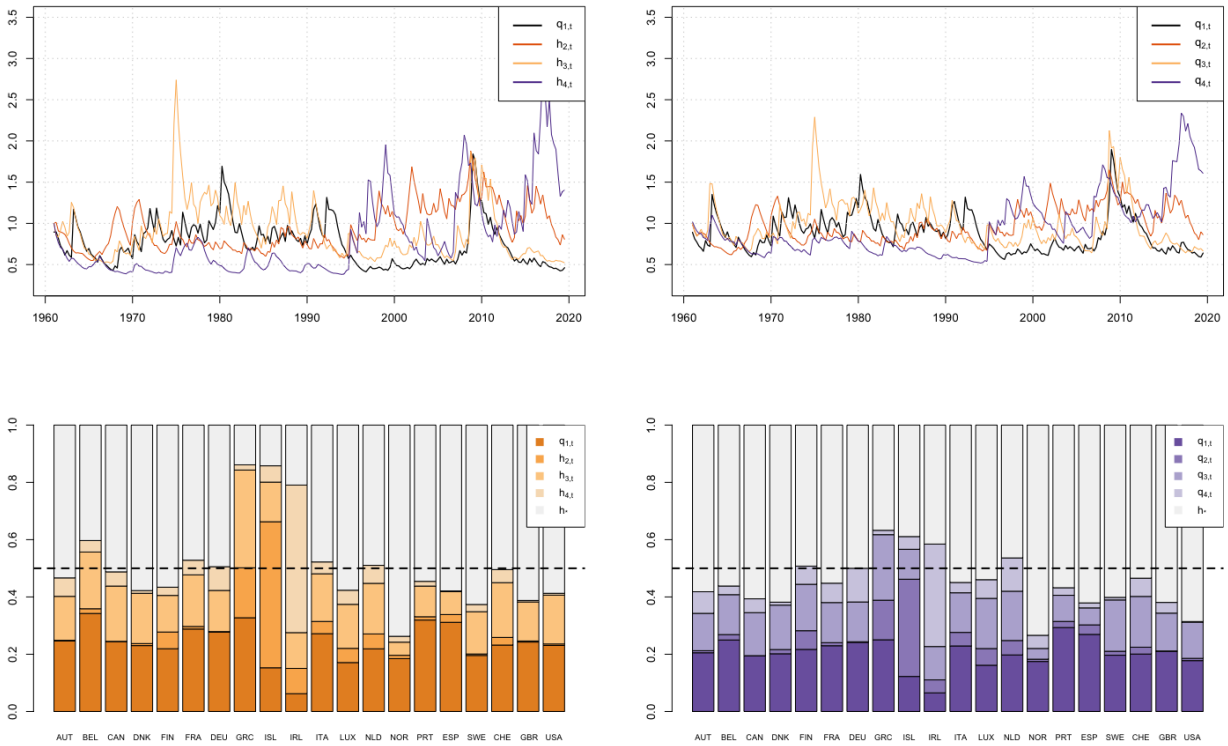

Figure 2.3: *Conditional volatilities for the four time-varying errors. Upper plots indicate volatilities evolution over time. Lower plot indicates averaged conditional volatility over T captured by each factor (coloured bars) divided by the total volatility, for each country. The dashed black line represents 0.5 of explained variability. Left is the CHDFM (orange) and right is 2SPCA (purple).*

all higher values of β and less on exogenous shocks, given the smaller α estimates. This behaviour is very similar to the one observed in the simulation framework of 1.6.1, with $T = 250, n = 50$. In particular, Figure 1.10 shows that 2SPCA tend to underestimate ϕ and α, while slightly overestimating β.

Standard errors for the parameters are shown in parenthesis and calculated using Shumway and Stoffer (2006) numerical procedure as described in 1.4.4 for the CHDFM, and through the common QMLE Hessian inversion for 2SPCA. (Francq and Zakoïan, 2010). The latter estimates, however, do not consider any uncertainty arising from the two-step approach. Factors are not treated as random variables but rather as fixed, thus potentially underestimating standard errors.

Since the factors are orthogonal it is possible to decompose every country's volatility as the sum of each factor's conditional volatility plus the idiosyncratic variability for each point in time. The bottom figures of 2.3 display how much of the country-conditional volatility can be explained by the model, with orange bars for the CHDFM and purple ones for the 2SPCA. Starting with the former, one can note that the main driving factor $q_{1,t}$, the

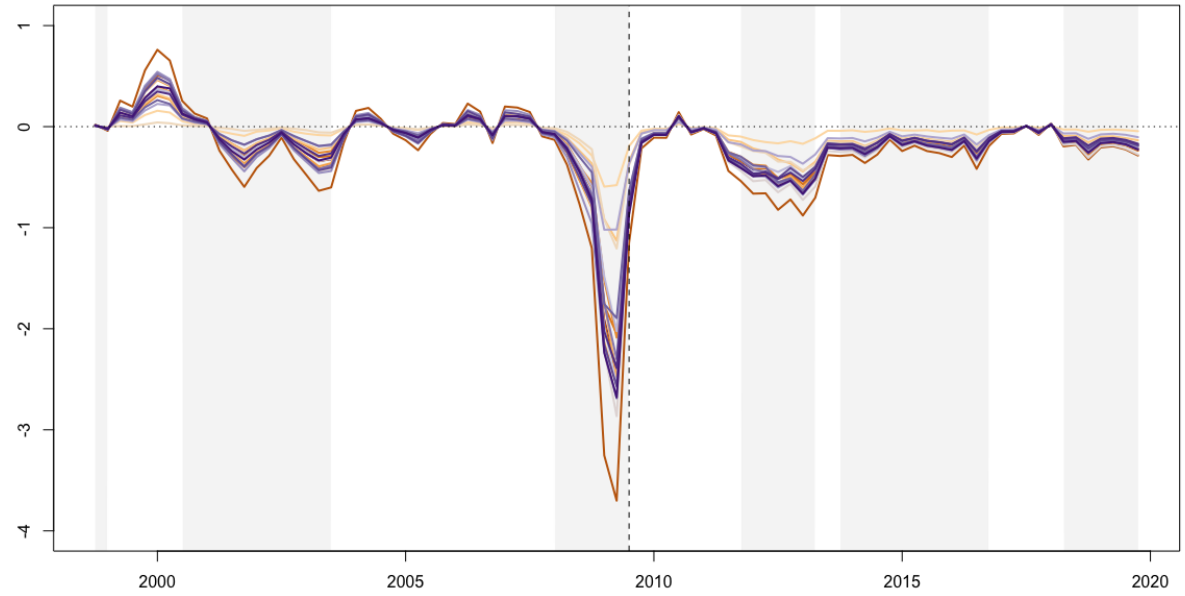

Figure 2.4: *One-step-ahead GDP growth forecast for every country in the sample. Shaded area indicates recession periods given by the OECD-based Recession Indicators.*

one with an autoregressive component, strongly represents countries such as Belgium and Greece (34% and 33%, respectively) but does so very little for countries such as Ireland and Iceland (6% and 15%, respectively). The factor-conditional variance is consistent across the whole time span and it is strongly affected by two major financial shocks: the early 1980s recession and the financial crisis of 2008 (top of Figure 2.3. The second factor volatility, $h_{2,t}$, mainly represents Iceland (51%) with minimal relation to the other countries. $h_{3,t}$ is spread uniformly over most of the OECD constituents, with Greece being the biggest weighting country. The conditional volatility plot also suggest that the bigger shocks occur during the late 1970, but the effect dissipates over time after until the surge of 2009. Finally, the fourth factor $h_{4,t}$ closely follows Ireland activity (51% of total variation) and puts emphasis on the volatility cluster that took place in the second half of the analysed data. Similar interpretations can be obtained for the 2SPCA model. Barplots in 2.6 and 2.5 showcase this decomposition in more details, with country average weightings for each decades instead of the whole period.

Figures 2.7 to 2.8 depict the GDP growth rates evolution over time for each country. Orange (purple) lines represent one-step-ahead mean prediction from CHDFM (2SPCA). Bottom grey lines indicate $GaR^{0.05}_{i,t|t+h}$ using EDF quantiles (light grey) and Normal quantiles (darker grey). Each country has around a 5% violation of $GaR^{0.05}_{i,t|t+h}$ as indicated by the black points. Backtesting results are outlined in Table 2.2. Results are very similar for both models. In terms of unconditional coverage, both models perform impressively, with none of the 20 countries in our sample exhibiting empirical coverage significantly lower than the 95 percent level, but with some increased clustered violations that indicate potential sequence dependence. Only one country (Italy) in the CHDFM approach reject the

independence hypothesis implied in the conditional test. Two countries, Italy and Canada, fail this test for the 2SPCA. The dynamic quantile test, which assesses the more general serial dependence structure of violations, performs better on the 2SPCA than the CHDFM with 50% and 45% cases not rejecting H_0, respectively. As for the absolute value of the tick loss, results are analogous. However, CHDFM demonstrate around a 9% improvement (reduction) in the loss function compared to the historical approach, while the 2SPCA approach showcase only a 6% reduction. Ireland, Belgium and Canada are the ones that benefit the most when heteroskedasticiy is taken into account.

2.1.7 Out-of-sample analysis

From 1961Q2 through 2019Q4, we iteratively estimate the two models and GARCH parameters using a rolling window of T= 150, and generate out-of-sample projections for the following quarter starting in 1998Q4. Our out-of-sample validation is based on about 65% of the available data since the forecasting exercise began in 1998Q4.
Figure 2.4 displays the one-step-ahead forecast of the GDP growth rates for each country. Given that the prediction step in the model is dictated only by the first factor whose $\phi_1 \neq 0$, each forecast GDP is proportional to F_1 by virtue of $\lambda_{1,i}$. The grey area in the graph represents periods of recession as given by the OECD-based Recession Indicators for OECD Economies.[7] As indicated by the graph, the model does a good job in recognizing world economy turning points. Table 2.3 showcases results of the backtesting analysis outlined in Section 2.1.4. Here, CHDFM demonstrates superior forecast ability. All countries pass conditional and unconditional tests. Additionally, the dynamic quantile test indicates no evidence of residual tail dynamics for each series after time-varying volatility has been accounted for. Tick loss is reduced by around 19% with respect to the historical benchmark.

[7]The time series is an interpretation of the OECD Composite Leading Indicators: Reference Turning Points and Component Series data, given by https://fred.stlouisfed.org/series/OECDNMERECDM. The dummy variable has a value of 1 in a recessionary period, while a value of 0 is an expansionary period.

	CHDFM					2SPCA				
	LR_{uc}	LR_{cc}	DQ	TL	Δ_{TL}	LR_{uc}	LR_{cc}	DQ	TL	Δ_{TL}
AUT	0.9406	0.8894	0.0427	0.1070	0.0202	0.9406	0.8894	0.0752	0.1077	0.0134
BEL	0.9406	0.0538	0.0032	0.0706	0.1801	0.9406	0.0538	0.0000	0.0757	0.1209
CAN	0.9406	0.0538	0.0001	0.0792	0.1653	0.9406	0.0000	0.0000	0.0836	0.1186
DNK	0.9406	0.8894	0.0976	0.1122	0.0416	0.9406	0.8894	0.0961	0.1123	0.0410
FIN	0.9406	0.3148	0.0794	0.1498	0.0606	0.9406	0.3148	0.1054	0.1513	0.0514
FRA	0.9406	0.8543	0.5787	0.0809	0.0966	0.9406	0.2723	0.1339	0.0820	0.0849
DEU	0.9406	0.8894	0.6171	0.1231	0.0295	0.9406	0.8894	0.6234	0.1231	0.0293
GRC	0.9406	0.0538	0.0070	0.2122	0.1216	0.9406	0.3148	0.2173	0.2191	0.0931
ISL	0.9406	0.5211	0.3818	0.2103	0.1251	0.9406	0.5211	0.2824	0.2180	0.0933
IRL	0.9406	0.8894	0.2163	0.1385	0.2888	0.9406	0.8894	0.0000	0.1751	0.1009
ITA	0.9406	0.0052	0.0002	0.0907	0.1325	0.9406	0.0052	0.0001	0.0926	0.1144
LUX	0.9406	0.3148	0.0000	0.1436	0.0714	0.9406	0.3148	0.0000	0.1458	0.0574
NLD	0.9406	0.5508	0.8463	0.1526	0.0293	0.9406	0.5508	0.8383	0.1544	0.0178
NOR	0.9406	0.8894	0.0171	0.1251	0.0045	0.9406	0.8894	0.0157	0.1257	-0.0004
PRT	0.9406	0.3148	0.0249	0.1087	0.1326	0.9406	0.3148	0.0006	0.1118	0.1083
ESP	0.9406	0.8894	0.1012	0.0875	0.1111	0.9406	0.8894	0.0525	0.0893	0.0925
SWE	0.9406	0.5211	0.2602	0.1406	0.0048	0.9406	0.5211	0.2562	0.1410	0.0017
CHE	0.9406	0.0538	0.0000	0.1147	0.0351	0.9406	0.0538	0.0000	0.1144	0.0378
GBR	0.9406	0.0538	0.0000	0.1041	0.0616	0.9406	0.0538	0.0000	0.1054	0.0494
USA	0.9406	0.3148	0.0006	0.0932	0.0530	0.9406	0.3148	0.0009	0.0945	0.0403
avg.	100%	**95%**	45%	**0.1222**	**0.0883**	100%	90%	**50%**	0.1261	0.0633

Table 2.2: Backtesting of in-sample GaR for CHDFM and 2SPCA. The first three columns indicate p-vaues of the respective statistics as described in Section 2.1.4. Fourth column describes the value of the tick loss, while the last one represents the improvement (in percentage reduction) with respect to historical benchmark GaR. Last row indicates how many countries have p-values over 5 %. Tick loss and percentage change in tick loss are displayed as averages across countries. Bold numbers indicate better metrics.

	CHDFM					2SPCA				
	LR_{uc}	LR_{cc}	DQ	TL	Δ_{TL}	LR_{uc}	LR_{cc}	DQ	TL	Δ_{TL}
AUT	0.0533	0.1526	0.4132	0.0820	0.1729	0.9000	0.8122	0.9479	0.0741	0.2523
BEL	0.2136	0.4394	0.9538	0.0801	0.2543	0.2087	0.2399	0.4918	0.0798	0.2565
CAN	0.5123	0.7219	0.4730	0.0800	0.2623	0.7162	0.5018	0.0001	0.0837	0.2280
DNK	0.0533	0.1526	0.3817	0.0927	0.1395	0.4109	0.4493	0.3884	0.0881	0.1819
FIN	0.5123	0.7219	0.4607	0.0861	0.1449	0.0944	0.0976	0.0003	0.1067	-0.0595
FRA	0.4109	0.5081	0.4141	0.0852	0.2629	0.0000	0.0000	0.0000	0.1313	-0.1363
DEU	0.2136	0.4394	0.0987	0.0829	0.1910	0.4109	0.5081	0.2321	0.0894	0.1276
GRC	0.0533	0.1526	0.4671	0.0935	0.1712	0.5123	0.1541	0.0011	0.0971	0.1396
ISL	0.2136	0.4394	0.8015	0.0988	0.1119	0.7162	0.0550	0.0002	0.1033	0.0712
IRL	0.9000	0.8122	0.0823	0.0938	0.2690	0.0383	0.0661	0.0000	0.1177	0.0833
ITA	0.9000	0.8122	0.7291	0.0831	0.2161	0.4109	0.5081	0.2677	0.0859	0.1896
LUX	0.0533	0.1526	0.3514	0.0927	0.1188	0.2087	0.3900	0.0048	0.0947	0.0999
NLD	0.0533	0.1526	0.4731	0.0805	0.1945	0.2087	0.1118	0.0153	0.0794	0.2051
NOR	0.0533	0.1526	0.4031	0.0987	0.0951	0.4109	0.0947	0.0114	0.0873	0.1997
PRT	0.0533	0.1526	0.3499	0.0794	0.2477	0.9000	0.8122	0.7755	0.0804	0.2383
ESP	0.4109	0.5081	0.3010	0.0832	0.2327	0.0944	0.0976	0.0000	0.0995	0.0828
SWE	0.2136	0.4394	0.1009	0.0884	0.1441	0.4109	0.0947	0.0000	0.0920	0.1089
CHE	0.9000	0.8122	0.7133	0.0834	0.1715	0.0944	0.0976	0.0087	0.0903	0.1026
GBR	0.0533	0.1526	0.3971	0.0807	0.1718	0.9000	0.3511	0.0175	0.0729	0.2518
USA	0.7162	0.6820	0.8400	0.0819	0.2193	0.0383	0.0661	0.0053	0.0965	0.0801
avg.	**100%**	**100%**	**100%**	**0.0863**	**0.1896**	95%	95%	30%	0.0925	0.1352

Table 2.3: Backtesting of out-of-sample GaR for CHDFM and 2SPCA. The first three columns indicate p-vaues of the respective statistics as described in Section 2.1.4. Fourth column describes the value of the tick loss, while the last one represents the improvement (in percentage reduction) with respect to historical benchmark GaR. Last row indicates how many countries have p-values over 5 %. Tick loss and percentage reduction in tick loss are displayed as averages across countries. Bold numbers indicate better metrics.

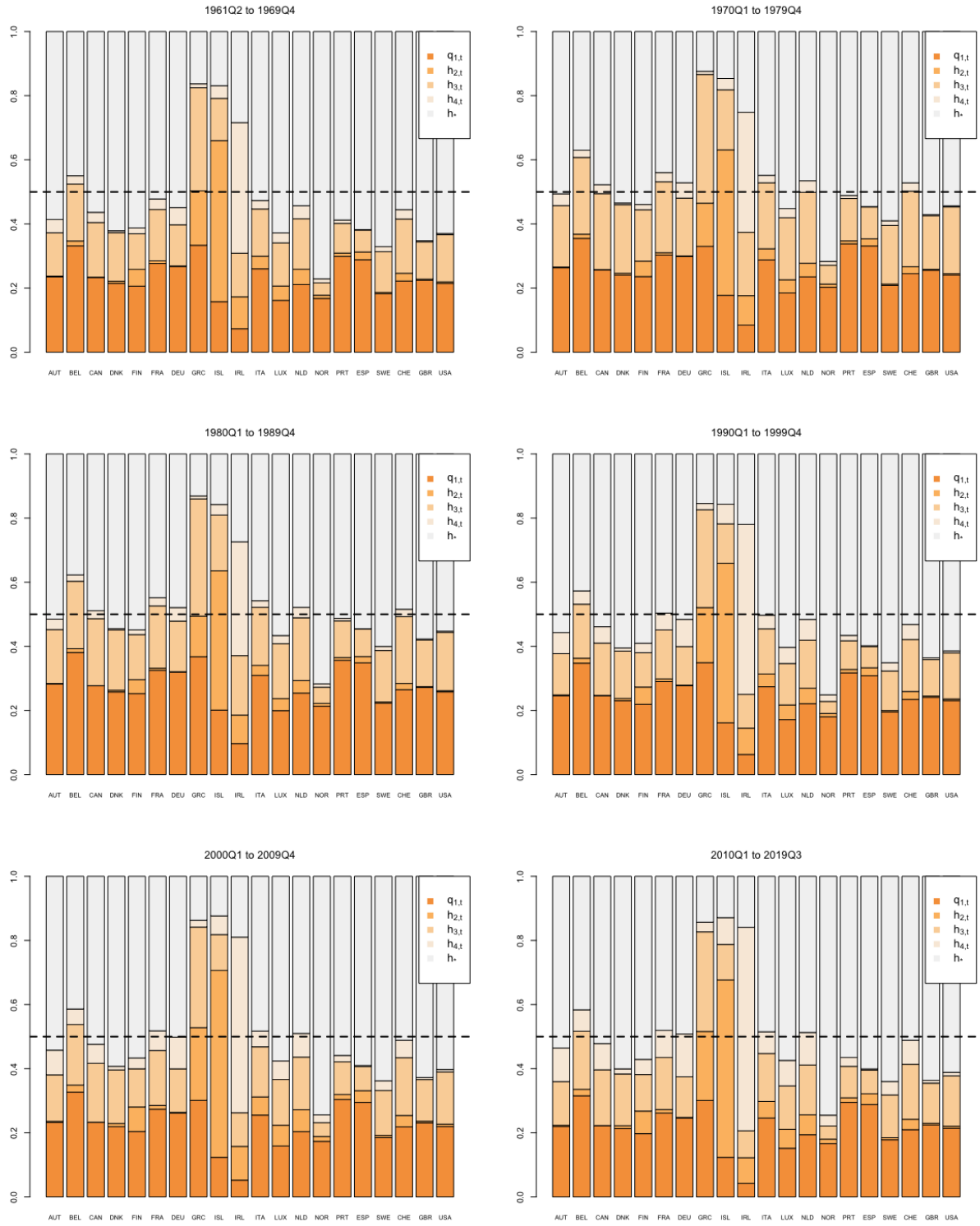

Figure 2.5: *Average percentage of explained variability in CHDFM by country. Orange bars indicates variability captured by each GARCH1(1,1) factor divided by the total volatility. Light purple bar shows constant idiosyncratic volatility. Quarterly data is aggregated in decades using sample means.*

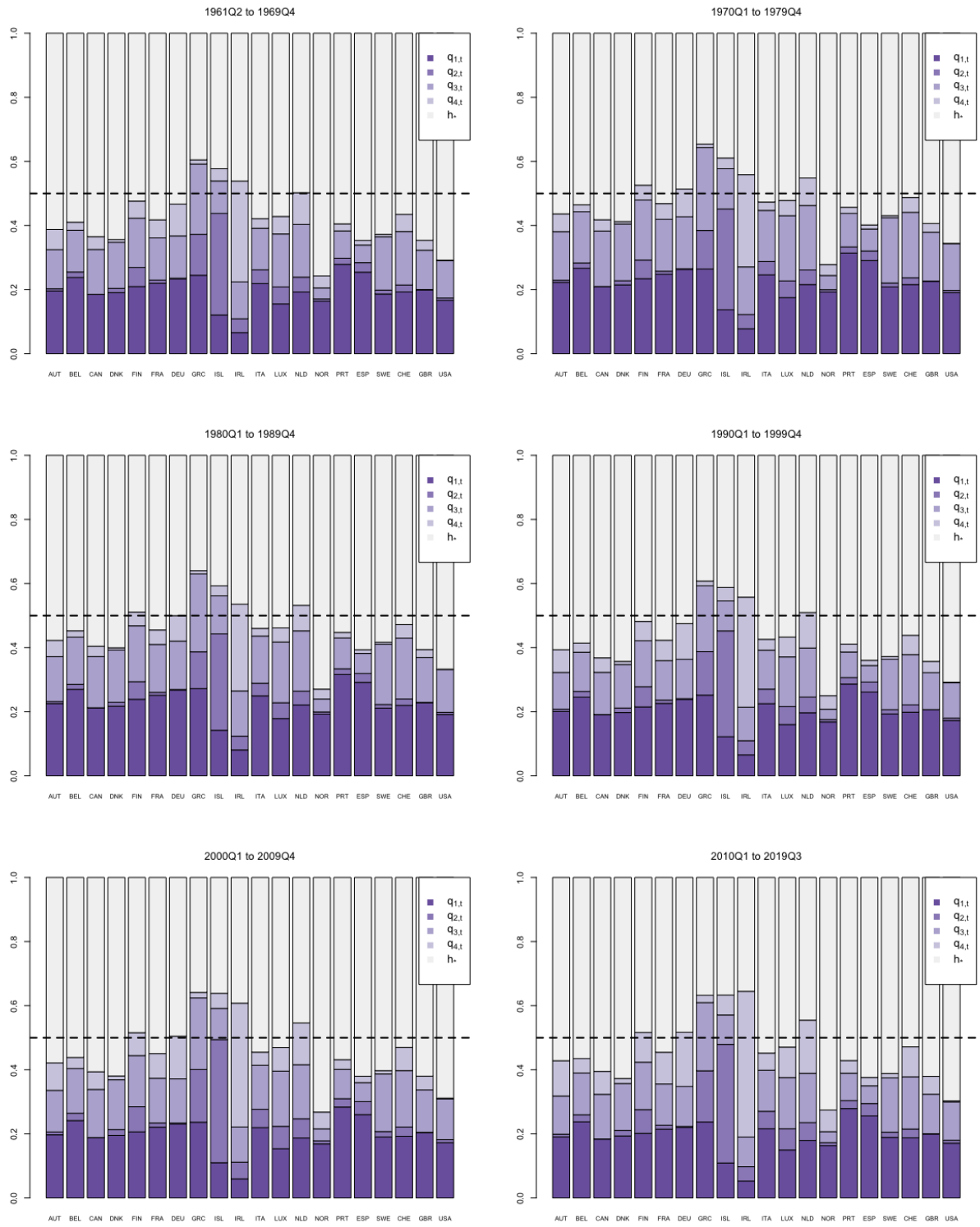

Figure 2.6: *Average percentage of explained variability in 2SPCA by country. Purple bars indicates variability captured by each GARCH1(1,1) factor divided by the total volatility. Light purple bar shows constant idiosyncratic volatility. Quarterly data is aggregated in decades using sample means.*

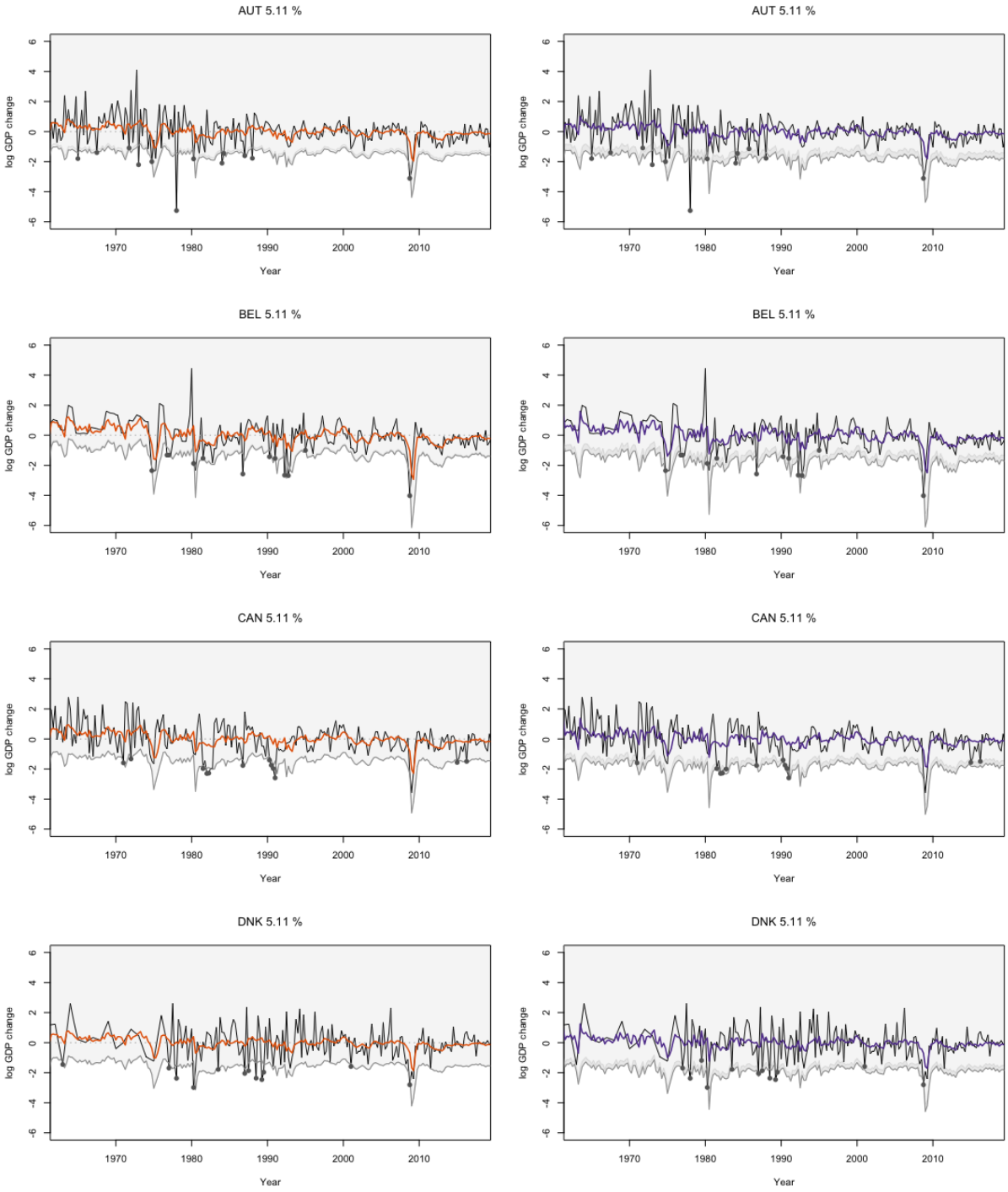

Figure 2.7: *GDP growth rate evolution over time for each country. Black lines are true values and orange (purple) lines represent one-step-ahead mean prediction from CHDFM (2SPCA). Bottom grey lines indicate* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles (light grey) and Normal quantiles (darker grey). Black points indicate violation of* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles. Numbers of total violations over the sample are indicated in the title in percentage form.*

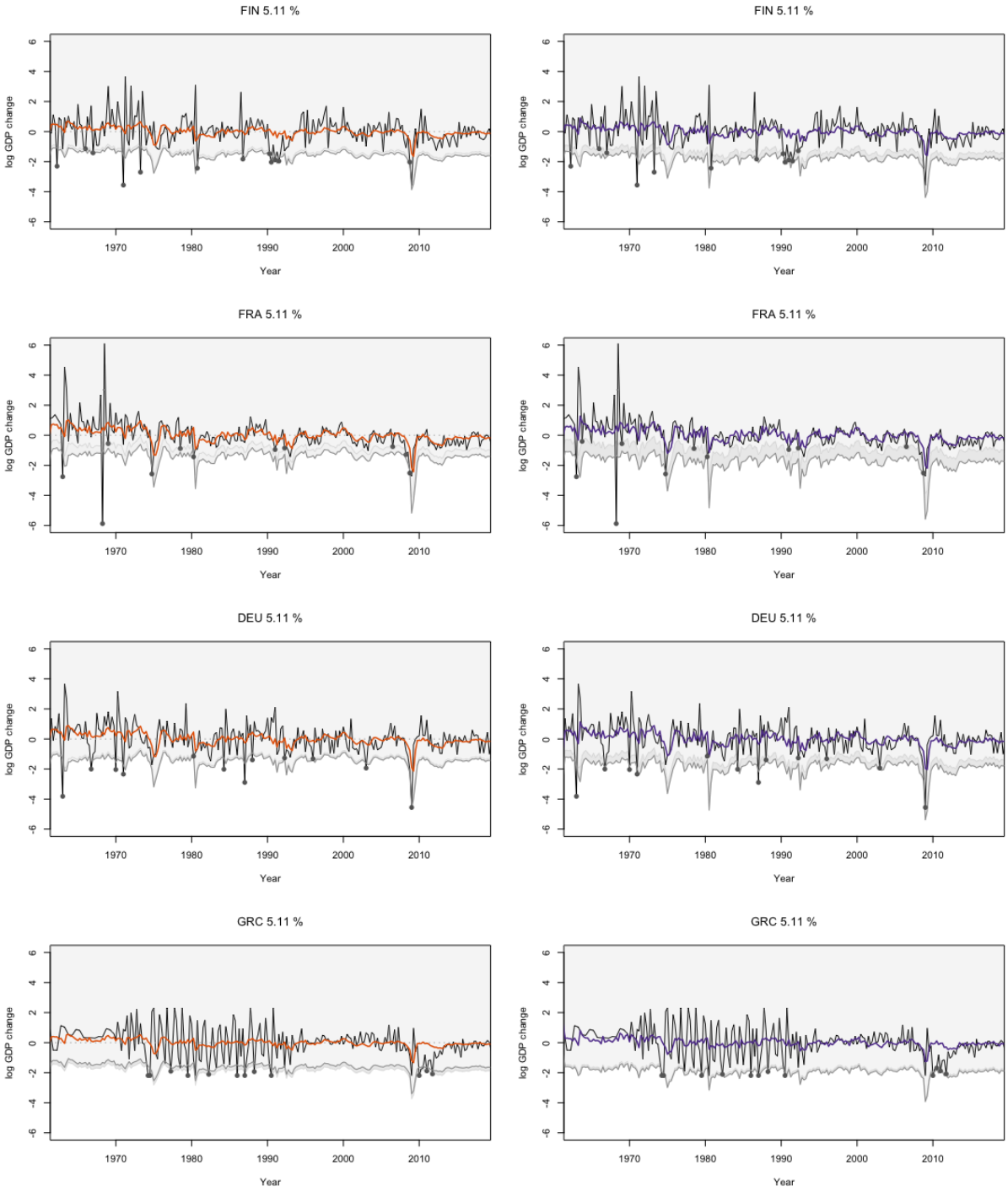

Figure 2.8: *GDP growth rate evolution over time for each country. Black lines are true values and orange (purple) lines represent one-step-ahead mean prediction from CHDFM (2SPCA). Bottom grey lines indicate* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles (light grey) and Normal quantiles (darker grey). Black points indicate violations of* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles. Numbers of total violations over the sample are indicated in the title in percentage form.*

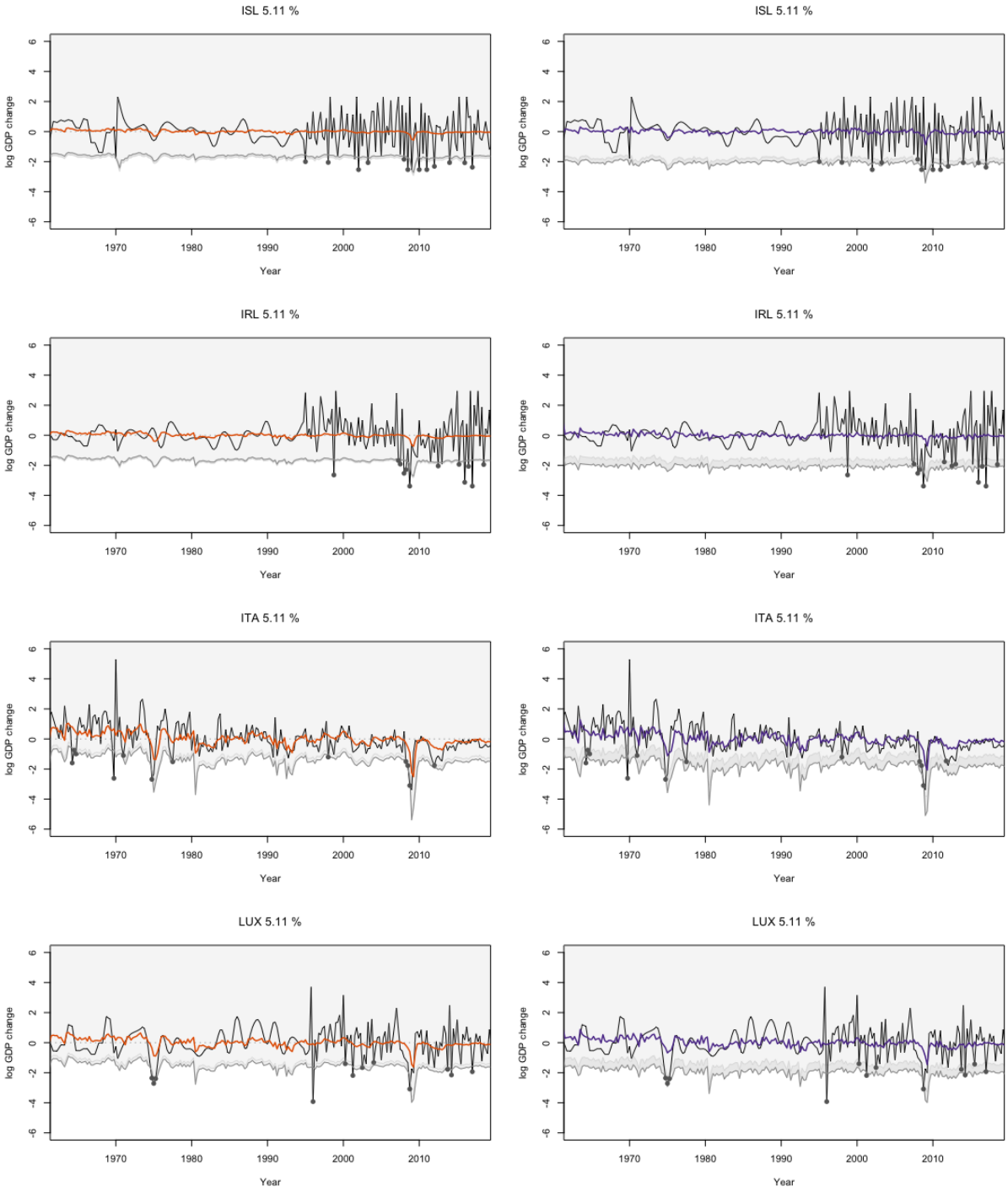

Figure 2.9: *GDP growth rate evolution over time for each country. Black lines are true values and orange (purple) lines represent one-step-ahead mean prediction from CHDFM (2SPCA). Bottom grey lines indicate* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles (light grey) and Normal quantiles (darker grey). Black points indicate violations of* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles. Numbers of total violations over the sample are indicated in the title in percentage form.*

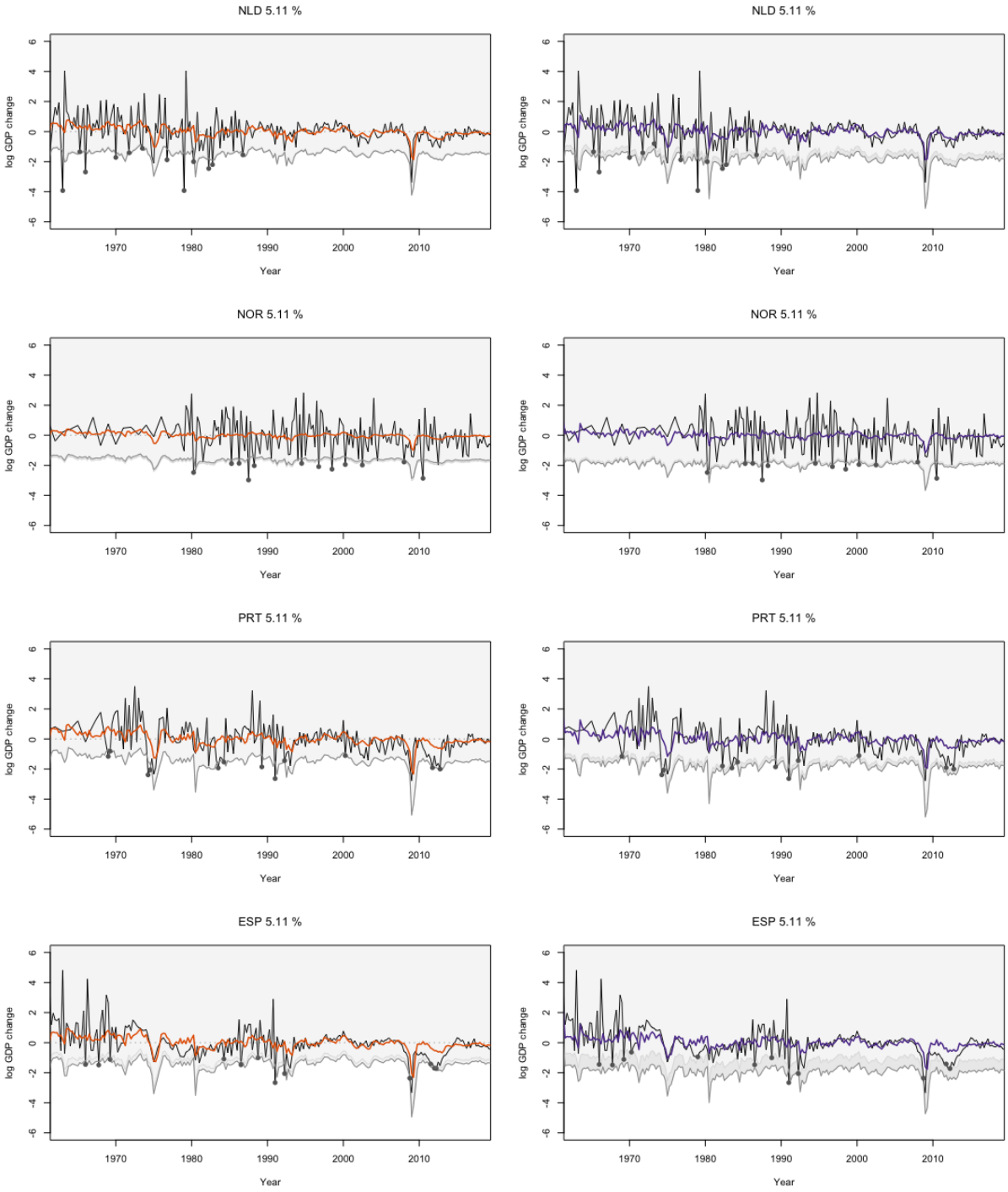

Figure 2.10: *GDP growth rate evolution over time for each country. Black lines are true values and orange (purple) lines represent one-step-ahead mean prediction from CHDFM (2SPCA). Bottom grey lines indicate* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles (light grey) and Normal quantiles (darker grey). Black points indicate violations of* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles. Numbers of total violations over the sample are indicated in the title in percentage form.*

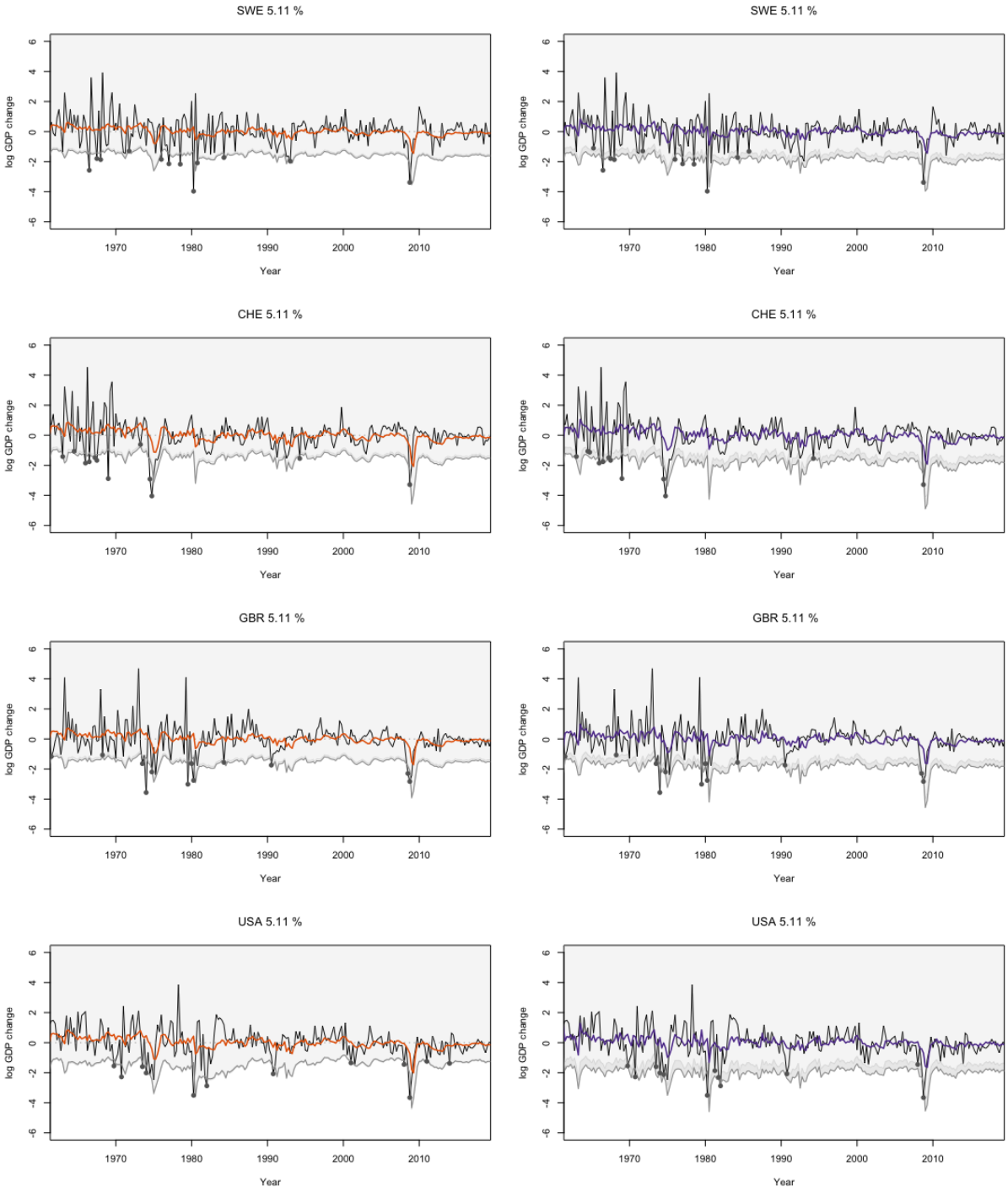

Figure 2.11: *GDP growth rate evolution over time for each country. Black lines are true values and orange (purple) lines represent one-step-ahead mean prediction from CHDFM (2SPCA). Bottom grey lines indicate* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles (light grey) and Normal quantiles (darker grey). Black points indicate violations of* $GaR^{0.05}_{i,t|t+h}$ *using EDF quantiles. Numbers of total violations over the sample are indicated in the title in percentage form.*

2.2 Minimum Variance Portfolio

2.2.1 Introduction

Most of the literature cited in Section 1.1 deals with theoretical specifications and different possible estimations of conditionally heteroskedastic factor models with a special focus on financial applications. Following the trend, in this second study, we employ the CHDFM to estimate the minimum variance portfolio (MVP). This issue has drawn considerable attention in recent years, especially within the factor analysis framework; see Fan et al. (2012), Trucíos et al. (2021), and Ding et al. (2021). Studying MVP is very important from an econometrics perspective because it provides a way to evaluate covariance matrix estimators and their forecasts. Furthermore, traditional mean–variance optimization requires knowledge of the mean, which is found to be more difficult to estimate than the covariance matrix.

Recent studies that deal with asset allocation, and specifically the MVP, rely on the acclaimed Dynamic Conditional Correlation (DCC) model of Engle and Sheppard (2001) to construct a time-varying covariance matrix. The approach consists of modelling separately the conditional variances and the conditional correlation matrix. The former are modelled using GARCH while the latter is modelled using the DCC model. A major component of the DCC is that the conditional correlation matrix is modelled as a function of the so-called *pseudo-correlation* matrix. This is a symmetric positive definite matrix that acts as a proxy of the true correlation matrix, but is not guaranteed to be unit-diagonal. The most efficient approach used in the literature is to rescale the pseudo-correlation matrix once it has been properly normalized in order to produce correlations. This drawback make it very difficult to define proper statistical proprieties and the asymptotic behaviour of the model (McAleer, 2019). Yet, the model is still used by practitioners for its wide applicability (De Nard et al., 2019; Engle et al., 2019; Brownlees and Engle, 2016), even though many modifications and corrections to the original specification have been developed over time (Tse and Tsui, 2002; Aielli, 2013; Brownlees and Llorens-Terrazas, 2022).

In some ways, DCC is less prone to the *curse of dimensionality* compared to multivariate GARCH but, in general, calculating the weight of a portfolio consisting of more than 30 assets is still a very demanding optimization problem. Together with shrinkage methods (Engle et al., 2019), factor models introduce a very natural way of tackling this issue. De Nard et al. (2019) study the usefulness of a factor structure in predicting large covariance matrices. The set of factors comprises both observed (Fama and French, 2015) and unobserved processes. The latter are, however, estimated through a two-step procedure which involves the extraction of latent factors from historical return data via PCA, a comparable approach to that of 2SPCA discussed in Chapter 1. A similar framework is adopted by Trucíos et al. (2021), in which the idiosyncratic conditional covariance is driven by a DCC and the common component is modelled as Mean GARCH (MGARCH). Estimation consists of multiple steps that take place in the frequency domain.

In this application we consider a modification of the GARCH-CHDFM, in which we allow the conditional covariance matrix of the state error to follow a DCC dynamic, i.e. to be

the product of two diagonal GARCH(1,1) standard deviation matrices and the DCC correlation matrix. Working with a pseudo-correlation matrix makes it infeasible to derive the Kalman filter prediction step of the correlation, as we did for the GARCH case. The issue and potential solution will be explored in the next section. Finally, the model is employed to forecast a covariance matrix and then construct optimal MVP weights.

This section further expands the theoretical framework from the previous chapter to better suit the CHDFM to a portfolio allocation problem. In particular, we relax the assumption of independent factors, allowing for a DCC-like structure. As a result, the unobserved components can be correlated and we can indeed analyse their time-varying conditional covariances and correlations. This exercise shows the high degree of flexibility of the model, and the higher potential performances with respect to the benchmark methodology.

This second part of the chapter will be structured as follow. Subsection 2.2.2 introduces the general model with Dynamic Conditional Correlation. Subsection 2.2.3 discusses the theoretical framework and examines some limitations of the approach together with potential solutions. Subsection 2.3.4 describes in more detail the structure of the model with two factors, outlining relevant numerical aspects. Subsection 2.3.5 illustrates portfolio allocation rule in the context of MVP and delineates optimal weights calculation. Subsection 2.3.5 explore the data set, which consists of more than 20 years of closing prices of the S&P500 constituents, and outlines the performance measure used to evaluate the different portfolios. Finally, Subsection 2.3.6 presents the main results and concludes.

2.2.2 CHDFM with Dynamic Conditional Correlation

Consider the approximate dynamic factor model as indicated in (1.5) - (1.6):

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\Lambda}\mathbf{F}_t + \boldsymbol{\xi}_t, \\ \mathbf{F}_t &= \boldsymbol{\Phi}\mathbf{F}_{t-1} + \boldsymbol{\eta}_t + \boldsymbol{\eta}_t^*. \end{aligned}$$

Conditional correlation models rely on the conditional covariance matrix decomposition so that standard deviations and correlation dynamics can be modelled separately. Following Engle (2002), the DCC structure can be introduced with the disturbance $\boldsymbol{\eta}_t$,

$$\boldsymbol{\eta}_t = \mathbf{S}_t^{1/2}\widetilde{\boldsymbol{\eta}}_{\boldsymbol{t}}, \tag{2.26}$$

$$\mathbf{S}_t = \mathbf{Q}_t^{1/2}\mathbf{C}_t\mathbf{Q}_t^{1/2}, \tag{2.27}$$

where $\mathbf{C}_t$ is the time-varying correlation matrix and $\mathbf{Q}_t^{1/2}$ is the diagonal matrix of standard deviations as defined in (1.4). Positive definiteness and restriction on the $\{-1,+1\}$ domain of $\mathbf{C}_t$ is achieved by modelling a proxy process, $\mathbf{R}_t$, as:

$$\mathbf{R}_t = (1-a-b)\overline{\mathbf{C}} + a\boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}^{'} + b\mathbf{R}_{t-1} \tag{2.28}$$

where $\boldsymbol{z}_t = [z_{1,t}, \cdots, z_{r,t}]'$ is the standardized disturbance vector calculated as $z_{i,t} = \eta_{i,t}/q_{i,t}$ and $\overline{\mathbf{C}}$ is a unit-diagonal positive definite matrix. Dividing each $\boldsymbol{\eta}_t$ by its conditional variance, we obtain unit standard deviation variables $\boldsymbol{z}_t$. In this way, modelling the correlations of $\boldsymbol{\eta}_t$ becomes equivalent to modelling the covariance of the standardized variables $\boldsymbol{z}_t$. The correlation matrix $\mathbf{C}_t$ is obtained by rescaling (2.28) such that,

$$\mathbf{C}_t = \text{diag}(\mathbf{R}_t)^{-1/2}\ \mathbf{R}_t\ \text{diag}(\mathbf{R}_t)^{-1/2}. \tag{2.29}$$

Within this framework, the correlation persistence parameters a and b are shared between factors. On the other hand, this does not imply that the level of correlation among factors (and eventually among observed variables) are the same at any time. Furthermore, persistence in volatility is unique for each $\mathbf{G}_t$.

2.2.3 A DFM setting for conditional correlation

Although DCC models are very much employed by practitioners when estimating dynamic conditional covariance (De Nard et al., 2019; Engle et al., 2019), some still question the statistical proprieties of the model; first, that the DCC does not have an underlying stochastic specification that could justify its derivation (see McAleer (2019) for both a discussion on the topics and references).
The first issue relates to the fact that $\mathbf{R}_t$ does not satisfy the definition of a correlation matrix and the authors refer to it as a *conditional pseudo-correlation matrix*, which is interpretable as the rescaled conditional covariance matrix of standardized residuals. For this reason, it is not possible to apply the Kalman Filter to calculate its moments in the same way as in the GARCH(1,1) case. Thus, we will refer to this model as an Approximate DCC-CHDFM and we will focus on its applicability, still achievable when $n \to \infty$, rather than its statistical proprieties.
Although there exists a comparable model for dynamic correlation by Tse and Tsui (2002) that directly computes $\mathbf{C}_t$, the recursion formula

$$\mathbf{C}_t = (1 - a - b)\overline{\mathbf{C}} + a\boldsymbol{\mathcal{C}}_{m,t-1} + b\mathbf{C}_{t-1}, \tag{2.30}$$

introduces $\boldsymbol{\mathcal{C}}_{m,t-1}$, the sample correlation matrix of $\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-m},\ m \geq r$, which is still extremely difficult to study, specifically when we deal with unobserved components. Furthermore, the fact that m is fixed and arbitrary and $\mathbf{C}_t$ is not the conditional expectation of $\mathcal{C}_{m,t}$, would make the analysis even more challenging.
Taking a look at (2.28), the first element that needs to be discussed is how to estimate $\boldsymbol{z}_t = \boldsymbol{\eta}_t/\boldsymbol{q}_t$. Here, we can adopt consistency results in Proposition 1 and Proposition 2 to calculate an approximate value of $\mathbb{E}[z_{i,t}|\mathcal{X}_t]$ given by

$$\dot{z}_{i,t|t} = \eta_{i,t|t}/q_{i,t|t-1}, \tag{2.31}$$

where $\eta_{i,t|t}$ and $q_{i,t|t-1}$ are the estimates obtained from the filter, as indicated in Section 1.3.1. It is important to note that the error is standardized by the conditional volatility

defined by $q_{i,t|t-1}$ and not $q_{i,t|t}$, as the latter quantity goes to 0 as $n \to \infty$, and is not representative of the true $q_{i,t}$. Furthermore, using the smoother, i.e. conditioning on $\mathcal{X}_T$, is not practically feasible as we need $z_{i,t|t}$ for the recursions.
Finally, the last matter relates to $\overline{\mathbf{C}}$. Aielli (2013) shows that the expression for $\overline{\mathbf{C}}$ does not match the unconditional variance of $\boldsymbol{z}_t$, thus it would be misleading to estimate the matrix using the sample second moment of the standardized return.[8] For this reason, we treat this as a unit-diagonal matrix, whose off-diagonal elements $\bar{\rho}_{i,j}$ are parameters to estimate together with a, b in the numerical optimization step in the ECME, i.e. $\bar{\rho}_{i,j} \in \boldsymbol{\theta}_{(\ell)}$ are the ones that maximize (1.29).

Some assumptions and identifying conditions set up in Section 1.2.2 and Section 1.4.1 need to be modified to take into account the correlation dynamic.

Assumption 1' (Dynamics) $\{\mathbf{F}_t\}$ *is a stationary process with* $\mathbb{E}[\mathbf{F}_t] = \mathbf{0}$ *and* $\mathrm{Var}[\mathbf{F}_t] < \infty$. *More specifically:* $\det(\mathbf{I}_r - \mathbf{\Phi} z) \neq 0$ *for all* $z \in \mathbb{C}$ *such that* $|z| \leq 1$. *Each element of* $\boldsymbol{\eta}_t = \mathbf{S}_t^{1/2} \widetilde{\boldsymbol{\eta}}_t$ *follows a DCC dynamic with* $\widetilde{\boldsymbol{\eta}}_t \sim NID(\mathbf{0}, \mathbf{I}_r)$ *and* $\mathbf{S}_t = \mathbf{Q}_t^{1/2} \mathbf{C_t} \mathbf{Q}_t^{1/2}$ *positive definite. Furthermore,* $\mathbf{Q}_t$ *is diagonal and* $\mathbf{C}_t$ *is unit-diagonal.* $\bar{\mathbf{R}}$ *is also assumed to be unit-diagonal. Finally,* $\omega_i, \alpha_i, \beta_i > 0$ *and* $\alpha_i + \beta_i < 1$, *for all* $i = 1, ..., r$, *and* $a, b > 0$ *and* $a + b < 1$.

Engle (2002) asserts that parameters a and b need to be constrained to ensure the system is stationary, thus we add it to the existing set of conditions.
Assumption (A1') also serves as an identifying restriction for the variance process. Assuming $\bar{\mathbf{R}}$ to be unit-diagonal is an over-identifying condition for the DCC, but comes naturally in this framework (Aielli, 2013).
Nonetheless, identifying condition (IC1) is not applicable in this setting since factors can now be cross-correlated and this requires a more stringent identification scheme to be imposed on $\mathbf{\Lambda}$ in order to relax restrictions on $\mathbf{\Omega}$. Specifically, we will leave factor variance unrestricted and operate on the loadings only.

Identifying condition 2 (IC2) $\mathbf{\Lambda} = \begin{bmatrix} \mathbf{I}_r & \mathbf{\Lambda}_0 \end{bmatrix}'$ *with* $\mathbf{I}_r$ *being the* $r \times r$ *identity matrix. Furthermore, the stochastic processes* $q_{i,t}$*'s for each* $i = 1, \cdots, r$ *are linearly independent, i.e.* $\nexists \boldsymbol{\delta} \in \mathbb{R}^r, \boldsymbol{\delta} \neq \mathbf{0} : \boldsymbol{\delta}' \boldsymbol{q}_t = 0 \ \forall t$.

This identification setup restricts the first $r \times r$ block of $\mathbf{\Lambda}$ to be an identity matrix. All r^2 restrictions are imposed on the loadings, leaving the factor process unrestricted. To guarantee factor existence, the only requirement is that $\mathbf{\Omega}$ is invertible. This structure,

[8] The author shows that, if $a + b < 1$ and that $\mathbb{E}[\mathbf{R}_t]$ and $\mathbb{E}[\boldsymbol{z}_t \boldsymbol{z}_t']$ are independent of t, then

$$\overline{\mathbf{C}} = \frac{1-b}{1-a-b} \mathbb{E}[\mathrm{diag}(\mathbf{R}_t)^{-1/2} \boldsymbol{z}_t \boldsymbol{z}_t' \mathrm{diag}(\mathbf{R}_t)^{-1/2}] - \frac{a}{1-a-b} \mathbb{E}[\boldsymbol{z}_t \boldsymbol{z}_t']. \tag{2.32}$$

however, affects the choice of the first r observed variables since they are going to be modelled as in the 'errors-in-variables' model of Pantula and Fuller (1986), $x_{it} = G_{it} + \xi_{it}$ for $i = 1, ..., r$. The reason for this choice is that this identifying scheme can easily be implemented in the PCA framework. In this way, we can have comparable estimates from a corresponding model based on PCA and, more importantly, we can initialize the ECME algorithm using estimators obtained from their PCA counterpart. Following Bai and Ng (2013), estimators for $\mathbf{\Lambda}$ and $\mathbf{F}_t$ can be obtained by

$$\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{\Lambda}}^{PCA}(\widehat{\mathbf{\Lambda}}_r^{PCA})^{-1}, \qquad \widehat{\mathbf{F}}_t = \widehat{\mathbf{F}}_t^{PCA}(\widehat{\mathbf{\Lambda}}_r^{PCA})', \tag{2.33}$$

where $\widehat{\mathbf{\Lambda}}^{PCA}$ and $\widehat{\mathbf{F}}_t^{PCA}$ are defined in (1.31) and $\widehat{\mathbf{\Lambda}}_r^{PCA}$ is the first $r \times r$ block of the loading matrix estimated through PCA.
In the second step, DCC parameters are estimated using Quasi Maximum Likelihood on the obtained $\widehat{\mathbf{F}}_t$

$$\widehat{\boldsymbol{\theta}}^{PCA} = \underset{\boldsymbol{\theta}}{\text{argmax}} \ \ell(\widehat{\mathbf{F}}_t; \widehat{\boldsymbol{\theta}}). \tag{2.34}$$

We remark, however, that the log-likelihood on $\widehat{\mathbf{F}}_t$ is not defined since $\widehat{\mathbf{F}}_t \neq \mathbf{F}_t$.

A final point to discuss is the multi-step-ahead forecasting for the DCC model. Differently from GARCH models, the DCC dynamic is a non-linear process since we have that $\mathbf{C}_t = \text{diag}(\mathbf{R}_t)^{-1/2}\ \mathbf{R}_t\ \text{diag}(\mathbf{R}_t)^{-1/2}$. For this reason, an h-step-ahead correlation forecast cannot be solved analytically to provide a proper forecast method. Engle and Sheppard (2001) make some simplifying assumptions to build a direct forecast of $\mathbf{C}_t$. In particular, assuming that $\overline{\mathbf{C}} \simeq \mathbf{C}$, where $\mathbf{C}$ is the unconditional correlation matrix of the factors and $\mathbb{E}[\mathbf{R}_{t+1}|\mathcal{X}_t] \simeq \mathbb{E}[\mathbf{C}_{t+1}|\mathcal{X}_t]$, then one can use the recursion

$$\begin{aligned} \mathbb{E}[\mathbf{C}_{T+h}|\mathcal{X}_t] = \mathbf{C}_{T+h|T} &= \overline{\mathbf{C}} + (a-b)(\mathbf{C}_{T+h-1|T} - \overline{\mathbf{C}}) && (2.35) \\ &= \overline{\mathbf{C}} + (a-b)^{h-1}(\mathbf{C}_{T+1|T} - \overline{\mathbf{C}}). && (2.36) \end{aligned}$$

Further simulation results for this approximation are presented in Engle (2009). Even though expression (2.36) presents an upward bias, this manifests over longer forecast horizon, generally $h > 100$. In our exercise we will focus on shorter horizon, with a multi-step forecast of a month, or 21 trading days, implying $h = 21$. For this reason, we expect the errors due to approximation to be small.

2.2.4 DCC-GARCH(1,1) with two correlated factors

The first equations follow the same structure as the previous application:

$$\begin{aligned}
\boldsymbol{x}_t &= \underbrace{\begin{bmatrix}\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2 & \mathbf{0} & \mathbf{0}\end{bmatrix}}_{\boldsymbol{\Lambda}^\dagger} \mathbf{F}_t^\dagger + \boldsymbol{\xi}_t \\
\mathbf{F}_t^\dagger &= \begin{bmatrix}F_{1,t}\\F_{2,t}\\\eta_{1,t}\\\eta_{2,t}\end{bmatrix} = \underbrace{\begin{bmatrix}\phi_1 & 0 & 0 & 0\\0 & \phi_2 & 0 & 0\\0 & 0 & 0 & 0\\0 & 0 & 0 & 0\end{bmatrix}}_{\boldsymbol{\Phi}^\dagger} \underbrace{\begin{bmatrix}F_{1,t-1}\\F_{2,t-1}\\\eta_{1,t-1}\\\eta_{2,t-1}\end{bmatrix}}_{\mathbf{F}_{t-1}^\dagger} + \underbrace{\begin{bmatrix}1 & 0 & 1 & 0\\0 & 1 & 0 & 1\\0 & 0 & 1 & 0\\0 & 0 & 0 & 1\end{bmatrix}}_{\boldsymbol{\Psi}^\dagger} \underbrace{\begin{bmatrix}\eta_{1,t}^*\\\eta_{2,t}^*\\\eta_{1,t}\\\eta_{2,t}\end{bmatrix}}_{\boldsymbol{\eta}_t^\dagger}
\end{aligned}$$

with $\boldsymbol{x}_t$ and $\boldsymbol{\xi}$ being $n \times 1$ vectors, $\boldsymbol{\Lambda}^\dagger$ an $n \times 4$ matrix, and $\boldsymbol{\Phi}^\dagger$ and $\boldsymbol{\Psi}^\dagger$ square matrices of dimension 4×4.
The estimated conditional variance of the states is given by

$$\mathrm{Var}[\boldsymbol{\eta}_{t+1}^\dagger|\mathcal{X}_t] = \begin{bmatrix}q_1^* & 0 & 0 & 0\\0 & q_2^* & 0 & 0\\0 & 0 & q_{1,t+1|t} & 0\\0 & 0 & 0 & q_{2,t+1|t}\end{bmatrix}^{1/2} \begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & \multicolumn{2}{c}{\multirow{2}{*}{$\mathbf{C}_{t+1|t}$}}\\0 & 0 & &\end{bmatrix} \begin{bmatrix}q_1^* & 0 & 0 & 0\\0 & q_2^* & 0 & 0\\0 & 0 & q_{1,t+1|t} & 0\\0 & 0 & 0 & q_{2,t+1|t}\end{bmatrix}^{1/2}$$

with q_1^*, q_2^* equal to 10^{-8}. Univariate variances in the lower-right block matrices follow the GARCH(1,1) dynamic

$$\begin{aligned}
q_{1,t+1|t} &= \omega_1 + \alpha_1(\eta_{1,t|t}^2 + P_{t|t}^{\eta_1}) + \beta_1 q_{1,t|t-1},\\
q_{2,t+1|t} &= \omega_2 + \alpha_2(\eta_{2,t|t}^2 + P_{t|t}^{\eta_2}) + \beta_2 q_{2,t|t-1},
\end{aligned}$$

while the 2×2 unit-diagonal conditional correlation matrix $\mathbf{C}_{t+1|t}$ is given by

$$\begin{aligned}
\mathbf{C}_{t+1|t} &= \begin{bmatrix}\rho_{1,t+1|t} & 0\\0 & \rho_{2,t+1|t}\end{bmatrix}^{-1/2} \begin{bmatrix}\rho_{1,t+1|t} & \rho_{1,2,t+1|t}\\\rho_{2,1,t+1|t} & \rho_{2,t+1|t}\end{bmatrix} \begin{bmatrix}\rho_{1,t+1|t} & 0\\0 & \rho_{2,t+1|t}\end{bmatrix}^{-1/2}\\
\mathbf{R}_{t+1|t} &= (1-a-b)\begin{bmatrix}1 & \bar{\rho}_{1,2}\\\bar{\rho}_{2,1} & 1\end{bmatrix} + a\begin{bmatrix}\dot{z}_{1,t|t}^2 & \dot{z}_{1,t|t}\dot{z}_{2,t|t}\\\dot{z}_{2,t|t}\dot{z}_{1,t|t} & \dot{z}_{2,t|t}^2\end{bmatrix} + b\begin{bmatrix}\rho_{1,t|t-1} & \rho_{1,2,t|t-1}\\\rho_{2,1,t|t-1} & \rho_{2,t|t-1}\end{bmatrix}
\end{aligned}$$

where $\dot{z}_{i,t|t} = \eta_{i,t|t}/q_{i,t|t-1}$ is the normalized state error. The initial state $\mathbf{F}_0$ is fixed at $\mathbf{0}$, while its initial state variance is given by:

$$\boldsymbol{\Omega}_0 = \begin{bmatrix}1 & 0 & 1-\phi_1^2 & 0\\0 & 1 & 0 & 1-\phi_2^2\\1-\phi_1^2 & 0 & 1-\phi_1^2 & 0\\0 & 1-\phi_2^2 & 0 & 1-\phi_2^2\end{bmatrix},$$

while other parameters are initialized using two-step PCA, first extracting factors and loadings through (2.33) and then estimating AR(1)-GARCH(1,1) parameters via QMLE.

$\bar{\rho}_{1,2}$ is initialized by calculating the sample correlation of the two factors.
In the DCC case, the ECME Algorithm contains an additional step for the correlation parameters. In this case, this is made up of 4 steps such that $\mathcal{S}_{\mathcal{Q}} = \{1\}$ and $\mathcal{S}_{\ell} = \{2, 3, 4\}$, thus the first maximizes the expected log-likelihood and the subsequent 3 steps maximize the actual one.

CM-Step 1. Same as GARCH(1,1). $\boldsymbol{\theta}_{(\mathcal{Q})} = [\text{vec}(\boldsymbol{\Lambda}), \text{vech}(\boldsymbol{\Gamma}), \text{vec}(\boldsymbol{\Phi})]$ are the analytical solutions to the maximized expected likelihood, as in (1.26) - (1.28). The only difference is with the loadings because of the identifying condition (IC2).

$$\begin{aligned}
\boldsymbol{\Lambda}^{(j)} &= \begin{bmatrix} \mathbf{I}_r & \boldsymbol{\Lambda}_0^{(j)} \end{bmatrix}' \\
\boldsymbol{\Lambda}_0^{(j)} &= \left(\sum_{t=1}^{T} [\boldsymbol{x}_t]_{3:n} \mathbf{F}'_{t|T} \right) \left(\sum_{t=1}^{T} \mathbf{F}_{t|T} \mathbf{F}'_{t|T} + \mathbf{P}^F_{t|T} \right)^{-1},
\end{aligned}$$

where $[\boldsymbol{x}_t]_{3:n}$ indicates all elements of the $\boldsymbol{x}_t$ vector except the first and the second.

CM-Step 2. Parameters $(\omega_1, \alpha_1, \beta_1)'$ are obtained via numerical optimization of (1.16) employing the BFGS algorithm subject to the following restrictions:

$$\begin{aligned}
&\alpha,\ \beta > 0, \\
&\alpha + \beta < 1, \\
&\omega = (1 - \alpha - \beta)(1 - \phi^2).
\end{aligned}$$

As with GARCH(1,1) the *sine* transformation is applied to α_1 and β_1 so that $\alpha_1^{\star} = 0.99\text{sin}(\alpha_1)^2$ and $\beta_1^{\star} = (0.99 - \alpha)\text{sin}(\beta_1)^2$ maintain the domain within (0,1).

CM-Step 3. Same as CM-Step 1, but with the focus on the second factor, i.e the parameters $(\omega_2, \alpha_2, \beta_2)'$.

CM-Step 4. This is the additional step required by the DCC, in which we optimize the actual likelihood (1.16) to obtain $(a, b, \bar{\rho}_{1,2})'$. The following restrictions are employed to preserve system stationarity and correlation proprieties:

$$\begin{aligned}
&a, b > 0, \\
&a + b < 1, \\
&-1 < \bar{\rho}_{1,2} < 1.
\end{aligned}$$

We apply the same *sine* transformation on a, b such that $a_1^{\star} = 0.99\text{sin}(a_1)^2$ and $b_1^{\star} = (0.99 - a)\text{sin}(b_1)^2$ as in the GARCH(1,1) case, plus the *inverse tangent* transformation on $\bar{\rho}_{1,2}$, i.e. $\bar{\rho}_{1,2}^{\star} = 2\text{arctan}(\bar{\rho}_{1,2})/\pi$ to preserve the $(-1, 1)$ domain.

The algorithm stops when the actual likelihood relative increase is smaller than the tolerance parameter $\varphi = 10^{-4}$.

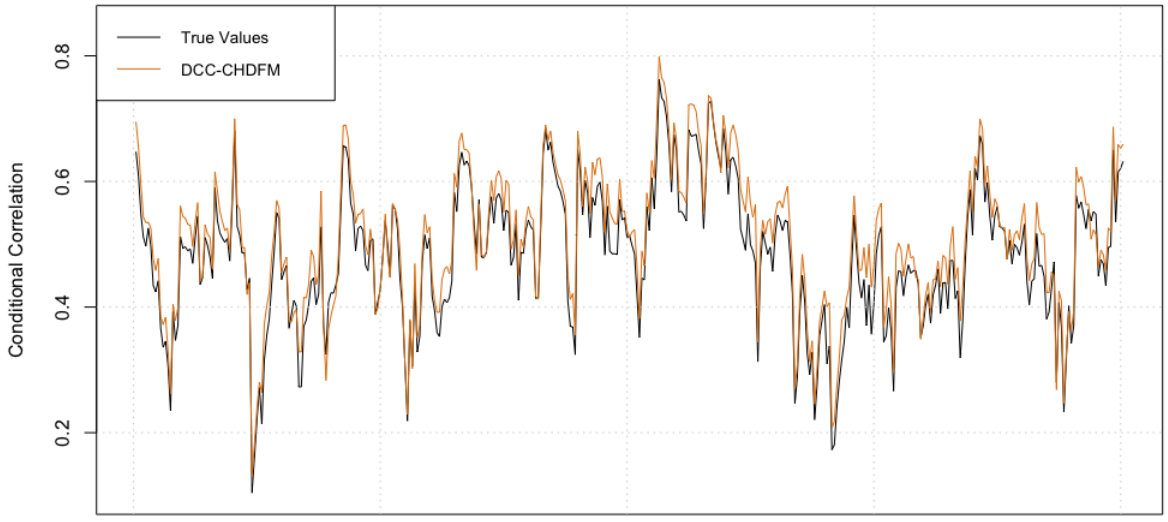

Figure 2.12: *Conditional Correlation of two factors (r = 2) for a simulated DCC-CHDFM, with correlation parameter* $\bar{\rho}_{1,2} = 0.5$. *The DCC parameters are* $a = 0.1$ *and* $b = 0.75$. $T = 1000$ *and* $n = 100$.

2.2.5 Portfolio Allocation Rules

The MVP is one of the most used investment strategies implemented by both academics and professionals. It has the lowest variance of all the optimal portfolios derived from Markowitz's mean-variance optimization problem (Markowitz, 1952) as it solves the following problem

$$\min_{\mathbf{w}} \mathbf{w}^{'} \boldsymbol{\Sigma}_t \mathbf{w} \qquad \text{subject to } \mathbf{w}^{'} \mathbf{1}_n = 1, \tag{2.37}$$

where $\mathbf{w}$ represents the $n \times 1$ vector of portfolio weights which determine the allocation of the portfolio, $\mathbf{1}_n$ is the $n-$dimensional vector of ones, and $\boldsymbol{\Sigma}_t$ is the $n \times n$ conditional covariance matrix of the returns $\boldsymbol{x}_t = (x_{1,t}, \cdots, x_{n,t})^{'}$.
In the absence of short-sale constraints, the optimization problem (2.37) has an analytical solution given by

$$\mathbf{w}^{\star} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_n}{\mathbf{1}_n^{'}\boldsymbol{\Sigma}^{-1}\mathbf{1}_n}. \tag{2.38}$$

MVP has some nice proprieties which make it different from other mean-variance optimal portfolios. First, the MVP portfolio's weights are independent of the mean vector of asset returns. Secondly, it occupies a unique spot on the set of optimal portfolios, which is a parabola in the mean-variance space and known as the efficient frontier. The mean and the variance of the MVP determine the vertex of this parabola. Implementation of (2.38)

requires knowledge of $\mathbf{\Sigma}_t$, which is not available in practice. The straightforward strategy is to replace the unknown $\mathbf{\Sigma}_t$ by an estimator $\widehat{\mathbf{\Sigma}}_t$ that yields a feasible portfolio with weights $\widehat{\mathbf{w}}$. Inaccurate estimators can have a significant influence on the built portfolio, sometimes even greater than the model uncertainty caused by the optimization issue. In particular, when the portfolio dimension $n \to \infty$, the effect becomes considerably stronger.
CHDFM is employed in this setting to benefit from the large cross-section dimension since the variance dynamic is confined to a few factors which drive the conditional heteroskedasticty. Finally, loadings project volatilities onto the observable space of returns. This can be implemented in two ways, either by assuming that factors are uncorrelated, so we will be using GARCH(1,1)-CHDFM, or by relaxing this assumption and employing the DCC-CHDFM.
In practice we will use an h-step-ahead forecast for the covariance estimator to calculate

$$\widehat{\mathbf{w}}_{t+h|t} = \frac{\widehat{\mathbf{\Sigma}}_{t+h|t}^{-1}\mathbf{1}_n}{\mathbf{1}_n'\widehat{\mathbf{\Sigma}}_{t+h|t}^{-1}\mathbf{1}_n}. \tag{2.39}$$

For each t, $\widehat{\mathbf{w}}_{t+h|t}$ represents the allocation weight of each stock at the portfolio construction date t, given the estimated h-step-ahead forecast of the covariance matrix $\widehat{\mathbf{\Sigma}}_{t+h|t}^{-1}$. The analysis relies on a rolling window estimation in which at each $t+h$ the portfolio is rebalanced and new weights are estimated. Thus, the empirical analysis revolves around the high-dimensional covariance matrices forecast.
We now compare the various estimation strategies included in the study:

1) **GARCH(1,1)-CHDFM**: *Conditionally Heteroskedastic Dynamic Factor Model* with uncorrelated factors and GARCH(1,1) dynamic. This is the work of Section 1.5.3.

2) **DCC(1,1)-CHDFM**: *Conditionally Heteroskedastic Dynamic Factor Model* with correlated factors via the DCC, as in Section 2.2.4.

3) **GARCH(1,1)-2SPCA**: *Two-step PCA* estimator, in which one extracts first the factor and loadings and subsequently estimates the GARCH(1,1) dynamic on the extracted latent processes. See Section (1.3.3).

4) **DCC(1,1)-2SPCA**: *Two-step PCA* estimator in which one calculates first the factor and loadings and subsequently estimates the DCC parameter. See Section (2.2.2).

5) **EWP** *Equal Weighted Portfolio.* No covariance is required in this case as each weight is constant and equal to $\widehat{\mathbf{w}}_{t+h|t} = \mathbf{w} = n^{-1}\mathbf{1}_n$.

The strategies include 4 dynamic conditional strategies and one constant strategy. We decided to include the equal-weighted portfolio suggested by DeMiguel et al. (2007) as a simple benchmark in addition to Markowitz portfolios based on MVP, because it has been claimed to be difficult to outperform. For the sake of the exercise we will treat the number of factors as known and equal to $r = 2$.

2.2.6 Data and Performance Measures

The empirical application revolves around portfolio construction in the context of the stock market. We focus our attention on the S&P500 and we download all the constituents daily closing prices from 01/01/1999 to 17/02/2021, for a total of 5,565 daily observations. We adopt the common convention that a 'month' is constituted by 21 consecutive trading days. The study consists in using $T = 1,260$ daily log-returns, about five years of past data, to estimate the parameters for the conditional variance matrix and then calculate the value of the forecast for the consecutive months, i.e $\widehat{\mathbf{\Sigma}}_{t+21|t}$, in order to compute $\widehat{\mathbf{w}}_{t+21|t}$. The weights are then rebalanced on a monthly basis, using the past 1,260 days to calculate portfolio optimal values.
The out-of-sample period spans from 01/12/2003, the month after the first 1,260 business days, through 17/02/2021, for a total of 4,305 days, or 205 months. In simple terms, we utilize a rolling window of $T = 1,260$ days with a sliding window of 21 days. Not all companies in the S&P500 span the whole period, so we eliminate the stocks that present missing values, bringing the total count of stocks to 358. Further filters are applied to data. First, we detect all stocks in the sample that have a correlation coefficient higher than 0.9 and we delete the one with the smaller average market capitalization. This brings the total number of stocks to 357. Secondly we replace extreme values, i.e values that are above 10 times the interquartile range in absolute value, with 0. Those outliers generally do not represent the intrinsic evolution of prices, but rather extraordinary events, such as stock splits.
We consider the portfolio size $n = 100$, and the investment universe is obtained by randomly sampling the 357 stocks to form a portfolio of 100 constituents.

As a performance measure we report the following quantities for each scenario:

- ◇ **AV**: the average return of the 4,305 out-of-sample return, annualized by multiplying the value for 252.
- ◇ **SD**: the standard deviation of the 4,305 out-of-sample return, multiplying by $\sqrt{252}$ to annualize.
- ◇ **IR**: the information ratio, calculated as IR = AV/SD.
- ◇ **TC**: turnover costs, calculated as $\text{TC} = T^{-1}\sum_{t=1}^{T}\|\widehat{\mathbf{w}}_t - \widehat{\mathbf{w}}_{t-1}\|$.
- ◇ **MW**: maximum weight, calculated as $\text{AMW} = T^{-1}\sum_{t=1}^{T}(\max_i \widehat{w}_{i,t})$. It indicates, on average, the biggest exposure of the portfolio.
- ◇ **SSC**: short selling costs, i.e the average exposure to negative positions in the portfolio, calculated as $\text{SSC} = (nT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{w}_{i,t}\mathbb{1}(\widehat{w}_{i,t} < 0)$.

The most relevant performance metric is the out-of-sample standard deviation, since the aim of the MVP is to minimize the volatility rather than maximize the portfolio expected

	AV	SD	SR	TC	MW	SSC
GARCH(1,1)-CHDFM	**12.65**	**13.83**	**0.92**	4.25	15.23	-1.88
GARCH(1,1)-2SPCA	11.62	14.68	0.79	3.49	12.07	-1.67
DCC(1,1)-CHDFM	11.85	13.87	0.85	6.18	13.51	-1.79
DCC(1,1)-2SPCA	11.75	14.56	0.81	3.64	12.24	-1.71
EWP	9.74	19.18	0.51	0.00	1.00	0.00

Table 2.4: Out-of-sample performance measure for each strategy. All values are multiplied by 100.

return or information ratio. Thus, SD measures the effectiveness of the portfolio to accomplish this objective. High out-of-sample average returns (AV) and out-of-sample information ratios (IR) are of course beneficial, but they should be viewed as being of secondary relevance when assessing the effectiveness of a covariance matrix estimator. The final three measures involving the weights give an idea of potential transaction costs, coming from high turnover and short selling arrangements.

2.2.7 Results and Further Research

The results for the different portfolio strategies are summarized in Table 2.4. We see that almost all dynamic strategies consistently outperform the EWP by a wide margin. Both DCC and GARCH-CHDFM produce better estimation performance than their 2SPCA counterparts, but the orthogonal factors provide the best results overall. With a 28% reduction in the portfolio volatility compared to the equally weighted one, it confirms a superior ability to predict the returns covariance matrix. It also features the highest average return and information ratio. DCC-CHDFM performs relatively well, but it may be reasonable to assume that a wider universe of asset class, such as fixed income or commodities, would produce a greater potential for conditional correlations.
Looking at the weights-related performance measure, we note that turnover costs are higher for the CHDFMs, especially for the DCC, since the weights reflect a more dynamic covariance matrix evolution (Figure 2.13). These models also tend to be more exposed to single stocks as confirmed by the average maximum weights.
Figure 2.14 shows the total wealth, *ex* transaction costs, that could be accumulated by investing \$1 on 01/12/2003 and rebalancing the portfolio on a monthly basis until 17/02/2022. GARCH and DCC-CHDFM would achieve a final wealth of \$7.28 and \$6.34, respectively, compared to the \$3.69 that would be obtained by the equally weighted strategy. MVPs are especially good in times of financial crisis, such as the one in 2008 or the one due to Covid-19 in 2021, as the weights are promptly adjusted to cut down exposure to highly volatile stocks. In particular, GARCH models are designed to take into consideration potential volatility clusters and adjust the weights correspondingly.
Although, this simple exercise showcases good forecasting performances for CHDFM, fur-

ther improvement and research could be carried out. First, we treated the number of factors as given, but an explanatory factor analysis could lead to the exact number of factors needed. For the stock market generally, the number of latent factors is just one, but this would have discredited the conditional correlation analysis. Secondly, the number of stocks and the selection of them is also fixed. A simulation exercise employing a different stock sample of the portfolios, with a different dimension, would produce more robust and quantifiable results. As mentioned above, including other asset classes would definitely be more meaningful for factor correlation. Finally, other estimation procedures, such as the one mentioned in De Nard et al. (2019), would provide a more exhaustive view and comparison of the dynamic covariance estimator.

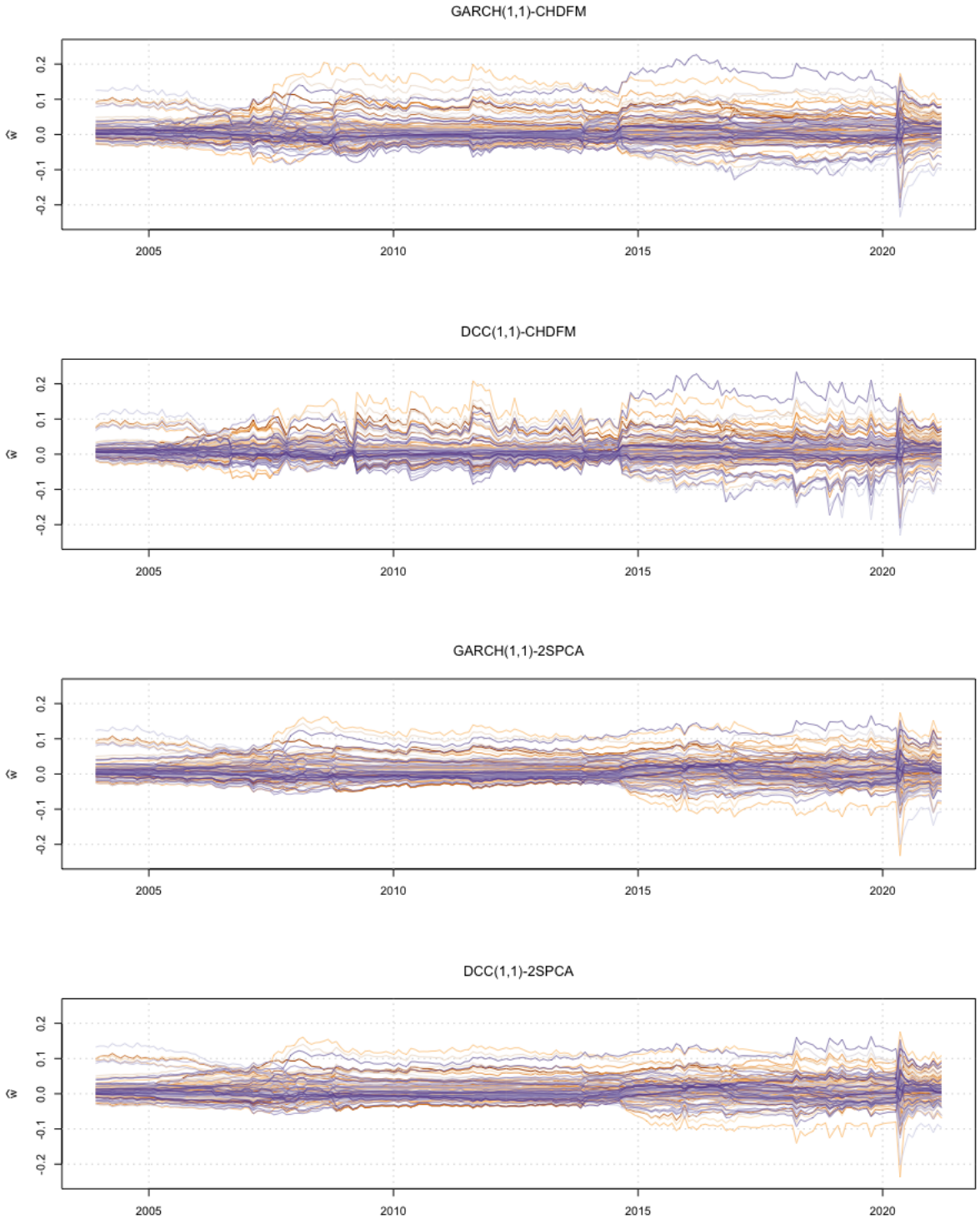

Figure 2.13: *Time series of optimal weights* $\widehat{\mathbf{w}}_{t+h|t}$ *of an* $n = 100$ *portfolio for the four dynamic strategies.*

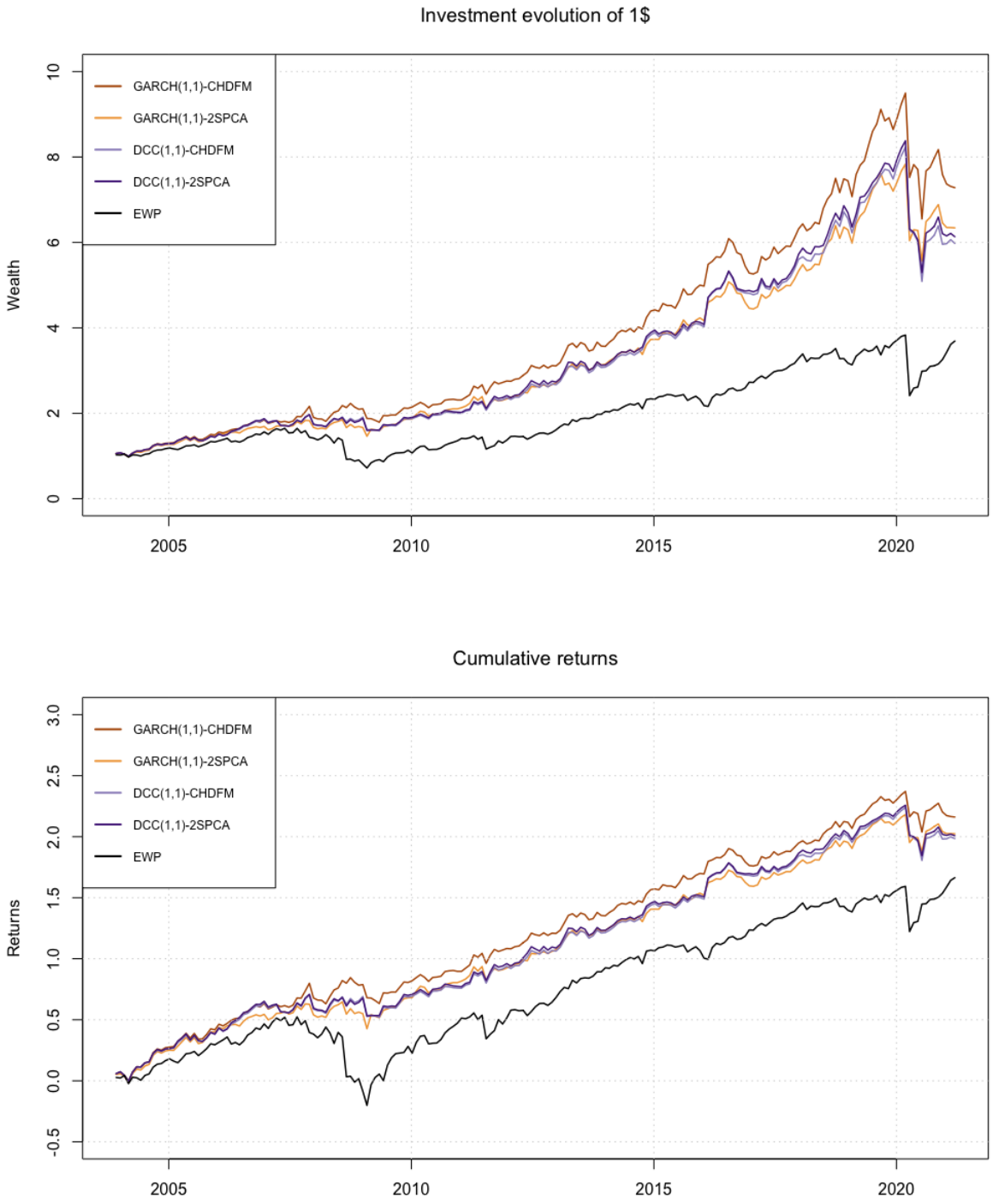

Figure 2.14: *Top: evolution of \$1 investment over time using optimal weights* $\widehat{\mathbf{w}}_{t+h|t}$ *for the five strategies. Bottom: cumulative sum of returns given the optimal set of weights of the different strategies.*

Chapter 3

Estimating Causal Effects of Interventions in Time Using Semiparametric Latent Factor Models

3.1 Introduction

Recently, many applications have been devoted to understanding and revealing *causal* rather than associative relations among variables. One approach in the context of time series is that of synthetic controls (Abadie and Gardeazabal, 2003) and various extensions. This is based on the idea of recovering the counterfactual outcome that would have been observed had an intervention not taken place.

This chapter expands and generalise this class of models, allowing for non-linearity through Semiparametric Latent Factor models (SLFM) (Teh et al., 2005; Alvarez and Lawrence, 2009). These models, which belong to the class of Multi-Output Gaussian Processes, have a high degree of flexibility in building the counterfactual outcome, since they employ all types of information without any limitations on the functional form. They also make it possible to assess the robustness of the synthetic controls, as we can use the posterior distributions of the Gaussian Processes to quantify uncertainty stemming from the functional form estimation. Lastly, as the models learn the relationships which prevail amongst all associated variables, there is no need to match the time series on a calendar basis, making the most of the available data.

To our best knowledge, the only paper that uses Gaussian processes in the context of potential outcomes is Alaa and van der Schaar (2017). The purpose of their article was to infer individualized treatment effects across a series of cross-sectional experiments. However, the bivariate setting arises from the use of the treated and control groups as dependent variables and no time component is exploited. There exists very recent literature that explores multitask causal learning using a Gaussian process exploiting Judea Pearl's Do-Calculus (Pearl, 1995). These papers (Aglietti et al., 2020), however, are mainly focused on understanding the main correlation structure of multiple continuous intervention functions

- defined with a directed acyclic graph (DAG) - as opposed to a single discrete intervention. Furthermore, the main data domain consists of cross-sectional experiments, not time series data.

The main contribution of this chapter is to offer a novel approach to causal inference by using Gaussian Processes. At first, this method removes any linearity assumption or, more generally, any need to specify a functional form. Then, as a fully Bayesian approach, it easily quantify uncertainty around the estimates. This promotes direct estimation of the causal effect estimands such as means and quantiles. Additionally, we develop a framework, based on a linear combination of different kernels, suited for time series and longitudinal data. Specifically, we decompose the whole data space using unobserved components, or factors, that enclose the dynamic structure and relationships of the panel data. This interpretation allows factor models to be generalized in a non linear way by replacing the resulting linear covariance with a non linear one. To test this methodology empirically, we estimate the effect of the UK's vaccination policy compared to that of other European countries. Thus, we learn how a faster vaccination schedule could have reduced the contagiousness and cumulative number of deaths occurred in the first half of 2021.

This chapter is structured as follows. In Section 3.2 we briefly introduce the causal framework and the synthetic control approach, presenting our main assumptions and the causal effect estimands. In Section 3.3 we define the proposed models based on Gaussian Processes. In Section 3.4, we present the estimation procedure. In Section 3.5 we present an illustrative empirical analysis of our approach to obtaining estimates of the causal effect of the UK's effective vaccination programme, introduced in January 2021, on deaths and the infection rate. Section 3.6 describes the main results of the analysis. Finally, we conclude in Section 3.7.

3.2 Causal Framework

The application we refer to throughout the paper serves as an illustrative example and is analysed in Section 3.5. It attempts to understand whether the early and intense vaccination campaign introduced in the UK affected the number of deaths and level of contagiousness of Covid-19 in the first semester of 2021. Formally, the treated unit is the UK and the treatment is the substantially accelerated vaccination schedule. Other EU countries, with other slower vaccination campaigns, will be used to construct the synthetic control for the UK. We would like to note here that we are comparing the UK with a non-treated counterfactual version of itself, using other European countries to create this counterfactual.
Each observation is denoted with $y_{i,t} \in \mathcal{Y}$, where $i = 1, \cdots, m$ is reserved for countries, and $t = 1, \cdots, T_i$ for observation times, and is associated with a set of d potentially time-varying predictors $\mathbf{x}_{i,t} \in \mathcal{X}^d$ such that

$$y_{i,t} = f(\mathbf{x}_{i,t}) + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \omega_i^2), \tag{3.1}$$

where $f(\cdot)$ is a generic function which expresses the input-output relationship and $\epsilon_{i,t}$ is the error term, having mean 0 and variance ω_i^2. The d-dimensional feature vector $\mathbf{x}_{i,t}$ is a set of time series specific to each unit i. In our application this includes mobility data and the number of tests for each country. The data span T_i periods and the first t_0 periods correspond to the data before the intervention, i.e. when the vaccination campaign began in the UK.

3.2.1 Synthetic Control Methods

Synthetic control methods have gained traction as a technique to estimate causal effects from variables that were subject to a single intervention or treatment in time. A traditional approach is the one based on the *difference-in-difference* (DD) method, a static linear regression model where the causal effect is estimated as the difference between the regression coefficient in the treated and the control groups. This is often implemented in a linear regression setting, with the quantity of interest being the interaction term of the dependent variable and the treatment group dummy variable. In this case, $f(\mathbf{x}_{i,t}) = \alpha + \mathbf{x}_{i,t}^{'}\boldsymbol{\beta} + \gamma D(t_0)$, where $D(t_0)$ is a dummy variable which takes values 1 for $t > t_0$ and $\alpha, \boldsymbol{\beta}, \gamma$ are the ordinary least square coefficients. However, DD methods suffer from two main drawbacks (Brodersen et al., 2014): the first one is that they assume that the data are independent and identically distributed, thus disregarding the temporal component; secondly, the pre- and post-intervention periods are captured solely by two time points Abadie et al. (2010); Abadie and Gardeazabal (2003) proposed models generalizing the DD as they allowed the effect of unit-specific unobserved variables to vary with time. In particular, they recover the counterfactual outcome by developing a control group that has a similar pattern in the pre-intervention period as the treated unit. To do so, they find a vector of weights $\{W_1, \cdots, W_{m-1}\}^{'}, W_j > 0, \sum W_j = 1$ which minimize the squared distance between the pre-intervention features (not time series) of the exposed region $\mathbf{x}_i$ and the the features for the unaffected regions $\{\mathbf{x}_j\}_{j \neq i}$. Then, the counterfactual outcome becomes $y_{i,t} = \sum_{i \neq j} W_j y_{j,t}$. However, this method has its own limitations. Indeed, it focuses only on possible convex combinations of control time series to match the treated variable. Furthermore, there is a non-negligible data loss in regards to the temporal component. First of all, only data in the pre-treatment period is used to fit the model and find the optimal weights of the counterfactual unit. Second, time series evolution and interaction over time is neglected, as data is aggregated over time or treated individually for each time period. An alternative class of models is identified by Brodersen et al. (2014), whose approach addresses many of the previous methods' limitations. The authors' approach relies on Bayesian state-space models which encompass the outcome's temporal evolution with exogenous regression components to efficiently build a counterfactual model. State-space models allow for flexibility when modelling a variable that is affected by external noise, distinguishing between a state equation which describes the transition of the latent variable from one point in time to the next one, and a measurement equation, which describes the

accuracy of the signal.[1] Being fully Bayesian makes it possible to (i) incorporate prior information about the model structures and parameters and (ii) have a posterior distribution, and thus a probabilistic uncertainty quantification of the causal impact of the intervention. Although the models focus on one outcome variable and multiple controls, an extension to a multivariate setting has been implemented using Multivariate Bayesian Structural Time Series (Menchetti and Bojinov, 2020), which is limited to linear relationships between outcomes and controls and subject to the Markovian assumption of the variables.

3.2.2 Assumptions

In this subsection, we set up the framework to estimate the causal effect of an intervention on the treated subject. Each subject $y_{i,t}$, i.e. each country in our application, is associated with a binary potential outcome $y_{i,t}(w_{i,t}) \in \mathbb{R}$ where $w_{i,t} \in \{0, 1\}$ is a treatment assignment indicator with '1' referring to the variable being treated (the UK) and '0' to the controls (other European countries). Furthermore, define $\boldsymbol{w}_{1:m,1:T} = \{\boldsymbol{w}_{1,1:T}, \cdots, \boldsymbol{w}_{m,1:T}\}$ as the assignment path up to time T of all units $i = 1, ..., m$ and denote $\mathrm{w}_{1:m,1:T}$ a realization of this path. As in Menchetti et al. (2021), throughout the paper we make a set of assumptions to guarantee that the differences in the potential outcome trajectories are a direct statistical consequence of the intervention. Since some of them can not be directly tested, we rely on the concept of *plausibility* in our empirical settings. In particular, we assume the following:

Assumption 1 (Single intervention) *Unit i received a single intervention if there exist a $t_0 \in \{1, ..., T\}$ such that $w_{i,t} = 0$ for all $t < t_0$ and $w_{i,t} = w_{i,s} = w_i$ for all $t, s > t_0$.*

This says that the treatment is single, i.e. it occurs at one point in time, and is persistent, i.e. has no disruptions. Then, we can ease the notation and drop the t subscript for $t > t_0$

Assumption 2 (Temporal no-interference) *For all $i \in \{1, ..., m\}$ and all $t > t_0$, the outcome of unit i at time t depends only on its own treatment path*

$$y_{i,t}(\mathbf{w}_{1:m,t_0+1:T}) = y_{i,t}(\mathbf{w}_{i,t_0+1:T}).$$

If it holds, one can also drop the subscript i from w_i as this assumption asserts that, whether or not other units receive the treatment at time t_0, this has no impact on other units' potential outcome. Units do not interfere with each other at any point in time. This is the time series equivalent of the cross-sectional Stable Unit Treatment Value Assumption (SUTVA) of Rubin (1974), also known as Temporal SUVTVA from the work of Bojinov and Shephard (2019). In our empirical study, each country's vaccination plan is confined by the country's border and does not affect other countries' rate of contagion or number of deaths. The main underlying observation around this assumption is that, during the period analysed, each country was isolated due to government restrictions. Thus, people's

[1] Formally, dropping the subscript i, $y_t = Z^{'}\alpha_t + \boldsymbol{\gamma}\mathbf{x}_t + \epsilon_t$ represents the *observation equation* and $\alpha_{t+1} = \Phi\alpha_t + R\eta_t$ is the *state equation*, where $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\eta_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$. See Brodersen et al. (2014) for more details.

mobility was prohibited or at least significantly limited. Even after the intervention, we likely expect other European countries' mobility not to affect the UK's number of deaths or contagion rate.
Assumptions 1 and 2 allow as to simplify the notation so that we can use $y_{i,t}(\mathrm{w})$ to indicate the potential outcome of a generic unit i at time t. Thus, the observed outcome for $t > t_0$ is $y_{i,t}(1)$, while $y_{i,t}(0)$ is the *unobserved* or *counterfactual* potential outcome which has to be estimated to measure the causal impact of the intervention.

Assumption 3 (Covariates-treatment independence) *Denote with $\mathbf{x}_{i,t}$ the vector of exogenous variables that are predictive of $y_{i,t}$. For $t > t_0$ those covariates are not affected by the intervention*

$$\mathbf{x}_{i,t}(1) = \mathbf{x}_{i,t}(0).$$

These covariates help improve the outcome prediction, but they produce an estimation bias if they are influenced by the treatment. For the analysis, we expect that an earlier vaccination, during the period considered, neither alters the the number of tests taken nor people's mobility, during the same period. We must remark that our treatment is indeed the quicker vaccination programme and not the programme just by itself. As a consequence, people could anticipate a mass vaccination taking place in the future months and adjust their mobility patterns, but we can assume that they would not travel more because of just being in a country with a higher vaccination rate.

Assumption 4 (Non-anticipating potential outcomes) *for all $i \in \{1, ..., m\}$, the outcome of the unit i at time $t < t_0$ is independent of the treatment that occurs in t_0*

$$y_{i,t}(\mathbf{w}_{1:m,1:T}) = y_{i,t}(\mathbf{w}_{1:m,1:t_0}).$$

This assumption is usually made in the literature (Bojinov and Shephard, 2019; Callaway and Sant'Anna, 2021) to affirm that the future intervention has no influence on pre-intervention statistical units, implying that there is no anticipation of the treatment effect before t_0. In the empirical application, although the government advertised the forthcoming vaccination campaign, the outcomes, such as the number of deaths, did not shift before the programme took place. Furthermore, people had no way to anticipate that the UK programme would have been significantly faster compared to other countries.

Assumption 5 (Non-anticipating Treatment) *The assignment mechanism at time t_0 for the unit i depends only on past covariates and past outcomes*

$$p(w_{i,t_0} = \mathrm{w}_{i,t_0} | \boldsymbol{w}_{1:t_0-1}, \mathbf{y}_{i,1:T}, X_{i,1:T}) = p(w_{i,t_0} = \mathrm{w}_{i,t_0} | \mathbf{y}_{i,1:t_0-1}, X_{i,1:t_0-1}). \tag{3.2}$$

This assumption is analogous of the *unconfounded assignment mechanism* (Imbens and Rubin, 2015) in a time series framework and ensures that, conditionally on past $\mathbf{y}_{i,1:t_0-1}, X_{i,1:t_0-1}$, any variations in the outcomes are to be attributed to the intervention. In our setting, the UK set up a faster vaccination campaign as just looking at the previous number of deaths and rate of reproduction shows.

3.2.3 Causal Estimands

Let $\delta_{i,t} = y_{i,t}(1) - y_{i,t}(0)$ be the individual level (UK) causal effect at time t, then the *additive causal* effect on the subject i at time t is the population average treatment effect and it is given by

$$\tau_{i,t} = \mathbb{E}\left(\delta_{i,t}|\mathbf{x}_{i,t}\right). \tag{3.3}$$

We are also interested in the uncertainty surrounding the treatment effect. This can either be measured through the variance

$$\varrho^2_{i,t} = \mathbb{V}\left(\delta_{i,t}|\mathbf{x}_{i,t}\right), \tag{3.4}$$

or by directly applying quantile functions to calculate credible regions

$$q^{\delta}_{i,t}(\alpha) = F^{-1}_{\alpha}\left(\delta_{i,t}|\mathbf{x}_{i,t}\right), \tag{3.5}$$

with levels of confidence generally set to $\alpha = 95\%$. We aim to estimate these values from a dataset $\mathcal{D} = \{X, \mathbf{y}, \mathrm{w}\}$, which involves $T = \sum_{i=1}^{m} T_i$ samples of different time series. The main challenge is that we only observe one of the potential outcomes for every subject i, which implies that the treatment effect is unobserved, so we cannot directly estimate $\tau_{i,t}$. In addition to its point-wise impact, we are interested in the cumulative effect of the intervention over time

$$\mathcal{T}_i = \sum_{t=t_0+1}^{T_i} \tau_{i,t} \tag{3.6}$$

where t_0 represents the time in which the intervention takes place. The cumulative sum is a suitable measure when $y_{i,t}$ is a *flow* variable which is measured over an interval of time (e.g number of deaths in a country). This quantity, however, loses its interpretability when $y_{i,t}$ is a *stock* variable, i.e. a quantity measured at a specific time, representing a quantity existing at that point in time (e.g rate of infectiousness). In this case, it is more meaningful to use the average treatment effect of the intervention

$$\overline{\tau}_i = \frac{1}{T_i - t_0} \sum_{t=t_0+1}^{T_i} \tau_{i,t} = \frac{\mathcal{T}_i}{T_i - t_0}. \tag{3.7}$$

This measure extends Sävje et al. (2019) to the time series framework of the average distributional shift effect since here it is averaged across time as opposed to units.

Within a Gaussian Processes (GP) framework, expected values and variances are straightforward to derive and we have that $\delta_{i,t} \sim \mathcal{N}(\tau_{i,t}, \varrho^2_{i,t})$. Sometimes, however, Gaussian likelihoods may not be appropriate and some mathematical transformations may be needed. For example, a random variable which takes only non-negative values (e.g. counts, lengths) would be better represented using a log-normal distribution. Thus, if a random variable

$\delta_{i,t} = \log(\delta^*_{i,t})$ has a Gaussian distribution, then $\delta^*_{i,t} \sim \log \mathcal{N}(\tau^*_{i,t}, \varrho^{*2}_{i,t})$ is log-normally distributed. By directly modelling the transformed variable one obtains the causal effect given by

$$\delta^*_{i,t} = \exp\left(\log y_{i,t}(1) - \log y_{i,t}(0)\right) = \frac{y_{i,t}(1)}{y_{i,t}(0)}. \tag{3.8}$$

Then, taking the expectation $\mathbb{E}[\delta^*_{i,t}|\mathbf{x}_{i,t}] = \tau^*_{i,t}$ and using the fact that $\delta^*_{i,t}$ is log-normal

$$\tau^*_{i,t} = \mathbb{E}\left(\frac{y_{i,t}(1)}{y_{i,t}(0)}|\mathbf{x}_{i,t}\right) = \exp\left(\tau_{i,t} + \frac{\varrho^2_{i,t}}{2}\right), \tag{3.9}$$

with related percentiles equations as in 3.5. This can be interpreted as a *multiplicative causal effect* with base 1. A ratio above (resp. below) 1 indicates a positive (resp. negative) effect of the treatment.
Generally, for the cumulative effect (3.6) and average effect (3.7), there is no closed form solution unless each $\delta_{i,t}$ is normally distributed and independent over time. In this case one can use samples from the posterior predictive distribution over the counterfactual variable to obtain samples from the posterior causal effect distribution, the quantity we are interested in. This method also works when using variable transformations, as one can convert it back to the original scale and then calculate the empirical cumulative (average) distribution, with given mean and quantiles.

3.3 Gaussian Processes

Most of the existing methods in the literature rely on a linear function $f(\cdot)$ in 3.1. In this paper, we aim to relax this linearity assumption and estimate $f(\cdot)$ in a non-parametric fashion, making few assumptions regarding its form. This is achieved using GPs that generally provide a powerful Bayesian method for regression and classification problems (Rasmussen and Williams, 2006). The function $f(\cdot)$ is treated as an unknown parameter and is assigned a suitable prior distribution defined by a user-specified kernel. Inference and prediction tasks are than carried out based on the corresponding posterior and predictive distribution, which also reflect the uncertainty of the estimation procedure.

3.3.1 Single-Output Gaussian Process

Let $\mathbf{y} = \{y_1, \cdots, y_T\} \in \mathbb{R}^T$ be the time series of the treated country (the UK) and $X = \{\mathbf{x}_1^{'}, \cdots \mathbf{x}_T^{'}\}^{'} \in \mathbb{R}^{T\times d}$ the matrix of the d associated covariates. Define as $Z = \{\mathbf{z}_1^{'}, \cdots \mathbf{z}_T^{'}\}^{'} \in \mathbb{R}^{T\times m-1}$ the matrix containing the time series of the relevant units for the synthetic control (the other European countries). Then $X^* = \{X^{'}, Z^{'}\}^{'}$ is the matrix of $d + m - 1$ the time-varying covariates, all with sample size T. A single-output GP (SOGP) takes the form

$$y_t = f(\mathbf{x}^*_t) + \epsilon_t, \;\; \epsilon_t \sim \mathcal{N}(0, \omega^2) \tag{3.10}$$

where ϵ_t is the independent and identically distributed (*i.i.d.*) noise which accounts for the model error. This process is completely defined by its mean, $\mu(\mathbf{x}_t^*) = \mathbb{E}[f(\mathbf{x}_t^*)]$, and covariance $k(\mathbf{x}_t^*, \mathbf{x}_s^*) = \mathbb{E}[(f(\mathbf{x}_t^*) - \mu(\mathbf{x}_t^*))(f(\mathbf{x}_s^*) - \mu(\mathbf{x}_s^*))]$ function for each $t, s = 1, \cdots, T$. Without loss of generality, we take the mean function to be 0 and consider a standard option for the kernel that defines the covariance.
As we will discuss in more detail in Section 3.6.1, this approach offers a robust and flexible alternative to Brodersen et al. (2014), but it is also similar in the sense that the $\mathbf{x}_t^*$s are not modelled and are assumed to be deterministic inputs. Potential further gains may be achieved if the joint distribution of $(y_t, \mathbf{x}_t^*)$, $\forall t$, is modelled. Such a model is presented in the next section using Multi-output Gaussian processes (MOGP), which generalize GPs in a multivariate framework.

3.3.2 Multi-Output Gaussian Process

MOPGs exploit correlations between multiple outputs and across the input space, thus providing the potential for better predictions, particularly in scenarios with noisy data or missing values (Bonilla et al., 2008). In this paper, we are going to focus on a class of models referred to as Semiparametric Latent Factor Models (SLFM), in which each output corresponds to a linear combination of one or more latent random functions. These shared processes help transfer the common information across units, without the need to specify a different kernel structure for each output. This is potentially useful in our context, as we want to incorporate knowledge from the other countries without making the model dependent on numerous state-specific parameters. Compared to the SOGP, the MOPG jointly models all the countries, each one with the appropriate set of covariates.

Define $\mathbf{y} = \{\mathbf{y}_1^{'}, \cdots, \mathbf{y}_m^{'}\}^{'}$, where $\mathbf{y}_i^{'} = \{y_{i,1}, \cdots, y_{i,T_i}\} \in \mathbb{R}^{T_i}$ is the time series vector of observed variables and $X = \{X_1^{'}, \cdots, X_m^{'}\}^{'}$with $X_i \in \mathbb{R}^{T_i \times d}$ the matrix of the d covariates associated with output i, and where i is the country. For the independent variables we assume a *heterotopic data* configuration (Liu et al., 2018), i.e. each output potentially has different time-varying covariates associated with it, $X_1 \neq, \cdots, \neq X_m$, each one with T_i samples such that $T = \sum_{i=1}^{m} T_i$. In this way, we can model the relationship occurring in each input-output set. The MOGP model is shown below,

$$y_{i,t} = f_{i,t}(\mathbf{x}_{i,t}) + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \omega_i^2), \tag{3.11}$$

for each $i = 1, \cdots, m$ and $t = 1, \cdots, T_i$ and where the *i.i.d.* noise $\epsilon_{i,t}$ accounts for the observation errors. The likelihood function for the m outputs is defined as

$$\mathbf{y}|\mathbf{f}(X), X, \Omega \sim \mathcal{N}(\mathbf{f}(X), \Omega), \tag{3.12}$$

where $\Omega = \text{diag}(\omega_1^2 \mathrm{I}_{T_1}, \cdots, \omega_m^2 \mathrm{I}_{T_m}) \in \mathbb{R}^{T \times T}$ and the outputs $\mathbf{f}(X) = \{f_1(X_1), \cdots f_m(X_m)\}^{'}$ are probability distributions in function space and represent the MOPG

$$\mathbf{f}(X) \sim \mathcal{GP}(\mu(X), \mathcal{K}(X, X)). \tag{3.13}$$

Without loss of generality, one can assume that $\mu(X) = \mathbf{0}$ while $\mathcal{K}(X, X) \in \mathbb{R}^{mT \times mT}$ is the multi-output positive semi-definite covariance matrix, defined as

$$\mathcal{K}(X, X) = \begin{bmatrix} K_{1,1}(X_1, X_1) & K_{1,2}(X_1, X_2) & \dots & K_{1,m}(X_1, X_m) \\ K_{2,1}(X_2, X_1) & K_{2,2}(X_2, X_2) & \dots & K_{2,m}(X_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ K_{m,1}(X_m, X_1) & K_{m,2}(X_m, X_2) & \dots & K_{m,m}(X_m, X_m) \end{bmatrix} \tag{3.14}$$

with $K_{i,j}(X_i, X_j) = K_{j,i}(X_j, X_i)' \ \forall i, j$ by symmetry. Taking a look at block matrices $K_{i,j}(X_i, X_j) \in \mathbb{R}^{T_i \times T_j}$, they are defined such that

$$K_{i,j}(X_i, X_j) = \begin{bmatrix} k(\mathbf{x}_{i,1}, \mathbf{x}_{j,1}) & \dots & k(\mathbf{x}_{i,1}, \mathbf{x}_{j,T_j}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_{i,T_i}, \mathbf{x}_{j,1}) & \dots & k(\mathbf{x}_{i,T_i}, \mathbf{x}_{j,T_j}) \end{bmatrix}, \quad i, j = 1, \dots, m.$$

The next step is to define the kernel for the covariance of each of the GPs $f(\mathbf{x}_j)$. Each kernel depends on a set of hyper-parameters $\boldsymbol{\phi}$ which determine its structure. For simplicity, let us focus first on the case of $i = j$, so we can drop the unit subscript.
The squared exponential kernel is a popular choice:

$$k_\phi(\mathbf{x}_s, \mathbf{x}_t) = \sigma^2 \exp\left(-\sum_{r=1}^{d} \frac{(x_{s,r} - x_{t,r})^2}{2\ell_r^2} \right),$$

where r is the r-th input and $\boldsymbol{\phi} = \{\ell_1, \cdots, \ell_d, \sigma^2\}'$. From this equation, we can see that the inverse of ℓ_r^2 regulates the how sensitive the kernel covariance is to changes in the r-th dimension of the input. For large values of ℓ_r^2, the inverse approaches zero, which will cause the value of the covariance to be invariant to the change in $(x_{s,r} - x_{t,r})^2$. This effect will then rule the relevance of the $t - r - th$ input to the kernel, hence, the name automatic relevance determination (ARD).

Another useful kernel is the Matérn kernel given by:

$$k_\phi(\mathbf{x}_s, \mathbf{x}_t) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} \left\| \mathbf{x}_s - \mathbf{x}_t \right\|^2 \right)^\nu J_\nu \left(\frac{\sqrt{2\nu}}{\ell} \left\| \mathbf{x}_s - \mathbf{x}_t \right\|^2 \right) \tag{3.15}$$

with $\boldsymbol{\phi} = \{\ell, \sigma^2\}'$ and where $J_\nu(\cdot)$ is the modified Bessel function and $\Gamma(\cdot)$ is the gamma function. When the dimension $d = 1$ we have that this kernel generates a continuous-time version of an AR(p) Gaussian process where $p = \nu - 1/2$. A particular case is achieved with $\nu = 1/2$, since the Matérn kernel reduces to the exponential kernel given by $k_\phi(\mathbf{x}_s, \mathbf{x}_t) = \exp(\left\| \mathbf{x}_s - \mathbf{x}_t \right\| / \ell)$, which is the covariance process of an Ornstein-Uhlenbeck (OU) process, the continuous-time analogue of an AR(1) process. Let us focus on the one-dimensional case with $\mathbf{x}_t = t$, i.e. the only covariate is time and let us call the lag between two time points Δ. It is shown that taking the Fourier transform of the power spectrum of an OU process on $\mathbb{R}$ with drift ϕ and diffusion σ gives

$$k(\Delta) = \frac{\sigma^2}{2\phi} e^{-\phi|\Delta|}. \tag{3.16}$$

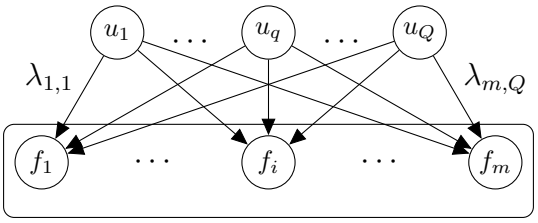

Figure 3.1: *Structure of an LCM. u_q and f_i represent the latent and observed function, respectively. $\lambda_{i,q}$ is the weight associated with each function.*

Thus, the exponential decay rate of the autocorrelation is captured using t instead of a lagged version of y_t. Both the above kernels are stationary, i.e. the covariance function depends on the relative positions of two inputs and not their absolute location.[2] If $k(x_1, x_1)$ and $k(x_2, x_2)$ are covariance functions over different spaces $\mathcal{X}_1$ and $\mathcal{X}_2$, then the direct sum $k(x, x) = k_1(x_1, x_1) + k_2(x_2, x_2)$ and the tensor product $k(x, x) = k_1(x_1, x_1) \cdot k_2(x_2, x_2)$ are also covariance functions (defined on the product space $\mathcal{X}_1 \times \mathcal{X}_2$), by virtue of the sum and product constructions. We can then flexibly sum or multiply all the kernels to have the type of interaction we need in the covariance matrix. As an example, a linear kernel plus a periodic one will generate a periodic kernel with a trend.

3.3.3 Multi-Output Kernels

Finally, in order to fully specify the distribution of $\mathbf{f}(X)$, which is a GP with multiple outputs, we need to make an assumption about the dependence between $f_i(\mathbf{x}_i)$s. The simplest case is to assume independence which will imply the following covariance structure $\mathcal{K} = \text{ diag}(K_{1,1}(X_1, X_1), ..., K_{m,m}(X_m, X_m))$. Alvarez et al. (2012) give a survey of several more flexible methods, including the intrinsic coregionalization model (ICM), the semiparametric latent factor model (SLFM), and the linear model of coregionalization (LMC). These models may be viewed as performing exploratory factor analysis on $\mathcal{K}$ with unobserved factors $u_q(X)$. In particular, let us consider the LMC. This specification, widely used in geostatistics, expresses the outputs as a linear combination of Q latent functions as

$$f_i(X_i) = \sum_{q=1}^{Q} \lambda_{i,q} u_q(X_i) \tag{3.17}$$

where $u_q(X_i)$ is itself a latent Gaussian process with mean $\mathbf{0}$ and $\text{Cov}[u_q(X_i), u_q(X_i)] = K_q(X_i, X_i)$, while $\lambda_{i,q}$s are the coefficients which measure output correlations. Furthermore, the model assumes that the latent processes, $u_q(X_i)$ and $u_p(X_i)$ for $p \neq q$, are independent and such that $\text{cov}[u_q(X_i), u_p(X_i)] = 0$. Then, cross-covariances between the output can be

[2] To allow some flexibility in the model, especially when data exhibit visible trends, one can introduce *non-stationary* kernels. The most simple example is the linear kernel. This is defined as $k_\phi(\mathbf{x}_s, \mathbf{x}_t) = \sigma_i^2 \mathbf{x}_s' \mathbf{x}_t$ where $\phi = \sigma^2$.

calculated by

$$K_{i,j}(X_i, X_j) = \sum_{q=1}^{Q} \lambda_{j,q}\lambda_{i,q}K_q(X_i, X_j). \tag{3.18}$$

Linear combinations of different kernels still result in a valid positive definite covariance matrix. This approach is defined as *separable* (Alvarez et al., 2012) due to the decoupled input-output structure of the covariance.
This model is defined as semiparametric, since it combines a nonparametric component (the Gaussian processes) with a parametric part (the linear mixing via $\lambda_{i,q}$).
For notational simplicity, let us assume an isotopic data configuration, i.e $X_\circ = X_1 = , \cdots , = X_m$. Then, define B_q as the positive semi-definite matrix such that $\mathrm{B}_q = \boldsymbol{\lambda}_q\boldsymbol{\lambda}_q^{'} + \boldsymbol{k}\mathrm{I}_q$ where $\boldsymbol{k} = \{k_{1,q}, \cdots , k_{m,q}\}$ and $k_{j,q}$ positive and $\boldsymbol{\lambda}_q = \{\lambda_{1,q}, \cdots , \lambda_{m,q}\}^{'}$. SLFM assumes that $\boldsymbol{\lambda}_q\boldsymbol{\lambda}_q^{'}$ has rank 1, but generalization with higher rank is possible using LMC. Then, one can write the multi-output covariance as

$$\mathcal{K}(X, X) = \sum_{q=1}^{Q} \mathrm{B}_q \otimes K_q(X_\circ, X_\circ). \tag{3.19}$$

The coregionalization matrix B defines the amount of *inter* and *intra* task transfers of learning among all the outputs. Thus, the latent kernel is shared across all the outputs but is scaled by a factor $\mathrm{B}_q^{[i,j]}$. For example, it is possible that in France, the rate of contagiousness is predicted better by the mobility data than it is in the remaining countries. In this case the B entry associated with France will be higher compared to other countries.[3] Denote by $X_\circ^{/t} \in \mathbb{R}^{T\times(d-1)}$ the matrix of d inputs minus *time*, which is denoted by t. Let us consider specifically the time series defined in (3.11). Using the structure of 3.19, we will focus on a particular specification given by

$$\mathcal{K} = \mathrm{B}_1 \otimes K_{rbf}(X_\circ^{/t}, X_\circ^{/t}) + \mathrm{B}_2 \otimes K_{Mat}(t_\circ, t_\circ), \tag{3.20}$$

where $K_{rbf}(\cdot), K_{Mat}(\cdot)$ are the squared exponential and Matérn kernels respectively. This combination captures both stationary trends between the output and the input and the autocorrelation structure for each output. Using the time trend as a separate covariate creates a distinctive kernel structure in which it's possible to see how $y_{i,t}$ is related to the other $y_{j,s}$, $\forall i, j, t, s$, something SOGPs cannot do, as shown in the second graph of Figures 3.4 and 3.6. If some country i has little to no relation to another country j the terms of $\mathrm{B}_2^{[i,j]}$ will be close to 0.

3.3.4 SOGP vs MOPG comparison

Although simpler in nature, SOGPs present some limitations when compared to MOPGs. First, there is a considerable data loss when applying a univariate approach as opposed

[3]The special case of $Q = 1$, generate the ICM. The computational complexity is largely reduced in exchange of a more restrictive architecture, as one latent process, with a specified kernel, becomes the only the source of variability among outputs.

to the multivariate *heterotopic* data configuration. We lose all information contained in the covariates of the control countries. Even including them, it won't benefit the model performance because we don't expect, for example, that the number of deaths in the UK is affected by *mobility data* of any other European country. And we would still face a high-dimensional input with potentially more parameters to estimate. Another crucial aspect of the univariate setting is that data is lost because the sample length of the training data is reduced. When we estimate the model we can only use as much data as the minimum sample length of each variable. Both the treated and control groups are restricted before the intervention.[4] In the multivariate case, we narrow the UK dataset up to t_0 but for the other countries we take advantage of the whole dataset. The goal of the multivariate case is to learn the common relationships between input-output, and the more data is fed into the system, the smaller the uncertainty surrounding prediction.

A second important aspect is variable time matching. Most of the studies involving causal estimates related to Covid-19 do not compare countries on a calendar-day basis since the virus hit each country at different dates (see Ghayda et al. (2020) for a meta-study on the comparison between the use of calendar date and days since the outbreak). For example, Born et al. (2020) use a common reference point to initialize observations for each country: $t = 1$ is the day when the number of total positive cases surpasses a threshold of one infection per one million people. This will ensure the effect of the pandemic is comparable across countries. In the MOGP setting, each country is temporally dependent on the others thanks to a separate kernel structure on time. This takes into consideration relative time distance, such as how many days it takes on average to pass from the peak number of deaths to half this number. This information is encoded into the kernel hyperparameters which define a rate of decay, such as ϕ in (3.16). The kernel then links each time point in the treated series to all points in the control time series. This type of architecture is displayed in the third picture of Figure 3.4 and Figure 3.6.

Finally, a feature that is related to SOGP when applied to causal analysis is that of input dimensionality. While, in the MOPG framework, each dependent variable has its own associated covariates, in SOGP the auxiliary outcomes are added to the pool of independent variables X. If the dimensionality of the input space is low, then learning the link function is a much easier problem than learning a high-dimensional function. Though GPs are capable of dealing with large-dimensional covariates in theory, there are several practical and computational issues to consider in their implementation (Tripathy et al., 2016). The inclusion of more input variables (not necessarily related to the response variable) can increase model fit quality, but may lead to poor future predictions due to variance-bias trade off (Cawley and Talbot, 2007). Furthermore, from a computational point of view, optimization becomes more challenging as too many covariates can result in a singular Hessian matrix (Djolonga et al., 2013).

[4]It is still possible to add lagged or forward versions of other countries, although this would inevitably impact input dimensionality.

3.4 Estimation

There are different approaches to GP's parameter and hyperparameters estimation. The Bayesian framework provides effective and consistent inference tools for the former issue. GPs can, indeed, be treated as hierarchical models, where the parameters are represented by the latent function $\mathbf{f}(X) = \mathbf{f}$, which in turn can be considered samples from a population characterized by hyper-parameters θs. In this case, $\boldsymbol{\theta} = \{\boldsymbol{\phi}^{'}, \boldsymbol{\lambda}^{'}, \boldsymbol{\omega}^{2'}\}^{'}$ contain the parameters of the kernel covariance functions $\boldsymbol{\phi}$, the components of the corregionalization matrix B, and likelihood variances $\boldsymbol{\omega}^2 = \text{diag}(\Omega)$. Given Bayes' rule, the posterior over the parameters is

$$p(\mathbf{f}|\mathbf{y}, X, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|\boldsymbol{\theta})}{\int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|\boldsymbol{\theta})\, d\mathbf{f}} \tag{3.21}$$

where $p(\mathbf{y}|\mathbf{f}, X)$ is the *likelihood*, $p(\mathbf{f}|\boldsymbol{\theta})$ is the prior, and the expression in the denominator is a normalizing constant, called *marginal likelihood*. We can then express the hyperparameters' posterior, making the marginal likelihood from above play the role of the likelihood so that

$$p(\boldsymbol{\theta}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|X, \boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}. \tag{3.22}$$

The main toolkit for the analysis and optimization is GPy, a Gaussian Process framework written in Python.[5]

3.4.1 Type II Maximum Likelihood

In practice, instead of maximizing the posterior in (3.22), one can instead maximize the marginal likelihood, with respect to the hyperparameters $\boldsymbol{\theta}$ (*Type II Maximum Likelihood*)

$$p(\mathbf{y}|X, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f}, X, \theta)p(\mathbf{f}|X, \boldsymbol{\theta})\, d\mathbf{f}. \tag{3.23}$$

The strength of GPs is the tractability of the integral over the parameters $\mathbf{f}$, since we know that $p(\mathbf{f}|X, \theta) = \mathcal{N}(\mathbf{f}|0, \mathcal{K})$. Furthermore, we have that $p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \Omega)$. Then, following Rasmussen and Williams (2006), one can perform the integration of the product of two normals which yields the log-marginal likelihood

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}'\Sigma\mathbf{y} - \frac{1}{2}\log|\Sigma| - \frac{T}{2}\log(2\pi), \tag{3.24}$$

where $\Sigma = \mathcal{K} + \Omega$ is the covariance matrix of the noisy outcome $\mathbf{y}$ and contains all *hyperparameters*. The first term is a data-fit term, as it is the only one involving y, the second one represents the complexity term, since it depends only on the covariance function, and the last one is a constant. Marginalizing out the Gaussian vector $\mathbf{f}$ moves up the Bayesian

[5] https://sheffieldml.github.io/GPy/. Following the authors' proposal, we will treat the *nugget* parameters $\boldsymbol{k}$ as a parameter to optimize in order to increase numerical stability. These parameters are in the parametrization of $\text{B}_q = \boldsymbol{\lambda}_q\boldsymbol{\lambda}_q^{'} + \boldsymbol{k}\text{I}_q$ to guarantee the positive definiteness of the kernel.

hierarchy by one level, thus reducing the odds of overfitting (Murphy, 2013). To maximize the marginal likelihood, we first find the derivative of the marginal likelihood with respect to the kernel hyperparameters

$$\frac{\partial}{\partial \theta_i} \log p(\mathbf{y}|X, \boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{'}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_i}\Sigma^{-1}\mathbf{y}^{'} - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_i}\right), \tag{3.25}$$

where $\frac{\partial \Sigma}{\partial \theta_i}$ depends on the structure of the kernel and the parameters we are taking derivatives of. The inversion of the $\mathcal{K}$ matrix requires $\mathcal{O}(n^3)$ by standard methods, and then $\mathcal{O}(n^2)$ time per hyperparameter to calculate the gradient. Given the minor relative computational cost of calculating derivatives, a gradient-based optimizer would be beneficial.[6] A very popular method is BFGS, named after its inventors Broyden, Fletcher, Goldfarb, and Shanno (Fletcher, 2000). As a Quasi-Newton procedure it approximates the Hessian using the differences of gradients over several iterations, thanks to a *secant* (Quasi-Newton) condition.

Algorithm 1 BFGS method

1: choose initial guess $\boldsymbol{\theta}_0$
2: choose B_0, the initial Hessian guess, e.g. $B_0 = \mathrm{I}$
3: **for** $k = 0, 1, 2, \cdots$ **do**
4: solve $B_k \boldsymbol{s}_k = -\nabla f(\boldsymbol{\theta}_k)$
5: $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \boldsymbol{s}_k$
6: $\mathbf{y}_k = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k)$
7: $B_{k+1}^{-1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^{'}}{\mathbf{y}_k^{'} \boldsymbol{s}_k} - \frac{B_k \boldsymbol{s}_k \boldsymbol{s}_k^{'} B_k}{\boldsymbol{s}_k^{'} B_k \boldsymbol{s}_k}$
8: **end for**

The standard BFGS method employs the full history of gradients to calculate the Hessian approximation. The limited memory BFGS, abbreviated as L-BFGS, uses only the most recent (usually 20) gradients to compute the product $B_k^{-1}\nabla f(\boldsymbol{\theta}_k)$. The main advantage of L-BFGS is that it requires less storage than the $n(n+1)/2$ elements required to store the Hessian estimate, requiring only $\mathcal{O}(sn)$ instead of $\mathcal{O}(n^2)$ (Nocedal and Wright, 2006).
The L-BFGS-B algorithm further extends L-BFGS to handle linear constraints on variables such that $l_i \leq \theta_i \leq u_i$, where l_i and u_i are constant lower and upper bounds for each θ_i. The algorithm separates fixed and unconstrained variables at each step by using the gradient method. Subsequently, it employs the L-BFGS method on the free variables to achieve higher accuracy.

[6] Generally, the objective function is non-convex and local minima exist and can make the the optimization procedure challenging. However, empirical studies with non-complex covariance functions seem to indicate that the issue is not extremely serious, as every local maxima correspond to a different interpretation of the data (Rasmussen and Williams, 2006).

3.4.2 Hamiltonian Monte Carlo

The most popular Bayesian methods rely on MCMC, which can be quite slow for a high-dimensional parameter space. It is possible to approximate the posterior over the latent functions and over the hyperparameters after setting the priors, using Hamiltonian Monte Carlo (HMC). Here, an additional *momentum* variable $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \mathrm{M})$ is introduced for each parameter $\boldsymbol{\theta}$, which is regarded as *position*. The covariance matrix M, called the mass matrix, rotates and scales the target distribution and it is generally set to the identity matrix, M = I, when no information is available on the target distribution. The joint density $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ defines the Hamiltonian

$$\begin{aligned} H(\boldsymbol{\phi}, \boldsymbol{\theta}) &= -\log p(\boldsymbol{\phi}, \boldsymbol{\theta}) && (3.26)\\ &= -\log p(\boldsymbol{\phi}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) && (3.27)\\ &= T(\boldsymbol{\phi}|\boldsymbol{\theta}) + V(\boldsymbol{\theta}). && (3.28) \end{aligned}$$

The first term, $T(\boldsymbol{\phi}|\boldsymbol{\theta}) = -\log p(\boldsymbol{\phi}|\boldsymbol{\theta})$, is called *kinetic energy* and it is equal to the square of the momentum since $-\log p(\boldsymbol{\phi}|\boldsymbol{\theta}) = \log p(\boldsymbol{\phi}) = 0.5\boldsymbol{\phi}'\boldsymbol{\phi}$, being the momentum density independent of the target density. The second term, $V(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$, is the *potential energy* and is related to the target distribution $p(\boldsymbol{\theta})$. This extended model then follows Hamiltonian dynamics through fictitious time, whose evolution depends on a set of differential equations:

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= +\frac{\partial H}{\partial \boldsymbol{\phi}} = \frac{\partial T}{\partial \boldsymbol{\phi}} && (3.29)\\ \frac{d\boldsymbol{\phi}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} = -\frac{\partial T}{\partial \boldsymbol{\theta}} - \frac{\partial V}{\partial \boldsymbol{\theta}} = -\frac{\partial V}{\partial \boldsymbol{\theta}}, && (3.30) \end{aligned}$$

since $\frac{\partial T}{\partial \boldsymbol{\theta}} = 0$ by independence. The solution to these differential equation is not available in closed form and must be computed numerically. The most popular numerical integrator, which preserves volume and reversibility of the system, is the *Leapfrog* integrator (Girolami and Calderhead, 2011). The leapfrog integrator takes L steps, each one of size ϵ, and iterates between a half step for the momentum and a full-step update for the position.

$$\begin{aligned} \boldsymbol{\phi}_{t+\frac{\epsilon}{2}} &= \boldsymbol{\phi}_t - \frac{\epsilon}{2}\frac{\partial V}{\partial \boldsymbol{\theta}_t} && (3.31)\\ \boldsymbol{\theta}_{t+\epsilon} &= \boldsymbol{\theta}_t - \epsilon M^{-1}\boldsymbol{\phi}_{t+\frac{\epsilon}{2}} && (3.32)\\ \boldsymbol{\phi}_{t+\epsilon} &= \boldsymbol{\phi}_{t+\frac{\epsilon}{2}} - \frac{\epsilon}{2}\frac{\partial V}{\partial \boldsymbol{\theta}_{t+\epsilon}}. && (3.33) \end{aligned}$$

The leapfrog discretization introduces small numerical errors in the total energy calculation. The correction takes the form of a Metropolis-Hastings step, in which the probability of accepting a proposal $(\boldsymbol{\phi}^*, \boldsymbol{\theta}^*)$ generated from $(\boldsymbol{\phi}, \boldsymbol{\theta})$ is $\min(1, \exp(H(\boldsymbol{\phi}, \boldsymbol{\theta}) - H(\boldsymbol{\phi}^*, \boldsymbol{\theta}^*)))$. In case of rejection, the previous values are used to initialize the new iterations. In practice, when using HMC two main parameters need to be tuned. Firstly, one needs to choose

the appropriate step size. Taking a look to the acceptance rate, it is possible to reduce or increase the value of ϵ. Smaller steps are more computationally expensive, but precision may improve. However, a very small ϵ makes it difficult to efficiently explore the target distribution. The best way to determine the appropriate length L of the simulation is to look at the parameters' auto-covariance function, increasing L to achieve more independent samples. Excessively long trajectories can erode computational effort, as the simulation exercise may generate loops, making the destination point the same as the initial one. Once reasonable values for ϵ and L have been determined, desired samples from the target distribution can be obtained.
Generally, for MOPG, the parameters of interest are the $\boldsymbol{\lambda}$ composing the corregional matrix B which defines the relationships among the outcomes.[7] The conditional posterior of $\boldsymbol{\lambda}$ can be computed as

$$p(\boldsymbol{\lambda}|\boldsymbol{\phi}, \boldsymbol{\omega}^2, \mathbf{y}, X) = \frac{p(\mathbf{y}|X, \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\omega}^2)p(\boldsymbol{\lambda}|\boldsymbol{\phi}, \boldsymbol{\omega}^2)}{p(\mathbf{y}|X, \boldsymbol{\phi}, \boldsymbol{\omega}^2)}. \tag{3.34}$$

By maximizing the denominator in (3.34) it is possible to obtain the maximum likelihood type II estimates,

$$\{\boldsymbol{\phi}^*, \boldsymbol{\omega}^{2*}\} = \arg\max_{\boldsymbol{\phi}, \boldsymbol{\omega}^2} p(\mathbf{y}|X, \boldsymbol{\phi}, \boldsymbol{\omega}^2). \tag{3.35}$$

Then, the marginal posterior of $\boldsymbol{\lambda}$ can be approximated by conditioning on the estimates obtained by ML-II optimization as in 3.24

$$p(\boldsymbol{\lambda}|X, \mathbf{y}) = \int p(\boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\omega}^2|X, \mathbf{y}) d\boldsymbol{\phi} d\boldsymbol{\omega}^2 \approx p(\boldsymbol{\lambda}|X, \mathbf{y}, \boldsymbol{\phi}^*, \boldsymbol{\omega}^{2*}). \tag{3.36}$$

In this way, one can focus the attention and computational burden only on the parameter of interest. As a robustness check, we will try to free up $\boldsymbol{\omega}^2$ as well, in order to account for the estimation uncertainty coming from the observational errors. The algorithm is performed using *GPy* and employing 5,000 samples, an identity mass matrix M = I, and a starting value of $\epsilon = 0.01$

3.4.3 Prior Specification

In order to adopt a Bayesian approach to inference we need to specify a prior distribution $p(\boldsymbol{\theta})$. We select a *weakly informative* prior distribution (Gelman et al., 2004), which incorporates enough information to regularize the posterior distribution. In this way, we keep the posterior within reasonable values without contributing actively to the knowledge of the underlying parameters. For the loadings, we employ a normal distribution,

$$\lambda_i \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2), \tag{3.37}$$

where $\mathcal{N}$ is the normal distribution with mean μ_λ and variance σ_λ^2. This former parameter expresses our expectation about the value of each element composing the corregional matrix. In practice, we set $\mu_\lambda = 0$ and $\sigma_\lambda = 10$. Furthermore, we want to specify the set of

[7] In the Bayesian estimation approach we will keep the parameters $\boldsymbol{k}$ fixed.

variance parameters that govern the observational errors. A typical prior distribution for such a variance is

$$\omega_i^2 \sim \mathcal{G}(a, b), \tag{3.38}$$

with $\mathcal{G}$ being the Gamma distribution with parameters a, b. Those parameters generally depend on our prior belief about the precision surrounding the collection of data. As a default, we set up $a = 0.1$ and $b = 1$. Overall, these specifications provide a useful, although wide, default while preserving flexibility in case a more specific prior information is available.

3.4.4 Posterior Predictive and Causal Estimates

Let us say we want to use observed data $\mathbf{x}$ to make predictions about data $\widetilde{\mathbf{x}}$. For example, as we will see in our application, $\mathbf{x}$ can be training data and $\widetilde{\mathbf{x}}$ test data. GPs are stochastic processes in which any finite subset of random variables follows a joint normal distribution. Thus, it is possible to determine the joint prior distribution of the observations $\mathbf{y}$ and the output $\widetilde{\mathbf{y}} = \mathbf{f}(\widetilde{X})$ at test points $\widetilde{X}$ as

$$\begin{bmatrix} \mathbf{y} \\ \widetilde{\mathbf{y}} \end{bmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathcal{K}(X, X) + \Omega & \mathcal{K}(X, \widetilde{X}) \\ \mathcal{K}(\widetilde{X}, X) & \mathcal{K}(\widetilde{X}, \widetilde{X}) \end{pmatrix} \right] \tag{3.39}$$

where $\mathcal{K}(X, \widetilde{X}) \in \mathbb{R}^{T \times \widetilde{T}}$ is the matrix of the covariances calculated at all pairs of training and test points, X and $\widetilde{X}$, respectively. Then, it is possible to analytically derive the posterior distribution of $\widetilde{\mathbf{y}}$, conditioned on $\mathbf{y}$, by using multivariate Gaussian proprieties.

$$\widetilde{\mathbf{y}}|\mathbf{y}, X, \widetilde{X} \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\Sigma}) \tag{3.40}$$

where $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\Sigma}$ are the predictive mean and predictive variance, given by

$$\widetilde{\boldsymbol{\mu}} = \mathcal{K}(\widetilde{X}, X)[\mathcal{K}(X, X) + \Omega]^{-1}\mathbf{y} \tag{3.41}$$

$$\widetilde{\Sigma} = \mathcal{K}(\widetilde{X}, \widetilde{X}) - \mathcal{K}(\widetilde{X}, X)[\mathcal{K}(X, X) + \Omega]^{-1}\mathcal{K}(X, \widetilde{X}). \tag{3.42}$$

Thus, the predictive uncertainty, $\widetilde{\Sigma}$, does not depend on $\mathbf{y}$, but only on the output dependencies given by the kernel structure of X and $\widetilde{X}$. However, when parameter uncertainty is accounted for, the distribution is no longer Gaussian as indicated in (3.40). Given the posterior distributions of the hyperparameters of the model $\boldsymbol{\theta}$ and the function $\mathbf{f}$, we can calculate the in-sample posterior predictive as

$$p(\widetilde{\mathbf{y}}|\widetilde{\mathbf{x}}, X, \mathbf{y}) = \int p(\widetilde{\mathbf{y}}|\widetilde{\mathbf{x}}, \mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f}|X, \mathbf{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|X, \mathbf{y}) d\mathbf{f} d\boldsymbol{\theta}. \tag{3.43}$$

For inference, we first simulate draws from the posterior of the hyperparameters, then we simulate GPs for the given set of hyperparameters to obtain prediction samples.

In causal impact analysis we are mainly concerned with the posterior predictive of the counterfactual time series i in the absence of an intervention. To do so, we need the out-of-sample forecasts from the distribution

$$p(\widetilde{y}_{i,t_0+h}|\widetilde{\mathbf{x}}_{i,t_0+h}, X_{i,1:t_0}, \mathbf{y}_{i,1:t_0}, \{X_{j\neq i,1:T_j}\}, \{\mathbf{y}_{j\neq i,1:T_j}\}) \tag{3.44}$$

for $h > 0$ and $j = 1, \ldots, m$. This is a special case of (3.43), only with $\mathbf{y}_i$ and X_i restricted for $t < t_0$. Then, we can obtain the counterfactual time series $\widetilde{\mathbf{y}}_i^{[k]}(0) = \{\widetilde{y}_{i,t_0+1}^{[k]}(0), \cdots, \widetilde{y}_{i,T_i}^{[k]}(0)\}'$, for $k = 1, \ldots, N$ samples. If Assumption 1 through Assumption 5 hold, these values are the realization of $y_{i,t}(0)$ defined in Section 3.2.3, i.e. the response that would have been observed after the intervention, had the intervention not taken place. It is worth noting that the posterior predictive density is conditional on the observed data of the treated country before the intervention as well as the data in all control countries both before and during the intervention. This is because we assumed temporal no-interference in Assumption 2, stating that the outcome of unit i at a time $t_0 + h$ depends solely on its own treatment path. Furthermore, through Bayesian model averaging, we integrate out all parameters (functions) and hyperparameters, so that the distribution does not depend on a particular choice of parameter estimates.
Finally, using the samples $\widetilde{y}_{i,t}^{[k]}(0)$, for $k = 1, \ldots, N$, it is possible to compute the posterior distribution of the pointwise impact

$$\widetilde{\delta}_{i,t}^{[k]} = y_{i,t}(1) - \widetilde{y}_{i,t}^{[k]}(0) \qquad t = t_0 + 1, \ldots, T_i, \tag{3.45}$$

where $y_{i,t}(1)$ is the observed outcome. As in Brodersen et al. (2014), the density in (3.43) is a joint distribution over all counterfactual data points, rather than a set of univariate predictive distributions. This ensures that we can correctly estimate the trajectory of the counterfactuals through the dynamic structure defined by the model. The samples are also employed to compute the cumulative and average impact (3.6) and (3.7).

3.5 Empirical Analysis

3.5.1 Covid-19 Vaccination programme

Evidence suggests that vaccination against Covid-19 reduces the risk of severe complications, including death, and slows down the transmission of infections (see Zheng et al. (2022) for a meta-analysis of numerous Covid-19 studies). We are still to determine how much of the observed slow down in the spread, and how many lives saved, are attributable to fast and effective inoculation policies such as those introduced by the UK, as opposed to more conservative programmes, such as the ones implemented, for instance, in France, Portugal, and Greece. The United Kingdom has delivered one of the world's fastest vaccination campaigns, giving the first shot to about 67% of the adult population and a second to 50% by the end of June 2021, potentially helping to reduce deaths and infection

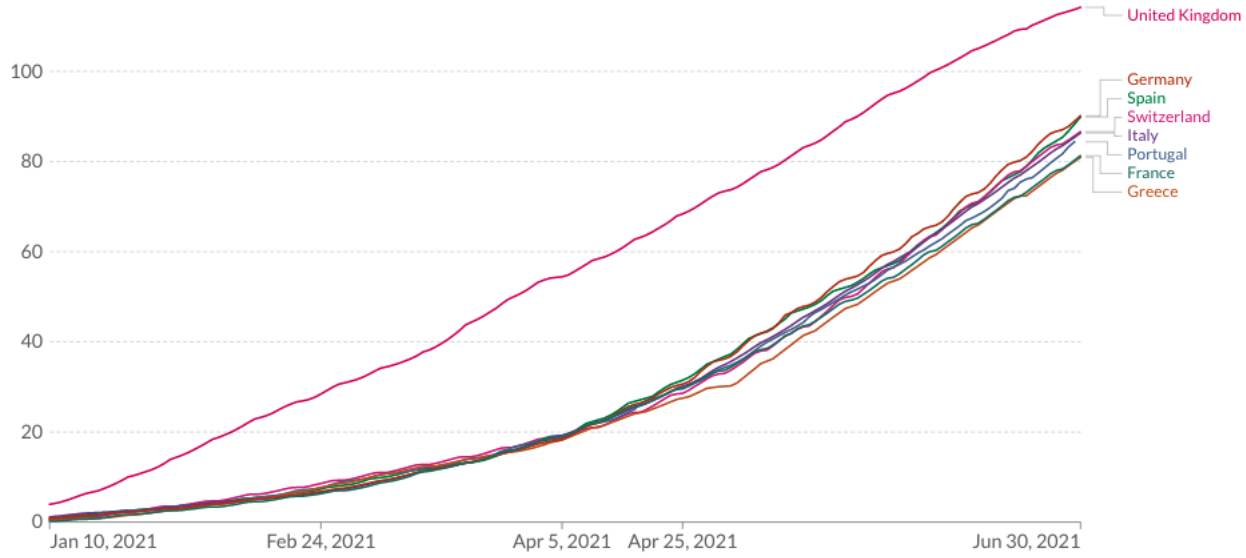

Figure 3.2: *Covid-19 vaccine doses administered per 100 people. Total number of doses administered, divided by the total population of the country. All doses, including boosters, are counted individually. Source: Our World in Data*

rates.[8] In this application, we seek to answer the question above (about the impact of the quick programme) by comparing the observed deaths and reproduction rate in the UK to a "counterfactual UK", a synthetic control constructed using EU countries with less ambitious inoculation programmes.

3.5.2 Data

Our data consist of weekly data points for different countries from 1^{st} March 2020 to 30^{th} June 2021. The date of the intervention t_0 is set to be the 31^{th} January 2021, since that is the first week in which the number of people that had the second dose surpassed 500,000. As mentioned above, the treated country for this study is the United Kingdom, as its policy differed from that of the other European countries.
We are going to focus our analysis on two different outcome (dependent) variables. The first one is the confirmed Covid-19 deaths per million people. The variable is divided by the population for each country to obtain a continuous variable and then it is converted on a log scale to better handle the asymmetry of the data arising from the absence of negative values. The second variable of interest is the estimate of the reproduction rate (R) of Covid-19. R measures the level of contagiousness of the virus and equals the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection.

[8]https://coronavirus.data.gov.uk/details/vaccinations

To respect the Covariates-treatment independence (Assumption 4), all covariates are considered unrelated to the intervention. They are: i) time trend, ii) Google Mobility Data and, iii) weekly number of Covid-19 tests. The first one is a variable that represents time, with $t = 1$ being the first time we have an observation. Since we are working with heterotopic data, each of the outcome variables is associated with different covariates, all with potentially different start times. For example, in the *number of deaths* study we have that Italy (the country whose reports are the earliest) first reported on '2020-03-15', thus this time value will be assigned as $t = 1$. The UK first reported on '2020-04-19' and this will be assigned a value $t = 5$, as this date is 5 weeks after $t = 1$. Although the time trend is linearly increasing, the relationship with the dependent variable is not necessarily linear thanks to the flexible structure of the kernel function on the input space. *Google Mobility Report* is a publicly available dataset that records how visits to different places changed overtime, compared to a baseline. The venues covered by the dataset are: grocery and pharmacy, parks, transit stations, retail, recreation, residential, and workplaces. Similarly to Her et al. (2022); Chatzilena et al. (2022) we reduce the dimension of the input space performing Principal Component Analysis (PCA) on each country and we use the first principal component as a single variable which represents country mobility. On average, more that 80% of the variability of the dataset is explained by this factor, meaning we can ignore other components. The last variable is the weekly average Covid-19 tests per 1,000 people.

3.5.3 Methodology

We now explain the methodology we employ to estimate the causal effects set out in Section 3.2.3. Our main concern is to ensure that the definitive model works well on the observed data before we apply it to creating the counterfactual. To assess the ability of each models to explain the data, we adopt a typical machine learning approach, splitting the data into train and test samples. In particular, we only use data from the period before the intervention as we want to avoid having data after the intervention as test samples. Evaluating the models based on data $\mathbf{y}$ that are contaminated by the intervention can be viewed as trying to minimize the distance between $y_{i,t}(1)$ and $y_{i,t}(0)$, thus introducing a downward bias on the causal impact estimate. The new dataset is split at time $t^* < t_0$ into two parts, the training and the test sets, which account for $2/3$ and $1/3$ of data, respectively. There are two main issues to address: which countries to choose as control series, and which type of kernel structure the model should have.

The first problem involves finding the combination of countries that achieve better predictive performance. Employing a too high number of countries in one model would increase exponentially the number of parameters to estimate. This leads to a greater model complexity, which would make the model prone to over-fitting. As a first step, we perform an early screening on the set by using dynamic time warping, or DTW (Giorgino, 2009). The algorithm produces a distance metric between two input time series. The similarity or dissimilarity of two time series is then calculated by converting the data into vectors and calculating the Euclidean distance between those points in vector space. Then, the

8 components which minimize the distance are chosen to be the set of potential control series of the experiment. This algorithm is particularly useful for dealing with sequences in which single components have characteristics that vary over time, not necessarily in sync. The second issue is related to the appropriate choice of the SLFM architecture and relevant kernel function. We compare different methods.

1) **2FGP**: *Two-Factor GP.* This is the model outlined in the Section 3.3.2 equation (3.20), in which the radial basis function (rbf) kernel is adopted on the input space given by the spatial covariates and the Matèrn Kernel on the time trend. This is to say that there are two unobserved and independent latent factors, one given by time, whose covariance structure resembles the continuous-time AR(p) process as outlined in Section 3.3.2, and one given by the non-linear relationship that occurs between the outcome and the covariates. Number of parameters: 31.

2) **INGP**: *Independent GPs.* While the coregionalized model shares information across outputs, the independent models cannot do that. In particular, we assume that in (3.19), $\boldsymbol{\lambda}_q = \mathbf{0}$ and $k_{j,q} = 1 \; \forall \; j, q$, i.e. $\mathrm{B}_q = \mathrm{I}_m$. In the regions where there is no training data specific to an output the independent models tend to revert to the prior assumptions. We want to test if there is a transfer of learning among all the outputs. Number of parameters: 11.

3) **1FGP**: *One Factor GP.* Instead of assuming two separate input spaces and kernels for time and the other covariates, as in (3.19), we combine the two kernels such that $\mathcal{K} = \mathrm{B}_1 \otimes (K_{rbf}(X_\circ, X_\circ) + K_{Mat}(X_\circ, X_\circ))$. In this way, we combine features of the *rbf* and *Matérn* kernels on a shared input space. This structure implies an unobserved factor common to all tasks, which does not differentiate between time and the other components. However, the model is simpler as it requires less parameters to estimate. Number of parameters: 21.

4) **2RBF**: *Two-RBF Factor GP.* The input space is divided into time trend and spatial covariates but the kernel function has the same structure (Radial Basis Function) for both t and $\overline{X}$. $\mathcal{K} = \mathrm{B}_1 \otimes K_{rbf}(\overline{X}, \overline{X}) + \mathrm{B}_2 \otimes K_{rbf}(t, t)$. This model tests if the *rbf* kernel structure can better describe time trends as opposed to the *Matèrn* kernel. Number of parameters 31.

5) **SOGP**: *Single-Output Gaussian Process.* The model is referenced in (3.10), in which the outcome variable is $\mathbf{y} = \mathbf{y}_i$ and the covariates are $X^* = (\{\mathbf{y}_j\}_{j \neq i}, X_i)$, i.e. all the other control variables and the relevant covariates for the treated subject i. We adopt a single rbf kernel on the whole input space without ARD to define the covariance structure. Number of parameters: 3.

6) **BCI**: *Bayesian Causal Impact.* The *local linear trend* outlined by Brodersen et al. (2014), with the same data structure employed by SOGP. The optimization is performed by using Kalman Filters and MCMC. In both the univariate cases we adopt

an *isotopic data* framework.[9] Number of parameters: 9.

Models 1 to 4 fall into the MOPGs framework defined in Section 3.3.2, Model 5 is the SOGP of Section 3.3, while Model 6 is a replica of the work of Brodersen et al. (2014). Once the models have been fitted on the training dataset we can evaluate the forecast performance calculating the distance from the observed values. We use different measures of dispersion:

- ◇ **MSE**: *Mean Squared Error* . It computes the average of the squared errors, calculated as the differences between the estimated values and the actual values

$$\text{MSE}_i = \frac{1}{t_0 - t^*} \sum_{t=t^*}^{t_0} (y_{i,t} - \widetilde{\mu}_{i,t})^2 \tag{3.46}$$

  where $\widetilde{\mu}_{i,t}$ is defined in (3.42). While simple, the MSE disregards the uncertainty of the predictions.

- ◇ **LogS**: *Log Score*. Forecasts are usually surrounded by uncertainty, and being able to quantify it is pivotal to good decision making. Consider the GP framework of Section 3.4.4, with the hyperparameters given by Type II MLE. The logarithmic score (Good, 1952) is defined as

$$\text{LogS}_i = -\frac{1}{t_0 - t^*} \sum_{t=t^*}^{t_0} \log \mathcal{N}(\widetilde{\mu}_{i,t}, \widetilde{\sigma}_{i,t}) \tag{3.47}$$

  where $\widetilde{\sigma}_{i,t}$ is the i^{th} diagonal element of $\widetilde{\Sigma}$ as in (3.42). Thus, the score is equal to the log of the predictive density of $\mathbf{y}_i$ given by (3.40). The measure also takes into consideration the variability of the point forecast. Since we are working with GP, the distribution is normal and available in closed form, making the calculation straightforward.[10]

- ◇ **ES**: *Energy Score*. This scoring rule is the multivariate extension of the continuous ranked probability score, CRSP (Matheson and Winkler, 1976). Let $\mathbf{y} = \{y_{t^*}, \cdots, y_{t_0}\}' \in \mathbb{R}^h$ be the values of the outcome i on the h-horizon test set where $h = t_0 - t^*$. Denote by $\widetilde{\mathbf{y}}$ the forecast distribution (3.40) on $\mathbb{R}^h$ with N samples $\{\widetilde{\mathbf{y}}^{[1]}, \cdots, \widetilde{\mathbf{y}}^{[N]}\}$ with $\widetilde{\mathbf{y}}^{[k]} = \{\widetilde{y}_{t^*}^{[k]}, \cdots, \widetilde{y}_{t_0}^{[k]}\}$ with $k = 1, \ldots, N$. Then, the energy score can be calculated as

$$\text{ES}_i = \frac{1}{N} \sum_{k=1}^{N} ||\mathbf{y}^{[k]} - \mathbf{y}|| - \frac{1}{2N^2} \sum_{k=1}^{N} \sum_{c=1}^{N} ||\mathbf{y}^{[k]} - \mathbf{y}^{[c]}|| \tag{3.48}$$

[9] For the SOGP and Bayesian Model of Brodersen et al. (2014) one can use only the number of points such that $T_i = \min(T_1, ...T_m)$, while for the Gaussian Process we have heterotopic data in which T_i may be different from T_j.

[10] The Gaussian distribution assumption does not hold in general, especially when parameter uncertainty is accounted for. In case of Bayesian estimation, such as in Section 3.4.2, the predictive distribution is no longer normal. Nonetheless, it is still possible to use the samples obtained from the HMC distribution to calculate the score (Jordan et al., 2019).

where $||\cdot||$ is the Euclidean norm on $\mathbb{R}^h$ and $c = 1, \ldots, N$. This function evaluates samples from a multivariate forecast and returns a single estimate.[11]

For a given model, the lower the score, the higher the accuracy of the forecast. All three measures are compared to see which model performs better. At this stage, given the high number of models to fit, we perform type II ML to optimize hyperparameters. Bayesian estimation is employed subsequently on the whole dataset to have a more exhaustive estimate of the *causal estimands*.
Before proceeding with the results, we sum up the procedure involved to get to the causal estimands.

1) We remove any European country which has no input or output data reported for the period under consideration, i.e from '01-01-2020' to '01-06-2021'.

2) We restrict the data before the intervention $t \leq t_0$ and split into train $t = 1, \cdots, t^* - 1$ and test $t = t^*, \cdots, t_0$ set.

3) We apply DTW on that period to restrict the number of available countries to 8.

4) For any given combination of 4 countries (plus the UK), we train the 6 models outlined in Section 3.5.3 and calculate the dispersion metrics. We have 70 combinations of countries, for a total of 420 models.

5) The combination which results in a lower Energy Score overall is selected as the best model.

6) We fit the selected model on the whole dataset, restricting only the UK before intervention date and calculate optimal $\boldsymbol{\theta}$, through type II ML.

7) We perform HMC with 5,000 samples on the parameters $\boldsymbol{\lambda}$ that define the correlation matrix B, leaving the others fixed to their MLE values.

8) Given the obtained samples of hyperparameters, we calculate the prediction distributions and related causal estimands.

The steps 1) to 5) take place before the intervention t_0 while steps 6) to 8) deal with the post-intervention analysis. The latter study is applied first to the weekly deaths per million people data and subsequently for the weekly infection rate R.

3.6 Results

3.6.1 Before Intervention: Model Comparison

We start by applying the DTW algorithm to select the countries that are most similar to the UK. This process restricts the pool of European candidates for each outcome. Given

[11] Sampling from the forecast distribution can be regarded as an approximation of the values of the proper scoring rules, for a sufficiently large N (Jordan et al., 2019).

	Weekly deaths per million people					
	2FGP	1FGP	2RBF	INGP	SOGP	BCI
MSE	0.8189	0.8274	1.2754	4.2811	**0.7263**	3.1693
logS	**0.2389**	0.4624	0.3451	1.1465	0.4008	4.7445
ES	**0.6776**	0.7791	0.8820	2.7745	0.7025	2.6261
	Weekly infection rate R					
	2FGP	1FGP	2RBF	INGP	SOGP	BCI
MSE	**0.3597**	0.4679	0.4028	2.1470	0.4905	0.5386
logS	-0.6960	-0.6815	**-0.8182**	0.7835	-0.6173	-0.5928
ES	**0.2883**	0.3354	0.2915	1.4496	0.3505	0.3794

Table 3.1: *Comparison of the different measures of prediction error outlined in Section 3.5.3 according to different models. Lowest values, which indicate good predictive performances, are indicated in bold.*

these countries, we focus on forecasting performance before the intervention t_0 to assess the performance of the various models. We fit the ones in Section 3.5.3 for each country combination and select those that achieve the lowest possible Energy Score. Taking a look at Table 3.1, we see that the 2FGP achieves the lowest Energy Score for both of the analyses. The other two GP-based models (1FGP and 2RBF) perform well overall, but slightly worse compared to the base model (2FGP). The INGP is the worst GP model since it cannot rely on the part of data after the training threshold $t > t^*$, thus converging to the dependent variable average values. The SOGP performance is solid on the Weekly Deaths application and on the infection rate. Although the dataset used is the same as BCI, the latter model employs a state-space framework in which the time layout is given by a random walk. Thus, in the case of the SOGP outperforming the BCI, it means that the non-linearity of variables provides a better ground for model prediction than modelling the time dynamics through a linear equation. Nonetheless, the two univariate models produce comparable trajectories, but in general BCI tends to underestimate the uncertainty, resulting in a lower Energy Score.[12] As mentioned above, MOPGs still possess a time structure without sacrificing non-linearity. Another comparison of the models' performance is displayed in Figure 3.3, where each model is represented by a different coloured line, and the crosses serve as the observations.

3.6.2 After Intervention: Causal Effect

Weekly deaths per million people

Once the model and the countries are established, it is possible to fit the model to the whole dataset, restricting only the UK in the period before intervention. For the analysis where the outcome is weekly deaths per million, the best results are achieved using five countries including the UK, namely 'Italy', 'Netherlands', 'reland', and 'Portugal'. Using type II Maximum likelihood as in (3.24) one can obtain the optimal hyperparameters for the kernels. The estimated kernel matrices are of dimension $mT \times mT$, where $T = \sum_{i=1}^{m} T_i$, and represent the variances and covariances of each data point. Understanding the kernel function of a GP is essential to interpreting the association not only between variables but also among different data points belonging to different variables. The kernel on the left of Figure 3.4 is $\mathcal{K}$, which is the sum of the outer products of each kernel times the coregionalization matrix B. Each of the m blocks represents the variance-covariance matrix of each process $f_j(\mathbf{x}_j)$. The upper left block is the one corresponding to the UK and it is smaller compared to the others, as only pre-intervention data is included. A first visual inspection of $\mathcal{K}$ shows that the time component accounts for the majority of the variability as values range from 0 to 2. Furthermore, we see that *mobility data* and *number of tests* were very important variables to explain Ireland (fourth country) but of almost no influence in the Netherlands (third country). However, we see how the independent variables of other countries explain very little for the UK (top row/column of the matrix). The time-domain coregionalization matrix is more homogeneous and we can see that UK weekly deaths followed a pattern more similar to Ireland, as we observe slightly higher values of $B_{2,[2,4]}$ compared to other countries. As mentioned in Section 3.6.1, all that matters is the relative distance in time between countries. Lockdown measures, vaccination programmes, etc. do not have to match, and the relationship between the main variable and lagged/forward version of the control units are still captured by the model, even if they are not linear.
Now, let us focus on the causal estimates. As shown in Figure 3.5, the GP model provides a close fit for the pre-intervention period. Following the beginning of the vaccination campaign, observations start to diverge from the counterfactual predictions: the actual number of weekly deaths, represented by the blue crosses, was consistently lower than what is predicted with slower vaccination rates. Subtracting observed from predicted data, as is shown on the right-hand part of Figure 3.5, produces the posterior estimate of the effect achieved by the campaign. The top-right graph gives an idea of the cumulative log number of deaths for the UK compared to control countries. Before the intervention, the cumulative difference was statistically non-significant, .i.e the differential number of deaths was the same among the countries. After t_0, however, the cumulative effect starts going down, reaching a value significantly lower than 0. In the UK, the counterfactual (log) number of deaths was higher than the actual values, demonstrating that the vaccination campaign was effective in saving lives. The bottom-right graph shows, instead, the point-wise estimate of

[12] Overall scores are far worse when including a linear trend as a covariate in the SOGP, so we decided to drop it.

$\tau^*_{i,t}$, the multiplicative causal effect with 95% credible regions. To understand the average effect right after the campaign, one can calculate the average effect as in (3.7). Over the whole period, on average, for every two Covid-19 deaths in the other 4 countries used, there was one Covid-19 death in the UK (that is, the ratio $\bar{\tau}^*_i$ is 51.41% [30.05%, 82.86%]). In a Bayesian fashion, we also take into account parameter uncertainty, especially the one deriving from the coregionalization matrix B. In Figure 3.5, the darker grey areas represent the supplemental uncertainty coming from parameter estimates, in particular due to the coregionalisation matrix.

Weekly infection rate R

The same analysis is run for the weekly infection rate, to measure if a different vaccination campaign is actually producing slower rates of infectiousness.[13] Half of the countries selected ('Portugal' and 'Ireland') are in common with the previous analysis, while 'France' and 'Denmark' are new control units. We can now employ type II Maximum likelihood on the whole dataset to find the optimal hyper-parameters of the kernel. Given that the reproduction rate is an estimate itself, it can be affected by many sources of variability induced by the data or the model used. To take into account this effect, we bound the likelihood variance (observation error) to a minimum value of around 0.01, which corresponds to 5% of countries' reproduction rate variance over the observed period. In the Bayesian setting, we set up a slightly more informative prior, $\omega_i^2 \sim \mathcal{G}(0.1, 1)$, to support the same decision. Without these restrictions the model converges to a data representation with an almost zero observational error. This reflects in a tight in-sample fit but poor out-of-sample performances.
In contrast to the weekly number of deaths, the time component does not seem to play a prevailing role in defining R dynamics. Looking at Figure 3.6, one can note that there is mainly a contemporaneous effect, as the lengthscale ℓ of the Matèrn kernel (3.15) is lower compared to the one estimated in Section 3.6.2. Decreasing the length parameter reduces the banding, as points further away from each other become less correlated. This means that data points have zero covariance with the lagged version of both the treated variable and control units. This effect is further minimized for 'France' (fifth country) as the estimate of $B_2^{[5,5]}$ is more than half that of other countries. In contrast, Google mobility data, i.e. which places people were visiting during the period, was effective in predicting how the contagiousness would change. However, as the variable is a principal component transformation, little interpretation can be given to individual venues. The causal effect is then estimated on the available dataset. As shown in Figure 3.7, before the vaccination campaign the model provides a good fit of the data. Afterwards, counterfactual predictions (orange lines) initially follow the main trend of observed data but then they start to deviate marginally. However, the variability surrounding the estimate is too broad to confirm any causal effect in the data. It is interesting to note that in the very final period of the analysis, starting May 2021 (the red line in Figure 3.7), the model predicts a

[13] We removed the number of test variables as it did not affect results, making the optimization more challenging.

stable reproduction rate. However, the data seems to diverge positively to higher values. We suspect that the reason for this discrepancy is the spread of the *delta* variant. This more contagious version of Covid-19 represented 73% of UK cases by the end of May 2021. Nevertheless, in the period following the vaccination campaign the average additive causal effect $\overline{\tau}_i$ equals 0.0063 with 95% credible interval [0.1653, -0.1582], thus no reduction is detected.[14]

3.7 Conclusion

A growing literature on applied causal inference indicates an increasing interest in evaluating the incremental impact of interventions and policies, especially during the Covid-19 pandemic (Li et al., 2021; Ma et al., 2022; Tang et al., 2022). With this paper, we propose a novel approach to obtain the counterfactual prediction of the unobserved outcome. We employ a Bayesian Machine Learning technique, based on Gaussian Processes, whose main features are discussed below.

The prevailing literature on dynamical Bayesian causal models revolves around Brodersen et al. (2014), whose approach is based on state-space models, which easily lend themselves to posterior inference. However, since closed-form solutions for the posterior are challenging to obtain, the authors resort to stochastic approximation, using MCMC.

In the general form of Gaussian Processes, causal effect posterior evaluation can be instead derived analytically. When this is not possible - for example, when using transformation of variables or cumulative measures - one can employ a GP sampling procedure. As a GP is fully characterized by its mean (generally 0) and its variance (the kernel), the process is straightforward.

Furthermore, state-space models impose some restrictions on the dynamic evolution of the states, notably, linearity. With GPs, the input $\mathbf{x}$ is transformed into a feature vector $f(\mathbf{x})$ through some non-linear mappings dictated by the structure of the kernel. In a time series, the degree of correlation between a variable and its lag is given by the relative time distance. When using a Matèrn kernel, it can be shown that the covariance matrix of the kernel gives rise to a particular form of a continuous-time AR(p) Gaussian process (Rasmussen and Williams, 2006). At the same time, nothing prevents us from using a more complex structure - such as periodic, linear, etc., or a combination thereof - to better fit the time curve. Furthermore, the linearity assumption with exogenous regressors embodied in state space can be relaxed by adopting an appropriate kernel architecture. Machine learning tools, such as cross-validation, can help decide which one better describes the data.

Another important improvement that GPs put forward is a *heterotopic* configuration of data, i.e. each output has a different training set with a potentially different number of samples. This approach plays a crucial role in causal analysis, since generally one has to discard all information after the intervention period to train the data, generating a non-negligible loss of data. In addition, no time matching is needed as the model understands

[14] On the contrary, it is slightly positive, although not statistically significant.

the relationship among the potential outcome and the explanatory variables for each unit, independently if these variables match in absolute time.
Lastly, using GPs one can easily quantify uncertainty around a measurement or prediction, since every data point possesses a defined distribution. This promotes direct estimation of the causal effect distribution, means, and quantiles.
To test this model in practice, we estimated the effect of the UK vaccination policy compared to other European countries. In particular, we analysed how the UK's faster inoculation campaign affected the cumulative number of deaths and the rate of contagiousness, as measured by the reproduction rate. The results suggested that vaccinations prevented deaths, since, on average, in the first semester of 2021, every death in the UK related to Covid-19 corresponded to two deaths in the rest of Europe. No statistically significant evidence was found to justify the proposition that vaccines reduce the number of cases directly caused by an infected individual. However, this has to be considered in light of the new and more infectious variants that started spreading over the continent at the end of our sample period.

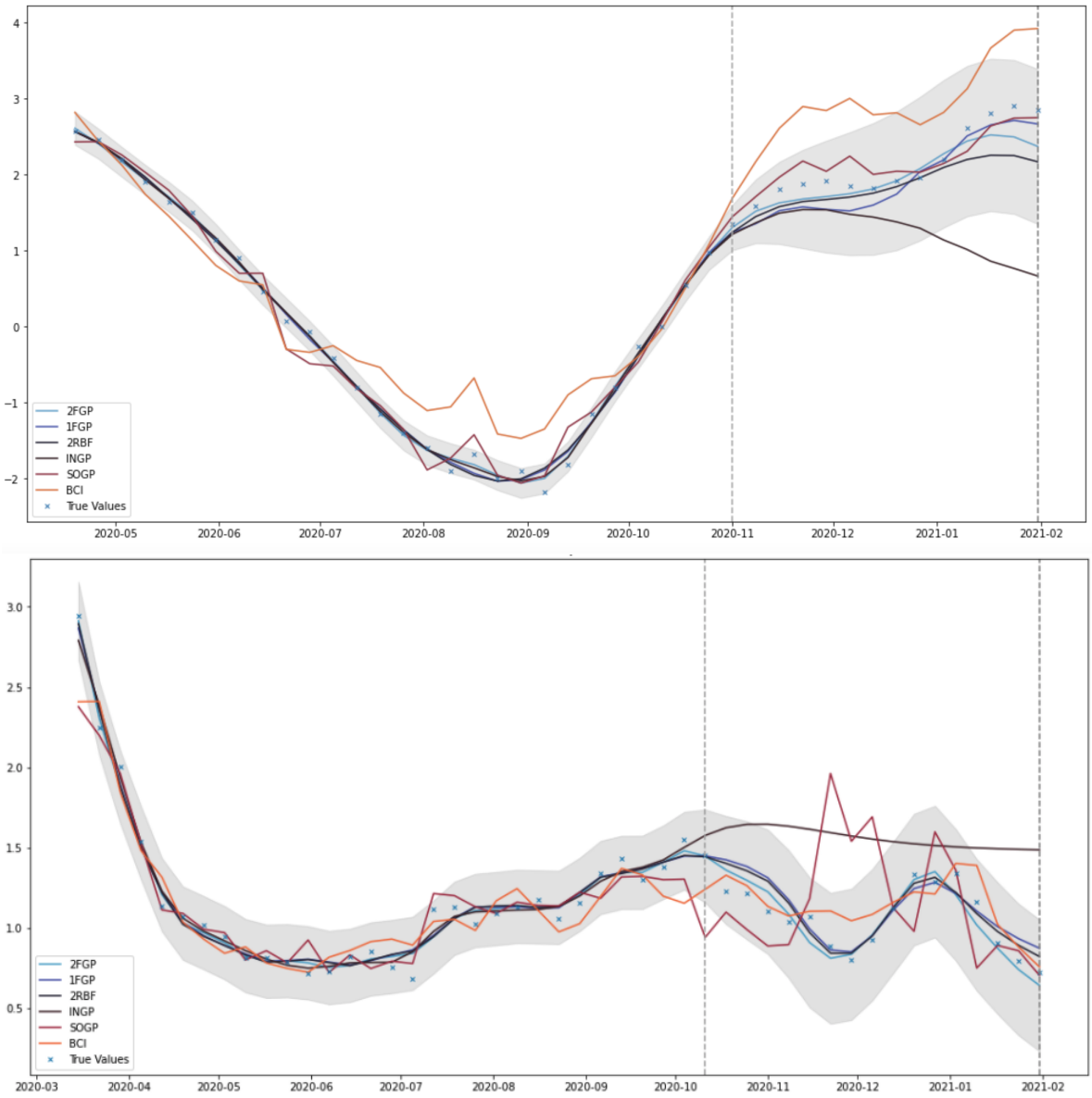

Figure 3.3: *Top: Predicted log weekly deaths per 1,000 people for the different models. Bottom: Predicted rate of infectiousness for the different models. The blue crosses represent observed data, and each line corresponds to a specific model. The first vertical grey dotted line is t^* which separates the training and test datasets. The second vertical dashed line identifies t_0. The grey shaded area represents the 95% prediction interval of the 2FGP, the base model.*

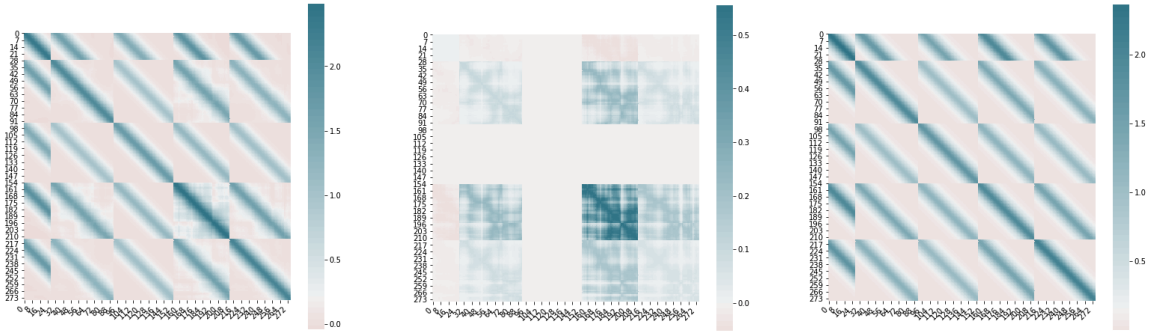

Figure 3.4: *The figure on the left is the total variance $\mathcal{K}$ as in* (3.20)*, while the other two represent the component kernel on the covariate space $K_{rbf}(X_\circ^{/t}, X_\circ^{/t})$ and time space $B_2 \otimes K_{Mat}(t_\circ, t_\circ)$,, respectively.*

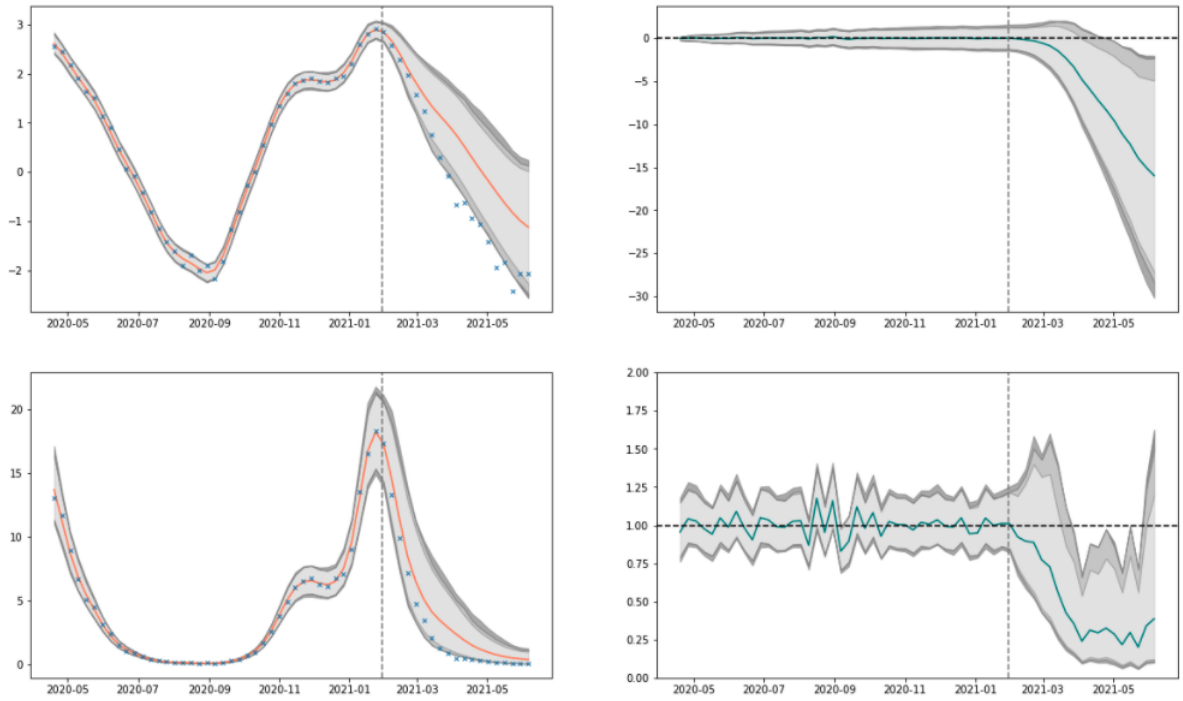

Figure 3.5: *The graphs on the left indicate the log weekly deaths on top and the level on the bottom. Data provided to train the model is at the left of t_0, the grey vertical line. Orange line represents model's predicted average of (log) weekly deaths. True values are in blue. On the right-hand part, the top graphs show the cumulative effect of log weekly deaths $\mathcal{T}_{i,t}$ while the bottom displays point-wise multiplicative causal effect $\tau^*_{i,t}$. Shaded area represents 95% credible intervals in grey. Light grey does not take into account parameter uncertainty, grey accounts only for $\boldsymbol{\lambda}$ and dark grey for $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ uncertainty.*

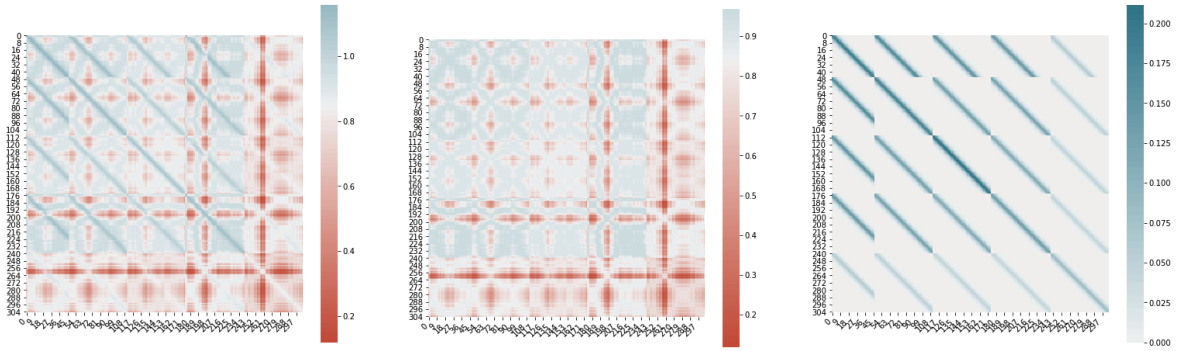

Figure 3.6: *The figure on the left is the total variance $\mathcal{K}$ as in* (3.20)*, while the other two represent the component kernel on the covariate space $K_{rbf}(X_\circ^{/t}, X_\circ^{/t})$ and time space $B_2 \otimes K_{Mat}(t_\circ, t_\circ)$,, respectively.*

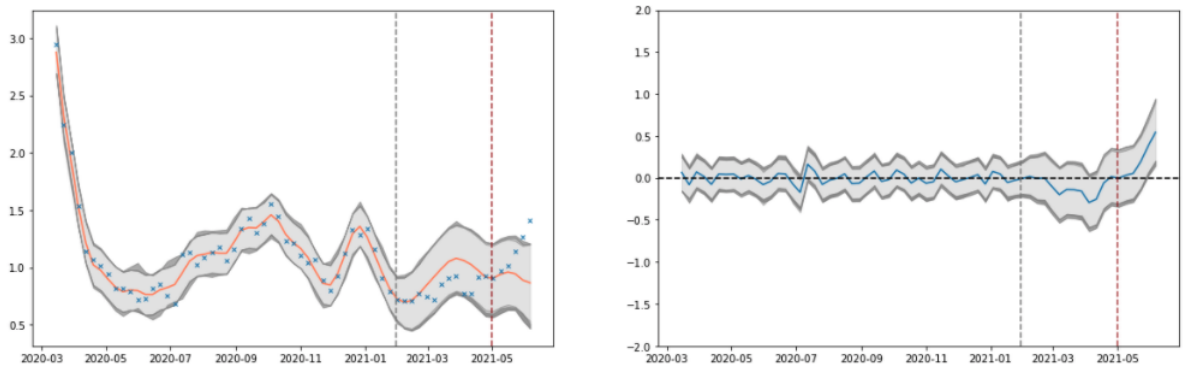

Figure 3.7: *The graph on the left indicates the evolution of the reproduction rate in the UK. Data provided to train the model is at the left of t_0, the grey vertical line. The orange line represents the predicted average of R with 95% credible intervals in grey. True values are in blue. On the right-hand part, the graph displays point-wise additive causal effect τ_t, with relative 95% bands in grey. The red vertical line at time '01-05-2020' indicates the start of the delta variant transmission in the UK.*

List of Acronyms

2SPCA: Two-step Principal Component Analysis.

ARMA: AutoRegressive Moving Average.

ARCH: AutoRegressive Conditional Heteroskedasticity.

ARD: Automatic Relevance Determination.

AV: AVerage.

BCI: Bayesian Causal Impact.

BFGS: Broyden–Fletcher–Goldfarb– Shanno.

CHDFM: Conditionally Heteroskedastic Dynamic Factor Model.

CM: Conditional Maximization.

DCC: Dynamic Conditional Correlation.

DD: Difference-in-Difference.

DFM: Dynamic Factor Model.

DQ: Dynamic Quantile.

DTW: Dynamic Time Warping.

EDF: Empirical Density Function.

ECDF: Empirical Cumulative Density Function.

ECME: Expectation Conditional Maximization Either.

EM: Expectation Maximization.

ES: Energy Score.

EWP: Equally Weighted Portfolio.

GaR: Growth at Risk.

GARCH: Generalized AutoRegressive Conditional Heteroskedasticity.

GDFM: Generalized Dynamic Factor Models.

GDP: Gross Domestic Product.

GP: Gaussian Process.

HMC: Hamiltonian Monte Carlo.

ICM: Intrinsic Coregionalization Model.

i.i.d.: Independent and Identically Distributed.

IR: Information Ratio.

L-BFGS: Limited memory Broyden–Fletcher–Goldfarb– Shanno.

LMC: Linear Model of Coregionalization.

LogS: Log-Score.

MCMC: Markov Chain Monte Carlo.

MGARCH: Mean Generalized AutoRegressive Conditional Heteroskedasticity.

MLE: Maximum Likelihood Estimation.

ML-II: Type II Maximum Likelihood .

MMSE: Miimum Mean Squared Error.

MOGP: Multi-Output Gaussian Process.

MSE: Mean Squared Error.

MVP: Minimum Variance Portfolio.

MW: Maximum Weight.

NID: Normal and Identically Distributed.

OECD: Organisation for Economic Co-operation and Development.

OU: Ornstein–Uhlenbeck.

PCA: Principal Component Analysis.

p.d.f.: Probability Density Function.

QMLE: Quasi Maximum Likelihood Estimation.

rbf: Radial Basis Function.

SLFM: Semiparametric Latent Factor Model.

SNR: Signal-to-Noise Ratio.

SOGP: Single-Output Gaussian Process.

SR: Sharpe Ratio.

SSC: Short Selling Costs.

SUVTVA: Stable Unit Treatment Value Assumption.

TC:Turnover Costs.

TL: Tick Loss.

VAR: Vector AutoRegression.

VTE: Variance Targeting Estimator.

Bibliography

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1):113–132.

Adrian, T., Boyarchenko, N., and Giannone, D. (2019). Vulnerable growth. *American Economic Review*, 109(4):1263–89.

Aglietti, V., Damoulas, T., Álvarez, M., and González, J. (2020). Multi-task causal learning with Gaussian processes. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6293–6304. Curran Associates, Inc.

Aguilar, M. (2009). A Latent Factor Model of Multivariate Conditional Heteroscedasticity. *Journal of Financial Econometrics*, 7(4):481–503.

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

Aielli, G. P. (2013). Dynamic conditional correlation: On properties and estimation. *Journal of Business & Economic Statistics*, 31(3):282–299.

Ait-Sahalia, Y. and Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, 201(2):384–399.

Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alvarez, M. A. and Lawrence, N. D. (2009). Sparse convolved multiple output Gaussian processes.

Alvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: a review.

Asai, M., McAleer, M., and Yu, J. (2006). Multivariate stochastic volatility: A review. *Econometric Reviews*, 25(2-3):145–175.

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

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40:436–465.

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

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 317:146–304.

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

Bai, J. and Wang, P. (2015). Identification and Bayesian estimation of dynamic factor models. *Journal of Business & Economic Statistics*, 33(2):221–240.

Barigozzi, M. and Hallin, M. (2017a). Generalized dynamic factor models and volatilities: estimation and forecasting. *Journal of Econometrics*, 201(2):307–321.

Barigozzi, M. and Hallin, M. (2017b). A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):581–605.

Barigozzi, M., Hallin, M., and Soccorsi, S. (2019). Identification of Global and Local Shocks in International Financial Markets via General Dynamic Factor Models. *Journal of Financial Econometrics*, 17(3):462–494.

Barigozzi, M. and Luciani, M. (2019). Qasi maximum likelihood estimation and inference of large approximate dynamic factor models via the EM algorithm.

Bojinov, I. and Shephard, N. (2019). Time series experiments and causal estimands: Exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bonilla, E. V., Chai, K., and Williams, C. (2008). Multi-task Gaussian process prediction. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Born, B., Dietrich, A., and Müller, G. (2020). The lockdown effect: A counterfactual for Sweden. CEPR Discussion Papers 14744, C.E.P.R. Discussion Papers.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2014). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274.

Brownlees, C. and Engle, R. F. (2016). SRISK: A Conditional Capital Shortfall Measure of Systemic Risk. *The Review of Financial Studies*, 30(1):48–79.

Brownlees, C. and Llorens-Terrazas, J. (2022). Projected Dynamic Conditional Correlations. *Available at SSRN: https://ssrn.com/abstract=3576985*.

Brownlees, C. and Souza, A. B. (2021). Backtesting global Growth-at-Risk. *Journal of Monetary Economics*, 118(C):312–330.

Burmeister, E., Wall, K. D., and Hamilton, J. D. (1986). Estimation of unobserved expected monthly inflation using Kalman filtering. *Journal of Business & Economic Statistics*, 4(2):147–160.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, 16:1190–1208.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230. Themed Issue: Treatment Effect 1.

Calzolari, G., Fiorentini, G., and Sentana, E. (2004). Constrained Indirect Estimation. *The Review of Economic Studies*, 71(4):945–973.

Carriero, A., Clark, T. E., and Marcellino, M. (2020). Capturing Macroeconomic Tail Risks with Bayesian Vector Autoregressions. Working Papers 20-02R, Federal Reserve Bank of Cleveland.

Cawley, G. C. and Talbot, N. L. C. (2007). Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research*, 8(31):841–861.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, pages 1281–1304.

Chatzilena, A., Demiris, N., and Kalogeropoulos, K. (2022). A modelling framework for the analysis of the transmission of sars-cov2.

Chicheportiche, R. and Bouchaud, J.-P. (2015). A nested factor model for non-linear dependencies in stock returns. *Quantitative Finance*, 15(11):1789–1804.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4):841–862.

Connor, G. and Korajczyk, R. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance*, 48(4):1263–91.

Connor, G., Korajczyk, R., and Linton, O. (2006). The common and specific components of dynamic volatility. *Journal of Econometrics*, 132(1):231–255.

De Nard, G., Ledoit, O., and Wolf, M. (2019). Factor Models for Portfolio Selection in Large Dimensions: The Good, the Better and the Ugly. *Journal of Financial Econometrics*, 19(2):236–257.

DeMiguel, V., Garlappi, L., and Uppal, R. (2007). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies*, 22(5):1915–1953.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 1–38.

Diebold, F. X. and Nerlove, M. (1989). The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model. *Journal of Applied Econometrics*, 4(1):1–21.

Ding, Y., Li, Y., and Zheng, X. (2021). High dimensional minimum variance portfolio estimation under statistical factor models. *Journal of Econometrics*, 222(1, Part B):502–515. Annals Issue:Financial Econometrics in the Age of the Digital Economy.

Djolonga, J., Krause, A., and Cevher, V. (2013). High-dimensional Gaussian process bandits. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Doz, C., Giannone, D., and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164:188–205.

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

Engle, R. (1987). *Multivariate Arch with Factor Structures: Cointegration in Variance*. Discussion paper- Department of Economics University of California San Diego. University of California Press.

Engle, R. (2002). Dynamic conditional correlation - a simple class of multivariate GARCH models. *Journal of Business and Economic Statistics*, 20:339–350.

Engle, R. (2009). *Anticipating Correlations: A New Paradigm for Risk Management*. Princeton University Press.

Engle, R., Ng, V. K., and Rothschild, M. (1990). Asset pricing with a factor-Arch covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics*, 45(1-2):213–237.

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

Engle, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics*, 22(4):367–381.

Engle, R. F. and Sheppard, K. (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Working Paper 8554, National Bureau of Economic Research.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fan, J., Li, Y., and Yu, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, 107(497):412–428.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:603–680.

Fan, J., Liao, Y., and Shi, X. (2015). Risks of large portfolios. *Journal of Econometrics*, 186(2):367–387.

Fiorentini, G., Sentana, E., and Shephard, N. (2004). Likelihood-based estimation of latent generalized arch structures. *Econometrica*, 72(5):1481–1517.

Fletcher, R. (2000). *Newton-Like Methods*, chapter 3, pages 44–79. John Wiley & Sons, Ltd.

Forni, M., Giovannelli, A., Lippi, M., and Soccorsi, S. (2018). Dynamic factor model with infinite-dimensional factor space: Forecasting. *Journal of Applied Econometrics*, 33(5):625–642.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The Generalized Dynamic Factor Model: Identification and estimation. *The Review of Economics and Statistics*, 82:540–554.

Forni, M. and Lippi, M. (2001). The Generalized Dynamic Factor Model: Representation theory. *Econometric Theory*, 17:1113–1141.

Francq, C., Horváth, L., and Zakoïan, J.-M. (2011). Merits and Drawbacks of Variance Targeting in GARCH Models. *Journal of Financial Econometrics*, 9(4):619–656.

Francq, C. and Zakoïan, J.-M. (2010). *Estimating GARCH Models by Qasi-Maximum Likelihood*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

Ghayda, R. A., Lee, K. H., Han, Y. J., Ryu, S., Hong, S. H., Yoon, S., Jeong, G. H., Lee, J., Lee, J. Y., Yang, J. W., Effenberger, M., Eisenhut, M., Kronbichler, A., Solmi, M., Li, H., Jacob, L., Koyanagi, A., Radua, J., Shin, J. I., and Smith, L. (2020). Estimation of global case fatality rate of coronavirus disease 2019 (COVID-19) using meta-analyses: Comparison between calendar date and days since the outbreak of the first confirmed case. *International Journal of Infectious Diseases*, 100:302–308.

Ghysels, E., Iania, L., and Striaukas, J. (2018). Quantile-based inflation risk models. NBB Working Paper 349, Brussels.

Giacomini, R. and Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics*, 23(4):416–431.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7):1–24.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

González-Rivera, G., Lee, T.-H., and Mishra, S. (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of Forecasting*, 20(4):629–645.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

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

Harvey, A., Ruiz, E., and Sentana, E. (1992). Unobserved component time series models with ARCH disturbances. *Journal of Econometrics*, 52:129–157.

Hayashi, T. (2018). On a norm inequality for a positive block-matrix. *Linear Algebra and its Applications*, 566:86–97.

Her, P. H., Saeed, S., Tram, K. H., and Bhatnagar, S. R. (2022). Novel mobility index tracks COVID-19 transmission following stay-at-home orders. *Scientific reports*, 12:7654.

Holmes, E. E. (2013). Derivation of an EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90:1–37.

King, M., Sentana, E., and Wadhwani, S. (1994). Volatility and links between national stock markets. *Econometrica*, 62(4):901–933.

Koopman, S. J., Mallee, M. I. P., and der Wel, M. V. (2010). Analyzing the term structure of interest rates using the dynamic Nelson–Siegel model with time-varying parameters. *Journal of Business & Economic Statistics*, 28(3):329–343.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84.

Laurent, S., Bauwens, L., and Rombouts, J. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, 21(1):79–109.

Li, Z., Xu, T., Zhang, K., Deng, H.-W., Boerwinkle, E., and Xiong, M. (2021). Causal analysis of health interventions and environments for influencing the spread of COVID-19 in the United States of America. *Frontiers in Applied Mathematics and Statistics*, 6.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81(4):633–648.

Liu, H., Cai, J., and Ong, Y.-S. (2018). Remarks on multi-output Gaussian process regression. *Knowledge-Based Systems*, 144:102–121.

Ma, J., Dong, Y., Huang, Z., Mietchen, D., and Li, J. (2022). Assessing the causal impact of COVID-19 related policies on outbreak dynamics: A case study in the US. WWW '22, page 2678–2686, New York, NY, USA. Association for Computing Machinery.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

McAleer, M. (2019). What they did not tell you about algebraic (non-) existence, mathematical (ir-)regularity, and (non-) asymptotic properties of the dynamic conditional correlation (DCC) model. *Journal of Risk and Financial Management*, 12(2).

Menchetti, F. and Bojinov, I. (2020). Estimating the effectiveness of permanent price reductions for competing products using multivariate Bayesian structural time series models. *Harvard Business School Working Paper*, 21-048.

Menchetti, F., Cipollini, F., and Mealli, F. (2021). Estimating the causal effect of an intervention in a time series setting: the C-ARIMA approach.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278.

Mhanna, A. (2015). Symmetric norm inequalities and positive semi-definite block-matrices.

Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.].

Ng, V., Engle, R. F., and Rothschild, M. (1992). A multi-dynamic-factor model for stock returns. *Journal of Econometrics*, 52(1):245–266.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York, NY, USA, second edition.

Normandin, M. (2004). Canadian and u.s. financial markets: Testing the international integration hypothesis under time-varying conditional volatility. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 37(4):1021–1041.

Normandin, M. and Phaneuf, L. (2004). Monetary policy shocks:: Testing identification conditions under time-varying conditional volatility. *Journal of Monetary Economics*, 51(6):1217–1243.

Pantula, S. and Fuller, W. (1986). Computational algorithms for the factor model. *Communications in Statistics. Simulation and Computation*, 15(1):227–259.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*. use R! Springer-Verlag, New York.

Plagborg-Møller, M., Reichlin, L., Ricco, G., and Hasenzagl, T. (2020). When is growth at risk? *Brookings Papers on Economic Activity*, 2020(Spring):167–229. Replication code and data (6 GB).

Prasad, A., Jeasakul, P., Alter, A., Lafarguette, R., Xiaochen Feng, A., Elekdag, S., and Wang, C. (2019). Growth at risk: Concept and application in IMF country surveillance. IMF Working Papers 2019/036, International Monetary Fund.

Radhakrishna Rao, C. (2000). Statistical proofs of some matrix inequalities. *Linear Algebra and its Applications*, 321(1):307 – 320. Eighth Special Issue on Linear Algebra and Statistics.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Sentana, E. (1992). Identification of multivariate conditionally heteroskedastic factor models. *Discussion Paper 139, Financial Markets Group, London School of Economics.*

Sentana, E., Calzolari, G., and Fiorentini, G. (2008). Indirect estimation of large conditionally heteroskedastic factor models, with an application to the Dow 30 stocks. *Journal of Econometrics*, 146(1):10–25.

Sentana, E. and Fiorentini, G. (2001). Identification, estimation and testing of conditionally heteroskedastic factor models. *Journal of Econometrics*, 102(2):143–164.

Shumway, R. H. and Stoffer, D. S. (2006). *Time series analysis and its applications: with R examples.* Springer.

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

Sánchez, A. C. and Röhn, O. (2016). How do policies influence GDP tail risks? OECD Economics Department Working Papers 1339, OECD Publishing.

Sävje, F., Aronow, P. M., and Hudgens, M. G. (2019). Average treatment effects in the presence of unknown interference.

Tang, W.-X., Li, H., Hai, M., and Zhang, Y. (2022). Causal analysis of impact factors of COVID-19 in China. *Procedia Computer Science*, 199:1483–1489. The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy after COVID-19.

Teh, Y. W., Seeger, M., and Jordan, M. I. (2005). Semiparametric latent factor models. In Cowell, R. G. and Ghahramani, Z., editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 333–340. PMLR. Reissued by PMLR on 30 March 2021.

Tripathi, G. (1999). A matrix extension of the cauchy-schwarz inequality. *Economics Letters*, 63(1):1–3.

Tripathy, R., Bilionis, I., and Gonzalez, M. (2016). Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *Journal of Computational Physics*, 321:191–223.

Trucíos, C., Mazzeu, J. H. G., Hallin, M., Hotta, L. K., Pereira, P. L. V., and Zevallos, M. (2021). Forecasting conditional covariance matrices in high-dimensional time series: A general dynamic factor approach. *Journal of Business & Economic Statistics*, 0(0):1–13.

Tse, Y. K. and Tsui, A. K. C. (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business & Economic Statistics*, 20(3):351–362.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103.

Zheng, C., Shao, W., Chen, X., Zhang, B., Wang, G., and Zhang, W. (2022). Real-world effectiveness of COVID-19 vaccines: a literature review and meta-analysis. *International Journal of Infectious Diseases*, 114:252–260.