Università
Ca'Foscari
Venezia

Corso di Dottorato di ricerca in Economia
Doctorat en Mathématiques Appliquées
cotutela con Université Paris I - Panthéon-Sorbonne

Ciclo XXX

Tesi di Ricerca

Essays on the econometric modelling of temporal networks

SSD: SECS-P/05

Coordinatore del Dottorato
prof. Giacomo Pasini

Primo Supervisore
prof. Monica Billio

Secondo Supervisore
prof. Dominique Guégan

Terzo Supervisore
prof. Roberto Casarin

Dottorando
Matteo Iacopini
Matricola 956154

Università
Ca'Foscari
Venezia

UNIVERSITÉ FRANCO ITALIENNE
UNIVERSITÀ ITALO FRANCESE

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

CA' FOSCARI UNIVERSITY OF VENICE
AND
UNIVERSITÉ PARIS I PANTHÉON-SORBONNE

DOCTORAL THESIS

Essays on the econometric modelling of temporal networks

Author:
Matteo IACOPINI

Supervisors:
Prof. Monica BILLIO
Prof. Dominique GUÉGAN
Prof. Roberto CASARIN

A thesis submitted in fulfilment of
the requirements for the degree of

PhD in Economics
Doctorat en Mathématiques Appliquées

in the

Department of Economics
and
Centre d'Économie de la Sorbonne

July 5, 2018

Declaration of Authorship

I, Matteo IACOPINI, declare that this thesis titled, "Essays on the econometric modelling of temporal networks" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at these Universities.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at these Universities or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"God used beautiful mathematics in creating the world."

Paul Dirac

"Mathematics is the language in which God has written the universe."

Galileo Galilei

"The most important questions of life are, for the most part, really only problems of probability."

Pierre Simon, Marquis de Laplace

CA' FOSCARI UNIVERSITY OF VENICE
AND
UNIVERSITÉ PARIS I PANTHÉON-SORBONNE

Abstract

Department of Economics

PhD in Economics
Doctorat en Mathématiques Appliquées

Essays on the econometric modelling of temporal networks

by Matteo IACOPINI

Graph theory has long been studied in mathematics and probability as a tool for describing dependence between nodes. However, only recently it has been implemented on data, givin birth to the statistical analysis of real networks.

The topology of economic and financial networks is remarkably complex: it is generally unobserved, thus requiring adequate inferential procedures for it estimation, moreover not only the nodes, but the structure of dependence itself evolves over time. Statistical and econometric tools for modelling the dynamics of change of the network structure are lacking, despite their increasing requirement in several fields of research. At the same time, with the beginning of the era of "*Big data*" the size of available datasets is becoming increasingly high and their internal structure is growing in complexity, hampering traditional inferential processes in multiple cases.

This thesis aims at contributing to this newborn field of literature which joins probability, economics, physics and sociology by proposing novel statistical and econometric methodologies for the study of the temporal evolution of network structures of medium-high dimension.

CA' FOSCARI UNIVERSITY OF VENICE
AND
UNIVERSITÉ PARIS I PANTHÉON-SORBONNE

Abstract

Department of Economics

PhD in Economics
Doctorat en Mathématiques Appliquées

Essays on the econometric modelling of temporal networks

by Matteo IACOPINI

La théorie des graphes a longtemps été étudiée en mathématiques et en probabilité en tant qu'outil pour décrire la dépendance entre les nœuds. Cependant, ce n'est que récemment qu'elle a été mise en œuvre sur des données, donnant naissance à l'analyse statistique des réseaux réels.

La topologie des réseaux économiques et financiers est remarquablement complexe: elle n'est généralement pas observée, et elle nécessite ainsi des procédures inférentielles adéquates pour son estimation, d'ailleurs non seulement les nœuds, mais la structure de la dépendance elle-même évolue dans le temps. Des outils statistiques et économétriques pour modéliser la dynamique de changement de la structure du réseau font défaut, malgré leurs besoins croissants dans plusieurs domaines de recherche. En même temps, avec le début de l'ère des "*Big data*", la taille des ensembles de données disponibles devient de plus en plus élevée et leur structure interne devient de plus en plus complexe, entravant les processus inférentiels traditionnels dans plusieurs cas.

Cette thèse a pour but de contribuer à ce nouveau champ littéraire qui associe probabilités, économie, physique et sociologie en proposant de nouvelles méthodologies statistiques et économétriques pour l'étude de l'évolution temporelle des structures en réseau de moyenne et haute dimension.

Contents

List of Figures

List of Tables

List of Symbols

List of Symbols

Variables

$\mathcal{X}$	N-dimensional tensor (or array), $N \geq 3$
$\mathbf{X}$	2-dimensional tensor (matrix)
$\mathbf{x}$	1-dimensional tensor (vector)
x	0-dimensional tensor (scalar)
$\mathbf{X}_{(\mathscr{R},\mathscr{C})}$	matricization of a tensor $\mathcal{X}$ according to the row and column index sets $\mathscr{R}$ and $\mathscr{C}$
$\mathbf{X}_{(n)}$	mode-n matricization of a tensor $\mathcal{X}$
$\mathcal{X}_{i_1,\ldots,i_N}$	entry $i_1, \ldots, i_N$ of the tensor $\mathcal{X}$
$\mathbf{X}_{i_1,i_2}$	entry i_1, i_2 of the matrix $\mathbf{X}$
$\mathbf{x}_{i_1}$	entry i_1 of the vector $\mathbf{x}$
α	parameter scalar
$\boldsymbol{\alpha}$	parameter vector
$\mathbf{A}$	parameter matrix
$\mathcal{A}$	parameter tensor
$\boldsymbol{\theta}$	set of the model's parameters
$\hat{\alpha}$	estimator of α
$\mathbf{I}_n$	identity matrix of dimensions $n \times n$
$\mathrm{diag}\,(\ldots)$	diagonal matrix with diagonal elements ...

Operators

$\cdot$	matrix product
$\otimes$	Kronecker product
$\odot$	Hadamard product
$\circ$	outer product
$\langle \cdot, \cdot \rangle$	scalar (inner) product
$\bar{\times}^n$	mode-n contracted product between two tensors
$\times$	Cartesian product
$\times^n$	mode-n product between a tensor and a matrix
$\times_n$	mode-n product between a tensor and a vector
$\mathrm{vec}\,(\cdot)$	vectorization operator for N-dimensional tensors, $N \geq 2$
$\mathbf{X}', \mathbf{x}'$	transpose of matrix, vector
$\mathbf{X}^{-1}$	inverse of matrix
$\mathcal{X}^{T(\sigma)}$	transpose of tensor according to the permutation σ

$\\|\cdot\\|_p$	L_p-norm of a vector, matrix, or tensor
$\lvert\mathbf{X}\rvert$	determinant of the matrix $\mathbf{X}$
$\lvert x\rvert$	absolute value of the scalar x
$\mathrm{tr}\,(\mathbf{X})$	trace of the matrix $\mathbf{X}$
Probability	
$\pi(x)$	prior probability density function of the random variable x
$p(x,y)$	joint probability density function of (x,y)
$p(x\mid y)$	conditional probability density function of x given y
$p(x)$	marginal probability density function of x
$C(u,v)$	copula distribution function of (u,v)
$c(u,v)$	copula probability density function of (u,v)
$L(\mathbf{x}\mid\boldsymbol{\theta})$	likelihood function of data $\mathbf{x}$, given the parameters $\boldsymbol{\theta}$
$\mathbb{P}(\cdot)$	probability distribution induced by the random variable x
$\mathbb{E}[\cdot]$	expected value
$\delta_A(\cdot)$	Dirac mass on the set A
$\sim$	"distributed as"
$\overset{iid}{\sim}$	"independent and identically distributed as"
$\Gamma(\cdot)$	univariate Gamma function
Distributions	
$\mathcal{N}(\mu,\sigma^2)$	Normal distribution
$\mathcal{N}_{I_1}(\boldsymbol{\mu},\Sigma)$	multivariate Normal distribution
$\mathcal{N}_{I_1,I_2}(\mathbf{M},\Sigma_1,\Sigma_2)$	matrix-variate Normal distribution
$\mathcal{N}_{I_1,\ldots,I_N}(\mathcal{M},\Sigma_1,\ldots,\Sigma_N)$	N-dimensional tensor Normal distribution
$\mathcal{G}a(\alpha,\beta)$	Gamma distribution, shape-rate formulation (mean α/β)
$\mathcal{IG}(\alpha,\beta)$	inverse Gamma distribution, shape-scale formulation
$\mathcal{E}xp(\lambda)$	exponential distribution (mean $1/\lambda$)
$\mathcal{B}e(\alpha,\beta)$	beta distribution
$\mathcal{U}(a,b)$	continuous uniform distribution on $[a,b]$
$\mathcal{D}ir(\alpha_1,\ldots,\alpha_D)$	D-dimensional Dirichlet distribution
$\mathrm{Laplace}(\mu,\sigma)$	Laplace distribution, location-scale formulation
$\mathrm{Logistic}(\mu,\sigma)$	logistic distribution, location-scale formulation
$\mathrm{GPD}(\mu,\sigma,\xi)$	generalized Pareto distribution, location-scale-shape formulation
$\mathrm{Lomax}(\lambda,\alpha)$	Lomax distribution, scale-shape formulation
$\mathrm{GiG}(p,a,b)$	generalized inverse Gaussian distribution
$\mathcal{B}ern(p)$	Bernoulli distribution
$\mathcal{W}_M(\nu,\Psi)$	M-dimensional Wishart distribution
$\mathcal{IW}_M(\nu,\Psi)$	M-dimensional inverse Wishart distribution
Sets	
$\mathbb{R}$	Euclidean space of the real numbers

$\mathbb{R}_+$	space of positive real numbers
$\bar{\mathbb{R}}$	extended real numbers, i.e. $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$
$\mathbb{R}^{I_1}$	I_1-dimensional Euclidean space or Euclidean space of the real-valued vectors of length I_1
$\mathbb{R}^{I_1 \times I_2}$	Euclidean space of the real-valued matrices of dimensions $I_1 \times I_2$
$\mathbb{R}^{I_1 \times \ldots \times I_N}$	Euclidean space of the real-valued matrices of dimensions $I_1 \times \ldots \times I_N$
$\mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2}$	tensor product of the spaces $\mathbb{R}^{I_1}, \mathbb{R}^{I_2}$
$\boldsymbol{\Theta}$	parameter space
$\{x_t\}_{t=1}^{T}$	set of values $\{x_1, \ldots, x_T$
$\mathbb{1}_A(\cdot)$	indicator function of the set A
#	cardinality of a set
$\in, \notin$	belongs to, does not belong to
$\cap$	intersection
$\cup$	union
$\subseteq, \subset$	subset of, proper subset of
$\mathbb{X} \cong_f \mathbb{Y}$	spaces $\mathbb{X}, \mathbb{Y}$ are isometric, with isometry $f : \mathbb{X} \to \mathbb{Y}$

To my Father,

for His love, His forgiveness and for being always with me

to my family,

for their support and for having believed in me,

to the people who love me

for the joy, case and love they gave me. . .

Chapter 1

Introduction

The secret of getting ahead is getting started.

MARK TWAIN

Statistics is the grammar of science.

KARL PEARSON

Mathematicians do not study objects, but relations between objects.

HENRI POINCARÉ

1.1 Preliminaries

This section is devoted to the introduction and description of some classes of complex or structured datasets, whose relevance to the econometric and statistical research has greatly increased in the last years. "*Structured*" data are those possessing an intrinsic or natural structure, which is reflected, for instance, in the way the data is collected (e.g., panel data) or in the type of object under study (e.g., images, input-output tables, or even three-dimensional objects). The term "*complex*" instead refers to the situation in which the data cannot be easily visualized, treated and interpreted by means of standard multivariate statistics tools.

1.1.1 Networks

Historically, the study of networks has been mainly the domain of a branch of discrete mathematics known as graph theory (see Bollobás (2012), Bollobás (2013) in mathematics, Lauritzen (1996), Whittaker (2009) in statistics, Jackson (2010) in economics and Diebold and Yilmaz (2015) in finance). In addition to the developments in mathematical graph theory, the research on networks has seen important achievements in some specialized contexts such as, economics, finance and social sciences, just to mention a few. The interconnections among economic agents have different interpretation according to the specific application, which ranges from trade between firms or countries, to personal relations among individuals, as well as to bilateral exposures among financial institutions.

The last decade has witnessed the birth of a new movement of interest and research in the study of complex networks (i.e. networks whose structure is irregular, complex and dynamically evolving in time), with the main focus moving from the analysis of small graphs to that of systems with thousands of nodes, and with a renewed attention on the properties of dynamic networks. The seminal papers of Watts and Strogatz (1998) and Barabási and Albert

(1999) have triggered this flurry of activity within the physics' community at the beginning of the 2000s, however it is right after the outbreak of the 2007 financial crisis that network analysis has attracted significant interest from several other areas of research, such as economics, econometrics, statistics and finance. In fact, the global financial crisis has shown that liquidity and valuation shocks may quickly propagate across the economic system through the linkages among the financial institutions operating in different markets, thus causing widespread losses with sizeable cascade effects. Consequently, understanding the dynamics of the linkages between institutions has become of paramount importance especially for policy-makers, both aiming at preventing an increase of the systemic risk and at improving the effectivness of forecasting tools about contagion and spillover effects (see Forbes and Rigobon (2001) and Forbes and Rigobon (2002) for a definition of financial contagion and brief review).

Many authors have found the complex interconnections between financial institutions or economic sectors to be the vulnerabilities responsible for the amplification of shocks, for instance, see Acemoglu et al. (2012), Gabaix (2011), Gai et al. (2011), Billio et al. (2012), Acemoglu et al. (2016), Billio et al. (2015b)). All these approaches exploit the notion of graphical structure, or network[1], to represent the interdependencies between financial or economic institutions. The need for understanding and mastering this class of objects is becoming crucial in economics and finance, as well as in many other fields, such as image and signal processing, biology and sociology.

The main motivation for the use of networks is that many real-world systems are too complex and intricate for humans to learn from, thus a compact yet flexible tool is needed for the identification of the main characteristics of such systems. The principal scope of networks is to provide a suitable framework for representing the relationships between a set of variables (or agents) in a complex system. Thanks to suitable visualization tools, graphics are able to provide an intuitive interpretation of these interactions by shading light on the direct and indirect connections among the agents. The knowledge of linkages is fundamental since they represent the channel through which phenomena hitting a single node subsequently propagate to the others, over space and time. Moreover, the knowledge of the network topology permits to precisely identify the role of the key linkages (and nodes) in this transmission process, as opposed to standard multivariate econometric tools. Considered together, these features represent the main drivers of the success of networks and graphical models especially in analysing contagion and systemic risk, and, more generally, in the description of complex dependence structures between economic variables.

Recently, network analysis has been further supported by a series of papers that have shown its in- and out-of-sample superior performance over traditional, correlation-based approaches (see Forbes and Rigobon (2001), Forbes and Rigobon (2002), Billio et al. (2012) and Diebold and Yilmaz (2014) among others). The analysis and modelling of economic and financial networks is currently a challenging area of research (e.g., see Schweitzer et al. (2009) and Diebold and Yilmaz (2015)). In economics, the seminal works by Acemoglu et al. (2012) and Gabaix (2011) have shed light on the role of sectoral interconnections in spreading the idiosyncratic shocks and, consequently, in generating significant fluctuations at the macro level. Network analysis has also proven to be a powerful tool for understanding the systemic vulnerabilities of the economic and financial systems. Acemoglu et al. (2015), Acemoglu et al. (2016) and Battiston et al. (2012) studied the role of the network topology in negative shock transmission (or contagion) and its contribution to the determination of systemic risk. Other authors have used graphical models for highlighting the structure of interconnections among financial markets and the fundamental macroeconomic sectors (e.g., see Kali and Reyes (2010)) and for explaining the relations among financial asset returns and volatilities (Diebold and Yilmaz (2014)). Several other networks have been analysed in

[1] The terms "graph" and "network" will be used interchangeably in the rest of the thesis, despite they pertain to different fields of study. Mathematics and probability use the word graph to refer to an abstract entity, whereas physics, applied statistics and social sciences adopt the term network and refer to a precise observed object, either directly observed or inferred from available data.

recent years: Battiston et al. (2007) studied the structure of the international financial network given by direct investments; Vitali et al. (2011) performed a micro-level analysis by exploiting data on corporate controls; networks with bipartite structure of the edges have been studied by Tumminello et al. (2011), whereas Bonanno et al. (2004) discussed the characteristics of a financial network inferred from equity data, rather than returns.

As many economic and financial networks are not observed, often the latent structure must be inferred from data. Early studies in econometrics thus focused on the definition of algorithms for the estimation of the graphical structure from financial from time series, by means of Granger non-causality or other suitable conditional independence testing procedures (e.g., see Billio et al. (2012), Barigozzi and Brownlees (2016)). Gaussian graphical models have been widely used in statistics as an instrument for both estimating and modelling a latent network. Thanks to the properties of the normal distribution, all the dependence between the random variables is encoded in the covariance matrix, thus reducing the process of network extraction to the estimation of the covariance (or the precision) matrix of the joint distribution. Several approaches have been proposed in the literature to solve this task, both in the frequentist and in the Bayesian framework (e.g., see Yuan and Lin (2007), Rajaratnam et al. (2008), Wang and West (2009), Carvalho and West (2007), Giudici and Spelta (2016), Jones and West (2005), Scott and Carvalho (2008), Cerchiello and Giudici (2016), Wang et al. (2011), Yoshida and West (2010)). With the aim of scaling with high-dimensional datasets, Friedman et al. (2008) and Meinshausen and Bühlmann (2006) proposed a graphical lasso estimator for inferring large but sparse covariance matrices (see Wang (2012) for a Bayesian approach), whereas Brownlees et al. (2017) developed a similar regularized estimation procedure for the realized precision matrix estimated from high-frequency data. On the other side, Carvalho et al. (2007) and Dellaportas et al. (2003) proposed two Bayesian estimation procedures applicable, respectively, under the assumption that the underlying graph is decomposable or not. The introduction of observed or inferred graphical structures has greatly improved the performance of financial econometric models both in terms of fitting and forecasting and has favoured thorough studies on financial contagion and systemic risk (e.g., see Carvalho and West (2007), Ahelegbey et al. (2016a), Ahelegbey et al. (2016b), Corsi et al. (2015), Hautsch et al. (2014), Billio et al. (2015b), Caporin et al. (2017)). Under similar assumptions structural instabilities in graphical models have been considered in Bianchi et al. (2018).

The range of empirical findings from this stream of literature is huge and varies according to the specific application under study. Nonetheless, several authors (e.g., see Fagiolo et al. (2010), Billio et al. (2012), Chinazzi et al. (2013), Billio et al. (2015a)) have found a characteristic that is common to many economic and financial networks: the temporal change of their topological structure. The implications of this stylized fact are remarkable. As the graph represents the set of dependence relations between the variables of interest, its change signals the variation of the interconnections and, consequently, of the financial and economic implications that it carries on. For instance, as shown in Acemoglu et al. (2015), the systemic risk varies according to the topology of the underlying network, thus a structural change of system of interconnections represented by the graph may imply a significant change of the systemic risk.

These findings have stressed the need for a set of suitable statistical frameworks able to describe its most relevant characteristics and their economic and financial externalities. There exist several approaches for network modelling in social sciences, starting from the Bernoulli random graph of Erdös and Rényi (1959) and including exponential random graph models, or ERGMs (Frank and Strauss (1986), Holland and Leinhardt (1981), Robins et al. (2007), Caimo and Friel (2011), Thiemichen et al. (2016)), latent space models (Handcock et al. (2007), Hoff et al. (2002), Rastelli et al. (2016), Friel et al. (2016)), stochastic block models (Nowicki and Snijders (2001), Wang and Wong (1987)) and their extension to mixed membership models (Airoldi et al. (2008)). The class exponential random graph models is concerned with the specification of a generative model for random graphs which is able to replicate a set of network characteristics as reported by user-defined statistics. Instead, stochastic block and mixed membership models focus on the identification of clusters of nodes, that is blocks of nodes which are characterized by high degree of intra-component connectivity and low inter-component connections. Finally, latent space approaches aim to embed network information into some (usually low dimensional) latent space, which is sometimes interpreted as the space of individual node's unobserved features. The low dimensionality of this space is exploited to perform the statistical network analysis before projecting the results back into the original space. A compelling review of statistical models for social networks can be found in Kolaczyk (2009), Goldenberg et al. (2010) and Jackson (2010), whereas de Paula (2017) presents a summary of the major models of network formation used in economics.

In a fundamental recent work, Caron and Fox (2017) defined a new model for sparse undirected random graphs, building on the notions of random measures used in Bayesian nonparametrics and on edge exchangeability (see also Veitch and Roy (2015), Borgs et al. (2016) and Cai et al. (2016) for other works founded on edge exchangeability). This remarkable contribution bridges in fact the gap between existing random graph models, which were suited for the generation and representation of dense graphs, and many real world networks which have been found to be sparse (see Caron and Fox (2017) for some examples). Subsequent extensions of this baseline model were aimed at introducing further structure in the generating mechanism, thus allowing the resulting graph to reproduce the main features of real world networks. Williamson (2016) extended the original model of Caron and Fox (2017) for allowing clustering of the edges, with the aim of improving forecasting performance, then she applied the method on the Enron e-mail dataset. The contemporaneous work by Todeschini and Caron (2016), instead, proposes a framework where nodes (i.e. airports) group together forming communities[2] (i.e. hubs). Finally, Palla et al. (2016) introduced temporal dependence in discrete time between two realizations of the random graph process, then applied the method for the study of three high-dimensional single-layer temporal networks.

The statistical approaches mentioned above provide effective methods for studying some specific network characteristics, but all of them share a common feature: they are designed for the study of static networks. They are useful in the analysis of a single, or an independent sequence of graphs, and do not give any tool for capturing the dynamic features of the underlying structure.

1.1.2 Temporal and multi-layer networks

A step forward in this direction has been done by a second generation of statistical network models, explicitly designed to account for the time varying topology (e.g., see Wehmuth et al. (2015) and Holme and Saramäki (2013) for a compelling review and Fig. 1.1 for a visual example). This stream of literature has originated in physics and focuses on the definition of network statistics able to describe the features of the graph which stem from its dynamical nature. The main interest here is the identification of the most important features of a temporally evolving graph, with the aim of proposing a generative random graph model able to reproduce synthetic characteristics of observed networks. Having this goal, many works such as Holme (2005) and Holme and Saramäki (2012) do not properly investigate the dynamics of the network, but consider the temporal dimension as an additional source of information to be exploited for defining more accurate network statistics.

In parallel, the economic community has started to study dynamic graphs from a different perspective, which is closer to the point of view of this thesis. Recent studies such as Zhou et al. (2010) and König et al. (2017) proposed a framework for inferring a time varying network structure from the data. This field of literature is mainly concerned with the estimation of a temporally evolving adjacency matrix that encodes the network structure (for example, Nakajima and West (2015), Bräuning and Koopman (2016), Giraitis et al. (2016), Kolar et al. (2010)). Building on the work of Frank and Strauss (1986), Robins and

[2] Their identification relies on external additional information.

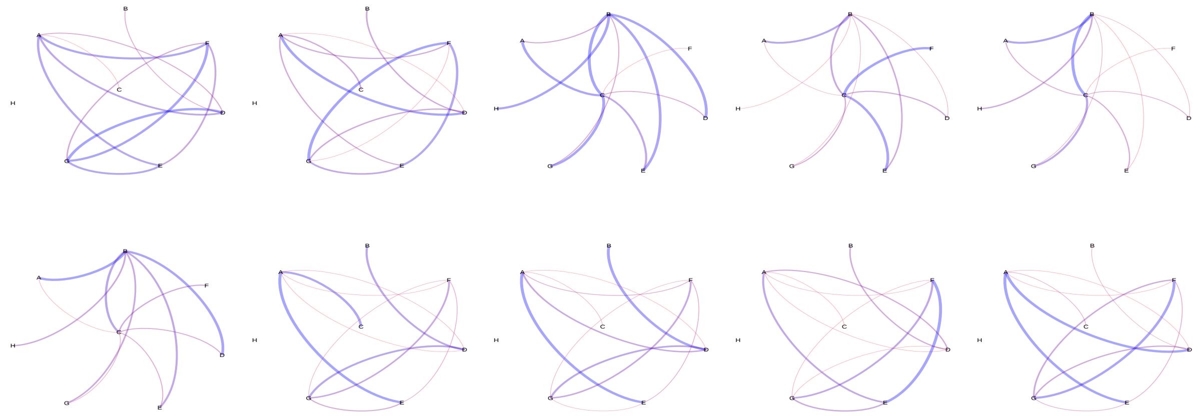


FIGURE 1.1: Temporal snapshots of a time varying single-layer network, starting from $t = 1$ (*top left*) to $t = 10$ (*bottom right*). Directed edges are clockwise oriented and coloured according to the corresponding weight (*red* for positive, blue for *negative*.

Pattison (2001) proposed a version of the p^* model for temporal random graphs, by using the network structure in one period as predictor for the next period's. Some authors have adopted a different perspective and have extended the class of exponential random graph models to encompass also temporal information though the definition of the family of temporal exponential random graph models, or TERGMs (e.g., see Hanneke et al. (2010), Krivitsky and Handcock (2014)). Another path has been followed by Sewell and Chen (2015), who generalized latent space models by allowing the nodes projected into the latent space to follow a temporal trajectory.

Recently, Mazzarisi et al. (2017) proposed a model which combines the intuitions of latent variables and autoregressive dynamics, by assuming a dynamic model of network formation where the edge's probability depends both on its existence during the previous period and on the unobserved nodes' features. Conversely, Betancourt et al. (2017) specified a model for the joint probability of couples of edges of a dynamic binary network in terms of individual edges' propensity to form a link.

The concept of network presented up to this point pertains a structure with a single layer (or stratum). However, in the real world there are situations where the same agents or entities are repeatedly observed according to a different criteria. For example, the friendship relations among the same collection of individuals may be traced on several on-line social networks, each one representing a specific layer. By collecting together all the layers (or strata) a multi-layer (or multiplex) network[3] is obtained (see Boccaletti et al. (2014), Kivelä et al. (2014) and Dickison et al. (2016) for a review). The degree of complexity of these structures can be significantly higher than the single-layer counterpart, due to the possible interconnections that may exist between the nodes or the edges belonging to different layers. For the same reason, however, multi-layer graphs are a very flexible tool apt for modelling complex real world phenomena in several, and apparently unrelated, disciplines.

Multiplex networks have been introduced in applied studies only in the last decade, nonetheless their usefulness as a statistical tool in modelling has fostered their rapid growth in popularity and currently multi-layer graphs are among the main instruments for representing and studying the behaviour of complex systems. One of the most important domain of application pertains the assessment of systemic risk (e.g., Montagna and Kok (2016), Aldasoro and Alves (2016) and Poledna et al. (2015)) and the identification of contagion channels (e.g., Mistrulli (2011)) in the interbank network, where the different layers represent several types of bilateral exposures. A different research question that has been addressed pertains the identification of hidden communities (e.g., Barigozzi et al. (2011), Gurtner et al.

[3] In the particular case where the subjects are the same over all layers, the multi-layer network is said to be node-aligned.

(2014), Bazzi et al. (2016)). Instead, Bargigli et al. (2015) and Cozzo et al. (2016) focused on the definition of suitable metrics and network statistics able to provide a synthetic description of the main features of multi-layer graphs and found that, in general, even the standard measures used in single-layer network analysis may yield significantly different results when applied in this context.

The introduction of several layers allows to answer many interesting research questions in other, related fields. A line of research which joins applied probability, finance and epidemiology is concerned with the study of diffusion (or spreading) processes along the paths identified by the links of a network. From this perspective, Kivelä et al. (2012) and De Domenico et al. (2016) studied the characteristics of a spreading process which percolates across the layers of a multiplex network and revealed some channels of contagion which single-layer graphs failed to detect.

Other scientific domains where the introduction of multiplex networks has represented a turning point of the research include neuroimaging and medicine. Here, multiplex networks provided a better description of human brain (e.g., see Battiston et al. (2017), Beckmann and Smith (2005), Estienne et al. (2001), Miwakeichi et al. (2004), Davidson et al. (2013), Damoiseaux et al. (2006), Rubinov and Sporns (2010)), transportation (Gallotti and Barthelemy (2015)) and social network analysis (e.g., Acar et al. (2006), Acar et al. (2005), Kolda et al. (2005), Murase et al. (2014)).

As for economic and financial networks, also their temporal and multi-layer counterparts are generally not observed, thus calling for statistical procedures for extracting networks from data. However, the additional features embedded by these two classes, namely time variations and multiple layers, call for different estimation techniques. In this context few results are available in the literature: Hanneke et al. (2010) and Pensky (2016) designed a procedure valid only for temporal exponential random graphs and graphons, respectively, whereas Oselio et al. (2014) and Stanley et al. (2016) considered multi-layer and multiplex stochastic block models, respectively. The two approaches which can have wider range of applicability have been suggested in the recent papers of Nakajima and West (2015) and Bianchi et al. (2018).

1.1.3 Multi-dimensional datasets

Most statistical methods are used to analyse the scores of objects (for example subjects, groups, countries) on a number of features, or variables, and the resulting data can be arranged in a 2-order array, or matrix. However, nowadays data are often far more complex. For example, the data may have been collected under a number of conditions and at several time stamps. When another profile or dimension is considered in addition to the previous two (i.e. subject and features), then the data become 3-order arrays, or more generally, multi-way arrays[4] (see Fig. 1.2). In such a case, there is a matrix for each condition, thus the data can be naturally arranged by stacking one matrix behind the other in to form an hyperrectangle, or 3-order tensor (see Appendix A.1 for a formal definition). As more dimensions are added to the data, it is possible to iterate the same procedure and obtain higher-order tensors or multi-way arrays (see Fig. 1.3 for an example). The collection of mathematical and statistical techniques designed to deal with and analyze multi-way data are referred to as tensor calculus (see Hackbusch (2012)) and multi-way methods (see Smilde et al. (2005)), respectively. We refer to Appendix A.1 for further details.

As opposed to tensor analysis (see Appendix A.1 for an introduction on tensor operators and decompositions), which deals directly with array-valued data without need of transforming them, both multivariate or matrix models require to firstly vectorize or matricize the array of data. In both cases the idea is to reshape the data into a long vector or a matrix, respectively, thus eliminating all the remaining dimensions. This has two main implications: first, it precludes the exploitation of any information naturally embedded in the structure of

[4] In this thesis we use the terms "tensor" and "array" interchangeably. The origins of the first term are related to mathematics and physics, while the latter is more commonly used in computer science.

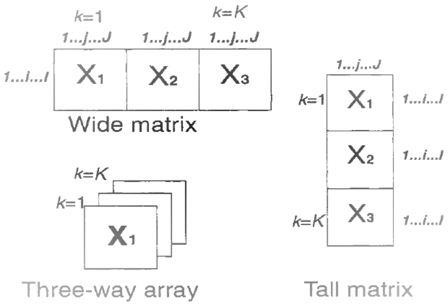


FIGURE 1.2: Example of 2-order (i.e. matrix or two-way) and 3-order (i.e. tensor or three-way) arrays.

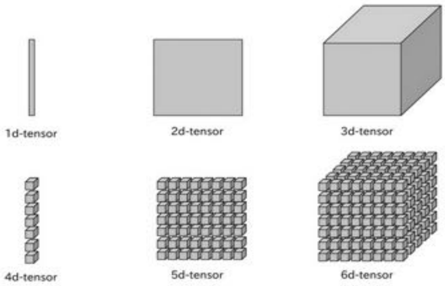


FIGURE 1.3: Example of tensor, from 1-order (*top left*) to 6-order (*bottom-right*).

the raw data; second, the transformed variables may still be too big for standard methods to apply.

Multi-country panels. Many economic data have been collected for years in the shape of long vectors or matrices. One of the most well-known example of matrix-variate dataset regards country-level sectoral input-output tables, which report the bilateral inflow and outflow of goods and services between any sector of a national economy. Recently, the higher computational capacity, the bigger storage capacities and the creation of the European Economic Union[5] have greatly fostered the development of this kind of data. The harmonization of the standards has permitted to merge together the information from several countries and to build *world* or *multi-country input-output* tables (see Timmer et al. (2015) for an introduction), which consist of a time series of sectoral input-output tables, for different countries. The nature of these datasets is clearly high-dimensional: the cross-sectional country level is divided according to the sectors of the economy and observed over time. Therefore, it seems evident that for effectively exploit the entire potential of these data it is first of all necessary to represent, store and manage these complex and high-dimensional objects in a proper way. Early studies using multi-country input-output tables (e.g., see Dietzenbacher et al. (2013), Lenzen et al. (2010), Lenzen et al. (2004), Wixted et al. (2006), Sanz Díaz et al. (2015)) have tried to exploit matrix tools for performing statistical analysis, but novel and more accurate and results are expected from the introduction and exploitation of appropriate structures for the variables.

The last decade has also been characterized by the increasingly availability of data on international trade (for example, from UN COMTRADE[6]), which collect bilateral import/export relations divided by commodity and over time, and international capital flows (for example, form the Bank of International Settlements[7]), that report bilateral inflows/outflows of financial capital divided by type and over time. These data have the intrinsic structure of a

[5] Nonetheless, analogous datasets are currently available for many countries worldwide (e.g. see http://www.oecd.org/sti/ind/inter-country-input-output-tables.htm).

[6] https://comtrade.un.org

[7] https://www.bis.org/index.htm

4-dimensional array (for example, a typical entry of the international trade dataset is represented by the tuple country, country, commodity, time). This is the main reason why models based on vectors and matrices are unable to provide an accurate representation, thus requiring the introduction of novel constructions for performing meaningful statistical analyses. During the last decade, these datasets have been subject to a thorough study aimed at uncovering its topological characteristics and their temporal change, at identifying hidden communities of countries and at defining new significative measure of synthesis of the network structure (for instance, see the papers by Barigozzi et al. (2011), Zhu et al. (2014), Fagiolo (2010), Schiavo et al. (2010), Fagiolo et al. (2008), Fagiolo et al. (2009), Hidalgo and Hausmann (2009) and Squartini et al. (2011)). The COMTRADE dataset has also been utilized for studying the resilience of the international trade network to external shocks and the temporal change of its topology (e.g., see Chinazzi et al. (2013), Fagiolo et al. (2010), Kharrazi et al. (2017) and Meyfroidt et al. (2010)).

Another relevant field of research in economics hinges on multi-country datasets for undertaking macroeconomic analysis and supporting policy-makers in forecasting. The class of multi-country panel vector autoregressive (VAR) models, introduced by Canova and Ciccarelli (2004), have become increasingly popular during the last decade and are currently one of the most relevant tools for macro-econometric studies (e.g., see Love and Zicchino (2006), Canova and Ciccarelli (2009), Canova et al. (2007), Canova et al. (2012), Canova and Ciccarelli (2013), Grossmann et al. (2014), Koop and Korobilis (2016), Lof and Malinen (2014), Billio et al. (2016)). Other researchers used multi-country panel data mainly for the sake of forecasting and time series analysis (e.g., see Korobilis (2016), Chudik et al. (2016), Sarantis and Stewart (2001), Weber and Matthews (2007)). All the mentioned models are able to exploit only a fraction of the available information, since they are essentially multivariate dynamic models, and are unable to cope with the high-dimensionality of the dataset. Furthermore, the process of vectorization of array-valued data implies that all infomation naturally embedded in its structure is lost, thus empowering the efficiency of the statistical analysis. The state-of-the-art on macroeconometric models consists in VAR models with deterministic (i.e. panel VAR) or stochastic (i.e. compressed VAR, see Koop et al. (2018), graphical VAR, see Ahelegbey et al. (2016a), stochastic search for VAR restrictions, see George et al. (2008)) restrictions aimed at reducing the dimension of the parameter space. Neither approach is fully satisfactory, since all are currently unable to fully exploit the original structure of the data.

Volatility surface. In the field of financial econometrics, the study of the implied volatility surface has attracted particular attention in the recent years (see Gatheral (2011) for a review). Implied volatility differs from historical (or realized) volatility. The latter is a direct measure of the movement of the corresponding underling's price over recent history; by contrast, implied volatility is determined by the market price of a derivative contract and not the underlying (thus there exists several implied volatilities for the same underlying). The implied volatility surface is the implied volatility of European options on a particular asset viewed as a function of strike price and time to maturity (Gatheral (2011)). Several parametric and semi-parametric methods have been proposed for estimating this bivariate function (see Homescu (2011) and Fengler (2012) for a review), alternatively nonparametric estimation via spline interpolation proved to be a valid alternative (Bliss and Panigirtzoglou (2002), Casarin et al. (2015)). In the following example, without loss of generality consider the latter approach (a similar reasoning holds also in the other cases). In all cases, the raw datasets consists of a discrete grid where each point is a tuple of implied volatility corresponding to a strike price at a given time to maturity. From this it is evident that the data is a three-dimensional object with a precise structure encoded by the definition of implied volatility surface.

In addition, the level of implied volatilities has been found to change over time (e.g., see Skiadopoulos et al. (2000)), thus continuously deforming the volatility surface. Nonetheless, recent studies (e.g., Cont and Da Fonseca (2002), Daglish et al. (2007), Fengler et al.

(2007), Vergote and Gutiérrez (2012), Casarin et al. (2015)) have ignored the natural tuple-structure of the volatility surface focusing only on the modelling in one or two dimensions the volatility.

Social networks. Complex data structures also arise in extracting relationships in social networks (see Scott (2017), Wasserman and Faust (1994), Jackson (2010)). The main purpose of this analysis is to discover and characterize the hidden structures in social networks (Borgatti et al. (2009)): for instance, extracting communication patterns among people or within organizations. Despite the different purpose and the fact that, in general, social networks are directly observed (or at least data on them can be collected more accurately then in finance), they share some similarities with the economic and financial counterparts previously presented. Most important of all, the fact that the raw data are collected in the same shape, that is of matrices or higher-order arrays. For example, Acar et al. (2005) and Acar et al. (2006) studied a chat room communication dataset, where each datum corresponds to the tuple users, keywords and time samples, thus representing a 3-order tensor as a whole. Similarly, Kolda et al. (2005) in studying web links has dealt with a three-dimensional array, with dimensions corresponding to web pages, web pages and anchor text, respectively.

The study of social networks has remarkable implications in several related fields of study, such as economics (e.g., Calvo-Armengol and Jackson (2004), Jackson (2010)) and medicine (e.g., Christakis and Fowler (2008)). Several studies have been proposed to uncover the mechanisms of link formation and the dynamics of the network in general (e.g., see Jackson and Watts (2002), Jackson and Wolinsky (1996)), mostly by modelling the behaviour of the single agents (i.e. the nodes of the network) and their relationships with each other (i.e. the edges). It is worthy to stress that social network data is generally richer than its financial counterpart, partially thanks to the observability of social relationships and individual features. This fact puts more in evidence the structured nature of social network data and, consequently, the increasing need for suitable objects able to couple with it.

Neuroimaging. Complex and structured data are often encountered also in neuroimaging and, more generally, in image processing. It has been shown in numerous studies in neuroscience that information contained in the data may not be accurately captured or uniquely identified by classical matrix analysis methods (e.g., see Estienne et al. (2001), Beckmann and Smith (2005), Damoiseaux et al. (2006), Davidson et al. (2013), Miwakeichi et al. (2004)). In particular, datasets composed by electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI) can consist of a sequence of 2-dimensional (2D) images of slices of the brain, or in the whole 3-dimensional (3D) brain volume. The temporal resolution of the data depends on the time between acquisitions of each individual volume: typically, brain volumes of dimensions $64 \times 64 \times 30$ are collected at several hundreds of time stamps. Moreover, the experiment is often repeated for many subjects. It is clear that fMRI data analysis is a time series analysis problem of massive proportions.

Each data point is characterized by a physical location in a 3-dimensional space and by the value of the response to the stimulus, thus yielding a tuple of four elements where each entry has a specific meaning which is not interchangeable. This is a key information that should always be taken into account. By contrast, the standard approach towards the statistical analysis of 3D images consists in limiting the study to a sub-sample consisting of a sequence of 2D slices (Lindquist (2008)). Therefore, it appears that this methodology is inefficient, due to the inability to effectively exploit all the available information. Recent developments (e.g., Beckmann and Smith (2005), Damoiseaux et al. (2006), Davidson et al. (2013), Miwakeichi et al. (2004)) in this field have proposed to tackle this issue by using high-dimensional tensors, which represent the natural shape of the raw data, for exploiting the available information, obtaining significantly better results.

A bridge between neuroimaging and economics has been provided by the recently born field of neuroeconomics (e.g., see Camerer et al. (2005) and Bossaerts and Murawski (2015)).

The key goals of the discipline are the determination of how choices were implemented biologically and the identification of the involved neural circuitry. Given that the reaction of human brain to stimuli is the core of the analysis, the main data sources used in this field consist of EEG and fMRI, the previous observations about the structure of the data still hold.

Finally, another stream of research where high-dimensional structured data are becoming increasingly popular is marketing. For example, Naik et al. (2008) analyzed an array made of 4-dimensional data arrays including variables, alternatives, subjects and time. The pressing need for models dealing with at least 3-dimensional data arrays has been pointed out by managers[8] as the very next future of marketing for the definition of adequate customer profiles. Similarly, Dessart et al. (2016) used a high-dimensional dataset for studying consumer engagement, whereas Balabanis and Diamantopoulos (2004) exploited a 4-dimensional array of consumer-level data for examining the preference patterns of U.K. consumers for domestic products and those originating from specific foreign countries for eight product categories. Overall, Erevelles et al. (2016) explicitly considers the adoption of novel variable formats able to accurately contain the data as one of the challenges that the era of "*Big Data*" is posing to marketing.

1.2 Motivation

This thesis is centred on the development of novel statistical and econometric frameworks for the analysis of time-varying networks, with a focus on medium-high dimensional cases[9]. This Section provides a brief overview of the main research questions which motivate the work of this thesis.

Overall, the state-of-the-art on statistical network modelling is at stack. Several empirical studies have proved that the topology of economic and financial networks is subject to non-negligible changes over time. Nonetheless, few attempts have been made for providing economically interpretable and computationally tractable statistical frameworks for explicitly modelling the time-varying nature of network processes.

Despite the main interest of this thesis resides in the study of the dynamics of networks, the techniques and models proposed have direct applicability also in other domains. The pouring of large and structured datasets in many fields (e.g., economics, biology and sociology) is calling for novel suitable statistical tools able to fully exploit the information embedded in the data.

1.2.1 Dynamic networks

During the last decade, the scientific community in econometrics, economics and finance has uncovered the role of networks in accounting for shock transmissions and in encoding the dependence structure of relevant variables. Nonetheless, this field of research is still at its infancy. Current widespread modelling efforts focus on the estimation of the latent topological structure of the network from time series data or on the inclusion of a graph in an econometric model for improving its performance. However, in most of these cases the network structure is inherently static, despite several studies have revealed the change of economic and financial networks over time (e.g., see Fagiolo et al. (2010), Billio et al. (2012), Chinazzi et al. (2013), Billio et al. (2015a)).

The bulk of current studies on dynamic graphs stems from physics and is devoted to the definition of novel measures and synthetic indicators which are able to describe the time varying features of the network (e.g., see Casteigts et al. (2012)). Another part of the literature concerns the development of random graph models able to replicate these temporal

[8] https://marketing.cioreview.com/cxoinsight/the-future-of-marketing-creating-3dimensional-customer-profiles-in-an-iot-frenzied-world-nid-24374-cid-51.html

[9] Citing Korobilis (2016): "*if a VAR for a single country has $G = 10$ variables, then this would be of medium size. Once we consider only, say, $N = 5$ such countries then the PVAR has 50 variables in total and can be considered large dimensional.*"

regularities (e.g., see Zhang et al. (2017)). In all these cases, the effects of the temporal evolution of the network are analysed by means of sufficient statistics and indicators, but the intrinsic dynamical process driving the change of the graphical structure is not studied.

Only few attempts have been made in very recent years to explicitly and directly model the process of temporal change of a network. There is an increasing interest in pushing forward the frontier of the research by providing suitable statistical models able to provide a meaningful description of the dynamics of the network structure. These models might have many interesting applications as complex networks are fundamental in several disciplines also outside economics, ranging from medicine to signal processing, from sociology to statistics. Nonetheless, both the statistical and econometric literature on time series analysis are still lacking of suitable tools for performing this kind of analysis. The current state of the research appears to be stuck: the scientific community is well aware of the dynamical properties of real world networks, but it is still looking for effective approaches to model these phenomena.

1.2.2 High-dimensionality

The last decade has been characterized by the exponential growth in the quantity and size of data, such that it is possible to call the current state of the world the era of "*Big Data*". This era has brought plenty of new challenges to the scientific community, ranging from statistics to computer science. High-dimensionality of data is one of the most important issues to be addressed, due to its direct implications in theoretical statistics, from the theoretical perspective, as well as in almost all the scientific domains, from an applied point of view. In this thesis the term *high-dimensionality* is used to refer to two different cases: large datasets, consisting of hundreds of variables observed in big sample sizes, and big data structures, to be intended as the case in which the variables of interest have an intrinsic complex structure (for example, they are naturally collected in the form of arrays).

The main challenge brought by high-dimensionality hinges on three strictly related issues. First, it hampers data analysis since standard statistical models are most often unable to deal with high-dimensional variables. Multivariate models, which represent the backbone of theoretical and applied statistics, as well as econometrics and machine learning, are not sufficiently flexible to cope with array-valued data or even unable to deal with them. This has fostered the use of dimensionality reduction techniques (see Fodor (2002), Camastra (2003) and Sorzano et al. (2014) for a review) aimed either at selecting the subset of most significant variables or at reshaping them into lower-dimensional objects. However, techniques such as principal component analysis (PCA), factor models and matrix decompositions, which are the building blocks of this stream of literature, may be inapt to deal with structured data. This has motivated a generalization of this methods to high-dimensional variables. Lastly, both large datasets and complex structured variables require significant computational effort even for performing simple analyses. Theoretical models must be coupled with efficient computational algorithms in order to provide an effective answer to the research question at hand. In many cases this may imply the infeasibility of current widespread practices, such as standard multivariate models, thus requiring the development of novel approaches.

Some remarks are in order. The first and most widespread attempts to model big data in a multivariate framework has relied upon penalized estimation encouraging sparsity. Though effective in cases where the number of variables to be accounted for is high as compared to the sample size, they fail to model even structured data. Conversely, models of data compression can achieve significant dimension reduction, but in general at the price of lack of interpretability of the output.

This provides the motivation for the adoption of a different approach, through the use of high-dimensional arrays, also called tensors. They generalize matrices to multi-dimensional cases, thus providing a suitable way to store and represent several types of structured data, such as multi-layer networks, multi-country panels, input-output and trade tables, but also functional MRI and EEG data. More than that, tensors have a significantly richer structure than matrices (which are 2-dimensional tensors) and plenty of operations and decompositions have been defined on them, extending and going beyond well-known matrix algebra tools. By means of these operators it is possible to define flexible tensor-valued processes that extend and generalize multivariate econometric models, whereas tensor decompositions allow for significant dimensionality reduction without big losses of information.

Building on these instruments recently introduced in the statistical literature, this thesis aims also at providing econometric models for high-dimensional, structured time series, represented by networks.

1.3 Contribution

The contributions of the thesis can be summarized as follows:

- definition of econometric models for real-valued tensor-variate time series processes. A new framework is presented for describing the autoregressive features of high-dimensional arrays. This encompasses real-valued networks as a special case of particular interest, but can be applied also to other frameworks where the object of interest is bi- or higher-dimensional.

- definition of econometric models for dynamic binary arrays. The probability of each entry is allowed to depend on a specific set of covariates, moreover both sparsity and temporal clustering of the entries are explicitly modelled. This framework permits to study the dynamics of edge formation and disruption in multi-layer binary networks with a big number of nodes. Moreover, the model is extended to jointly model a multivariate series of relevant variables.

- proposal of computationally efficient algorithms for Bayesian estimation of the proposed models, which scale well up to medium-high sizes. Even though MCMC algorithms are not parallelizable, it is possible to exploit tensor decompositions in order to parallelize computations inside each iteration, thus providing a remarkable speed-up. In addition, several simulation studies are performed to demonstrate the scalability and accuracy of the proposed methods.

- definition of a statistical framework for time-varying probability density functions. By applying the methodology to copulas, it is possible to flexibly describe the dynamics of the dependence structure between variables. Moreover, since density functions are a particular class of constrained functions, the method can be extended also to other cases.

- applications of the proposed models to time-varying economic and financial networks. Datasets on networks of different size and nature are analysed from a temporal point of view, with the aim of uncovering the determinants of their evolution as well as for forecasting and impulse-response analysis.

1.4 Outline of the thesis

The structure of the thesis is as follows. In Section 1.1 we present the state of the art of the literature on temporal network, with a focus on statistical and econometric modelling.

Chapter 2 presents a novel Bayesian econometric model that extends current widespread multivariate models used in time series analysis. It is introduced the use of tensors in an econometric model as a way for coupling with both the structure and the high-dimensionality of the data. The general framework is first presented, then applied for studying the autoregressive nature of the topological structure of a real-valued network. Extensive simulations are carried to prove the accuracy and computation efficiency of the sampler in high-dimensional settings. The main contribution is the definition of a model able to capture the dynamics of matrix-valued time series processes, of which real networks are a particular case, while retaining computational efficiency and economic interpretability.

Chapter 3 describes a non-linear dynamical model for time-varying binary, directed networks. This framework allows for temporal clustering of the network structure, according to a hidden Markov chain process, which drives both the observed sparsity patterns and the individual edges' probabilities. The main contribution is the definition of a model capable of describing the temporal evolution of the structure of a network both globally, in terms of its sparsity, and individually, explicitly modelling the dynamics of each edge's probability. Moreover, the use of tensors together with a Bayesian approach for the inference allow to couple with high-dimensional networks with more than one hundred of nodes.

Chapter 4 presents a statistical framework for the nonparametric prediction of bivariate probability density functions, focusing in particular on bivariate copulas. We follow the frequentist approach for inference. The main contribution is the definition of a process able to flexibly describe the temporal evolution of the dependence structure, thus of the link in an undirected graph. Following the literature on vine copulas, it is possible to interpret a set of random variables as nodes of a network, where each edge is described by a specific bivariate copula function. From this perspective, the proposed methodology can be used for modelling the temporal evolution of linkages of an undirected network.

Chapter 2

Bayesian Dynamic Tensor Regression

The art of doing mathematics consists in finding that special case which contains all the germs of generality.

DAVID HILBERT

Everything should be made as simple as possible, but not simpler.

ALBERT EINSTEIN

2.1 Introduction

The increasing availability of large sets of time series data with complex structure, such as EEG (e.g., Li and Zhang (2017)), neuroimaging (e.g., Zhou et al. (2013)), two or multidimensional tables (e.g., ,Balazsi et al. (2015), Carvalho and West (2007)), multilayer networks (e.g., Aldasoro and Alves (2016), Poledna et al. (2015)) has put forward some limitations of the existing multivariate econometric models. In the era of "*Big Data*", mathematical representations of information in terms of vectors and matrices have some non-negligible drawbacks, the most remarkable being the difficulty of accounting for the structure of the data, their nature and the way they are collected (e.g., contiguous pixels in an image, cells of matrix representing a geographical map). As such, if this information is neglected in the modelling the econometric analysis might provide misleading results.

When the data are gathered in the form of matrices (i.e. 2-dimensional arrays), or more generally as tensors, that is multi-dimensional arrays, a statistical modelling approach can rely on vectorizing the object of interest by stacking all its elements in a column vector, then resorting to standard multivariate analysis techniques. The vectorization of an array does not preserve the structural information encrypted in its original format. In other words, the physical characteristics of the data (e.g, the number of dimensions and the length of each of them) matter, since a cell is highly likely to depend on a subset of its contiguous cells. Collapsing the data into a 1-dimensional array does not allow to preserve this kind of information, thus making this statistical approach unsuited for modelling tensors. The development of novel methods capable to deal with 2- or multi-dimensional arrays avoiding their vectorization is still an open challenging question in statistics and econometrics.

Many results for 1-dimensional random variables in the exponential families have been extended to the 2-dimensional case (i.e. matrix-variate, see Gupta and Nagar (1999) for a compelling review). Conversely, tensors have been recently introduced in statistics (see Hackbusch (2012), Kroonenberg (2008), Cichocki et al. (2009)), providing the background for more efficient algorithms in high dimensions especially in handling *Big Data* (e.g. Cichocki (2014)). However, a compelling statistical approach to multi-dimensional random objects is lacking and constitutes a promising field of research.

Recently, the availability of 3-dimensional datasets (e.g., medical data) has fostered the use of tensors in many different fields of theoretical and applied statistics. The main purpose of this article is to contribute to this growing literature by proposing an extension of standard multivariate econometric models to tensor-variate response and covariates.

Matrix models in econometrics have been employed over the past decade, especially in time series analysis where they have been widely used for providing a state space representation (see Harrison and West (1999)). However, only recently the attention of the academic community has moved towards the study of this class of models. Within the time series analysis literature, matrix-variate models have been used for defining dynamic linear models (e.g., Carvalho and West (2007) and Wang and West (2009)), whereas Carvalho et al. (2007) exploited Gaussian graphical models for studying matrix-variate time series. In a different context, matrix models have also been used for classification of longitudinal datasets in Viroli (2011) and Viroli and Anderlucci (2013).

Viroli (2012) the author presented a first generalization of the multivariate regression by introducing a matrix-variate regression where both response and covariate are matrices. Ding and Cook (2016) propose a bilinear multiplicative matrix regression model whose vectorised form is a VAR(1) with restrictions on the covariance matrix. The main shortcoming in using bilinear models (either in the additive or multiplicative form) is the difficulty in introducing sparsity constrains. Imposing a zero restriction on a subset of the reduced form coefficients implies a zero restriction on the structural coefficients[1]. Ding and Cook (2016) proposed a generalization of the envelope method of Cook et al. (2010) for achieving sparsity and increasing efficiency of the regression. Further studies which have used matrices as either the response or a covariate include Durante and Dunson (2014b), who considered tensors and Bayesian nonparametric frameworks and Hung and Wang (2013), who defined a logistic regression model with a matrix-valued covariate.

Following the model specification strategy available in the existing literature, there are two main research streams. In the first one, Zhou et al. (2013), Zhang et al. (2014) and Xu et al. (2013) propose a linear regression models with a real-valued N-order tensor $\mathcal{X}$ of data to explain a one-dimensional response, by means of the scalar product with a tensor of coefficients $\mathcal{B}$ of the same size. More in detail, Zhang et al. (2014) propose a multivariate model with tensor covariate for longitudinal data analysis; whereas Zhou et al. (2013) uses a generalized linear model with exponential link and tensor covariate for analysing image data. Finally, the approach of Xu et al. (2013) exploits a logistic link function with a tensor covariate to predict a binary scalar response.

In the second and more general stream of the literature (e.g., Hoff (2015) and Li and Zhang (2017)) both response and covariate of a regression model are tensor-valued. From a modelling point of view, there are different strategies. Hoff (2015) regresses a N-order array on an array of the same order but with smaller dimensions by exploiting the Tucker product, and follows the Bayesian approach for the estimation. Furthermore, Bayesian nonparametric approaches for models with a tensor covariate have been formulated by Zhao et al. (2013), Zhao et al. (2014) and Imaizumi and Hayashi (2016). They exploited Gaussian processes with a suitable covariance kernel for regressing a scalar on a multidimensional data array. Conversely, Li and Zhang (2017) defines a model where response and covariates are multidimensional arrays of possibly different order, and subsequently uses the envelope method coupled with an iterative maximum likelihood method for inference.

We propose a new dynamic linear regression modelling framework for tensor-valued response and covariates. We show that our framework admits as special cases Bayesian VAR models (Sims and Zha (1998)), Bayesian panel VAR models (proposed by Canova and Ciccarelli (2004), see Canova and Ciccarelli (2013) for a review) and Multivariate Autoregressive models (i.e. MAR, see Carriero et al. (2016)), as well as univariate and matrix regression models. Furthermore, we exploit the PARAFAC decomposition for reducing the number of parameters to estimate, thus making inference on network models feasible.

[1] The phenomenon is worse in the bilinear multiplicative model, given that each reduced form coefficient is given by the product of those in the structural equation.

We also contribute to the empirical analysis of tensor data in two ways. First, we provide an original study of time-varying economic and financial networks and show that our model can be successfully used to carry out forecast and impulse response analysis in this high-dimensional setting. Few attempts have been made to model time-evolving networks (for example, Holme and Saramäki (2012), Kostakos (2009), Barrat et al. (2013), Anacleto and Queen (2017) and references in Holme and Saramäki (2013)), and this field of research, which stems from physics, has focused on providing a representation and a description of temporally evolving graphs. Second, we show how tensor regression con be applied to macroeconomic panel data, where standard vectorized models cannot be used.

The structure of the chapter is as follows. Section 2.2 is devoted to a brief introduction to tensor calculus and to the presentation of the new modelling framework. The details of the estimation procedure are given in Section 2.3. In Section 2.4 we test proposed model on simulated datasets and in Section 2.5 we present some empirical applications.

2.2 A Tensor Regression Model

We introduce multi-dimensional arrays (i.e. tensors) and some basic notions of tensor algebra which will be used in this paper. Moreover, we present a general tensor regression model and discuss some special cases.

2.2.1 Tensor Calculus and Decompositions

The use of tensors is well established in physics and mechanics (see Synge and Schild (1969), Adler et al. (1975), Malvern (1986), Lovelock and Rund (1989), Aris (2012) and Abraham et al. (2012)), but very few references can be found in the literature outside these disciplines. For a general introduction to the algebraic properties of tensor spaces we refer to Hackbusch (2012). A noteworthy introduction to tensors and corresponding operations is in Lee and Cichocki (2016), while we make reference to Kolda and Bader (2009) and Cichocki et al. (2009) for a review on tensor decompositions. In the rest of the paper we will use the terms tensor decomposition and tensor representation interchangeably, even though the latter one is more suited to our approach.

A N-order tensor is a N-dimensional array (whose dimensions are also called modes). The number of dimensions is the *order* of the tensor. Vectors and matrices are examples of first- and second-order tensors, respectively, while one may think about a third order tensor as a series of matrices of the same size put one in front of the other one, forming a parallelepiped. In the rest of the paper we will use lower-case letters for scalars, lower-case bold letters for vectors, capital letters for matrices and calligraphic capital letters for tensors. When dealing with matrices, in order to select a column (or row) we adopt the symbol ":". The same convention is used for tensors when considering all elements of a given mode. For example, let $A \in \mathbb{R}^{m \times n}$ be a matrix and $\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ an array of order N, then $A_{i,j}$ and $\mathcal{B}_{i_1,\ldots,i_N}$ indicate the (i,j)-th and $(i_1,\ldots,i_N)$-th element of A and $\mathcal{B}$, respectively, and:

(i) $A_{(i,:)}$ is the i-th row of A, $\forall i \in \{1,\ldots,m\}$;

(ii) $A_{(:,j)}$ is the j-th column of A, $\forall j \in \{1,\ldots,n\}$;

(iii) $\mathcal{B}_{(i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_N)}$ is the mode-k fiber of $\mathcal{B}$, $\forall k \in \{1,\ldots,N\}$

(iv) $\mathcal{B}_{(i_1,\ldots,i_{k-1},:,:,i_{k+2},\ldots,i_N)}$ is the mode-$k,k+1$ slice of $\mathcal{B}$, $\forall k \in \{1,\ldots,N-1\}$

The mode-k fiber is the equivalent of rows and columns in a matrix, more precisely it is the vector obtained by fixing all but the k-th index of the tensor. Instead, slices (i.e. bi-dimensional fibers of matrices) or generalizations of them, by keeping fixed all but two or more dimensions (or modes) of the tensor.

The *mode-n matricization* (or unfolding), denoted by $\mathbf{X}_{(n)}$, is the operation of transforming a N-dimensional array $\mathcal{X}$ into a matrix. It consists in re-arranging the mode-n fibers of the tensor to be the columns of the matrix $\mathbf{X}_{(n)}$, which has size $I_n \times \bar{I}_{(-n)}$ with $\bar{I}_{(-n)} = \prod_{i \neq n} I_i$. The mode-$n$ matricization of $\mathcal{X}$ maps the $(i_1, \ldots, i_N)$ element of $\mathcal{X}$ to the (i_n, j) element of $\mathbf{X}_{(n)}$, where:

$$j = 1 + \sum_{m \neq n} (i_m - 1) \prod_{p \neq n}^{m-1} I_p \tag{2.1}$$

For some numerical examples, see Kolda and Bader (2009) and Appendix A.1. The mode-1 unfolding is of interest for providing a visual representation of a tensor: for example, when $\mathcal{X}$ be a third-order tensor, its mode-1 unfolding $\mathbf{X}_{(1)}$ is a matrix of size $I_1 \times I_2 I_3$ obtained by horizontally stacking the frontal slices of the tensor. The *vectorization* operator stacks all the elements in direct lexicographic order, forming a vector of length $\bar{I} = \prod_i I_i$. However, notice that other orderings are possible (as for the vectorisation of matrices), since the ordering of the elements is not important as long as it is consistent across the calculations. The mode-n matricization can also be used to vectorize a tensor $\mathcal{X}$, by exploiting this relationship:

$$\mathrm{vec}\,(\mathcal{X}) = \mathrm{vec}\left(\mathbf{X}_{(1)}\right), \tag{2.2}$$

where $\mathrm{vec}\left(\mathbf{X}_{(1)}\right)$ stacks vertically into a vector the columns of the matrix $\mathbf{X}_{(1)}$. Many product operations have been defined for tensors (e.g., see Lee and Cichocki (2016)), but here we constrain ourselves to the operators used in this work. Concerning the basic product operations, the *scalar product* between two tensors $\mathcal{X}, \mathcal{Y}$ of equal order and same dimensions, $I_1, \ldots, I_N$, is defined as:

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1,\ldots,i_N} \mathcal{Y}_{i_1,\ldots,i_N} = \sum_{i_1,\ldots,i_N} \mathcal{X}_{i_1,\ldots,i_N} \mathcal{Y}_{i_1,\ldots,i_N}. \tag{2.3}$$

For the ease of notation, we will use the multiple-index summation for indicating the sum over all the corresponding indices.

The *mode-M contracted product* of two tensors $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_M}$ and $\mathcal{Y} \in \mathbb{R}^{J_1 \times \ldots \times J_N}$ with $I_M = J_M$, denoted $\mathcal{X} \times^M \mathcal{Y}$, yields a tensor $\mathcal{Z} \in \mathbb{R}^{I_1 \times \ldots \times I_{M-1} \times J_1 \times \ldots \times J_{N-1}}$ of order $M + N - 2$, with entries:

$$\mathcal{Z}_{i_1,\ldots,i_{M-1},j_1,\ldots,j_{N-1}} = (\mathcal{X} \times^M \mathcal{Y})_{i_1,\ldots,i_{M-1},j_1,\ldots,j_{N-1}} = \sum_{i_M=1}^{I_M} \mathcal{X}_{i_1,\ldots,i_M} \mathcal{Y}_{j_1,\ldots,i_M,\ldots,j_N}. \tag{2.4}$$

Therefore, it is a generalization of the matrix product. The notation $\times^{1\ldots M}$ is used to denote a sequence of mode-m contracted products, with $m = 1, \ldots, M$.

The *mode-n product* between a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ and a matrix $\mathbf{A} \in \mathbb{R}^{J \times I_n}$, $1 \leq n \leq N$, is denoted by $\mathcal{X} \bar{\times}_n A$ and yields a tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \ldots, I_{n-i}, J, I_{n+1}, \ldots \times I_N}$ of the same order of $\mathcal{X}$, with the n-th mode's length changed. Each mode-n fiber of the tensor is multiplied by the matrix A, which yields element-wise:

$$\mathcal{Y}_{i_1,\ldots,i_{n-1},j,i_{n+1},\ldots,i_N} = (\mathcal{X} \bar{\times}_n \mathbf{A})_{i_1,\ldots,i_{n-1},j,i_{n+1},\ldots,i_N} = \sum_{i_n=1}^{I_n} \mathcal{X}_{i_1,\ldots,i_N} \mathbf{A}_{j,i_n}. \tag{2.5}$$

Analogously, the *mode-n product between a tensor and a vector*, i.e. between $\mathcal{X}$ and $\mathbf{v} \in \mathbb{R}^{I_n}$, yields a lower order tensor, since the n-th mode is suppressed as a consequence of the product. It is given, element-wise, by:

$$\mathcal{Y}_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_N} = (\mathcal{X} \times_n \mathbf{v})_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_N} = \sum_{i_n=1}^{I_n} \mathcal{X}_{i_1,\ldots,i_n,\ldots,i_N} \mathbf{v}_{i_n} , \tag{2.6}$$

with $\mathcal{Y} \in \mathbb{R}^{I_1 \times \ldots, I_{n-i}, I_{n+1}, \ldots \times I_N}$. It is clear that, as for the matrix dot product, the order of the elements in the multiplication matters and both products are not commutative.

The *Hadamard product* $\odot$ is defined in the same usual way as for matrices, i.e. the element-wise multiplication. Formally, for $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$, $\mathcal{Y} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ and $\mathcal{Z} \in \mathbb{R}^{I_1^{\times} \ldots \times I_N}$ it holds:

$$\mathcal{Z}_{i_1,\ldots,i_N} = (\mathcal{X} \odot \mathcal{Y})_{i_1,\ldots,i_N} = \mathcal{X}_{i_1,\ldots,i_N} \mathcal{Y}_{i_1,\ldots,i_N} . \tag{2.7}$$

Finally, let $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_M}$ and $\mathcal{Y} \in \mathbb{R}^{J_1 \times \ldots \times J_N}$. The *outer product* $\circ$ of two tensors is the tensor $\mathcal{Z} \in \mathbb{R}^{I_1 \times \ldots \times I_M \times J_1 \times \ldots \times J_N}$ whose entries are:

$$\mathcal{Z}_{i_1,\ldots,i_M,j_1,\ldots,j_N} = (\mathcal{X} \circ \mathcal{Y})_{i_1,\ldots,i_M,j_1,\ldots,j_N} = \mathcal{X}_{i_1,\ldots,i_M} \mathcal{Y}_{j_1,\ldots,j_N} . \tag{2.8}$$

For example, the outer product of two vectors is a matrix, while the outer product of two matrices is a fourth order tensor.

Tensor decompositions represent the core of current statistical models dealing with multidimensional variables since many of them allow to represent a tensor as a function of lower dimensional variables, such as matrices of vectors, linked by suitable multidimensional operations. We now define two tensor decompositions, the Tucker and the parallel factor (PARAFAC), which are useful in our applications because the elements of the decomposition are generally low dimensional and easier to handle than the original tensor. Let R be the rank of the tensor $\mathcal{X}$, that is minimum number of rank-1 tensors whose linear combination yields $\mathcal{X}$. A N-order tensor is of rank 1 when it is the outer product of N vectors.

The Tucker decomposition is a higher-order generalization of the Principal Component Analysis (PCA): a tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ is decomposed into the product (along the corresponding modes) of a "core" tensor $\mathcal{G} \in \mathbb{R}^{g_1 \times \ldots \times g_N}$ and N factor matrices $\mathbf{A}^{(m)} \in \mathbb{R}^{I_m \times J_m}$, $m = 1, \ldots, N$:

$$\mathcal{B} = \mathcal{G} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \times_3 \ldots \times_N \mathbf{A}^{(N)} = \sum_{i_1=1}^{g_1} \sum_{i_2=1}^{g_2} \ldots \sum_{i_N=1}^{g_N} \mathcal{G}_{i_1,i_2,\ldots,i_N} \mathbf{a}_{i_1}^{(1)} \circ \mathbf{a}_{i_2}^{(2)} \circ \ldots \circ \mathbf{a}_{i_N}^{(N)} , \tag{2.9}$$

where $\mathbf{a}_{i_l}^{(m)} \in \mathbb{R}^{g_m}$ is the m-th column of the matrix $\mathbf{A}^{(m)}$. As a result, each entry of the tensor is obtained as:

$$\mathcal{B}_{j_1,\ldots,j_N} = \sum_{i_1=1}^{g_1} \sum_{i_2=1}^{g_2} \ldots \sum_{i_N=1}^{g_N} \mathcal{G}_{i_1,i_2,\ldots,i_N} \cdot \mathbf{A}_{i_1,j_1}^{(1)} \cdots \mathbf{A}_{i_N,j_N}^{(N)} \quad j_m = 1, \ldots, I_m, \; m = 1, \ldots, N . \tag{2.10}$$

A special case of the Tucker decomposition, called PARAFAC(R)[2], is obtained when the core tensor is the identity tensor and the factor matrices have all the same number of columns, R. A graphical representation of this decomposition for a third-order tensor is shown in Fig. (2.1). More precisely, the PARAFAC(R) is a low rank decomposition which represents a tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ as a finite sum of R rank-1 tensors obtained as the outer products of N

[2] See Harshman (1970). Some authors (e.g. Carroll and Chang (1970) and Kiers (2000)) use the term CODECOMP or CP instead of PARAFAC.

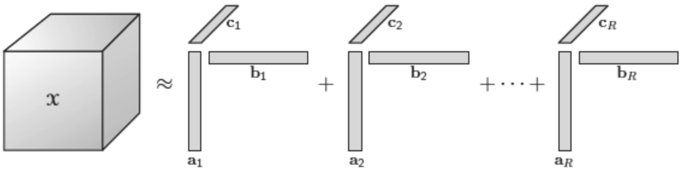


FIGURE 2.1: PARAFAC decomposition of $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, with $\mathbf{a}_r \in \mathbb{R}^{I_1}$, $\mathbf{b}_r \in \mathbb{R}^{I_2}$ and $\mathbf{c}_r \in \mathbb{R}^{I_3}$, $r = 1, \dots, R$. Figure from Kolda and Bader (2009).

vectors, also called PARAFAC marginals[3] $\boldsymbol{\beta}_j^{(r)} \in \mathbb{R}^{I_j}$, $j = 1, \dots, N$:

$$\mathcal{B} = \sum_{r=1}^{R} \mathcal{B}_r = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \dots \circ \boldsymbol{\beta}_N^{(r)} . \tag{2.11}$$

Remark 2.2.1
There exists a one-to-one relation between the mode-n product between a tensor and a vector and the vectorisation and matricization operators. Consider a N-order tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ for which is specified a PARAFAC(R) decomposition, a $(N-1)$-order tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \dots \times I_{N-1}}$ and a vector $\mathbf{x} \in \mathbb{R}^{I_N}$. Then:

$$\mathcal{Y} = \mathcal{B} \times_N \mathbf{x} \iff \operatorname{vec}(\mathcal{Y}) = \mathbf{B}'_{(N)}\mathbf{x} \iff \operatorname{vec}(\mathcal{Y})' = \mathbf{x}'\mathbf{B}_{(N)} \tag{2.12}$$

and, denoting $\boldsymbol{\beta}_j^{(r)}$, for $j = 1, \dots, N$ and $r = 1, \dots, R$, the marginals of the PARAFAC(R) decomposition of $\mathcal{B}$ we have:

$$\mathbf{B}_{(N)} = \sum_{r=1}^{R} \boldsymbol{\beta}_N^{(r)} \operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \dots \circ \boldsymbol{\beta}_{N-1}^{(r)}\right)' . \tag{2.13}$$

These relations allows to establish a link between operators defined on tensors and operators defined on matrices, for which plenty of properties are known from linear algebra.

Remark 2.2.2
For two vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^m$ the following relations hold between the outer product, the Kronecker product $\otimes$ and the vectorisation operator:

$$\mathbf{u} \otimes \mathbf{v}' = \mathbf{u} \circ \mathbf{v} = \mathbf{u}\mathbf{v}' \tag{2.14}$$
$$\mathbf{u} \otimes \mathbf{v} = \operatorname{vec}(\mathbf{v} \circ \mathbf{u}) . \tag{2.15}$$

2.2.2 A General Dynamic Model

The new model we propose, in its most general formulation is:

$$\mathcal{Y}_t = \mathcal{A}_0 + \sum_{j=1}^{p} \mathcal{A}_j \times_{N+1} \operatorname{vec}\left(\mathcal{Y}_{t-j}\right) + \mathcal{B} \times_{N+1} \operatorname{vec}(\mathcal{X}_t) + \mathcal{C} \times_{N+1} \mathbf{z}_t + \mathcal{D} \times_n \mathbf{W}_t + \mathcal{E}_t, \tag{2.16}$$
$$\mathcal{E}_t \overset{iid}{\sim} \mathcal{N}_{I_1,\dots,I_N}(0, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_N) ,$$

[3] An alternative representation may be used, if all the vectors $\boldsymbol{\beta}_j^r$ are normalized to have unitary length. In this case the weight of each component r is captured by the r-th component of the vector $\boldsymbol{\lambda} \in \mathbb{R}^R$:

$$\mathcal{B} = \sum_{r=1}^{R} \lambda_r \left(\boldsymbol{\beta}_1^{(r)} \circ \dots \circ \boldsymbol{\beta}_N^{(r)}\right)$$

.

where the tensor response and noise $\mathcal{Y}_t$, $\mathcal{E}_t$ are N-order tensors of sizes $I_1 \times \ldots \times I_N$, while the covariates include a M-order tensor $\mathcal{X}_t$ of sizes $J_1 \times \ldots \times J_M$, a matrix $\mathbf{W}_t$ with dimensions $I_n \times K$ and a vector $\mathbf{z}_t$ of length Q.

The coefficients are all tensors of suitable order and sizes: $\mathcal{A}_j$ have dimensions $I_1 \times \ldots \times I_N \times I^*$, with $I^* = \prod_i I_i$, $\mathcal{B}$ has dimensions $I_1 \times \ldots \times I_N \times J^*$, with $J^* = \prod_j J_j$, $\mathcal{C}$ has dimensions $I_1 \times \ldots \times I_N \times Q$ and $\mathcal{D}$ has sizes $I_1 \times \ldots \times I_{n-1} \times K \times I_{n+1} \ldots \times I_N$. The symbol $\times_n$ stands for the mode-n product between a tensor and a vector defined in eq. (A.4). The reason for the use of tensors coefficients, as opposed to scalars and vectors, is twofold: first, this permits each entry of each covariate to exert a different effect on each entry of the response variable; second, the adoption of tensors allows to exploit the various decompositions, which are fundamental for providing a parsimonious and flexible parametrization of the statistical model.

The noise is assumed to follow a tensor normal distribution (see Ohlson et al. (2013), Manceur and Dutilleul (2013), Arashi (2017)), a generalization of the multivariate normal distribution. Let $\mathcal{X}$ and $\mathcal{M}$ be two N-order tensors of dimensions $I_1, \ldots, I_N$. Define $I^* = \prod_{j=1}^N I_j$, $I^*_{-i} = \prod_{j \neq i} I_j$ and let $\times^{1 \ldots N}$ be a sequence of mode-j contracted products, $j = 1, \ldots, N$, between the $(K+N)$-order tensor $\mathcal{X}$ and the $(N+M)$-order tensor $\mathcal{Y}$ of conformable dimensions, defined as follows:

$$\left(\mathcal{X} \times^{1\ldots N} \mathcal{Y}\right)_{j_1,\ldots,j_K,h_1,\ldots,h_M} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \mathcal{X}_{j_1,\ldots,j_K,i_1,\ldots,i_N} \mathcal{Y}_{i_N,\ldots,i_1,h_1,\ldots,h_M} . \tag{2.17}$$

Finally, let $\mathbf{U}_j \in \mathbb{R}^{I_j \times I_j}$, $j \in \{1, \ldots, N\}$ be positive definite matrices. The probability density function of a N-order tensor normal distribution with mean array $\mathcal{M}$ and positive definite covariance matrices $U_1, \ldots, U_N$, is given by:

$$f_{\mathcal{X}}(\mathcal{X}) = (2\pi)^{-\frac{d^*}{2}} \prod_{j=1}^{N} \left|U_j\right|^{-\frac{I^*_{-j}}{2}} \exp\left\{ -\frac{1}{2} (\mathcal{X} - \mathcal{M}) \times^{1\ldots N} \left(\circ_{j=1}^{N} \mathbf{U}_j^{-1} \right) \times^{1\ldots N} (\mathcal{X} - \mathcal{M}) \right\} . \tag{2.18}$$

The tensor normal distribution can be rewritten as a multivariate normal distribution with separable covariance matrix for the vectorized tensor, more precisely it holds (see Ohlson et al. (2013)) $\mathcal{X} \sim \mathcal{N}_{I_1,\ldots,I_N}(\mathcal{M}, \mathbf{U}_1, \ldots, \mathbf{U}_N) \iff \mathrm{vec}\left(\mathcal{X}\right) \sim \mathcal{N}_{I_1 \cdots I_N}(\mathrm{vec}\left(\mathcal{M}\right), \mathbf{U}_N \otimes \ldots \otimes \mathbf{U}_1)$. The restriction imposed by the separability assumption allows to reduce the number of parameters to estimate with respect to the unrestricted vectorized from, while allowing both within and between mode dependence.

The unrestricted model in eq. (2.16) cannot be estimated, as the number of parameters greatly outmatches the available data. We address this issue by assuming a PARAFAC(R) decomposition for the tensor coefficients, which makes the estimation feasible by reducing the dimension of the parameter space. For example, let $\mathcal{B}$ be a N-order tensor of sizes $I_1 \times \ldots \times I_N$ and rank R, then the number of parameters to estimate in the unrestricted case is given by $\prod_{i=1}^N I_i$ while in the PARAFAC(R) restricted model is $R \sum_{i=1}^N I_i$.

Example 2.2.1

For the sake of exposition, consider the model in eq. (2.16) where the response is a third-order tensor $\mathcal{Y}_t \in \mathbb{R}^{k \times k \times k^2}$ and the covariates include only a constant term, that is a coefficient tensor $\mathcal{A}_0$ of the same size. Define by $k_{\mathcal{E}}$ the number of parameters of the noise distribution. As a result, the total number of parameters to estimate in the unrestricted case is given by:

$$\prod_{i=1}^{3} I_i + k_{\mathcal{E}} = \mathcal{O}(k^4) , \tag{2.19}$$

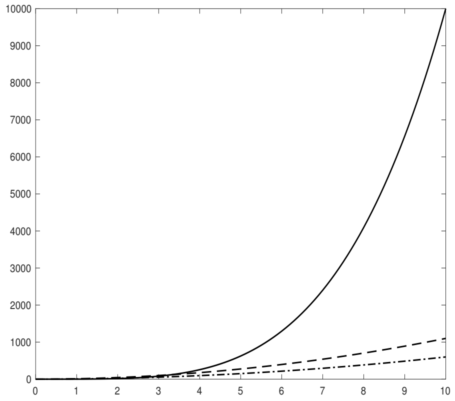


FIGURE 2.2: Number of parameters (*vertical axis*) as function of the response dimension (*horizontal axis*) for unconstrained (*solid*) and PARAFAC(R) with $R = 10$ (*dashed*) and $R = 5$ (*dotted*).

while assuming a PARAFAC(R) decomposition on $\mathcal{A}_0$ it reduces to:

$$\sum_{r=1}^{R}\sum_{i=1}^{3} I_i + k_{\mathcal{E}} = \mathcal{O}(k^2)\,. \tag{2.20}$$

The magnitude of this reduction is illustrated in Fig. (2.2), for two different values of the rank.

A well-known issue is that a low rank decomposition is not unique. In a statistical model this translates into an identification problem for the PARAFAC marginals $\boldsymbol{\beta}_j^{(r)}$ arising from three sources:

(i) *scale* identification, because replacing $\boldsymbol{\beta}_j^{(r)}$ with $\lambda_{jr}\boldsymbol{\beta}_j^{(r)}$ for $\prod_{j=1}^{N}\lambda_{jr} = 1$ does not alter the outer product;

(ii) *permutation* identification, since for any permutation of the indices $\{1,\ldots,R\}$ the outer product of the original vectors is equal to that of the permuted ones;

(iii) *orthogonal transformation* identification, due to the fact that multiplying two marginals by an orthonormal matrix $\mathbf{Q}$ leaves unchanged the outcome $\boldsymbol{\beta}_j^{(r)}\mathbf{Q} \circ \boldsymbol{\beta}_k^{(r)}\mathbf{Q} = \boldsymbol{\beta}_j^{(r)} \circ \boldsymbol{\beta}_k^{(r)}$.

In our framework these issues do not hamper the inference as our interest is only in the coefficient tensor, which is exactly identified. In fact, we use the PARAFAC decomposition as a practical modelling tool without attaching any interpretation to its marginals.

2.2.3 Important special cases

The model in eq. (2.16) is a generalization of several well-known econometric models, as shown in the following remarks.

Remark 2.2.3 (Univariate)
If we set $I_j = 1$ for $j = 1,\ldots,N$, then the model in eq. (2.16) reduces to a univariate regression:

$$y_t = \mathcal{A} + \mathcal{B}'\operatorname{vec}(\mathbf{X}_t) + \mathcal{C}'\mathbf{z}_t + \epsilon_t, \qquad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0,\sigma^2), \tag{2.21}$$

where the coefficients reduce to $\mathcal{A} = \bar{\alpha} \in \mathbb{R}$, $\mathcal{B} = \boldsymbol{\beta} \in \mathbb{R}^Q$ and $\mathcal{C} = \boldsymbol{\gamma} \in \mathbb{R}^J$. See Appendix B.1 for further details.

Remark 2.2.4 (SUR)
If we set $I_j = 1$ for $j = 2, \ldots, N$ and define the unit vector $\boldsymbol{\iota} \in \mathbb{R}^{I_1}$, then the model in eq. (2.16) reduces to a multivariate regression which is interpretable as a Seemingly Unrelated Regression (SUR) model (Zellner (1962)):

$$\mathbf{y}_t = \mathcal{A} + \mathcal{B} \times_2 \mathbf{z}_t + \mathcal{C} \times_2 \mathrm{vec}\left(\mathbf{X}_t\right) + \mathcal{D} \times_1 \mathrm{vec}\left(W_t\right) + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \overset{iid}{\sim} \mathcal{N}_m\left(0, \boldsymbol{\Sigma}\right), \tag{2.22}$$

where the tensors of coefficients can be expressed as: $\mathcal{A} = \boldsymbol{\alpha} \in \mathbb{R}^m$, $\mathcal{B} = \bar{\mathbf{B}} \in \mathbb{R}^{m \times J}$, $\mathcal{C} = \mathbf{C} \in \mathbb{R}^{m \times Q}$ and $\mathcal{D} = \mathbf{d} \in \mathbb{R}^m$. See Appendix B.1 for further details.

Remark 2.2.5 (VARX and Panel VAR)
Consider the setup of the previous Remark 2.2.4. If we choose $\mathbf{z}_t = \mathbf{y}_{t-1}$ we end up with an (unrestriced) VARX(1) model. Notice that another vector of regressors $\mathbf{w}_t = \mathrm{vec}\left(\mathbf{W}_t\right) \in \mathbb{R}^q$ may enter the regression (2.22) pre-multiplied (along mode-3) by a tensor $\mathcal{D} \in \mathbb{R}^{m \times n \times q}$. Since we are not putting any kind of restrictions on the covariance matrix $\boldsymbol{\Sigma}$ in (2.22), the general model (2.16) encompasses as a particular case also the panel VAR models of Canova and Ciccarelli (2004), Canova et al. (2007), Canova and Ciccarelli (2009) and Canova et al. (2012).

Remark 2.2.6 (VECM)
It is possible to interpret the model in eq. (2.16) as a generalisation of the Vector Error Correction Model (VECM) widely used in multivariate time series analysis (see Engle and Granger (1987), Schotman and Van Dijk (1991)). A standard K-dimensional VAR(1) model reads:

$$\mathbf{y}_t = \boldsymbol{\Pi} \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}) \,. \tag{2.23}$$

Defining $\Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ and $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $K \times R$ matrices of rank $R < K$, we obtain the VECM used for studying the cointegration relations between the components of $\mathbf{y}_t$:

$$\Delta \mathbf{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t \,. \tag{2.24}$$

Since $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}' = \sum_{r=1}^R \boldsymbol{\alpha}_{:,r} \boldsymbol{\beta}'_{:,r} = \sum_{r=1}^R \tilde{\boldsymbol{\beta}}_1^{(r)} \circ \tilde{\boldsymbol{\beta}}_2^{(r)}$, we can interpret the VECM model in the previous equation as a particular case of the model in eq. (2.16) where the coefficient $\mathcal{B}$ is the matrix $\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}'$. Furthermore by writing $\Pi = \sum_{r=1}^R \tilde{\boldsymbol{\beta}}_1^{(r)} \circ \tilde{\boldsymbol{\beta}}_2^{(r)}$ we can interpret this relation as a rank-R PARAFAC decomposition of Π. Thus we can interpret the rank of the PARAFAC decomposition for the matrix of coefficients as the cointegration rank and, in presence of cointegrating relations, the vectors $\tilde{\boldsymbol{\beta}}_1^{(r)}$ are the mean-reverting coefficients and $\tilde{\boldsymbol{\beta}}_2^{(r)} = (\tilde{\beta}_{2,1}^{(r)}, \ldots, \tilde{\beta}_{2,K}^{(r)})$ are the cointegrating vectors. In fact, the PARAFAC(R) decomposition for matrices corresponds to a low rank (R) matrix approximation (see Eckart and Young (1936)). We make reference to Appendix B.1 for further details.

Remark 2.2.7 (Tensor AR)
By removing all the covariates from eq. (2.16) except the lags of the dependent variable, we obtain a tensor autoregressive model:

$$\mathcal{Y}_t = \mathcal{A}_0 + \sum_{j=1}^p \mathcal{A}_j \times_{D+1} \mathcal{Y}_{t-j} + \mathcal{E}_t \qquad \mathcal{E}_t \overset{iid}{\sim} \mathcal{N}_{I_1, \ldots, I_N}(0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N) \,. \tag{2.25}$$

2.3 Bayesian Inference

In this section, without loss of generality, we present the inference procedure for a special case of the model in eq. (2.16), given by:

$$\mathbf{Y}_t = \mathcal{B} \times_3 \mathrm{vec}\left(\mathbf{X}_t\right) + \mathbf{E}_t, \quad \mathbf{E}_t \overset{iid}{\sim} \mathcal{N}_{I_1, I_2}(\mathbf{0}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \,, \tag{2.26}$$

which can also be rewritten in vectorized form as:

$$\operatorname{vec}(\mathbf{Y}_t) = \mathbf{B}'_{(3)} \operatorname{vec}(\mathbf{X}_t) + \operatorname{vec}(\mathbf{E}_t), \quad \operatorname{vec}(E_t) \overset{iid}{\sim} \mathcal{N}_{I_1 I_2}(\mathbf{0}, \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1). \tag{2.27}$$

Here $\mathbf{Y}_t \in \mathbb{R}^{I_1 \times I_2}$ is a matrix response, $\mathbf{X}_t \in \mathbb{R}^{I_1 \times I_2}$ is a covariate matrix of the same size of $\mathbf{Y}_t$ and $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times I_1 I_2}$ is a coefficient tensor. The noise term $\mathbf{E}_t \in \mathbb{R}^{I_1 \times I_2}$ is distributed according to a matrix variate normal distribution, with zero mean and covariance matrices $\mathbf{\Sigma}_1 \in \mathbb{R}^{I_1 \times I_1}$ and $\mathbf{\Sigma}_2 \in \mathbb{R}^{I_2 \times I_2}$ accounting for the covariance between the columns and the rows, respectively. This distribution is a particular case of the tensor Gaussian introduced in eq. (2.18) whose probability density function is given by:

$$f_X(\mathbf{X}) = (2\pi)^{-\frac{I_1 I_2}{2}} |\mathbf{U}_2|^{-\frac{I_1}{2}} |\mathbf{U}_1|^{-\frac{I_2}{2}} \exp\left\{ -\frac{1}{2} \mathbf{U}_2^{-1} (\mathbf{X} - \mathbf{M})' \mathbf{U}_1^{-1} (\mathbf{X} - \mathbf{M}) \right\} \tag{2.28}$$

where $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2}$, $\mathbf{M} \in \mathbb{R}^{I_1 \times I_2}$ is the mean matrix and the covariance matrices are $\mathbf{U}_j \in \mathbb{R}^{I_j \times I_j}$, $j = 1, 2$, where index 1 represents the rows and index 2 stands for the columns of the variable $\mathbf{X}$.

The choice the Bayesian approach for inference is motivated by the fact that the large number of parameters may lead to an over-fitting problem, especially when the samples size is rather small. This issue can be addressed by the indirect inclusion of parameter restrictions through a suitable specification of the corresponding prior distribution. Considering the unrestricted model in eq. (2.26), it would be necessary to define a prior distribution on the three-dimensional array $\mathcal{B}$. The literature on this topic is scarce: though Ohlson et al. (2013) and Manceur and Dutilleul (2013) presented the family of elliptical array-valued distributions, which include the tensor normal and tensor t, the latter are rather inflexible as imposing some structure on a subset of the entries of the array is very complicated.

We assume a PARAFAC(R) decomposition on the coefficient tensor for achieving two goals: first, by reducing the parameter space this assumption makes estimation feasible; second, the decomposition transforms a multidimensional array into the outer product of vectors, we are left we the choice of a prior distribution on vectors, for which many constructions are available. In particular, we can incorporate sparsity beliefs by specifying a suitable shrinkage prior directly on the marginals of the PARAFAC. Indirectly, this introduces *a priori* sparsity on the coefficient tensor.

2.3.1 Prior Specification

The choice of the prior distribution on the PARAFAC marginals is crucial for recovering the sparsity pattern of the coefficient tensor and for the efficiency of the inference. In the Bayesian literature the global-local class of prior distributions represent a popular and successful structure for providing shrinkage and regularization in a wide range of models and applications. These priors are based on scale mixtures of normal distributions, where the different components of the covariance matrix produce desirable shrinkage properties of the parameter. By construction, global-local priors are not suited for recovering an exact zero (differently from spike-and-slab priors, see Mitchell and Beauchamp (1988a), George and McCulloch (1997), Ishwaran and Rao (2005)), instead they can be recovered via post-estimation thresholding (see Park and Casella (2008)). However, spike-and-slab priors become intractable as the dimensionality of the parameter grows. By contrast, the global-local shrinkage priors have greater scalability and thus represent a desirable choice in high-dimensional models, such as our framework. Motivated by these arguments, we adopt the hierarchical specification forwarded by Guhaniyogi et al. (2017) in order to define adequate global-local shrinkage priors for the marginals[4].

[4] This class of shrinkage priors has been firstly proposed by Bhattacharya et al. (2015) and Zhou et al. (2015).

The global-local shrinkage prior for each PARAFAC marginal $\boldsymbol{\beta}_j^{(r)}$ of the coefficient tensor $\mathcal{B}$ is defined as a scale mixture of normals centred in zero, with three components for the covariance. The global component τ is drawn from a gamma distribution[5]. The vector of component-level (shared by all marginals in the r-th component of the decomposition) variances $\boldsymbol{\phi}$ is sampled from a R-dimensional Dirichlet distribution with parameter $\boldsymbol{\alpha} = \alpha \boldsymbol{\iota}_R$, where $\boldsymbol{\iota}_R$ is the vector of ones of length R. Finally, the local component of the variance is a diagonal matrix $\mathbf{W}_{j,r} = \text{diag}(\mathbf{w}_{j,r})$ whose entries are exponentially distributed with hyper-parameter $\lambda_{j,r}$. The latter is a key parameter for driving the shrinkage to zero of the marginals and is drawn from a gamma distribution. Summarizing, for $p = 1, \ldots, I_j$, $j = 1, \ldots, 3$ and $r = 1, \ldots, R$ we have the following hierarchical prior structure for each vector of the PARAFAC(R) decomposition in eq. (A.14):

$$\pi(\boldsymbol{\phi}) \sim \mathcal{D}ir(\alpha \boldsymbol{\iota}_R) \tag{2.29a}$$

$$\pi(\tau) \sim \mathcal{G}a(a_\tau, b_\tau) \tag{2.29b}$$

$$\pi(\lambda_{j,r}) \sim \mathcal{G}a(a_\lambda, b_\lambda) \tag{2.29c}$$

$$\pi(w_{j,r,p}|\lambda_{j,r}) \sim \mathcal{E}xp(\lambda_{j,r}^2/2) \tag{2.29d}$$

$$\pi\left(\boldsymbol{\beta}_j^{(r)} \middle| \mathbf{W}_{j,r}, \boldsymbol{\phi}, \tau\right) \sim \mathcal{N}_{I_j}(\mathbf{0}, \tau \phi_r \mathbf{W}_{j,r}) \,. \tag{2.29e}$$

Several shrinking prior distributions have been proposed in the literature for dealing with large dimensional models (e.g., SUR and VAR models), among the most frequently used in econometrics we mention the stochastic search variable selection (SSVS) introduced by George and McCulloch (1993). It relies on a spike and slab prior distribution (see Mitchell and Beauchamp (1988b), George and McCulloch (1997)) and has been extended in various ways. See Dellaportas et al. (2002) for a review and George et al. (2008), Jochmann et al. (2010) and Wang (2010) for applications in econometrics and time series analysis. Other classes of priors are the Bayesian lasso of Park and Casella (2008) and the Bayesian elastic-net (e.g., see Gefang (2014), Korobilis (2013a), Korobilis (2013b)).

By definition, hierarchical priors based on spike and slab mixtures are able to recover exact zeros of the parameters thanks to the Dirac mass at zero (the "spike" of the mixture), whereas global-local priors based on scale mixtures of Normals are only capable of shrinking the parameter towards zero. They cannot recover exact zeros, unless a particular choice of the scale distributions assigning positive mass at zero is made. However, when a global-local prior is chosen, exact zero values for the estimated coefficients can be obtained by post-estimation thresholding, following Park and Casella (2008).

The main reason motivating our choice in favour of a global-local shrinkage prior is the computational gain it allows as compared to the other hierarchical priors: the former avoids the need to introduce latent allocation variables for each coefficient, thus resulting in a lower dimensional parameter space and higher mixing of the chain as compared to spike and slab shrinkage priors. Recently Ročková and George (2014) proposed a feasible implementation of the SSVS in high dimensions based on the expectation maximization (EM) algorithm.

Concerning the covariance matrices for the noise term in eq. (2.16), the Kronecker structure does not allow to separately identify the scale of the covariance matrices U_n, thus requiring the specification of further restrictions. Wang and West (2009) and Dobra (2015) adopt independent hyper-inverse Wishart prior distributions (Dawid and Lauritzen (1993)) for each U_n, then impose the identification restriction $U_{n,11} = 1$ for $n = 2, \ldots, N$. Instead, Hoff (2011) suggests to introduce dependence between the Inverse Wishart prior distribution

[5] We use the shape-rate formulation for the gamma distribution:

$$x \sim \mathcal{G}a(a,b) \qquad \Longleftrightarrow \qquad f(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \qquad a > 0, b > 0$$

.

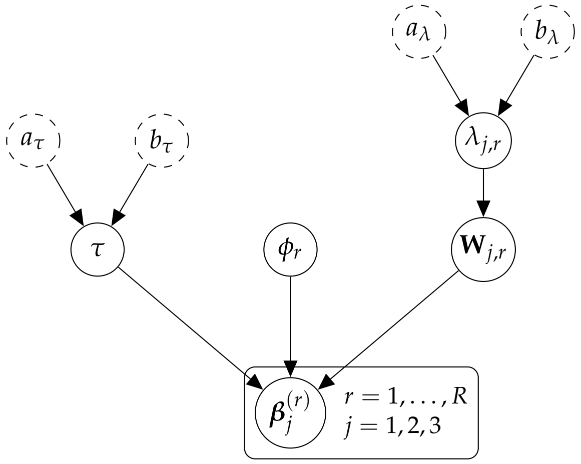


FIGURE 2.3: Hierarchical shrinkage prior for the PARAFAC marginals. White circles with continuous border represent the parameters, white circles with dashed border represent fixed hyper-parameters.

$\mathcal{IW}(\nu_n, \gamma\mathbf{\Psi}_n)$ of each U_n, $n = 1, \dots, N$, via a hyper-parameter $\gamma \sim \mathcal{G}a(a, b)$ affecting the scale of each location matrix parameter. Finally, the hard constraint $\mathbf{\Sigma}_n = \mathbf{I}_{I_n}$ (where $\mathbf{I}_k$ is the identity matrix of size k), for all but one n, implicitly imposes that the dependence structure within different modes is the same, but there is no dependence between modes. To account for marginal dependence, it is possible to add a level of hierarchy by introducing a hyper-parameter in the spirit of Hoff (2011). Following Hoff (2011), we assume conditionally independent inverse Wishart prior distributions for the covariance matrices of the error term $\mathbf{E}_t$ and add a level of hierarchy via the hyper-parameter γ which governs the scale of the covariance matrices:

$$\pi(\gamma) \sim \mathcal{G}a(a_\gamma, b_\gamma) \tag{2.30a}$$
$$\pi(\mathbf{\Sigma}_1|\gamma) \sim \mathcal{IW}_{I_1}(\nu_1, \gamma\mathbf{\Psi}_1) \tag{2.30b}$$
$$\pi(\mathbf{\Sigma}_2|\gamma) \sim \mathcal{IW}_{I_2}(\nu_2, \gamma\mathbf{\Psi}_2)\,. \tag{2.30c}$$

Defining the vector of all parameters as $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\phi}, \tau, \Lambda, \mathbf{W}, \mathcal{B}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2\}$, with $\mathbf{\Lambda} = \{\lambda_{j,r} : j = 1, \dots, 3,\ r = 1, \dots, R\}$ and $\mathbf{W} = \{W_{j,r} : j = 1, \dots, 3,\ r = 1, \dots, R\}$, the joint prior distribution is given by:

$$\pi(\boldsymbol{\theta}) = \pi(\mathcal{B}|\mathbf{W}, \boldsymbol{\phi}, \tau)\pi(\mathbf{W}|\mathbf{\Lambda})\pi(\boldsymbol{\phi})\pi(\tau)\pi(\mathbf{\Lambda})\pi(\mathbf{\Sigma}_1|\gamma)\pi(\mathbf{\Sigma}_2|\gamma)\pi(\gamma). \tag{2.31}$$

The directed acyclic graphs (DAG) of the hierarchical shrinkage prior on the PARAFAC marginals $\boldsymbol{\beta}_j^{(r)}$ and the overall prior structure are given in Figs. 2.3-2.4, respectively.

2.3.2 Posterior Computation

The likelihood function of the model in eq. (2.26) is given by:

$$L\left(\mathbf{Y}_1, \dots, \mathbf{Y}_T|\boldsymbol{\theta}\right) = \prod_{t=1}^{T}(2\pi)^{-\frac{I_1 I_2}{2}}\left|\mathbf{\Sigma}_2\right|^{-\frac{I_1}{2}}\left|\mathbf{\Sigma}_1\right|^{-\frac{I_2}{2}}\exp\left\{-\frac{1}{2}\mathbf{\Sigma}_2^{-1}(\mathbf{Y}_t - \mathcal{B}\times_3\mathbf{x}_t)'\mathbf{\Sigma}_1^{-1}(\mathbf{Y}_t - \mathcal{B}\times_3\mathbf{x}_t)\right\}, \tag{2.32}$$

where $\mathbf{x}_t = \mathrm{vec}\left(\mathbf{X}_t\right)$. Since the posterior distribution is not tractable in closed form, we adopt an MCMC procedure based on Gibbs sampling. The computations and technical details of the derivation of the posterior distributions are given in Appendix B.3. As a consequence of the hierarchical structure of the prior, we can articulate the sampler in three main blocks:

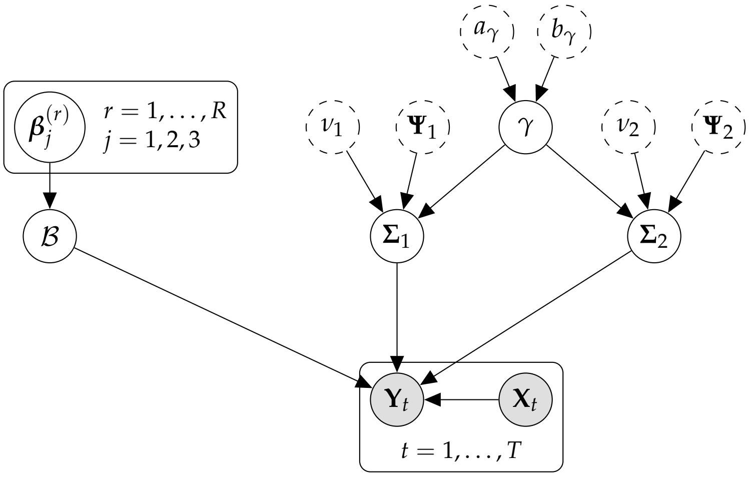


FIGURE 2.4: Overall prior structure. Gray circles represent observable variables, white circles with continuous border represent the parameters, white circles with dashed border represent fixed hyper-parameters.

I) sample the hyper-parameters of the global and component-level variance for the marginals, according to:

$$p(\boldsymbol{\phi}, \tau | \mathcal{B}, \mathbf{W}) = p(\boldsymbol{\phi}, \tau | \mathcal{B}, \mathbf{W}) \tag{2.33}$$

(i) sample independently the auxiliary variable ψ_r, for $r = 1, \dots, R$, from:

$$p(\psi_r | \mathcal{B}, \mathbf{W}) \propto GiG\left(\alpha - \frac{I_0}{2}, 2b_\tau, 2C_r\right) \tag{2.34}$$

then, for $r = 1, \dots, R$:

$$\phi_r = \frac{\psi_r}{\sum_{l=1}^{R} \psi_l}. \tag{2.35}$$

(ii) finally, sample τ from:

$$p(\tau | \mathcal{B}, \mathbf{W}, \boldsymbol{\phi}) \propto GiG\left(a_\tau - \frac{RI_0}{2}, 2b_\tau, 2\sum_{r=1}^{R} \frac{C_r}{\phi_r}\right). \tag{2.36}$$

II) define $\mathbf{Y} = \{\mathbf{Y}_t\}_{t=1}^{T}$, then sample from the posterior of the hyper-parameters of the local component of the variance of the marginals and the marginals themselves, as follows:

$$\begin{aligned} p\left(\boldsymbol{\beta}_j^{(r)}, \mathbf{W}_{j,r}, \lambda_{j,r} \middle| \boldsymbol{\phi}, \tau, \mathbf{Y}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2\right) &= p\left(\lambda_{j,r} | \boldsymbol{\beta}_j^{(r)}, \phi_r, \tau\right) p\left(w_{j,r,p} | \lambda_{j,r}, \phi_r, \tau, \boldsymbol{\beta}_j^{(r)}\right) \\ &\quad \cdot p\left(\boldsymbol{\beta}_j^{(r)} | \boldsymbol{\beta}_{-j}^{(r)}, \mathcal{B}_{-r}, \phi_r, \tau, \mathbf{Y}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2\right) \end{aligned} \tag{2.37}$$

(i) for $j = 1, 2, 3$ and $r = 1, \dots, R$ sample independently:

$$p\left(\lambda_{j,r} | \boldsymbol{\beta}_j^{(r)}, \phi_r, \tau\right) \propto \mathcal{G}a\left(a_\lambda + I_j, b_\lambda + \frac{\left\|\boldsymbol{\beta}_j^{(r)}\right\|_1}{\sqrt{\tau \phi_r}}\right). \tag{2.38}$$

(ii) for $p = 1, \dots, I_j$, $j = 1, 2, 3$ and $r = 1, \dots, R$ sample:

$$p\left(w_{j,r,p} | \lambda_{j,r}, \phi_r, \tau, \boldsymbol{\beta}_j^{(r)}\right) \propto GiG\left(\frac{1}{2}, \lambda_{j,r}^2, \frac{\beta_{j,k}^{(r)^2}}{\tau \phi_r}\right) \tag{2.39}$$

(iii) define $\boldsymbol{\beta}_{-j}^{(r)} = \left\{\boldsymbol{\beta}_i^{(r)} : i \neq j\right\}$ and $\mathcal{B}_{-r} = \{B_i : i \neq r\}$, where $B_r = \boldsymbol{\beta}_1^{(r)} \circ \dots \circ \boldsymbol{\beta}_N^{(r)}$. For $r = 1, \dots, R$ sample the PARAFAC marginals from:

$$p\left(\boldsymbol{\beta}_1^{(r)} | \boldsymbol{\beta}_{-1}^{(r)}, \mathcal{B}_{-r}, \phi_r, \tau, \mathbf{Y}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\right) \propto \mathcal{N}_{I_1}(\bar{\boldsymbol{\mu}}_{\beta_1}, \bar{\boldsymbol{\Sigma}}_{\beta_1}) \tag{2.40}$$

$$p\left(\boldsymbol{\beta}_2^{(r)} | \boldsymbol{\beta}_{-2}^{(r)}, \mathcal{B}_{-r}, \phi_r, \tau, \mathbf{Y}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\right) \propto \mathcal{N}_{I_2}(\bar{\boldsymbol{\mu}}_{\beta_2}, \bar{\boldsymbol{\Sigma}}_{\beta_2}) \tag{2.41}$$

$$p\left(\boldsymbol{\beta}_3^{(r)} | \boldsymbol{\beta}_{-3}^{(r)}, \mathcal{B}_{-r}, \phi_r, \tau, \mathbf{Y}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\right) \propto \mathcal{N}_{I_3}(\bar{\boldsymbol{\mu}}_{\beta_3}, \bar{\boldsymbol{\Sigma}}_{\beta_3}) \, . \tag{2.42}$$

III) sample the covariance matrices from their posterior:

$$p(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \gamma | \mathcal{B}, \mathbf{Y}) = p(\boldsymbol{\Sigma}_1 | \mathcal{B}, \mathbf{Y}, \boldsymbol{\Sigma}_2, \gamma) p(\boldsymbol{\Sigma}_2 | \mathcal{B}, \mathbf{Y}, \boldsymbol{\Sigma}_1, \gamma) p(\gamma | \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \tag{2.43}$$

(i) sample the row covariance matrix:

$$p(\boldsymbol{\Sigma}_1 | \mathcal{B}, \mathbf{Y}, \boldsymbol{\Sigma}_2, \gamma) \propto \mathcal{IW}_{I_1}(\nu_1 + I_1, \gamma \boldsymbol{\Psi}_1 + S_1) \tag{2.44}$$

(ii) sample the column covariance matrix:

$$p(\boldsymbol{\Sigma}_2 | \mathcal{B}, \mathbf{Y}, \boldsymbol{\Sigma}_1, \gamma) \propto \mathcal{IW}_{I_2}(\nu_2 + I_2, \gamma \boldsymbol{\Psi}_2 + S_2) \, . \tag{2.45}$$

(iii) sample the scale hyper-parameter:

$$p(\gamma | \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \propto \mathcal{G}a\left(\nu_1 I_1 + \nu_2 I_2, \operatorname{tr}\left(\boldsymbol{\Psi}_1 \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Psi}_2 \boldsymbol{\Sigma}_2^{-1}\right)\right) \, . \tag{2.46}$$

For improving the mixing of the algorithm, it is possible to substitute the draw from the full conditional distribution of the global variance parameter τ or of the PARAFAC marginals with a Hamiltonian Monte Carlo (HMC) step (see Neal (2011)).

2.4 Simulation Results

We report the results of a simulation study where we have tested the performance of the proposed sampler on synthetic datasets of matrix-valued sequences $\{\mathbf{Y}_t, \mathbf{X}_t\}_{t=1}^T$, where $\mathbf{Y}_t, \mathbf{X}_t$ have different size across simulations. The methods described in this paper can be rather computationally intensive, nevertheless thanks to the tensor decomposition we used allows the estimation to be carried out on a laptop. All the simulations were run on an Apple MacBookPro with a 3.1GHz Intel Core i7 processor, RAM 16GB, using MATLAB r2017b with the aid of the Tensor Toolbox v.2.6[6], taking about 30h for a short run of the highest-dimensional case (i.e. $I_1 = I_2 = 50$).

For different sizes ($I_1 = I_2$) of the response and covariate matrices, we generated a matrix-variate time series $\{\mathbf{Y}_t, \mathbf{X}_t\}_{t=1}^T$ by simulating each entry of $\mathbf{X}_t$ from:

$$x_{ij,t} - \mu = \alpha_{ij}(x_{ij,t-1} - \mu) + \eta_{ij,t} \, , \qquad \eta_{ij,t} \sim \mathcal{N}(0, 1) \tag{2.47}$$

[6] Available at: http://www.sandia.gov/ tgkolda/TensorToolbox/index-2.6.html

and a matrix-variate time series $\{\mathbf{Y}_t\}_t$ according to:

$$\mathbf{Y}_t = \mathcal{B} \times_3 \operatorname{vec}(\mathbf{X}_t) + \mathbf{E}_t , \qquad \mathbf{E}_t \sim \mathcal{N}_{I_1,I_2}(\mathbf{0}, \mathbf{\Sigma}_1, \mathbf{I}_{I_2}) . \tag{2.48}$$

where $\mathbb{E}[\eta_{ij,t}\eta_{kl,v}] = 0$, $\mathbb{E}[\eta_{ij,t}E_v] = 0$, $\forall (i,j) \neq (k,l)$, $\forall t \neq v$, and $\alpha_{ij} \sim \mathcal{U}(-1,1)$. We randomly draw $\mathcal{B}$ by using the PARAFAC representation in eq. (A.14), with rank $R = 5$ and marginals sampled from the prior distribution in eq. (2.29e).

The response and covariate matrices in the simulated datasets have the following sizes:

(I) $I_1 = I_2 = I = 10$, for $T = 60$;

(II) $I_1 = I_2 = I = 20$, for $T = 60$;

(III) $I_1 = I_2 = I = 30$, for $T = 60$;

(IV) $I_1 = I_2 = I = 40$, for $T = 60$;

(V) $I_1 = I_2 = I = 50$, for $T = 60$.

We initialized the Gibbs sampler by setting the PARAFAC marginals $\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \boldsymbol{\beta}_3^{(r)}$, $r = 1, \ldots, R$ (with $R = 5$), with the output of a simulated annealing algorithm (see Appendix B.2) and run the algorithm for $N = 10000$ iterations. We present the results for the case $\mathbf{\Sigma}_2 = \mathbf{I}_{I_2}$. Since they are similar, we omit the results for unconstrained $\mathbf{\Sigma}_2$, estimated with the Gibbs in Section 2.3.

Also, we provide a deeper study of the properties of the proposed sampler by presenting in Appendix B.7 the details of the convergence properties of the algorithm in the cases (I)-(II)-(III).

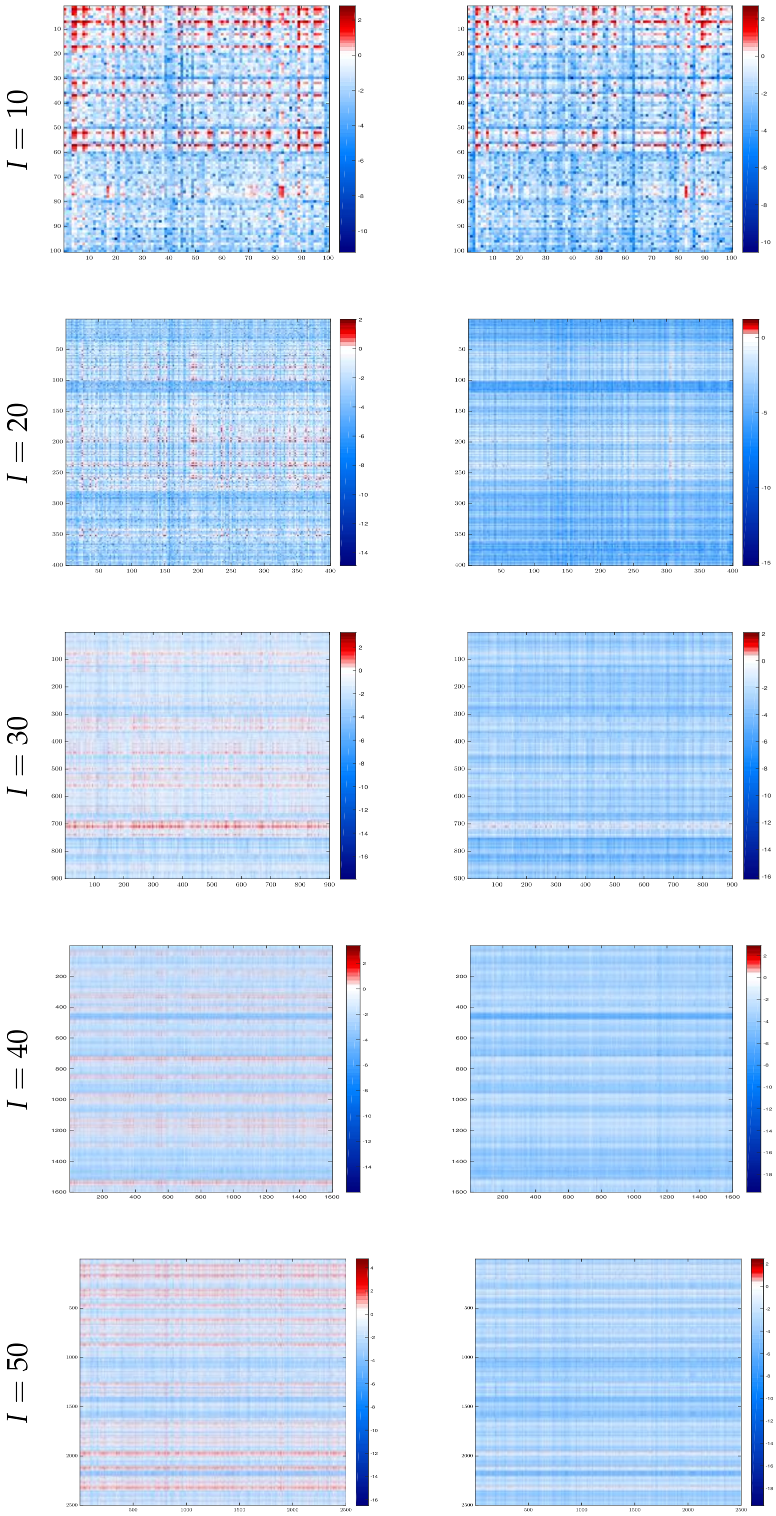


FIGURE 2.5: Logarithm of the absolute value of the coefficient tensors: true $\mathcal{B}$ (*left*) and estimated $\hat{\mathcal{B}}$ (*right*).

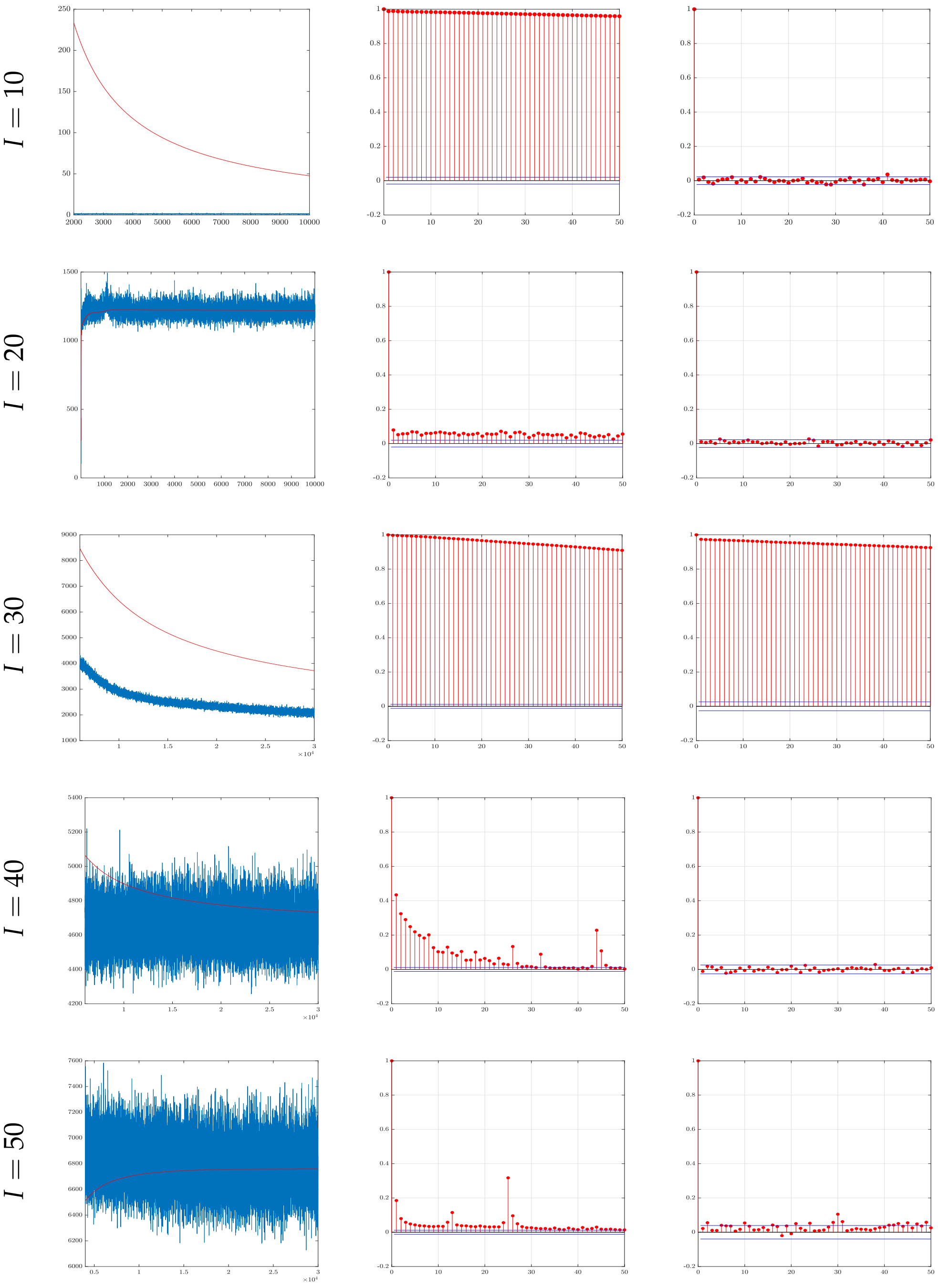


FIGURE 2.6: MCMC output (*left*), autocorrelation function for the entire sample (*middle* plot) and after burn-in (*right* plot) of the Frobenious norm of the difference between the true and the estimated covariance matrix $\mathbf{\Sigma}_1$.

The results are reported in Figs. 2.5-2.6, for the different simulated datasets. Fig. 2.5 shows the good accuracy of the sampler in estimating the coefficient tensor, whose number of entries ranges from 10^4 in the first to 50^4 in the last simulation setting. The estimation error is mainly due to the over-shrinking to zero of large signals. This well-known drawback of global-local hierarchical prior distributions (e.g., see Carvalho et al. (2010)) is related to its sensitivity to the hyper-parameters setting. Fig. 2.6 plots the MCMC output of the Frobenious norm (i.e. the L_2 norm) of the covariance matrix of the error term. After a graphical inspection of the trace plots (first column) we chose a burn-in period of 2000 iterations. Due to autocorrelation in the sample (second column plots) we applied thinning and selected every 10th iteration. In most of the cases, after removing burn-in iterations and performing thinning, the autocorrelation wipes out.

We refer the reader to Appendix B.5 for additional details on the simulation experiments, such as trace plots and autocorrelation functions for tensor entries and individual hyper-parameters.

2.5 Application

As put forward by Schweitzer et al. (2009), the analysis of economic networks is one of the most recent and complex challenges that the econometric community is facing nowadays. We contribute to the econometric literature about complex networks by applying the methodology proposed in Section 2.3 to the study of the temporal evolution of the international trade network (ITN). This economic network has been previously studied by several authors (e.g., see Hidalgo and Hausmann (2009), Fagiolo et al. (2009), Kharrazi et al. (2017), Meyfroidt et al. (2010), Zhu et al. (2014), Squartini et al. (2011)), who have analysed its topological properties and identified its main communities. However, to the best of our knowledge, this is the first attempt to model the temporal evolution of the network as a whole.

The raw trade data come from the United Nations COMTRADE database, a publicly available resource[7]. The particular dataset we use is a subset of the whole COMTRADE database and consists of yearly observations from 1998 to 2016 of total imports and exports between $I_1 = I_2 = I = 10$ countries. In order to remove possible sources of non-linearities in the data, we use a logarithmic transform of the variables of interest. We thus consider the international trade network at each time stamp as one observation from a real-valued matrix-variate stochastic process. Fig. 2.7 shows the whole network sequence in our dataset.

We estimate the model in eq. (2.26) setting $\mathbf{X}_t = \mathbf{Y}_{t-1}$, thus obtaining a matrix-variate autoregressive model. Each matrix $\mathbf{Y}_t$ is the $I \times I$ real-valued weighted adjacency matrix of the corresponding international trade network in year t, whose entry (i, j) contains the total exports of country i vis-à-vis country j, in year t. The series $\{\mathbf{Y}_t\}_t$, $t = 1, \ldots, T$, has been standardized (over the temporal dimension). We run the Gibbs sampler for $N = 10,000$ iterations. The output is reported below.

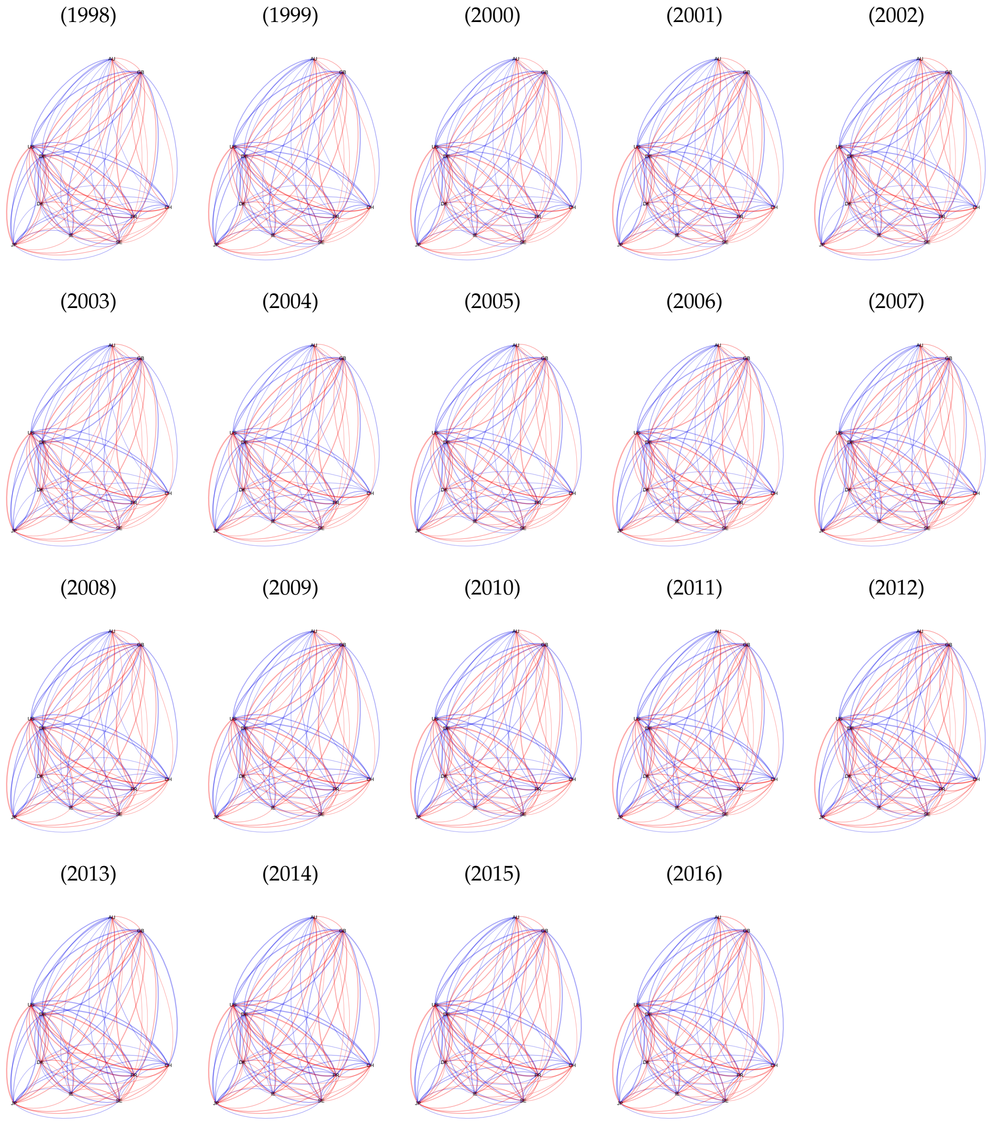
The mod-3 matricization of the estimated coefficient tensor is shown in the left panel of Fig. 2.8, each column corresponds to the effects of a lag one edge (horizontal axis) on all the contemporaneous edges (vertical axis). Positive effects in red and negative effects in blue. Fig. 2.9 shows the estimated covariance matrices of the noise term, that is $\hat{\mathbf{\Sigma}}_1, \hat{\mathbf{\Sigma}}_2$. As regards the estimated coefficient tensor, we find that a significant degree of heterogeneity in the estimated coefficients which points against parameter pooling assumptions. Furthermore, there are patterns showing that groups of edges (i.e. bilateral trade flows) with mainly positive (red) or negative (blue) effect on all the other edges: this may suggest that there are some countries playing a key role (either as exporters of as importers) for them.

The distribution of the entries of the estimated coefficient tensor (middle panel) confirms the evidence of heterogeneity. The distribution is right-skewed and leptokurtic with mode at zero, which is a consequence of the shrinkage of the estimated coefficient.

Moreover, in order to assess the stationarity of the model, we computed the eigenvalues of the mode-3 matricization of the estimated coefficient tensor and the right panel of Fig. 2.8 plots the logarithm of their modulus. All the estimated eigenvalues are strictly lower than one in modulus, thus indicating that the process describing the evolution of the trade network is stationary.

Concerning the estimated covariance matrices of the noise term (Fig. 2.9), we find that in both cases (i.e. $\hat{\mathbf{\Sigma}}_1, \hat{\mathbf{\Sigma}}_2$) the highest estimated values correspond to individual variances, while the estimated covariances are lower in magnitude and heterogeneous. In addition, there is evidence of heterogeneity in the dependence structure, since $\hat{\mathbf{\Sigma}}_1$, which captures the covariance between exporting countries (i.e., rows of the matrix $\mathbf{Y}_t$), differs from $\hat{\mathbf{\Sigma}}_2$,

[7] https://comtrade.un.org



(1998) (1999) (2000) (2001) (2002)

(2003) (2004) (2005) (2006) (2007)

(2008) (2009) (2010) (2011) (2012)

(2013) (2014) (2015) (2016)

FIGURE 2.7: Commercial trade network evolving over time from 1998 (*top left*) to 2016 (*bottom right*). Nodes represent countries, red and blue colored edges stand for exports and imports between two countries, respectively. Edge thickness represents the magnitude of the flow.

which describes the covariance between importing countries (i.e., columns of $\mathbf{Y}_t$) and the dependence between exporting countries is higher, on average, than that between importing countries.

For assessing the convergence of the MCMC chain, Fig. 2.9 shows the trace plot and autocorrelation functions (without thinning) of the Frobenious norm of each estimated matrix. Both sets of plots show a good mixing of the chain.

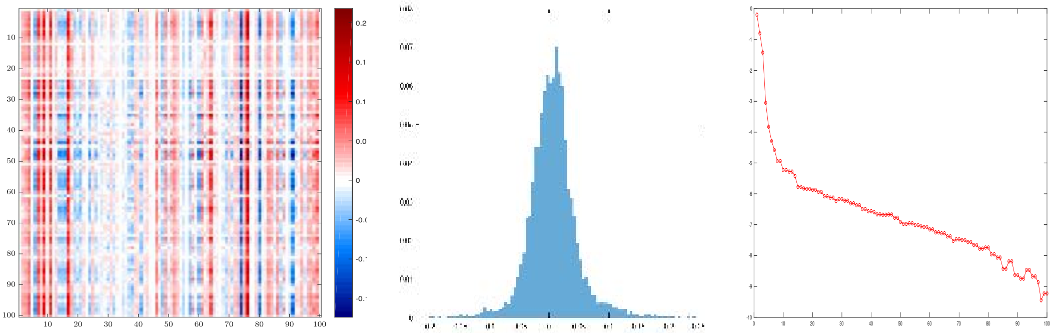


FIGURE 2.8: *Left:* Transpose of the mode-3 matricized estimated coefficient tensor, $\hat{\mathbf{B}}'_{(3)}$. *Middle:* distribution of the estimated entries of $\hat{\mathbf{B}}_{(3)}$. *Right:* logarithm of the modulus of the eigenvalues of $\hat{\mathbf{B}}_{(3)}$, in decreasing order.

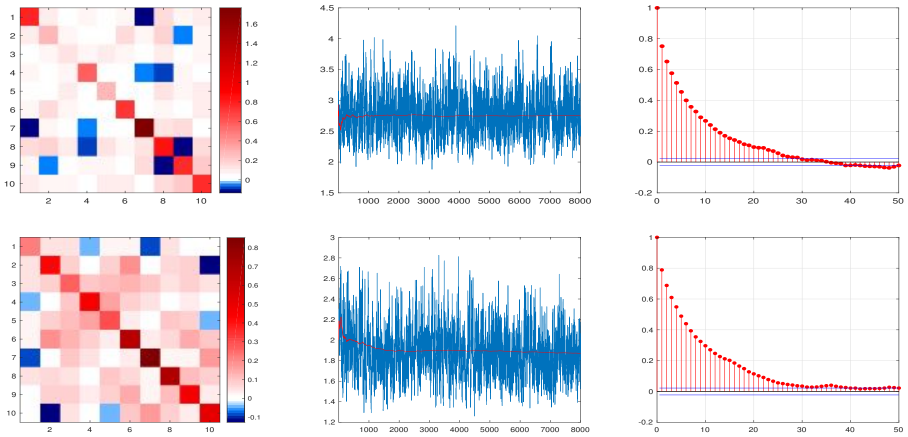


FIGURE 2.9: Estimated covariance matrix of the noise term (*first*), posterior distributions (*second*), MCMC output (*third*) and autocorrelation functions (*fourth*) of the Frobenious norm of the covariance matrix of the noise term. *First row*: $\mathbf{\Sigma}_1$, *second row*: $\mathbf{\Sigma}_2$.

2.5.1 Impulse response analysis

For understanding the role exerted by the various links of the network, Fig. 2.10 shows the sum of the entries of the corresponding positive and negative entries of the estimated coefficient tensor in red and blue, respectively.

We find that edges' impact tend to cluster, that is, those with high positive cumulated effects have very low negative cumulated effects and vice-versa. Thus, the bottom panel of Fig. 2.10 shows the sum of the absolute values of all corresponding entries of the estimated coefficient tensor, which can be interpreted as a measure of the importance of the edge in the network. Based on this statistic, we plot the position of the 10 most and least relevant edges in the network (in red and blue, respectively) in Fig 2.11. The picture has a heterogeneous structure: first, no single country seems to exert a key role, neither as exporter nor as importer; second, the most and least relevant edges are evenly distributed between the exporting and the importing side.

We study the effects of the propagation of a shock on a single and a group of edges in the network by means of the impulse response function obtained as follows. Define the reverse of the vectorization operator $\operatorname{vec}(\cdot)$ by $\operatorname{vecr}(\cdot)$ and let $\tilde{E}$ be a binary matrix of shocks such that each non zero entry (i,j) of $\tilde{E}$ corresponds to a unitary shock on the edge (i,j). Then the matrix-valued impulse response function is obtained from the recursion:

$$Y_1 = \mathcal{B} \times_3 \operatorname{vec}\left(\tilde{E}\right) = \operatorname{vecr}\left(\mathbf{B}'_{(3)} \cdot \operatorname{vec}\left(\tilde{E}\right)\right) \tag{2.49}$$

$$Y_2 = \mathcal{B} \times_3 \operatorname{vec}\left(\operatorname{vecr}\left(\mathbf{B}'_{(3)} \cdot \operatorname{vec}\left(\tilde{E}\right)\right)\right) = \operatorname{vecr}\left(\mathbf{B}'_{(3)} \cdot \mathbf{B}'_{(3)} \cdot \operatorname{vec}\left(\tilde{E}\right)\right) \tag{2.50}$$

$$= \operatorname{vecr}\left([\mathbf{B}'_{(3)}]^2 \cdot \operatorname{vec}\left(\tilde{E}\right)\right) , \tag{2.51}$$

which, for the horizon $h > 0$, generalizes to:

$$Y_h = \operatorname{vecr}\left([\mathbf{B}'_{(3)}]^h \cdot \operatorname{vec}\left(\tilde{E}\right)\right) . \tag{2.52}$$

This equation shows that it is possible to study the joint effect that a contemporaneous shock on a subset of the edges of the network has on the whole network over time.

We define the most relevant edges in the network as those which exert impact on the others and Fig 2.11 shows the locations of the 10 most relevant edges in the network according to different criteria: highest total positive effects, highest total negative effects and highest total net effects.

Fig. 2.12 and 2.13, respectively, plot the impulse response function of a unitary shock on the 10 most relevant and the 10 least relevant edges (determined by ranking according to the sum of the absolute values of the entries of the estimated coefficient tensor), for $h = 1, \ldots, 14$ periods. Figs. 2.14-2.15 show the effects of a unitary shock to the most and least influential edges, respectively.

We find that the effects are remarkably different: both the magnitude and the persistence of the impact of a shock to the most relevant edges is significantly greater than that obtained by hitting the least relevant edges.

Moreover, a shock to the most relevant edge is more persistent than a shock on the least relevant and the magnitude of the effects is significantly higher. However, compared to the effects of a shock on 10 edges, both persistence and magnitude are remarkably lower. Furthermore, a shock to a single edge affects almost all the others because of the high degree of interconnection of the network, which is responsible for the propagation both in the space (i.e., cross-section) and over time.

The joint analysis of the impulse response functions and the distribution of the most and least influential links in Fig. 2.11 points out the key role of the network structure in the propagation of shocks.

We refer to Appendix B.6 for additional plots of the estimation.

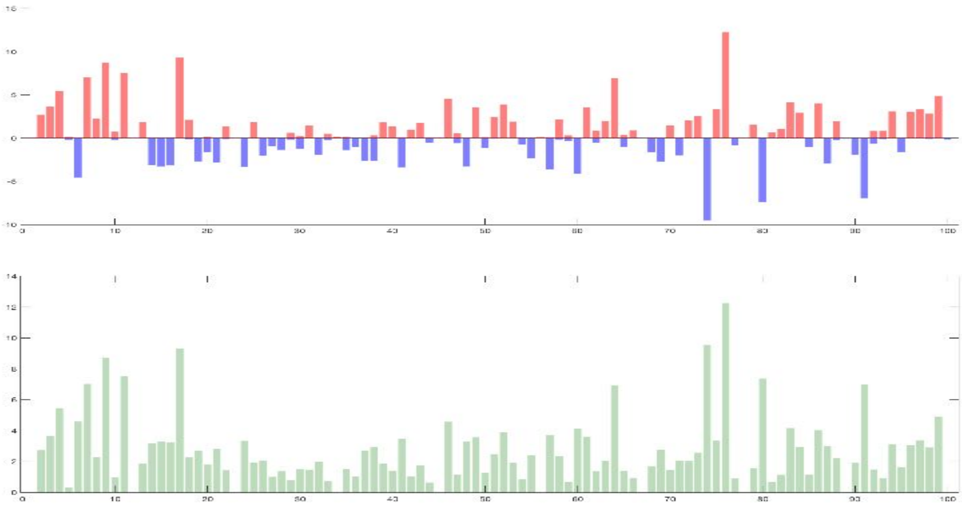


FIGURE 2.10: Sum of positive entries (*red,top*), negative entries (*blue,top*) and of absolute values of all entries (*dark green,bottom*) of the estimated coefficient tensor (*y-axis*), per each edge (*x-axis*).

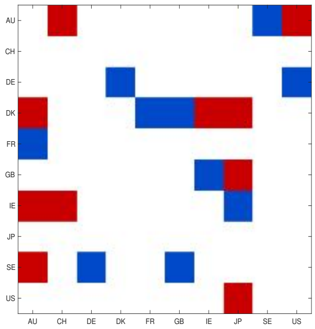


FIGURE 2.11: Position in the network of the 10 most relevant (*red*) and least relevant (*blue*) edges, according to the sum of the absolute values. Countries' labels on both axes.

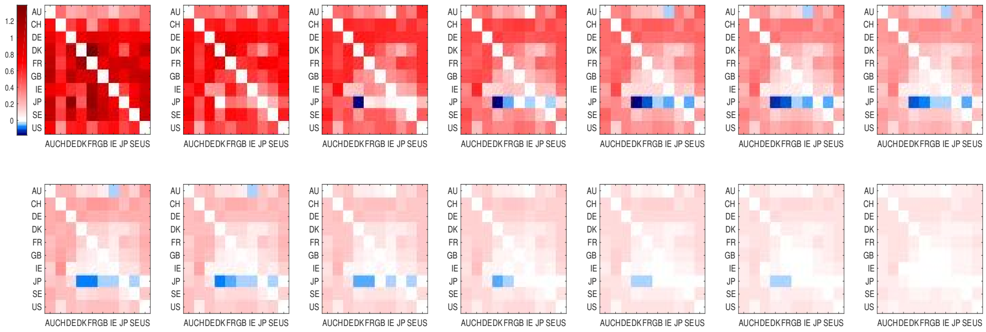


FIGURE 2.12: Impulse response for $h = 1, \ldots, 14$ periods. Unitary shock on the 10 most relevant edges (sum of absolute values of all coefficients). Countries' labels on both axes.

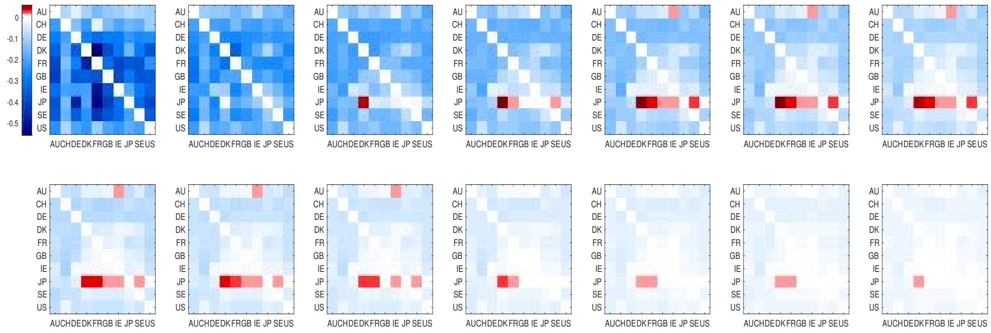


FIGURE 2.13: Impulse response for $h = 1, \ldots, 14$ periods. Unitary shock on the 10 least relevant edges (sum of absolute values of all coefficients). Countries' labels on both axes.

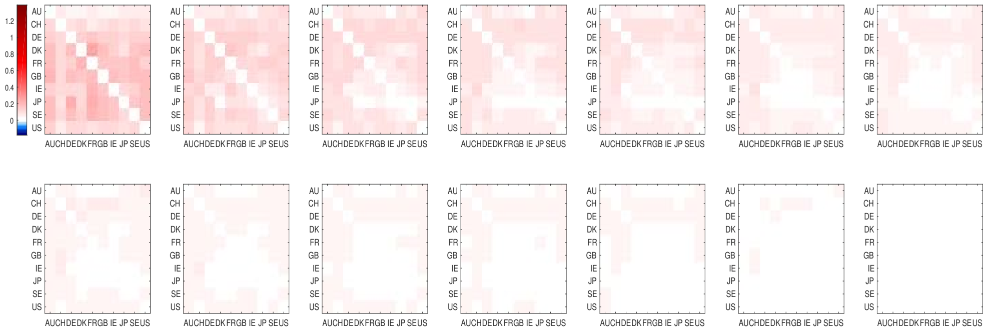


FIGURE 2.14: Impulse response for $h = 1, \ldots, 14$ periods. Unitary shock on the most relevant edge (sum of absolute values of all coefficients). Countries' labels on both axes.

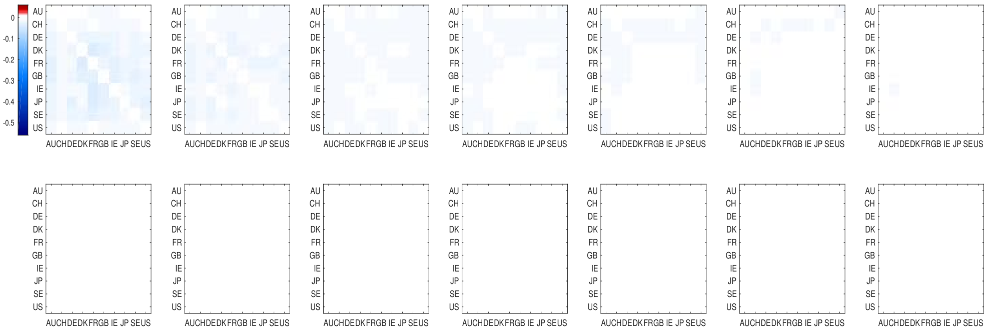


FIGURE 2.15: Impulse response for $h = 1, \ldots, 14$ periods. Unitary shock on the least relevant edge (sum of absolute values of all coefficients). Countries' labels on both axes.

2.6 Conclusions

We defined a new statistical framework for dynamic tensor regression. It is a generalisation of many models frequently used in time series analysis, such as VAR, panel VAR, SUR and matrix regression models. The PARAFAC decomposition of the tensor of regression coefficients allows to reduce the dimension of the parameter space but also permits to choose flexible multivariate prior distributions, instead of multidimensional ones. Overall, this allows to encompass sparsity beliefs and to design efficient algorithm for posterior inference.

We tested the Gibbs sampler algorithm on synthetic matrix-variate datasets with matrices of different sizes, obtaining good results in terms of both the estimation of the true value of the parameter and the efficiency.

The proposed methodology has been applied to the analysis of temporal evolution of a subset of the international trade networks. We found evidence of (i) wide heterogeneity in the sign and magnitude of the estimated coefficients; (ii) stationarity of the network process.

Acknowledgements

This research has benefited from the use of the Scientific Computation System of Ca' Foscari University of Venice (SCSCF) for the computational for the implementation of the estimation procedure.

Chapter 3

Bayesian Markov Switching Tensor Regression for Time-varying Networks

In mathematics the art of proposing a question must be held of higher value than solving it.

GEORG CANTOR

It is far better to foresee even without certainty than not to foresee at all.

HENRI POINCARÉ

3.1 Introduction

The analysis of large sets of binary data is a central issue in many applied fields such as biostatistics (e.g. Schildcrout and Heagerty (2005), Wilbur et al. (2002)), image processing (e.g. Yue et al. (2012)), machine learning (e.g. Banerjee et al. (2008), Koh et al. (2007)), medicine (e.g. Christakis and Fowler (2008)), text analysis (e.g. Taddy (2013), Turney (2002)) and theoretical and applied statistics (e.g. Ravikumar et al. (2010), Sherman et al. (2006), Visaya et al. (2015)). Without loss of generality, in this paper we focus on binary series representing time-evolving networks.

From the outbreak of the financial crisis of 2007 there has been an increasing interest in financial network analysis. The fundamental questions on the role of agents' connections, the dynamic process of link formation and destruction, the diffusion process within the economic and/or financial system of external and internal shocks have attracted an increasing interest from the scientific community (e.g., Billio et al. (2012) and Diebold and Yilmaz (2014)).

Despite the wide economic and financial literature exploiting networks in theoretical models (e.g. Acemoglu et al. (2012), Di Giovanni et al. (2014), Chaney (2014), Mele (2017), Graham (2017)), the econometric analysis of networks and of their dynamical properties is at its infancy and many research questions are still striving for an answer. This paper contributes at filling this gap addressing some important questions in building statistical models for network data.

The first issue concerns measuring the impact of a given set of covariates on the dynamic process of link formation. We propose a parsimonious model that can be successfully used to this aim, building on a novel research domain on tensor calculus in statistics. This new literature (see, e.g. Kolda and Bader (2009), Cichocki et al. (2015) and Cichocki et al. (2016) for a review) proposes a generalisation of matrix calculus to higher dimensional arrays, called tensors. The main advantage in using tensors is the possibility of dealing with the complexity of novel data structures which are becoming increasingly available, such as networks, multi-layer networks, three-way tables, spatial panels with multiple series observed for each unit (e.g., municipalities, regions, countries). The use of tensors prevents the reshaping and manipulation of the data, thus allowing to preserve the intrinsic structure. Another advantage of tensors stems from their numerous decompositions and approximations, which provide a representation of the model in a lower-dimensional space (see (Hackbusch, 2012, ch.7-8)). In this paper we exploit the PARAFAC decomposition for reducing the number of parameters to estimate, thus making inference on network models feasible.

Another issue regards the time stability of the dependence structure between variables. For example, Billio et al. (2012), Billio et al. (2015a), Ahelegbey et al. (2016a), Ahelegbey et al. (2016b) and Bianchi et al. (2018) showed empirically that the network structure of the financial system has experienced a rather long period of stability in the early 2000s and a significantly increasing connectivity before the outbreak of the financial crisis. Starting from these stylized facts, we provide a new Markov switching model for structural changes of the network topology. After the seminal paper of Hamilton (1989), the existing Markov switching models at the core of the Bayesian econometrics literature consider VAR models (e.g., Sims and Zha (2006), Sims et al. (2008)), factor models (e.g., Kaufmann (2000), Kim and Nelson (1998)) or dynamic panels (e.g., Kaufmann (2015), Kaufmann (2010)) and have been extended allowing for stochastic volatility (Smith (2002), Chib et al. (2002)), ARCH and GARCH effects (e.g., see Hamilton and Susmel (1994), Haas et al. (2004), Klaassen (2002) and Dueker (1997), among the others) and stochastic correlation (Casarin et al. (2018)). We contribute to this literature by applying Markov switching dynamics to tensor-valued data.

Motivated by the observation that financial networks are generally sparse, with sudden abrupt changes in the level of sparsity across time, we define a framework which allows us to tackle the issue of time-varying sparsity. To accomplish this task, we compose the proposed Markov switching dynamics with a zero-inflated logit model. In this sense, we contribute to the network literature on modelling edges' probabilities (e.g., Durante and Dunson (2014a) and Wang et al. (2017)), by considering a time series of networks with multiple layers and varying sparsity patterns.

Finally, another relevant question concerns the study of the joint evolution of a network and a set of economic variables of interest. To the best of our knowledge, there is no previous work providing a satisfactory econometric framework to solve this problem. Within the literature on joint modelling discrete and continuous variables Dueker (2005) used the latent variable interpretation of the binary regression and built a VAR model for unobserved continuous-valued variables and quantitative observables. Instead, Taddy (2010) assumes the continuous variable follows a dynamic linear model and the discrete outcome follows a Poisson process with intensity driven by the continuous one. Our contribution to this literature consists in a new joint model for binary tensors and real-valued vectors.

The model we propose is presented in Section 3.2. We go through the details of the Bayesian inferential procedure in Sections 3.3-3.4 while in Section 3.5 we study the performance of the MCMC procedure on synthetic datasets. Finally, we apply the methodology to a real dataset and discuss the results in Section 3.6 and draw the conclusions in Section 3.7.

3.2 A Markov switching model for networks

A relevant object in our modelling framework is a a D-order tensor, that is a D-dimensional array, element of the tensor product of D vector spaces, each one endowed with a coordinate system. See (Hackbusch, 2012, ch.3) for an introduction to algebraic tensor spaces. A tensor can be thought of as the multidimensional extension of a matrix (which is a 2-order tensor), where each dimension is called mode. Other objects of interest are the slice of a tensor, that is a matrix obtained by fixing all but two of the indices of the multidimensional array, and the tube, or fiber, that is a vector resulting from keeping fixed all indices but one. Matrix operations and results from linear algebra can be generalized to tensors (see Hackbusch (2012) or Kroonenberg (2008)). Here we define only the mode-n product between a tensor and a vector and refer the reader to Appendix A.1 for further details. For a D-order tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$ and a vector $\mathbf{v} \in \mathbb{R}^{d_n}$, the mode-$n$ product between them is a $(D-1)$-order tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times \ldots \times d_{n-1} \times d_{n+1} \times \ldots \times d_D}$ whose entries are defined by:

$$\mathcal{Y}_{(i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_D)} = (\mathcal{X} \times_n \mathbf{v})_{(i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_D)} = \sum_{i_n=1}^{d_n} \mathcal{X}_{i_1,\ldots,i_n,\ldots,i_D} \mathbf{v}_{i_n} . \tag{3.1}$$

Let $\{\mathcal{X}_t\}_{t=1}^T$ and $\{\mathcal{X}_t^*\}_{t=1}^T$ be two sequences of binary and real 3-order tensors of size $I \times J \times K$, respectively. In our multilayer network application, $\mathcal{X}_t$ is an adjacency tensor and each of its frontal slices, $\mathbf{X}_{k,t}$, represents the adjacency matrix of k-th layer. See Boccaletti et al. (2014) and Kivelä et al. (2014) for an introduction to multilayer networks. Let $\{\mathbf{y}_t\}_{t=1}^T$ be a sequence of real-valued vectors $\mathbf{y}_t = (y_{t,1}, \ldots, y_{t,M})'$ representing a set of relevant economic or financial indicators. Our model consists of two systems of equations whose parameters switch over time according to a hidden Markov chain process.

The first set of equations pertains the model for the temporal network. One of the most recurrent features of observed networks is edge sparsity, which in random graph theory is defined to be the case in which the number of edges of a graph grows about linearly with the number of nodes (see (Diestel, 2012, ch.7)). For a finite graph size, we consider a network to be sparse when the fraction of edges over the square of nodes, or total degree density, is below 10%. Moreover, the sparsity pattern of many real networks is not time homogeneous. To describe its dynamics we assume that the probability of observing an edge in each layer of the network is a mixture of a Dirac mass at 0 and a Bernoulli distribution, where both the mixing probability and the probability of success are time-varying. Consequently, each entry $x_{ijk,t}$ of the tensor $\mathcal{X}_t$ (that is, each edge of the corresponding network) is distributed as a zero-inflated logit:

$$x_{ijk,t} | \rho(t), \mathbf{g}_{ijk}(t) \sim \rho(t)\delta_{\{0\}}(x_{ijk,t}) + (1-\rho(t))\mathcal{B}ern\left(x_{ijk,t} \middle| \frac{\exp\{\mathbf{z}_{ijk,t}' \mathbf{g}_{ijk}(t)\}}{1+\exp\{\mathbf{z}_{ijk,t}' \mathbf{g}_{ijk}(t)\}} \right) . \tag{3.2}$$

Notice that this model admits an alternative representation as:

$$x_{ijk,t} | \rho(t), \mathbf{g}_{ijk}(t) \sim \rho(t)\delta_{\{0\}}(x_{ijk,t}) + (1-\rho(t))\delta_{\{d_{ijk,t}\}}(x_{ijk,t}) \tag{3.3}$$

$$d_{ijk,t} = \mathbb{1}_{\mathbb{R}_+}(x_{ijk,t}^*) \tag{3.4}$$

$$x_{ijk,t}^* = \mathbf{z}_{ijk,t}' \mathbf{g}_{ijk}(t) + \varepsilon_{ijk,t} \qquad \varepsilon_{ijk,t} \overset{iid}{\sim} \text{Logistic}(0,1) . \tag{3.5}$$

where $\mathbf{z}_{ijk,t} \in \mathbb{R}^Q$ is a vector of edge-specific covariates and $\mathbf{g}_{ijk}(t) \in \mathbb{R}^Q$ is a time-varying edge-specific vector of parameters. This specification allows to classify the zeros (i.e. absence of edge) into "structural" and "random", conditionally on arising from the atomic mass, or due to the randomness described in eqs. (3.4)-(3.5), respectively. The parameter $\rho(t)$ is thus the time-varying probability of observing a structural zero. In the following, without loss of generality, we focus on the case of common set of covariates, that is $\mathbf{z}_{ijk,t} = \mathbf{z}_t$, for $t = 1, \ldots, T$.

The second set of equations regards the vector of economic variables and is given by:

$$y_{m,t} = \mu_{m,t} + \varpi_{m,t} \qquad \varpi_{m,t} \sim \mathcal{N}(0, \sigma_{m,t}^2) , \tag{3.6}$$

for $m = 1, \ldots, M$ and $t = 1, \ldots, T$. In vector form, we denote the mean vector and the covariance matrix by $\boldsymbol{\mu}(t)$ and $\boldsymbol{\Sigma}(t)$, respectively.

The specification of the model is completed with the assumption that the time variation of the parameters $\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t), \rho(t), \mathbf{g}_{ijk}(t)$ are driven by a hidden homogeneous Markov chain

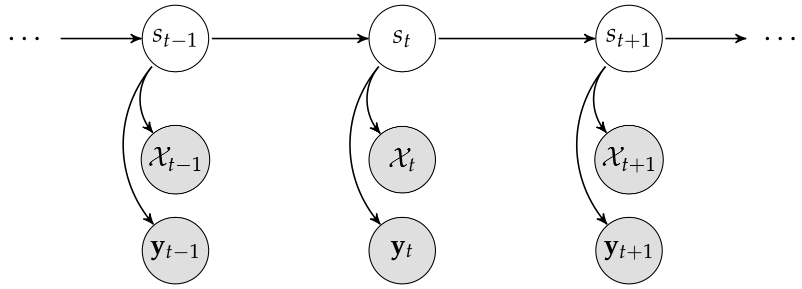


FIGURE 3.1: Directed acyclic graph (DAG) of the model in eq. (3.8a)-(3.8c). Gray circles represent observable variables and white circles latent variables. Directed arrows indicate the direction of causality.

$\{s_t\}_{t=1}^T$ with discrete, finite state space $\{1,\dots,L\}$, that is $\boldsymbol{\mu}(t) = \boldsymbol{\mu}_{s_t}$, $\boldsymbol{\Sigma}(t) = \boldsymbol{\Sigma}_{s_t}$, $\rho(t) = \rho_{s_t}$ and $\mathbf{g}_{ijk}(t) = \mathbf{g}_{ijk,s_t}$. The transition matrix of the chain $\{s_t\}_t$ is assumed to be time-invariant and denoted by $\Xi = (\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_L)'$, where each $\boldsymbol{\xi}_l = (\xi_{l,1},\dots,\xi_{l,L})'$ is a probability vector and the transition probability from state i to state j is $\mathbb{P}(\{s_t = j\}|\{s_{t-1} = i\}) = \xi_{i,j}$, $i,j = 1,\dots,L$.

The causal structure of the model is given in Fig. 3.1, whereas the description of the systems follows.

In order to give a compact representation of the general model, define $\mathbb{X}^d = \{\mathcal{X} \in \mathbb{R}^{i_1\times\ldots\times i_d}\}$ the set of real-valued d-order tensors of size $(i_1 \times \ldots \times i_d)$ and $\mathbb{X}^d_{0,1} = \{\mathcal{X}\in\mathbb{R}^{i_1\times\ldots\times i_d} : \mathcal{X}_{i_1,\dots,i_d} \in \{0,1\}\} \subset \mathbb{X}^d$ the set of adjacency tensors of size $(i_1 \times \ldots \times i_d)$. Define a linear operator between these two sets by $\Psi : \mathbb{X}^d \to \mathbb{X}^d_{0,1}$ such that $\mathcal{X}^* \mapsto \Psi(\mathcal{X}^*) \in \{0,1\}^{i_1\times\ldots\times i_d}$. Denote the indicator function for the set A by $\mathbb{1}_A(x)$, which takes value 1 if $x \in A$ and 0 otherwise, and let $\mathbb{R}_+$ be the positive real half-line. For a matrix $\mathbf{X}^*_{k,t} \in \mathcal{X}^{I,J}$ it is possible to write the first equation of the model in matrix form by $\Psi(\mathbf{X}^*_{k,t}) = (\mathbb{1}_{\mathbb{R}_+}(x^*_{ijk,t}))_{i,j}$, for each slice k of $\mathcal{X}^*_t$. Eq. (3.5) postulates that each edge x_{ijk} admits an individual set of coefficients $\mathbf{g}_{ijk}(t)$. By collecting all these vectors along the indices i, j, k, we can rewrite eq. (3.5) in compact form by means of a fourth-order tensor $\mathcal{G}(t) \in \mathbb{R}^{I\times J\times K\times Q}$, thus obtaining:

$$\mathcal{X}^*_t = \mathcal{G}(t) \times_4 \mathbf{z}_t + \mathcal{E}_t\,, \tag{3.7}$$

where $\mathcal{E}_t \in \mathbb{R}^{I\times J\times K}$ is a third-order tensor with entries $\varepsilon_{ijk,t} \sim \text{Logistic}(0,1)$ and $\times_n$ stands for the mode-n product between a tensor and a vector previously introduced.

The statistical framework we propose for a time series $\{\mathcal{X}_t, \mathbf{y}_t\}_{t=1}^T$ is given by the following system of equations:

$$\begin{cases} \mathcal{X}_t = \mathcal{B}(t) \odot \Psi(\mathcal{X}^*_t) & b_{ijk}(t) \overset{iid}{\sim} \mathcal{B}ern(1-\rho(t)) \quad (3.8a)\\ \mathcal{X}^*_t = \mathcal{G}(t) \times_4 \mathbf{z}_t + \mathcal{E}_t & \quad (3.8b)\\ \mathbf{y}_t = \boldsymbol{\mu}(t) + \boldsymbol{\omega}_t & \boldsymbol{\omega}_t \overset{iid}{\sim} \mathcal{N}_M(\mathbf{0}, \boldsymbol{\Sigma}(t)) \quad (3.8c) \end{cases}$$

where $\mathcal{B}(t)$ is a tensor of the same size of $\mathcal{X}_t$ whose entries are independent and identically distributed (iid) Bernoulli random variables with probability of success $1-\rho(t)$ and $\odot$ stands for the element-by-element Hadamard product (see (Magnus and Neudecker, 1999, ch.3)).

This model can be represented as a SUR (see Zellner (1962)) and also admits an interpretation as a factor model. To this aim, let $\otimes$ denote the Kronecker product (see (Magnus and Neudecker, 1999, ch.3)) and define $\mathbf{z}_t = (1, \tilde{\mathbf{z}}_t)'$, where $\tilde{\mathbf{z}}_t$ denotes the covariates and $\boldsymbol{\Sigma}^{1/2}$ is a matrix satisfying $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$. In addition, let $\{\tilde{\mathbf{u}}_t\}_t$ be a martingale difference process and $\bar{\boldsymbol{\xi}}_t = (\mathbb{1}_{\{1\}}(s_t), \ldots, \mathbb{1}_{\{L\}}(s_t))'$. Then we obtain:

$$\begin{cases} \mathcal{X}_t = \mathcal{B}(t) \odot \Psi(\mathcal{X}_t^*) & b_{ijk}(t) \overset{iid}{\sim} \mathcal{B}ern(1-\rho(t)) \\ \mathcal{X}_t^* = \mathcal{G} \times_4 (\bar{\boldsymbol{\xi}}_t \otimes \mathbf{z}_t)' + \mathcal{E}_t = \mathcal{G} \times_4 (\bar{\boldsymbol{\xi}}_t, \bar{\boldsymbol{\xi}}_t \otimes \tilde{\mathbf{z}}_t)' + \mathcal{E}_t & \varepsilon_{ijk,t} \overset{iid}{\sim} \text{Logistic}(0,1) \\ \mathbf{y}_t = (\bar{\boldsymbol{\xi}}_t \otimes \boldsymbol{\mu}) + (\bar{\boldsymbol{\xi}}_t \otimes \boldsymbol{\Sigma}^{1/2})\boldsymbol{\omega}_t^* & \boldsymbol{\omega}_t^* \overset{iid}{\sim} \mathcal{N}_M(\mathbf{0}_M, \mathbf{I}_M) \\ \bar{\boldsymbol{\xi}}_{t+1} = \Xi \bar{\boldsymbol{\xi}}_t + \tilde{\mathbf{u}}_t & \mathbb{E}[\tilde{\mathbf{u}}_t | \tilde{\mathbf{u}}_{t-1}] = 0 \end{cases} \tag{3.9}$$

3.3 Bayesian Inference

To derive the likelihood function of the model in eqs. (3.8a) to (3.8c) and develop an efficient inferential process, it is useful to start from eq. (3.2), which describes the statistical model for the likelihood of each edge as a zero-inflated logit model. Starting from the seminal work of Lambert (1992), who proposed a modelling framework for count data with a great proportion of zeros, zero-inflated models have been applied to settings where the response variable is not integer-valued. Binary responses have been considered by Harris and Zhao (2007), who dealt with an ordered probit model. This is the closest approach to ours, though the specification in eq. (3.2) substantially differs in two aspects. First, we use of a logistic link function, which is known to have slightly fatter tails than the cumulative normal distribution used in probit models. Second, differently from the majority of the literature which assumes a constant mixing probability, the parameter $\rho(t)$ is evolving according to a latent process.

From eq. (3.2) we derive the probability of observing or not an edge, respectively, as:

$$\mathbb{P}(x_{ijk,t} = 1 | \rho(t), \mathbf{g}_{ijk}(t)) = (1-\rho(t)) \frac{\exp\{\mathbf{z}_t' \mathbf{g}_{ijk}(t)\}}{1+\exp\{\mathbf{z}_t' \mathbf{g}_{ijk}(t)\}} \tag{3.10a}$$

$$\mathbb{P}(x_{ijk,t} = 0 | \rho(t), \mathbf{g}_{ijk}(t)) = \rho(t) + (1-\rho(t)) \left(1 - \frac{\exp\{\mathbf{z}_t' \mathbf{g}_{ijk}(t)\}}{1+\exp\{\mathbf{z}_t' \mathbf{g}_{ijk}(t)\}} \right). \tag{3.10b}$$

This allows us to exploit different types of tensor representations (see Kolda and Bader (2009) for a review), in particular for the sake of parsimony, we assume a PARAFAC decomposition with rank R (assumed fixed and known) for the tensor $\mathcal{G}(t)$:

$$\mathcal{G}(t) = \sum_{r=1}^{R} \gamma_1^{(r)}(t) \circ \gamma_2^{(r)}(t) \circ \gamma_3^{(r)}(t) \circ \gamma_4^{(r)}(t), \tag{3.11}$$

where for each value of the state s_t the vectors $\gamma_h^{(r)}(t) = \gamma_{h,s_t}^{(r)}$, $h \in \{1,2,3,4\}$, $r = 1, \ldots, R$, are the marginals of the PARAFAC decomposition and have length I, J, K and Q, respectively. By the same argument, we denote $\mathcal{G}(t) = \mathcal{G}_{s_t}$ and $\mathbf{g}_{ijk}(t) = \mathbf{g}_{ijk,s_t}$. This specification permits us to achieve two distinct but fundamental goals: (i) parsimony of the model, since for each value of the state s_t the dimension of the parametric space is reduced from $IJKQ$ to $R(I+J+K+Q)$ parameters; (ii) sparsity of the tensor coefficient, through a suitable choice of the prior distribution for the marginals.

We are given a sample $\{\mathcal{X}_t, \mathbf{y}_t\}_{t=1}^T$ and adopt the notation: $\boldsymbol{\mathcal{X}} = \{\mathbf{X}_t\}_{t=1}^T$, $\mathbf{y} = \{\mathbf{y}_t\}_{t=1}^T$, $\mathbf{s} = \{\mathbf{s}_t\}_{t=0}^T$, $\mathcal{D} = \{\mathcal{D}_t\}_{t=1}^T$ and $\boldsymbol{\Omega} = \{\boldsymbol{\Omega}_t\}_{t=1}^T$. Define $\mathcal{T}_l = \{t : s_t = l\}$ and $T_l = \#\mathcal{T}_l$, for each regime $l = 1,2$. Then, in order to write down the analytic form of the complete data likelihood, we introduce the latent variables $\{s_t\}_{t=1}^T$, taking values $s_t = l$, $l \in \{1,2\}$ and evolving according to a discrete Markov chain with transition matrix $\Xi \in \mathbb{R}^{L \times L}$. Finally, denote the whole set of parameters by $\boldsymbol{\theta}$.

The inference is carried out following the Bayesian paradigm and exploiting a data augmentation strategy (Tanner and Wong (1987)). The Pólya-Gamma scheme for models with binomial likelihood proposed by Polson et al. (2013) has been proven to outperform existing schemes for Bayesian inference in logistic regression models in terms of computational speed and higher effective sample size. Furthermore, given a normal prior of the vector of parameters, a Pólya-Gamma prior on latent variables leads to a conjugate posteriors: the full conditional for the parameter vector is normal while that of the latent variable follows a Pólya-Gamma. This allows to use a Gibbs sampler instead of a Metropolis-Hastings algorithm, thus avoiding the need to choose and adequately tune the proposal distribution. Among recent uses of this data augmentation scheme, Wang et al. (2017) used it in a similar framework for network-response regression model, while Holsclaw et al. (2017) exploited it in a time-inhomogeneous hidden Markov model.

The likelihood function is:

$$L(\boldsymbol{\mathcal{X}},\mathbf{y}|\boldsymbol{\theta}) = \sum_{s_1,\ldots,s_T} \prod_{t=1}^{T} p(\mathcal{X}_t,\mathbf{y}_t|s_t,\boldsymbol{\theta})p(s_t|s_{t-1})\,, \tag{3.12}$$

where the index $l \in \{1,\ldots,L\}$ represents the regime. Through the introduction of a latent variables $\mathbf{s} = \{s_t\}_{t=0}^{T}$, we obtain the data augmented likelihood:

$$L(\boldsymbol{\mathcal{X}},\mathbf{y},\mathbf{s}|\boldsymbol{\theta}) = \prod_{t=1}^{T}\prod_{l=1}^{L}\prod_{h=1}^{L}\left[p(\mathcal{X}_t,\mathbf{y}_t|s_t=l,\boldsymbol{\theta})p(s_t=l|s_{t-1}=h,\Xi)\right]^{\mathbb{1}(s_t=l)\mathbb{1}(s_{t-1}=h)}\,. \tag{3.13}$$

The conditional distribution of the observation given the latent variable and marginal distribution of s_t are given by, respectively:

$$p(\mathcal{X}_t,\mathbf{y}_t|s_t=l,\boldsymbol{\theta}) = f_l(\mathcal{X}_t,\mathbf{y}_t|\boldsymbol{\theta}_l) \tag{3.14}$$

$$p(s_t=l|s_{t-1}=h,\Xi) = p_h\,. \tag{3.15}$$

Considering the observation model in eq. (3.2) and defining $\mathcal{T}_l = \{t : s_t = l\}$ for each $l = 1,\ldots,L$, we can rewrite eq. (C.2) as:

$$\begin{aligned}
L(\boldsymbol{\mathcal{X}},\mathbf{y},\mathbf{s}|\boldsymbol{\theta}) &= \prod_{t=1}^{T}\prod_{l=1}^{L}\left[p(\mathcal{X}_t|s_t=l,\boldsymbol{\theta})p(\mathbf{y}_t|s_t=l,\boldsymbol{\theta})\right]^{\mathbb{1}(s_t=l)}\prod_{h=1}^{L}\left[p(s_t=l|s_{t-1}=h,\Xi)\right]^{\mathbb{1}(s_t=l)\mathbb{1}(s_{t-1}=h)} \\
&= \prod_{l=1}^{L}\prod_{t\in\mathcal{T}_l}\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\left[(1-\rho_l)\frac{\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,l}\}}{1+\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,l}\}}\right]^{x_{ijk,t}}\left[\rho_l+(1-\rho_l)\frac{1}{1+\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,l}\}}\right]^{1-x_{ijk,t}} \\
&\cdot\prod_{l=1}^{L}\prod_{t\in\mathcal{T}_l}(2\pi)^{-m/2}\left|\Sigma_l\right|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}(\mathbf{y}_t-\boldsymbol{\mu}_l)\right\} \\
&\cdot\prod_{t=1}^{T}\prod_{l=1}^{L}\prod_{h=1}^{L}p_h^{\mathbb{1}(s_t=l)\mathbb{1}(s_{t-1}=h)}\,.
\end{aligned} \tag{3.16}$$

Since the function cannot be expressed as a series of products due to the sum in the rightmost term, we choose to further augment the data via the through the introduction of latent allocation variables $\mathcal{D} = \{\mathcal{D}_l\}_{l=1}^{L}$, with $\mathcal{D}_l = (d_{ijk,l})$ for $i = 1,\ldots,I$, $j = 1,\ldots,J$ and $k = 1,\ldots,K$. Finally, we perform another augmentation as in Polson et al. (2013), for dealing with the logistic part of the model. When the hidden chain is assumed to be first order Markov, with two possible states, that is $L = 2$, the complete data likelihood is given by:

$$\begin{aligned}
L(\boldsymbol{\mathcal{X}},\mathbf{y},\mathcal{D},\boldsymbol{\Omega},\mathbf{s}|\boldsymbol{\theta}) &= p(\boldsymbol{\mathcal{X}},\mathcal{D},\boldsymbol{\Omega}|\mathbf{s},\boldsymbol{\theta})p(\mathbf{y}|\mathbf{s},\boldsymbol{\theta})p(\mathbf{s}|\boldsymbol{\theta}) \\
&= \prod_{t=1}^{T}p(\mathcal{X}_t,\mathcal{D}_t,\boldsymbol{\Omega}_t|s_t,\boldsymbol{\theta})p(\mathbf{y}_t|s_t,\boldsymbol{\theta})p(s_t|\boldsymbol{\theta})
\end{aligned} \tag{3.17a}$$

$$= \left[\prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} \prod_{i=1}^{I} \prod_{j=i}^{J} \prod_{k=1}^{K} \underbrace{p(x_{ijk,t}, d_{ijk,t}, \omega_{ijk,t} | s_t = l, \rho_l, \mathcal{G}_l)}_{I} \right] \tag{3.17b}$$

$$\cdot \left[\prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} \underbrace{p(\mathbf{y}_t | s_t = l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}_{II} \right] \cdot \left[\underbrace{p(\mathbf{s} | \Xi)}_{III} \right] \tag{3.17c}$$

where we have exploited the conditional independence of $\boldsymbol{\mathcal{X}}$ and $\mathbf{y}$ given the hidden chain $\mathbf{s}$. We augment the model by introducing the latent allocation $d_{ijk,l} \in \{0, 1\}$ for $l = 1, \ldots, L$. Secondly, we use a further data augmentation step via the introduction of the latent variables $\omega_{ijk,t}$ following Polson et al. (2013), for dealing with the logistic part of the mixture.

The complete data likelihood for $\boldsymbol{\mathcal{X}}$ is given by:

$$L(\boldsymbol{\mathcal{X}}, \mathbf{y}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s} | \boldsymbol{\theta}) =$$

$$= \prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \rho_l^{d_{ijk,t}} \cdot \delta_{\{0\}}(x_{ijk,t})^{d_{ijk,t}} \cdot \left(\frac{1-\rho_l}{2} \right)^{1-d_{ijk,t}} \cdot \exp\left\{ -\frac{\omega_{ijk,t}}{2} (\mathbf{z}_t' \mathbf{g}_{ijk,l})^2 + \kappa_{ijk,t} (\mathbf{z}_t' \mathbf{g}_{ijk,l}) \right\}$$

$$\cdot \prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} (2\pi)^{-m/2} |\boldsymbol{\Sigma}_l|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y}_t - \boldsymbol{\mu}_l)' \boldsymbol{\Sigma}_l^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_l) \right\}$$

$$\cdot \left[\prod_{t=1}^{T} \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} p(\omega_{ijk,t}) \right] \cdot \left[\prod_{g=1}^{L} \prod_{l=1}^{L} \xi_{g,l}^{N_{gl}(\mathbf{s})} \right] \cdot p(s_0 | \Xi), \tag{3.18}$$

where $d_{ij,t}$ is a latent allocation variable and $\omega_{ij,t}$ is a Pólya-Gamma latent variable. See Appendix C.2 for the details of the data augmentation strategy and the derivation of the complete data likelihood.

A well-known identification issue for mixture models is the label switching problem (see, for example, Celeux (1998)), which stems from the fact that the likelihood function is invariant to relabeling of the mixture components. This may represent a problem for Bayesian inference, especially when the unobserved components are not well separated, since the associated labels may wrongly change across iterations. Several proposals have been made for solving this identification issue (see Frühwirth-Schnatter (2006) for a review). The permutation sampler proposed by Frühwirth-Schnatter (2001) can be applied under the assumption of exchangeability of the posterior distribution, which is satisfied when the prior distribution for the transition probabilities of the hidden Markov chain is symmetric. Alternatively, there are situations when the particular application provides meaningful restrictions on the value of some parameters. These restrictions generally stem from theoretical results, or interpretation of the different regimes, which is the reason why they are widely used in macroeconomics and finance.

Following this second approach, we can use as identification constraint for the regimes the mixing probability of the zero-inflated logit in eq. (3.3). This can be interpreted as the likelihood of a "structural" absent edge, therefore by sorting the regimes in decreasing order, from "sparse" to "dense", we impose: $\rho_1 > \rho_2 > \ldots > \rho_L$.

As regards the prior distributions for the parameters of interest, we choose the following specifications. Denote $\boldsymbol{\iota}_n$ the n-dimensional vector of ones. We assume an independent prior on $\gamma_{h,l}^{(r)}$ for each regime $l = 1, \ldots, L$, thus representing the *a priori* ignorance of the different value of these parameters for varying l. In particular, for each $r = 1, \ldots, R$, each $h = 1, \ldots, 4$ and each $l = 1, \ldots, L$ we specify the global-local shrinkage prior:

$$\pi(\gamma_{h,l}^{(r)} | \overline{\zeta}_{h,l}^{r}, \tau, \phi_r, w_{h,r,l}) \sim \mathcal{N}_{n_h}(\overline{\zeta}_{h,l}^{r},\ \tau \phi_r w_{h,r,l} \mathbf{I}_{n_h}) \tag{3.19}$$

where $\mathbf{n} = (I, J, Q)'$ is a vector containing the length of each vector $\gamma_{h,l}^{(r)}$ and the prior mean is set to $\overline{\zeta}_{h,l}^{r} = 0$ for each $h = 1, \dots, 4$, $l = 1, \dots, L$, $r = 1, \dots, R$. The parameter τ represents the global component of the variance, common to all marginals, ϕ_r is level component (specific for each $r = 1, \dots, R$) and $w_{h,r}$ is the local component. The choice of a global-local shrinkage prior, as opposed to a spike-and-slab distribution, is motivated by the reduced computational complexity and the capacity to handle high-dimensional settings.

In addition, for allowing greater flexibility, we assume the following hyper-priors for the variance components[1]:

$$\pi(\tau) \sim \mathcal{G}a(\overline{a}^{\tau}, \overline{b}^{\tau}) \tag{3.20}$$

$$\pi(\boldsymbol{\phi}) \sim \mathcal{D}ir(\overline{\boldsymbol{\alpha}}) \qquad \overline{\boldsymbol{\alpha}} = \overline{\alpha}\boldsymbol{\iota}_R \tag{3.21}$$

$$\pi(w_{h,r,l}|\lambda_l) \sim \mathcal{E}xp(\lambda_l^2/2) \qquad \forall\, h, r, l \tag{3.22}$$

$$\pi(\lambda_l) \sim \mathcal{G}a(\overline{a}_l^{\lambda}, \overline{b}_l^{\lambda}) \qquad \forall\, l\,. \tag{3.23}$$

The further level of hierarchy for the local components $w_{h,r,l}$ is added with the aim of favouring information sharing across local components of the variance (indices h, r) within a given regime l. This hierarchical prior induces the following marginal prior on the vector $\mathbf{w}_l = (w_{1,1,l}, \dots, w_{4,R,l})'$:

$$\begin{aligned}
\pi(\mathbf{w}_l) &= \int_{\mathbb{R}_+} \prod_{r=1}^{R} \prod_{h=1}^{4} \pi(w_{h,r,l}|\lambda_l)\pi(\lambda_l)\, \mathrm{d}\lambda_l \\
&= \int_{\mathbb{R}_+} \frac{(\overline{b}_l^{\lambda})^{\overline{a}_l^{\lambda}}}{2\Gamma(\overline{a}_l^{\lambda})} \lambda_l^{\overline{a}_l^{\lambda}+8R-1} \exp\left\{ -\overline{b}_l^{\lambda}\lambda_l - \left(\sum_{r=1}^{R} \sum_{h=1}^{4} w_{h,r,l} \right) \frac{\lambda_l^2}{2} \right\} \mathrm{d}\lambda_l\,.
\end{aligned} \tag{3.24}$$

The marginal prior for a generic entry $w_{h,r,l}$ is a compound gamma distribution[2], that is $p(w_{h,r,l}) \sim \mathrm{CoGa}(1, \overline{a}_l^{\lambda}, 1, \overline{b}_l^{\lambda})$, with $\overline{a}_l^{\lambda} > -1$. In the univariate case (i.e $H = 1$, $R = 1$ and $L = 1$), we obtain a generalized Pareto distribution[3] $\pi(w) = gP(0, a_{\lambda}, b_{\lambda}/a_{\lambda})$.

The specification of an exponential distribution for the local component of the variance of

[1] We use the shape-rate formulation for the gamma distribution, that is for $\alpha > 0, \beta > 0$:

$$x \sim \mathcal{G}a(\alpha, \beta) \iff f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \in (0, +\infty)\,.$$

[2] Alternatively, following (Johnson et al., 1995, p.248), this is called generalized beta prime distribution or generalized beta distribution of the second kind $\mathcal{B}e_2(\alpha, \beta, p, q)$, whose probability density function (with $B(\alpha, \beta)$ being the usual beta function) is given by:

$$p(x|\alpha, \beta, p, q) = \frac{1}{qB(\alpha, \beta)} p \left(\frac{x}{q}\right)^{\alpha p - 1} \left[1 + \left(\frac{x}{q}\right)^{p} \right]^{-(\alpha+\beta)} \qquad x \in \mathbb{R}_+,\ \alpha, \beta, p, q \in \mathbb{R}_+\,. \tag{3.25}$$

In our case, the probability density function is defined by a mixture of two gamma distributions (see also Dubey (1970)):

$$p(x|\alpha, \beta, 1, q) = \int_0^{\infty} \mathcal{G}a(x|\alpha, p)\mathcal{G}a(p|\beta, q)\, \mathrm{d}p = \frac{q^{\beta} x^{\alpha-1}(q+x)^{\alpha+\beta}}{B(\alpha, \beta)} \qquad x \in \mathbb{R}_+,\ \alpha, \beta, q \in \mathbb{R}_+\,. \tag{3.26}$$

In our case, the parametrisation is $(1, a_{\lambda}, 1, b_{\lambda})$. This special case is also called a Lomax(a, b) distribution with parameters $(a_{\lambda}, b_{\lambda})$.

[3] The probability density function of the generalized Pareto distribution is:

$$p(x|\mu, \xi, \sigma) = \frac{1}{\sigma} \left(1 + \frac{(x-\mu)}{\xi\sigma} \right)^{-(\xi+1)} \qquad x \in \mathbb{R}_+,\ \mu, \xi \in \mathbb{R},\ \sigma \in \mathbb{R}_+\,. \tag{3.27}$$

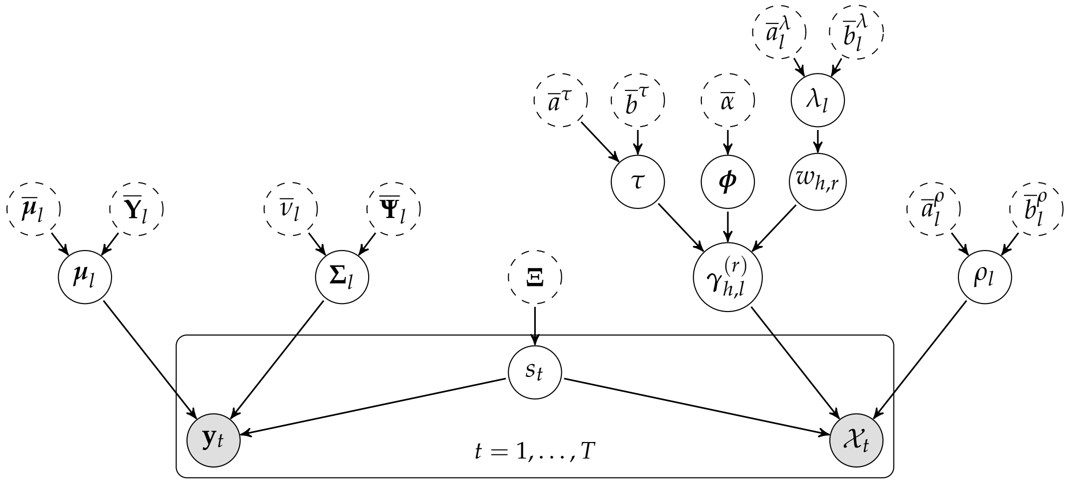


FIGURE 3.2: DAG of the model in eq. (3.8a)-(3.8c) and prior structure in eq. (3.19)-(3.31). Gray circles denote observable variables, white circles with continuous border indicate parameters, white circles with dashed border indicate fixed hyper-parameters.

the $\gamma_{h,l}^{(r)}$ yields a Laplace distribution for each component of the vectors once the $w_{h,r,l}$ is integrated out, that is $\gamma_{h,l,i}^{(r)} | \lambda_l, \tau, \phi_r \sim \text{Laplace}(\mathbf{0}, \lambda_l / \sqrt{\tau \phi_r})$ for all $i = 1, \dots, n_h$. The marginal distribution of each entry, integrating all remaining random components, is instead a generalized Pareto distribution.

In probit and logit models it is not possible to identify the coefficients of the latent regression equation as well as the variance of the noise (e.g., see Wooldridge (2010)). As a consequence, we make the usual identifying restriction by imposing unitary variance for each $\epsilon_{ijk,t}$.

The mixing probability of the observation model is assumed beta distributed:

$$\pi(\rho_l) \sim \mathcal{B}e(\bar{a}_l^{\rho}, \bar{b}_l^{\rho}) \qquad \forall\, l = 1, \dots, L\,. \tag{3.28}$$

Concerning the parameters of the second equation (vector $\mathbf{y}_t \in \mathbb{R}^m$), we assume the priors:

$$\pi(\boldsymbol{\mu}_l) \sim \mathcal{N}_M(\bar{\boldsymbol{\mu}}_l, \bar{\boldsymbol{\Upsilon}}_l) \quad \forall\, l = 1, \dots, L \tag{3.29}$$

$$\pi(\boldsymbol{\Sigma}_l) \sim \mathcal{IW}_M(\bar{\nu}_l, \bar{\boldsymbol{\Psi}}_l) \quad \forall\, l = 1, \dots, L\,. \tag{3.30}$$

Finally, each row of the transition matrix of the Markov chain process $\mathbf{s}_t$ is assumed to be *a priori* distributed according to a Dirichlet distribution:

$$\pi(\boldsymbol{\xi}_{l,:}) \sim \mathcal{D}ir(\bar{\mathbf{c}}_{l,:}) \quad \forall\, l = 1, \dots, L\,. \tag{3.31}$$

The overall structure of the hierarchical prior distribution is represented graphically by means of the directed acyclic graph (DAG) in Fig. 3.2.

3.4 Posterior Approximation

For explanatory purposes, in this section we focus on single layer graphs (i.e. $k = 1$), which is a special case of the model in eqs. (3.8a)-(3.8c). In Appendix C.3 we present the computational details for the general case with multi-layer network observations (i.e. $K > 1$).

For reducing the burden of the notation, we define $\mathcal{G} = \{\mathcal{G}_l\}_{l=1}^{L}$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_l\}_{l=1}^{L}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_l\}_{l=1}^{L}$ and $\boldsymbol{\rho} = \{\rho_l\}_{l=1}^{L}$. Moreover, denote by $\mathbf{W} \in \mathbb{R}^{3 \times R \times L}$ the matrix whose elements $(w_{h,r_l})_{h,r,l}$ are the components of the marginal-specific variance. Combining the complete data likelihood with the prior distributions yields a posterior sampling scheme consisting of four blocks (see Appendix C.3 for the derivation of the posterior full conditional distributions).

In the first block (I) the sampler draws the latent variables from the full conditional distribution:

$$p(\mathbf{s},\mathcal{D},\mathbf{\Omega}|\boldsymbol{\mathcal{X}},\mathbf{y},\mathcal{G},\boldsymbol{\mu},\mathbf{\Sigma},\Xi,\boldsymbol{\rho}) = p(\mathbf{s}|\boldsymbol{\mathcal{X}},\mathbf{y},\mathcal{G},\boldsymbol{\mu},\mathbf{\Sigma},\Xi,\boldsymbol{\rho}) \tag{3.32}$$

$$\cdot \prod_{ijt} p(\omega_{ij,t}|x_{ij,t},s_t,\mathcal{G}_{s_t}) p(d_{ij,t}|x_{ij,t},s_t,\mathcal{G}_{s_t},\rho_{s_t}) \,. \tag{3.33}$$

Samples of $\mathbf{s}$ are obtained via the multi-move Forward-Filtering-Backward-Sampler (see Frühwirth-Schnatter (2006)). The latent variables $\omega_{ij,t}$ are sampled independently for each $i = 1,\ldots,I$, $j = 1,\ldots,J$ and $t = 1,\ldots,T$ from:

$$p(\omega_{ij,t}|x_{ij,t},s_t,\mathcal{G}_{s_t}) \propto PG(1,\mathbf{z}_t'\mathbf{g}_{ijk,s_t}) \,, \tag{3.34}$$

The latent variables $\omega_{ij,t}$ are sampled in block for each t. This is done by sampling $\mathbf{u}_t = \text{vec}\,(\mathbf{\Omega}_t)$ from the vectorised version of the PG random number generator, then reshaping $\mathbf{\Omega}_t = \text{vecr}(\mathbf{u}_t)$. The latent variables $d_{ij,t}$ are sampled independently for each $i = 1,\ldots,I$, $j = 1,\ldots,J$ and $t = 1,\ldots,T$ from:

$$p(d_{ij,t} = 1|x_{ij,t},s_t,\mathcal{G}_{s_t},\rho_{s_t}) \propto \rho_{s_t}\delta_{\{0\}}(x_{ij,t}) \tag{3.35a}$$

$$p(d_{ij,t} = 0|x_{ij,t},s_t,\mathcal{G}_{s_t},\rho_{s_t}) \propto (1-\rho_{s_t})\frac{\exp\{(\mathbf{z}_t'\mathbf{g}_{ijk,s_t})x_{ij,t}\}}{1+\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,s_t}\}} \,. \tag{3.35b}$$

Block (II) regards the hyper-parameters which control the variance of the PARAFAC marginals, and have full conditional distribution:

$$p(\tau,\boldsymbol{\phi},\mathbf{W}|\{\gamma_{h,l}^{(r)}\}_{h,l,r}) = p(\boldsymbol{\phi}|\{\gamma_{h,l}^{(r)}\}_{h,l,r},\mathbf{W})p(\tau|\{\gamma_{h,l}^{(r)}\}_{h,l,r},\mathbf{W},\boldsymbol{\phi})p(\mathbf{W}|\{\gamma_{h,l}^{(r)}\}_{h,l,r},\boldsymbol{\phi},\tau) \,. \tag{3.36}$$

The auxiliary variables ψ_r are sampled independently for $r = 1,\ldots,R$ from:

$$p(\psi_r|\{\gamma_{h,1}^{(r)}\}_{h,l},\mathbf{w}_r) \propto \text{GiG}\left(2\bar{b}^{\tau}, \sum_{h=1}^{3}\sum_{l=1}^{L}\frac{\gamma_{h,l}^{(r)'}\gamma_{h,l}^{(r)}}{w_{h,r}}, \bar{\alpha}-n\right) \tag{3.37}$$

then, for each $r = 1,\ldots,R$ define:

$$\phi_r = \frac{\psi_r}{\sum_{v=1}^{R}\psi_v} \,. \tag{3.38}$$

The parameters $\boldsymbol{\phi}$ are sampled in a separate block since they all enter the full conditionals of $w_{h,r,l}$ and $\gamma_{h,l}^{(r)}$. The global variance parameter τ is drawn from:

$$p(\tau|\{\gamma_{h,l}^{(r)}\}_{h,l,r},\mathbf{W},\boldsymbol{\phi}) \propto \text{GiG}\left(2\bar{b}^{\tau}, \sum_{r=1}^{R}\sum_{h=1}^{3}\sum_{l=1}^{L}\frac{\gamma_{h,l}^{(r)'}\gamma_{h,l}^{(r)}}{\phi_r w_{h,r}}, (\bar{\alpha}-n)R\right) \,. \tag{3.39}$$

The local variance parameters $w_{h,r,l}$ are independently drawn for each $h = 1,2,3$, $r = 1,\ldots,R$ and $l = 1,\ldots,L$ from:

$$p(w_{h,r,l}|\gamma_{h,l}^{(r)},\phi_r,\tau,\lambda_l) \propto \text{GiG}\left(\lambda_l^2, \frac{\gamma_{h,l}^{(r)'}\gamma_{h,l}^{(r)}}{\tau\phi_r}, 1-\frac{n_h}{2}\right) \,. \tag{3.40}$$

Finally, denoting $\mathbf{w}_l$ the collection of all $w_{h,r,l}$ for a given l, the parameters λ_l are independently drawn for each $l = 1, \dots, L$ from:

$$p(\lambda_l|\mathbf{w}_l) \propto \lambda_l^{\bar{a}_l^{\lambda}+6R-1} \cdot \exp\left\{-\lambda_l \bar{b}_l^{\lambda}\right\} \cdot \left\{-\frac{\lambda_l^2}{2} \sum_{r=1}^{R} \sum_{h=1}^{3} w_{h,r,l}\right\} . \tag{3.41}$$

The third block (III) concerns the marginals of the PARAFAC decomposition for the tensors $\mathcal{G}_l$ for every $l = 1, \dots, L$. The vectors $\gamma_{h,l}^{(r)}$ are sampled independently for all $h = 1, 2, 3$ and every $r = 1, \dots, R$ from:

$$p(\gamma_{h,l}^{(r)}|\mathcal{X}, \mathbf{W}, \boldsymbol{\phi}, \tau, \mathbf{s}, \mathcal{D}, \boldsymbol{\Omega}) \propto \mathcal{N}_{n_h}\left(\tilde{\boldsymbol{\zeta}}_{h,l}^{r}, \tilde{\boldsymbol{\Lambda}}_{h,l}^{r}\right) . \tag{3.42}$$

Finally, in block (IV) are drawn the mixing probability, the transition matrix and the main parameters of the second equation. The mixing probability is sampled for every $l = 1, \dots, L$ from:

$$p(\rho_l|\mathcal{D}, \mathbf{s}) \propto \mathcal{B}e(\tilde{a}_l, \tilde{b}_l) . \tag{3.43}$$

Each row of the transition matrix is independently drawn for every $l = 1, \dots, L$ from:

$$p(\boldsymbol{\xi}_{l,:}) \propto \mathcal{D}ir(\tilde{\mathbf{c}}) . \tag{3.44}$$

The mean and covariance matrix of the second equation are sampled independently for every $l = 1, \dots, L$, respectively from:

$$p(\boldsymbol{\mu}_l|\mathbf{y}, \mathbf{s}, \boldsymbol{\Sigma}_l) \propto \mathcal{N}_M(\tilde{\boldsymbol{\mu}}_l, \tilde{\boldsymbol{\Upsilon}}_l) \tag{3.45}$$

and:

$$p(\boldsymbol{\Sigma}_l|\mathbf{y}, \mathbf{s}, \boldsymbol{\mu}_l) \propto \mathcal{IW}_M(\tilde{\nu}_l, \tilde{\boldsymbol{\Psi}}_l) . \tag{3.46}$$

Blocks (I) and (II) are Rao-Blackwellized Gibbs steps: in block (I) we have marginalised over both $(\mathcal{D}, \boldsymbol{\Omega})$ in the full joint conditional distribution of the state $\mathbf{s}$ and $\mathcal{D}$ (together with $\boldsymbol{\rho}$) in the full conditional of $\boldsymbol{\Omega}$, while in (II) we have integrated out τ from the full conditional of $\boldsymbol{\phi}$ (see sec. C.3.2). Blocks (III) and (IV) are standard Gibbs steps, concerned with sampling from the full conditional (eventually exploiting conditional independence relations).

3.5 Simulation Results

We consider three simulation settings for the model in eqs. (3.8a)-(3.8c) corresponding to different sizes I and J, with $I = J$, of the adjacency matrix $\mathbf{X}_t$. The other parameters indicated below are kept fixed across settings. The three synthetic datasets used to check the efficiency of the proposed Gibbs sampler share the same hyper-parameters' values, but differ in the size of the adjacency matrices. We consider:

(I) $I = J = 100$, with $Q = 3$ common covariates and $M = 2$ exogenous variables;

(II) $I = J = 150$, with $Q = 3$ common covariates and $M = 2$ exogenous variables;

(III) $I = J = 200$, with $Q = 3$ common covariates and $M = 2$ exogenous variables.

We generated a sample of size $T = 60$ and at each time step we simulate a square matrix $\mathbf{X}_t$, a vector $\mathbf{y}_t$ of length $m = 2$ and a set of $Q = 3$ covariates $\mathbf{z}_t$. The covariates have been generated from a stationary Gaussian VAR(1) process with entries of the coefficient matrix i.i.d. from a truncated standard normal distribution. We considered two regimes (i.e. $L = 2$)

and generated the trajectory of the Markov chain $\{s_t\}_{t=1}^T$ setting:

$$\Xi = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix} \qquad p(s_0) = \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix} . \tag{3.47}$$

For each regime $l = 1, 2$, we generated the marginals of the PARAFAC decomposition (rank $R = 5$) of the tensor $\mathcal{G}_l$, the mixing probability in the first equation and the mean and covariance in the second equation of the model according to:

$$\begin{aligned} &\rho_1 = 0.8 && \rho_2 = 0.2 \\ &\gamma_{h,1}^{(r)} \overset{iid}{\sim} \mathcal{N}_{n_h}(\mathbf{0}_{n_h}, \mathbf{I}_{n_h}) \;\; \forall h, r && \gamma_{h,2}^{(r)} \overset{iid}{\sim} \mathcal{N}_{n_h}(\boldsymbol{\iota}_{n_h}, \mathbf{I}_{n_h}) \;\; \forall h, r \\ &\boldsymbol{\mu}_1 = [2, 2]' && \boldsymbol{\mu}_2 = [-2, -2]' \\ &\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix} && \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \end{aligned} \tag{3.48}$$

We initialised the marginals of the PARAFAC decomposition of the tensor of coefficients $\mathcal{G}_l$ at the output of the Simulated Annealing algorithm (see (Robert and Casella, 2004, pp. 163-173)) and we kept the same value for each $l = 1, \ldots, L$. The other parameters $(\boldsymbol{\rho}, \mathbf{W}, \boldsymbol{\phi}, \tau, \Xi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ have been initialised by sampling from their prior. Finally, we have chosen the following values for the hyper-parameters:

$$\begin{array}{llll} \bar{\alpha} = 0.5 & \bar{b}_\tau = 2 & \bar{\lambda} = 4 & \underline{\zeta}_{h,l}^r = \mathbf{0}_{n_h} \;\; \forall h, l, r \\ \bar{a}_1 = 5 & \bar{b}_1 = 2 & \bar{a}_2 = 2 & \bar{b}_2 = 5 \\ \bar{\boldsymbol{\mu}}_1 = \mathbf{0}_m & \bar{\boldsymbol{\mu}}_2 = \mathbf{0}_m & \bar{\mathbf{Y}}_1 = \mathbf{I}_m & \bar{\mathbf{Y}}_2 = \mathbf{I}_m \\ \bar{\nu}_1 = m & \bar{\nu}_2 = m & \bar{\boldsymbol{\Psi}}_1 = \mathbf{I}_m & \bar{\boldsymbol{\Psi}}_2 = \mathbf{I}_m \\ \bar{\mathbf{c}}_1 = [8, 4] & \bar{\mathbf{c}}_2 = [4, 8] & & \end{array} \tag{3.49}$$

For each simulation setting, we evaluate the mean square error of the estimated coefficient tensor:

$$MSE = \frac{1}{2}(MSE_1 + MSE_2) = \frac{1}{2IJK} \left(\left\| \mathcal{G}_1^* - \hat{\mathcal{G}}_1 \right\|_2^2 + \left\| \mathcal{G}_2^* - \hat{\mathcal{G}}_2 \right\|_2^2 \right) , \tag{3.50}$$

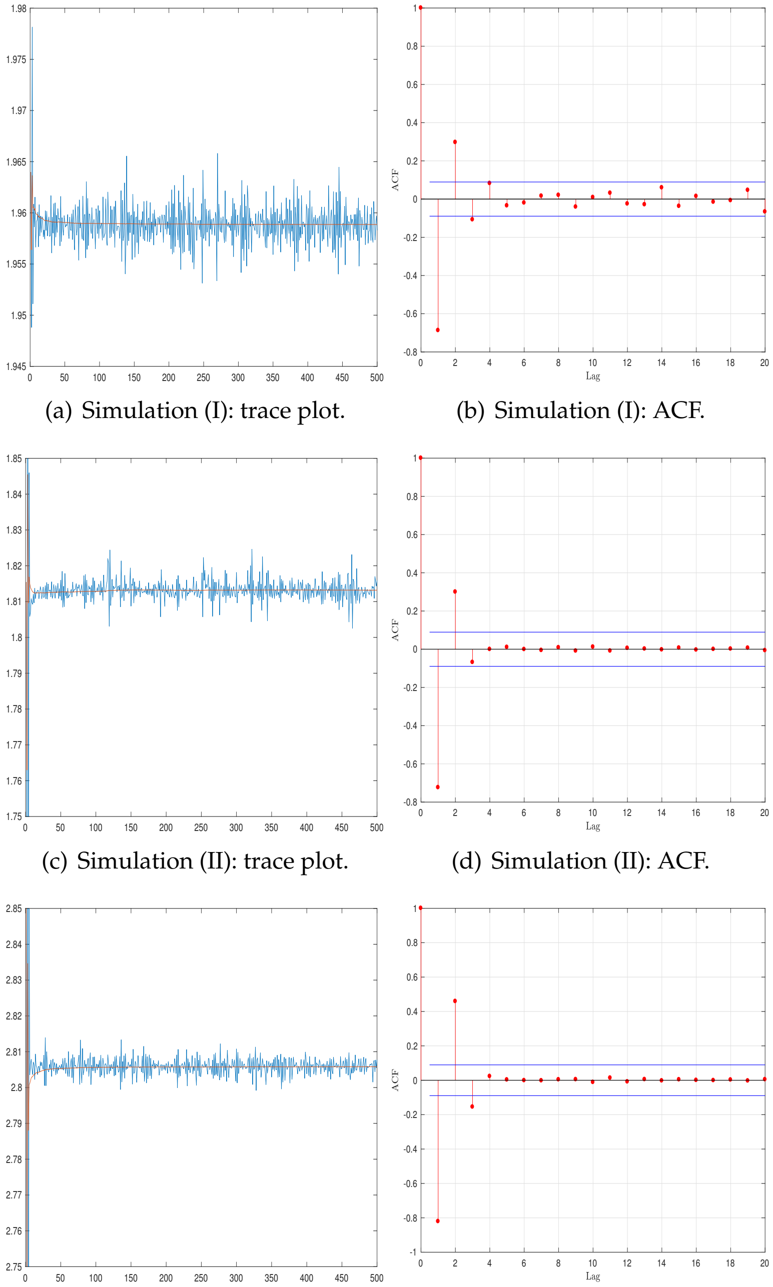
where $\|\cdot\|_2$ is the Frobenious norm for tensors, i.e.:

$$\left\| \mathcal{G}_\ell^* - \hat{\mathcal{G}}_\ell \right\|_2^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (g_{ijk,\ell}^* - \hat{g}_{ijk,\ell})^2 . \tag{3.51}$$

All simulations have been performed using MATLAB r2016b with the aid of the Tensor Toolbox v.2.6[4].

Figs. 3.3(a)-3.3(c)-3.3(e) report the trace plots of the error, for each of the three simulations, respectively, while Figs. 3.3(b)-3.3(d)-3.3(f) plot the corresponding autocorrelation functions. All these graphs show that the estimated total error series rapidly stabilises around a small value, meaning that the sampler is able to recover the true value of the tensor parameter. Furthermore, from the analysis of Figs. 3.3(b)-3.3(d)-3.3(f) we can say that the autocorrelation of the posterior draws of the total error vanishes after three lags, thus representing a first indicator of the efficiency of the sampler. We remind to Appendix C.5 for further details and plots about the performance of the sampler in each simulated example.

[4] Available at: `http://www.sandia.gov/~tgkolda/TensorToolbox/index-2.6.html`



(a) Simulation (I): trace plot.

(b) Simulation (I): ACF.

(c) Simulation (II): trace plot.

(d) Simulation (II): ACF.

(e) Simulation (III): trace plot.

(f) Simulation (III): ACF.

FIGURE 3.3: *Left:* Trace plots (blue line) with superimposed progressive mean (orange line) of the total error in estimating the tensor of regression coefficients, for each simulation. *Right:* corresponding autocorrelation function, for each simulation.

Table (3.1) reports the effective sample size (ESS) in the formulation provided by Gelman et al. (2014):

$$ESS = \frac{N}{1 + 2\sum_{l=1}^{\infty} \hat{\varrho}_l}, \tag{3.52}$$

where $\hat{\varrho}_l$ is the sample autocorrelation function at lag l and N is the sample size (i.e., the length of the simulation). For computational reasons, the infinite sum is truncated at $L = \min\{l : \hat{\varrho}_l < 10^{-4}\}$. The ESS is interpreted as an efficiency index: it represents the number of simulated draws that can be interpreted as iid draws from the posterior distribution (in fact, in presence of exact iid sampling schemes we have $ESS = N$). The results in Tab. (3.1) show that in all three simulation settings the effective sample size is about half of the length of the simulation.

Simulation	ESS	ACF(1)	ACF(5)
I	245		
II			
III			

TABLE 3.1: Convergence diagnostic statistics for the total error, for each simulated case. ESS is rounded to the smallest integer.

3.6 Applications

3.6.1 Data description

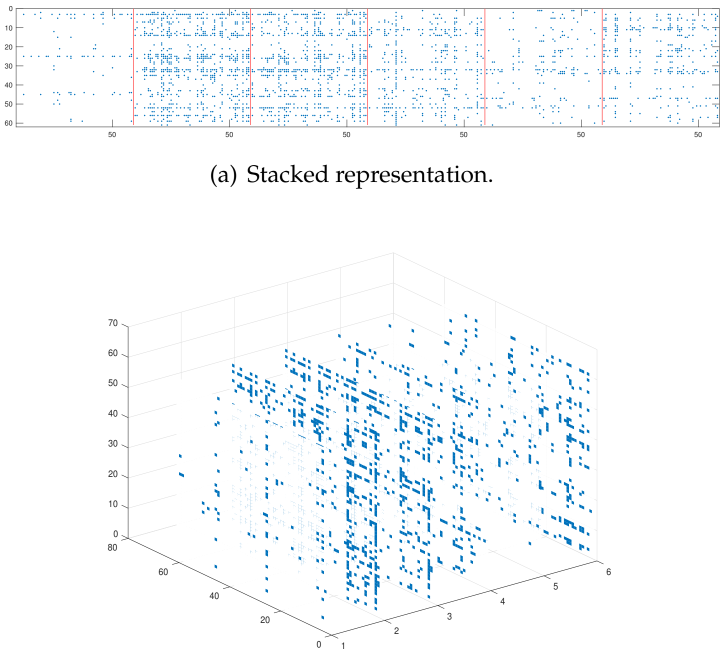
We apply the proposed methodology to the well-known dataset of financial networks of Billio et al. (2012), Ahelegbey et al. (2016b), Ahelegbey et al. (2016b), Bianchi et al. (2018). The dataset consists of $T = 110$ monthly binary, directed networks estimated via the Granger causality approach, where the nodes are European financial institutions. Other methods for extracting the network structure from data can be used, as this is not relevant for our econometric framework, which applies to any sequence of binary tensors.

The original dataset is composed by the daily closing price series at a daily frequency from 29th December 1995 to 16th January 2013 of all the European financial institutions active and dead in order to cope with survivorship bias. It covers a total of 770 European financial firms which are traded in 10 European financial markets (core and peripheral). The pairwise Granger causalities are estimated on daily returns using a rolling window approach with a length for each window of 252 observations (approximately 1 year). We obtain a total of 4197 adjacency matrices during the period from 8th January 1997 to 16th January 2013.

Then, we define a binary adjacency matrix for each month by setting an entry to 1 only if the corresponding Granger-causality link existed for the whole month (i.e. for each trading day of the corresponding month), and setting the entry to 0 otherwise. Since the panel is unbalanced due to entry and exit of financial institutions from the sample over time, we consider a subsample of length $T = 110$ months (from December 2003 to January 2013) made of 61 financial institutions.

We can visualize a sequence of adjacency matrices representing a time series of networks in several ways. Fig. 3.4(a) shows a stacked representation of a subsample composed by six adjacency matrices, while Fig. 3.4(b) plots a 3-dimensional array representation of the same data. In the first case, all matrices are stacked horizontally. Instead, the 3-dimensional representation plots each matrix in front of the other, as frontal slices of an array. It is possible to interpret the two plots as equivalent representations of a third-order tensor: in this case, Fig. 3.4(a) shows the matricised form (along mode 1) of the tensor, while Fig. 3.4(b) plots its frontal slices. Finally, Fig. 3.5 plots the graph associated to two of these adjacency matrices. Though this representation allows for visualising the topology of a network, it is impractical for giving a compact representation of the whole time series of networks. Thus, we provide in Fig. 3.6 the stacked representation of the whole network sequence. Each row plots twelve time-consecutive adjacency matrices, starting from the top-left corner.

The most striking features emerging from Fig. 3.6 are the time-varying degree distribution and the temporal clustering of sparse and dense networks.



(a) Stacked representation.

(b) 3D representation.

FIGURE 3.4: Stacked (a) and 3-dimensional (b) representations of a subsample of adjacency matrices (months $t = 65, 69, 73, 77, 81, 85$). Blue dots are existing edges, white dots are absent edges. A red line is used to separate each matrix (or tensor slice).

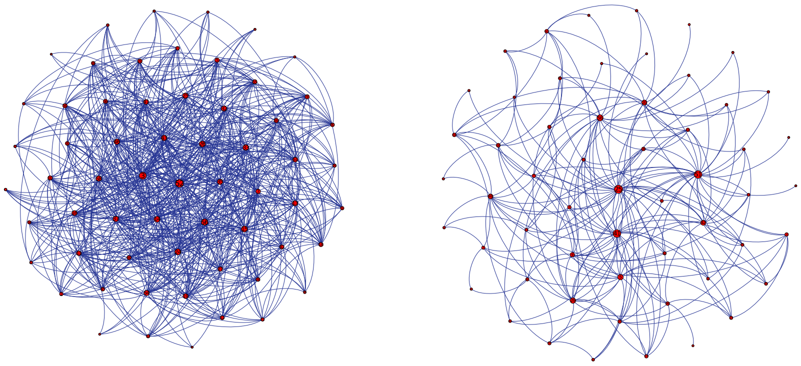


FIGURE 3.5: Graphical representation of networks at time $t = 69$ (dense case) and $t = 77$ (sparse case), respectively. The size of the each node is proportional to its total degree. Edges are clockwise directed.

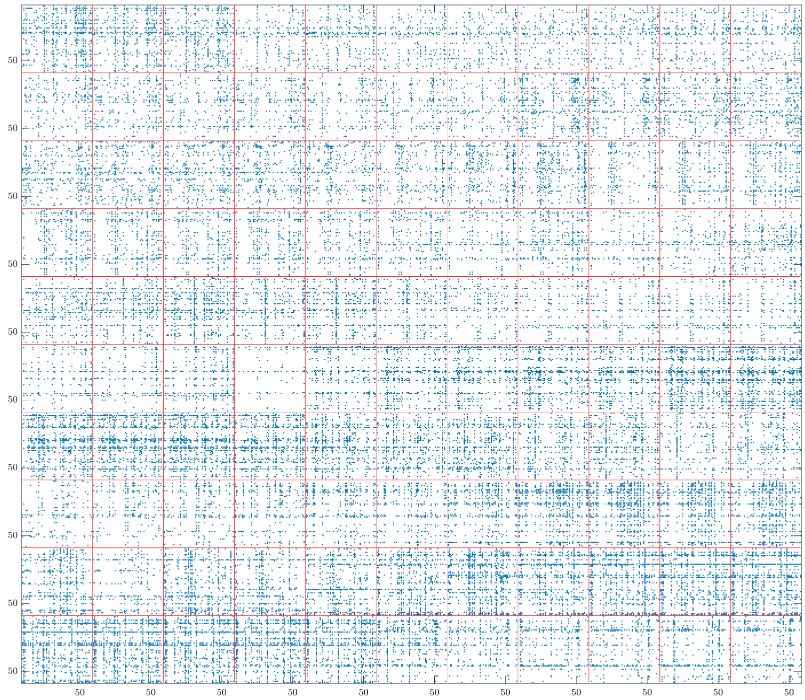


FIGURE 3.6: Full network dataset, with $I = J = 61$ nodes and sample size of $T = 110$ months. In each red box there is an adjacency matrix, starting from top-left at time $t = 1$, the first row contains matrices from time $t = 1$ to $t = 11$, the second row from $t = 12$ to $t = 22$ and so on. Blue dots are existing edges, white dots are absent edges. Red lines are used to delimit the matrices. Labels on the horizontal and vertical axes stand for the corresponding node of the network.

The set of covariates $\mathbf{z}_t$ used to explain each edge's probability includes a constant term and:

- the network total degree (dtd), defined as the total number of edges in the network at time $t = 1, \dots, T$;
- the monthly change of the VSTOXX index (DVX), which is the volatility index for the STOXX50 (and may considered the counterpart of the VIX for Europe);
- the monthly log-returns on the STOXX50 index (STX), taken as a European equivalent to the US S&P500 index;
- the credit spread (crs), defined as the difference between BAA and AAA indices provided by Moody's;
- the term spread (trs), defined as the difference between the 10-year returns of reference Government bonds and the 6-months EURIBOR;
- the momentum factor (mom), obtained from Kenneth French's website[5], provides a measure of th tendency for rising asset prices to rise further and falling prices to keep falling.

All covariates have been standardised and included with one lag of delay, except DVX which is contemporaneous to the response, following the standard practice in volatility modelling (e.g., see, Corsi et al. (2013), Delpini and Bormetti (2015) Majewski et al. (2015)).

[5] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

3.6.2 Results

We estimated a stripped-down version of the general model presented in Section (3.2) consisting only of eq. (3.8a). We run the Gibbs sampler for $N = 10000$ iterations, after having initialised the latent state variables $\{\mathbf{s}_t\}_t$ according to suitable network statistics and the marginals of the tensor decomposition in both regimes (see Supplementary material for further details). We estimate the model with tensor rank $R = 5$ and discuss the main empirical findings (the analysis has been performed also for $R = 8$, obtaining similar results).

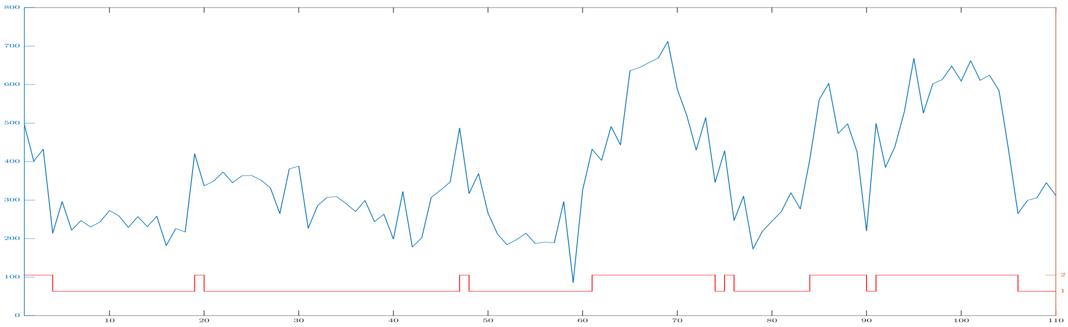


FIGURE 3.7: Total degree of the observed network time series (blue) against the estimated hidden Markov chain (red).

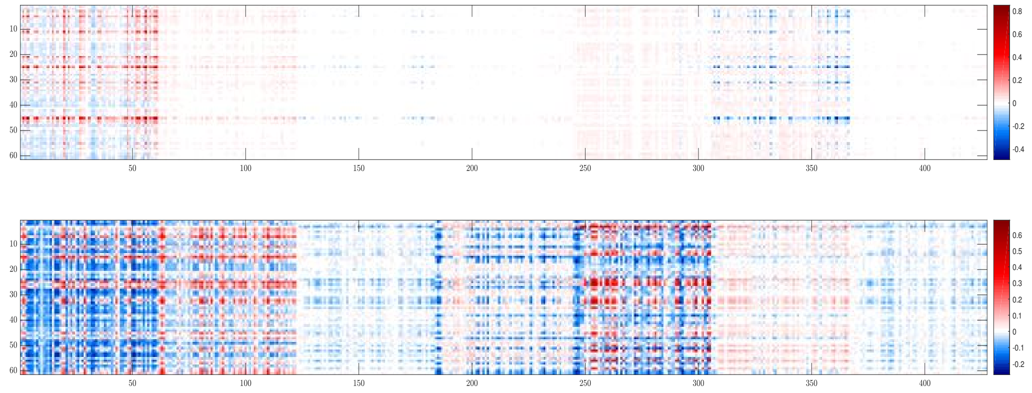


FIGURE 3.8: Posterior mean of tensor of coefficients, in matricised form, in the first (*top*) and second (*bottom*) state of the Markov switching process. For all the slices of each tensor we used the same color scale. Red, blue and white colors indicates positive, negative and zero values of the coefficients, respectively.

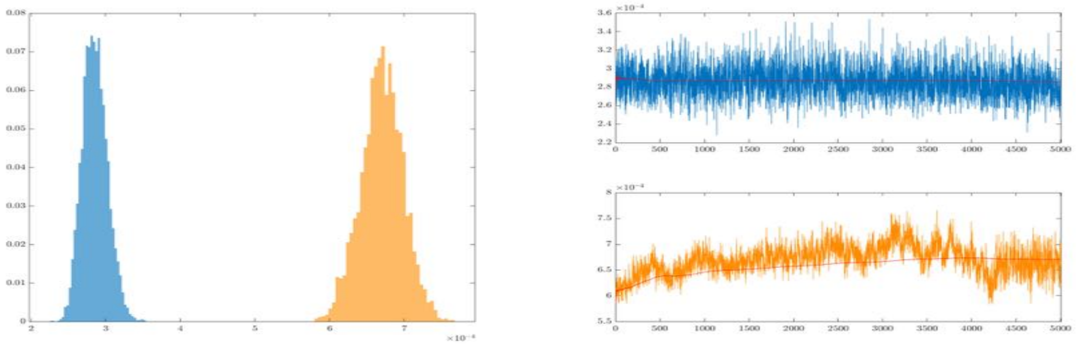


FIGURE 3.9: Posterior distribution (*left* plot) and MCMC output (*right* plots) of the quadratic norm of the tensor of coefficients, in regime 1 (*blue*) and regime 2 (*orange*).

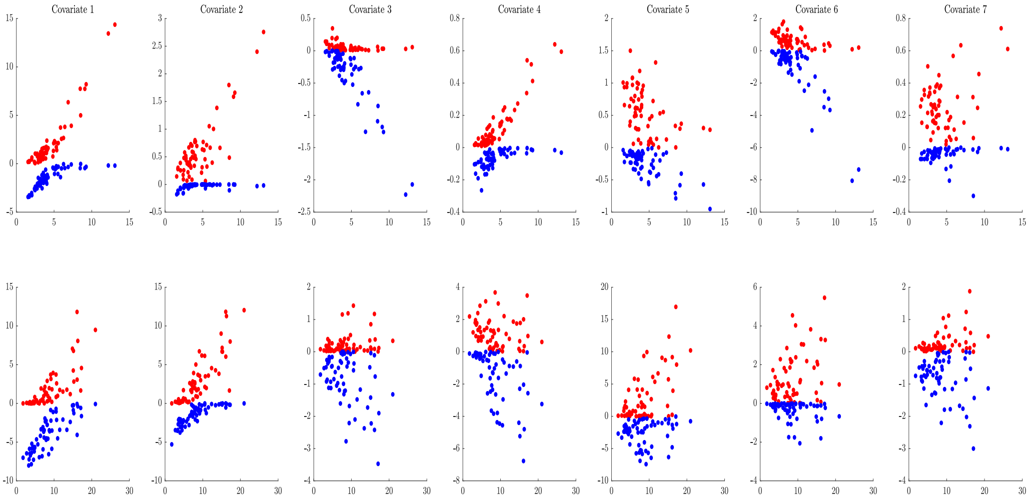


FIGURE 3.10: Scatter plots of total node degree averaged over networks within regime (*x-axis*) versus the sum of positive (*y-axis*, *red*) and the sum of negative (*y-axis*, *blue*) entries of each slice of the coefficient tensor, in regime 1 (*top*) and regime 2 (*bottom*). Coefficients corresponding to incoming edges' probability.

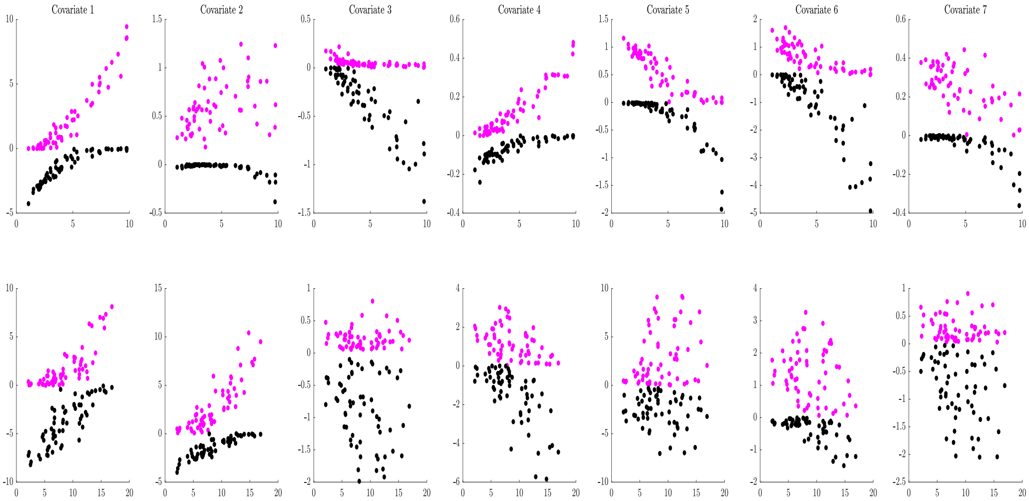


FIGURE 3.11: Scatter plots of total node degree averaged over networks within regime (*x-axis*) versus the sum of positive (*y-axis*, *magenta*) and the sum of negative (*y-axis*, *black*) entries of each slice of the coefficient tensor, in regime 1 (*top*) and regime 2 (*bottom*). Coefficients corresponding to outgoing edges' probability.

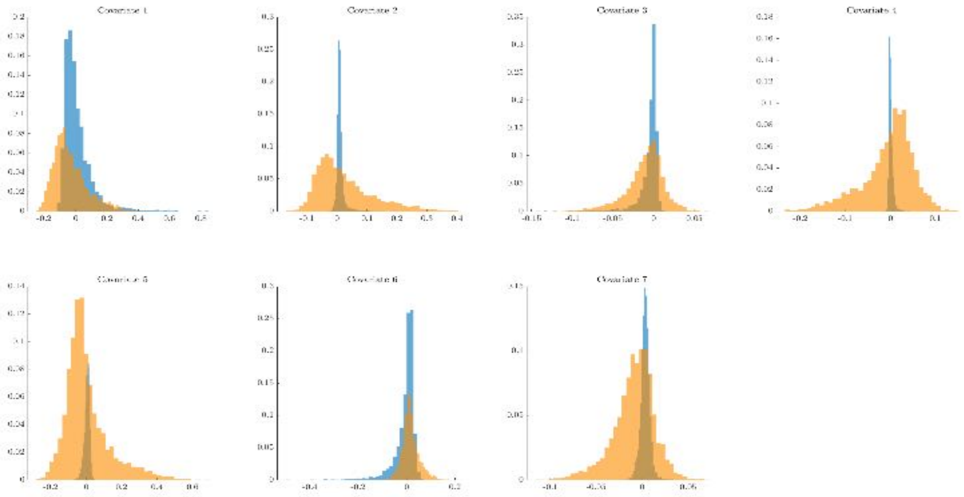


FIGURE 3.12: Distribution of the entries of each slice (different plots) of the estimated coefficient tensor, in regime 1 (*blue*) and regime 2 (*orange*).

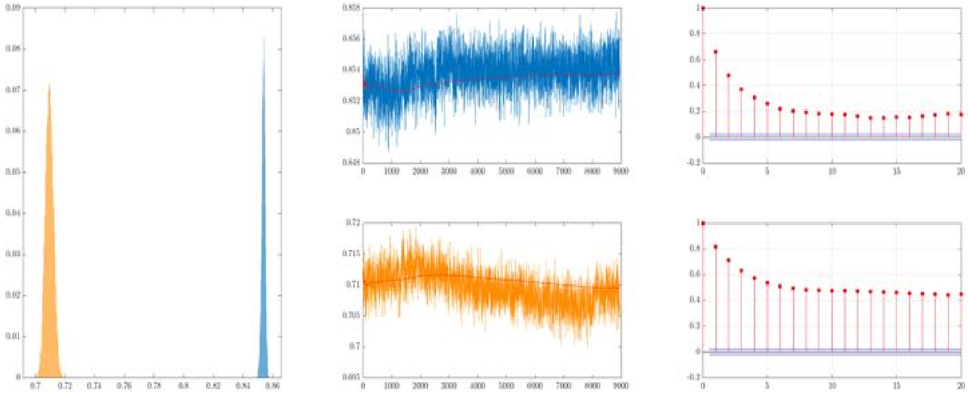


FIGURE 3.13: Posterior distribution (*left* plot), MCMC output (*middle* plots) and autocorrelation functions (*right* plots) of the mixing coefficient ρ_l in the two regimes. Regime 1 in blue and in top panels, regime 2 in orange and in bottom panels.

The estimated hidden Markov chain is plotted in Fig. 3.7 together with the total degree of the observed network time series, using label 1 for the sparse regime and label 2 for the dense regime. The algorithm associates to the dense state in periods when the total degree of the network is remarkably above the average.

There is substantial heterogeneity in the effect of covariates across edges, within the same regime, as reported in Fig. 3.8. Here, the estimated tensor is plotted in matricized form along mode 1 (using two different color scales in the two figures): on the vertical axis we have 61 nodes, while on the horizontal there are $61 \cdot 7$ nodes (the number of covariates including the constant), corresponding to 7 matrices, one for each covariate, horizontally stacked. The entry (i, j) of matrix in position k reports the coefficient of the k-th covariate on the probability of observing the edge (i, j). Thus, within the same regime we observe a significant variation of both the sign and the magnitude of the effect of a covariate on the probability of observing an edge. In words, there is not a single covariate able to explain (and predict) an edge's probability by itself, but several indicators are required. Moreover, a model with pooled time series fails to capture such heterogeneity. The posterior mean is 1.56 in regime 1 and 4.63 in regime 2, but in both cases it is not significantly different from zero, that is, it lies inside a 95% credible interval around zero (see Fig. C.31 in Appendix C.6). This is in contrast with our model, where the fraction of tensor entries statistically different from zero is 12% and 36% in regime 1 and 2, respectively. Thus, we conclude that a pooled model is not suited for describing the heterogeneous effects of the various covariates on the different edges, whereas our model is able to capture them.

We find substantial evidence in favour of major changes of the effects that the covariates exert on the edges' probability in the two regimes. By comparing the two matricized tensors in Fig. 3.8 we note that both the sign and the magnitude of the coefficients differ in the two states. The interpretation is that, according to the regime, the probability of observing a link between two nodes is driven by a different set of variables but also the qualitative influence (i.e. the sign of the coefficient) of the same regressor varies. For example, on average, the credit spread seems to exert a positive effect on the probability of observing an edge in the sparse regime, while its effect in the dense regime has higher magnitude and acts in the opposite direction.

Fig. 3.10 reports, for each regime and covariate, the scatter plot of the total degree of each node (horizontal axis), averaged within regime, against the sum of all negative and positive coefficients' values for the probability of observing an incoming edge. Similarly, Fig. 3.11 shows the same plot considering the effects on the probability of observing an outgoing edge. Together, these plots allow to detect the existence of a relation between the overall positive and negative magnitude of the effects of the covariates on the probability of observing an edge, conditionally on the total degree of the node to which the link is attached. The results show that for several covariates such an association exists: on average, more central nodes (i.e. those with higher total degree) tend to have higher probability of establishing an edge, either incoming or outgoing. This is due to the upward sloping shape of the scatter plot. It is remarkable to notice that for different covariates, such as the momentum factor, there is a different relation for negative and positive effects: for increasingly central nodes, both sums tend to more extreme values. Moreover, by comparing the results in Figs. 3.10-3.11 we see that the results are similar if we look either at incoming or outgoing edges. Finally, between regimes there seems to be no change except for the strength of the relation, which appears stronger in the second one (corresponding to the dense state of the network).

In Fig. 3.12 we plot the distribution of the entries of each slice (over the edges), for every regime, for a more qualitative analysis of the change of the coefficients' values between regimes. There is a different dispersion in the cross-sectional distribution of the coefficients' estimates. In particular, all distributions appear more concentrated around zero in the sparse state, while in the dense regime the mean value is different (and varies according to the covariate) and all distributions show fatter tails than in sparse state.

As a summary statistic, Fig. 3.9 reports the distribution and the trace plots of the quadratic norm of the tensor coefficients in each regime. The two distributions are well separated, with the norm in the first regimes concentrated around smaller values than in the second regime. This implies that, on average, in the sparse state there is a higher probability that the zeros (i.e. absence of edges) are due to the structural component (that is, the Dirac mass in eq. (3.3)), moreover the probability of success of the Bernoulli distribution is smaller than in the dense regime.

Finally, Fig. 3.13 shows the posterior distributions of the regime-dependent probabilities of observing a structural zero, in the two regimes. The distributions are well separated, with posterior means around 0.85 and 0.71 in the sparse and dense state, respectively.

Additional plots regarding the hyper-parameters of the model are shown in Appendix C.6.

3.7 Conclusions

We presented an econometric framework for modelling of a time series of binary tensors, which we interpret as representing multi-layer networks. We proposed a parsimonious parametrization based on the PARAFAC decomposition of the parameter tensor. Moreover, the parameters of the model can switch between multiple regimes, thus allowing to capture the time-varying topology of financial networks and economic indicators. We specified a zero-inflated logit model for the probability of each entry of the observed tensor, which permits to capture the varying sparsity patterns of the observed time series. The proposed framework is also able to jointly model a temporal sequence of binary arrays and a vector of economic variables.

We followed the Bayesian paradigm in the inferential process and developed a Gibbs sampler via multiple data augmentation steps for estimating the parameters of interest. The performance of the algorithm has been tested on simulated datasets with networks of different sizes, ranging from medium (i.e. 100 nodes) to big (i.e. 200 nodes). The results of the estimation procedure are encouraging in all simulations.

Finally, we estimated a stripped-down version the model on a real dataset of networks between European financial institutions. The results suggest the existence of two different regimes associated to the degree density of the network. Moreover, in each regime the most degree central nodes tend to be more sensitive to the effect of covariates (either positive or negative) on their probability to link to other nodes. Overall, the probability of forming an edge is not depending on a single covariate, but a combination of several financial indicators is needed to explain and predict it. Finally, nature of the absent edges is estimated to be different, with the sparse regime having a high probability of structural zeros, as compared to the dense regime.

Acknowledgements

This research has benefited from the use of the Scientific Computation System of Ca' Foscari University of Venice (SCSCF) for the computational for the implementation of the inferential procedure.

Chapter 4

Nonparametric forecasting of multivariate probability density functions

> *The infinite! No other question has ever moved so profoundly the spirit of man.*
>
> David Hilbert

> *Divide each difficulty into as many parts as is feasible and necessary to resolve it.*
>
> René Descartes

4.1 Introduction

One of the most relevant research fields in theoretical and applied statistics is devoted to the study of the dependence between random variables. In finance, the analysis of the dependence patterns is a challenging problem and its understanding serves several purposes: control of risk clustering, credit, market and systemic risk measurement, pricing and hedging of credit sensitive instruments (such as collateralized debt obligations or CDOs) and credit portfolio management. The analysis of the relationships between economic and financial variables is crucial for the identification of causality relations (e.g., see Granger (1988), White and Lu (2010)). From a statistical perspective, the main purpose is the development of models able to describe and forecast the joint dynamic behaviour of financial variables. Moreover, these models may provide an effective support for the financial regulator (for example, in predicting and counteracting an increase of the systemic risk).

Firstly developed by Sklar (1959), copula functions have attracted significant attention over the last decade, particularly within the financial and econometric communities, as a flexible instrument for modelling the joint distribution of random variables (see Joe (1997), Nelsen (2013) and Durante and Sempi (2015) for an introduction and a compelling review). Let $(X_1, \ldots, X_d)$ be a random vector with continuous marginal cumulative distribution functions (cdf) $F_i(\cdot)$ and probability density function (pdf) $f_i(\cdot)$. The random vector $(U_1, \ldots, U_d) = (F_1(X_1), \ldots, F_d(X_d))$, obtained by application of the probability integral transform, has uniform marginals. The copula of $(X_1, \ldots, X_d)$ is defined as the joint cumulative distribution function $C : [0, 1]^d \to [0, 1]$ of $(U_1, \ldots, U_d)$, that is $C(u_1, \ldots, u_d) = \mathbb{P}(U_1 \leq u_1, \ldots, U_d \leq u_d)$. Moreover, denoting $F(\cdot)$ the joint cumulative distribution of $(X_1, \ldots, X_d)$ and by $f(\cdot)$ its probability density function, Sklar's theorem (Sklar (1959)) states that there exists a unique copula $C(\cdot)$ with probability density function $c : [0, 1]^d \to \mathbb{R}_+$ such that $F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$ and $f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i)$.

The use of a copulas permits to separately deal with the marginal distributions and the dependence structure among a set of random variables, thus providing a high degree of

Copula family	**Upper λ_U**	**Lower λ_L**
Gaussian(ρ)	1 if $\rho = 1$	1 if $\rho = 1$
t-student(ν, ρ)	$\lambda > 0$ if $\rho > 1$	$\lambda > 0$ if $\rho > 1$
Gumbel(θ)	$2 - 2^{1/\theta}$	0
Clayton(θ)	0	$2^{-1/\theta}$
Frank(θ)	0	0
Fréchet(p, q)	q	q

TABLE 4.1: Upper and lower tail dependence for some copula families. In brackets the parameters of the copula family.

flexibility in modelling the corresponding joint distribution. The literature on quantitative finance and financial econometrics has widely recognized the importance of this instrument, as documented by the review of Patton (2012) and the textbooks by Cherubini et al. (2004) and Cherubini et al. (2011).

A standard practice in financial econometrics, motivated by the fact that multivariate data (e.g., returns of a portfolio of assets) have non-Gaussian marginal distributon, consists in assuming a heavy-tailed distribution for the marginals (or to estimate them nonparametrically) and to join them with a parametric copula function, which fully describes the dependence structure (Deheuvels (1978), Deheuvels (1979)) through its parameters. This approach allows a parsimonious description of the dependence between two variables by means of the few parameters of a copula function.

This method has some undesirable shortcomings. First, each parametric copula family is designed to describe only a specific dependence pattern (for example, see Table 4.1), which makes the selection of the family a crucial aspect of every modelling strategy. To this aim, we recall the definition of upper (lower) tail dependence from Cherubini et al. (2004). This concept is used to describe situations where high (low) values of a variables are likely to be observed together with high (low) values of the other. In terms of the copula pdf, this means that the probability is concentrated to the top-right (bottom-left) corner. Formally, given two random variables $X \sim G_X$ and $Y \sim G_Y$ with bivariate copula $C(\cdot)$, the upper and lower tail dependence coefficients (upper TDC and lower TDC, respectively) are given by:

$$\lambda_U = \lim_{u \to 1^-} \mathbb{P}(G_X(X) > u | G_Y(Y) > u) = \lim_{u \to 1^-} \frac{1 - 2u + C(u,u)}{1-u}, \tag{4.1}$$

$$\lambda_L = \lim_{u \to 0^+} \mathbb{P}(G_X(X) < u | G_Y(Y) < u) = \lim_{u \to 0^+} \frac{C(u,u)}{u}, \tag{4.2}$$

which are asymptotically equivalent to:

$$\lambda_U = 2 - \lim_{u \to 1^-} \frac{\log(C(u,u))}{\log(u)}, \qquad \lambda_L = 2 - \lim_{u \to 0^+} \frac{\log(1 - 2u + C(u,u))}{\log(1-u)}. \tag{4.3}$$

The copula $C(\cdot)$ is said to have upper (lower) tail dependence when $\lambda_U \neq 0$ ($\lambda_L \neq 0$). Table 4.1 reports the tail dependence coefficients for some commonly used copula families (see Cherubini et al. (2004)). Only some of them (e.g., Gaussian, t-student and Fréchet, for some values of their parameters) have simultaneously upper and lower tail dependence: this happens only for some values of the parameter of the copula and, in any case, the tail dependence coefficients are equal, thus implying that the tail dependence is symmetric.

Second, when the copula parameter is assumed to be fixed, these constructions are able to identify only the overall, static relations and fail to account for any kind of spatial or temporal change. This constraint is particularly restrictive in time series analysis of financial data, as pointed out by Fermanian and Scaillet (2004). In fact, very often the relations between financial variables are non linear and change over time.

To address this shortcoming, Patton (2006a) and Patton (2006b) introduced dynamic copula models by assuming that the parameters of the copula function are driven by an autoregressive process. Instead, Fermanian and Wegkamp (2012) allowed the parameters to depend on past realizations of the observables. These seminal works, have opened a new avenue to research (see Manner and Reznikova (2012) for a review) and has brought outstanding improvements to the econometrician's toolbox. For example, So and Yeung (2014) and Jondeau and Rockinger (2006) incorporate dynamics into a copula-GARCH model improving its forecasting performance, Dias and Embrechts (2004) and Van Den Goorbergh et al. (2005) exploited dynamic copulas in modelling high-frequency data and option pricing, respectively, whereas Oh and Patton (2017) has recently applied this methodology to the study of systemic risk. Other relevant empirical contributions exploiting this construction include Almeida et al. (2016), Bartram et al. (2007), Weiß and Supper (2013), Hu (2010), Hafner and Reznikova (2010), Hafner and Manner (2012), Guidolin and Timmermann (2006).

Despite the greater flexibility, dynamic copulas may fail to account for the characteristics of the dependence structure among financial data. Since each copula family is constructed for describing a specific dependence pattern, the change of its parameters may not be enough to capture other types of dependence.

The recent paper by Guégan and Zhang (2010) found empirical evidence supporting this theory. They developed a strategy for testing the null of a static copula against a dynamic copula, then upon rejection they tested for the change of the copula family over different temporal windows. The main results is that the dependence structure between the S&P500 and NASDAQ indices experienced a great variability over time, thus stressing the need for a dynamic model; nonetheless the null hypothesis of equal copula function family was rejected for several windows. This suggests that, in this dataset, the evolution of the dynamic copula parameter is insufficient to account for the variation of the dependence structure and different copula families should be used for different temporal windows.

To overcome these limitations, we propose a methodology that has the whole function as the object of interest, instead of a finite-dimensional parameter vector. We do this by exploiting some results developed in the literature on functional data analysis, which we briefly introduce in the following.

Starting from the seminal work of Bosq (2000), functional data analysis (see Ramsay and Silverman (2005) and Ferraty and Vieu (2006) for a thorough treatment) has found applications also in finance and financial econometrics (see Hörmann and Kokoszka (2012), Kokoszka (2012) and Kidziński (2015) for an introduction to the topic). Different models have been proposed for time series of functional data: for example, Sen and Klüppelberg (2015) assumed stationarity and estimated a VARMA functional process for electricity spot prices, conversely Liebl (2010) used the same data but proposed a method for dealing with non-stationary data and applied it in Liebl (2013) for forecasting. Within the same stream of research, Horváth et al. (2010) and Horváth et al. (2014) designed a testing procedure for detecting non-stationarity. Aue et al. (2015) and Kargin and Onatski (2008), instead, used time series functional for improving on the forecasting performance of multivariate on forecasting. More recently, Kidziński et al. (2016) and Klepsch et al. (2017) extended the theory of univariate ARMA models to the functional framework, by introducing also seasonal effects. Finally, Petris (2013) and Canale and Ruggiero (2016) developed an inferential procedure following the Bayesian paradigm, following a parametric and non-parametric approach, respectively.

The literature on functional time series modelling can be partitioned into two main classes, according to the methodology developed. The parametric framework, firstly introduced by Bosq (2000), hinges on the linear functional autoregressive model (FAR) which can be considered an infinite-dimensional analogue of vector autoregressive (VAR) processes, widely used in times series analysis. By contrast, the non-parametric approach (see Ferraty and Vieu (2006) and Ramsay and Silverman (2005) for an overview) relies on functional principal component analysis (fPCA). See Appendix D.1 for a short introduction.

Unfortunately, none of the previously mentioned approaches is suited for dealing with constrained functions, such as probability density functions (pdfs), which must be positive (on their support) and have unit integral. A statistical model for the analysis of a time series of pdfs should include a mechanism for dealing with these constraints. In the literature, three main approaches have been proposed to address this issue. One possibility consists in ignoring the constraints and treating the pdfs as an unrestricted functions, then after the estimation and forecasting steps, the output is re-normalized in order to get a probability density function. Sen and Ma (2015) adopted this approach for studying a time series of pdfs of financial data from the S&P500 and the Bombay Stock Exchange.

More appealing alternatives do not need to post-process the output and allow to perform the analysis taking into account the constraints. The seminal works by Egozcue et al. (2006), van der Boogaart et al. (2010), van der Boogaart et al. (2014) and Egozcue and Pawlowsky-Glahn (2015) introduced the notion of Bayes space, that is a Hilbert space of probability density functions. They borrowed from compositional data analysis and Aitchison's geometry (see Aitchison (1986)) and interpreted probability density functions as infinite-dimensional compositional vectors. They replaced the pointwise sum and multiplication by the operations of perturbation and powering which, for $f(\cdot), g(\cdot) \in \mathbb{D}(I)$, $I \subset \mathbb{R}^n$, and $\alpha \in \mathbb{R}$, are defined by:

$$f(\mathbf{x}) \oplus g(\mathbf{x}) = \frac{f(\mathbf{x})g(\mathbf{x})}{\int_I f(\mathbf{x})g(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x})}, \qquad \alpha \odot f(\mathbf{x}) = \frac{f(\mathbf{x})^\alpha}{\int_I f(\mathbf{x})^\alpha\,\mu(\mathrm{d}\mathbf{x})}. \tag{4.4}$$

Instead, the analogue of subtraction is given by $f(\cdot) \ominus g(\cdot) = f(\cdot) \oplus [-1 \odot g(\cdot)]$. They showed that the tuple $(\mathbb{D}(I), \oplus, \odot)$ is a space and that the subset $\mathbb{D}^*(I) \subset \mathbb{D}(I)$ of probability density functions whose logarithm is square integrable is a Hilbert space. These remarkable results permit to conduct the analysis directly in $\mathbb{D}^*(I)$, provided that it is possible to re-define the necessary statistical models by means of the new operations $\oplus, \odot$. van der Boogaart et al. (2014) proved that $\mathbb{D}^*(I)$ is isomorphic to the space $\mathcal{L}_2^*(I)$ of functions on I with square integrable logarithm (written $\mathbb{D}^*(I) \cong_{\mathrm{clr}} \mathcal{L}_2^*(I)$) via the centred log-ratio isometry defined as follows (see Section 4.2.1 for the notation).

Definition 4.1.1 (Centred log-ratio)
Let ν be a measure on $\mathbb{R}^n$ and $f : I \to \mathbb{R}_+$ be a probability density function supported on a set $I \subset \mathbb{R}^n$ of finite ν-measure, that is $\nu(I) < \infty$ and $\nu(I) \neq 0$. The centred log-ratio (clr) transformation is an invertible map is defined as:

$$\mathrm{clr}(f)(\mathbf{x}) = g(\mathbf{x}) = \log(f)(\mathbf{x}) - \frac{1}{\nu(I)} \int_I \log(f)(\mathbf{y})\,\nu(\mathrm{d}\mathbf{y}), \tag{4.5}$$

with inverse given by:

$$\mathrm{clr}^{-1}(g)(\mathbf{x}) = f(\mathbf{x}) = \frac{\exp(g)(\mathbf{x})}{\int_I \exp(g)(\mathbf{y})\,\nu(\mathrm{d}\mathbf{y})}. \tag{4.6}$$

Consequently, by definition 4.1.1 it follows that the clr transform of a pdf supported on I has to satisfy the following constraint (which we will call zero integral constraint in the rest of this paper):

$$\int_I \mathrm{clr}(f)(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) = \int_I \log(f)(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) - \int_I \frac{1}{\mu(I)} \left[\int_I \log(f)(\mathbf{y})\,\mu(\mathrm{d}\mathbf{y})\right] \mu(\mathrm{d}\mathbf{x}) = 0. \tag{4.7}$$

The spaces $\mathbb{D}(I), \mathbb{D}^*(I), \mathcal{L}_1(I), \mathcal{L}_2(I), \mathcal{L}_2^*(I)$ are defined with respect to a reference measure ν and contain equivalence classes of functions which are proportional each other, that is we implicitly defined $\mathbb{D}(I) = \mathbb{D}_\nu(I)$, $\mathbb{D}^*(I) = \mathbb{D}_\nu^*(I)$, $\mathcal{L}_{1,\nu}(I) = \mathcal{L}_{1,\nu}(I)$, $\mathcal{L}_2(I) = \mathcal{L}_{2,\nu}(I)$, $\mathcal{L}_2^*(I) = \mathcal{L}_{2,\nu}^*(I)$. In this paper we consider to be the Lebesgue reference measure, i.e. $\nu = \mu$.

In order to single out a specific element it is necessary to normalize the reference measure. This can be easily done if the set I is ν-finite, whereas if $\nu(I) = \infty$, normalization can be done using the centring procedure (see van der Boogaart et al. (2014)). Moreover, the following relations hold: $\mathbb{D}^*(I) \subset \mathbb{D}(I) \subset \mathcal{L}_1(I)$ and $\mathbb{D}^*(I) \cong_{\text{clr}} \mathcal{L}_2^*(I) \subset \mathcal{L}_2(I) \subset \mathcal{L}_1(I)$.

It is possible to use the clr transform to project a pdf into the Hilbert space $\mathcal{L}_2^*(I)$ (provided that its logarithm is square integrable), which is a space embedded with the operations of pointwise addition and multiplication. We can perform the statistical analysis in this space and then project the output back into $\mathbb{D}^*(I)$ by the inverse map $\text{clr}^{-1}(\cdot)$. This strategy via the centred log-ratio map has been proposed by Hron et al. (2016) for performing fPCA on univariate pdfs with compact support. Canale and Vantini (2016) developed a different isometric, bijective function which maps constrained functions into a pre-Hilbert space, then estimated a FAR model on this space and transformed back the result into the original space. Though general, this framework is not explicitly designed for dealing with pdfs, but only bounded and monotonic functions, thus preventing from its use in the current setting. The same idea of transforming pdfs into $\mathcal{L}_2(I)$ via an invertible map has been followed by Petersen and Müller (2016), who defined two different transformations satisfying this property: the log hazard and the log quantile density transformations, respectively. Despite their strong theoretical properties, both maps have the shortcoming of not having an equivalent transformation applicable in the multivariate case, which makes them unsuited for the analysis of multivariate probability density functions.

The empirical finding by Guégan and Zhang (2010) represents the key stylized fact motivating our work. Given that a dynamic copula model may not be sufficiently flexible to describe the time varying dependence between financial variables, we contribute to this active field of research by proposing a different statistical framework for forecasting multivariate probability density functions. To address the issues related with modelling pdfs, we extend the procedure of Hron et al. (2016) who build on the previous work by van der Boogaart et al. (2010, 2014). The idea is to map the space of probability density functions to the space of integrable functions through an isometry, perform the analysis in this space (which has nicer properties), then use the inverse mapping to get the solution in the original space. Our contribution is also related to the studies of Liebl (2013) and Hays et al. (2012), who developed dynamic models for forecasting functional time series of electricity prices on the basis of fPCA. However, our focus is on the modelling of probability density functions, which call for the adoption of more complex tools than that of unrestricted functions. Finally, we contribute to the literature on dynamic dependence modelling in finance by providing a tool able to forecast the temporal evolution of the dependence pattern between the S&P500 and NASDAQ indices.

We propose a nonparametric framework for forecasting multivariate probability density functions by extending existing models for the analysis of cross sectional, univariate probability density functions with compact support. We focus on bivariate copula pdfs because of their great importance in finance, however the proposed methodology is flexible and general, thus permitting to deal with the prediction of general pdfs with bounded or, under some conditions, unbounded support. Thanks to the fact that a copula pdf encompasses all information on the dependence structure, we can interpret our approach as a general framework for modelling the (temporally evolving) dependence patterns between random variables.

The reminder of the chapter is as follows. In Section 4.2 we introduce the notation as well as the fundamental concepts that will be used throughout the paper. Section 4.3 presents the details of the proposed baseline methodology, whereas Section 4.4 provides insights on potential issues and extensions. Section 4.5 provides an overview of the financial dataset used and presents the results of the empirical analysis. Finally, Section 4.6 concludes and describes some extensions of the current work and lines of future research.

4.2 Preliminaries

In this section we describe the proposed methodology after having introduced the main notation that will be used throughout the rest of the paper.

4.2.1 Notation

Throughout the paper, if not differently specified, greek letters denote unknown quantities to be estimated, whereas latin letters any other variable. We denote scalars with lower-case letters, vectors with boldface lower-case letters and matrices with boldface upper-case letters. We use the shorthand $f(\cdot)$ for denoting a function, regardless of the number of arguments it takes, moreover we denote the composition of functions by $(g \circ f)(\cdot) = g(f(\cdot)) = g(f)(\cdot)$. The inner product between two functions $f(\cdot), g(\cdot)$ supported on $I \subseteq \mathbb{R}^n$ is defined in the standard way by $\langle f(\cdot), g(\cdot)\rangle = \int_I f(\mathbf{x})g(\mathbf{x})\,\mathrm{d}\mathbf{x}$. The integer part of the scalar x is denoted $\lfloor x \rfloor$.

We use the notation $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_T]$ to denote a collection of T matrices $\mathbf{A}_t$ of equal size $N \times M$. The symbol $\mathbf{I}_k$ is used for the identity matrix of size $k \times k$, whereas $\mathbf{0}_k$ for the $k \times 1$ column vector of zeros. Moreover, empty spaces in the matrices stand for zero entries. Let $\mathbf{L}_N$ be the matrix representation of the first difference operator L, that is the $N \times (N+1)$ matrix which post-multiplied by a $(N+1)$-dimensional vector $\mathbf{a}$ yields a vector of size N, $\mathbf{La}$, whose entries are the first differences of the elements of $\mathbf{a}$. The Moore-Penrose pseudo-inverse of the $N \times M$ matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{\dagger}$. If $\mathbf{A}$ is positive definite, we define by $\mathbf{A}^{1/2}$ its (unique) principal square root.

In Section 4.3 we will refer to the spaces of functions described as follows. We define $\mathbb{F}_+(I)$ to be the space of non-negative, integrable functions on $I \subseteq \mathbb{R}^n$, whose general element is the function $f : I \to \mathbb{R}_+$, and we let $\mathbb{F}_0(I)$ be the space of functions on I with zero integral. $\mathbb{D}(I)$ denotes the set of probability density functions with support I and we define $\mathbb{D}^*(I)$ to be the space of probability density functions with support I whose logarithm is square integrable. We denote by $\mu(\cdot)$ the Lebesgue measure on $\mathbb{R}^n$, for $n \geq 1$. In the case $n = 1$ we also use the shorthand notation $\mathrm{d}x = \mu(\mathrm{d}x)$, whereas for $n > 1$ we define $\mathrm{d}\mathbf{x} = \mu(\mathrm{d}\mathbf{x})$. Consequently, if $I = [a, b]$ then $\mu(I) = b - a$. All integrability definitions are made using the Lebesgue measure as reference measure, if not differently specified. Let $\mathcal{L}_p(I)$ be the space of p-integrable functions supported on I and let $\mathcal{L}_p^*(I)$ be the space of functions on I whose logarithm is p-integrable. The n-dimensional unit simplex is defined as $\mathcal{S}^n = \{\mathbf{x} \in \mathbb{R}^n : x_i > 0,\ i = 1, \ldots, n, \text{ and } \sum_{i=1}^n x_i = 1\}$, whereas $\mathcal{S}_0 = \{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\}$ is the subspace of n-dimensional vectors whose elements have zero sum. We define $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ be the canonical basis of the space of $N \times N$ matrices. For two spaces X, Y we use the notation $X \cong_f Y$ to indicate that they are isomorphic through the isometric isomorphism $f : X \to Y$.

Let $\mathbf{x} = (x_1, \ldots, x_n)'$ be a vector of observations. In Section 4.3.1 we denote the empirical (marginal) cumulative distribution function of $\mathbf{x}$ by $F^n(\mathbf{x}) = (F^n(x_1), \ldots, F^n(x_n))'$. Moreover, define the rank transformation of $\mathbf{x}$ to be the function that maps each element x_i of $\mathbf{x}$ to:

$$R_i = \sum_{j=1}^{n} \mathbb{1}(x_j \leq x_i)\,. \tag{4.8}$$

Denote with $\mathbf{u} = (u_1, \ldots, u_n)'$ with $u_i \in [0, 1]$, $i = 1, \ldots, n$, the vector of pseudo-observations associated to the observations $\mathbf{x}$, used for the estimation of the empirical copula in Section 4.3.1. Each pseudo-observation is defined as:

$$u_i = \frac{1}{n} R_i\,. \tag{4.9}$$

In Section 4.3.2 and Appendix D.1, we will use the following notation in performing functional principal component analysis (fPCA). We define a sample of observed functions by $\mathbf{f}(\cdot) = (f_1(\cdot), \ldots, f_T(\cdot))'$, with $f_t : I \to \mathbb{R}$, for some domain $I \subseteq \mathbb{R}^n$. Moreover, we let $\check{\mathbf{f}}(\cdot) = (\check{f}_1(\cdot), \ldots, \check{f}_T(\cdot))'$ denote the approximation of the observed functions obtained as an outcome of the fPCA. The principal component functions are denoted by $\boldsymbol{\xi}(\cdot) = (\xi_1(\cdot), \ldots, \xi_J(\cdot))'$ and the scores associated to the function $\check{f}_t(\cdot)$ are $\boldsymbol{\beta}_t = (\beta_{t,1}, \ldots, \beta_{t,J})'$, moreover let $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T)$. The corresponding estimated quantities are denoted by $\widehat{\boldsymbol{\xi}}(\cdot)$, $\widehat{\boldsymbol{\beta}}_t$ and $\widehat{\mathbf{B}}$.

In Section 4.3.2 we also use spline functions, for which we adopt the following notation. We denote by $\mathfrak{I}_n$ a n-dimensional index set whose elements are the n-tuple $(i_1, \ldots, i_n)$ with entries $i_j \in [1, I_j]$, where I_j is a positive integer for each $j = 1, \ldots, n$. We denote with $\boldsymbol{\lambda}$ the vector with entries $\lambda_0 < \ldots < \lambda_{g+1}$ representing the points of the knot sequence used for the spline functions. For a m-order spline, define the extended knot sequence as the vector $\bar{\boldsymbol{\lambda}}$ of length $2m + g + 2$ whose entries satisfy the relations:

$$\underbrace{\lambda_{-m} = \ldots = \lambda_{-1}}_{m} = \underbrace{\lambda_0 < \ldots < \lambda_{g+1}}_{g+2} = \underbrace{\lambda_{g+1+1} = \ldots = \lambda_{g+m+1}}_{m} .$$

An extended knot sequence for a bivariate spline is defined by $\bar{\boldsymbol{\lambda}}^{x,y} = \bar{\boldsymbol{\lambda}}^x \otimes \bar{\boldsymbol{\lambda}}^y$, where $\bar{\boldsymbol{\lambda}}^x, \bar{\boldsymbol{\lambda}}^y$ are the extended knot sequences along each axis and the generic entry is the couple $\bar{\lambda}_{i,j}^{x,y} = (\bar{\lambda}_i^x, \bar{\lambda}_j^y)$. $B_i^k(x)$ denotes the univariate (basis) B-spline function of degree $m - 1$, with knot sequence indexed by i and $D_x^\ell[f](x)$, $\ell \leq m - 1$, is the partial derivative operator of order ℓ, applied to the function f with respect to the variable x. For univariate splines of degree m, let $\mathbf{b} = (b_{-m}, \ldots, b_g)'$ be the $(g + m + 1)$ coefficient vector, whereas for bivariate splines of the same degree we define the $(m + g + 1) \times (m + g + 1)$ coefficient matrix by $\bar{\mathbf{B}}$. Moreover, let $\mathbf{C}^{m+1}(\mathbf{x}^n)$ be the $n \times (g + m + 1)$ collocation matrix of B-spline functions evaluated at the observation points $\mathbf{x}^n = (x_1, \ldots, x_n)'$:

$$\mathbf{C}^{m+1}(\mathbf{x}^n) = \begin{bmatrix} B_{-m}^{m+1}(x_1) & \ldots & B_g^{m+1}(x_1) \\ \vdots & \ddots & \vdots \\ B_{-m}^{m+1}(x_n) & \ldots & B_g^{m+1}(x_n) \end{bmatrix} \tag{4.10}$$

Following (De Boor, 2001, ch.10), a univariate spline function of degree k and the corresponding partial derivative of order ℓ are given by:

$$s_m(x) = \sum_{i=-m}^{g} b_i B_i^{m+1}(x) , \tag{4.11}$$

$$D_x^\ell[s_m](x) = s_m^{(\ell)}(x) = \sum_{i=-m}^{g} b_i^\ell B_i^m(x) , \tag{4.12}$$

Given an extended knot sequence $\bar{\boldsymbol{\lambda}}$ and evaluation points $\mathbf{x}^n = (x_1, \ldots, x_n)'$, are given by:

$$s_k(\mathbf{x}^n) = \sum_{i=-k}^{g} b_i B_i^{k+1}(\mathbf{x}^n) = \mathbf{C}^{k+1}(\mathbf{x}^n)\mathbf{b} , \tag{4.13}$$

$$D_x^\ell[s_m](\mathbf{x}^n) = s_k^{(\ell)}(\mathbf{x}^n) = \mathbf{C}^{m+1-\ell}(\mathbf{x}^n)\mathbf{b}^{(\ell)} = \mathbf{C}^{m+1-\ell}(\mathbf{x}^n)\mathbf{S}_\ell \mathbf{b} , \tag{4.14}$$

where $\mathbf{C}^{m+1}(\mathbf{x})$ is a matrix of B-splines evaluated at the points $\mathbf{x}$, $\mathbf{b}$ is the coefficient vector and $\mathbf{S}_\ell$ is a matrix transforming the coefficient vectors of splines of degrees m to those of their derivatives of degree ℓ. The direct link between the coefficients of a spline and its derivative are due to the property that the derivative of a spline is another spline of lower degree (see De Boor (2001)), that is:

$$s_{m+1}(x) = \int s_m(x)\,\mathrm{d}x\,. \tag{4.15}$$

Similarly, we define the d-dimensional tensor product spline function by the tensor product between univariate splines (see Schumaker (2007)):

$$s_m(x_1,\ldots,x_d) = \sum_{i_1}\cdots\sum_{i_d} b_{i_1,\ldots,i_d} B^m_{i_1}(x_1)\cdots B^m_{i_d}(x_d) = \sum_{i_1,\ldots,i_d} b_{i_1,\ldots,i_d} B^m_{i_1}(x_1)\cdots B^m_{i_d}(x_d)\,, \tag{4.16}$$

with $b_{i_1,\ldots,i_d} \in \mathbb{R}$ for $(i_1,\ldots,i_d) \in \mathfrak{I}_d$. Notice that the coefficients $b_{i_1,\ldots,i_d}$, with $(i_1,\ldots,i_d) \in \mathfrak{I}_d$, can be represented as a vector of length $\prod_{j=1}^d I_j$ or, equivalently, as a d-order array (or tensor) with dimensions $I_1 \times \ldots \times I_d$. The partial derivatives of the multivariate spline in eq. (4.16) are given by Schumaker (2007) (and can be easily computed via Algorithm 5.11 in Schumaker (2007)):

$$D^{\ell_1}_{x_1}\cdots D^{\ell_d}_{x_d}[s_k](x_1,\ldots,x_d) = \sum_{i_1,\ldots,i_d} b^{\ell_1,\ldots,\ell_d}_{i_1,\ldots,i_d} B^{m-\ell_1}_{i_1}(x_1)\cdots B^{m-\ell_d}_{i_d}(x_d)\,. \tag{4.17}$$

Finally, in Section 4.3.3 we use the notation $\widetilde{\mathbf{B}} = (\widetilde{\boldsymbol{\beta}}'_{T+1},\ldots,\widetilde{\boldsymbol{\beta}}_{T+H})'$, where $\widetilde{\boldsymbol{\beta}}_{T+h}$ is the forecast for the vector $\widehat{\boldsymbol{\beta}}_T$ at horizon $h = 1,\ldots,H$. The corresponding forecast for the fPCA approximate functions are denoted by $\widetilde{\mathbf{f}}_{T+H}(\cdot) = (\widetilde{f}_{T+1}(\cdot),\ldots,\widetilde{f}_{T+H}(\cdot))'$ whereas $\widetilde{\mathbf{c}}_{T+H}(\cdot) = (\widetilde{c}_{T+1}(\cdot),\ldots,\widetilde{c}_{T+H}(\cdot))'$ denotes the forecast of the copula probability density functions.

4.2.2 Related literature

Given an observed bivariate time series of relevant economic or financial variables $(\mathbf{x},\mathbf{y}) = \{x_{t,i}, y_{t,i})\}_{ti}$, with $i = 1,\ldots,N$ for each $t = 1,\ldots,T$, with unknown time varying copula probability density function $c_t(\cdot)$, the purpose of this methodology is to obtain a h-step ahead forecast the copula pdf $\widetilde{c}_{T+h}(\cdot)$, $h = 1,\ldots,H$. In order to achieve this result, we will borrow some concepts from various streams of literature. The core of the methodology is grounded on functional data analysis (FDA), in particular the technique of functional principal component analysis (fPCA), and on the centred log-ratio isometry between $\mathbb{D}^*(I)$ and $\mathcal{L}^*_2(I)$. Furthermore, we exploit several concepts from the literature on nonparametric estimation of copula functions and, finally, we use standard techniques for multivariate time series analysis.

The use of functional autoregressive processes (FAR) proposed by Bosq (2000) is prevented by the constraints holding on pdfs and the fact that $\mathbb{D}(I)$ is not closed under pointwise addition and multiplication. In situations like this, post-processing techniques are necessary for mapping the output of a given procedure into the desired space. However, this procedure is suboptimal as there are guarantees that all the information is preserved by this mapping. By contrast, an efficient forecasting model for probability density functions should yield a consistent output, that is the predicted function must be pdfs.

In second instance, as functions are infinite-dimensional objects, the original forecasting problem would require to work with infinite-dimensional spaces. Though natural, this brings in a significant degree of complexities that a similar problem in finite-dimensional spaces (i.e., Euclidean spaces). Clearly, a direct matching between an infinite-dimensional problem and a finite-dimensional one, does not exist. Moreover, naïve techniques for moving from an infinite-dimensional problem into one a finite one via discretization of the functions could be too rough and lose too much information. Nonetheless, under suitable assumptions it is possible to approximate the infinite-dimensional functional forecasting problem by a simpler one in which the parameters of interest are finite-dimensional vectors. Moreover, under certain conditions this approximation is optimal (according to a weighted least squares criterion).

In order to avoid post-processing and rough approximations, in Section 4.3.2 we exploit the centred log-ratio isometry between the spaces $\mathbb{D}^*(I)$, $\mathcal{L}_2^*(I)$ and define a factor model for approximating the clr-transformed densities (that is, $\text{clr}(c_t)(\cdot) \approx \check{f}_t(\cdot)$):

$$\check{f}_t(\cdot) = \boldsymbol{\beta}_t' \boldsymbol{\xi}(\cdot) = \sum_{j=1}^{J} \beta_{t,j} \xi_j(\cdot) \,, \tag{4.18}$$

where $\xi_j(\cdot)$ are the principal component functions (or factors) and the coefficients $\beta_{t,j} \in \mathbb{R}$ are the principal component scores, both estimated by means of the functional principal component analysis. The factor model defines an approximation of the functions $\text{clr}(c_t)(\cdot)$ by means of a finite linear combination of common, time-invariant factors with component specific time-varying scores. The optimality criterion given by the quadratic distance $||\text{clr}(c_t)(\cdot) - \check{f}_t(\cdot)||_2$ is minimized by the choice of the principal component functions that maximize the explained variability of the series $\text{clr}(c_t)(\cdot)$, $t = 1, \ldots, T$ (see Ramsay and Silverman (2005)).

Functional data analysis is a growing field of research and the existing results dealing with probability density functions are scarce. A possible interpretation of fPCA, in analogy with multivariate PCA, identifies the principal component functions with the eigenfunctions of their covariance operator of the observed functions. Following this interpretation, Hörmann et al. (2015) provided a remarkable extension of fPCA to time series functional data. They worked on the frequency domain using the techniques of Brillinger (2001) for estimating the dynamic principal component functions, which account for the temporal dependence among functional observations. Unfortunately, their results are not straightforwardly extendible to pdfs.

In fact, when dealing with pdfs, the estimation of the principal component functions $\boldsymbol{\xi}(\cdot)$ poses some issues which call for the development of specific procedures. Egozcue et al. (2006) has proved the analogy between probability density functions and compositional vectors, which are vectors belonging to the n-dimensional unit simplex $\mathcal{S}^n$ representing fractions or proportions and constitute the cornerstone of compositional data analysis (see Aitchison (1986)). Egozcue et al. (2006) interpreted pdfs as infinite-dimensional compositional vectors and translated into the functional domain the main results of compositional data analysis: this includes the definition of the operations of perturbation and powering (analogue of addition and scalar multiplication), $\oplus, \odot$, that make $(\mathbb{D}(I), \oplus, \odot)$ a space. van der Boogaart et al. (2010) and van der Boogaart et al. (2014) proved that $(\mathbb{D}(I), \oplus, \odot)$ is indeed a Hilbert space and showed that the centred log-ratio widely used in compositional data analysis is an isometry (i.e. an isometric isomorphism) between the spaces $\mathbb{D}^*(I), \mathcal{L}_2^*(I)$.

These results opened new possibilities to the functional analysis of pdfs. While it is possible to exploit the operations $\oplus, \odot$ for performing statistical analyses directly on $\mathbb{D}(I)$, the need for re-definition of standard techniques by $\oplus, \odot$ has lead the researchers to prefer the use of isometries. For the sake of working out fPCA of (transformed) univariate pdfs in $\mathcal{L}_2(I)$, Petersen and Müller (2016) proposed two isometries (the log-hazard and the log-quantile transforms) between $\mathbb{D}(K)$ and $\mathbb{D}(K)$, for K a compact subset of $\mathbb{R}$, whereas Hron et al. (2016) exploited the clr map. Other contributions in this area include Salazar et al. (2015), who proposed a forecasting model for univariate pdfs, and Menafoglio et al. (2014), who studied the problem of interpolation via the kriging method for probability density functions. The common strategy consists in three steps: the transformation of the pdfs into a suitable Hilbert subspace of $\mathcal{L}_2(I)$, where the statistical model is defined and the analysis is undertaken. Finally, the use of the inverse of the isometry for mapping the result back into $\mathbb{D}(I)$. Finally, the contribution of Machalovà et al. (2016) (see also Machalovà (2002a), Machalovà (2002b)) is based on the interpretation of fPCA as an eigenproblem, which the authors solved proposing a solution within the class of spline functions (see Appendix D.1 for more details on fPCA and related solution methods). This allowed for the inclusion of the zero integral constraint as a constraint on the coefficients of the basis spline functions.

In this paper we follow van der Boogaart et al. (2014) and use the centred log-ratio transform to map pdfs into the Hilbert space $\mathcal{L}_2^*(I)$. Then, we extend to the multivariate framework the strategy developed by Machalovà et al. (2016) for dealing with the integral constraint for univariate pdfs with compact support, thus obtaining a way to account for the constraint eq. (4.7) in the estimation of the principal component factors and scores.

The most appealing feature of the factor model in eq. (4.18) we specify is that all the information about the temporal dependence between the functions is carried by the scores, which form a vector-valued time series. Therefore, a forecast for the approximated function $\widetilde{f}_{T+h}(\cdot)$ at horizon $h \geq 1$, can be obtained by plugging-in a forecast for the scores, computed by well-known methods (e.g., VAR models). Then we get a forecast for the pdf, $\widetilde{c}_{T+h}(\cdot)$, by simply applying the inverse centred log-ratio map in eq. (4.6).

Our strategy shares some similarities with Liebl (2013) and Hron et al. (2016), but the methodologies differ in some key aspects. First and most important, we are interested in forecasting pdfs, which complicates the analysis with respect to the unrestricted case of Liebl (2013). Moreover, we extend the analysis of Hron et al. (2016) to the bivariate case (though the methodology generalizes easily to multidimensionality). Finally, we provide some remarks about how to deal with the case of densities with unbounded support.

4.3 Methodology

We propose a strategy for estimating the factor model in eq. (4.18), then forecasting the clr-transformed functions $\widetilde{f}_{T+h}(\cdot)$ and the corresponding pdfs $\widetilde{c}_{T+h}(\cdot)$, $h = 1, \dots, H$. The methodology focuses on the forecast of bivariate copula probability density functions, however, the method is general and can be applied without structural changes to general multivariate pdfs with bounded support as well as to pdfs with unbounded support that satisfy an additional constraint (see Section 4.4). The modelling framework can be summarized as follows:

- in Section 4.3.1 we partition the raw dataset in sub-samples corresponding to different periods t, then for each of them we estimate the copula probability density function (or, in the generally case, the multivariate pdf).
  When dealing with copula pdfs, we use the nonparametric density estimator proposed by Chen (1999) for avoiding the boundary bias (for general multivariate pdfs we suggest standard product kernel estimators).

- next, in Section 4.3.2 we estimate the factor model in eq. (4.18) by a modified version of the functional principal component analysis. In this section we combine the centred log-ratio transform and spline functions for estimating the principal component functions and the scores ($\widehat{\boldsymbol{\xi}}(\cdot)$ and $\widehat{\mathbf{B}}$, respectively) such that the resulting functions $\breve{f}_t(\cdot)$ (approximating $\text{clr}(c_t)(\cdot)$) satisfy the restrictions of probability density functions.
  We generalize the strategy proposed by Machalovà et al. (2016) and its application in Hron et al. (2016) to the multivariate (and potentially unbounded) case.

- finally, in Section 4.3.3 we estimate a VAR(p) process for the time series of scores previously estimated and forecast the scores h steps ahead, $h = 1, \dots, H$. Then, we get the forecast of the approximated function $\breve{f}_{T+h}(\cdot)$ and by applying the inverse centred

log-ratio transform we obtain a predicted copula probability density function (or the multivariate pdf) $\widetilde{c}_{T+h}(\cdot)$.
The forecasting strategy extends Liebl (2013) from univariate unconstrained functions to multivariate pdfs.

Algorithm 1 synthetically represents the proposed strategy. Each block is described in detail in the following subsections.

Algorithm 1 Methodology

```
function COPULAESTIM($\mathbf{x}, \mathbf{y}$)
    a) split data $\{x_{t'}, y_{t'}\}_{t'=1}^{TN}$ into $T$ sub-samples $\{x_{t,i}, y_{t,i}\}_{i=1}^{N}$        ▷ data for 2-steps
    for $t = 1, \dots, T$ do
        b) compute pseudo-obs $\{u_{t,i}, v_{t,i}\}_{i=1}^{N}$ from $\{x_{t,i}, y_{t,i}\}_{i=1}^{N}$
        c) estimate copula $\widehat{c}_t(u, v)$        ▷ Beta kernel
        d) compute clr transform of copula values $\mathrm{clr}(\widehat{c}_t)(u_{t,i}, v_{t,i})$        ▷ clr
    end for
    return $(\mathbf{U}, \mathbf{V}, \mathbf{C}) = \{u_{t,i}, v_{t,i}, \mathrm{clr}(\widehat{c}_t)(u_{t,i}, v_{t,i})\}_{t,i}$
end function

function MOD_FPCA($\mathbf{U}, \mathbf{V}, \mathbf{C}, \bar{\boldsymbol{\lambda}}^{u,v}$)
    for $t = 1, \dots, T$ do
        a) $(\mathbf{d}_t, \boldsymbol{\phi}) \leftarrow$ solve constrained optimal smoothing problem $(\mathbf{U}, \mathbf{V}, \mathbf{C}, \bar{\boldsymbol{\lambda}}^{x,y})$
    end for
    b) $(\widehat{\boldsymbol{\xi}}, \widehat{\mathbf{B}}) \leftarrow$ solve eigenproblem $(\mathbf{D}, \boldsymbol{\phi})$
    return $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_T)$
end function

function PREDICTION($\widehat{\mathbf{B}}$)
    a) estimate VAR($p$) for $\{\widehat{\boldsymbol{\beta}}_t\}_t$
    b) forecast scores $\widetilde{\mathbf{B}} = (\widetilde{\boldsymbol{\beta}}_{T+1}, \dots, \widetilde{\boldsymbol{\beta}}_{T+H})$
    c) forecast transformed pdfs $\widetilde{\mathbf{f}}_{T+H}(\cdot) = (\widetilde{f}_{T+1}(\cdot), \dots, \widetilde{f}_{T+H}(\cdot))'$
    d) forecast pdfs $\widetilde{\mathbf{c}}_{T+H}(\cdot) = (\widetilde{c}_{T+1}(\cdot), \dots, \widetilde{c}_{T+H}(\cdot))'$        ▷ inverse clr
    return $\widetilde{\mathbf{c}}_{T+H}(\cdot)$
end function
```

4.3.1 Step 1 - Copula estimation

After the introduction of the empirical copula (Deheuvels (1978), Deheuvels (1979)), which is a nonparametric estimator for the copula cumulative distribution function, several nonparamteric techniques for the estimation of a copula pdf and cdf have been proposed. We follow Chen (1999) and Charpentier et al. (2007) and estimate the copula pdfs from raw data via a product Beta kernel estimator. Among the main advantages of this approach we remark the greater flexibility with respect to parametric methods, the smoothness of the estimated function (as opposed to the empirical copula) and the absence of boundary bias.

Consider a sample of observations $\{x_{t'}, y_{t'}\}_{t'}$ of size T' (for instance, with daily frequency). First of all, we fix the reference period t (i.e., year, quarter) and split the raw sample accordingly into T sub-samples of size N (to be interpreted, for instance, as T years of N daily observations), $\{x_{t,i}, y_{t,i}\}_{i,t}$ of size N, for $t = 1, \dots, T$. The reference period coincides with the frequency of the functional time series we want to analyse, whereas the intra-period observations are interpreted as noisy measurements of the discretized continuous function of interest $c_t(\cdot)$. Consequently, we are going to use the N data points in each period t to estimate the function $c_t(\cdot)$, then we use the resulting functional time series for performing forecasts of the probability density through a modified fPCA algorithm.

We exploit the intra-period information (i.e. N observations, for fixed period t) for estimating a copula pdf for each period, $\widehat{c}_t(\cdot)$. Recall that a copula probability density has uniformly distributed marginals representing the marginal cumulative distributions. As the latter are unknown, it is necessary to estimate them as first step. In practice, we compute the pseudo-observations (see Nelsen (2013), Cherubini et al. (2004)) defined as follows:

$$(u_{t,i}, v_{t,i}) = \left(F_x^N(x_{t,i}), F_y^N(y_{t,i}) \right) . \tag{4.19}$$

The pseudo-observations can be obtained in two ways. One methods consists in estimating the marginals $F_x^N(x_{t,i}), F_y^N(y_{t,i})$ via the empirical cumulative distribution function, then evaluating them at $(x_{t,i}, y_{t,i})$. Alternatively, the pseudo-observations can be obtained directly through the rank transformed data.

We choose the second method, as it is computationally faster and provides distributions closer to the uniform. The rank transformation generates pseudo-observations according to (similarly for y):

$$R_{t,i}^x = \sum_{j=1}^N \mathbb{1}(x_{t,j} \leq x_{t,i}) , \qquad u_{t,i} = \frac{1}{N} R_{t,i}^x . \tag{4.20}$$

Given the pseudo observations, we estimate the copula probability density function by using a nonparametric kernel density estimator obtained as the product of univariate Beta kernels. Given a sample $\{x_t\}_{t=1}^T$, the Beta kernel density estimator (Chen (1999), Charpentier et al. (2007)) is defined as:

$$\widehat{f_m}(x) = \frac{1}{T} \sum_{t=1}^T \mathcal{K}_\beta \left(x_t; 1 + \frac{x}{m}, 1 + \frac{1-x}{m} \right) , \tag{4.21}$$

where $\mathcal{K}_\beta(\cdot; a, b)$ is the pdf of a Beta distribution with parameters (a, b) and m is the bandwidth. Alternative nonparametric methods for the estimation of a probability density function, such as the kernel estimator of Fermanian and Scaillet (2003), do not fit well the current framework because of the inadequateness of the methods to deal with the compact support. This causes a boundary bias problem if an unbounded kernel (such as the Gaussian) is used or a lack of smoothness, if the derivatives of the empirical copula distribution are chosen. Both shortcomings are instead solved by the product Beta kernel estimator. The price to pay is the lack of adequate rules of thumb for the specification of the bandwidth, which must be tuned case-by-case. The estimated smooth functions $(\widehat{c}_1(\cdot), \ldots, \widehat{c}_T(\cdot))$ are used to compute the values of the copula function at specific couples of pseudo-observations, that is $\widehat{c}_t(u_{t,i}, v_{t,j})$, for $i, j = 1, \ldots, N$, $t = 1, \ldots, T$. Finally, we apply the clr transform in eq. (4.5) to obtain $\mathbf{C}_t = (\mathrm{clr}(\widehat{c}_t)(u_{t,i}, v_{t,j}))_{ij}$, for $i, j = 1, \ldots, N$ and $t = 1, \ldots, T$. In compact notation, denote the matrices of pseudo-observations $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_T)$, $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_T)$ and the associated collection of matrices of clr-transformed copula pdf values $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_T]$. The series of matrices $\mathbf{C}$ is required for estimating the constrained spline functions in Section 4.3.2.

In the general case, when the interest lies on multivariate pdfs with unbounded support, we propose to estimate the density via product kernel estimators, with standard choices of the univariate kernels such as Gaussian or Epanechnikov.

Concerning the interpretation of the method, we make the following remarks:

- functions are infinite-dimensional objects, thus from a computational perspective it is impossible to deal with them directly, but a discretization step is in order. Functional data in a strict sense do no exist, instead available data can be defined as noisy observations of discretized functions. Each discretized functional data point, broadly speaking, consists of a pair of location and value of the function at location (where location has no particular meaning). For instance, for univariate functions the location is the point on the x-axis, whereas the value at the location is the corresponding value of the function (on the y-axis). Therefore, in the current framework, we may think of the N rank transformed observations $\{u_{t,i}, v_{t,i}\}_i$, for a given t, as a set of location points whereas $\widehat{c}_t(u_{t,i}, v_{t,i})$ represents the value of the copula function at those locations (i.e., a discretized version of the underlying smooth function). We remark that this is a standard interpretation in functional data analysis (e.g., see Ramsay and Silverman (2005)) and is unrelated to the procedure developed here.

- From a financial point of view, the copula pdfs are a flexible instrument providing all the information about the dependence between the marginal series $\mathbf{x}_t = \{x_{t,i}\}_i$ and $\mathbf{y}_t = \{y_{t,i}\}_i$ at time t. They are remarkably richer than a single scalar parameter: in addition to the extreme cases of independence (corresponding to a product copula) and perfect dependence (diagonal copula), they permit to study several particular forms of dependence, such as tail dependence (i.e. the probability of comovements of the variables in the upper/lower tail of the distribution, see Joe (1997), Nelsen (2013)).
  From this perspective, the availability of a (estimated) time series of copula pdfs permits to have information on different forms of dependence across several periods. Instead of limiting to a descriptive analysis on the variation of (finite-dimensional) synthetic statistics built from each function $c_t(\cdot)$, we aim at characterizing how the whole dependence pattern evolves over time.

4.3.2 Step 2 - Modified fPCA

Starting from $\mathbf{C}$, the time series of clr-transformed pdfs values estimated in Section 4.3.1, our goal in this section is to estimate the factor model in eq. (4.18) using the tools from functional principal component analysis (fPCA). In words, we estimate the function $\check{f}_t(\cdot)$ that approximates the centred log-ratio transform of the pdf $c_t(\cdot)$, for $t = 1, \dots, T$. In this section we are considering bivariate copula pdf, whose support is compact $[0,1]^d$, with $d = 2$. See Section 4.4 for a discussion about the general frameworks when the pdfs have unbounded support or are multivariate with $d > 2$. The strategy does not impose any assumption except that the decay of the pdf at infinity must be such that its logarithm is square integrable. Moreover, given that probability density functions represent a special case of constrained functions, the proposed methodology can be applied as well for forecasting multivariate square integrable functions.

The outcome of this step is a vector of (time invariant) estimated factors $\widehat{\boldsymbol{\xi}}(\cdot)$ and a vector-valued time series of scores $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_T)$ which will be used in Section 4.3.3 for building a forecast of the pdf $\widetilde{c}_{T+h}(\cdot)$, with $h = 1, \dots, H$. Appendix D.1 provides a summary of the results from functional data analysis used in this paper, we refer to Ramsay and Silverman (2005), Ferraty and Vieu (2006) for a more detailed presentation.

We present the outline of the strategy and the results, referring to Appendix D.2 for detailed computations. Ordinary fCPA is designed for the analysis of unconstrained functions, however in our framework the object of interest are pdfs, that is functions constrained to be positive on their support and to have unit integral. This calls for a modification of standard fPCA in order to account for the constraints without the need to post-process the output. We propose a strategy for addressing this issue consisting in the exploitation of the centred log-ratio transform and spline functions. The clr transform allows the analysis to be carried out in the space $\mathcal{L}_2^*(I)$, which is preferred over $\mathbb{D}^*(I)$ due to its nicer properties that make easier ordinary calculus. Then, we are left with the estimation of the factor model in eq. (4.18), which we interpret as an eigenproblem. A first approach consists in the discretization of the functions involved and the solution of the resulting multivariate problem: despite being intuitive, this approach easily breaks down as the dimension increases because of the number of points necessary for providing a good discrete grid. Instead, we choose to express both the target function to be approximated by the factors and the factors themselves by a finite linear combination of pre-specified basis functions. This implicitly reduces the infinite-dimensional problem to an eigenproblem for the vector coefficients of the basis expansion. Following Machalovà et al. (2016), we choose a B-spline basis as it allows to analytically solve the resulting eigenproblem taking into account the integral constraint in eq. (4.7).

More formally, we propose to estimate the factor model in eq. (4.18) by interpreting the functions $\boldsymbol{\xi}(\cdot)$ as the eigenfunctions of the covariance operator of the functions $\check{\mathbf{f}}(\cdot)$, denoted G. For each period $t = 1, \ldots, T$ and $k = 1, 2, \ldots$, this yields the eigenproblem:

$$\begin{cases} \int G(\mathbf{x}, \mathbf{y}) \xi_k(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \beta_{t,k} \xi_k(\mathbf{y}) & \text{(4.22a)} \\ \int \xi_k(\mathbf{x}) \xi_k(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1 & \text{(4.22b)} \end{cases}$$

subject to the additional constraints $\langle \xi_k, \xi_j \rangle = 0$, for $k \neq j$ and $\int \xi_j \, \mu(\mathrm{d}\mathbf{x}) = 0$ for $j = 1, 2, \ldots$. Then, we look for a solution within the class of tensor product, bivariate spline functions (see (Ramsay and Silverman, 2005, ch.8) for a review of alternative solution methods), which allows to include the zero integral constraint as a linear constraint on the coefficients of the basis spline functions, thanks to the relation between splines with their derivatives.

Since a spline function can be expressed as a linear combination of known basis B-splines (see Section 4.2.1 for the notation), we need to solve a finite dimensional optimization problem for the coefficient vector of the spline. The constrained optimal smoothing problem, for each period $t = 1, \ldots, T$, is:

$$\begin{cases} \min_{s_m} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \left[s_m^{(\ell_1, \ell_2)}(u, v) \right]^2 \mathrm{d}v \, \mathrm{d}u + \alpha \left[\sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j} \left(\mathrm{clr}(\widehat{c}_t)(u_{t,i}, v_{t,j}) - s_m(u_{t,i}, v_{t,j}) \right)^2 \right] \\ \text{s.t.} \int_{a_1}^{b_1} \int_{a_2}^{b_2} s_m(u, v) \, \mathrm{d}v \, \mathrm{d}u = 0 \end{cases} \tag{4.23}$$

where $s_m(\cdot, \cdot)$ is a spline of degree m, ℓ_1, ℓ_2 are the degree of the partial derivatives with respect to u, v, respectively, $\{u_{t,i}, v_{t,j}\}_{ij}$ with $i, j = 1, \ldots, N$, are the evaluation points and $\{\mathrm{clr}(\widehat{c}_t)(u_{t,i}, v_{t,j})\}_{ij}$ is the corresponding value of the clr-transformed pdf. Notice that N is number of observations allocated to each period $t = 1, \ldots, T$. $\{w_{i,j}\}_{ij}$ is a sequence of point-specific weights, whereas α is the global weight of the least squares component in the smoothing problem. Finally, the interval $(a_1, b_1) \times (a_2, b_2)$ is support of the original function and of the spline. In the following we assume: $a_1 = a_2 = 0$, $b_1 = b_2 = 1$, $\ell_1 = \ell_2 = 2$, meaning that we look for a solution in the class of cubic splines on the interval $[0, 1]^2$. Moreover, we consider an extended knot sequence given by the regular grid $\bar{\boldsymbol{\lambda}}^{u,v} = \bar{\boldsymbol{\lambda}}^{u} \otimes \bar{\boldsymbol{\lambda}}^{v}$, with:

$$\bar{\boldsymbol{\lambda}}^{u} = (\lambda^u_{-m}, \lambda^u_{-m+1}, \ldots, \lambda^u_{g+m+1})', \tag{4.24a}$$
$$\bar{\boldsymbol{\lambda}}^{v} = (\lambda^v_{-m}, \lambda^v_{-m+1}, \ldots, \lambda^v_{g+m+1})' \tag{4.24b}$$

with:

$$\lambda^u_{-m} = \ldots = \lambda^u_0 < \ldots < \lambda^u_{g+1} = \ldots = \lambda^u_{g+m+1} \tag{4.25a}$$
$$\lambda^v_{-m} = \ldots = \lambda^v_0 < \ldots < \lambda^v_{g+1} = \ldots = \lambda^v_{g+m+1} \,. \tag{4.25b}$$

This is a square grid with the same knots along both directions (that is, the x-axis and the y-axis, respectively), however we may choose a different number of interpolation knots for each dimension. We have decided to use the same number of knots and the same location because we are interpolating a copula probability density function with support $[0,1]^2$.

Lemma 4.3.0.1
Define $\phi_k^{m+1}(\cdot)$, $k = 1,\dots,K$ the B-spline basis functions of order m. The optimal spline function solving the problem in eq. (4.23) is given by:

$$s_m(u,v) = \mathbf{C}^{m+1}(u,v)\mathbf{d} = \sum_{k=1}^{K} d_k \psi_k^{m+1}(u,v)\,. \tag{4.26}$$

See Appendix D.2.1 for the detailed computations.

The spline functions in eq. (4.26) represent an interpolated multivariate probability density function, with evaluation points $(u_{t,i}, v_{t,i})_i$ and values $\mathrm{clr}(\widehat{c}_t)(u_{t,i}, v_{t,i})$, for $i = 1,\dots,N$. By repeating this procedure for each sub-sample $(\mathbf{u}_t, \mathbf{v}_t)_t$, with $t = 1,\dots,T$, we end up with a series of T multivariate spline functions satisfying the zero integral constraint. With a slight abuse of notation, define $\breve{f}_t(\cdot) = s_m(\cdot)$ the spline in eq. (4.26) estimated using the sub-sample $(\mathbf{u}_t, \mathbf{v}_t)_t$, for each period $t = 1,\dots,T$. Therefore, we can write in compact notation:

$$\breve{f}_t(\cdot) = \sum_{k=1}^{K} d_{t,k} \psi_k^{m+1}(\cdot) = \mathbf{d}_t' \boldsymbol{\psi}(\cdot)\,, \tag{4.27}$$

where $\mathbf{d}_t = (d_{t,1},\dots,d_{t,K})'$ and $\boldsymbol{\psi}(\cdot) = (\psi_1^{m+1}(\cdot),\dots,\psi_K^{m+1}(\cdot))'$. It is now possible to solve the eigenproblem in eq. (4.22a) using the same B-spline functions $\boldsymbol{\psi}(\cdot)$ as a basis for the principal component functions $\xi_j(\cdot)$, $j = 1,2,\dots$:

$$\xi_j(\cdot) = \sum_{k=1}^{K} a_{j,k} \psi_k^{m+1}(\cdot) = \mathbf{a}_j' \boldsymbol{\psi}(\cdot)\,, \tag{4.28}$$

where $\mathbf{a}_j = (a_{j,1},\dots,a_{j,K})'$. From this basis expansion, the infinite-dimensional eigenproblem in eq. (4.22a) reduces to a finite-dimensional optimization problem for the coefficient vectors $\mathbf{a}_j$, for $j = 1,\dots,J$. For selecting the number of principal components J, we sort the estimated eigenvalues in decreasing order and compute the proportion of total variability explained by $v_j = \rho_j / \sum_k \rho_k$, for $j = 1,2,\dots$. Then, we retain the first J factors accounting for a given share $\bar{d}$ of the total variability, that is $J = \arg\min_j \{\sum_j v_j \geq \bar{d}\}$. The solution of this multivariate eigenproblem is obtained by first finding the optimal $\mathbf{u}_j$ satisfying (see Appendix D.2.2 for detailed computations):

$$T^{-1}\mathbf{M}^{1/2}\mathbf{D}'\mathbf{D}\mathbf{M}^{1/2}\mathbf{u}_j = \rho_j \mathbf{u}_j\,, \tag{4.29}$$

then transforming $\widehat{\mathbf{a}}_j = \mathbf{M}^{1/2}\widehat{\mathbf{u}}_j$, for $j = 1,2,\dots$. The solution of eq. (4.29) yields an estimate of the principal component functions by plugging $\widehat{\mathbf{a}}_j$ in eq. (4.28):

$$\widehat{\xi}_j(\cdot) = \widehat{\mathbf{a}}_j' \boldsymbol{\psi}(\cdot)\,. \tag{4.30}$$

Since the eigenvectors are not uniquely identified, we follow Liebl (2013) and transform them by applying the VARIMAX orthonormal rotation (see Kaiser (1958), Abdi (2003)). The eigenvalues provide an estimate for the scores $\widehat{\boldsymbol{\beta}}_t = (\widehat{\beta}_{t,1},\dots,\widehat{\beta}_{t,J})'$, for each period $t =$

$1, \dots, T$. This coincide with (see Ramsay and Silverman (2005)):

$$\widehat{\boldsymbol{\beta}}_t = \begin{bmatrix} \langle \widehat{\xi}_1, \widehat{\xi}_1 \rangle & \dots & \langle \widehat{\xi}_1, \widehat{\xi}_J \rangle \\ \vdots & \ddots & \vdots \\ \langle \widehat{\xi}_J, \widehat{\xi}_1 \rangle & \dots & \langle \widehat{\xi}_J, \widehat{\xi}_J \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle \check{f}_t, \widehat{\xi}_1 \rangle \\ \vdots \\ \langle \check{f}_t, \widehat{\xi}_J \rangle \end{bmatrix} . \tag{4.31}$$

As final output of this step we obtain the estimated time series of scores $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_T)$.

Each estimated eigenfunction can be seen as a continuous function of the clr-transformed functions, that is $\widehat{\xi}_j(\cdot) = g(\check{f}_1(\cdot), \dots, \check{f}_T(\cdot))$. Hence, by the continuous mapping theorem, the estimator of the eigenfunction is consistent provided that the estimators for the clr-transformed functions are consistent too. Recall that each $\check{f}_t$ corresponds to the centred log-ratio (continuous and smooth) transformation of a copula pdf, and it is estimated via a spline. It is known (see Schumaker (2007)) that splines approximate arbitrarily well continuous smooth functions on a bounded interval.

Consequently, from the consistency of splines in approximating a smooth function (as is $\check{f}_t(\cdot), t = 1, \dots, T$ in our case) it descends the consistency of the estimator for each eigenfunction $\widehat{\xi}_j(\cdot)$ and, by another application of the continuous mapping theorem, the consistency of the estimator of the associated scores $\widehat{\boldsymbol{\beta}}_t$.

4.3.3 Step 3 - Prediction

In this last step, we aim at obtaining a H steps ahead forecast $\widetilde{c}_{T+H}(\cdot)$ of the pdf $c_t(\cdot)$. The task is accomplished in three steps: first, we estimate a VAR(p) process on the time series of estimated principal component scores from Section 4.3.2, $\{\widehat{\boldsymbol{\beta}}_t\}_{t=1}^T$, then we use the fitted values for obtaining a forecast of the scores $\widetilde{\boldsymbol{\beta}}_{T+h}$, $h = 1, \dots, H$. Next, for $h = 1, \dots, H$ we derive a forecast for the approximated function $\widetilde{f}_{T+h}(\cdot)$ by plugging-in eq. (4.18) and finally we get the forecast of the pdf $\widetilde{c}_{T+h}(\cdot)$ by applying the inverse clr transform to $\widetilde{f}_{T+h}(\cdot)$.

The estimated scores from the Section 4.3.2 for a vector-valued time series, where each vector has length J. We propose to model the time series through a VAR(p), as follows:

$$\widehat{\boldsymbol{\beta}}_t = \boldsymbol{\phi}_{const} + \boldsymbol{\phi}_{trend} t + \sum_{l=1}^{p} \boldsymbol{\Phi}_l \widehat{\boldsymbol{\beta}}_{t-l} + \boldsymbol{\epsilon}_t , \qquad \boldsymbol{\epsilon}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_J) . \tag{4.32}$$

Denoting the estimated coefficients by $(\widehat{\boldsymbol{\phi}}_{const}, \widehat{\boldsymbol{\phi}}_{trend}, \widehat{\boldsymbol{\Phi}}_1, \dots, \widehat{\boldsymbol{\Phi}}_p)$, we perform forecasts for each $h = 1, \dots, H$ steps ahead in the usual way:

$$\widetilde{\boldsymbol{\beta}}_{T+h} = \widehat{\boldsymbol{\phi}}_{const} + \widehat{\boldsymbol{\phi}}_{trend}(T+h) + \sum_{l=1}^{p} \widehat{\boldsymbol{\Phi}}_l \widehat{\boldsymbol{\beta}}_{T-l} . \tag{4.33}$$

Then, we obtain the predicted clr-transformed function $\widetilde{f}_{T+h}(\cdot) \in \mathcal{L}_2^*(I)$ by substituting $\widetilde{\boldsymbol{\beta}}_{T+h}$ and the estimated principal components $\widehat{\boldsymbol{\xi}}(\cdot)$ into eq. (4.18), thus obtaining for $h = 1, \dots, H$:

$$\widetilde{f}_{T+h}(\cdot) = \widetilde{\boldsymbol{\beta}}_{T+h}' \widehat{\boldsymbol{\xi}}(\cdot) = \sum_{j=1}^{J} \widetilde{\beta}_{T+h,j} \widehat{\xi}_j(\cdot) . \tag{4.34}$$

Finally, in order to compute the predicted probability density function $\widetilde{c}_{T+h}(\cdot) \in \mathbb{D}^*(I)$ we apply the inverse centred log-ratio transformation, for $h = 1, \ldots, H$:

$$\widetilde{c}_{T+h}(\cdot) = \operatorname{clr}^{-1}(\widetilde{f}_{T+h})(\cdot) = \frac{\exp\left\{\widetilde{f}_{T+h}(\cdot)\right\}}{\int \exp\left\{\widetilde{f}_{T+1}(\cdot)\right\}}. \tag{4.35}$$

The final outcome of the whole procedure is the set of forecasts of the multivariate pdf $\widetilde{\mathbf{c}}_{T+H}(\cdot) = (\widetilde{c}_{T+1}(\cdot), \ldots, \widetilde{c}_{T+H}(\cdot))'$.

The size of the VAR process in eq. (4.32) corresponds to the number of principal components selected in the fPCA, J and is generally small. Therefore, the dimensionality of the VAR does not hamper the estimation procedure even though the length T of the time series is not really long. This is a consequence of the dimensionality reduction brought by the fPCA, interpreted as a factor model here.

Nonetheless, in higher-dimensional settings it may be still possible to estimate the coefficient matrix in eq. (4.32) by adding a regularization term. The recent contributions, Nicholson et al. (2016) and Nicholson et al. (2017) designed and implemented[1] several types of penalized regression for large VARX models (including the LASSO case) allowing up to $d = 130$ marginal series.

As regards the numerical implementation of the procedure, the core of the proposed methodology relies on standard linear algebra operations, for which computationally efficient algorithms are available. Moreover, the dimensionality reduction brought by the fPCA has the additional advantage of reducing the size of the coefficient matrix of the VAR process do be estimated. Overall, the entire procedure represented in Algorithm 1 is quite fast (see the details for the application in Section 4.5).

4.4 Extensions

Here we briefly discuss some possible extensions of the methodology discussed in Section 4.3.

4.4.1 Unbounded support

The results in van der Boogaart et al. (2010), van der Boogaart et al. (2014) hold also for pdfs with unbounded support, provided that they are absolutely continuous with respect to a measure with finite total mass. This requirement is a direct consequence of the formula for the centred log-ratio in eq. (4.5), which involves at the denominator the total mass of the support. In fact, the problem when dealing with pdfs defined on an unbounded region is that the Lebesgue measure of the whole domain is not finite, hence it would be necessary to choose a different, finite the reference measure of the spaces $\mathbb{D}(I), \mathbb{D}^*(I), \mathcal{L}_2(I), \mathcal{L}_2^*(I)$. If the new reference measure ν is absolutely continuous with respect to the Lebesgue measure, i.e. $\mathrm{d}\nu = g(\cdot)\,\mathrm{d}\mu$, then for $h(\cdot) \in \mathbb{D}_\nu(I)$ it holds $f(\cdot) = h(\cdot)g(\cdot) \in \mathbb{D}_\mu(I)$. Therefore in the particular case $\nu \ll \mu$ performing the analysis of the original pdf series $h_t(\cdot)$ under the reference measure ν is equivalent to perform the analysis of the modified series $h_t(\cdot)g(\cdot)$ under the Lebesgue measure.

Example 4.4.1 (Alternative reference measure)
Let $I = \mathbb{R}^n$ and let $g = \mathrm{d}\mathbb{P}_\mathcal{N}/\mathrm{d}\mu$ to be the Radon-Nikodym derivative of the finite standard Gaussian measure $\mathbb{P}_\mathcal{N}$ with respect to the n-dimensional Lebesgue measure on $\mathbb{R}^n$ (thus, g is the pdf of a

[1] Estimation can be carried out using the `R` (https://cran.r-project.org) package "`BigVAR`" (https://cran.r-project.org/web/packages/BigVAR/index.html), see Nicholson et al. (2017).

standard normal distribution), the change of measure yields:

$$\int_{\mathbb{R}^n} f(\mathbf{x})\, \mathrm{d}\mu = \int_{\mathbb{R}^n} f(\mathbf{x}) \frac{\mathrm{d}\mu}{\mathrm{d}\mathbb{P}_{\mathcal{N}}} \mathrm{d}\mathbb{P}_{\mathcal{N}} = \int_{\mathbb{R}^n} \frac{f(\mathbf{x})}{g(\mathbf{x})}\, \mathrm{d}\mathbb{P}_{\mathcal{N}} = \int_{\mathbb{R}^n} h(\mathbf{x})\, \mathrm{d}\mathbb{P}_{\mathcal{N}}\,.$$

If $\log(h)(\mathbf{x})$ *is square integrable, then all the previous results can be applied, since* $\mathbb{P}_{\mathcal{N}}(\mathbb{R}^n) = 1$*. If instead* $I = \mathbb{R}_+$ *one may use the measure* ω *induced by a Gamma distribution, since* $\omega(\mathbb{R}_+) = 1$.

Lemma 4.4.0.1 (seevan der Boogaart et al. (2014))
Let $\eta \in \mathbb{D}^*(I)$ *be a probability measure with unbounded support* I *and density* $f(\cdot) = \frac{\mathrm{d}\eta}{\mathrm{d}\nu}(\cdot)$ *with respect to the reference measure* ν*, with* $\nu(I) = \infty$*. If* $\exists\, \mu$ *measure such that:*

(i) $\mu \ll \nu$ *and* $\nu \ll \mu$*, with density* $g(\cdot) = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\cdot)$*;*

(ii) $\mu(I) < \infty$*;*

(iii) $\log(f/g)(\cdot)$ *is* μ*-integrable;*

then:

- $\nexists\, \mathrm{clr}_\nu(\eta)$*;*
- $\mathrm{clr}_\mu(\eta)$ *exists and is equal to*

$$\begin{aligned} \mathrm{clr}_\mu(\eta)(\cdot) &= \log\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\right)(\cdot) - \frac{1}{\mu(I)} \int_I \log\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\right)(\mathbf{y})\, \mu(\mathrm{d}\mathbf{y}) \\ &= \log\left(\frac{\mathrm{d}\eta}{\mathrm{d}\nu}\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)(\cdot) - \frac{1}{\mu(I)} \int_I \log\left(\frac{\mathrm{d}\eta}{\mathrm{d}\nu}\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)(\mathbf{y})\, \mu(\mathrm{d}\mathbf{y}) \\ &= \log(f/g)(\cdot) - \frac{1}{\mu(I)} \int_I \log(f/g)(\mathbf{y})\, \mu(\mathrm{d}\mathbf{y})\,. \end{aligned} \tag{4.36}$$

Example 4.4.2 (Clr with unbounded support)
Let $p_0 = \mathrm{d}\mathbb{P}_{\mathcal{N}}/\mathrm{d}\mu$ *be the density of the standard Gaussian measure with respect to the Lebesgue measure on* $\mathbb{R}$*. Let* ν *be a measure and* $p_\nu = \mathrm{d}\nu/\mathrm{d}\mu$ *be its density with respect to the Lebesgue measure. Let* $g = \mathrm{d}\nu/\mathrm{d}\mathbb{P}_{\mathcal{N}}$ *be the density of* ν *with respect to the Gaussian measure. Since* $\mu(\mathbb{R}) = \infty$*, the centred log-ratio for* g *is not defined. However, by changing measure from* μ *to* $\mathbb{P}_{\mathcal{N}}$ *we obtain:*

$$\begin{aligned} \mathrm{clr}(g)(\cdot) &= \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mathbb{P}_{\mathcal{N}}}\right)(\cdot) - \frac{1}{\mathbb{P}_{\mathcal{N}}(\mathbb{R})} \int_{\mathbb{R}} \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mathbb{P}_{\mathcal{N}}}\right)(u)\, \mathrm{d}\mathbb{P}_{\mathcal{N}}(u) \\ &= \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\mathbb{P}_{\mathcal{N}}}\right)(\cdot) - \int_I \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\mathbb{P}_{\mathcal{N}}}\right)(u)\, \frac{\mathrm{d}\mathbb{P}_{\mathcal{N}}}{\mathrm{d}\mu}\, \mathrm{d}\mu(u) \\ &= \log\left(\frac{p_\nu}{p_0}\right)(\cdot) - \int_I \log\left(\frac{p_\nu}{p_0}\right)(u)\, p_0(u)\, \mathrm{d}\mu(u)\,. \end{aligned} \tag{4.37}$$

Notice that the integral on the last line is an expectation with respect to the probability measure $\mathbb{P}$*, also Monte Carlo methods for numerical integration can be applied if the density* p_0 *can be easily sampled from, as is, for example, when* p_0 *is the pdf of a normal distribution.*

Once a new reference measure has been chosen and the clr transform has been applied accordingly, the unbounded support is no more of concern for the methodology. In fact, the B-spline basis functions are defined also on unbounded regions and are computed for a given, finite knot sequence. The location of the knots would depend on the fatness of the tails of the densities, since fatter tails would require the knot sequence to be more scattered for having the resulting spline interpolating well the pdf. For example, a standard normal random variable has unbounded support, but almost the 95% of the mass in the interval $[-2, 2]$.

Consequently, the unboundedness of the support of the pdfs affect the spaces $\mathbb{D}_\nu^*(I)$, $\mathcal{L}_{2,\nu}^*(I)$ to which the functions belong, but does not require a modification of the other parts of the procedure, since the basic constructions behind the result in eq. (4.26) are left unchanged.

4.4.2 Multivariate case: $d > 2$

The proposed methodology can be easily extended to deal with d-dimensional ($d > 2$) probability density functions. The change would be involve the size of the sparse block diagonal matrices described in Appendix D.2.

The only concern that arises when $d > 2$ is the curse of dimensionality, as is typical in nonparametric statistics. In the proposed model this occurs through the need for an increasingly high number of observations $\{x_{1,t,n}, \ldots, x_{d,t,n}\}_n$ for each period t in order to provide a good kernel estimation of the copula probability density function. In addition, if the high dimension is associated to a high degree of complexity of the dependence structure, it may be necessary also to increase the number of principal components to keep, J. This in turn results in a higher dimensionality of the VAR model for the scores in Section 4.3.3. However, we do not expect this to be a significant obstacle, as compared to the previous issue which represents the true bottleneck to high-dimensional applications.

4.5 Application

The dataset is composed by daily observations of S&P500 and NASDAQ indices from 1st January 1980 to 31st December 2017, for a total of $10,032$ observations over 38 years. We make the following assumptions. We start by taking first differences of the two series in order to remove non-stationarity, then for each period $t = 1, \ldots, T$ we assume to observe a sample $\{x_{t,i}, y_{t,i}\}_{i=1}^N$, with $N = 247$, of intra-period observations $(x_{t,i}, y_{t,i}) \in I = [0, 1]^2$. We compute the copula pseudo-observations $(u_{t,i}, v_{t,i}) = (F_x^N(x_{t,i}), F_y^N(y_{t,i}))$, $i = 1, \ldots, N$, for each $t = 1, \ldots, T$ via the rank of the observations. Then, the empirical copula probability density function is estimated non-parametrically with the Beta kernel density estimator (Charpentier et al. (2007), Chen (1999)), using a diagonal bandwidth. The choice of the bandwidth for Beta kernel estimators is tricky since no rules of thumb are available for its optimal choice. We choose $m = 0.0251$, which is the mean of the optimal bandwidths (one for each period $t = 1, \ldots, T$) obtained by minimizing the least squares cross validation criterion (see Silverman (1986), Wand and Jones (1994)) in each period $t = 1, \ldots, T$. See Appendix D.4 for the results using different values of m.

The choice of this splitting of the sample into $T = 38$ years allows us to estimate the function $c_t(\cdot)$, for each t, using up to $N = 248$ data points, while keeping a time series of estimated functions of length $T = 38$, thus providing a good balance of the data between the intra-period and the temporal dimensions. We used the augmented Dickey-Fuller test for testing the null hypothesis of the presence of a unit root in each of the series, resulting in the non rejection of the null both when the whole sample is considered, and with reference to each period t. Therefore, we take first differences of the raw data (see Appendix D.3 for additional plots), thus reducing the size of each sub-sample to $N = 247$.

In the following we assume the stationarity of the estimated copula pdfs. We are not aware of statistical procedures for testing the stationarity of multivariate functional time series. The closest approach by Horváth et al. (2014) proposes a test for univariate functional time series. We leave as future work the testing of the assumption of stationarity for the time series of multivariate functions, eventually by extending the results put forward by Horváth et al. (2014).

Lags	Model A	Model B	Model C	Model D
1	1823.5	1799.6	1821.5	1797.8
2	1831.7	1804.0	1827.0	1799.3
3	1834.2	1813.6	1826.3	1804.4
4	1820.7	1804.2	1801.2	**1769.1**

TABLE 4.2: BIC for different VAR specifications of the VAR(p) model in eq. (4.32). Model A: no constant, no trend; model B: constant, no trend; model C: trend, no constant; model D: constant and trend. The best model according to BIC is in bold.

We choose the following values of the parameters:

$$\begin{array}{ccccc} T = 38 & N = 247 & H = 10 & m = 0.0251 & \bar{d} = 0.92 \\ g = 4 & m = 3 & \ell = 2 & \alpha = 0.8 & \mathbf{W} = \mathbf{I}_{N^2} \end{array} \tag{4.38}$$

After having estimated the copula pdfs $\widehat{c}_1(\cdot), \ldots, \widehat{c}_{38}(\cdot)$, we de-meaned them using the perturbation and powering operations defined in Section 4.2, obtaining $\widehat{\widehat{c}}_t(\cdot) = \widehat{c}_t(\cdot) \ominus \bar{c}(\cdot)$, where $\bar{c}(\cdot) = 1/T \odot \bigoplus_{t=1}^{T} \widehat{c}_t(\cdot)$, which has been used as input for the step 2 of Algorithm 1.

The number of eigenfunctions to take has been estimated as described in Section 4.3.2, by $J = \arg\min_j \{\sum_j \widehat{\rho}_j \geq \bar{d}\}$, yielding $J = 4$. Values of α lower (greater) than unity imply higher (lower) relative weight of the smoothing component with respect to the least squares in the constrained optimal smoothing problem in eq. (4.23). We found that $\alpha = 0.8$ provides a good balance between the two. As robustness check, we performed the analysis with different values of $\bar{d}$ (thus implying different number of eigenfunctions J) without significant changes. We run Algorithm 1 on an Apple MacBookPro with a 3.1GHz Intel Core i7 processor, RAM 16GB, using MATLAB r2017b without exploiting parallel calculus. This required around ten minutes of computation, with step 1 being the most computational intensive part.

The value of the BIC for several specifications of the VAR(p) model in eq. (4.32) for the time series of scores are reported in Table 4.2 and suggest to choose a VAR(4) model including a constant and a time trend. All the estimated VAR models are stationary.

For comparing the results, the estimated copula pdfs $\widehat{c}_t(\cdot)$, $t = 1, \ldots, T$ (respectively, the forecasted copula pdfs $\widetilde{c}_{T+h}(\cdot)$, $h = 1, \ldots, H$) have been computed by applying the inverse clr map to the functions $\widehat{\widetilde{f}}_t(\cdot)$ ($\widetilde{\widetilde{f}}_{T+h}(\cdot)$) estimated (forecasted) from the factor model in eq. (4.18), using J eigenfunctions. Fig. 4.1 shows the contour plot of the time series of the estimated bivariate copula pdfs $\widehat{c}_t(\cdot)$, $t = 1, \ldots, T$, whereas Fig. 4.2 reports the contour and 3D density plots of the forecasted pdfs $\widetilde{c}_{T+h}(\cdot)$, $h = 1, \ldots, H$. We found that:

- there is evidence of significant temporal changes of the estimated pdfs $\widehat{c}_t(\cdot)$. Periods (i.e. years) where the joint probability is concentrated around the bottom-left corner, meaning stronger lower tail dependence, alternate with periods where also upper tail dependence appears. There are two main implications of this stylized fact:
  - it signals that none of the copula families considered in Table 4.1, which are the most commonly used in econometrics, is able to account for the varying dependence over the whole time span of the sample, not even by letting the copula parameter vary over time. The reason is that all of them have either only one type of tail dependence (upper or lower), or both but in symmetric way. Moreover, the same conclusion holds even if a dynamic copula model is specified by allowing the copula parameter to change over time.
  - it is consistent with the results of Guégan and Zhang (2010) discussed in Section 4.1, who found that a dynamic copula model for the whole sample is not satisfactory and that different parametric copula families should be used for modelling different temporal windows.

- Fig. D.2 in Appendix D.3 shows the time series of the fPCA scores along with their forecasts (with 95% confidence intervals). For all series we do not reject the null hypothesis of stationarity (using the ADF test). Moreover, by comparing Fig. 4.1 with Fig. D.2 we find that smooth evolutions of the fPCA scores of the clr-transformed pdfs are able to generate significant changes of the pdf.
- the forecasts of the bivariate copula pdf in Fig. 4.2 are smoothly varying over the forecasting horizon. The forecasts seem to be able to generate heterogeneous tail dependence patterns, as is the case for the observed data. Consequently, we find that the proposed methodology is able to provide non-flat forecasts which can capture and describe the temporal evolution of the bivariate time series.

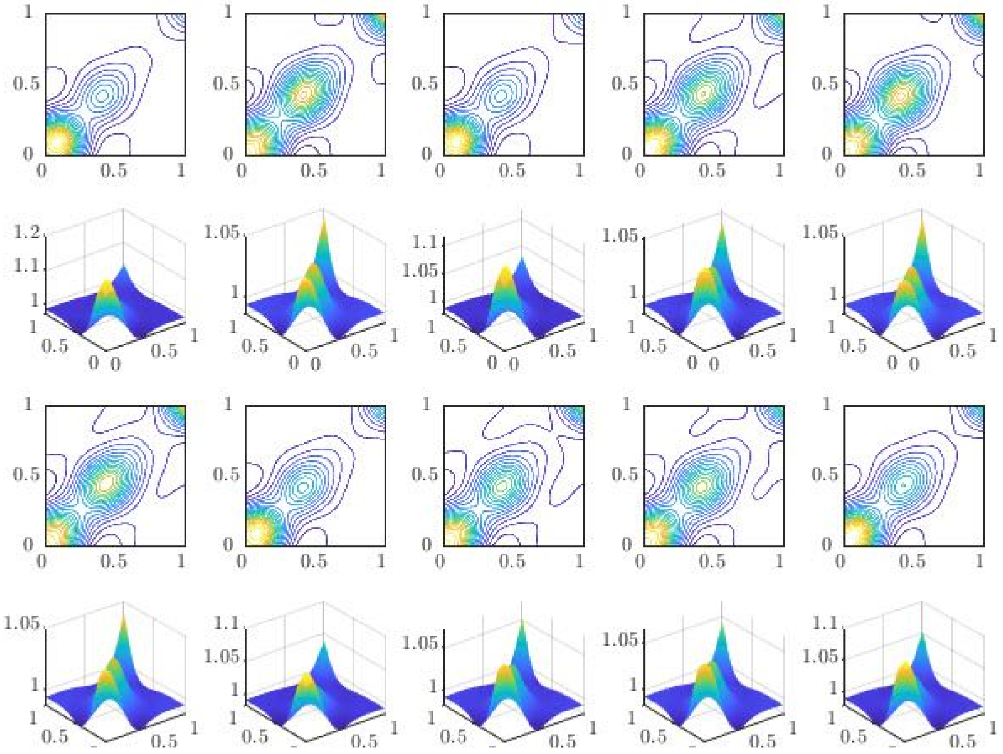


FIGURE 4.2: Contour plots (*first* and *third* row) and the corresponding 3D density plot (*second* and *fourth* row) of the forecasted bivariate copula pdfs, approximated via fPCA, for each horizon $h = 1, \ldots, 5$ (first and second rows) and $h = 6, \ldots, 10$ (third and fourth rows), starting from the top-left panel.

Several parametric and nonparametric estimators for the TDC λ_U, λ_L defined in eq. (4.1)-(4.3) have been proposed in the literature. Here we use the non-parametric estimator obtained from eq. (4.3). Let $u \in [0,1]$ be an arbitrarily small threshold and let $\hat{C}_N(\cdot)$ be the empirical copula cumulative probability function, then the estimator is defined by (see Frahm et al. (2005)):

$$\widehat{\lambda}_U = 2 - \frac{\log(\hat{C}_N(1-u, 1-u))}{\log(1-u)}, \qquad \widehat{\lambda}_L = 2 - \frac{\log(1 - 2u + \hat{C}_N(u,u))}{\log(1-u)}. \tag{4.39}$$

Fig. 4.3 shows the estimated tail dependence coefficients for the sample observations, for each period $t = 1, \ldots, 38$, using a grid of 20 equally spaced threshold values between 0.01 and 0.20. Instead, Fig. 4.4 plots only the case for the median value of the threshold values, i.e. $u = 0.10$. We find significant variation of both the upper and lower tail dependence coefficients over time, which are always different from zero. In addition, the values of the upper TDC differ from those of the lower TDC, thus highlighting an asymmetric tail dependence. The threshold parameter seems to exert a minor role, as almost all the trajectories

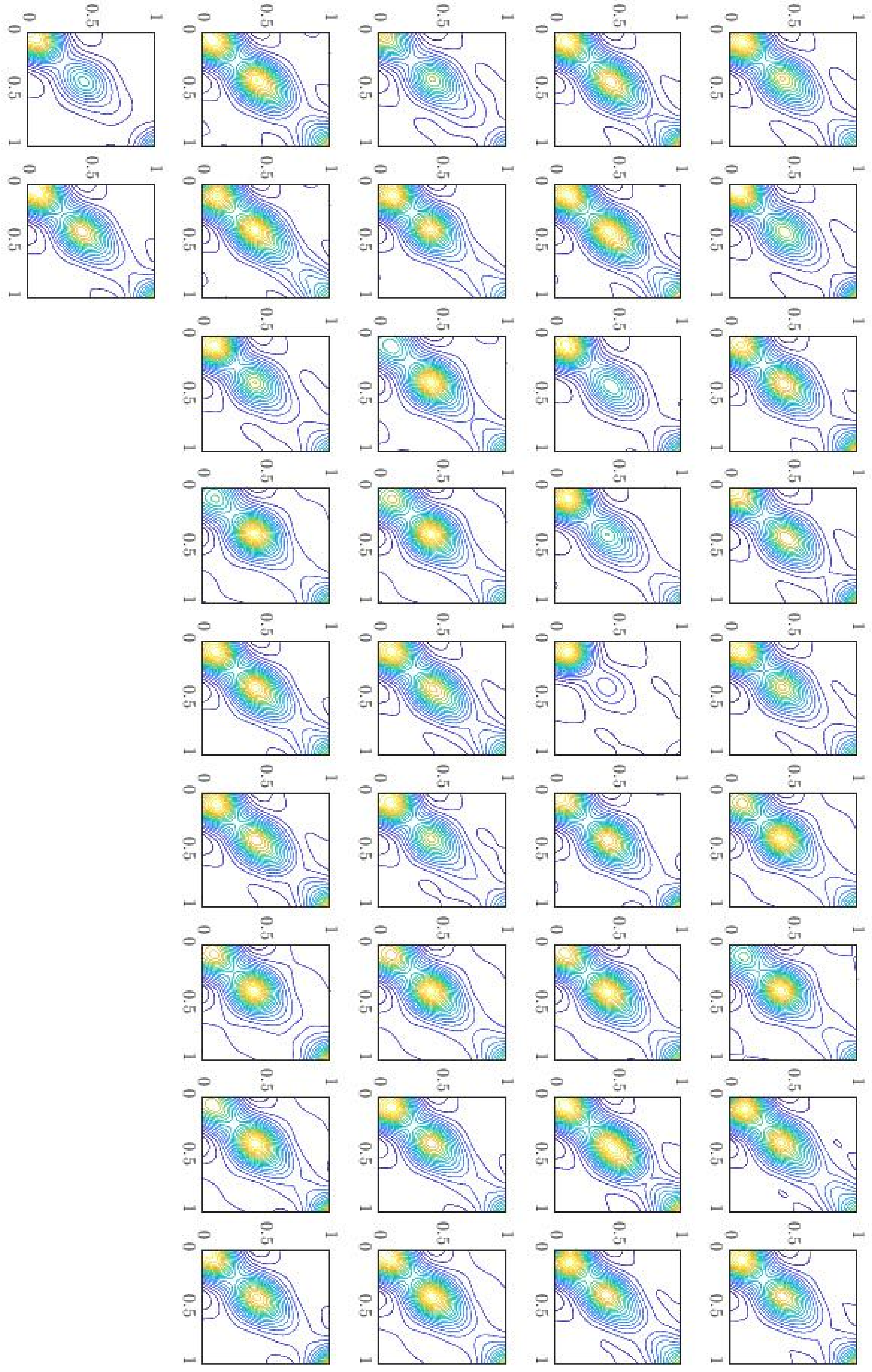


FIGURE 4.1: Contour plot of time series of bivariate copula pdfs, approximated via fPCA, for each year $t = 1, \ldots, T$, starting from $t = 1$ in the top-left panel.

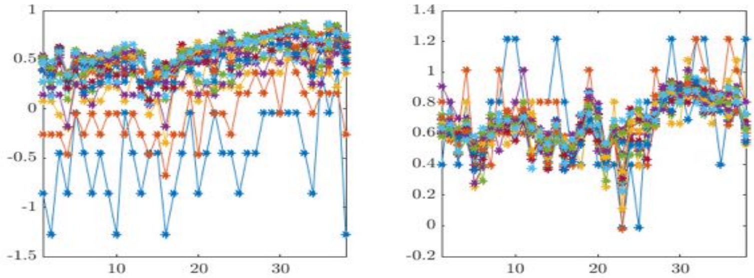


FIGURE 4.3: Upper (*left*) and lower (*right*) tail dependence coefficients of the bivariate time series $(\mathbf{x}_t, \mathbf{y}_t)$, for $t = 1, \dots, 38$ (*x-axis*). Each curve corresponds to a different threshold $u = 0.01, 0.02, \dots, 0.20$.

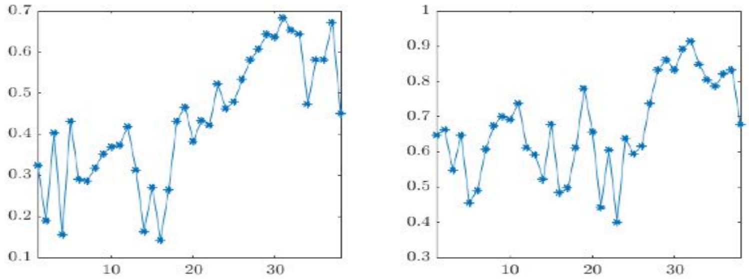


FIGURE 4.4: Upper (*left*) and lower (*right*) tail dependence coefficients of the bivariate time series $(\mathbf{x}_t, \mathbf{y}_t)$, for $t = 1, \dots, 38$ (*x-axis*), threshold $u = 0.10$.

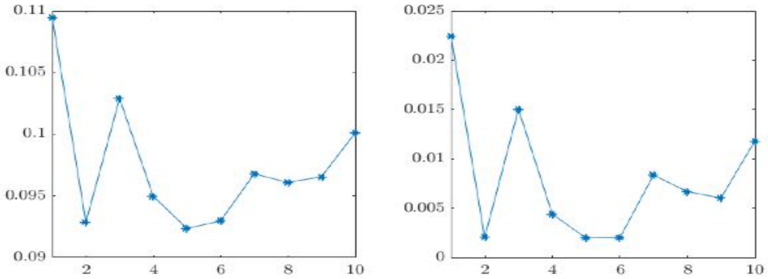


FIGURE 4.5: Upper (*left*) and lower (*right*) tail dependence coefficients of the forecasted bivariate copula pdf $c_{T+h}(\cdot)$, for $h = 1, \dots, 10$ (*x-axis*), threshold $u = 0.10$.

of both $\widehat{\lambda}_U$, $\widehat{\lambda}_L$ remain quite close to each other, except for few values of u, thus indicating robustness of the results (with respect to u). Moreover, the range of variation of the lower TDC is slightly higher than that of the upper TDC, in line with the previous findings in the financial econometrics literature (see Cherubini et al. (2004)).

Fig. 4.5 shows the estimated tail coefficients for the forecasted copula pdfs, for each horizon $h = 1, \dots, 10$, using the threshold value $u = 0.10$. The results are in line with the findings of the sample data: both estimated coefficients are different from zero (however, we did not obtain the standard errors necessary for testing this hypothesis) and asymmetric between the upper and lower case, furthermore they change over time. Considered together, the findings in and out of sample estimated TDC points towards the rejection the use of copula families has either only one type of tail dependence or symmetric tail dependence, as show in Table 4.1.

Remark 4.5.1 (Interpretation)
The proposed methodology, as opposed to standard (semi)parametric dynamic copula models allows to visualize and quantify the temporal evolution of both the upper and lower tail dependence between bivariate time series, as well as to estimate the associated TDC. These findings suggest that the use of this methodology can improve the state-of-the-art on risk modelling due to its flexibility in modelling

the dynamics of the dependence between random variables, which is the cornerstone for definition of adequate risk measures.

4.6 Conclusions

The time varying nature of the dependence pattern between financial variables is a challenging issue in statistics and econometrics. Common methods based on the specification of a dynamic copula model are not enough flexible to describe the temporal change, because each copula family has only a specific kind of tail dependence.

We contribute to this literature by proposing a nonparametric model for forecasting multivariate probability density functions with bounded or unbounded support. The methodology is used for studying the temporal evolution of the copula probability density function encrypting the dependence structure between the S&P500 and the NASDAQ indices. We found evidence of time varying tail dependence which cannot be captured by commonly used econometric models based on dynamic copulas, whereas the model we propose is able to account for these changes. The forecasts highlight smooth but significant variation of the bivariate copula pdf.

The proposed methodology is quite general and can be applied also to other domains. An appealing framework deserving further research concerns the definition of time varying graphical models through dynamic vine copulas (Bedford and Cooke (2002), Joe and Kurowicka (2011)), which combine a tree-like graphical structure, for representing the conditional independence relationships among a set of variables, with bivariate copulas, which describe the pairwise conditional dependence. Here, the method can be used for (separately) modelling the temporal evolution of each bivariate copula characterizing the edges of the network. Our methodology can be parallelized over the edges, for coupling with the issue of dimensionality.

Another stream of research worth further investigation regards the empirical analysis of multivariate (with dimension $d > 2$) pdfs with unbounded support, such as multivariate normal distributions, which are the building block of many well-known econometric models.

Chapter 5

Conclusions

> *We can only see a short distance ahead, but we can see plenty there that needs to be done.*
>
> Alan Turing

> *The mistakes and unresolved difficulties of the past in mathematics have always been the opportunities of its future.*
>
> Eric Temple Bell

The development of suitable models for studying high-dimensional, complex datasets is one of the main challenges that statistics and econometrics are currently facing. The literature on these topics is still at its infancy and represents a promising field of study. This thesis contributes to this growing stream of literature by proposing different statistical approaches for modelling the temporal dependence between complex data structures. Though the empirical studied have focused on dynamic networks, the applicability of the methods to extends to several other contexts.

In the following we summarise the main findings of the thesis and suggest some fields where the proposed methodologies can be applied.

Chapter 2 presents a Bayesian econometric model for real-valued tensor-variate data. The model allows to preserve and exploit the information about the complex structure of the data, as opposed to mainstream vector-variate models. One of the main advantages of the method is its ability to recover the significant heterogeneity of the coefficients. This suggests that complex data structures are also characterised by an intricate, heterogeneous system of interconnections which needs to be accounted for in empirical analyses. Therefore models based on pooling coefficients may have substantial limitations in these cases.

When applied to the study of time series of network data, the methodology admits an interpretation as a reduced-form network autoregressive model, for which tools used in VAR models, such as impulse response functions, are available.

Finally, the computational efficiency of the procedure has been tested, finding good performance even in estimation problems with high-dimensional parameter spaces with several hundreds of parameters.

Chapter 3 describes a Markov switching model for time-varying binary arrays. With respect to the previous model, this framework allows to capture an additional layer of complexity, by allowing both cross-sectional and temporal heterogeneity of the coefficients driving the probability of an occurrence in the array. In particular, it is able to recover the temporal clustering of the coefficients as well as to estimate the different impact that a set of covariates has on each entry of the response.

The main results in financial networks analysis pertain the ability of capturing the temporal evolution of the topology of the network, both in terms of time varying sparsity of the whole structure and in terms of the dynamics of each edge's probability.

From a computational perspective, the model scales well and it permits to estimate the heterogeneity in individual entry's probability even with data arrays containing thousands of entries.

Chapter 4 proposes a nonparametric model for forecasting multivariate probability density functions. The method treats the constraints imposed by probability density functions directly, without the need for post-processing the output. It is highly flexible and requires minimal stationarity assumptions on the underlying time series.

The main advantage of the model is the ability to provide forecast of the whole function of interest without any parametric assumption on the function itself. Moreover, the forecasts made by this method are able to account for significant changes over time of the functions.

The empirical analysis on bivariate copula probability density functions has highlighted the limitations of parametric and semi-parametric models, and shown the advantages of the nonparametric approach in providing accounting for remarkably different patterns of tail dependence over time.

All the methodologies we have developed in this thesis are quite general may be applied also for studying phenomena outside the domain of temporal networks. In particular, dynamic tensor regressions developed in Chapter 2 can be used as a model for general real-values matrix- and tensor-variate time series, which are becoming popular in the literature thanks to the spreading of complex data.

A complementary framework is provided in Chapter 3, where a time series of binary arrays is of interest. This is an extension of multivariate binary regression models which can account for complex cross-sectional structures as well as temporal dynamics. Alternative applications may include the analysis of micro-level panel data on employment relations between firms and individuals, the illness status of workers in different industries and the existence of certain financial relations between banks and households or industries.

The techniques proposed in Chapter 4 are promising for studying time series or cross-sectional data of multivariate probability density functions. This is a rather recent field of research, whose potential is still largely unexplored. The availability of great amount of data at high frequencies, especially in finance, suggests that point forecasts based on vector-valued series may be improved by forecasting the entire function representing the behaviour of the series in a suitable time interval. The generality of the methodology proposed in Chapter 4 makes it appealing for this kind of applications, including, for example, forecasting of the volatility surface, or of high-frequency financial data.

Appendix A

Appendix A

A.1 Background material on tensor calculus

In this section we introduce some operators for multilinear arrays (i.e. tensors): in particular, we consider operations acting on tensors and between tensors and lower-dimensional objects (such as matrices and vectors) as well as some representation (decomposition/approximation) results for tensors. A noteworthy introduction to tensors and corresponding operations is in Lee and Cichocki (2016), while a remarkable reference for tensor decomposition methods is Kolda and Bader (2009). We use the following notation: matrices are represented by boldface upper-case letters, vectors by boldface lower-case letters, scalars by lower-case letters and, finally, calligraphic letters denote tensors, if not differently specified.

A N-order tensor is an element of the tensor product of N vector spaces. Since there exists a isomorphism between two vector spaces of dimensions N and $M < N$, it is possible to define a one-to-one map between their elements, that is, between a N-order tensor and a M-order tensor. We call this tensor reshaping and give its formal definition below.

Definition A.1.1 (Tensor reshaping)
Let $V_1, \ldots, V_N$ and $U_1, \ldots, U_M$ be vector subspaces $V_n, U_m \subseteq \mathbb{R}$ and $\mathcal{X} \in \mathbb{R}^{I_1 \times \ldots \times I_N} = V_1 \otimes \ldots \otimes V_N$ be a N-order real tensor of dimensions $I_1, \ldots, I_N$. Let $(\mathbf{v}_1, \ldots, \mathbf{v}_N)$ be a canonical basis of $\mathbb{R}^{I_1 \times \ldots \times I_N}$ and let Π_S be the projection defined as:

$$\begin{aligned} \Pi_S : & V_1 \otimes \ldots \otimes V_N \to V_{s_1} \otimes \ldots \otimes V_{s_k} \\ & \mathbf{v}_1 \otimes \ldots \otimes \mathbf{v}_N \mapsto \mathbf{v}_{s_1} \otimes \ldots \otimes \mathbf{v}_{s_k} , \end{aligned}$$

with $S = \{s_1, \ldots, s_k\} \subset \{1, \ldots, N\}$. Let $(S_1, \ldots, S_M)$ be a partition of $\{1, \ldots, N\}$. The $(S_1, \ldots, S_M)$ tensor reshaping of $\mathcal{X}$ is defined as:

$$\begin{aligned} \mathcal{X}_{(S_1,\ldots,S_M)} &= (\Pi_{S_1}\mathcal{X}) \otimes \ldots \otimes (\Pi_{S_M}\mathcal{X}) \\ &\in \left(\bigotimes_{s \in S_1} V_s \right) \otimes \ldots \otimes \left(\bigotimes_{s \in S_M} V_s \right) \\ &= U_1 \otimes \ldots \otimes U_M . \end{aligned}$$

It can be proved that the mapping is an isomorphism between $V_1 \otimes \ldots \otimes V_N$ and $U_1 \otimes \ldots \otimes U_M$.

The operation of converting a tensor into a matrix can be seen as a particular case of tensor reshaping, where a N-order tensor is mapped to a 2-order tensor. In practice, it consists in choosing the modes of the array to map with the rows and columns of the resulting matrix, then permuting the tensor and reshaping it, accordingly. The formal follows.

Definition A.1.2
Let $\mathcal{X}$ be a N order tensor with dimensions $I_1, \ldots, I_N$. Let the ordered sets $\mathscr{R} = \{r_1, \ldots, r_L\}$ and $\mathscr{C} = \{c_1, \ldots, c_M\}$ be a partition of $\mathbf{N} = \{1, \ldots, N\}$ and let $I_{\mathbf{N}} = \{I_1, \ldots, I_N\}$. The matricized

tensor is specified by:

$$\mathbf{X}_{(\mathscr{R}\times\mathscr{C}:\mathbf{I_N})} \in \mathbb{R}^{J\times K} \qquad J=\prod_{n\in\mathscr{R}} I_n \quad K=\prod_{n\in\mathscr{C}} I_n\,. \tag{A.1}$$

Indices of $\mathscr{R},\mathscr{C}$ are mapped to the rows and the columns, respectively. More precisely:

$$\left(\mathbf{X}_{(\mathscr{R}\times\mathscr{C}:\mathbf{I_N})}\right)_{j,k} = \mathcal{X}_{i_1,i_2,\dots,i_N} \tag{A.2}$$

with:

$$j = 1+\sum_{l=1}^{L}\left[(i_{r_l}-1)\prod_{l'=1}^{l-1} I_{r'_l}\right] \qquad k = 1+\sum_{m=1}^{M}\left[(i_{c_m}-1)\prod_{m'=1}^{m-1} I_{c'_m}\right] \tag{A.3}$$

Many product operations have been defined for tensors, but here we constrain ourselves to the operator used in this work and we point to Lee and Cichocki (2016) for a summary of other operators. The *mode−n product* between a tensor $\mathcal{X}$ and a vector $\mathbf{v}\in\mathbb{R}^{d_n}$ can be interpreted as the standard Euclidean inner product between the vector and each mode-n fiber of the tensor. Consequently, this operator suppresses one dimension of the tensor and results in a lower order tensor. It is defined, element-wise, by:

$$\mathcal{Y}_{(i_1,\dots,i_{n-1},i_{n+1},\dots,i_D)} = (\mathcal{X}\times_n \mathbf{v})_{(i_1,\dots,i_{n-1},i_{n+1},\dots,i_D)} = \sum_{i_n}^{d_n} \mathcal{X}_{i_1,\dots,i_D}\mathbf{v}_{i_n}, \tag{A.4}$$

with $\mathcal{Y}\in\mathbb{R}^{d_1\times\dots,d_{n-i},d_{n+1},\dots\times d_D}$. Notice that this product is not commutative, since the order of the elements in the multiplication is relevant.

Finally, let $\mathcal{Y}\in\mathbb{R}^{d_1^Y\times\dots\times d_M^Y}$ and $\mathcal{X}\in\mathbb{R}^{d_1^X\times\dots\times d_N^X}$. The *outer product* $\circ$ of two tensors[1] is the tensor $\mathcal{Z}\in\mathbb{R}^{d_1^Y\times\dots\times d_M^Y\times d_1^X\times\dots\times d_N^X}$ whose entries are:

$$\mathcal{Z}_{i_1,\dots,i_M,j_1,\dots,j_N} = (\mathcal{Y}\circ\mathcal{X})_{i_1,\dots,i_M,j_1,\dots,j_N} = \mathcal{Y}_{i_1,\dots,i_M}\mathcal{X}_{j_1,\dots,j_N}. \tag{A.5}$$

For example, the outer product of two vectors is a matrix, while the outer product of two matrices is a tensor of order 4. As a special case, the outer product of two column vectors **a**, **b** can be equivalently represented by means of the Kronecker product $\otimes$:

$$\mathbf{a}\circ\mathbf{b} = \mathbf{b}\otimes\mathbf{a} = \mathbf{a}\cdot\mathbf{b}'\,. \tag{A.6}$$

In the following, we introduce two multilinear operators acting on tensors, see Kolda (2006) for more details.

Definition A.1.3 (Tucker operator)
Let $\mathcal{Y}\in\mathbb{R}^{J_1\times\dots\times J_N}$ and $\mathbf{N}=\{1,\dots,N\}$. Let $\{\mathbf{A}_n\}_n$ be a collection of N matrices such that $\mathbf{A}_n\in\mathbb{R}^{I_n\times J_n}$ for $n\in\mathbf{N}$. The Tucker operator is defined as:

$$[\![\mathcal{Y};\mathbf{A}_1,\dots,\mathbf{A}_N]\!] = \mathcal{Y}\times_1\mathbf{A}_1\times_2\mathbf{A}_2\dots\times_N\mathbf{A}_N\,, \tag{A.7}$$

and the resulting tensor has size $I_1\times\dots\times I_N$.

Definition A.1.4 (Kruskal operator)
Let $\mathbf{N}=\{1,\dots,N\}$ and $\{\mathbf{A}_n\}_n$ be a collection of N matrices such that $\mathbf{A}_n\in\mathbb{R}^{I_n\times R}$ for $n\in\mathbf{N}$. Let $\mathcal{I}$ be the identity tensor of size $R\times\dots\times R$, i.e. a tensor having ones along the superdiagonal and zeros elsewhere. The Kruskal operator is defined as:

$$\mathcal{X} = [\![\mathbf{A}_1,\dots,\mathbf{A}_N]\!] = [\![\mathcal{I};\mathbf{A}_1,\dots,\mathbf{A}_N]\!]\,, \tag{A.8}$$

[1] This operator still applies to vectors and matrices, as they are special cases of tensors of order 1 and 2, respectively.

with $\mathcal{X}$ a tensor of size $I_1 \times \ldots \times I_N$. An alternative representation is obtained by defining $\mathbf{a}_n^{(r)}$ the r-th column of the matrix A_n and using the outer product:

$$\mathcal{X} = [\![\mathbf{A}_1, \ldots, \mathbf{A}_N]\!] = \sum_{r=1}^{R} \mathbf{a}_1^{(r)} \circ \ldots \circ \mathbf{a}_N^{(r)} . \tag{A.9}$$

By exploiting the Khatri-Rao product $\odot$ (i.e. the column-wise Kronecker product for $\mathbf{A} \in \mathbb{R}^{I \times K}$, $\mathbf{B} \in \mathbb{R}^{J \times K}$ defined as $\mathbf{A} \odot \mathbf{B} = [\mathbf{a}_{:,1} \otimes \mathbf{b}_{:,1}, \ldots, \mathbf{a}_{:,K} \otimes \mathbf{b}_{:,K}]$) in combination with the mode-n matricization and the vectorization operators, we get the following additional representations of $\mathcal{X} = [\![\mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$:

$$\mathbf{X}_{(n)} = \mathbf{A}_n \left(\mathbf{A}_N \odot \ldots \odot \mathbf{A}_{n+1} \odot \mathbf{A}_{n-1} \odot \ldots \odot \mathbf{A}_1\right)' \tag{A.10}$$

$$\mathrm{vec}\left(\mathcal{X}\right) = \left(\mathbf{A}_N \odot \ldots \odot \mathbf{A}_1\right) \mathbf{1}_R , \tag{A.11}$$

where $\mathbf{1}_R$ is a vector of ones of length R.

We now define two tensor representations, or decompositions, which are useful in two respects: (i) the algebraic objects that form the decomposition are generally low dimensional and more easily tractable than the tensor; (ii) they can be used to provide a good approximation of the original array. Also, let us denote with R^* be the rank of tensor $\mathcal{X}$, the abstraction of the notion of matrix rank.

The Tucker decomposition can be thought of as a higher-order generalization of Principal Component Analysis (PCA): a tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$ is decomposed into (more precisely, it is approximated by) the product (along the corresponding mode) of a "core" tensor $\mathcal{Y} \in \mathbb{R}^{y_1 \times \ldots \times y_D}$ and D factor matrices $\mathbf{A}^{(l)} \in \mathbb{R}^{d_l \times y_l}$, $1 \leq l \leq D$. Following the notation in Kolda and Bader (2009):

$$\mathcal{X} = \mathcal{Y} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \times_3 \ldots \times_D \mathbf{A}^{(D)} = \sum_{i_1=1}^{y_1} \sum_{i_2=1}^{y_2} \ldots \sum_{i_D=1}^{y_D} y_{i_1,i_2,\ldots,i_D} \mathbf{a}_{i_1}^{(1)} \circ \mathbf{a}_{i_2}^{(2)} \circ \ldots \circ \mathbf{a}_{i_D}^{(D)}. \tag{A.12}$$

Here $\mathbf{a}_{i_l}^{(l)} \in \mathbb{R}^{g_l \times 1}$ is the l-th column of the matrix $\mathbf{A}^{(l)}$. As a result, each entry of the tensor is obtained as:

$$\mathcal{X}_{j_1,\ldots,j_D} = \sum_{i_1=1}^{y_1} \sum_{i_2=1}^{y_2} \ldots \sum_{i_D=1}^{y_D} y_{i_1,i_2,\ldots,i_D} a_{i_1,j_1}^{(1)} a_{i_2,j_2}^{(2)} \ldots a_{i_D,j_D}^{(D)} \qquad 1 \leq j_l \leq d_l,\, 1 \leq l \leq D. \tag{A.13}$$

A special case of the Tucker decomposition is obtained when the core tensor collapses to a scalar and the factor matrices reduce to a single column vector each one is called PARAFAC(R)[2]. More precisely, the PARAFAC(R) decomposition allows to represent a D-order tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$ as the sum of R rank one tensors, that is, of outer products (denoted by $\circ$) of vectors (also called marginals in this case)[3]:

$$\mathcal{X} = \sum_{r=1}^{R} \mathcal{X}_r = \sum_{r=1}^{R} \mathbf{x}_1^{(r)} \circ \ldots \circ \mathbf{x}_D^{(r)}, \tag{A.14}$$

[2] See Harshman (1970). Some authors (e.g., Carroll and Chang (1970) and Kiers (2000)) use the term CODECOMP or CP instead of PARAFAC.

[3] An alternative representation may be used, if all the vectors $\mathbf{x}_j^r$ are normalized to have unitary length. In this case the weight of each component r is captured by the $r - th$ component of the vector $\lambda \in \mathbb{R}^R$:

$$\mathcal{X} = \sum_{r=1}^{R} \lambda_r \left(\mathbf{x}_1^{(r)} \circ \ldots \circ \mathbf{x}_D^{(r)}\right)$$

.

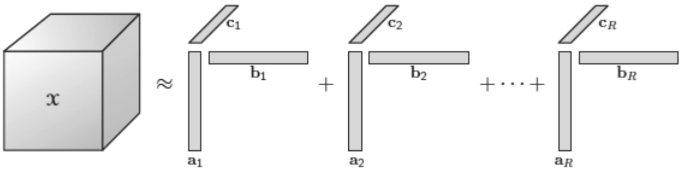


FIGURE A.1: PARAFAC decomposition of $\mathcal{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$, with $\mathbf{a}_r \in \mathbb{R}^{d_!}$, $\mathbf{b}_r \in \mathbb{R}^{d_2}$ and $\mathbf{c}_r \in \mathbb{R}^{d_3}$, $1 \leq r \leq R$. Figure from Kolda and Bader (2009).

with $\mathbf{x}_j^{(r)} \in \mathbb{R}^{d_j}$ $\forall j = 1, \ldots, D$. For a tensor of arbitrary order, the determination of the rank is a $NP-$hard problem (Kolda and Bader (2009)), as a consequence, in applied works, one generally fixes R, uses a PARAFAC(R) approximation, and then run a sensitivity analysis of the results with respect to R. The higher the value of R, the better is the approximation. Alternatively, whenever it is possible to define a measure for the approximation accuracy one may define a grid of values $\{R_i\}_{i=1}^{\bar{R}}$ at which evaluate the accuracy, then choose the value of the grid which yields the best approximation.

The rest of the section contains some results relating the operators we have just defined.

Proposition A.1.0.1 (4.3 in Kolda (2006))
Let $\mathcal{Y} \in \mathbb{R}^{J_1\times\ldots\times J_N}$ and $\mathbf{N} = \{1, \ldots, N\}$ and let $\mathbf{A} \in \mathbb{R}^{I_n\times J_n}$ for all $n \in \mathbf{N}$. If $\mathscr{R} = \{r_1, \ldots, r_L\}$ and $\mathscr{C} = \{c_1, \ldots, c_M\}$ partition $\mathbf{N}$, then:

$$\mathcal{X} = [\![\mathcal{Y}; \mathbf{A}_1, \ldots, \mathbf{A}_N]\!] \iff \mathbf{X}_{(\mathscr{R}\times\mathscr{C}:J_\mathbf{N})} = \left(\mathbf{A}^{(r_L)} \otimes \ldots \otimes \mathbf{A}^{(r_1)}\right) \mathbf{Y}_{(\mathscr{R}\times\mathscr{C}:J_\mathbf{N})} \left(\mathbf{A}^{(c_M)} \otimes \ldots \otimes \mathbf{A}^{(c_1)}\right)' \tag{A.15}$$

where $\mathcal{X} = [\![\mathcal{Y}; \mathbf{A}_1, \ldots, \mathbf{A}_N]\!] = \mathcal{Y} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \ldots \times_n \mathbf{A}_N$ denotes the Tucker product between the tensor $\mathcal{Y}$ and the collection of matrices $\{\mathbf{A}_n\}_{n=1}^N$. The Kruskal operator is a special case of the Tucker operator, obtained when the tensor $\mathcal{Y} = \mathcal{I}$ is an identity tensor of dimensions $R \times \ldots \times R$ and the matrices $\{\mathbf{A}_n\}_{n=1}^N$ have dimension $\mathbf{A}_n \in \mathbb{R}^{I_n\times R}$. Therefore, we can represent the product using the outer product representation, as follows. Consider the collection of vectors $\{\mathbf{a}^{(n)}\}_{n=1}^N$, of length $\mathbf{a}^{(n)} \in \mathbb{R}^{I_n}$, formed by the columns of the matrices $\mathbf{A}_n$. Then:

$$\begin{aligned} \mathcal{X} = [\![\mathcal{I}; \mathbf{A}_1, \ldots, \mathbf{A}_N]\!] = \circ_{n=1}^N \mathbf{a}^{(n)} \iff \\ \mathbf{X}_{(\mathscr{R}\times\mathscr{C}:J_\mathbf{N})} = \left(\mathbf{a}^{(r_L)} \otimes \ldots \otimes \mathbf{a}^{(r_1)}\right) \mathbf{I}_{(\mathscr{R}\times\mathscr{C}:J_\mathbf{N})} \left(\mathbf{a}^{(c_M)} \otimes \ldots \otimes \mathbf{a}^{(c_1)}\right)' . \end{aligned} \tag{A.16}$$

Remark A.1.1 (Contracted product – vectorization)
Let $\mathcal{X} \in \mathbb{R}^{I_1\times\ldots\times I_N}$ and $\mathcal{Y} \in \mathbb{R}^{J_1\times\ldots\times J_N\times J_{N+1}\times\ldots\times J_{N+P}}$. Let $(\mathscr{S}_1, \mathscr{S}_2)$, with $\mathscr{S}_1 = \{1, \ldots, N\}$, $\mathscr{S}_2 = \{N+1, \ldots, N+P\}$, be a partition of $\{1, \ldots, N+P\}$. The following results hold:

a) if $P = 0$ and $I_n = J_n$ for $n = 1, \ldots, N$, then:

$$\mathcal{X} \times^{1\ldots N} \mathcal{Y} = \langle \mathcal{X}, \mathcal{Y} \rangle = \mathrm{vec}\,(\mathcal{X})' \cdot \mathrm{vec}\,(\mathcal{Y}) \quad \in \mathbb{R} . \tag{A.17}$$

b) if $P > 0$ and $I_n = J_n$ for $n = 1, \ldots, N$, then:

$$\mathcal{X} \times^{1\ldots N} \mathcal{Y} = \mathrm{vec}\,(\mathcal{X}) \times^1 \mathcal{Y}_{(\mathscr{S}_1, \mathscr{S}_2)} \quad \in \mathbb{R}^{j_1\times\ldots\times j_P} \tag{A.18}$$

$$\mathcal{Y} \times^{1\ldots N} \mathcal{X} = \mathcal{Y}_{(\mathscr{S}_1, \mathscr{S}_2)} \times^1 \mathrm{vec}\,(\mathcal{X}) \quad \in \mathbb{R}^{j_1\times\ldots\times j_P} . \tag{A.19}$$

c) if $P = N$ and $I_n = J_n = J_{N+n}$, $n = 1, \ldots, N$, then:

$$\mathcal{X} \times^{1\ldots N} \mathcal{Y} \times^{1\ldots N} \mathcal{X} = \mathrm{vec}\,(\mathcal{X})' \cdot \mathbf{Y}_{(\mathscr{R}\times\mathscr{C})} \cdot \mathrm{vec}\,(\mathcal{X}) \quad \in \mathbb{R} . \tag{A.20}$$

Proof. Case a). By definition of contracted product and tensor scalar product:

$$\begin{aligned}\mathcal{X} \times^{1\dots N} \mathcal{Y} &= \sum_{i_1=1}^{I_1} \dots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1,\dots,i_N} \cdot \mathcal{Y}_{i_1,\dots,i_N} \\ &= \sum_{i_1,\dots,i_N} \mathcal{X}_{i_1,\dots,i_N} \cdot \mathcal{Y}_{i_1,\dots,i_N} = \langle \mathcal{X}, \mathcal{Y} \rangle = \operatorname{vec}(X)' \cdot \operatorname{vec}(Y) \ .\end{aligned}$$

Case b). Define $I^* = \prod_{n=1}^{N} I_n$ and $k = 1 + \sum_{j=1}^{N} (i_j - 1) \prod_{m=1}^{j-1} I_m$. By definition of contracted product and tensor scalar product:

$$\begin{aligned}\mathcal{X} \times^{1\dots N} \mathcal{Y} &= \sum_{i_1=1}^{I_1} \dots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1,\dots,i_N} \cdot \mathcal{Y}_{i_1,\dots,i_N,j_{N+1},\dots,j_{N+P}} \\ &= \sum_{k=1}^{I^*} \mathcal{X}_k \cdot \mathcal{Y}_{k,j_{N+1},\dots,j_{N+P}} \ .\end{aligned}$$

Notice that the one-to-one correspondence established by the mapping between k and $(i_1, \dots, i_N)$ corresponds to that of the vectorization of a tensor of size N and dimensions $I_1, \dots, I_N$. Moreover, it also corresponds to the mapping established by the tensor reshaping of a tensor of order $N + P$ with dimensions $I_1, \dots, I_N, J_{N+1}, \dots, J_{N+P}$ into another tensor of order $1 + P$ and dimensions $I^*, J_{N+1}, \dots, J_{N+P}$. Define $S = \{1, \dots, N\}$, such that $(S, N + 1, \dots, N + P)$ is a partition of $\{1, \dots, N + P\}$. Then:

$$\mathcal{X} \times^{1\dots N} \mathcal{Y} = \operatorname{vec}(X) \times^1 \mathcal{Y}_{(S,N+1,\dots,N+P)} \ .$$

Similarly, defining $S = \{P + 1, \dots, N + P\}$ yields the second part of the result.

Case c). We follow the same strategy adopted in case b). Define $S_1 = \{1, \dots, N\}$ and $S_2 = \{N + 1, \dots, N + P\}$, such that $(S - 1, S_2)$ is a partition of $\{1, \dots, N + P\}$. Let k, k' be defined as in case b). Then:

$$\begin{aligned}\mathcal{X} \times^{1\dots N} \mathcal{Y} \times^{1\dots N} \mathcal{X} &= \sum_{i_1=1}^{I_1} \dots \sum_{i_N=1}^{I_N} \sum_{i'_1=1}^{I_1} \dots \sum_{i'_N=1}^{I_N} \mathcal{X}_{i_1,\dots,i_N} \cdot \mathcal{Y}_{i_1,\dots,i_N,i'_1,\dots,i'_N} \cdot \mathcal{X}_{i'_1,\dots,i'_N} \\ &= \sum_{k=1}^{I^*} \sum_{i'_1=1}^{I_1} \dots \sum_{i'_N=1}^{I_N} \mathcal{X}_k \cdot \mathcal{Y}_{k,i'_1,\dots,i'_N} \cdot \mathcal{X}_{i'_1,\dots,i'_N} \\ &= \sum_{k=1}^{I^*} \sum_{k'=1}^{I^*} \mathcal{X}_k \cdot \mathcal{Y}_{k,k'} \cdot \mathcal{X}_{k'} \\ &= \operatorname{vec}(\mathcal{X})' \cdot \mathcal{Y}_{(S_1,S_2)} \cdot \operatorname{vec}(\mathcal{X}) \ .\end{aligned}$$

□

In the following we define a relation between the matricization of a tensor resulting from the outer product of matrices and the Kronecker product.

Remark A.1.2 (Kronecker - matricization)
Let $X_1, \dots, X_N$ be square matrices of size $I_n \times I_n$, $n = 1, \dots, N$ and let $\mathcal{X} = X_1 \circ \dots \circ X_N$ denote the N-order tensor with dimensions $(J_1, \dots J_{2N}) = (I_1, \dots, I_N, I_1, \dots, I_N)$ obtained as the outer product of the matrices $\{U_n\}$. Let $(\mathscr{S}_1, \mathscr{S}_2)$, with $\mathscr{S}_1 = \{1, \dots, N\}$ and $\mathscr{S}_2 = \{N + 1, \dots, N\}$, be a partition of $I_{\mathbf{N}} = \{1, \dots, 2N\}$. Then:

$$\mathcal{X}_{(\mathscr{S}_1,\mathscr{S}_2)} = \mathbf{X}_{(\mathscr{R}\times\mathscr{C}:I_{\mathbf{N}})} = (\mathbf{X}_N \otimes \dots \otimes \mathbf{X}_1) \ . \tag{A.21}$$

Proof. Use the pair of indices (i_n, i'_n) for the entries of the matrix $\mathbf{X}_n$, $n = 1, \dots, N$. By definition of outer product:

$$(\mathbf{X}_1 \circ \ldots \circ \mathbf{X}_N)_{i_1,i_2,\ldots,i_N,i'_1,i'_2,\ldots,i'_N} = (\mathbf{X}_1)_{i_1,i'_1} \cdot (\mathbf{X}_2)_{i_2,i'_2} \cdots (\mathbf{X}_N)_{i_N,i'_N} .$$

From the definition of matricization, $\mathcal{X}_{(\mathscr{S}_1,\mathscr{S}_2)} = \mathbf{X}_{(\mathscr{R}\times\mathscr{C}:I_{\mathbf{N}})}$. Moreover:

$$\left(\mathcal{X}_{(\mathscr{S}_1,\mathscr{S}_2)}\right)_{h,k} = \mathcal{X}_{i_1,\ldots,i_{2N}}$$

with:

$$h = \sum_{p=1}^{N} (i_{S_{1,p}} - 1) \prod_{q=1}^{p-1} J_{S_{1,p}} \qquad k = \sum_{p=1}^{N} (i_{S_{2,p}} - 1) \prod_{q=1}^{p-1} J_{S_{2,p}} .$$

By definition of the Kronecker product we have: that the entry (h', k') of $(X_N \otimes \ldots \otimes X_1)$ is given by:

$$(X_N \otimes \ldots \otimes X_1)_{h',k'} = (X_N)_{i'_N,i'_N} \cdots (X_1)_{i_1,i'_1}$$

where:

$$h' = \sum_{p=1}^{N} (i_{S_{1,p}} - 1) \prod_{q=1}^{p-1} J_{S_{1,p}} \qquad k' = \sum_{p=1}^{N} (i_{S_{2,p}} - 1) \prod_{q=1}^{p-1} J_{S_{2,p}} .$$

Since $h = h'$ and $k = k'$ and the associated elements of $\mathcal{X}_{(\mathscr{S}_1,\mathscr{S}_2)}$ and $(X_N \otimes \ldots \otimes X_1)$ are the same, the result follows. □

Remark A.1.3
Let $\mathcal{X}$ be a N-order tensor of dimensions $I_1 \times \ldots \times I_N$ and let $I^* = \prod_{i=1}^{N} I_i$. Then there exists a vec-permutation (or commutation) matrix $K_{1\to n}$ of size $I^* \times I^*$ such that:

$$K_{1\to n} \operatorname{vec}(\mathcal{X}) = K_{1\to n} \operatorname{vec}\left(\mathbf{X}_{(1)}\right) = \operatorname{vec}\left(\mathbf{X}_{(n)}\right) . \tag{A.22}$$

Moreover, it holds:

$$\operatorname{vec}\left(\mathbf{X}_{(n)}\right) = \operatorname{vec}\left(\mathbf{X}_{(1)}^{T_\sigma}\right) = \operatorname{vec}\left(\mathcal{X}^{T_\sigma}\right) , \tag{A.23}$$

where

$$\mathbf{X}_{(1)}^{T_\sigma} = \left(\mathcal{X}^{T_\sigma}\right)_{(1)} = \mathbf{X}_{(n)} , \tag{A.24}$$

is the mode-1 matricization of the transposed tensor $\mathcal{X}^{T_\sigma}$ according to the permutation σ which exchanges modes 1 and n, leaving the others unchanged. That is, for $i_j \in \{1, \ldots, I_j\}$ and $j = 1, \ldots, N$:

$$\sigma(i_j) = \begin{cases} 1 & j = n \\ n & j = 1 \\ i_j & j \neq 1, n . \end{cases}$$

Remark A.1.4
Let $\mathcal{X}$ be a N-order random tensor with dimensions $I_1, \ldots, I_N$ and let $\mathbf{N} = \{1, \ldots, N\}$ be partitioned by the index sets $\mathscr{R} = \{r_1, \ldots, r_m\} \subset \mathbf{D}$ and $\mathscr{C} = \{c_1, \ldots, c_p\} \subset \mathbf{N}$, i.e. $\mathbf{N} = \mathscr{R} \cup \mathscr{C}$, $\mathscr{R} \cap \mathscr{C} = \emptyset$ and $N = m + p$. Then:

$$\mathcal{X} \sim \mathcal{N}_{I_1,\ldots,I_N}(\mathcal{M}, \mathbf{U}_1, \ldots, \mathbf{U}_N) \iff \mathbf{X}_{(\mathscr{R}\times\mathscr{C})} \sim \mathcal{N}_{m,p}(\mathbf{M}_{(\mathscr{R}\times\mathscr{C})}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) , \tag{A.25}$$

with:

$$\mathbf{\Sigma}_1 = \mathbf{U}_{r_m} \otimes \ldots \otimes \mathbf{U}_{r_1} \qquad \mathbf{\Sigma}_2 = \mathbf{U}_{c_p} \otimes \ldots \otimes \mathbf{U}_{c_1} . \tag{A.26}$$

Proof. We demonstrate the statement for $\mathscr{R} = \{n\}$, $n \in \mathbf{N}$, however the results follows from the same steps also in the general case $\#\mathscr{R} > 1$. The strategy it to demonstrate that the probability density functions of the two distributions coincide. To this aim consider separately the exponent and the normalizing constant. Define $I_{-j} = \prod_{i=1,n\neq j}^{N} I_i$ and $I_{\mathbf{N}} = \{I_1, \ldots, I_N\}$, then for the normalizing constant we have:

$$(2\pi)^{-\frac{\prod_i I_i}{2}} |U_1|^{-\frac{I_{-1}}{2}} \cdots |\mathbf{U}_n|^{-\frac{I_{-n}}{2}} \cdots |\mathbf{U}_N|^{-\frac{I_{-N}}{2}} = \tag{A.27}$$

$$= (2\pi)^{-\frac{\prod_i I_i}{2}} |\mathbf{U}_1|^{-\frac{I_{-1}}{2}} \cdots |\mathbf{U}_{n-1}|^{-\frac{I_{-(n-1)}}{2}} \cdot |\mathbf{U}_{n+1}|^{-\frac{I_{-(n+1)}}{2}} \cdots |\mathbf{U}_N|^{-\frac{I_{-N}}{2}} \cdot |\mathbf{U}_n|^{-\frac{I_{-n}}{2}}$$

$$= (2\pi)^{-\frac{\prod_i I_i}{2}} |\mathbf{U}_N \otimes \ldots \otimes \mathbf{U}_{n-1} \otimes \mathbf{U}_{n+1} \otimes \ldots \otimes \mathbf{U}_N|^{-\frac{n}{2}} \cdot |\mathbf{U}_n|^{-\frac{I_{-n}}{2}} . \tag{A.28}$$

Concerning the exponent, let $\mathbf{i} = (i_1, \ldots, i_N)$ and, for ease of notation, define $\mathcal{Y} = \mathcal{X} - \mathcal{M}$ and $\mathcal{U} = (\mathbf{U}_N^{-1} \circ \ldots \circ \mathbf{U}_1^{-1})$. By the definition of contracted product it holds:

$$\mathcal{Y} \times^{1\ldots N} \mathcal{U} \times^{1\ldots N} \mathcal{Y} = \tag{A.29}$$

$$= \sum_{i_1,\ldots,i_N} \sum_{i'_1,\ldots,i'_N} y_{i_1,\ldots,i_n,\ldots,i_N} \cdot u^{-1}_{i_1,i'_1} \cdots u^{-1}_{i_n,i'_n} \cdots u^{-1}_{i_N,i'_N} \cdot y_{i'_1,\ldots,i_n,\ldots,i'_N} .$$

Define $\mathbf{j} = \sigma(\mathbf{i})$, where σ is the permutation defined above exchanging i_1 with i_n, $n \in \{2, \ldots, N\}$. Then the previous equation can be rewritten as:

$$= \sum_{j_1,\ldots,j_N} \sum_{j'_1,\ldots,j'_N} y_{j_n,\ldots,j_1,\ldots,i_N} \cdot u^{-1}_{j_n,j'_n} \cdots u^{-1}_{j_1,j'_1} \cdots u^{-1}_{i_N,i'_N} \cdot y_{j'_n,\ldots,j'_1,\ldots,i'_N}$$

$$= \mathcal{Y}^{\sigma} \times^{1\ldots N} \left(\mathbf{U}_1^{-1} \circ \ldots \circ \mathbf{U}_N^{-1}\right)^{\sigma} \times^{1\ldots N} \mathcal{Y}^{\sigma} ,$$

where $\mathcal{Y}^{\sigma}$ is the transpose tensor of $\mathcal{Y}$ (see Pan (2014)) obtained by permuting the first and the n-th modes and similarly for the 6-order tensor $(\mathbf{U}_1^{-1} \circ \ldots \circ \mathbf{U}_N^{-1})^{\sigma}$. Let $(\mathscr{S}_1, \mathscr{S}_2)$, with $\mathscr{S}_1 = \{1, \ldots, N\}$ and $\mathscr{S}_2 = \{N+1, \ldots, 2N\}$, be a partition of $\{1, \ldots, 2N\}$. By vectorizing eq. (A.29) and exploiting the results in A.1.1 and A.1.2, we have:

$$\mathcal{Y} \times^{1\ldots N} \mathcal{U} \times^{1\ldots N} \mathcal{Y} = \mathrm{vec}(\mathcal{Y})' \cdot \mathcal{U}_{(\mathscr{S}_1,\mathscr{S}_2)} \cdot \mathrm{vec}(\mathcal{Y}) \tag{A.30}$$

$$\begin{aligned}
&= \mathrm{vec}(\mathcal{Y})' \cdot \left(\mathbf{U}_N^{-1} \otimes \ldots \otimes \mathbf{U}_n^{-1} \otimes \ldots \otimes \mathbf{U}_1^{-1}\right) \cdot \mathrm{vec}(\mathcal{Y}) \\
&= \mathrm{vec}(\mathcal{Y}^{\sigma})' \cdot \left(\mathbf{U}_N^{-1} \otimes \ldots \otimes \mathbf{U}_1^{-1} \otimes \mathbf{U}_n^{-1}\right) \cdot \mathrm{vec}(\mathcal{Y}^{\sigma}) \\
&= \mathrm{vec}\left(\mathbf{Y}_{(n)}\right)' \cdot \left(\mathbf{U}_N^{-1} \otimes \ldots \otimes \mathbf{U}_1^{-1} \otimes \mathbf{U}_n^{-1}\right) \cdot \mathrm{vec}\left(\mathbf{Y}_{(n)}\right) \\
&= \mathrm{vec}\left(\mathbf{Y}_{(n)}\right)' \cdot \mathrm{vec}\left(\mathbf{U}_n^{-1} \cdot \mathbf{Y}_{(n)} \cdot \left[\mathbf{U}_N^{-1} \otimes \ldots \otimes \mathbf{U}_1^{-1}\right]\right) \\
&= \mathrm{tr}\left(\mathbf{Y}'_{(n)} \cdot \mathbf{U}_n^{-1} \cdot \mathbf{Y}_{(n)} \cdot \left[\mathbf{U}_N^{-1} \otimes \ldots \otimes \mathbf{U}_1^{-1}\right]\right) \\
&= \mathrm{tr}\left(\left[\mathbf{U}_N^{-1} \otimes \ldots \otimes \mathbf{U}_1^{-1}\right] \left[\mathbf{X}_{(n)} - \mathbf{M}_{(n)}\right]' \cdot \mathbf{U}_n^{-1} \cdot \left[\mathbf{X}_{(n)} - \mathbf{M}_{(n)}\right]\right) .
\end{aligned} \tag{A.31}$$

Since the term in (A.27) and (A.30) are the normalizing constant and the exponent of the tensor normal distribution, whereas (A.28) and (A.31) are the corresponding expressions for the desired matrix normal distribution, the result is proved for the case $\#\mathscr{R} = 1$. In the general case $\#\mathscr{R} = r > 1$ the proof follows from the same reasoning, by substituting the permutation σ with another permutation σ' which exchanges the modes of the tensor such that the first r modes of the transpose tensor $\mathcal{Y}^{\sigma'}$ correspond to the elements of $\mathscr{R}$. $\square$

Appendix B

Appendix B

B.1 Proofs of the results in Section 2.2

Proof of result in Remark 2.2.3. By assuming $I_j = 1$, for $j = 1, \dots, N$, in mode (2.16), then:

$$\mathcal{Y}_t, \mathcal{A}, \mathbf{E}_t \in \mathbb{R}^{I_1 \times \ldots \times I_N} \to \mathbb{R} \tag{B.1}$$
$$\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N \times J} \to \mathbb{R}^J \tag{B.2}$$
$$\mathcal{C} \in \mathbb{R}^{I_1 \times \ldots \times I_N \times Q} \to \mathbb{R}^Q \tag{B.3}$$

where the matrix $\mathbf{W}_t$ has been removed as covariate. In order to keep it, it would be necessary either to vectorize it (then $\mathcal{D}$ would follow the same change as $\mathcal{B}$) or to assume an inner product (here $\mathcal{D}$ would reduce to a matrix of the same dimension of $\mathbf{W}_t$). Notice that a N-order tensor whose modes have all unitary length is essentially a scalar. As a consequence, the error term distribution reduces to a univariate Gaussian, with 0 mean and variance σ^2. Finally, also the mode-3 product reduces to the standard inner product between vectors.

The PARAFAC(R) decomposition still holds in this case. Consider only $\mathcal{A}$ and $\mathcal{B}$, as the other tensors behave in the same manner. For ease of notation we drop the index t, since it does not affect the result:

$$\mathcal{A} = \sum_{r=1}^{R} \alpha_1^{(r)} \circ \ldots \circ \alpha_N^{(r)} = \sum_{r=1}^{R} \alpha_1^{(r)} \cdot \ldots \cdot \alpha_N^{(r)} = \sum_{r=1}^{R} \tilde{\alpha}_r = \bar{\alpha} \in \mathbb{R}. \tag{B.4}$$

Here, $\tilde{\alpha}_r = \prod_{j=1}^{N} \alpha_j^{(r)}$. Since each mode of $\mathcal{A}$ has unitary length, each of the marginals of the PARAFAC(R) decomposition is a scalar, therefore the outer product reduces to the ordinary product and the outcome is a scalar too (obtained by R sums of D products). Concerning $\mathcal{B}$, we apply the same way of reasoning, with the only exception that in this case one of the modes (the last, in the formulation of eq. (2.16)) has length $J > 1$, implying that the corresponding marginal is a vector of the same length. The result is a vector, as stated:

$$\mathcal{B} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \ldots \circ \beta_N^{(r)} \circ \boldsymbol{\beta}_{D+1}^{(r)} = \sum_{r=1}^{R} \beta_1^{(r)} \cdot \ldots \cdot \beta_N^{(r)} \cdot \boldsymbol{\beta}_{N+1}^{(r)} \tag{B.5}$$
$$= \sum_{r=1}^{R} \tilde{\beta}_r \boldsymbol{\beta}_{N+1}^{(r)} = \boldsymbol{\beta} \in \mathbb{R}^Q , \tag{B.6}$$

where $\tilde{\beta}_r = \prod_{j=1}^{N} \beta_j^{(r)}$. By an analogous proof, one gets:

$$\mathcal{C} = \gamma \in \mathbb{R}^J . \tag{B.7}$$

which completes the proof. □

Proof of result in Remark 2.2.5. Without loss of generality, let $J_j = 1$, for $j = 2, \dots, N$ in model (2.16), then:

$$\mathcal{Y}_t, \mathcal{A}, \mathbf{E}_t \in \mathbb{R}^{I_1 \times \ldots \times I_N} \to \mathbb{R}^m \tag{B.8}$$

$$\mathcal{B} \in \mathbb{R}^{I_1 \times \ldots \times I_N \times J} \to \mathbb{R}^{m \times J} \tag{B.9}$$

$$\mathcal{C} \in \mathbb{R}^{I_1 \times \ldots \times I_N \times Q} \to \mathbb{R}^{m \times Q} \tag{B.10}$$

$$\mathcal{D} \in \mathbb{R}^{I_1 \times \ldots \times I_{n-1} \times K \times I_{n+1} \ldots \times I_N} \to \mathbb{R}^{m \times K}\,, \tag{B.11}$$

where it is necessary to assume that $\mathbf{W}_t \in \mathbb{R}^{m \times K}$. The two mode-$N+1$ products become mode-2 products and the distribution of the error term reduces to the multivariate (n-dimensional) Gaussian, with a unique covariance matrix ($m \times m$).

As the PARAFAC(R) approximation is concerned, the result for $\mathcal{A}$ follows from the second part of the previous proof and yields $\mathcal{A} = \boldsymbol{\alpha} \in \mathbb{R}^m$. For the remaining tensors, it holds (dropping the index for notational ease):

$$\mathcal{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \beta_2^{(r)} \circ \ldots \circ \beta_N^{(r)} = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \left(\beta_2^{(r)} \cdot \ldots \cdot \beta_N^{(r)}\right) \circ \boldsymbol{\beta}_{N+1}^{(r)} \tag{B.12}$$

$$= \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_{N+1}^{(r)} \cdot \tilde{\beta}_r = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)\prime} \boldsymbol{\beta}_{N+1}^{(r)} \cdot \tilde{\beta}_r \tag{B.13}$$

$$= \sum_{r=1}^{R} B^{(r)} \cdot \tilde{\beta}_r = \tilde{B} \in \mathbb{R}^{m \times J} \tag{B.14}$$

where $\tilde{\beta}_r = \prod_{j=2}^{N} \beta_j^{(r)}$. The same result holds for the tensor $\mathcal{C}$, which is equal to $C \in \mathbb{R}^{m \times Q}$, with the last mode's length changed from J to Q. Finally, concerning $\mathcal{D}$:

$$\mathcal{D} = \sum_{r=1}^{R} \delta_1^{(r)} \circ \ldots \circ \delta_{n-1}^{(r)} \circ \boldsymbol{\delta}_n^{(r)} \circ \delta_{n+1}^{(r)} \circ \ldots \circ \delta_N^{(r)} = \sum_{r=1}^{R} \left(\delta_1^{(r)} \cdot \ldots \cdot \delta_{n-1}^{(r)}\right) \cdot \boldsymbol{\delta}_n^{(r)} \cdot \left(\delta_{n+1}^{(r)} \cdot \ldots \cdot \delta_N^{(r)}\right) \tag{B.15}$$

$$= \sum_{r=1}^{R} \boldsymbol{\delta}_n^{(r)} \cdot \left(\delta_1^{(r)} \cdot \ldots \cdot \delta_{n-1}^{(r)}\right) \cdot \left(\delta_{n+1}^{(r)} \cdot \ldots \cdot \delta_N^{(r)}\right) = \sum_{r=1}^{R} \boldsymbol{\delta}_n^{(r)} \cdot \tilde{\delta}_r = \mathbf{d} \in \mathbb{R}^m\,, \tag{B.16}$$

with $\tilde{\delta}_r = \prod_{j \neq n}^{N} \delta_j^{(r)}$. Notice that the resulting mode-n product reduces to an ordinary dot product between the matrix W and the vector $\bar{\mathbf{d}}$.

It remains to prove that the structure imposed by standard VARX and Panel VAR models holds also in the model of eq. (2.16). Notice that the latter does not impose any restriction on the coefficients, other than the PARAFAC(R) decomposition. It must be stressed that it is not possible to achieve the desired structure of the coefficients, in terms of the location of the zeros, by means of an accurate choice of the marginals. In fact, the decomposition we are assuming does not allow to create a particular structure on the resulting tensor.

Nonetheless, it is still possible to achieve the desired result by a slight modification of the model in eq. (2.16). For example, consider the coefficient tensor $\mathcal{B}$, then to create a tensor whose entries are non-zero only in some pre-specified (hence *a-priori* known) cells, it suffices to multiply $\mathcal{B}$ by a binary tensor (i.e. one where all entries are either 0 or 1) via the Hadamard product. In formulas, let $\mathcal{H} \in \{0,1\}^{I_1 \times \ldots \times I_N \times J}$, such that it has 0 only in those cells which are known to be null. Then:

$$\tilde{\mathcal{B}} = \mathcal{H} \odot \mathcal{B}$$

will have the desired structure. The same way of reasoning holds for any coefficient tensor as well as for the covariance matrices.

To conclude, in Panel VAR models one generally has as regressors in each equation a function of the endogenous variables (for example their average). Since this does not affect the coefficients of the model, it is possible to re-create it in our framework by simply rearranging the regressors in eq. (2.16) accordingly. In terms of the model, none of the issues described invalidates the formulation of eq. (2.16), which is able to encompass all of them by suitable rearrangements of the covariates and/or the coefficients, which are consistent with the general model. □

Remark B.1.1 (follows from 2.2.6)
From the VECM in eq. (2.24) and denoting $\mathbf{y}_{t-1} = \operatorname{vec}(\mathbf{Y}_{t-1})$ we can obtain an explicit form for the long run equilibrium (or cointegrating) relations, as follows:

$$\begin{aligned} \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1} &= \left(\sum_{r=1}^{R} \tilde{\boldsymbol{\beta}}_1^{(r)} \circ \tilde{\boldsymbol{\beta}}_2^{(r)} \circ \tilde{\boldsymbol{\beta}}_3^{(r)} \right) \times_3 \mathbf{y}_{t-1} && \text{(B.17a)} \\ &= \sum_{r=1}^{R} \left(\tilde{\boldsymbol{\beta}}_1^{(r)} \circ \tilde{\boldsymbol{\beta}}_2^{(r)} \right) \cdot \langle \tilde{\boldsymbol{\beta}}_3^{(r)}, \mathbf{y}_{t-1} \rangle && \text{(B.17b)} \\ &= \sum_{r=1}^{R} \tilde{B}_{12}^{(r)} \cdot \langle \tilde{\boldsymbol{\beta}}_3^{(r)}, \mathbf{y}_{t-1} \rangle , && \text{(B.17c)} \end{aligned}$$

with $\tilde{B}_{12}^{(r)} = \tilde{\boldsymbol{\beta}}_1^{(r)} \circ \tilde{\boldsymbol{\beta}}_2^{(r)}$ being a $K \times K$ matrix of loadings for each $r = 1, \ldots, R$, while the inner product $\langle \tilde{\boldsymbol{\beta}}_3^{(r)}, \mathbf{y}_{t-1} \rangle$ defines the cointegrating relations. Notice that for a generic entry $y_{ij,t}$, the previous long run relation is defined in terms of all the entries of the lagged matrix $\mathbf{Y}_{t-1}$, each one having a long run coefficient (in the r-th relation) $\tilde{\beta}_{3,k}^{(r)}$, where k can be obtained from (i, j) via a one-to-one mapping corresponding to the reshaping of the $K \times K$ matrix $\mathbf{Y}_{t-1}$ into the $K^2 \times 1$ vector $\mathbf{y}_{t-1}$.

Finally, as the cointegrating relations are not unique, that is $\boldsymbol{\beta}$ in eq. (2.24) is not identified, the same is true for the tensor model, as noted in Section 2.2.

B.2 Initialisation details

It is well known that the Gibbs sampler algorithm is highly sensitive to the choice of the initial value. From this point of view, the most difficult parameters initialise in the proposed model are the margins of the tensor of coefficients, that is the set of vectors: $\{\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \boldsymbol{\beta}_3^{(r)}\}_{r=1}^{R}$. Due to the high complexity of the parameter space, we have chosen to perform an initialisation scheme which is based on the Simulated Annealing (SA) algorithm (see Robert and Casella (2004) and Press et al. (2007) for a thorough discussion). This algorithm is similar to the Metropolis-Hastings one, and the idea behind it is to perform a stochastic optimisation by proposing random moves from the current state which are always accepted when improving the optimum and have positive probability of acceptance even when they are not improving. This is used in order to allow the algorithm to escape from local optima. Denoting the objective function to be minimised by $f(\boldsymbol{\theta})$, the Simulated Annealing method accepts a move from the current state $\boldsymbol{\theta}^(i)$ to the proposed one $\boldsymbol{\theta}^{new}$ with probability given by the Bolzmann-like distribution:

$$p(\Delta f, T) = \exp\left\{ -\frac{\Delta f}{T} \right\} . \tag{B.18}$$

Here $\Delta f = f(\boldsymbol{\theta}^{new}) - f(\boldsymbol{\theta}^{(i)})$ and T is a parameter called temperature. The key of the SA method is in the cooling scheme, which describes the deterministic, decreasing evolution of the temperature over the iterations of the algorithm: it has been proved that under sufficiently slow decreasing schemes, the SA yields a global optimum.

We propose to use the SA algorithm for minimising the objective function:

$$f(\{\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \boldsymbol{\beta}_3^{(r)}\}_{r=1}^R) = \kappa_N \psi_N + \kappa_3 \psi_3 \,, \tag{B.19}$$

where κ_N is an overall penalty given by the Frobenius norm of the tensor constructed from simulated margins, while κ_3 is the penalty of the sum (over r) of the norms of the marginals $\boldsymbol{\beta}_3^{(r)}$. In formulas:

$$\psi_N = \left\| \mathcal{B}^{SA} \right\|_2 \qquad \psi_3 = \sum_{r=1}^R \left\| \boldsymbol{\beta}_3^{(r)} \right\|_2 . \tag{B.20}$$

The proposal distribution for each margin is a normal $\mathcal{N}_{I_j}(\mathbf{0}, \sigma I)$, independent from the current state of the algorithm. Finally, we have chosen a logarithmic cooling scheme which updates the temperature at each iteration of the SA:

$$T_i = \frac{k}{1 + \log(i)} \qquad i = 1, \ldots, I \,, \tag{B.21}$$

where $k > 0$ is a tuning parameter, which can be interpreted as the initial value of the temperature. In order to perform the initialisation of the margins, we run the SA algorithm for $I = 1000$ iterations, then we took the vectors which gave the best fit in terms of minimum value of the objective function.

In the tensor case, the initialization of the PARAFAC marginals $\{\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \boldsymbol{\beta}_3^{(r)}, \boldsymbol{\beta}_4^{(r)}\}_{r=1}^R$ follows the same line, with ψ_3 in eq. (B.20) replaced by:

$$\psi_4 = \sum_{r=1}^R \left\| \boldsymbol{\beta}_4^{(r)} \right\|_2 . \tag{B.22}$$

B.3 Computational details - matrix case

In this section we will follow the convention of denoting the prior distributions with $\pi(\cdot)$. In addition, let $\mathbf{W} = \{\mathbf{W}_{j,r}\}_{j,r}$ be the collection of all (local variance) matrices $\mathbf{W}_{j,r}$, for $j = 1, 2, 3$ and $r = 1, \ldots, R$; $I_0 = \sum_{j=1}^3 I_j$ the sum of the length of each mode of the tensor $\mathcal{B}$ and $\mathbf{Y} = \{\mathbf{Y}_t, \mathbf{X}_t\}_t$ the collection of observed variables.

B.3.1 Full conditional distribution of $\boldsymbol{\phi}$

In order to derive this posterior distribution, we make use of Lemma 7.9 in Guhaniyogi et al. (2017). Recall that: $a_\tau = \alpha R$, $b_\tau = \alpha (R)^{1/N}$ and $I_0 = \sum_{j=1}^N I_j$. The prior for $\boldsymbol{\phi}$ is $\pi(\boldsymbol{\phi}) \sim \mathcal{D}ir(\boldsymbol{\alpha})$.

$$p(\boldsymbol{\phi} | \mathcal{B}, \mathbf{W}) \propto \pi(\boldsymbol{\phi}) p(\mathcal{B} | \mathbf{W}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi}) \int_0^{+\infty} p(\mathcal{B} | \mathbf{W}, \boldsymbol{\phi}, \tau) \pi(\tau) \mathrm{d}\tau \,. \tag{B.23}$$

By plugging in the prior distributions for τ, $\boldsymbol{\phi}$, $\boldsymbol{\beta}_j^{(r)}$ we obtain[1]:

$$p(\boldsymbol{\phi} | \mathcal{B}, \mathbf{W}) \propto \prod_{r=1}^R \phi_r^{\alpha - 1} \int_0^{+\infty} \left[\prod_{r=1}^R \prod_{j=1}^N (\tau \phi_r)^{-I_j/2} \left| \mathbf{W}_{j,r} \right|^{-1/2} \exp\left\{ -\frac{1}{2\tau\phi_r} \boldsymbol{\beta}_j^{(r)\prime} \mathbf{W}_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)} \right\} \right]$$

[1] We have used the property of the determinant: $\det(kA) = k^n \det(A)$, for A square matrix of size n and k scalar.

$$
\begin{aligned}
&\cdot \tau^{a_\tau-1} \exp\left\{-b_\tau \tau\right\} \mathrm{d}\tau \\
&\propto \prod_{r=1}^{R} \phi_r^{\alpha-1} \int_0^{+\infty} \left[\prod_{r=1}^{R} (\tau\phi_r)^{-I_0/2} \exp\left\{ -\frac{1}{2\tau\phi_r} \sum_{j=1}^{N} \boldsymbol{\beta}_j^{(r)'} \mathbf{W}_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)} \right\} \right] \\
&\cdot \tau^{a_\tau-1} \exp\left\{-b_\tau \tau\right\} \mathrm{d}\tau \,. \qquad \text{(B.24)}
\end{aligned}
$$

Define $C_r = \frac{1}{2}\sum_{j=1}^{N} \boldsymbol{\beta}_j^{(r)'} \mathbf{W}_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)}$, then group together the powers of τ and ϕ_r as follows:

$$
\begin{aligned}
p(\boldsymbol{\phi}|\mathcal{B},\mathbf{W}) &\propto \prod_{r=1}^{R} \phi_r^{\alpha-1-\frac{I_0}{2}} \int_0^{+\infty} \tau^{a_\tau-1-\frac{RI_0}{2}} \exp\left\{-b_\tau \tau\right\} \left[\prod_{r=1}^{R} \exp\left\{ -\frac{1}{2\tau\phi_r} C_r \right\} \right] \mathrm{d}\tau \\
&= \prod_{r=1}^{R} \phi_r^{\alpha-1-\frac{I_0}{2}} \int_0^{+\infty} \tau^{a_\tau-1-\frac{RI_0}{2}} \exp\left\{ -b_\tau\tau - \sum_{r=1}^{R} \frac{C_r}{2\tau\phi_r} \right\} \mathrm{d}\tau \,. \qquad \text{(B.25)}
\end{aligned}
$$

Recall that the probability density function of a Generalized Inverse Gaussian in the parametrization with three parameters ($a > 0$, $b > 0$, $p \in \mathbb{R}$), with $x \in (0, +\infty)$, is given by:

$$
x \sim GiG(a,b,p) \;\; \Rightarrow \;\; p(x|a,b,p) = \frac{\left(\frac{a}{b}\right)^{\frac{p}{2}}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left\{ -\frac{1}{2}\left(ax + \frac{b}{x} \right) \right\} , \qquad \text{(B.26)}
$$

with $K_p(\cdot)$ a modified Bessel function of the second type. Our goal is to reconcile eq. (B.73) to the kernel of this distribution. Since by definition $\sum_{r=1}^{R} \phi_r = 1$, it holds that $\sum_{r=1}^{R} (b_\tau \tau \phi_r) = (b_\tau \tau) \sum_{r=1}^{R} \phi_r = b_\tau \tau$. This allows to rewrite the exponential as:

$$
\begin{aligned}
p(\boldsymbol{\phi}|\mathcal{B},\mathbf{W}) &\propto \prod_{r=1}^{R} \phi_r^{\alpha-1-\frac{I_0}{2}} \int_0^{+\infty} \tau^{\left(a_\tau - \frac{RI_0}{2}\right)-1} \exp\left\{ -\sum_{r=1}^{R} \left(\frac{C_r}{2\tau\phi_r} + b_\tau\tau\phi_r \right) \right\} \mathrm{d}\tau \\
&= \int_0^{+\infty} \left(\prod_{r=1}^{R} \phi_r^{\alpha-\frac{I_0}{2}-1} \right) \tau^{\left(\alpha R - \frac{RI_0}{2}\right)-1} \exp\left\{ -\sum_{r=1}^{R} \left(\frac{C_r}{2\tau\phi_r} + b_\tau\tau\phi_r \right) \right\} \mathrm{d}\tau \,, \qquad \text{(B.27)}
\end{aligned}
$$

where we expressed $a_\tau = \alpha R$. According to the results in Appendix A and Lemma 7.9 of Guhaniyogi et al. (2017), the function in the previous equation is the kernel of a generalized inverse Gaussian for $\psi_r = \tau\phi_r$, which yields the distribution of ϕ_r after normalization. Hence, for $r = 1, \ldots, R$, we first sample :

$$
p(\psi_r|\mathcal{B},\mathbf{W},\tau,\alpha) \sim GiG\left(\alpha - \frac{I_0}{2}, 2b_\tau, 2C_r \right) \qquad \text{(B.28)}
$$

then, renormalizing, we obtain:

$$
\phi_r = \frac{\psi_r}{\sum_{l=1}^{R} \psi_l} \,. \qquad \text{(B.29)}
$$

B.3.2 Full conditional distribution of τ

The posterior distribution of the global variance parameter, τ, is derived by simple application of Bayes' Theorem:

$$
p(\tau|\mathcal{B},\mathbf{W},\boldsymbol{\phi}) \propto \pi(\tau) p(\mathcal{B}|\mathbf{W},\boldsymbol{\phi},\tau)
$$

$$
\begin{aligned}
&\propto \tau^{a_\tau - 1} \exp\left\{-b_\tau \tau\right\} \left[\prod_{r=1}^{R} (\tau\phi_r)^{-\frac{I_0}{2}} \exp\left\{ -\frac{1}{2\tau\phi_r} \sum_{j=1}^{N} \boldsymbol{\beta}_j^{(r)'} (\mathbf{W}_{j,r})^{-1} \boldsymbol{\beta}_j^{(r)} \right\} \right] \\
&\propto \tau^{a_\tau - \frac{RI_0}{2} - 1} \exp\left\{ -b_\tau \tau - \left(\sum_{r=1}^{R} \frac{C_r}{\phi_r} \frac{1}{\tau} \right) \right\} .
\end{aligned}
\tag{B.30}
$$

This is the kernel of a generalized inverse Gaussian:

$$
p(\tau | \mathcal{B}, \mathbf{W}, \boldsymbol{\phi}) \sim GiG\left(a_\tau - \frac{RI_0}{2}, 2b_\tau, 2\sum_{r=1}^{R} \frac{C_r}{\phi_r} \right) .
\tag{B.31}
$$

B.3.3 Full conditional distribution of $\lambda_{j,r}$

Start by observing that, for $j = 1, 2, 3$ and $r = 1, \ldots, R$, the prior distribution on the vector $\boldsymbol{\beta}_j^{(r)}$ defined in eq. (2.29e) implies that each component follows a double exponential distribution:

$$
\beta_{j,p}^{(r)} \sim D\mathbf{E}\left(0, \frac{\lambda_{j,r}}{\sqrt{\tau\phi_r}} \right)
\tag{B.32}
$$

with probability density function, for $j = 1, 2, 3$:

$$
\pi(\beta_{j,p}^{(r)} | \lambda_{j,r}, \phi_r, \tau) = \frac{\lambda_{j,r}}{2\sqrt{\tau\phi_r}} \exp\left\{ -\frac{\left|\beta_{j,p}^{(r)}\right|}{(\lambda_{j,r}/\sqrt{\tau\phi_r})^{-1}} \right\} .
\tag{B.33}
$$

Then, exploiting the prior distribution $\pi(\lambda_{j,r}) \sim \mathcal{G}a(a_\lambda, b_\lambda)$ and eq. (B.33):

$$
\begin{aligned}
p\left(\lambda_{j,r} | \boldsymbol{\beta}_j^{(r)}, \phi_r, \tau\right) &\propto \pi(\lambda_{j,r}) p\left(\boldsymbol{\beta}_j^{(r)} | \lambda_{j,r}, \phi_r, \tau\right) \\
&\propto \lambda_{j,r}^{a_\lambda - 1} \exp\left\{-b_\lambda \lambda_{j,r}\right\} \prod_{p=1}^{I_j} \frac{\lambda_{j,r}}{2\sqrt{\tau\phi_r}} \exp\left\{ -\frac{\left|\beta_{j,p}^{(r)}\right|}{(\lambda_{j,r}/\sqrt{\tau\phi_r})^{-1}} \right\} \\
&= \lambda_{j,r}^{a_\lambda - 1} \left(\frac{\lambda_{j,r}}{2\sqrt{\tau\phi_r}} \right)^{I_j} \exp\left\{-b_\lambda \lambda_{j,r}\right\} \exp\left\{ -\frac{\sum_{p=1}^{I_j} \left|\beta_{j,p}^{(r)}\right|}{\sqrt{\tau\phi_r}/\lambda_{j,r}} \right\} \\
&\propto \lambda_{j,r}^{(a_\lambda + I_j) - 1} \exp\left\{ -\left(b_\lambda + \frac{\left\|\boldsymbol{\beta}_j^{(r)}\right\|_1}{\sqrt{\tau\phi_r}} \right) \lambda_{j,r} \right\} .
\end{aligned}
\tag{B.34}
$$

This is the kernel of a gamma distribution, hence for $j = 1, 2, 3$, $r = 1, \ldots, R$:

$$
p(\lambda_{j,r} | \mathcal{B}, \phi_r, \tau) \sim \mathcal{G}a\left(a_\lambda + I_j, b_\lambda + \frac{\left\|\boldsymbol{\beta}_j^{(r)}\right\|_1}{\sqrt{\tau\phi_r}} \right) .
\tag{B.35}
$$

B.3.4 Full conditional distribution of $w_{j,r,p}$

We sample independently each component $w_{j,r,p}$ of the matrix $\mathbf{W}_{j,r} = \mathrm{diag}(\mathbf{w}_{j,r})$, for $p = 1, \dots, I_j$, $j = 1,2,3$ and $r = \dots, R$, from the full conditional distribution:

$$\begin{aligned} p\left(w_{j,r,p} | \boldsymbol{\beta}_j^{(r)}, \lambda_{j,r}, \phi_r, \tau\right) &\propto p\left(\beta_{j,p}^{(r)} | w_{j,r,p}, \phi_r, \tau\right) \pi(w_{j,r,p} | \lambda_{j,r}) \\ &= (\tau\phi_r)^{-\frac{1}{2}} w_{j,r,p}^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\tau\phi_r} \beta_{j,p}^{(r)2} w_{j,r,p}^{-1} \right\} \frac{\lambda_{j,r}^2}{2} \exp\left\{ -\frac{\lambda_{j,r}^2}{2} w_{j,r,p} \right\} \\ &\propto w_{j,r,p}^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda_{j,r}^2}{2} w_{j,r,p} - \frac{\beta_{j,p}^{(r)2}}{2\tau\phi_r} w_{j,r,p}^{-1} \right\}, \end{aligned} \tag{B.36}$$

where the second row comes from the fact that $w_{j,r,p}$ influences only the p-th component of the vector $\boldsymbol{\beta}_j^{(r)}$. For $p = 1, \dots, I_j$, $j = 1,2,3$ and $r = 1, \dots, R$ we get:

$$p\left(w_{j,r,p} | \boldsymbol{\beta}_j^{(r)}, \lambda_{j,r}, \phi_r, \tau\right) \sim \left(\frac{1}{2}, \lambda_{j,r}^2, \frac{\beta_{j,p}^{(r)2}}{\tau\phi_r} \right). \tag{B.37}$$

B.3.5 Full conditional distribution of the PARAFAC marginals $\boldsymbol{\beta}_j^{(r)}$, for $j = 1,2,3$

For $r = 1, \dots, R$ we sample the PARAFAC marginals $(\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \boldsymbol{\beta}_3^{(r)})$ fro their full conditional distribution, since their joint distribution is not available in closed form. First, it is necessary to rewrite the likelihood function in a suitable way. To this aim, for $j = 1,2,3$ and $r = 1, \dots, R$ define $\boldsymbol{\beta}_{-j}^{(r)} = \left\{ \boldsymbol{\beta}_i^{(r)} : i \neq j \right\}$, $\mathcal{B}_r = \boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}$ and $\mathcal{B}_{-r} = \{\mathcal{B}_i : i \neq r\}$. By properties of the mode-n product:

$$\mathcal{B} \times_3 \mathbf{x}_t = \left(\sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \times_3 \mathbf{x}_t = \left(\sum_{\substack{s=1 \\ s \neq r}}^{R} \boldsymbol{\beta}_1^{(s)} \circ \boldsymbol{\beta}_2^{(s)} \circ \boldsymbol{\beta}_3^{(s)} \right) \times_3 \mathbf{x}_t + \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \times_3 \mathbf{x}_t . \tag{B.38}$$

Since our interest is in $\boldsymbol{\beta}_j^{(r)}$ for $j = 1,2,3$, we focus on the second term of eq. (B.38):

$$\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \times_3 \mathbf{x}_t = \sum_{i_3=1}^{I_3} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \cdot \beta_{3,i_3}^{(r)} x_{t,i_3} = \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \cdot \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle . \tag{B.39}$$

The equality comes from the definition of mode-n product given in eq. (A.4). It holds:

$$\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \times_3 \mathbf{x}_t = \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \cdot \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle = \boldsymbol{\beta}_1^{(r)} \circ \left(\boldsymbol{\beta}_2^{(r)} \cdot \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) \tag{B.40}$$

$$= \left(\boldsymbol{\beta}_1^{(r)} \cdot \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) \circ \boldsymbol{\beta}_2^{(r)}. \tag{B.41}$$

We exploited the fact that the outcome of the inner product is a scalar, then the result follows by linearity of the outer product.

Given a sample of length T and assuming that the distribution at time $t = 0$ is known (as standard practice in time series analysis), the likelihood function is given by:

$$\begin{aligned} L\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\right) &= \prod_{t=1}^{T} \left[(2\pi)^{-\frac{k^2}{2}} \left|\boldsymbol{\Sigma}_2\right|^{-\frac{k}{2}} \left|\boldsymbol{\Sigma}_1\right|^{-\frac{k}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1} \left(\mathbf{Y}_t - \mathcal{B} \times_3 \mathbf{x}_t\right)' \boldsymbol{\Sigma}_1^{-1} \left(\mathbf{Y}_t - \mathcal{B} \times_3 \mathbf{x}_t\right) \right) \right\} \right] \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1} \tilde{\mathbf{E}}_t' \boldsymbol{\Sigma}_1^{-1} \tilde{\mathbf{E}}_t \right) \right\} , \end{aligned} \tag{B.42}$$

with:

$$\tilde{\mathbf{E}}_t = \left(\mathbf{Y}_t - \mathcal{B}_{-r} \times_3 \mathbf{x}_t - \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) . \tag{B.43}$$

Now, we can focus on a specific r and $j = 1, 2, 3$ and derive the full conditionals of each marginal vector of the tensor $\mathcal{B}$. To make computations clear:

$$L\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\right) \propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \operatorname{tr}\left(a_{1t} + a_{2t} + b_{1t} + b_{2t} + c_t \right) \right\} , \tag{B.44}$$

where:

$$a_{1t} = -\boldsymbol{\Sigma}_2^{-1} \mathbf{Y}_t' \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \tag{B.45a}$$

$$a_{2t} = -\boldsymbol{\Sigma}_2^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \boldsymbol{\Sigma}_1^{-1} \mathbf{Y}_t \tag{B.45b}$$

$$b_{1t} = \boldsymbol{\Sigma}_2^{-1} \left(\mathcal{B}_{-r} \times_3 \mathbf{x}_t \right)' \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \tag{B.45c}$$

$$b_{2t} = \boldsymbol{\Sigma}_2^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \boldsymbol{\Sigma}_1^{-1} \left(\mathcal{B}_{-r} \times_3 \mathbf{x}_t \right) \tag{B.45d}$$

$$c_t = \boldsymbol{\Sigma}_2^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle . \tag{B.45e}$$

Exploiting linearity of the trace operator and the property $\operatorname{tr}\left(A'\right) = \operatorname{tr}\left(A\right)$, one gets:

$$\begin{aligned} p\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\right) &\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \left(\operatorname{tr}\left(a_{1t}\right) + \operatorname{tr}\left(a_{2t}\right) + \operatorname{tr}\left(b_{1t}\right) + \operatorname{tr}\left(b_{2t}\right) + \operatorname{tr}\left(c_t\right) \right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \left(2\operatorname{tr}\left(a_{1t}\right) + 2\operatorname{tr}\left(b_{1t}\right) + \operatorname{tr}\left(c_t\right) \right) \right\} . \end{aligned} \tag{B.46}$$

Consider now each term in the sum at the exponent, and exploit the property $\operatorname{tr}\left(ABC\right) = \operatorname{tr}\left(CAB\right) = \operatorname{tr}\left(BCA\right)$:

$$\begin{aligned} &\sum_{t=1}^{T} 2\operatorname{tr}\left(-\boldsymbol{\Sigma}_2^{-1} \mathbf{Y}_t' \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) + 2\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1} \left(\mathcal{B}_{-r} \times_3 \mathbf{x}_t \right)' \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) \\ &+ \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) \\ &= 2\operatorname{tr}\left(-\boldsymbol{\Sigma}_2^{-1} \left(\sum_{t=1}^{T} \mathbf{Y}_t' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \right) \\ &+ 2\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1} \left(\sum_{t=1}^{T} \left(\mathcal{B}_{-r} \times_3 \mathbf{x}_t \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \right) \right) \end{aligned}$$

$$
\begin{aligned}
&+ \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\left(\sum_{t=1}^{T}\langle\boldsymbol{\beta}_3^{(r)},\mathbf{x}_t\rangle^2\right)\right) \\
&= 2\operatorname{tr}\left(-\boldsymbol{\Sigma}_2^{-1}\left(\sum_{t=1}^{T}\mathbf{Y}_t'\langle\boldsymbol{\beta}_3^{(r)},\mathbf{x}_t\rangle\right)\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\right) \\
&+ 2\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\sum_{t=1}^{T}\left(\mathcal{B}_{-r}\times_3\mathbf{x}_t\right)'\langle\boldsymbol{\beta}_3^{(r)},\mathbf{x}_t\rangle\right)\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\right) \\
&+ \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\boldsymbol{\beta}_3^{(r)'}VV'\boldsymbol{\beta}_3^{(r)}\right), \qquad \text{(B.47)}
\end{aligned}
$$

where $V = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T] \in \mathbb{R}^{k^2 \times T}$. Hence the likelihood function is proportional to:

$$
\begin{aligned}
L\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\right) \propto \exp\Bigg\{-\frac{1}{2}\Bigg[&2\operatorname{tr}\left(-\boldsymbol{\Sigma}_2^{-1}\left(\sum_{t=1}^{T}\mathbf{Y}_t'\langle\boldsymbol{\beta}_3^{(r)},\mathbf{x}_t\rangle\right)\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\right) \\
&+ 2\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\sum_{t=1}^{T}\left(\mathcal{B}_{-r}\times_3\mathbf{x}_t\right)'\langle\boldsymbol{\beta}_3,^{(r)}\mathbf{x}_t\rangle\right)\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\right) \\
&+ \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\boldsymbol{\beta}_3^{(r)'}VV'\boldsymbol{\beta}_3^{(r)}\right)\Bigg]\Bigg\}. \qquad \text{(B.48)}
\end{aligned}
$$

It is now possible to proceed and derive the full conditional distributions of the PARAFAC marginals $\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \boldsymbol{\beta}_3^{(r)}$, for fixed r.

Full conditional distribution of $\boldsymbol{\beta}_1^{(r)}$

From eq. (B.48):

$$
\begin{aligned}
L(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2) \propto \exp\Bigg\{-\frac{1}{2}\Bigg[&-2\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\sum_{t=1}^{T}\mathbf{Y}_t'\langle\boldsymbol{\beta}_3^{(r)},\mathbf{x}_t\rangle\right)\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\right) \\
&+ 2\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\sum_{t=1}^{T}\left(\mathcal{B}_{-r}\times_3\mathbf{x}_t\right)'\langle\boldsymbol{\beta}_3^{(r)},\mathbf{x}_t\rangle\right)\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\right) \\
&+ \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\boldsymbol{\beta}_3^{(r)'}VV'\boldsymbol{\beta}_3^{(r)}\right)\Bigg]\Bigg\} \\
= \exp\Bigg\{-\frac{1}{2}\Bigg[&\boldsymbol{\beta}_3^{(r)'}VV'\boldsymbol{\beta}_3^{(r)}\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\right) \\
&-2\operatorname{tr}\left(\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\right)\boldsymbol{\Sigma}_2^{-1}\sum_{t=1}^{T}\left(\mathbf{Y}_t'-\left(\mathcal{B}_{-r}\times_3\mathbf{x}_t\right)'\right)\langle\boldsymbol{\beta}_3^{(r)},\mathbf{x}_t\rangle\right)\Bigg]\Bigg\}. \qquad \text{(B.49)}
\end{aligned}
$$

For the posterior of $\boldsymbol{\beta}_1^{(r)}$ as well as for that of $\boldsymbol{\beta}_2^{(r)}$, define:

$$\begin{aligned}\tilde{a} &= \boldsymbol{\beta}_3^{(r)\prime} V V' \boldsymbol{\beta}_3^{(r)} \\ \tilde{\mathbf{E}} &= \sum_{t=1}^{T} \left(\mathbf{Y}_t' - (\mathcal{B}_{-r} \times_3 \mathbf{x}_t)' \right) \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \, .\end{aligned}$$

In addition, exploit the fact that $\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} = \boldsymbol{\beta}_1^{(r)} \boldsymbol{\beta}_2^{(r)\prime}$. As a result, eq. (B.49) becomes:

$$\begin{aligned} L(\mathbf{Y}|\mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) &\propto \exp\left\{ -\frac{1}{2} \left[\tilde{a} \operatorname{tr}\left(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\beta}_1^{(r)} \boldsymbol{\beta}_2^{(r)\prime} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\beta}_2^{(r)} \boldsymbol{\beta}_1^{(r)\prime} \right) - 2 \operatorname{tr}\left(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\beta}_1^{(r)} \boldsymbol{\beta}_2^{(r)\prime} \boldsymbol{\Sigma}_2^{-1} \tilde{\mathbf{E}} \right) \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[\tilde{a} \left(\boldsymbol{\beta}_1^{(r)\prime} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\beta}_1^{(r)} \right) \left(\boldsymbol{\beta}_2^{(r)\prime} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\beta}_2^{(r)} \right) - 2 \boldsymbol{\beta}_2^{(r)\prime} \boldsymbol{\Sigma}_2^{-1} \tilde{\mathbf{E}} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\beta}_1^{(r)} \right] \right\} , \end{aligned} \tag{B.50}$$

where the last equality comes from the use of the previously mentioned properties of the trace as well as by recognizing that the trace of a scalar is the scalar itself (all the terms in brackets in the last expression are scalars).

Equation (B.50) serves as a basis for the derivation of both the posterior of $\boldsymbol{\beta}_1^{(r)}$ and $\boldsymbol{\beta}_2^{(r)}$. With reference to the first one, the likelihood function in eq. (B.50) can be rearranged as to form the kernel of a Gaussian. For ease of notation define $\tilde{a}_1 = \tilde{a} \left(\boldsymbol{\beta}_2^{(r)\prime} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\beta}_2^{(r)} \right)$, then from eq. (B.50) it holds:

$$L(\mathbf{Y}|\mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}_1^{(r)\prime} \left(\frac{\boldsymbol{\Sigma}_1}{\tilde{a}_1} \right)^{-1} \boldsymbol{\beta}_1^{(r)} - 2 \boldsymbol{\beta}_2^{(r)\prime} \boldsymbol{\Sigma}_2^{-1} \tilde{\mathbf{E}} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\beta}_1^{(r)} \right] \right\} . \tag{B.51}$$

By Bayes' Theorem we obtain:

$$\begin{aligned} & p\left(\boldsymbol{\beta}_1^{(r)} | \boldsymbol{\beta}_{-1}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{1,r}, \phi_r, \tau, \mathbf{Y}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \right) \propto \pi\left(\boldsymbol{\beta}_1^{(r)} | \mathbf{W}_{j,r}, \phi_r, \tau \right) L\left(\mathbf{Y} | \mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \right) \\ & \propto \exp\left\{ -\tfrac{1}{2} \boldsymbol{\beta}_1^{(r)\prime} \left(\mathbf{W}_{1,r} \phi_r \tau \right)^{-1} \boldsymbol{\beta}_1^{(r)} \right\} \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}_1^{(r)\prime} \left(\frac{\boldsymbol{\Sigma}_1}{\tilde{a}_1} \right)^{-1} \boldsymbol{\beta}_1^{(r)} - 2 \boldsymbol{\beta}_2^{(r)\prime} \boldsymbol{\Sigma}_2^{-1} \tilde{\mathbf{E}} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\beta}_1^{(r)} \right] \right\} \\ & \propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}_1^{(r)\prime} \left(\left(\mathbf{W}_{1,r} \phi_r \tau \right)^{-1} + \left(\frac{\boldsymbol{\Sigma}_1}{\tilde{a}_1} \right)^{-1} \right) \boldsymbol{\beta}_1^{(r)} - 2 \boldsymbol{\beta}_2^{(r)\prime} \boldsymbol{\Sigma}_2^{-1} \tilde{\mathbf{E}} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\beta}_1^{(r)} \right] \right\} . \end{aligned} \tag{B.52}$$

This is the kernel of a normal distribution, therefore for $r = 1, \dots, R$:

$$p\left(\boldsymbol{\beta}_1^{(r)} | \boldsymbol{\beta}_{-1}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{1,r}, \phi_r, \tau, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \mathbf{Y} \right) \sim \mathcal{N}_{I_1}\left(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_1}, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_1} \right) , \tag{B.53}$$

where:

$$\begin{aligned} \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_1} &= \left[\left(\mathbf{W}_{1,r} \phi_r \tau \right)^{-1} + \tilde{a}_1 \boldsymbol{\Sigma}_1^{-1} \right]^{-1} \\ \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_1} &= \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_1} \boldsymbol{\Sigma}_1^{-1} \tilde{\mathbf{E}}' \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\beta}_2^{(r)} \, . \end{aligned}$$

Full conditional distribution of $\beta_2^{(r)}$

Consider the likelihood function in eq. (B.50) and define $\tilde{a}_2 = \tilde{a}\left(\beta_1^{(r)\prime}\boldsymbol{\Sigma}_1^{-1}\beta_1^{(r)}\right)$. By algebraic manipulation we obtain the proportionality relation:

$$L(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2) \propto \exp\left\{-\frac{1}{2}\left[\beta_2^{(r)\prime}\left(\frac{\boldsymbol{\Sigma}_2}{\tilde{a}_2}\right)^{-1}\beta_2^{(r)} - 2\beta_2^{(r)\prime}\boldsymbol{\Sigma}_2^{-1}\tilde{\mathbf{E}}\boldsymbol{\Sigma}_1^{-1}\beta_1^{(r)}\right]\right\} . \tag{B.54}$$

Then, Bayes' theorem yields:

$$\begin{aligned}
&p\left(\beta_2^{(r)}|\beta_{-2}^{(r)},\mathcal{B}_{-r},\mathbf{W}_{2,r},\phi_r,\tau,\mathbf{Y},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\right) \propto \pi\left(\beta_1^{(r)}|\mathbf{W}_{j,r},\phi_r,\tau\right) L\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\right) \\
&\propto \exp\left\{-\frac{1}{2}\beta_2^{(r)\prime}\left(\mathbf{W}_{2,r}\phi_r\tau\right)^{-1}\beta_2^{(r)}\right\}\exp\left\{-\frac{1}{2}\left[\beta_2^{(r)\prime}\left(\frac{\boldsymbol{\Sigma}_2}{\tilde{a}_2}\right)^{-1}\beta_2^{(r)} - 2\beta_2^{(r)\prime}\boldsymbol{\Sigma}_2^{-1}\tilde{\mathbf{E}}\boldsymbol{\Sigma}_1^{-1}\beta_1^{(r)}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\beta_2^{(r)\prime}\left(\left(\mathbf{W}_{2,r}\phi_r\tau\right)^{-1} + \left(\frac{\boldsymbol{\Sigma}_2}{\tilde{a}_2}\right)^{-1}\right)\beta_2^{(r)} - 2\beta_2^{(r)\prime}\boldsymbol{\Sigma}_2^{-1}\tilde{\mathbf{E}}\boldsymbol{\Sigma}_1^{-1}\beta_1^{(r)}\right]\right\} .
\end{aligned} \tag{B.55}$$

Which, for $r = 1,\ldots,R$, is the kernel of a normal distribution:

$$p\left(\beta_2^{(r)}|\beta_{-2}^{(r)},\mathcal{B}_{-r},\mathbf{W}_{2,r},\phi_r,\tau,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\mathbf{Y}\right) \sim \mathcal{N}_{I_2}\left(\bar{\mu}_{\beta_2},\bar{\boldsymbol{\Sigma}}_{\beta_2}\right) , \tag{B.56}$$

where:

$$\begin{aligned}
\bar{\boldsymbol{\Sigma}}_{\beta_2} &= \left[\left(\mathbf{W}_{2,r}\phi_r\tau\right)^{-1} + \tilde{a}_2\boldsymbol{\Sigma}_2^{-1}\right]^{-1} \\
\bar{\mu}_{\beta_2} &= \bar{\boldsymbol{\Sigma}}_{\beta_2}\boldsymbol{\Sigma}_2^{-1}\tilde{\mathbf{E}}\boldsymbol{\Sigma}_1^{-1}\beta_1^{(r)} .
\end{aligned}$$

Full conditional distribution of $\beta_3^{(r)}$

For ease of notation, define:

$$\begin{aligned}
A &= \boldsymbol{\Sigma}_1^{-1}\left(\beta_1^{(r)}\circ\beta_2^{(r)}\right)\boldsymbol{\Sigma}_2^{-1} \\
\tilde{A} &= A\left(\beta_1^{(r)}\circ\beta_2^{(r)}\right)' .
\end{aligned}$$

Define $\tilde{V} = V\cdot(\mathrm{tr}(\tilde{A}))^{\frac{1}{2}}$, then eq. (B.48) becomes:

$$\begin{aligned}
&L(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2) \propto \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\,\mathrm{tr}\left(A\sum_{t=1}^{T}\mathbf{Y}_t'\langle\beta_3^{(r)},\mathbf{x}_t\rangle\right) + 2\,\mathrm{tr}\left(A\sum_{t=1}^{T}\left(\mathcal{B}_{-r}\times_3\mathbf{x}_t\right)'\langle\beta_3^{(r)},\mathbf{x}_t\rangle\right) + \beta_3^{(r)\prime}VV'\beta_3^{(r)}\,\mathrm{tr}\left(\tilde{A}\right)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-2\,\mathrm{tr}\left(A\sum_{t=1}^{T}\mathbf{Y}_t'\langle\beta_3^{(r)},\mathbf{x}_t\rangle\right) + 2\,\mathrm{tr}\left(A\sum_{t=1}^{T}\left(\mathcal{B}_{-r}\times_3\mathbf{x}_t\right)'\langle\beta_3^{(r)},\mathbf{x}_t\rangle\right) + \beta_3^{(r)\prime}\tilde{V}\tilde{V}'\beta_3^{(r)}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\beta_3^{(r)\prime}\tilde{V}\tilde{V}'\beta_3^{(r)} - 2\,\mathrm{tr}\left(A\sum_{t=1}^{T}\mathbf{Y}_t'\langle\beta_3^{(r)},\mathbf{x}_t\rangle - A\sum_{t=1}^{T}\left(\mathcal{B}_{-r}\times_3\mathbf{x}_t\right)'\langle\beta_3^{(r)},\mathbf{x}_t\rangle\right)\right]\right\} .
\end{aligned} \tag{B.57}$$

Then, focus on the second term in square brackets:

$$\operatorname{tr}\left(A \sum_{t=1}^{T} \mathbf{Y}_t' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle - A \sum_{t=1}^{T} \left(\mathcal{B}_{-r} \times_3 \mathbf{x}_t \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right)$$
$$= \operatorname{tr}\left(A \left(\sum_{t=1}^{T} \mathbf{Y}_t' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle - \left(\mathcal{B}_{-r} \times_3 \mathbf{x}_t \right)' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) \right) = \operatorname{tr}\left(A \sum_{t=1}^{T} \left(\mathbf{Y}_t' - \left(\mathcal{B}_{-r} \times_3 \mathbf{x}_t \right)' \right) \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) . \tag{B.58}$$

For ease of notation, define $\tilde{\tilde{Y}}_t = \mathbf{Y}_t' - (\mathcal{B}_{-r} \times_3 \mathbf{x}_t)'$, then by linearity of the trace operator:

$$= \operatorname{tr}\left(A \sum_{t=1}^{T} \tilde{\tilde{Y}}_t' \langle \boldsymbol{\beta}_3^{(r)}, \mathbf{x}_t \rangle \right) = \operatorname{tr}\left(\sum_{t=1}^{T} \left(A \tilde{\tilde{Y}}_t' \right) \left(\boldsymbol{\beta}_3^{(r)\prime} \mathbf{x}_t \right) \right) = \sum_{t=1}^{T} \operatorname{tr}\left(A \tilde{\tilde{Y}}_t \right) \left(\boldsymbol{\beta}_3^{(r)\prime} \mathbf{x}_t \right)$$
$$= \sum_{t=1}^{T} \tilde{y}_t \left(\boldsymbol{\beta}_3^{(r)\prime} \mathbf{x}_t \right) = \tilde{\mathbf{y}}' V' \boldsymbol{\beta}_3^{(r)} , \tag{B.59}$$

where we defined $\tilde{y}_t = \operatorname{tr}(A \tilde{\tilde{Y}}_t)$. As a consequence, rewrite eq. (B.57) as:

$$L(\mathbf{Y} | \mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}_3^{(r)\prime} (\tilde{V} \tilde{V}') \boldsymbol{\beta}_3^{(r)} - 2 \tilde{\mathbf{y}}' V' \boldsymbol{\beta}_3^{(r)} \right] \right\} . \tag{B.60}$$

We can now recover the full conditional posterior distribution of $\boldsymbol{\beta}_3^{(r)}$ by applying Bayes' Theorem:

$$p\left(\boldsymbol{\beta}_3^{(r)} | \boldsymbol{\beta}_{-3}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{3,r}, \phi_r, \tau, \mathbf{Y}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \right) \propto \pi\left(\boldsymbol{\beta}_3^{(r)} | \mathbf{W}_{3,r}, \phi_r, \tau \right) L\left(\mathbf{Y} | \mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \right)$$
$$\propto \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}_3^{(r)\prime} \left(\mathbf{W}_{3,r} \phi_r \tau \right)^{-1} \boldsymbol{\beta}_3^{(r)} \right\} \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}_3^{(r)\prime} (\tilde{V} \tilde{V}') \boldsymbol{\beta}_3^{(r)} - 2 \tilde{\mathbf{y}}' V' \boldsymbol{\beta}_3^{(r)} \right] \right\}$$
$$\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}_3^{(r)\prime} \left(\left(\mathbf{W}_{3,r} \phi_r \tau \right)^{-1} + \tilde{V} \tilde{V}' \right) \boldsymbol{\beta}_3^{(r)} - 2 \tilde{\mathbf{y}}' V' \boldsymbol{\beta}_3^{(r)} \right] \right\} , \tag{B.61}$$

which is the kernel of a normal distribution. As a consequence, defining:

$$\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_3} = \left[\left(\mathbf{W}_{3,r} \phi_r \tau \right)^{-1} + \tilde{V} \tilde{V}' \right]^{-1}$$
$$\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_3} = \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_3} V \tilde{\mathbf{y}} ,$$

we get, for $r = 1, \ldots, R$:

$$p\left(\boldsymbol{\beta}_3^{(r)} | \boldsymbol{\beta}_{-3}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{j,r}, \phi_r, \tau, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \mathbf{Y} \right) \sim \mathcal{N}_{I_1 \cdot I_2}\left(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_3}, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_3} \right) . \tag{B.62}$$

B.3.6 Full conditional distribution of $\boldsymbol{\Sigma}_1$

Given a inverse Wishart prior, the posterior full conditional distribution for $\boldsymbol{\Sigma}_1$ is conjugate:

$$p(\boldsymbol{\Sigma}_1 | \mathcal{B}, \mathbf{Y}, \boldsymbol{\Sigma}_2, \gamma) \propto L(\mathbf{Y} | \mathcal{B}, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_1) \pi(\boldsymbol{\Sigma}_1)$$
$$\propto |\boldsymbol{\Sigma}_1|^{-\frac{T I_2}{2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1} \left(\mathbf{Y}_t - \mathcal{B} \times_3 \mathbf{x}_t \right)' \boldsymbol{\Sigma}_1^{-1} \left(\mathbf{Y}_t - \mathcal{B} \times_3 \mathbf{x}_t \right) \right) \right\}$$

$$
\begin{aligned}
&\cdot |\mathbf{\Sigma}_1|^{-\frac{\nu_1+I_1+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\gamma\mathbf{\Psi}_1\mathbf{\Sigma}_1^{-1}\right)\right\} \\
&\propto |\mathbf{\Sigma}_1|^{-\frac{\nu_1+I_1+TI_2+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_1\mathbf{\Sigma}_1^{-1}\right)+\operatorname{tr}\left(\sum_{t=1}^{T}\left(\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)\mathbf{\Sigma}_2^{-1}\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)'\right)\mathbf{\Sigma}_1^{-1}\right)\right]\right\} .
\end{aligned}
\tag{B.63}
$$

The last row comes from exploiting two ties the linearity of the trace operator. For ease of notation, define $S_1 = \sum_{t=1}^{T}\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)\mathbf{\Sigma}_2^{-1}\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)'$, obtaining:

$$
\begin{aligned}
p(\mathbf{\Sigma}_1|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_2,\gamma) &\propto |\mathbf{\Sigma}_1|^{-\frac{\nu_1+I_1+TI_2+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_1\mathbf{\Sigma}_1^{-1}\right)+\operatorname{tr}\left(S_1\mathbf{\Sigma}_1^{-1}\right)\right]\right\} \\
&\propto |\mathbf{\Sigma}_1|^{-\frac{(\nu_1+TI_2)+I_1+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(\gamma\mathbf{\Psi}_1+S_1\right)\mathbf{\Sigma}_1^{-1}\right)\right\} ,
\end{aligned}
\tag{B.64}
$$

where we have used again the linearity of the trace operator. As a consequence:

$$
p(\mathbf{\Sigma}_1|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_2,\gamma) \sim \mathcal{IW}_{I_1}\left(\nu_1+TI_2,\gamma\mathbf{\Psi}_1+S_1\right) . \tag{B.65}
$$

B.3.7 Full conditional distribution of $\mathbf{\Sigma}_2$

By the same reasoning of $\mathbf{\Sigma}_1$, the posterior full conditional distribution of $\mathbf{\Sigma}_2$ is conjugate and follows from:

$$
\begin{aligned}
&p(\mathbf{\Sigma}_2|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_1,\gamma) \propto L(\mathbf{Y}|\mathcal{B},\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)\pi(\mathbf{\Sigma}_2|\gamma) \\
&\propto |\mathbf{\Sigma}_2|^{-\frac{TI_1}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left(\mathbf{\Sigma}_2^{-1}\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)'\mathbf{\Sigma}_1^{-1}\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)\right)\right\} \\
&\cdot |\mathbf{\Sigma}_2|^{-\frac{\nu_2+I_2+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\gamma\mathbf{\Psi}_2\mathbf{\Sigma}_2^{-1}\right)\right\} \\
&\propto |\mathbf{\Sigma}_2|^{-\frac{\nu_2+I_2+TI_1+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_2\mathbf{\Sigma}_2^{-1}\right)+\operatorname{tr}\left(\sum_{t=1}^{T}\left(\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)'\mathbf{\Sigma}_1^{-1}\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)\right)\mathbf{\Sigma}_2^{-1}\right)\right]\right\} .
\end{aligned}
\tag{B.66}
$$

The last row comes from exploiting two ties the linearity of the trace operator. For ease of notation, define $S_2 = \sum_{t=1}^{T}\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)'\mathbf{\Sigma}_1^{-1}\left(\mathbf{Y}_t-\mathcal{B}\times_3\mathbf{x}_t\right)$, obtaining:

$$
\begin{aligned}
p(\mathbf{\Sigma}_2|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_1,\gamma) &\propto |\mathbf{\Sigma}_2|^{-\frac{\nu_2+I_2+I_1k+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_2\mathbf{\Sigma}_2^{-1}\right)+\operatorname{tr}\left(S_2\mathbf{\Sigma}_2^{-1}\right)\right]\right\} \\
&\propto |\mathbf{\Sigma}_2|^{-\frac{(\nu_2+TI_1)+I_2+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(\gamma\mathbf{\Psi}_2+S_2\right)\mathbf{\Sigma}_2^{-1}\right)\right\} ,
\end{aligned}
\tag{B.67}
$$

where we have used again the linearity of the trace operator. As a consequence:

$$
p(\mathbf{\Sigma}_2|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_1) \sim \mathcal{IW}_{I_2}\left(\nu_2+TI_1,\gamma\mathbf{\Psi}_2+S_2\right) . \tag{B.68}
$$

B.3.8 Full conditional distribution of γ

Using a gamma prior distribution we have:

$$
p(\gamma|\mathbf{\Sigma}_1,\mathbf{\Sigma}_2) \propto p(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2|\gamma)\pi(\gamma)
$$

$$\begin{aligned}
&\propto |\gamma \boldsymbol{\Psi}_1|^{-\frac{\nu_1}{2}} |\gamma \boldsymbol{\Psi}_2|^{-\frac{\nu_2}{2}} \exp\left\{-\frac{1}{2} \operatorname{tr}\left(\gamma \boldsymbol{\Psi}_1 \boldsymbol{\Sigma}_1^{-1}\right)\right\} \exp\left\{-\frac{1}{2} \operatorname{tr}\left(\gamma \boldsymbol{\Psi}_2 \boldsymbol{\Sigma}_2^{-1}\right)\right\} \gamma^{a_\gamma - 1} \exp\{-b_\gamma \gamma\} \\
&\propto \gamma^{-\frac{\nu_1 I_1 + \nu_2 I_2}{2}} \exp\left\{-\frac{1}{2}\gamma \operatorname{tr}\left(\boldsymbol{\Psi}_1 \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Psi}_2 \boldsymbol{\Sigma}_2^{-1}\right) - b_\gamma \gamma\right\} \gamma^{a_\gamma - 1} \\
&\propto \gamma^{a_\gamma - \frac{\nu_1 I_1 + \nu_2 I_2}{2} - 1} \exp\left\{-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Psi}_1 \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Psi}_2 \boldsymbol{\Sigma}_2^{-1}\right) - b_\gamma \gamma\right\}, \qquad \text{(B.69)}
\end{aligned}$$

thus:

$$p(\gamma|\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \sim \mathcal{G}a\left(a_\gamma + \frac{1}{2}(\nu_1 I_1 + \nu_2 I_2), b_\gamma + \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Psi}_1 \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Psi}_2 \boldsymbol{\Sigma}_2^{-1}\right)\right). \qquad \text{(B.70)}$$

B.4 Computational details - tensor case

In this section we will follow the convention of denoting the prior distributions with $\pi(\cdot)$. In addition, let $\mathbf{W} = \{\mathbf{W}_{j,r}\}_{j,r}$ be the collection of all (local variance) matrices $\mathbf{W}_{j,r}$, for $j = 1, 2, 3, 4$ and $r = 1, \ldots, R$; $I_0 = \sum_{j=1}^{4} I_j$ the sum of the length of each mode of the tensor $\mathcal{B}$ and $\mathbf{Y} = \{\mathcal{Y}_t, \mathcal{X}_t\}_t$ the collection of observed variables.

B.4.1 Full conditional distribution of $\boldsymbol{\phi}$

In order to derive this posterior distribution, we make use of Lemma 7.9 in Guhaniyogi et al. (2017). Recall that: $a_\tau = \alpha R$, $b_\tau = \alpha (R)^{1/N}$ and $I_0 = \sum_{j=1}^{N} I_j$. The prior for $\boldsymbol{\phi}$ is $\pi(\boldsymbol{\phi}) \sim \mathcal{D}ir(\boldsymbol{\alpha})$.

$$p(\boldsymbol{\phi}|\mathcal{B}, \mathbf{W}) \propto \pi(\boldsymbol{\phi}) p(\mathcal{B}|\mathbf{W}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi}) \int_0^{+\infty} p(\mathcal{B}|\mathbf{W}, \boldsymbol{\phi}, \tau) \pi(\tau) \mathrm{d}\tau. \qquad \text{(B.71)}$$

By plugging in the prior distributions for τ, $\boldsymbol{\phi}$, $\boldsymbol{\beta}_j^{(r)}$ we obtain[2]:

$$\begin{aligned}
p(\boldsymbol{\phi}|\mathcal{B}, \mathbf{W}) &\propto \prod_{r=1}^{R} \phi_r^{\alpha-1} \int_0^{+\infty} \left[\prod_{r=1}^{R} \prod_{j=1}^{N} (\tau \phi_r)^{-I_j/2} \left|\mathbf{W}_{j,r}\right|^{-1/2} \exp\left\{-\frac{1}{2\tau\phi_r} \boldsymbol{\beta}_j^{(r)\prime} \mathbf{W}_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)}\right\}\right] \\
&\quad \cdot \tau^{a_\tau - 1} \exp\{-b_\tau \tau\} \mathrm{d}\tau \\
&\propto \prod_{r=1}^{R} \phi_r^{\alpha-1} \int_0^{+\infty} \left[\prod_{r=1}^{R} (\tau \phi_r)^{-I_0/2} \exp\left\{-\frac{1}{2\tau\phi_r} \sum_{j=1}^{N} \boldsymbol{\beta}_j^{(r)\prime} \mathbf{W}_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)}\right\}\right] \\
&\quad \cdot \tau^{a_\tau - 1} \exp\{-b_\tau \tau\} \mathrm{d}\tau. \qquad \text{(B.72)}
\end{aligned}$$

Define $C_r = \sum_{j=1}^{N} \boldsymbol{\beta}_j^{(r)\prime} \mathbf{W}_{j,r}^{-1} \boldsymbol{\beta}_j^{(r)}$, then group together the powers of τ and ϕ_r as follows:

$$\begin{aligned}
p(\boldsymbol{\phi}|\mathcal{B}, \mathbf{W}) &\propto \prod_{r=1}^{R} \phi_r^{\alpha - 1 - \frac{I_0}{2}} \int_0^{+\infty} \tau^{a_\tau - 1 - \frac{R I_0}{2}} \exp\{-b_\tau \tau\} \left[\prod_{r=1}^{R} \exp\left\{-\frac{1}{2\tau\phi_r} C_r\right\}\right] \mathrm{d}\tau \\
&= \prod_{r=1}^{R} \phi_r^{\alpha - 1 - \frac{I_0}{2}} \int_0^{+\infty} \tau^{a_\tau - 1 - \frac{R d_0}{2}} \exp\left\{-b_\tau \tau - \sum_{r=1}^{R} \frac{C_r}{2\tau\phi_r}\right\} \mathrm{d}\tau. \qquad \text{(B.73)}
\end{aligned}$$

[2] We have used the property of the determinant: $\det(kA) = k^n \det(A)$, for A square matrix of size n and k scalar.

Recall that the probability density function of a Generalized Inverse Gaussian in the parametrization with three parameters ($a > 0$, $b > 0$, $p \in \mathbb{R}$), with $x \in (0, +\infty)$, is given by:

$$x \sim GiG(a, b, p) \;\Rightarrow\; p(x|a, b, p) = \frac{\left(\frac{a}{b}\right)^{\frac{p}{2}}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left\{ -\frac{1}{2}\left(ax + \frac{b}{x} \right) \right\}, \tag{B.74}$$

with $K_p(\cdot)$ a modified Bessel function of the second type. Our goal is to reconcile eq. (B.73) to the kernel of this distribution. Since by definition $\sum_{r=1}^R \phi_r = 1$, it holds that $\sum_{r=1}^R (b_\tau \tau \phi_r) = (b_\tau \tau) \sum_{r=1}^R \phi_r = b_\tau \tau$. This allows to rewrite the exponential as:

$$\begin{aligned} p(\boldsymbol{\phi}|\mathcal{B}, \mathbf{W}) &\propto \prod_{r=1}^R \phi_r^{\alpha - 1 - \frac{I_0}{2}} \int_0^{+\infty} \tau^{\left(a_\tau - \frac{RI_0}{2}\right) - 1} \exp\left\{ -\sum_{r=1}^R \left(\frac{C_r}{2\tau\phi_r} + b_\tau \tau \phi_r \right) \right\} \mathrm{d}\tau \\ &= \int_0^{+\infty} \left(\prod_{r=1}^R \phi_r^{\alpha - \frac{I_0}{2} - 1} \right) \tau^{\left(\alpha R - \frac{RI_0}{2}\right) - 1} \exp\left\{ -\sum_{r=1}^R \left(\frac{C_r}{2\tau\phi_r} + b_\tau \tau \phi_r \right) \right\} \mathrm{d}\tau, \end{aligned} \tag{B.75}$$

where we expressed $a_\tau = \alpha R$. According to the results in Appendix A and Guhaniyogi et al. (2017), the function in the previous equation is the kernel of a generalized inverse Gaussian for $\psi_r = \tau\phi_r$, which yields the distribution of ϕ_r after normalization. Hence, for $r = 1, \ldots, R$, we first sample :

$$p(\psi_r|\mathcal{B}, \mathbf{W}, \tau, \alpha) \sim GiG\left(\alpha - \frac{I_0}{2}, 2b_\tau, 2C_r \right) \tag{B.76}$$

then, renormalizing, we obtain (see Kruijer et al. (2010)):

$$\phi_r = \frac{\psi_r}{\sum_{l=1}^R \psi_l}. \tag{B.77}$$

B.4.2 Full conditional distribution of τ

The posterior distribution of the global variance parameter, τ, is derived by simple application of Bayes' Theorem:

$$\begin{aligned} p(\tau|\mathcal{B}, \mathbf{W}, \boldsymbol{\phi}) &\propto \pi(\tau) p(\mathcal{B}|\mathbf{W}, \boldsymbol{\phi}, \tau) \\ &\propto \tau^{a_\tau - 1} \exp\{-b_\tau \tau\} \left[\prod_{r=1}^R (\tau\phi_r)^{-\frac{I_0}{2}} \exp\left\{ -\frac{1}{2\tau\phi_r} \sum_{j=1}^4 \boldsymbol{\beta}_j^{(r)\prime} (\mathbf{W}_{j,r})^{-1} \boldsymbol{\beta}_j^{(r)} \right\} \right] \\ &\propto \tau^{a_\tau - \frac{RI_0}{2} - 1} \exp\left\{ -b_\tau \tau - \left(\sum_{r=1}^R \frac{C_r}{\phi_r} \frac{1}{\tau} \right) \right\}. \end{aligned} \tag{B.78}$$

This is the kernel of a generalized inverse Gaussian:

$$p(\tau|\mathcal{B}, \mathbf{W}, \boldsymbol{\phi}) \sim GiG\left(a_\tau - \frac{RI_0}{2}, 2b_\tau, 2\sum_{r=1}^R \frac{C_r}{\phi_r} \right). \tag{B.79}$$

B.4.3 Full conditional distribution of $\lambda_{j,r}$

Start by observing that, for $j = 1,2,3,4$ and $r = 1,\ldots,R$, the prior distribution on the vector $\boldsymbol{\beta}_j^{(r)}$ defined in eq. (2.29e) implies that each component follows a double exponential distribution:

$$\beta_{j,p}^{(r)} \sim D\mathbf{E}\left(0, \frac{\lambda_{j,r}}{\sqrt{\tau\phi_r}}\right) \tag{B.80}$$

with probability density function, for $j = 1,2,3,4$ and $r = 1,\ldots,R$, given by:

$$\pi(\beta_{j,p}^{(r)}|\lambda_{j,r},\phi_r,\tau) = \frac{\lambda_{j,r}}{2\sqrt{\tau\phi_r}} \exp\left\{-\frac{\left|\beta_{j,p}^{(r)}\right|}{(\lambda_{j,r}/\sqrt{\tau\phi_r})^{-1}}\right\}. \tag{B.81}$$

Then, exploiting the prior $\pi(\lambda_{j,r}) \sim \mathcal{G}a(a_\lambda, b_\lambda)$ and eq. (B.81):

$$\begin{aligned}
p\left(\lambda_{j,r}|\boldsymbol{\beta}_j^{(r)},\phi_r,\tau\right) &\propto \pi(\lambda_{j,r})p\left(\boldsymbol{\beta}_j^{(r)}|\lambda_{j,r},\phi_r,\tau\right) \\
&\propto \lambda_{j,r}^{a_\lambda-1}\exp\left\{-b_\lambda\lambda_{j,r}\right\}\prod_{p=1}^{I_j}\frac{\lambda_{j,r}}{2\sqrt{\tau\phi_r}}\exp\left\{-\frac{\left|\beta_{j,p}^{(r)}\right|}{(\lambda_{j,r}/\sqrt{\tau\phi_r})^{-1}}\right\} \\
&= \lambda_{j,r}^{a_\lambda-1}\left(\frac{\lambda_{j,r}}{2\sqrt{\tau\phi_r}}\right)^{I_j}\exp\left\{-b_\lambda\lambda_{j,r}\right\}\exp\left\{-\frac{\sum_{p=1}^{I_j}\left|\beta_{j,p}^{(r)}\right|}{\sqrt{\tau\phi_r}/\lambda_{j,r}}\right\} \\
&\propto \lambda_{j,r}^{(a_\lambda+I_j)-1}\exp\left\{-\left(b_\lambda+\frac{\left\|\boldsymbol{\beta}_j^{(r)}\right\|_1}{\sqrt{\tau\phi_r}}\right)\lambda_{j,r}\right\}.
\end{aligned} \tag{B.82}$$

Thus, for $j = 1,2,3,4$, $r = 1,\ldots,R$, the full conditional distribution of $\lambda_{j,r}$ is given by:

$$p(\lambda_{j,r}|\mathcal{B},\phi_r,\tau) \sim \mathcal{G}a\left(a_\lambda + I_j, b_\lambda + \frac{\left\|\boldsymbol{\beta}_j^{(r)}\right\|_1}{\sqrt{\tau\phi_r}}\right). \tag{B.83}$$

B.4.4 Full conditional distribution of $w_{j,r,p}$

We sample independently each component $w_{j,r,p}$ of the matrix $\mathbf{W}_{j,r} = \mathrm{diag}(\mathbf{w}_{j,r})$, for $p = 1,\ldots,I_j$, $j = 1,2,3,4$ and $r = \ldots,R$, from the full conditional distribution:

$$\begin{aligned}
p\left(w_{j,r,p}|\boldsymbol{\beta}_j^{(r)},\lambda_{j,r},\phi_r,\tau\right) &\propto p\left(\beta_{j,p}^{(r)}|w_{j,r,p},\phi_r,\tau\right)\pi(w_{j,r,p}|\lambda_{j,r}) \\
&= (\tau\phi_r)^{-\frac{1}{2}}w_{j,r,p}^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\tau\phi_r}\beta_{j,p}^{(r)2}w_{j,r,p}^{-1}\right\}\frac{\lambda_{j,r}^2}{2}\exp\left\{-\frac{\lambda_{j,r}^2}{2}w_{j,r,p}\right\} \\
&\propto w_{j,r,p}^{-\frac{1}{2}}\exp\left\{-\frac{\lambda_{j,r}^2}{2}w_{j,r,p}-\frac{\beta_{j,p}^{(r)2}}{2\tau\phi_r}w_{j,r,p}^{-1}\right\},
\end{aligned} \tag{B.84}$$

where the second row comes from the fact that $w_{j,r,p}$ influences only the p-th component of the vector $\boldsymbol{\beta}_j^{(r)}$. For $p = 1, \ldots, I_j$, $j = 1, 2, 3, 4$ and $r = 1, \ldots, R$ we get:

$$p\left(w_{j,r,p} | \boldsymbol{\beta}_j^{(r)}, \lambda_{j,r}, \phi_r, \tau\right) \sim \left(\frac{1}{2}, \lambda_{j,r}^2, \frac{\beta_{j,p}^{(r)^2}}{\tau\phi_r}\right) . \tag{B.85}$$

B.4.5 Full conditional distributions of PARAFAC marginals $\boldsymbol{\beta}_j^{(r)}$, for $j = 1, 2, 3, 4$

Define $\boldsymbol{\alpha}_1 \in \mathbb{R}^I$, $\boldsymbol{\alpha}_2 \in \mathbb{R}^J$ and $\boldsymbol{\alpha}_3 \in \mathbb{R}^K$ and let $\mathcal{A} = \text{vec}\left(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2 \circ \boldsymbol{\alpha}_3\right)$. Then it holds:

$$\begin{align}
\text{vec}\left(\mathcal{A}\right) &= \text{vec}\left(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2 \circ \boldsymbol{\alpha}_3\right) = \boldsymbol{\alpha}_3 \otimes \text{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right) \notag \\
&= \boldsymbol{\alpha}_3 \otimes \left(\boldsymbol{\alpha}_2 \otimes \mathbf{I}_I\right) \text{vec}\left(\boldsymbol{\alpha}_1\right) = \left(\boldsymbol{\alpha}_3 \otimes \boldsymbol{\alpha}_2 \otimes \mathbf{I}_I\right) \boldsymbol{\alpha}_1 \tag{B.86} \\
&= \boldsymbol{\alpha}_3 \otimes \left[\left(\mathbf{I}_J \otimes \boldsymbol{\alpha}_1\right) \text{vec}\left(\boldsymbol{\alpha}_2'\right)\right] = \left(\boldsymbol{\alpha}_3 \otimes \mathbf{I}_J \otimes \boldsymbol{\alpha}_1\right) \boldsymbol{\alpha}_2 \tag{B.87} \\
&= \text{vec}\left(\text{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right) \boldsymbol{\alpha}_3'\right) = \left(\mathbf{I}_K \otimes \text{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right)\right) \text{vec}\left(\boldsymbol{\alpha}_3'\right) \notag \\
&= \left(\mathbf{I}_K \otimes \text{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right)\right) \boldsymbol{\alpha}_3 = \left(\mathbf{I}_K \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_1\right) \boldsymbol{\alpha}_3 . \tag{B.88}
\end{align}$$

Consider the model in eq. (2.26), it holds:

$$\begin{align}
\mathcal{Y}_t &= \mathcal{B} \times_4 \mathbf{x}_t + \mathbf{E}_t \notag \\
\text{vec}\left(\mathcal{Y}_t\right) &= \text{vec}\left(\mathcal{B} \times_4 \mathbf{x}_t + \mathbf{E}_t\right) \notag \\
&= \text{vec}\left(\mathcal{B}_{-r} \times_4 \mathbf{x}_t\right) + \text{vec}\left(\mathcal{B}_r \times_4 \mathbf{x}_t\right) + \text{vec}\left(\mathbf{E}_t\right) \notag \\
&\propto \text{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\right) \cdot \mathbf{x}_t' \boldsymbol{\beta}_4^{(r)} . \tag{B.89}
\end{align}$$

It is then possible to make explicit the dependence on each PARAFAC marginal by exploiting the results in eq. (B.86)-(B.88), as follows:

$$\begin{align}
\text{vec}\left(\mathcal{Y}_t\right) &\propto \text{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\right) \cdot \mathbf{x}_t' \boldsymbol{\beta}_4^{(r)} = \mathbf{b}_4 \boldsymbol{\beta}_4^{(r)} \tag{B.90} \\
&\propto \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \left(\boldsymbol{\beta}_3^{(r)} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \mathbf{I}_I\right) \boldsymbol{\beta}_1^{(r)} = \mathbf{b}_1 \boldsymbol{\beta}_1^{(r)} \tag{B.91} \\
&\propto \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \left(\boldsymbol{\beta}_3^{(r)} \otimes \mathbf{I}_J \otimes \boldsymbol{\beta}_1^{(r)}\right) \boldsymbol{\beta}_2^{(r)} = \mathbf{b}_2 \boldsymbol{\beta}_2^{(r)} \tag{B.92} \\
&\propto \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \left(\mathbf{I}_K \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right) \boldsymbol{\beta}_3^{(r)} = \mathbf{b}_3 \boldsymbol{\beta}_3^{(r)} . \tag{B.93}
\end{align}$$

Given a sample of length T and assuming that the distribution at time $t = 0$ is known (as standard practice in time series analysis), the likelihood function is given by:

$$\begin{align}
L\left(\mathbf{Y} | \mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3\right) &= \prod_{t=1}^{T} (2\pi)^{-\frac{k^2 q}{2}} \left|\boldsymbol{\Sigma}_3\right|^{-\frac{k^2}{2}} \left|\boldsymbol{\Sigma}_2\right|^{-\frac{kq}{2}} \left|\boldsymbol{\Sigma}_1\right|^{-\frac{kq}{2}} \notag \\
&\quad \cdot \exp\left\{-\frac{1}{2}(\mathcal{Y}_t - \mathcal{B} \times_4 \mathbf{x}_t) \times^{1\ldots3} \left(\circ_{j=1}^{3} \boldsymbol{\Sigma}_j^{-1}\right) \times^{1\ldots3} (\mathcal{Y}_t - \mathcal{B} \times_4 \mathbf{x}_t)\right\} \tag{B.94} \\
&\propto \exp\left\{-\frac{1}{2} \sum_{t=1}^{T} \tilde{\mathbf{E}}_t \times^{1\ldots3} (\boldsymbol{\Sigma}_1^{-1} \circ \boldsymbol{\Sigma}_2^{-1} \circ \boldsymbol{\Sigma}_3^{-1}) \times^{1\ldots3} \tilde{\mathbf{E}}_t\right\} , \tag{B.95}
\end{align}$$

with:

$$\tilde{\mathbf{E}}_t = \left(\mathcal{Y}_t - \mathcal{B}_{-r} \times_4 \mathbf{x}_t - \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \right) . \tag{B.96}$$

Alternatively, by exploiting the relation between the tensor normal distribution and the multivariate normal distribution, we have:

$$\begin{aligned} L\left(\mathbf{Y}|\mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3\right) &= \prod_{t=1}^{T} (2\pi)^{-\frac{k^2 q}{2}} \left| \boldsymbol{\Sigma}_3 \otimes \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1 \right|^{-\frac{1}{2}} \\ &\cdot \exp\left\{ -\frac{1}{2} \mathrm{vec}\left(\mathcal{Y}_t - \mathcal{B} \times_4 \mathbf{x}_t \right)' \left(\boldsymbol{\Sigma}_3^{-1} \otimes \boldsymbol{\Sigma}_2^{-1} \otimes \boldsymbol{\Sigma}_1^{-1} \right) \mathrm{vec}\left(\mathcal{Y}_t - \mathcal{B} \times_4 \mathbf{x}_t \right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \mathrm{vec}\left(\tilde{\mathbf{E}}_t \right)' \left(\boldsymbol{\Sigma}_3^{-1} \otimes \boldsymbol{\Sigma}_2^{-1} \otimes \boldsymbol{\Sigma}_1^{-1} \right) \mathrm{vec}\left(\tilde{\mathbf{E}}_t \right) \right\} , \end{aligned} \tag{B.97}$$

where: with:

$$\begin{aligned} \mathrm{vec}\left(\tilde{\mathbf{E}}_t \right) &= \mathrm{vec}\left(\mathcal{Y}_t - \mathcal{B}_{-r} \times_4 \mathbf{x}_t - \left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \right) \\ &= \mathrm{vec}\left(\mathcal{Y}_t \right) - \mathrm{vec}\left(\mathcal{B}_{-r} \times_4 \mathbf{x}_t \right) - \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \\ &\propto \mathrm{vec}\left(\mathcal{Y}_t \right) - \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle . \end{aligned} \tag{B.98}$$

Thus, defining $\mathbf{y}_t = \mathrm{vec}\left(\mathcal{Y}_t \right)$ and $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_3^{-1} \otimes \boldsymbol{\Sigma}_2^{-1} \otimes \boldsymbol{\Sigma}_1^{-1}$, one gets:

$$\begin{aligned} L\left(\mathbf{Y}|\mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3\right) &\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \mathrm{vec}\left(\tilde{\mathbf{E}}_t \right)' \left(\boldsymbol{\Sigma}_3^{-1} \otimes \boldsymbol{\Sigma}_2^{-1} \otimes \boldsymbol{\Sigma}_1^{-1} \right) \mathrm{vec}\left(\tilde{\mathbf{E}}_t \right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \left[\mathrm{vec}\left(\mathcal{Y}_t \right) - \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \right]' \boldsymbol{\Sigma}^{-1} \right. \\ &\left. \cdot \left[\mathrm{vec}\left(\mathcal{Y}_t \right) - \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \right] \right\} \\ &= \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \mathbf{y}_t' \boldsymbol{\Sigma}^{-1} \mathbf{y}_t - \mathbf{y}_t' \boldsymbol{\Sigma}^{-1} \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \right. \\ &- \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right)' \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \boldsymbol{\Sigma}^{-1} \mathbf{y}_t \\ &\left. + \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right)' \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \boldsymbol{\Sigma}^{-1} \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} -2\mathbf{y}_t' \boldsymbol{\Sigma}^{-1} \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \right. \\ &\left. + \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right)' \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \boldsymbol{\Sigma}^{-1} \mathrm{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)} \right) \langle \boldsymbol{\beta}_4^{(r)}, \mathbf{x}_t \rangle \right\} . \end{aligned} \tag{B.99}$$

Now, we focus on a specific $j = 1, 2, 3, 4$ and derive proportionality results which will be necessary to obtain the posterior full conditional distributions of the PARAFAC marginals of the tensor $\mathcal{B}$. Consider the case $j = 1$. By exploiting eq. (B.91) we get:

$$\begin{aligned}
L\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_3\right) \propto \exp\Bigg\{ & -\frac{1}{2}\sum_{t=1}^{T} -2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\right)\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)} \\
& + \left(\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\right)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\right)'\boldsymbol{\Sigma}^{-1}\left(\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\right)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\right)\Bigg\} \\
= \exp\Bigg\{ & -\frac{1}{2}\sum_{t=1}^{T} -2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\right)\boldsymbol{\beta}_1^{(r)} \\
& + \left[\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\right)\boldsymbol{\beta}_1^{(r)}\right]'\boldsymbol{\Sigma}^{-1}\left[\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\right)\boldsymbol{\beta}_1^{(r)}\right]\Bigg\} \\
= \exp\Bigg\{ & -\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\beta}_1^{(r)'}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2\left(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\right)\boldsymbol{\beta}_1^{(r)} \\
& - 2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\right)\Bigg\}\boldsymbol{\beta}_1^{(r)} \\
= \exp & \left\{-\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\beta}_1^{(r)'}\mathbf{S}_1^L(t)\boldsymbol{\beta}_1^{(r)} - 2\mathbf{m}_1^L(t)\boldsymbol{\beta}_1^{(r)}\right\}, \qquad \text{(B.100)}
\end{aligned}$$

with:

$$\mathbf{S}_1^L(t) = \langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2\left(\boldsymbol{\beta}_3^{(r)'}\otimes\boldsymbol{\beta}_2^{(r)'}\otimes\mathbf{I}_{I_1}\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\right) \qquad \text{(B.101)}$$

$$\mathbf{m}_1^L(t) = \mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\boldsymbol{\beta}_3^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\otimes\mathbf{I}_{I_1}\right). \qquad \text{(B.102)}$$

Consider the case $j = 2$. From eq. (B.92) we get:

$$\begin{aligned}
L\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_3\right) \propto \exp\Bigg\{ & -\frac{1}{2}\sum_{t=1}^{T} -2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)}\otimes\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\right)\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)} \\
& + \left(\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\right)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\right)'\boldsymbol{\Sigma}^{-1}\left(\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)}\circ\boldsymbol{\beta}_2^{(r)}\circ\boldsymbol{\beta}_3^{(r)}\right)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\right)\Bigg\} \\
= \exp\Bigg\{ & -\frac{1}{2}\sum_{t=1}^{T} -2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\circ\boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\beta}_2^{(r)} \\
& + \left[\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\beta}_2^{(r)}\right]'\boldsymbol{\Sigma}^{-1}\left[\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\beta}_2^{(r)}\right]\Bigg\} \\
= \exp\Bigg\{ & -\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\beta}_2^{(r)'}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2\left(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}_3^{(r)}\otimes\mathbf{I}_{I_2}\otimes\boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\beta}_2^{(r)}
\end{aligned}$$

$$
\begin{aligned}
&- 2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle \left(\boldsymbol{\beta}_3^{(r)} \otimes \mathbf{I}_{I_2} \otimes \boldsymbol{\beta}_1^{(r)}\right) \Bigg\}\boldsymbol{\beta}_2^{(r)} \\
&= \exp\left\{ -\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\beta}_2^{(r)\prime}\mathbf{S}_2^L(t)\boldsymbol{\beta}_2^{(r)} - 2\mathbf{m}_2^L(t)\boldsymbol{\beta}_2^{(r)} \right\}, \qquad \text{(B.103)}
\end{aligned}
$$

with:

$$\mathbf{S}_2^L(t) = \langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2 \left(\boldsymbol{\beta}_3^{(r)\prime} \otimes \mathbf{I}_{I_2} \otimes \boldsymbol{\beta}_1^{(r)\prime}\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}_3^{(r)} \otimes \mathbf{I}_{I_2} \otimes \boldsymbol{\beta}_1^{(r)}\right) \tag{B.104}$$

$$\mathbf{m}_2^L(t) = \mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle \left(\boldsymbol{\beta}_3^{(r)} \otimes \mathbf{I}_{I_2} \otimes \boldsymbol{\beta}_1^{(r)}\right). \tag{B.105}$$

Consider the case $j = 3$, by exploiting eq. (B.93) we get:

$$
\begin{aligned}
L\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_3\right) \propto \exp&\Bigg\{ -\frac{1}{2}\sum_{t=1}^{T} -2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\right)\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)} \\
&+ \left(\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\right)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\right)'\boldsymbol{\Sigma}^{-1}\left(\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\right)\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\right)\Bigg\} \\
&= \exp\Bigg\{ -\frac{1}{2}\sum_{t=1}^{T} -2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\beta}_3^{(r)} \\
&+ \left[\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\beta}_3^{(r)}\right]'\boldsymbol{\Sigma}^{-1}\left[\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\beta}_3^{(r)}\right]\Bigg\} \\
&= \exp\Bigg\{ -\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\beta}_3^{(r)\prime}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2\left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right)\boldsymbol{\beta}_3^{(r)} \\
&- 2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle\left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right)\Bigg\}\boldsymbol{\beta}_3^{(r)} \\
&= \exp\left\{ -\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\beta}_3^{(r)\prime}\mathbf{S}_3^L(t)\boldsymbol{\beta}_3^{(r)} - 2\mathbf{m}_3^L(t)\boldsymbol{\beta}_3^{(r)} \right\}, \qquad \text{(B.106)}
\end{aligned}
$$

with:

$$\mathbf{S}_3^L(t) = \langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle^2 \left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)\prime} \otimes \boldsymbol{\beta}_1^{(r)\prime}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right) \tag{B.107}$$

$$\mathbf{m}_3^L(t) = \mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\langle\boldsymbol{\beta}_4^{(r)},\mathbf{x}_t\rangle \left(\mathbf{I}_{I_3} \otimes \boldsymbol{\beta}_2^{(r)} \otimes \boldsymbol{\beta}_1^{(r)}\right). \tag{B.108}$$

Finally, in the case $j = 4$. From eq. (B.99) we get:

$$
\begin{aligned}
L\left(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_3\right) \propto \exp&\Bigg\{ -\frac{1}{2}\sum_{t=1}^{T} -2\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\right)\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)} \\
&+ \boldsymbol{\beta}_4^{(r)\prime}\mathbf{x}_t\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\right)'\boldsymbol{\Sigma}^{-1}\operatorname{vec}\left(\boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \boldsymbol{\beta}_3^{(r)}\right)\mathbf{x}_t'\boldsymbol{\beta}_4^{(r)}\Bigg\}
\end{aligned} \tag{B.109}
$$

$$= \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \beta_4^{(r)'} \mathbf{S}_4^L(t) \beta_4^{(r)} - 2\mathbf{m}_4^L(t) \beta_4^{(r)} \right\} , \tag{B.110}$$

with:

$$\mathbf{S}_4^L(t) = \mathbf{x}_t \operatorname{vec}\left(\beta_1^{(r)} \circ \beta_2^{(r)} \circ \beta_3^{(r)}\right)' \mathbf{\Sigma}^{-1} \operatorname{vec}\left(\beta_1^{(r)} \circ \beta_2^{(r)} \circ \beta_3^{(r)}\right) \mathbf{x}_t' \tag{B.111}$$

$$\mathbf{m}_4^L(t) = \mathbf{y}_t' \mathbf{\Sigma}^{-1} \operatorname{vec}\left(\beta_1^{(r)} \circ \beta_2^{(r)} \circ \beta_3^{(r)}\right) \mathbf{x}_t' . \tag{B.112}$$

It is now possible to derive the full conditional distributions for the PARAFAC marginals $\beta_1^{(r)}, \beta_2^{(r)}, \beta_3^{(r)}, \beta_4^{(r)}$, for $r = 1, \dots, R$, as shown in the following.

Full conditional distribution of $\beta_1^{(r)}$

The posterior full conditional distribution of $\beta_1^{(r)}$ is obtained by combining the prior distribution in eq. (2.29e) and the likelihood in eq. (B.100) as follows:

$$\begin{aligned}
&p(\beta_1^{(r)} | \beta_{-1}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{1,r}, \phi_r, \tau, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3, \mathbf{Y}) \propto L(\mathbf{Y} | \mathcal{B}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3) \pi(\beta_1^{(r)} | \mathbf{W}_{1,r}, \phi_r, \tau) \\
&\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \beta_1^{(r)'} \mathbf{S}_1^L(t) \beta_1^{(r)} - 2\mathbf{m}_1^L(t) \beta_1^{(r)} \right\} \cdot \exp\left\{ -\frac{1}{2} \beta_1^{(r)'} (\mathbf{W}_{1,r} \phi_r \tau)^{-1} \beta_1^{(r)} \right\} \\
&= \exp\left\{ -\frac{1}{2} \left[\sum_{t=1}^{T} \beta_1^{(r)'} \mathbf{S}_1^L(t) \beta_1^{(r)} - 2\mathbf{m}_1^L(t) \beta_1^{(r)} + \beta_1^{(r)'} (\mathbf{W}_{1,r} \phi_r \tau)^{-1} \beta_1^{(r)} \right] \right\} \\
&= \exp\left\{ -\frac{1}{2} \left[\beta_1^{(r)'} \left(\sum_{t=1}^{T} \mathbf{S}_1^L(t) + (\mathbf{W}_{1,r} \phi_r \tau)^{-1} \right) \beta_1^{(r)} - 2 \left(\sum_{t=1}^{T} \mathbf{m}_1^L(t) \right) \beta_1^{(r)} \right] \right\} \\
&= \exp\left\{ -\frac{1}{2} \left[\beta_1^{(r)'} \bar{\mathbf{\Sigma}}_{\beta_1^r}^{-1} \beta_1^{(r)} - 2 \bar{\boldsymbol{\mu}}_{\beta_1^r} \beta_1^{(r)} \right] \right\} ,
\end{aligned}$$

where:

$$\begin{aligned}
\bar{\mathbf{\Sigma}}_{\beta_1^r} &= \left[(\mathbf{W}_{1,r} \phi_r \tau)^{-1} + \sum_{t=1}^{T} \mathbf{S}_1^L(t) \right]^{-1} \\
\bar{\boldsymbol{\mu}}_{\beta_1^r} &= \bar{\mathbf{\Sigma}}_{\beta_1^r} \left[\sum_{t=1}^{T} \mathbf{m}_1^L(t) \right]' .
\end{aligned}$$

Thus the posterior full conditional distribution of $\beta_1^{(r)}$, for $r = 1, \dots, R$, is given by:

$$p(\beta_1^{(r)} | \beta_{-1}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{1,r}, \phi_r, \tau, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3, \mathbf{Y}) \sim \mathcal{N}_{I_1}(\bar{\boldsymbol{\mu}}_{\beta_1^r}, \bar{\mathbf{\Sigma}}_{\beta_1^r}) . \tag{B.113}$$

Full conditional distribution of $\beta_2^{(r)}$

The posterior full conditional distribution of $\beta_2^{(r)}$ is obtained by combining the prior distribution in eq. (2.29e) and the likelihood in eq. (B.103) as follows:

$$p(\beta_2^{(r)} | \beta_{-2}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{2,r}, \phi_r, \tau, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3, \mathbf{Y}) \propto L(\mathbf{Y} | \mathcal{B}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3) \pi(\beta_2^{(r)} | \mathbf{W}_{2,r}, \phi_r, \tau)$$

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\beta}_2^{(r)\prime}\mathbf{S}_2^L(t)\boldsymbol{\beta}_2^{(r)} - 2\mathbf{m}_2^L(t)\boldsymbol{\beta}_2^{(r)}\right\}\cdot\exp\left\{-\frac{1}{2}\boldsymbol{\beta}_2^{(r)\prime}(\mathbf{W}_{2,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_2^{(r)}\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\sum_{t=1}^{T}\boldsymbol{\beta}_2^{(r)\prime}\mathbf{S}_2^L(t)\boldsymbol{\beta}_2^{(r)} - 2\mathbf{m}_2^L(t)\boldsymbol{\beta}_2^{(r)} + \boldsymbol{\beta}_2^{(r)\prime}(\mathbf{W}_{2,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_2^{(r)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_2^{(r)\prime}\left(\sum_{t=1}^{T}\mathbf{S}_2^L(t) + (\mathbf{W}_{2,r}\phi_r\tau)^{-1}\right)\boldsymbol{\beta}_2^{(r)} - 2\left(\sum_{t=1}^{T}\mathbf{m}_2^L(t)\right)\boldsymbol{\beta}_2^{(r)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_2^{(r)\prime}\bar{\boldsymbol{\Sigma}}_{\beta_2^r}^{-1}\boldsymbol{\beta}_2^{(r)} - 2\bar{\boldsymbol{\mu}}_{\beta_2^r}\boldsymbol{\beta}_2^{(r)}\right]\right\},
\end{aligned}
$$

where:

$$
\begin{aligned}
\bar{\boldsymbol{\Sigma}}_{\beta_2^r} &= \left[(\mathbf{W}_{2,r}\phi_r\tau)^{-1} + \sum_{t=1}^{T}\mathbf{S}_2^L(t)\right]^{-1} \\
\bar{\boldsymbol{\mu}}_{\beta_2^r} &= \bar{\boldsymbol{\Sigma}}_{\beta_2^r}\left[\sum_{t=1}^{T}\mathbf{m}_2^L(t)\right]'.
\end{aligned}
$$

Thus the posterior full conditional distribution of $\boldsymbol{\beta}_2^{(r)}$, for $r = 1, \ldots, R$, is given by:

$$
p(\boldsymbol{\beta}_2^{(r)}|\boldsymbol{\beta}_{-2}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{2,r}, \phi_r, \tau, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \mathbf{Y}) \sim \mathcal{N}_{I_2}(\bar{\boldsymbol{\mu}}_{\beta_2^r}, \bar{\boldsymbol{\Sigma}}_{\beta_2^r}). \tag{B.114}
$$

Full conditional distribution of $\boldsymbol{\beta}_3^{(r)}$

The posterior full conditional distribution of $\boldsymbol{\beta}_3^{(r)}$ is obtained by combining the prior distribution in eq. (2.29e) and the likelihood in eq. (B.106) as follows:

$$
\begin{aligned}
&p(\boldsymbol{\beta}_3^{(r)}|\boldsymbol{\beta}_{-3}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{3,r}, \phi_r, \tau, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \mathbf{Y}) \propto L(\mathbf{Y}|\mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3)\pi(\boldsymbol{\beta}_3^{(r)}|\mathbf{W}_{3,r}, \phi_r, \tau) \\
&\propto \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\beta}_3^{(r)\prime}\mathbf{S}_3^L(t)\boldsymbol{\beta}_3^{(r)} - 2\mathbf{m}_3^L(t)\boldsymbol{\beta}_3^{(r)}\right\}\cdot\exp\left\{-\frac{1}{2}\boldsymbol{\beta}_3^{(r)\prime}(\mathbf{W}_{3,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_3^{(r)}\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\sum_{t=1}^{T}\boldsymbol{\beta}_3^{(r)\prime}\mathbf{S}_3^L(t)\boldsymbol{\beta}_3^{(r)} - 2\mathbf{m}_3^L(t)\boldsymbol{\beta}_3^{(r)} + \boldsymbol{\beta}_3^{(r)\prime}(\mathbf{W}_{3,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_3^{(r)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_3^{(r)\prime}\left(\sum_{t=1}^{T}\mathbf{S}_3^L(t) + (\mathbf{W}_{3,r}\phi_r\tau)^{-1}\right)\boldsymbol{\beta}_3^{(r)} - 2\left(\sum_{t=1}^{T}\mathbf{m}_3^L(t)\right)\boldsymbol{\beta}_3^{(r)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_3^{(r)\prime}\bar{\boldsymbol{\Sigma}}_{\beta_3^r}^{-1}\boldsymbol{\beta}_3^{(r)} - 2\bar{\boldsymbol{\mu}}_{\beta_3^r}\boldsymbol{\beta}_3^{(r)}\right]\right\},
\end{aligned}
$$

where:

$$
\begin{aligned}
\bar{\boldsymbol{\Sigma}}_{\beta_3^r} &= \left[(\mathbf{W}_{3,r}\phi_r\tau)^{-1} + \sum_{t=1}^{T}\mathbf{S}_3^L(t)\right]^{-1} \\
\bar{\boldsymbol{\mu}}_{\beta_3^r} &= \bar{\boldsymbol{\Sigma}}_{\beta_3^r}\left[\sum_{t=1}^{T}\mathbf{m}_3^L(t)\right]'.
\end{aligned}
$$

Thus the posterior full conditional distribution of $\boldsymbol{\beta}_3^{(r)}$, for $r = 1, \dots, R$, is given by:

$$p(\boldsymbol{\beta}_3^{(r)}|\boldsymbol{\beta}_{-3}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{3,r}, \phi_r, \tau, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \mathbf{Y}) \sim \mathcal{N}_{I_3}(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_3^r}, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_3^r}) \,. \tag{B.115}$$

Full conditional distribution of $\boldsymbol{\beta}_4^{(r)}$

The posterior full conditional distribution of $\boldsymbol{\beta}_4^{(r)}$ is obtained by combining the prior distribution in eq. (2.29e) and the likelihood in eq. (B.110) as follows:

$$\begin{aligned}
&p(\boldsymbol{\beta}_4^{(r)}|\boldsymbol{\beta}_{-4}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{4,r}, \phi_r, \tau, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \mathbf{Y}) \propto L(\mathbf{Y}|\mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3)\pi(\boldsymbol{\beta}_4^{(r)}|\mathbf{W}_{4,r}, \phi_r, \tau) \\
&\propto \exp\left\{-\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\beta}_4^{(r)\prime}\mathbf{S}_4^L(t)\boldsymbol{\beta}_4^{(r)} - 2\mathbf{m}_4^L(t)\boldsymbol{\beta}_4^{(r)}\right\} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{\beta}_4^{(r)\prime}(\mathbf{W}_{4,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_4^{(r)}\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\sum_{t=1}^{T} \boldsymbol{\beta}_4^{(r)\prime}\mathbf{S}_4^L(t)\boldsymbol{\beta}_4^{(r)} - 2\mathbf{m}_4^L(t)\boldsymbol{\beta}_4^{(r)} + \boldsymbol{\beta}_4^{(r)\prime}(\mathbf{W}_{4,r}\phi_r\tau)^{-1}\boldsymbol{\beta}_4^{(r)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_4^{(r)\prime}\left(\sum_{t=1}^{T}\mathbf{S}_4^L(t) + (\mathbf{W}_{4,r}\phi_r\tau)^{-1}\right)\boldsymbol{\beta}_4^{(r)} - 2\left(\sum_{t=1}^{T}\mathbf{m}_4^L(t)\right)\boldsymbol{\beta}_4^{(r)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_4^{(r)\prime}\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_4^r}^{-1}\boldsymbol{\beta}_4^{(r)} - 2\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_4^r}\boldsymbol{\beta}_4^{(r)}\right]\right\} ,
\end{aligned}$$

where:

$$\begin{aligned}
\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_4^r} &= \left[(\mathbf{W}_{4,r}\phi_r\tau)^{-1} + \sum_{t=1}^{T}\mathbf{S}_4^L(t)\right]^{-1} \\
\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_4^r} &= \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_4^r}\left[\sum_{t=1}^{T}\mathbf{m}_4^L(t)\right]' \,.
\end{aligned}$$

Thus the posterior full conditional distribution of $\boldsymbol{\beta}_4^{(r)}$, for $r = 1, \dots, R$, is given by:

$$p(\boldsymbol{\beta}_4^{(r)}|\boldsymbol{\beta}_{-4}^{(r)}, \mathcal{B}_{-r}, \mathbf{W}_{4,r}, \phi_r, \tau, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \mathbf{Y}) \sim \mathcal{N}_{I_1I_2I_3}(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_4^r}, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_4^r}) \,. \tag{B.116}$$

B.4.6 Full conditional distribution of $\boldsymbol{\Sigma}_1$

Given a inverse Wishart prior, the posterior full conditional distribution for $\boldsymbol{\Sigma}_1$ is conjugate. For ease of notation, define $\tilde{\mathbf{E}}_t = \mathcal{Y}_t - \mathcal{B} \times_4 \mathbf{x}_t$, $\tilde{\mathbf{E}}_{(1),t}$ the mode-1 matricization of $\tilde{\mathbf{E}}_t$ and $\mathbf{Z}_1 = \boldsymbol{\Sigma}_3^{-1} \otimes \boldsymbol{\Sigma}_2^{-1}$. By exploiting the relation between the tensor normal distribution and the multivariate normal distribution and the properties of the vectorization and trace operators, we obtain:

$$\begin{aligned}
&p(\boldsymbol{\Sigma}_1|\mathcal{B}, \mathbf{Y}, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \gamma) \propto L(\mathbf{Y}|\mathcal{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3)\pi(\boldsymbol{\Sigma}_1|\gamma) \\
&\propto |\boldsymbol{\Sigma}_1|^{-\frac{TI_2I_3}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\mathrm{vec}\left(\mathcal{Y}_t - \mathcal{B}\times_4\mathbf{x}_t\right)'\left(\boldsymbol{\Sigma}_3^{-1}\otimes\boldsymbol{\Sigma}_2^{-1}\otimes\boldsymbol{\Sigma}_1^{-1}\right)\mathrm{vec}\left(\mathcal{Y}_t - \mathcal{B}\times_4\mathbf{x}_t\right)\right\} \\
&\cdot |\boldsymbol{\Sigma}_1|^{-\frac{\nu_1+I_1+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\gamma\boldsymbol{\Psi}_1\boldsymbol{\Sigma}_1^{-1}\right)\right\}
\end{aligned}$$

$$
\begin{aligned}
&\propto |\boldsymbol{\Sigma}_1|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_1\boldsymbol{\Sigma}_1^{-1}\right)+\sum_{t=1}^{T}\operatorname{vec}\left(\tilde{\mathbf{E}}_t\right)'(\mathbf{Z}_1\otimes\boldsymbol{\Sigma}_1^{-1})\operatorname{vec}\left(\tilde{\mathbf{E}}_t\right)\right]\right\}\\
&\propto |\boldsymbol{\Sigma}_1|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_1\boldsymbol{\Sigma}_1^{-1}\right)+\sum_{t=1}^{T}\operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)'(\mathbf{Z}_1\otimes\boldsymbol{\Sigma}_1^{-1})\operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)\right]\right\}\\
&\propto |\boldsymbol{\Sigma}_1|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_1\boldsymbol{\Sigma}_1^{-1}\right)+\sum_{t=1}^{T}\operatorname{tr}\left(\operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)'\operatorname{vec}\left(\boldsymbol{\Sigma}_1^{-1}\tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_1\right)\right)\right]\right\}\\
&\propto |\boldsymbol{\Sigma}_1|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_1\boldsymbol{\Sigma}_1^{-1}\right)+\sum_{t=1}^{T}\operatorname{tr}\left(\tilde{\mathbf{E}}'_{(1),t}\boldsymbol{\Sigma}_1^{-1}\tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_1\right)\right]\right\}\\
&\propto |\boldsymbol{\Sigma}_1|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_1\boldsymbol{\Sigma}_1^{-1}\right)+\sum_{t=1}^{T}\operatorname{tr}\left(\tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_1\tilde{\mathbf{E}}'_{(1),t}\boldsymbol{\Sigma}_1^{-1}\right)\right]\right\}\,. \qquad \text{(B.117)}
\end{aligned}
$$

For ease of notation, define $S_1 = \sum_{t=1}^{T} \tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_1\tilde{\mathbf{E}}'_{(1),t}$. Then:

$$
\begin{aligned}
p(\boldsymbol{\Sigma}_1|\mathcal{B},\mathbf{Y},\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_3) &\propto |\boldsymbol{\Sigma}_1|^{-\frac{\nu_1+I_1+TI_2I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_1\boldsymbol{\Sigma}_1^{-1}\right)+\operatorname{tr}\left(S_1\boldsymbol{\Sigma}_1^{-1}\right)\right]\right\}\\
&\propto |\boldsymbol{\Sigma}_1|^{-\frac{(\nu_1+TI_2I_3)+I_1+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(\gamma\boldsymbol{\Psi}_1+S_1\right)\boldsymbol{\Sigma}_1^{-1}\right)\right\}\,, \qquad \text{(B.118)}
\end{aligned}
$$

Therefore, the posterior full conditional distribution of $\boldsymbol{\Sigma}_1$ is given by:

$$
p(\boldsymbol{\Sigma}_1|\mathcal{B},\mathbf{Y},\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_3,\gamma) \sim \mathcal{IW}_{I_1}\left(\nu_1+TI_2I_3,\gamma\boldsymbol{\Psi}_1+S_1\right)\,. \qquad \text{(B.119)}
$$

B.4.7 Full conditional distribution of $\boldsymbol{\Sigma}_2$

Given a inverse Wishart prior, the posterior full conditional distribution for $\boldsymbol{\Sigma}_2$ is conjugate. For ease of notation, define $\tilde{\mathbf{E}}_t = \mathcal{Y}_t - \mathcal{B}\times_4\mathbf{x}_t$ and $\tilde{\mathbf{E}}_{(2),t}$ the mode-2 matricization of $\tilde{\mathbf{E}}_t$. By exploiting the relation between the tensor normal distribution and the matrix normal distribution and the properties of the Kronecker product and of the vectorization and trace operators we obtain:

$$
\begin{aligned}
&p(\boldsymbol{\Sigma}_2|\mathcal{B},\mathbf{Y},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_3,\gamma) \propto L(\mathbf{Y}|\mathcal{B},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_3)\pi(\boldsymbol{\Sigma}_2|\gamma)\\
&\propto |\boldsymbol{\Sigma}_2|^{-\frac{TI_1I_3}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}(\mathcal{Y}_t-\mathcal{B}\times_4\mathbf{x}_t)\times^{1\ldots3}(\boldsymbol{\Sigma}_1^{-1}\circ\boldsymbol{\Sigma}_2^{-1}\circ\boldsymbol{\Sigma}_3^{-1})\times^{1\ldots3}(\mathcal{Y}_t-\mathcal{B}\times_4\mathbf{x}_t)\right\}\\
&\cdot|\boldsymbol{\Sigma}_2|^{-\frac{\nu_2+I_2+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}_2\boldsymbol{\Sigma}_2^{-1}\right)\right\}\\
&\propto |\boldsymbol{\Sigma}_2|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_2\boldsymbol{\Sigma}_2^{-1}\right)+\sum_{t=1}^{T}\tilde{\mathbf{E}}_t\times^{1\ldots3}(\boldsymbol{\Sigma}_1^{-1}\circ\boldsymbol{\Sigma}_2^{-1}\circ\boldsymbol{\Sigma}_3^{-1})\times^{1\ldots3}\tilde{\mathbf{E}}_t\right]\right\}\\
&\propto |\boldsymbol{\Sigma}_2|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_2\boldsymbol{\Sigma}_2^{-1}\right)+\sum_{t=1}^{T}\operatorname{tr}\left(\tilde{\mathbf{E}}'_{(2),t}(\boldsymbol{\Sigma}_3^{-1}\otimes\boldsymbol{\Sigma}_1^{-1}\otimes\boldsymbol{\Sigma}_2^{-1})\tilde{\mathbf{E}}_{(2),t}\right)\right]\right\}\\
&\propto |\boldsymbol{\Sigma}_2|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\boldsymbol{\Psi}_2\boldsymbol{\Sigma}_2^{-1}\right)+\sum_{t=1}^{T}\operatorname{tr}\left((\boldsymbol{\Sigma}_3^{-1}\otimes\boldsymbol{\Sigma}_1^{-1})\tilde{\mathbf{E}}'_{(2),t}\boldsymbol{\Sigma}_2^{-1}\tilde{\mathbf{E}}_{(2),t}\right)\right]\right\}
\end{aligned}
$$

$$\propto |\mathbf{\Sigma}_2|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_2\mathbf{\Sigma}_2^{-1}\right) + \operatorname{tr}\left(\sum_{t=1}^{T}\tilde{\mathbf{E}}_{(2),t}(\mathbf{\Sigma}_3^{-1}\otimes\mathbf{\Sigma}_1^{-1})\tilde{\mathbf{E}}'_{(2),t}\mathbf{\Sigma}_2^{-1}\right)\right]\right\}$$
$$\propto |\mathbf{\Sigma}_2|^{-\frac{\nu_2+I_2+TI_1I_3+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\gamma\mathbf{\Psi}_2\mathbf{\Sigma}_2^{-1} + S_2\mathbf{\Sigma}_2^{-1}\right)\right\},$$

where for ease of notation we defined $S_2 = \sum_{t=1}^{T}\tilde{\mathbf{E}}_{(2),t}(\mathbf{\Sigma}_3^{-1}\otimes\mathbf{\Sigma}_1^{-1})\tilde{\mathbf{E}}'_{(2),t}$. Therefore, the posterior full conditional distribution of $\mathbf{\Sigma}_2$ is given by:

$$p(\mathbf{\Sigma}_2|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_1,\mathbf{\Sigma}_3) \sim \mathcal{IW}_{I_2}\left(\nu_2 + TI_1I_3, \gamma\mathbf{\Psi}_2 + S_2\right). \tag{B.120}$$

B.4.8 Full conditional distribution of $\mathbf{\Sigma}_3$

Given a inverse Wishart prior, the posterior full conditional distribution for $\mathbf{\Sigma}_3$ is conjugate. For ease of notation, define $\tilde{\mathbf{E}}_t = \mathcal{Y}_t - \mathcal{B}\times_4\mathbf{x}_t$, $\tilde{\mathbf{E}}_{(1),t}$ the mode-1 matricization of $\tilde{\mathbf{E}}_t$ and $\mathbf{Z}_3 = \mathbf{\Sigma}_2^{-1}\otimes\mathbf{\Sigma}_1^{-1}$. By exploiting the relation between the tensor normal distribution and the multivariate normal distribution and the properties of the vectorization and trace operators, we obtain:

$$p(\mathbf{\Sigma}_3|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_1,\mathbf{\Sigma}_2,\gamma) \propto L(\mathbf{Y}|\mathcal{B},\mathbf{\Sigma}_1,\mathbf{\Sigma}_2,\mathbf{\Sigma}_3)\pi(\mathbf{\Sigma}_3|\gamma)$$
$$\propto |\mathbf{\Sigma}_3|^{-\frac{TI_1I_2}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\operatorname{vec}\left(\mathcal{Y}_t - \mathcal{B}\times_4\mathbf{x}_t\right)'(\mathbf{\Sigma}_3^{-1}\otimes\mathbf{\Sigma}_2^{-1}\otimes\mathbf{\Sigma}_1^{-1})\operatorname{vec}\left(\mathcal{Y}_t - \mathcal{B}\times_4\mathbf{x}_t\right)\right\}$$
$$\cdot |\mathbf{\Sigma}_3|^{-\frac{\nu_3+I_3+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\gamma\mathbf{\Psi}_3\mathbf{\Sigma}_3^{-1}\right)\right\}$$
$$\propto |\mathbf{\Sigma}_3|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_3\mathbf{\Sigma}_3^{-1}\right) + \sum_{t=1}^{T}\operatorname{vec}\left(\tilde{\mathbf{E}}_t\right)'(\mathbf{\Sigma}_3^{-1}\otimes\mathbf{Z}_3)\operatorname{vec}\left(\tilde{\mathbf{E}}_t\right)\right]\right\}$$
$$\propto |\mathbf{\Sigma}_3|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_3\mathbf{\Sigma}_3^{-1}\right) + \sum_{t=1}^{T}\operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)'(\mathbf{\Sigma}_3^{-1}\otimes\mathbf{Z}_3)\operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)\right]\right\}$$
$$\propto |\mathbf{\Sigma}_3|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_3\mathbf{\Sigma}_3^{-1}\right) + \sum_{t=1}^{T}\operatorname{tr}\left(\operatorname{vec}\left(\tilde{\mathbf{E}}_{(1),t}\right)'\operatorname{vec}\left(\mathbf{Z}_3\tilde{\mathbf{E}}_{(1),t}\mathbf{\Sigma}_3^{-1}\right)\right)\right]\right\}$$
$$\propto |\mathbf{\Sigma}_3|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_3\mathbf{\Sigma}_3^{-1}\right) + \sum_{t=1}^{T}\operatorname{tr}\left(\tilde{\mathbf{E}}'_{(1),t}\mathbf{Z}_3\tilde{\mathbf{E}}_{(1),t}\mathbf{\Sigma}_3^{-1}\right)\right]\right\}. \tag{B.121}$$

For ease of notation, define $S_3 = \sum_{t=1}^{T}\tilde{\mathbf{E}}_{(1),t}\mathbf{Z}_3\tilde{\mathbf{E}}'_{(1),t}$. Then:

$$p(\mathbf{\Sigma}_3|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_1,\mathbf{\Sigma}_2) \propto |\mathbf{\Sigma}_3|^{-\frac{\nu_3+I_3+TI_1I_2+1}{2}} \exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\gamma\mathbf{\Psi}_3\mathbf{\Sigma}_3^{-1}\right) + \operatorname{tr}\left(S_3\mathbf{\Sigma}_3^{-1}\right)\right]\right\}$$
$$\propto |\mathbf{\Sigma}_3|^{-\frac{(\nu_3+TI_1I_2)+I_3+1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(\gamma\mathbf{\Psi}_3 + S_3\right)\mathbf{\Sigma}_3^{-1}\right)\right\}, \tag{B.122}$$

Therefore, the posterior full conditional distribution of $\mathbf{\Sigma}_3$ is given by:

$$p(\mathbf{\Sigma}_3|\mathcal{B},\mathbf{Y},\mathbf{\Sigma}_1,\mathbf{\Sigma}_2) \sim \mathcal{IW}_{I_3}\left(\nu_3 + TI_1I_2, \gamma\mathbf{\Psi}_3 + S_3\right). \tag{B.123}$$

B.4.9 Full conditional distribution of γ

Using a gamma prior distribution we have:

$$\begin{aligned} p(\gamma|\mathbf{\Sigma}_1,\mathbf{\Sigma}_2,\mathbf{\Sigma}_3) &\propto p(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2,\mathbf{\Sigma}_3|\gamma)\pi(\gamma) \\ &\propto \prod_{i=1}^{3} |\gamma\mathbf{\Psi}_i|^{-\frac{\nu_i}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\gamma\mathbf{\Psi}_i\mathbf{\Sigma}_i^{-1}\right)\right\}\gamma^{a_\gamma-1}\exp\{-b_\gamma\gamma\} \\ &\propto \gamma^{a_\gamma-\frac{\sum_{i=1}^{3}\nu_i I_i}{2}-1}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\sum_{i=1}^{3}\mathbf{\Psi}_i\mathbf{\Sigma}_i^{-1}\right)-b_\gamma\gamma\right\}, \end{aligned} \tag{B.124}$$

thus:

$$p(\gamma|\mathbf{\Sigma}_1,\mathbf{\Sigma}_2,\mathbf{\Sigma}_3) \sim \mathcal{G}a\left(a_\gamma+\frac{1}{2}\sum_{i=1}^{3}\nu_i I_i, b_\gamma+\frac{1}{2}\operatorname{tr}\left(\sum_{i=1}^{3}\mathbf{\Psi}_i\mathbf{\Sigma}_i^{-1}\right)\right). \tag{B.125}$$

B.5 Additional simulations' output

B.5.1 Simulation 10x10

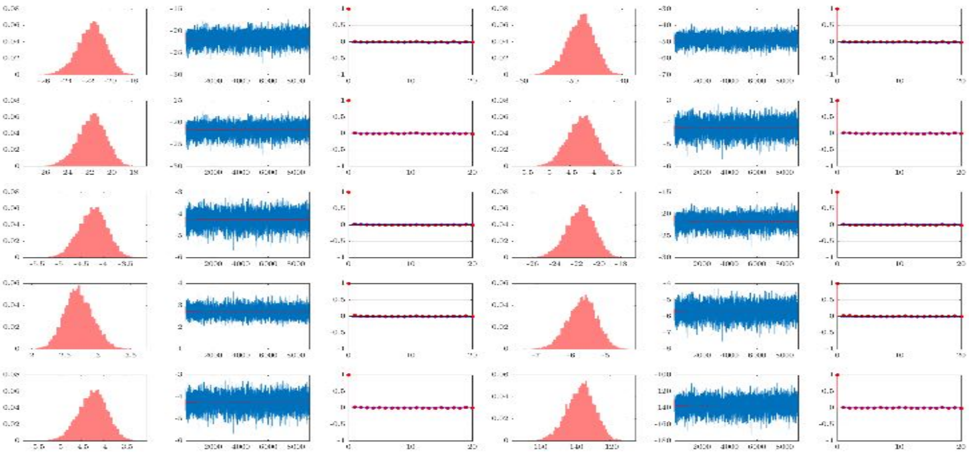


FIGURE B.1: Posterior distribution (*first, fourth* columns), MCMC output (*second, fifth* columns) and autocorrelation function (*third, sixth* columns) of some entries of the estimated covariance matrix $\mathbf{\Sigma}_1$.

B.5.2 Simulation 20x20

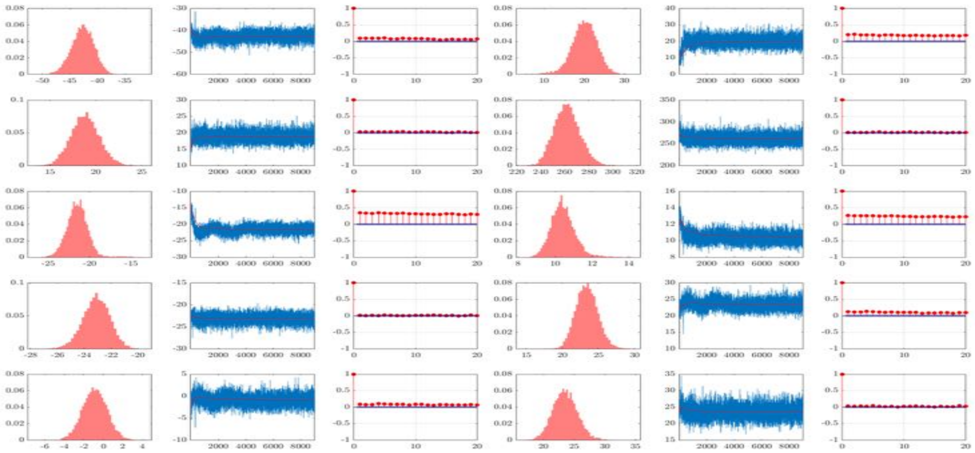


FIGURE B.2: Posterior distribution (*first*, *fourth* columns), MCMC output (*second*, *fifth* columns) and autocorrelation function (*third*, *sixth* columns) of some entries of the estimated covariance matrix $\mathbf{\Sigma}_1$.

B.6 Additional application's output

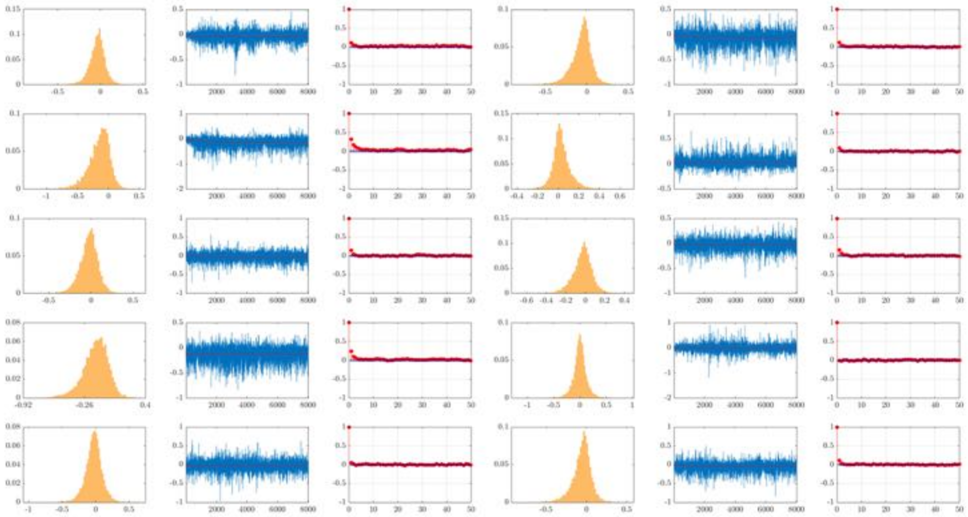


FIGURE B.3: Posterior distribution (*first*, *fourth* columns), MCMC output (*second*, *fifth* columns) and autocorrelation function (*third*, *sixth* columns) of some entries of the estimated coefficient tensor.

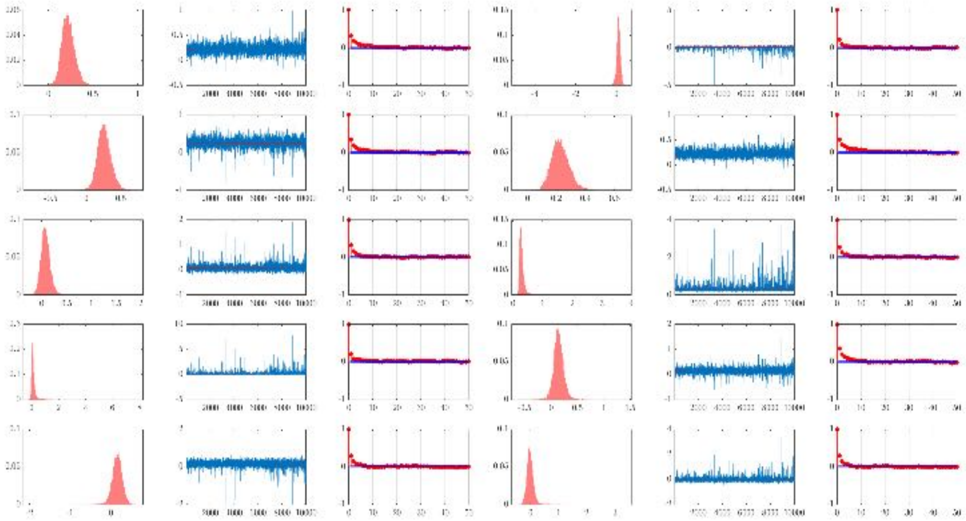


FIGURE B.4: Posterior distribution (*first, fourth* columns), MCMC output (*second, fifth* columns) and autocorrelation function (*third, sixth* columns) of some entries of the estimated error covariance matrix $\mathbf{\Sigma}_1$.

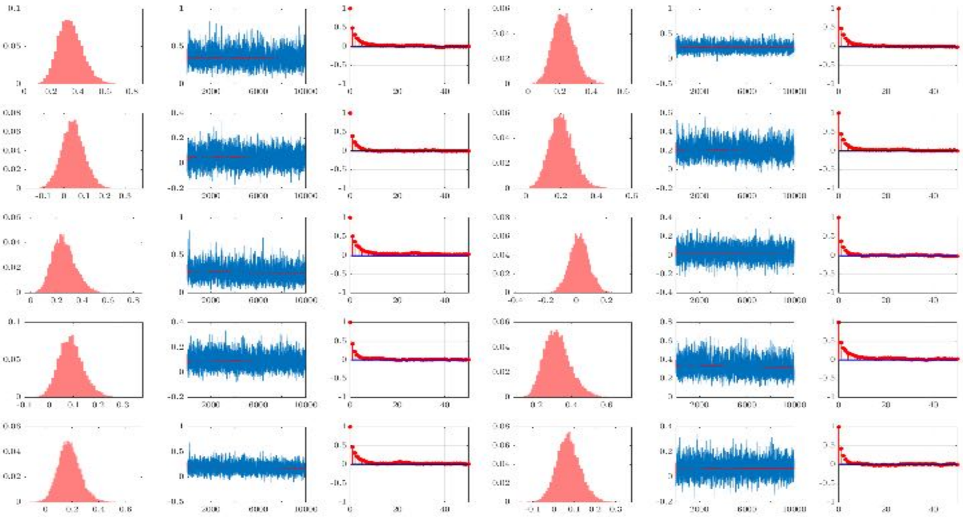


FIGURE B.5: Posterior distribution (*first, fourth* columns), MCMC output (*second, fifth* columns) and autocorrelation function (*third, sixth* columns) of some entries of the estimated error covariance matrix $\mathbf{\Sigma}_2$.

B.7 Convergence diagnostics

The inefficiency factor (INEF) is given by the ratio of the variance of the mean of MCMC draws over the variance of the mean of sample of the same size, under the assumption of independent sampling. Since MCMC draws are dependent by construction, we have that $INF \geq 1$, with values close to 1 meaning lower dependence between the elements of the chain (as measured by the autocorrelation). Let ρ_j be the autocorrelation at lag $j \geq 1$ of the MCMC draws of a given parameter θ and let N be the number of MCMC iterations. Then the INEF is given by:

$$INEF = 1 + 2\sum_{j=1}^{\infty} \rho_j \,. \tag{B.126}$$

The effective sample size (ESS) applies a correction to the number of MCMC draws by accounting for the autocorrelation between them. It provides an estimates of the number of posterior draws which can be considered independent. Clearly, $ESS \leq N$, with higher values implying better mixing of the chain. It is strictly related to the INEF and is defined as:

$$ESS = \frac{N}{1 + 2\sum_{j=1}^{\infty} \rho_j} = \frac{N}{INEF} \,. \tag{B.127}$$

The Geweke test statistic (Geweke (1991)) compares the location of the sampled parameter on two different intervals of the chain. If the mean values of the parameter in the two intervals are statistically equal, then one concludes that the two samples come from the same distribution. Let $A = \{n : 1 \leq n \leq N_A\}$ and $B = \{n : N - N_B \leq n \leq N\}$ be the initial and terminal intervals of a chain of length N, with $N_A < N$, $N_B < N$ and $(N_A + N_B)/N < 1$. Let $\overline{\boldsymbol{\theta}}_A, \overline{\boldsymbol{\theta}}_B$ be the mean values of the parameter $\boldsymbol{\theta}$ computed in the intervals A and B, respectively. Let $S_A(0), S_B(0)$ be the associated standard deviations estimated via the spectral densities at frequency zero. It is defined by:

$$Z_N = \frac{\overline{\boldsymbol{\theta}}_A - \overline{\boldsymbol{\theta}}_B}{\sqrt{N_A^{-1} S_A(0) + N_B^{-1} S_B(0)}} \xrightarrow[N\to\infty]{\mathscr{L}} \mathcal{N}(0,1) \,. \tag{B.128}$$

The Gelman and Rubin test statistic (Gelman and Rubin (1992)) is based on multiple chains run in parallel, with different starting values (possibly over-dispersed relative to the posterior distribution). The idea to test whether the chains have forgotten their initial values such that the output from all chains is indistinguishable. The test statistic is based a comparison of within-chain and between-chain variances, as follows:

$$G = \sqrt{\frac{(d+3)V}{(d+1)W}} \,, \tag{B.129}$$

where W is the mean of the empirical variance within each chain, V is the sample mean of all chains combined and $d = 2V^2/\mathbb{V}ar(V)$ are the degrees of freedom. Values substantially above 1 indicate lack of convergence.

B.7.1 Simulation $I_1 = I_2 = 10$

In the following we report the convergence diagnostic criteria previously described, for the model with $I_1 = I_2 = 10$. We run 30 parallel chains, using the same dataset and hyper-parameter set-up, with different starting points for the Gibbs sampler. The analysis has been performed using `R`[3] and the `CODA package`[4]. For the statistics based on a single chain, we report the mean and standard deviation over all the 30 parallel chains.

We assess the convergence of the chain by testing four different statistics:

- *Vnorm*, is the sum of quadratic distances (posterior mean minus true value) of the variance hyper-parameters for all the PARAFAC marginals:

$$Lnorm = (R(I + J + IJ) + R + 1)^{-1} \left[\sum_{r=1}^{R} \sum_{j=1}^{3} \left\| w^*_{j,r} \phi^*_r \tau - \widehat{w}_{j,r} \widehat{\phi}_r \widehat{\tau} \right\|_2 \right] . \tag{B.130}$$

[3] https://cran.r-project.org

[4] https://cran.r-project.org/web/packages/coda/index.html

- *Lnorm*, is the sum of quadratic distances (posterior mean minus true value) of each entry of the hyper-parameter $\boldsymbol{\lambda}$:

$$Lnorm = (3R)^{-1} \sum_{r=1}^{R} \sum_{j=1}^{3} \left\| \lambda_{j,r}^{*} - \widehat{\lambda}_{j,r} \right\|_2 . \tag{B.131}$$

- *Snorm*, is the sum of the quadratic distances (posterior mean minus true value) of each noise covariance matrix:

$$Snorm = (IJ)^{-2} \left[\left\| \Sigma_1^{*} - \widehat{\Sigma}_1 \right\|_2 + \left\| \Sigma_2^{*} - \widehat{\Sigma}_2 \right\|_2 \right] . \tag{B.132}$$

- *distT*, is the quadratic distance (posterior mean minus true value) of the coefficient tensor:

$$distT = (IJ)^{-2} \left\| \mathcal{B}^{*} - \widehat{\mathcal{B}} \right\|_2 . \tag{B.133}$$

Table B.1 reports the mean and standard deviations (over 30 chains) of the ESS, INEF and Geweke tests as well as the Gelman-Rubin test (based on 30 chains) for each of the four statistics.

The first row of Figs. B.6-B.9 shows the trace plot (together with 90% credible intervals) and autocorrelation function of the mean (over 30 chains) of each statistic, using all MCMC iterations. By contrast, the second row of Figs. B.6-B.9 reports the same plots, after having removed the burn-in iterations and having performed thinning.

Statistic	**ESS**		**INEF**		**Gwk**		**Gel-Rub**
	mean	std	mean	std	mean	std	
Vnorm	242.41	65.97	111.95	34.29	-0.61	1.51	1.02
Lnorm	224.01	110.04	334.06	483.34	-0.48	1.11	1.03
Snorm	920.60	98.27	27.46	2.99	-0.53	1.31	1.00
distT	177.54	97.58	3517.60	6779.69	0.92	4.46	1.00

TABLE B.1: Convergence diagnostics, for the case $I_1 = I_2 = 10$. For *ESS* (effective sample size), *INEF* (inefficiency factor) and *Gwk* (Geweke statistic) we report means and standard deviations over 30 chains run in parallel, with same setting and different starting points, whereas *Gel-Rub* is the Gelman-Rubin statistic considering all chains.

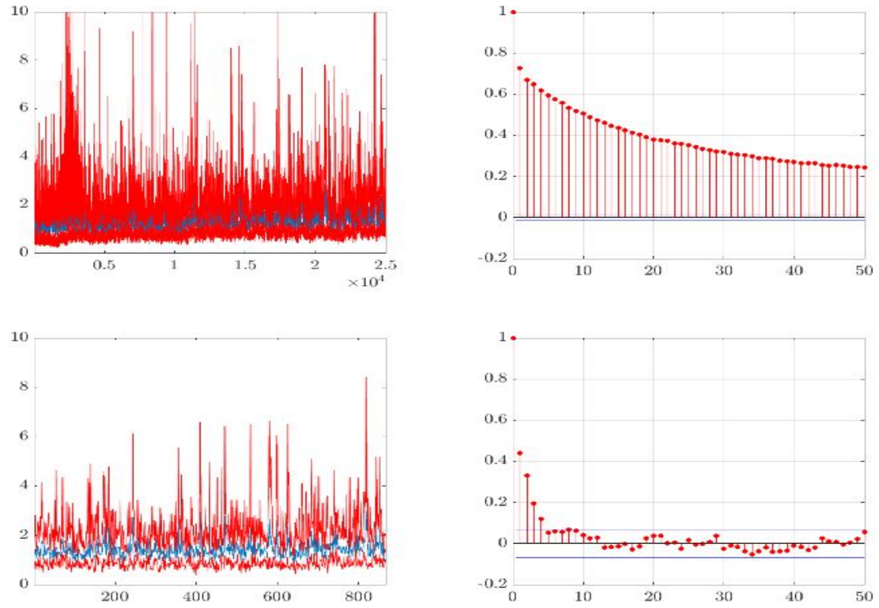


FIGURE B.6: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the sum of quadratic norms of the differences $\mathbf{w}^*_{i,r}\phi^*_r\tau^* - \widehat{\mathbf{w}}_{i,r}\widehat{\phi}_r\widehat{\tau}$, for $i = 1, 2, 3$ and $r = 1, \ldots, R$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (12000) and thinning by 15.

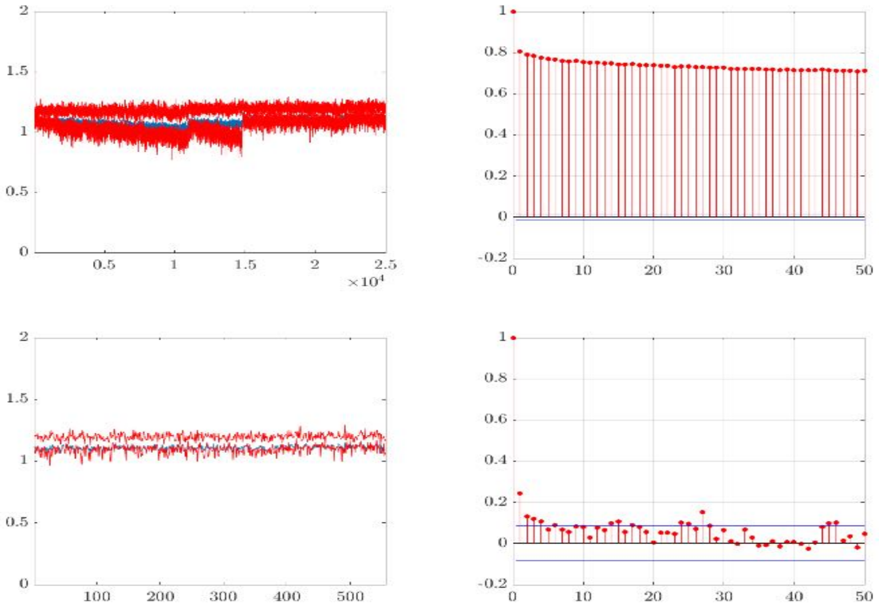


FIGURE B.7: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the quadratic norm of the difference $\boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (15000) and thinning by 18.

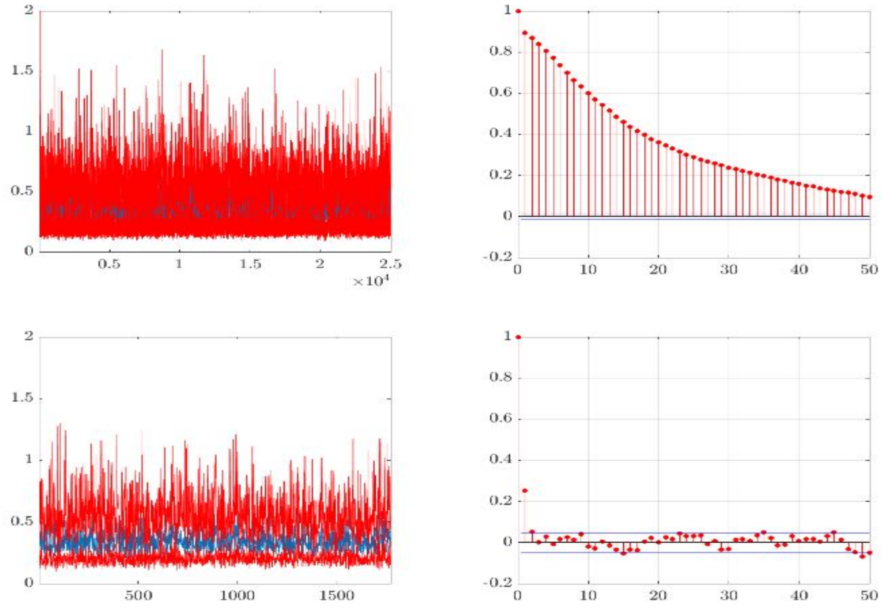


FIGURE B.8: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the sum of quadratic norms of the differences $\Sigma_1^* - \widehat{\Sigma}_1$ and $\Sigma_2^* - \widehat{\Sigma}_2$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (2000) and thinning by 13.

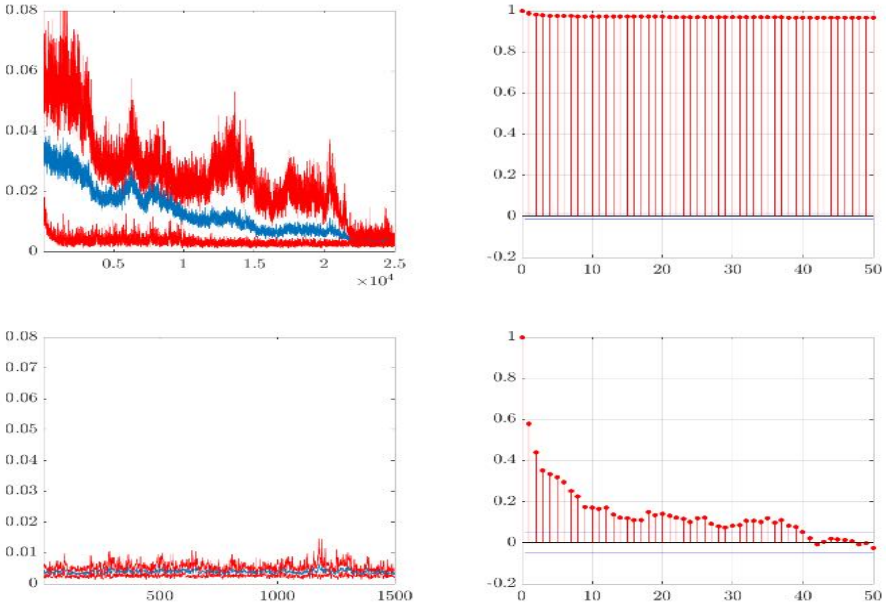


FIGURE B.9: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of quadratic norm of the difference $\mathcal{B}^* - \widehat{\mathcal{B}}$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (20000) and thinning by 2.

B.7.2 Simulation $I_1 = I_2 = 20$

In the following we report the convergence diagnostic criteria previously described, for the model with $I_1 = I_2 = 20$. We run 30 parallel chains, using the same dataset and hyper-parameter set-up, with different starting points for the Gibbs sampler. For the statistics based on a single chain, we report the mean and standard deviation over all the 30 parallel chains.

Table B.2 reports the mean and standard deviations (over 30 chains) of the ESS, INEF and Geweke tests as well as the Gelman-Rubin test (based on 30 chains) for each of the four statistics.

The first row of Figs. B.10-B.13 shows the trace plot (together with 90% credible intervals) and autocorrelation function of the mean (over 30 chains) of each statistic, using all MCMC iterations. By contrast, the second row of Figs. B.10-B.13 reports the same plots, after having removed the burn-in iterations and having performed thinning.

Statistic	**ESS**		**INEF**		**Gwk**		**Gel-Rub**
	mean	std	mean	std	mean	std	
Vnorm	195.79	128.48	827.98	1894.26	1.58	2.63	2.45
Lnorm	728.95	486.87	88.35	111.57	3.48	4.48	3.52
Snorm	403.29	152.58	204.25	482.37	0.46	4.66	1.17
distT	166.03	184.72	505.61	678.13	0.57	2.57	6.05

TABLE B.2: Convergence diagnostics, for the case $I_1 = I_2 = 20$. For *ESS* (effective sample size), *INEF* (inefficiency factor) and *Gwk* (Geweke statistic) we report means and standard deviations over 30 chains run in parallel, with same setting and different starting points, whereas *Gel-Rub* is the Gelman-Rubin statistic considering all chains.

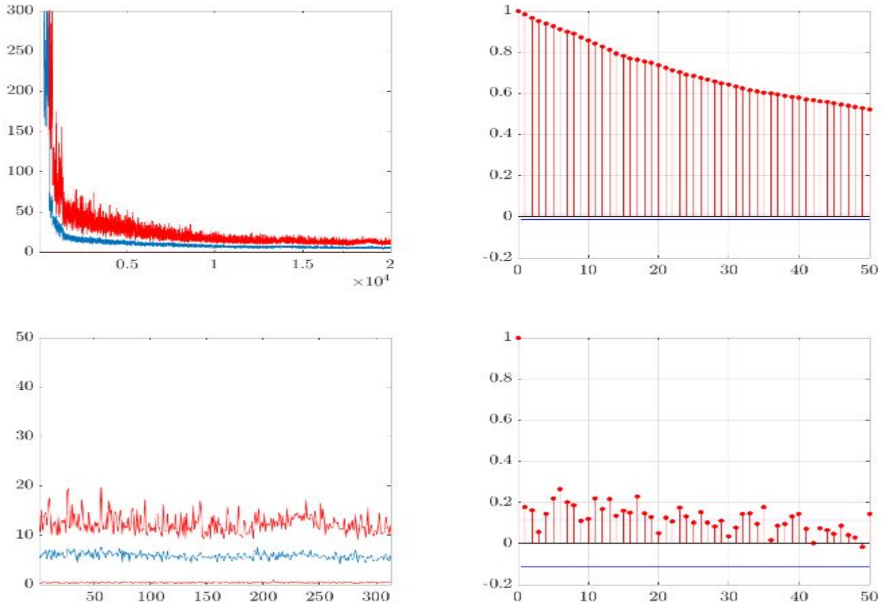


FIGURE B.10: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the sum of quadratic norms of the differences $\mathbf{w}^*_{i,r}\phi^*_r\tau^* - \widehat{\mathbf{w}}_{i,r}\widehat{\phi}_r\widehat{\tau}$, for $i = 1, 2, 3$ and $r = 1, \ldots, R$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (15000) and thinning by 16.

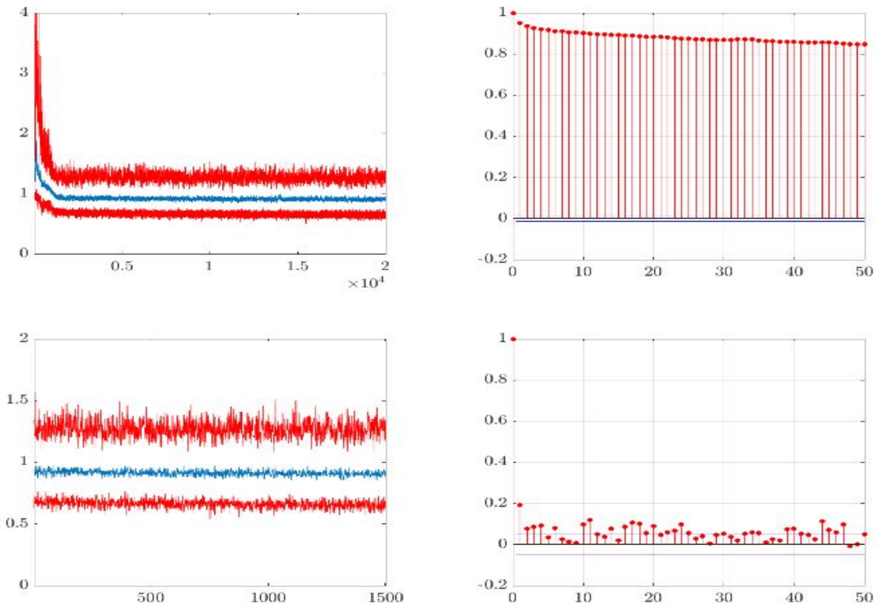


FIGURE B.11: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the quadratic norm of the difference $\boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (2000) and thinning by 12.

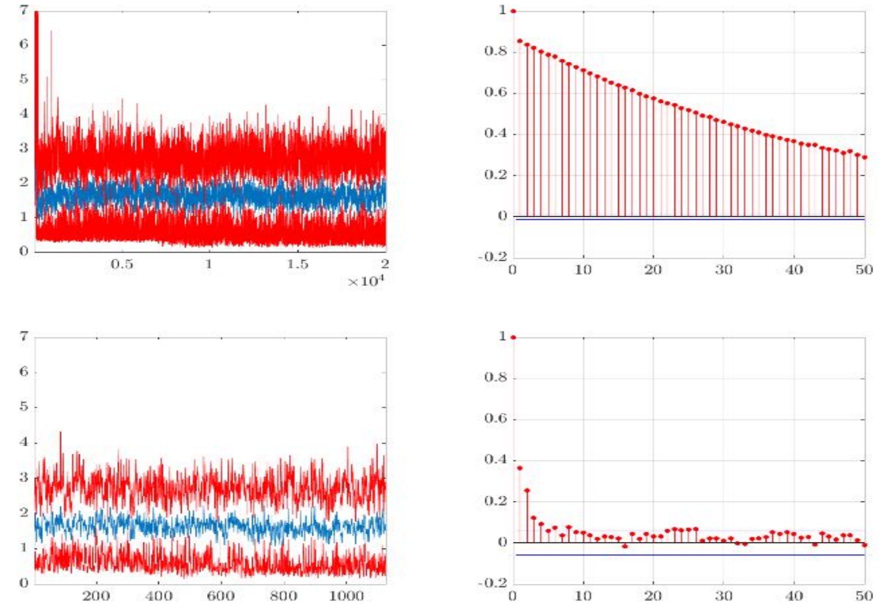


FIGURE B.12: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the sum of quadratic norms of the differences $\Sigma_1^* - \widehat{\Sigma}_1$ and $\Sigma_2^* - \widehat{\Sigma}_2$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (2000) and thinning by 16.

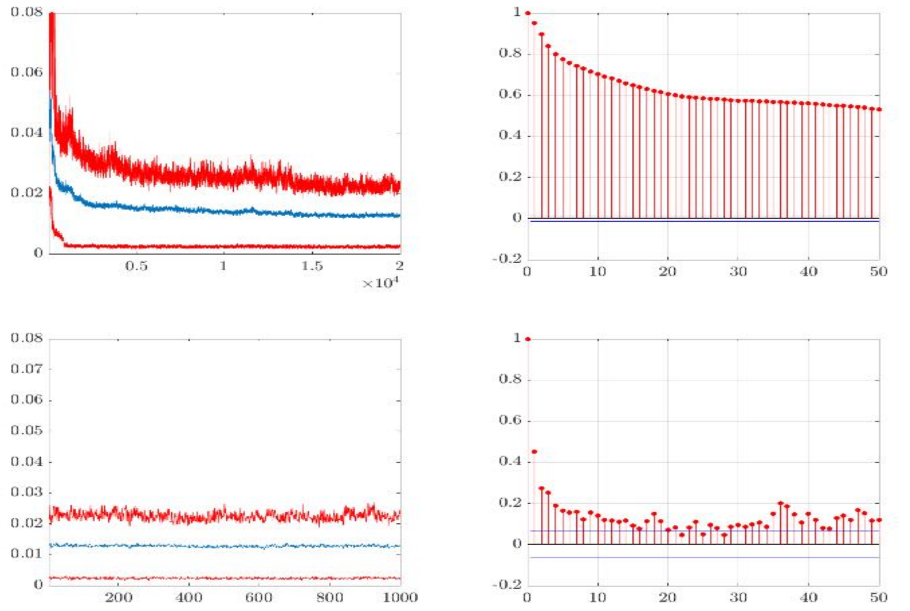


FIGURE B.13: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of quadratic norm of the difference $\mathcal{B}^* - \widehat{\mathcal{B}}$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (14000) and thinning by 6.

B.7.3 Simulation $I_1 = I_2 = 30$

In the following we report the convergence diagnostic criteria previously described, for the model with $I_1 = I_2 = 30$. We run 30 parallel chains, using the same dataset and hyper-parameter set-up, with different starting points for the Gibbs sampler. For the statistics based on a single chain, we report the mean and standard deviation over all the 30 parallel chains.

Table B.3 reports the mean and standard deviations (over 30 chains) of the ESS, INEF and Geweke tests as well as the Gelman-Rubin test (based on 30 chains) for each of the four statistics.

The first row of Figs. B.14-B.17 shows the trace plot (together with 90% credible intervals) and autocorrelation function of the mean (over 30 chains) of each statistic, using all MCMC iterations. By contrast, the second row of Figs. B.14-B.17 reports the same plots, after having removed the burn-in iterations and having performed thinning.

Statistic	**ESS**		**INEF**		**Gwk**		**Gel-Rub**
	mean	std	mean	std	mean	std	
Vnorm	20.40	27.80	4856.39	3280.00	8.18	5.38	9.03
Lnorm	297.99	283.79	279.79	223.58	1.17	4.24	1.35
Snorm	2074.57	2054.96	23.36	20.46	0.05	2.92	1.29
distT	81.62	90.23	2234.62	3627.81	2.55	5.84	6.57

TABLE B.3: Convergence diagnostics, for the case $I_1 = I_2 = 30$. For *ESS* (effective sample size), *INEF* (inefficiency factor) and *Gwk* (Geweke statistic) we report means and standard deviations over 30 chains run in parallel, with same setting and different starting points, whereas *Gel-Rub* is the Gelman-Rubin statistic considering all chains.

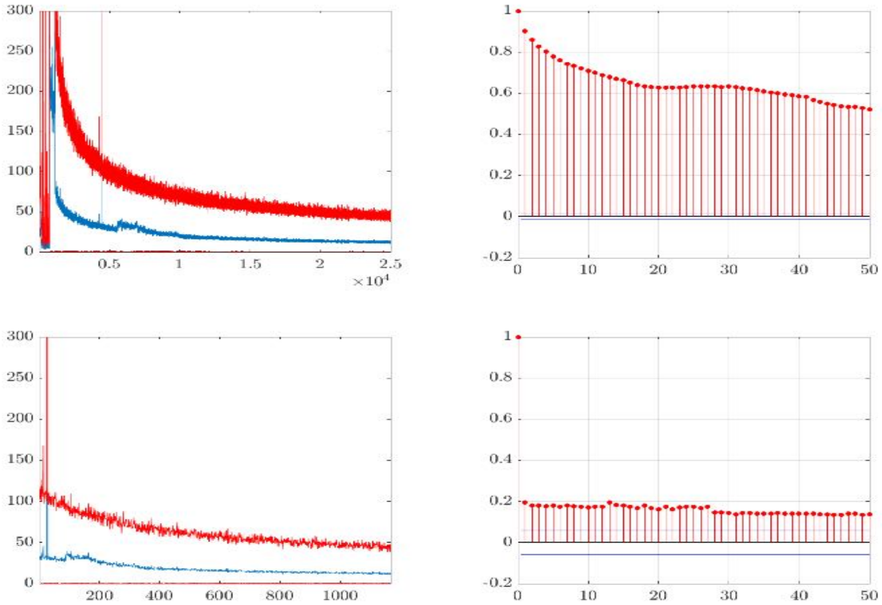


FIGURE B.14: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the sum of quadratic norms of the differences $\mathbf{w}^*_{i,r}\phi^*_r\tau^* - \widehat{\mathbf{w}}_{i,r}\widehat{\phi}_r\widehat{\tau}$, for $i = 1, 2, 3$ and $r = 1, \ldots, R$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (4000) and thinning by 18.

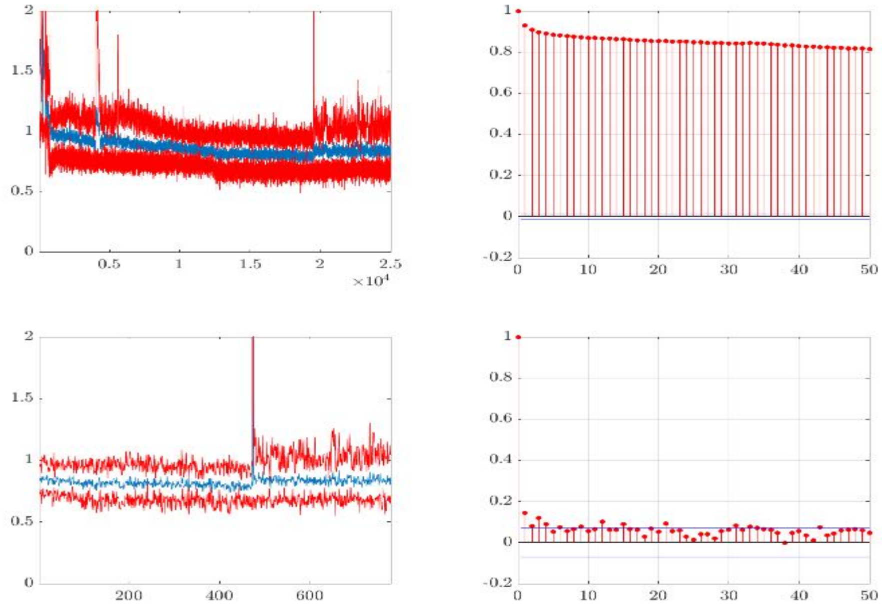


FIGURE B.15: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the quadratic norm of the difference $\boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (11000) and thinning by 18.

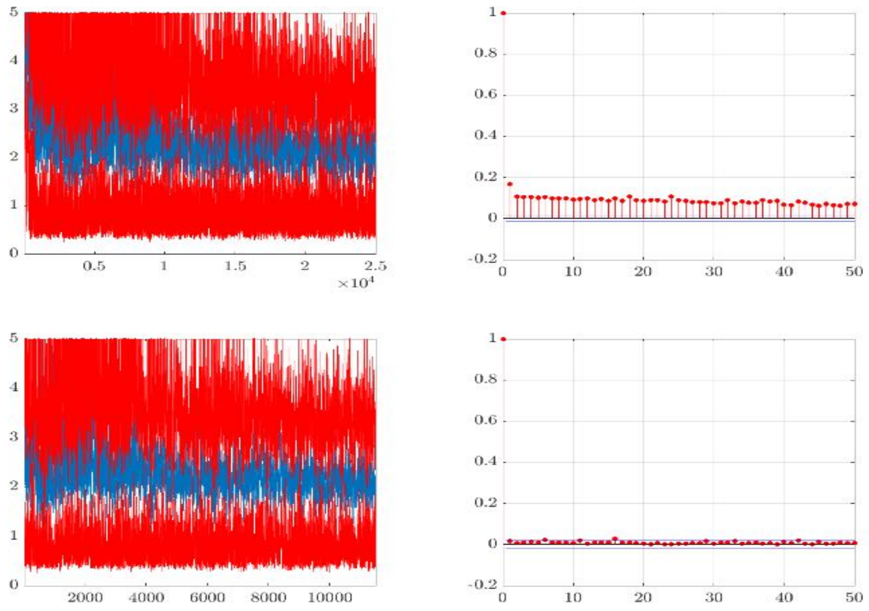


FIGURE B.16: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of the sum of quadratic norms of the differences $\Sigma_1^* - \widehat{\Sigma}_1$ and $\Sigma_2^* - \widehat{\Sigma}_2$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (2000) and thinning by 2.

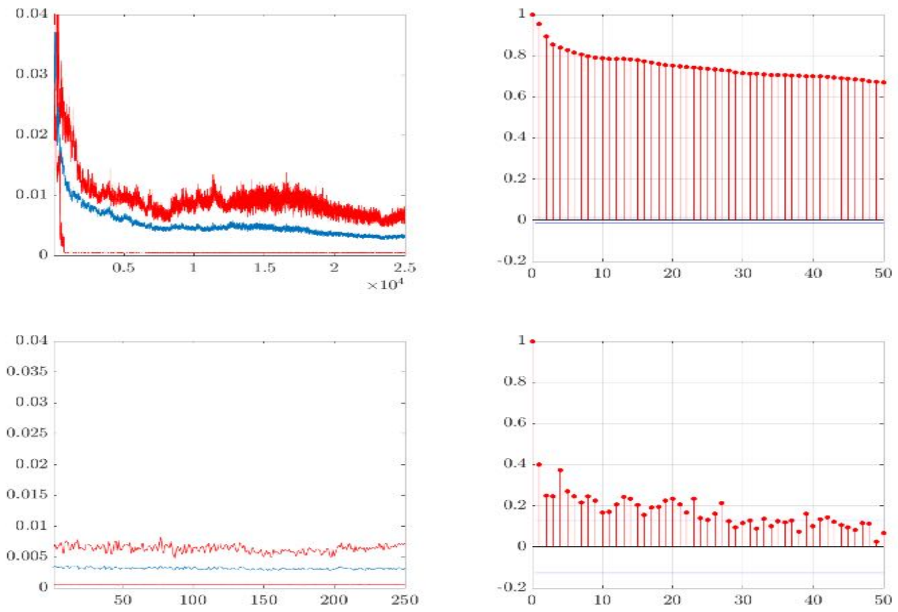


FIGURE B.17: Trace plots (mean in blue, 90% credible intervals in red) and autocorrelation functions of quadratic norm of the difference $\mathcal{B}^* - \widehat{\mathcal{B}}$. *First row*: all MCMC iterations. *Second row*: MCMC iterations after burn-in (21000) and thinning by 16.

B.7.4 Model selection for model $I_1 = J_2 = 10$

The choice of the tensor rank parameter R in the PARAFAC(R) decomposition can be interpreted as a model selection problem. Let $\mathcal{M}_j$ denote the statistical model where $R = j$, that is, the case where a PARAFAC(j) decomposition is assumed for the coefficient tensor $\mathcal{B}$. Several methods exists in the Bayesian literature for model comparison and selection. The most commonly used is based on posterior model probabilities, computed through Bayes factors (e.g., see Kass and Raftery (1995), Lopes (2014)). The posterior odds of model $\mathcal{M}_j$

over model $\mathcal{M}_{j_0}$ is defined as the product of the prior odds $p(\mathcal{M}_j/p(\mathcal{M}_{j_0})$ and the Bayes factor (BF):

$$BF(j_0;j) = \frac{p(\mathbf{Y}|\mathcal{M}_j)}{p(\mathbf{Y}|\mathcal{M}_{j_0})} \cdot \frac{p(\mathcal{M}_j)}{p(\mathcal{M}_{j_0})} = \frac{p(\mathcal{M}_j|\mathbf{Y})}{p(\mathcal{M}_{j_0}|\mathbf{Y})}\,, \tag{B.134}$$

where $p(\mathbf{Y}|\mathcal{M}_j) = \int_\Theta p(\mathbf{Y}|\boldsymbol{\theta},\mathcal{M}_j)p(\boldsymbol{\theta}|\mathcal{M}_j)\,\mathrm{d}\boldsymbol{\theta}$ is the marginal likelihood for the model $\mathcal{M}_j$.

The computation of the marginal likelihood (for each model) is the most relevant issue in Bayesian model comparison based on Bayes factors, since no simple method is currently available for numerically evaluating the multi-dimensional integral. The standard Monte Carlo estimator based on drawing samples from the prior distribution has very large variance, which hampers its use in practice. One of the most widespread approaches uses the harmonic mean estimator (Kass and Raftery (1995)) for estimating the marginal likelihood. This estimator admits an interpretation as an importance sampling estimator, when the posterior distribution of the parameters is used as importance distribution. The estimator is defined as:

$$\widehat{p}(\mathbf{Y}|\mathcal{M}_j) = \left[\frac{1}{N}\sum_{n=1}^{N}\left(p(\mathbf{Y}|\boldsymbol{\theta}^{(n)},\mathcal{M}_j)\right)^{-1}\right]^{-1}\,, \tag{B.135}$$

where $\boldsymbol{\theta}^{(n)}$ represents the value of the parameters at the n-th iteration of the chain and N is the total number of MCMC iterations. However, as documented in the literature (e.g., see Neal (1994), Robert and Wraith (2009)), this estimator may suffer from infinite variance as a consequence of light tails of the likelihood function. This is indeed the case when the likelihood is Normal, as in the proposed model. Analytically integrating out one of the covariance matrices of the noise term, i.e. Σ_1, Σ_2, yields a matrix-t distribution and results in a Rao-Blackwellized estimator, but does not lead to substantial improvement in approximating the marginal likelihood.

Recently Walker (2014) proposed a method for estimating the marginal likelihood through data augmentation procedure. In practice the procedure requires to update multiple chains at each iteration of the MCMC algorithm and necessitates of a proposal distribution for starting a new chain from the existing ones. Despite the promising results in simple cases, the computational complexity of this method in high dimensional frameworks prevents its application to the model we are considering.

Therefore, we have chosen to perform model comparison on the basis of the Bayesian (or Schwartz) information criterion (BIC), which has an interpretation as the Bayes factor for a particular choice of the prior distribution for the parameters (referred to as the unit information prior, see Bollen et al. (2012)). For a sample of size T, the BIC of model $\mathcal{M}_j$ is defined as:

$$BIC_j = BIC(\mathcal{M}_j) = \log\left(L(\mathbf{Y}|\boldsymbol{\theta}_j,\mathcal{M}_j)\right) - \frac{d_j}{2}\log(T)\,.,\tag{B.136}$$

with d_j representing the total number of parameters of model $\mathcal{M}_j$. An advantage of the BIC over the Bayes factor previously defined, is that the former is invariant to the choice of the prior distribution of the parameters, while the latter is highly sensitive to it. Fig. B.18 shows the results for the comparison of 8 models, defined by varying $R = j$, with $j = 2,\ldots,9$, in the matrix regression case with $I_1 = I_2 = 10$. The two graphs represent the BIC and the root mean squared error (RMSE), respectively. Both are in favour of the model with $R = 6$, which is close to the true tensor rank chosen in this simulation, that is 5.

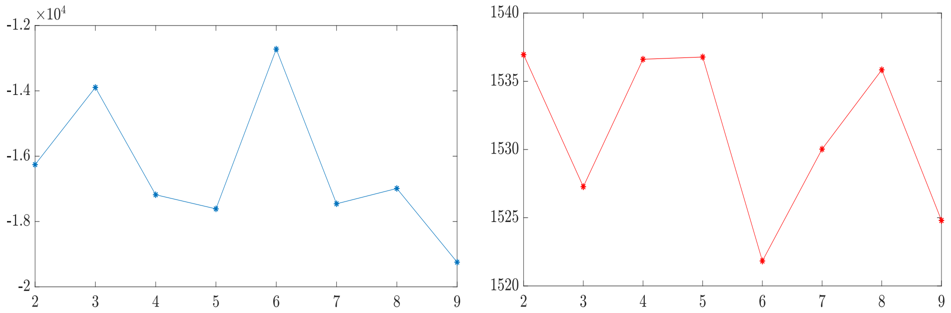


FIGURE B.18: Bayesian (or Schwartz) information criterion (*left*) and root mean squared error (*right*) for each model $\mathcal{M}_j$, with $j = 2, \ldots, 9$. Each model index corresponds to a value of tensor rank R, i.e. model $\mathcal{M}_2$ corresponds to the model with $R = 2$.

Appendix C

Appendix C

C.1 Prior distribution on tensor entries

The assumed hierarchical prior distribution on the marginals of the PARAFAC(R) decomposition assumed for the tensor of coefficients in each regime induces a prior distribution on each single entry of the tensor which is not normal. Fig. C.1-C.3 show the empirical distribution of two randomly chosen entries of a tensor $\mathcal{Y} \in \mathbb{R}^{100\times100\times3}$ whose PARAFAC decomposition is assumed with $R = 5$ and $R = 10$, respectively. Compared to the standard normal distribution and the standard Laplace distribution[1] the prior distribution induced on the single entries of the tensor is still symmetric, but has heavier tails.

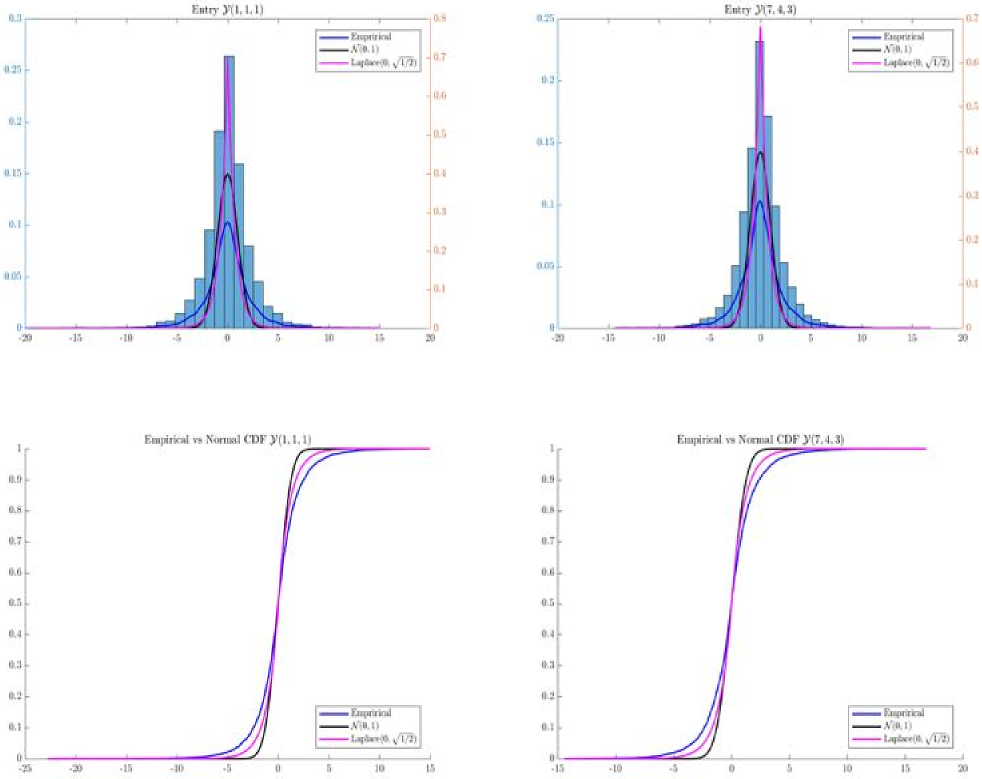


FIGURE C.1: Monte Carlo approximation of prior distribution (with $R = 5$) of an element of the tensor (histogram and dark blue line) against the standard Normal distribution (black) and the standard Laplace distribution (magenta).

[1] The probability density function of the Laplace distribution with mean μ and variance $2b^2$ is given by:

$$f(x|\mu,b) = \frac{1}{2b}\exp\left\{-\frac{|x-\mu|}{2b}\right\} \qquad x \in \mathbb{R},\ \mu \in \mathbb{R},\ b > 0$$

and has kurtosis equal to 6.

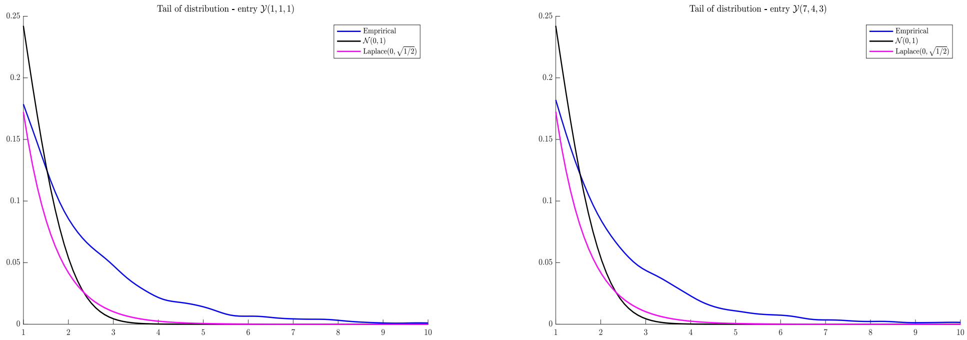


FIGURE C.2: Monte Carlo approximation of the right tail of the prior distribution (with $R = 5$) of an element of the tensor (histogram and dark blue line) against the standard Normal distribution (black) and the standard Laplace distribution (magenta).

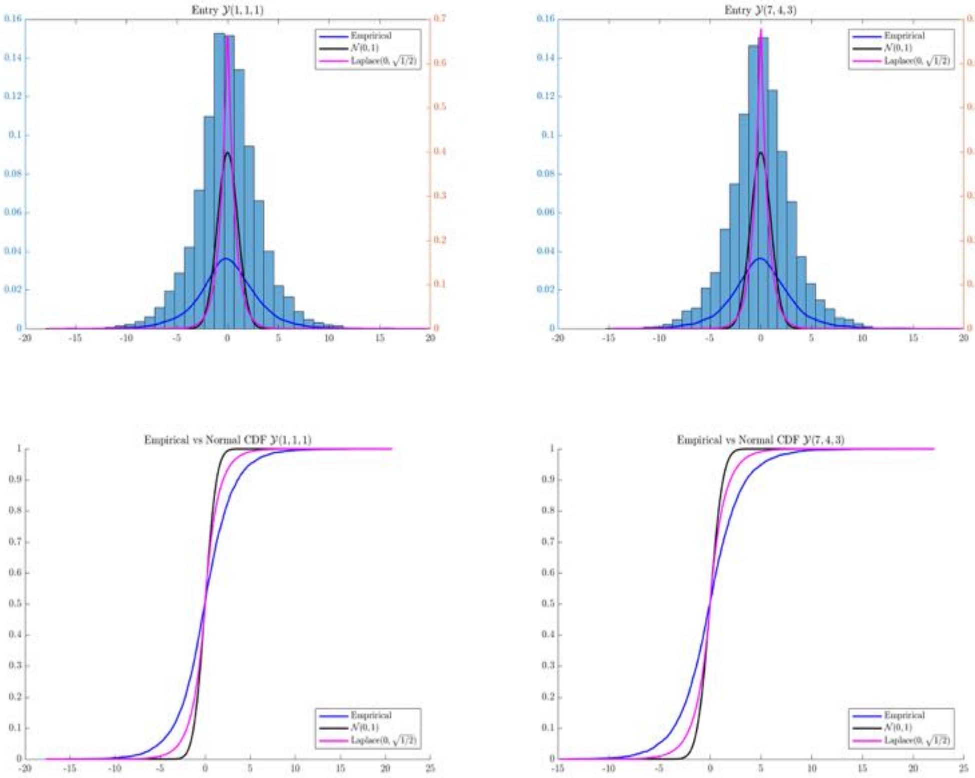


FIGURE C.3: Monte Carlo approximation of prior distribution (with $R = 10$) of an element of the tensor (histogram and dark blue line) against the standard Normal distribution (black) and the standard Laplace distribution (magenta).

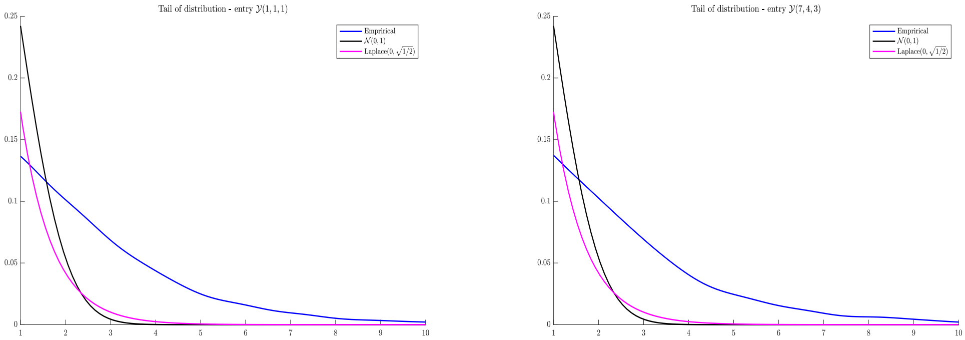


FIGURE C.4: Monte Carlo approximation of the right tail of the prior distribution (with $R = 10$) of an element of the tensor (histogram and dark blue line) against the standard Normal distribution (black) and the standard Laplace distribution (magenta).

The analytical formula for the prior distribution of the generic entry $g_{ijkp,l}$ of the fourth-order tensor $\mathcal{G}_l \in \mathbb{R}^{I \times J \times K \times P}$ can be obtained from the PARAFAC(R) decomposition in eq. (A.14) and the hierarchical prior on the marginals in eq. (3.19), (3.20), (3.21), (3.22):

$$\pi(g_{ijkp,l}) = \int_{\mathbb{R}_+} \int_{\mathcal{S}^R} \int_{(\mathbb{R}_+ \times \mathbb{R}_+)^{4R}} \pi(g_{ijkp,l} | \tau, \boldsymbol{\phi}, \mathbf{w}) \pi(\tau) \pi(\boldsymbol{\phi}) \pi(\mathbf{w}) \, \mathrm{d}\tau \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\mathbf{w} \,, \tag{C.1}$$

where $\mathcal{S}^R$ is the standard R-simplex. The entry $g_{ijkp,l}$ can be expressed in terms of the tensor marginals $\{\gamma_{h,l}^{(r)}\}_{hrl}$ as follows:

$$g_{ijkp,l} = \sum_{r=1}^{R} \gamma_{1,i,l}^{(r)} \cdot \gamma_{2,j,l}^{(r)} \cdot \gamma_{3,k,l}^{(r)} \cdot \gamma_{4,p,l}^{(r)} \,. \tag{C.2}$$

By exploiting the conditional independence relations in the hierarchical prior of the marginals in eq. (3.19), we can thus rewrite the conditional distribution $\pi(g_{ijkp,l} | \tau, \boldsymbol{\phi}, \mathbf{w})$ in eq. (C.1) as:

$$\pi(g_{ijkp,l} | \tau, \boldsymbol{\phi}, \mathbf{w}) = \mathbb{P} \left(\sum_{r=1}^{R} \gamma_{1,i,l}^{(r)} \cdot \gamma_{2,j,l}^{(r)} \cdot \gamma_{3,k,l}^{(r)} \cdot \gamma_{4,p,l}^{(r)} \right) \,, \tag{C.3}$$

which is the distribution of a finite sum of independent, univariate normal distributions, centred in zero, but with individual-specific variance. The distribution of each of these products has been characterised by Springer and Thompson (1970), who proved the following theorem.

Theorem C.1.1 (4 in Springer and Thompson (1970))
The probability density function of the product $z = \prod_{j=1}^{J} x_j$ of J independent Normal random variables $x_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 1, \ldots, J$, is a Meijer G-function multiplied by a normalising constant H:

$$p(z | \{\sigma_j^2\}_{j=1}^{J}) = H \cdot G_{J,0}^{J,0} \left(z^2 \cdot \prod_{j=1}^{J} \frac{1}{2\sigma_j} \middle| \mathbf{0} \right) \,, \tag{C.4}$$

where

$$H = \left[(2\pi)^{J/2} \cdot \prod_{j=1}^{J} \sigma_j \right]^{-1} \tag{C.5}$$

and $G^{m,n}_{p,q}(\cdot|\cdot)$ *is a Meijer G-function (with* $c \in \mathbb{R}$ *and* $s \in \mathbb{C}$*):*

$$G^{m,n}_{p,q}\left(z \middle| \begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix}\right) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} z^{-s} \frac{\prod_{j=1}^{m} \Gamma(s+b_j) \cdot \prod_{j=1}^{n} \Gamma(1-a_j-s)}{\prod_{j=n+1}^{p} \Gamma(s+a_j) \cdot \prod_{j=m+1}^{q} \Gamma(1-b_j-s)} \, \mathrm{d}s \,. \tag{C.6}$$

The integral is taken over a vertical line in the complex plane. Notice that in the special case $J = 2$ we have $z \sim c_1 P_1 - c_2 P_2$, with $P_1, P_2 \sim \chi^2_1$ and $c_1 = \mathbb{V}(x_1 + x_2)/4$, $c_2 = \mathbb{V}(x_1 - x_2)/4$.

C.2 Data augmentation

The likelihood function is:

$$L(\boldsymbol{\mathcal{X}}, \mathbf{y}|\boldsymbol{\theta}) = \sum_{s_1,\ldots,s_T} \prod_{t=1}^{T} p(\mathcal{X}_t, \mathbf{y}_t | s_t, \boldsymbol{\theta}) p(s_t|s_{t-1}) \,, \tag{C.1}$$

where the index $l \in \{1, \ldots, L\}$ represents the regime. Through the introduction of a latent variables $\mathbf{s} = \{s_t\}_{t=0}^{T}$, we obtain the data augmented likelihood:

$$L(\boldsymbol{\mathcal{X}}, \mathbf{y}, \mathbf{s}|\boldsymbol{\theta}) = \prod_{t=1}^{T} \prod_{l=1}^{L} \prod_{h=1}^{L} \left[p(\mathcal{X}_t, \mathbf{y}_t | s_t = l, \boldsymbol{\theta}) p(s_t = l | s_{t-1} = h, \Xi) \right]^{\mathbb{1}(s_t=l)\mathbb{1}(s_{t-1}=h)} \,. \tag{C.2}$$

The conditional distribution of the observation given the latent variable and marginal distribution of s_t are given by, respectively:

$$p(\mathcal{X}_t, \mathbf{y}_t | s_t = l, \boldsymbol{\theta}) = f_l(\mathcal{X}_t, \mathbf{y}_t | \boldsymbol{\theta}_l) \tag{C.3}$$

$$p(s_t = l | s_{t-1} = h, \Xi) = p_h \,. \tag{C.4}$$

Considering the observation model in eq. (3.2) and defining $\mathcal{T}_l = \{t : s_t = l\}$ for each $l = 1, \ldots, L$, we can rewrite eq. (C.2) as:

$$\begin{aligned} L(\boldsymbol{\mathcal{X}}, \mathbf{y}, \mathbf{s}|\boldsymbol{\theta}) &= \prod_{t=1}^{T} \prod_{l=1}^{L} \left[p(\mathcal{X}_t | s_t = l, \boldsymbol{\theta}) p(\mathbf{y}_t | s_t = l, \boldsymbol{\theta}) \right]^{\mathbb{1}(s_t=l)} \prod_{h=1}^{L} \left[p(s_t = l | s_{t-1} = h, \Xi) \right]^{\mathbb{1}(s_t=l)\mathbb{1}(s_{t-1}=h)} \\ &= \prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \left[(1-\rho_l) \frac{\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}}{1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}} \right]^{x_{ijk,t}} \left[\rho_l + (1-\rho_l) \frac{1}{1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}} \right]^{1-x_{ijk,t}} \\ &\cdot \prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} (2\pi)^{-m/2} |\Sigma_l|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y}_t - \boldsymbol{\mu}_l)' \boldsymbol{\Sigma}_l^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_l) \right\} \\ &\cdot \prod_{t=1}^{T} \prod_{l=1}^{L} \prod_{h=1}^{L} p_h^{\mathbb{1}(s_t=l)\mathbb{1}(s_{t-1}=h)} \,. \end{aligned} \tag{C.5}$$

Since the function cannot be expressed as a series of products due to the sum in the rightmost term, we choose to further augment the data via the through the introduction of latent allocation variables $\mathcal{D} = \{\mathcal{D}_l\}_{l=1}^{L}$, with $\mathcal{D}_l = (d_{ijk,l})$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Finally, we perform another augmentation as in Polson et al. (2013), for dealing with the logistic part of the model. When the hidden chain is assumed to be first order Markov, with two possible states, that is $L = 2$, the complete data likelihood is given by:

$$L(\boldsymbol{\mathcal{X}}, \mathbf{y}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s}|\boldsymbol{\theta}) = p(\boldsymbol{\mathcal{X}}, \mathcal{D}, \boldsymbol{\Omega}|\mathbf{s}, \boldsymbol{\theta}) p(\mathbf{y}|\mathbf{s}, \boldsymbol{\theta}) p(\mathbf{s}|\boldsymbol{\theta})$$

$$= \prod_{t=1}^{T} p(\mathcal{X}_t, \mathcal{D}_t, \boldsymbol{\Omega}_t | s_t, \boldsymbol{\theta}) p(\mathbf{y}_t | s_t, \boldsymbol{\theta}) p(s_t | \boldsymbol{\theta}) \tag{C.6a}$$

$$= \left[\prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} \prod_{i=1}^{I} \prod_{j=i}^{J} \prod_{k=1}^{K} \underbrace{p(x_{ijk,t}, d_{ijk,t}, \omega_{ijk,t} | s_t = l, \rho_l, \mathcal{G}_l)}_{I} \right] \tag{C.6b}$$

$$\cdot \left[\prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} \underbrace{p(\mathbf{y}_t | s_t = l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}_{II} \right] \cdot \left[\underbrace{p(\mathbf{s} | \Xi)}_{III} \right] \tag{C.6c}$$

where we have exploited the conditional independence of $\boldsymbol{\mathcal{X}}$ and $\mathbf{y}$ given the hidden chain $\mathbf{s}$. We start by analysing in detail the first term (I). The joint distribution of the observation $x_{ijk,t}$ and the latent variables $(d_{ijk,t}, \omega_{ijk,t})$ is obtained from the marginal distribution of the observation in two steps. First, we augment the model by introducing the latent allocation $d_{ijk,l} \in \{0, 1\}$ for $l = 1, \ldots, L$. Via this data augmentation step we are able to factorise the summation in eq. (3.2) for each regime $l = 1, \ldots, L$. In words, the allocation latent variable is used to identify the component of the mixture in eq. (3.2) from which the observation $x_{ijk,t}$ is drawn. Secondly, we use a further data augmentation step via the introduction of the latent variables $\omega_{ijk,t}$ following Polson et al. (2013), for dealing with the logistic part of the mixture.

By introducing the allocation variable $d_{ijk,t}$ in eq. (3.2), for each $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$ and $t = 1, \ldots, T$, we obtain:

$$\begin{aligned} & p(x_{ijk,t} | d_{ijk,t}, s_t = l, \rho_l, \mathcal{G}_l) \\ & = \left[\delta_{\{0\}} \right]^{\mathbb{1}\{d_{ijk,t}=1\}} \cdot \left[\mathcal{B}ern(x_{ijk,t} | \eta_{ijk,t}) \right]^{\mathbb{1}\{d_{ijk,t}=0\}} \\ & = \left[\delta_{\{0\}}(x_{ijk,t}) \right]^{d_{ijk,t}} \cdot \left[\left(\frac{\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}}{1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}} \right)^{x_{ijk,t}} \left(1 - \frac{\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}}{1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}} \right)^{1 - x_{ijk,t}} \right]^{1 - d_{ijk,t}} \\ & = \left[\delta_{\{0\}}(x_{ijk,t}) \right]^{d_{ijk,t}} \cdot \frac{\left(\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\} \right)^{x_{ijk,t}(1 - d_{ijk,t})}}{\left(1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\} \right)^{(1 - d_{ijk,t})}} . \end{aligned} \tag{C.7}$$

$$\begin{aligned} & p(x_{ijk,t}, d_{ijk,t} | s_t = l, \rho_l, \mathcal{G}_l) \\ & = \rho_l^{\mathbb{1}\{d_{ijk,t}=1\}} \cdot \left[\delta_{\{0\}}(x_{ijk,t}) \right]^{\mathbb{1}\{d_{ijk,t}=1\}} \cdot (1 - \rho_l)^{\mathbb{1}\{d_{ijk,t}=0\}} \cdot \left[\mathcal{B}ern(x_{ijk,t} | \eta_{ijk,t}) \right]^{\mathbb{1}\{d_{ijk,t}=0\}} \\ & = \rho_l^{d_{ijk,t}} \cdot \left[\delta_{\{0\}}(x_{ijk,t}) \right]^{d_{ijk,t}} \cdot (1 - \rho_l)^{1 - d_{ijk,t}} \cdot \frac{\left(\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\} \right)^{x_{ijk,t}(1 - d_{ijk,t})}}{\left(1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\} \right)^{(1 - d_{ijk,t})}} . \end{aligned} \tag{C.8}$$

The marginal distribution of the allocation variable is:

$$p(d_{ijk,t} | s_t) = \mathcal{B}ern(\rho_{s_t}) , \tag{C.9}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$ and $t = 1, \ldots, T$.

By Theorem 1 in Polson et al. (2013), it is possible to decompose the ratio in the right hand side of eq. (C.8) as follows:

$$\frac{\left(\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,l}\}\right)^{x_{ijk,t}(1-d_{ijk,t})}}{\left(1+\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,l}\}\right)^{(1-d_{ijk,t})}} = 2^{-(1-d_{ijk,t})} \int_0^\infty \exp\left\{-\frac{\omega_{ijk,t}}{2}(\mathbf{z}_t'\mathbf{g}_{ijk,l})^2 + \kappa_{ijk,t}(\mathbf{z}_t'\mathbf{g}_{ijk,l})\right\} p(\omega_{ijk,t})\,\mathrm{d}\omega_{ijk,t}\,, \tag{C.10}$$

where for every $l = 1,\dots,L$, $i = 1,\dots,I$, $j = 1,\dots,J$, $k = 1,\dots,K$ and $t = 1,\dots,T$:

$$\kappa_{ijk,t} = x_{ijk,t}(1-d_{ijk,t}) - \frac{1-d_{ijk,t}}{2} = (1-d_{ijk,t})\left(x_{ijk,t} - \frac{1}{2}\right). \tag{C.11}$$

Therefore we get the following conditional and joint distributions, respectively:

$$\begin{aligned} &p(x_{ijk,t}, d_{ijk,t} | \omega_{ijk,t}, s_t = l, \rho_l, \mathcal{G}_l) = \\ &= \rho_l^{d_{ijk,t}} \cdot \left(0^{x_{ijk,t}} 1^{1-x_{ijk,t}}\right)^{d_{ij,t}} \cdot \left(\frac{1-\rho_l}{2}\right)^{1-d_{ijk,t}} \cdot \exp\left\{-\frac{\omega_{ijk,t}}{2}(\mathbf{z}_t'\mathbf{g}_{ijk,l})^2 + \kappa_{ijk,t}(\mathbf{z}_t'\mathbf{g}_{ijk,l})\right\}. \end{aligned} \tag{C.12}$$

$$\begin{aligned} &p(x_{ijk,t}, d_{ijk,t}, \omega_{ijk,t} | s_t = l, \rho_l, \mathcal{G}_l) = \\ &= \rho_l^{d_{ijk,t}} \cdot \left(0^{x_{ijk,t}} 1^{1-x_{ijk,t}}\right)^{d_{ijk,t}} \cdot \left(\frac{1-\rho_l}{2}\right)^{1-d_{ijk,t}} \cdot \exp\left\{-\frac{\omega_{ijk,t}}{2}(\mathbf{z}_t'\mathbf{g}_{ijk,l})^2 + \kappa_{ijk,t}(\mathbf{z}_t'\mathbf{g}_{ijk,l})\right\} p(\omega_{ijk,t}). \end{aligned} \tag{C.13}$$

Finally, the marginal distribution of each latent variable $\omega_{ijk,t}$ from the data augmentation scheme follows a Pólya-Gamma distribution:

$$\omega_{ijk,t} \sim PG(1,0). \tag{C.14}$$

A continuous random variable $x \in [0,+\infty)$ has a Pólya-Gamma distribution with parameters $b > 0$, $c \in \mathbb{R}$ if the following stochastic representation holds (where $\overset{D}{=}$ stands for equality in distribution):

$$x \sim PG(b,c) \quad \Longleftrightarrow \quad x \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)} \tag{C.15}$$

where $g_k \sim \mathcal{G}a(b,1)$ are i.i.d. random variables. See Polson et al. (2013) for further details. The part (II) of eq. (C.6c) is the likelihood of a multivariate normal mean regression, hence:

$$p(\mathbf{y}_t | s_t = l, \boldsymbol{\theta}) = (2\pi)^{-m/2} |\boldsymbol{\Sigma}_l|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_l)\right\}. \tag{C.16}$$

The last term in eq. (C.6c), according to the assumption of first order time homogeneous Markov chain, factors as[2]:

$$p(s_t|\boldsymbol{\theta}) = p(s_0|\Xi) \prod_{v=1}^{t} p(s_v|s_{v-1},\Xi) = p(s_0|\Xi) \prod_{v=1}^{t} \xi_{s_{v-1},s_v} = p(s_0|\Xi) \prod_{g=1}^{L}\prod_{l=1}^{L} \xi_{g,l}^{N_{gl}(\mathbf{s}^t)} \tag{C.17}$$

where $\mathbf{s}^t = (s_0,\dots,s_t)'$ and $N_{gl}(\mathbf{s}^t)$ is a function counting the number of transitions from

[2] See (Frühwirth-Schnatter, 2006, ch.11) for more details.

state g to state l in the vector $\mathbf{s}^t$, that is (symbol # denotes the cardinality of a set): $N_{gl}(\mathbf{s}^t) = \#\{s_{t-1} = g, s_t = l\}, \forall g, l = 1, \ldots, L$. The complete data likelihood for $\mathcal{X}$ is thus obtained by plugging, for each $l = 1, \ldots, L$, eq. (C.13), eq. (C.16) and eq. (C.17) in eq. (C.6c):

$$
\begin{aligned}
&L(\boldsymbol{\mathcal{X}}, \mathbf{y}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s}|\boldsymbol{\theta}) = \\
&= \left[\prod_{l=1}^{L}\prod_{t\in\mathcal{T}_l}\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \rho_l^{d_{ijk,t}} \cdot \delta_{\{0\}}(x_{ijk,t})^{d_{ij,t}} \cdot \left(\frac{1-\rho_l}{2}\right)^{1-d_{ijk,t}} \cdot \exp\left\{-\frac{\omega_{ijk,t}}{2}(\mathbf{z}_t'\mathbf{g}_{ijk,l})^2 + \kappa_{ijk,t}(\mathbf{z}_t'\mathbf{g}_{ijk,l})\right\}\right] \\
&\cdot \left[\prod_{l=1}^{L}\prod_{t\in\mathcal{T}_l}(2\pi)^{-m/2}\,|\boldsymbol{\Sigma}_l|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}(\mathbf{y}_t-\boldsymbol{\mu}_l)\right\}\right] \\
&\cdot \left[\prod_{t=1}^{T}\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} p(\omega_{ijk,t})\right] \cdot \left[\prod_{g=1}^{L}\prod_{l=1}^{L}\xi_{g,l}^{N_{gl}(\mathbf{s})}\right] \cdot p(s_0|\Xi)\,.
\end{aligned}
\tag{C.18}
$$

C.3 Computational Details

C.3.1 Gibbs sampler

The structure of the partially collapsed Gibbs sampler (Van Dyk and Park (2008)) is as follows:

$$
\begin{gathered}
p(\mathbf{s}|\boldsymbol{\mathcal{X}}, \mathcal{G}, \boldsymbol{\rho}, \Xi) \\
p(\mathbf{D}|\mathbf{s}, \mathcal{G}, \boldsymbol{\rho}) \\
p(\boldsymbol{\Omega}|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}) \\
p(\boldsymbol{\psi}|\mathcal{G}, \mathbf{W}) \\
p(\tau|\mathcal{G}, \mathbf{W}, \boldsymbol{\phi}) \\
p(w_{h,r,l}|\gamma_{h,l}^{(r)}, \lambda_l, \boldsymbol{\phi}, \tau) \\
p(\lambda_l|w_{1,1,l}, \ldots, w_{4,R,l}) \\
p(\gamma_{1,l}^{(r)}|\gamma_{-1,l}^{(r)}, \mathcal{G}_{-r,l}, \boldsymbol{\phi}, \tau, \mathbf{W}, \boldsymbol{\mathcal{X}}, \mathbf{s}) \\
p(\gamma_{2,l}^{(r)}|\gamma_{-2,l}^{(r)}, \mathcal{G}_{-r,l}, \boldsymbol{\phi}, \tau, \mathbf{W}, \boldsymbol{\mathcal{X}}, \mathbf{s}) \\
p(\gamma_{3,l}^{(r)}|\gamma_{-3,l}^{(r)}, \mathcal{G}_{-r,l}, \boldsymbol{\phi}, \tau, \mathbf{W}, \boldsymbol{\mathcal{X}}, \mathbf{s}) \\
p(\gamma_{4,l}^{(r)}|\gamma_{-4,l}^{(r)}, \mathcal{G}_{-r,l}, \boldsymbol{\phi}, \tau, \mathbf{W}, \boldsymbol{\mathcal{X}}, \mathbf{s}) \\
p(\boldsymbol{\rho}|\mathbf{s}, \mathbf{D}) \\
p(\Xi|\mathbf{s}) \\
p(\boldsymbol{\mu}_l|\mathbf{y}, \mathbf{s}, \boldsymbol{\Sigma}_l) \sim \mathcal{N}_M(\tilde{\boldsymbol{\mu}}_l, \tilde{\mathbf{Y}}_l) \\
p(\boldsymbol{\Sigma}_l|\mathbf{y}, \mathbf{s}, \boldsymbol{\mu}_l) \sim \mathcal{IW}_M(\tilde{\nu}_l, \tilde{\boldsymbol{\Psi}}_l)\,.
\end{gathered}
$$

Step 1. sample latent variables from

$$
p(\mathbf{s}, \mathbf{D}, \boldsymbol{\Omega}|\boldsymbol{\mathcal{X}}, \mathcal{G}, \boldsymbol{\rho}, \Xi) = p(\mathbf{s}|\boldsymbol{\mathcal{X}}, \mathcal{G}, \boldsymbol{\rho}, \Xi) \cdot p(\mathbf{D}|\mathbf{s}, \mathcal{G}, \boldsymbol{\rho}) \cdot p(\boldsymbol{\Omega}|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G})
$$

- $p(\mathbf{s}|\boldsymbol{\mathcal{X}}, \mathcal{G}, \boldsymbol{\rho}, \Xi)$ via FFBS (Frühwirth-Schnatter (2006))
- $p(d_{ijk,t}|s_t, \mathcal{G}_t, \rho_t) \sim \mathcal{B}ern(\tilde{p}_{d_{ijk,t}})$
- $p(\omega_{ijkv,t}|x_{ijk,t}, s_t, \mathcal{G}_t) \sim \mathcal{PG}(1,\, \mathbf{z}_t'\mathbf{g}_{ijk,s_t})$

Step 2. sample variance hyper-parameters from

$$p(\boldsymbol{\phi},\tau,W|\mathcal{G}) = \underbrace{p(\boldsymbol{\phi}|\mathcal{G},\mathbf{W})}_{\text{collapse }\tau} \cdot p(\tau|\mathcal{G},\boldsymbol{\phi},\mathbf{W}) \cdot p(\mathbf{W}|\mathcal{G},\boldsymbol{\lambda},\boldsymbol{\phi},\tau)p(\boldsymbol{\lambda}|\mathbf{W})$$

- $p(\psi_r|\mathcal{G}^{(r)},\mathbf{w}_r) \sim GiG\left(2\bar{b}^{\tau}, \sum_{h=1}^{4}\sum_{l=1}^{L}\frac{\gamma_{h,l}^{(r)\prime}\gamma_{h,l}^{(r)}}{w_{h,r}}, \bar{\alpha}-n\right)$ then $\phi_r = \psi_r/\sum_i \psi_i$
- $p(\tau|\mathcal{G},\mathbf{W},\boldsymbol{\phi}) \sim GiG\left(2\bar{b}^{\tau}, \sum_{r=1}^{R}\sum_{h=1}^{4}\sum_{l=1}^{L}\frac{\gamma_{h,l}^{(r)\prime}\gamma_{h,l}^{(r)}}{\phi_r w_{h,r}}, (\bar{\alpha}-n)R\right)$
- $p(w_{h,r,l}|\gamma_{h,l}^{(r)},\phi_r,\tau,\lambda_l) \sim \mathrm{GiG}\left(\lambda_l^2, \frac{\gamma_{h,l}^{(r)\prime}\gamma_{h,l}^{(r)}}{\tau\phi_r}, 1-\frac{n_h}{2}\right)$
- $p(\lambda_l|\mathbf{w}_l) \propto \lambda_l^{\bar{a}_l^{\lambda}+8R-1}\exp\left\{-\lambda_l\bar{b}_l^{\lambda}\right\}\cdot\exp\left\{-\frac{\lambda_l^2}{2}\sum_{r=1}^{R}\sum_{h=1}^{4}w_{h,r,l}\right\}$

Step 3. sample tensor marginals from

$$p\left(\mathcal{G}|\boldsymbol{\mathcal{X}},\mathbf{s},\boldsymbol{\phi},\tau,\mathbf{W}\right) = \prod_{l=1}^{L} p\left(\{\gamma_{1,l}^{(r)},\gamma_{2,l}^{(r)},\gamma_{3,l}^{(r)},\gamma_{4,l}^{(r)}\}_{r=1}^{R}\middle|\boldsymbol{\mathcal{X}},\mathbf{s},\boldsymbol{\phi},\tau,\mathbf{W}\right)$$

- $p(\gamma_{1,l}^{(r)}|\gamma_{-1,l}^{(r)},\mathcal{G}_{-r,l},\boldsymbol{\phi},\tau,\mathbf{W},\boldsymbol{\mathcal{X}},\mathbf{s}) \sim \mathcal{N}_{d_1}(\boldsymbol{\mu}_{\gamma_{1,l}},\Sigma_{\gamma_{1,l}})$
- $p(\gamma_{2,l}^{(r)}|\gamma_{-2,l}^{(r)},\mathcal{G}_{-r,l},\boldsymbol{\phi},\tau,\mathbf{W},\boldsymbol{\mathcal{X}},\mathbf{s}) \sim \mathcal{N}_{d_2}(\boldsymbol{\mu}_{\gamma_{2,l}},\Sigma_{\gamma_{2,l}})$
- $p(\gamma_{3,l}^{(r)}|\gamma_{-3,l}^{(r)},\mathcal{G}_{-r,l},\boldsymbol{\phi},\tau,\mathbf{W},\boldsymbol{\mathcal{X}},\mathbf{s}) \sim \mathcal{N}_{d_3}(\boldsymbol{\mu}_{\gamma_{3,l}},\Sigma_{\gamma_{3,l}})$
- $p(\gamma_{4,l}^{(r)}|\gamma_{-4,l}^{(r)},\mathcal{G}_{-r,l},\boldsymbol{\phi},\tau,\mathbf{W},\boldsymbol{\mathcal{X}},\mathbf{s}) \sim \mathcal{N}_{d_4}(\boldsymbol{\mu}_{\gamma_{4,l}},\Sigma_{\gamma_{4,l}})$

Step 4. sample switching parameters and transition matrix from

$$p(\rho_l,\xi_{l,l}|\mathbf{s},\mathbf{D}) = p(\rho_l|\mathbf{s},\mathbf{D})\cdot p(\xi_{l,l}|\mathbf{s})$$

- $p(\rho_l|\mathbf{s},\mathbf{D}) \sim \mathcal{B}e(\tilde{a}_l,\tilde{b}_l)$
- $p(\boldsymbol{\xi}_{l,:}|\mathbf{s}) \sim \mathcal{D}ir(\tilde{\mathbf{c}})$

Step 5. sample the parameters of the second equation from

$$p(\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l|\mathbf{y},\mathbf{s}) = p(\boldsymbol{\mu}_l|\mathbf{y},\mathbf{s},\boldsymbol{\Sigma}_l)p(\boldsymbol{\Sigma}_l|\mathbf{y},\mathbf{s},\boldsymbol{\mu}_l)$$

- $p(\boldsymbol{\mu}_l|\mathbf{y},\mathbf{s},\boldsymbol{\Sigma}_l) \sim \mathcal{N}_M(\tilde{\boldsymbol{\mu}}_l,\tilde{\mathbf{Y}}_l)$
- $p(\boldsymbol{\Sigma}_l|\mathbf{y},\mathbf{s},\boldsymbol{\mu}_l) \sim \mathcal{IW}_M(\tilde{\nu}_l,\tilde{\boldsymbol{\Psi}}_l)$

The derivation of the full conditional distribution is illustrated in the following subsections.

C.3.2 Full conditional distribution of ϕ_r

The full conditional of the common (over r) component of the variance of the marginals from the PARAFAC, for each $r = 1, \ldots, R$, can be obtained in closed form collapsing τ. This can be done by exploiting a result in Guhaniyogi et al. (2017), which states that the posterior full conditional of each ϕ_r can be obtained by normalising Generalised Inverse Gaussian distributed random variables ψ_r, where $\psi_r = \tau\phi_r$:

$$p(\phi_r | \mathcal{G}^{(r)}, \mathbf{w}_r) = \frac{\psi_r}{\sum_{i=1}^{R} \psi_i} \qquad \forall\, r \tag{C.1}$$

where for every $r = 1, \ldots, R$:

$$\psi_r \sim \mathrm{GiG}\left(2\overline{b}^{\tau}, \sum_{h=1}^{4} \sum_{l=1}^{L} \frac{\gamma_{h,l}^{(r)'} \gamma_{h,l}^{(r)}}{w_{h,r,l}}, \overline{\alpha} - n \right) . \tag{C.2}$$

In the previous notation, $GiG(\cdot)$ stands for the Generalized Inverse Gaussian distribution. The Generalized Inverse Gaussian probability density function with three parameters $a > 0$, $b > 0$, $p \in \mathbb{R}$, for the random variable $x \in (0, +\infty)$, is given by:

$$x \sim GiG(a, b, p) \;\; \Rightarrow \;\; p(x|a, b, p) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left\{ -\frac{1}{2}\left(ax + \frac{b}{x} \right) \right\} \tag{C.3}$$

with $K_p(\cdot)$ a modified Bessel function of the second type.
The computation necessary for obtaining this result are as follows:

$$p(\boldsymbol{\phi}|\mathcal{G}, \mathbf{W}) \propto p(\boldsymbol{\phi}) \int_0^\infty p(\mathcal{G}|\mathbf{W}, \boldsymbol{\phi}, \tau) p(\tau) \,\mathrm{d}\tau \tag{C.4a}$$

$$\propto \prod_{r=1}^{R} \phi_r^{\overline{\alpha}-1} \int_0^\infty \prod_{r=1}^{R} \prod_{h=1}^{4} \prod_{l=1}^{L} (\tau\phi_r w_{h,r,l} \mathbf{I}_{n_h})^{-1/2} \exp\left\{ -\frac{1}{2} \gamma_{h,l}^{(r)'} (\tau\phi_r w_{h,r,l} \mathbf{I}_{n_h})^{-1} \gamma_{h,l}^{(r)} \right\}$$
$$\cdot\, \tau^{a_\tau - 1} \exp\left\{ -\overline{b}^{\tau} \tau \right\} \,\mathrm{d}\tau \tag{C.4b}$$

$$= \int_0^\infty \prod_{r=1}^{R} \phi_r^{\overline{\alpha}-1} \prod_{h=1}^{4} (\tau\phi_r w_{h,r,l} \mathbf{I}_{n_h})^{-1} \exp\left\{ -\frac{1}{2} \sum_{l=1}^{L} (\tau\phi_r w_{h,r,l})^{-1} \gamma_{h,l}^{(r)'} \gamma_{h,l}^{(r)} \right\}$$
$$\cdot\, \tau^{\overline{a}^{\tau} - 1} \exp\left\{ -\overline{b}^{\tau} \tau \right\} \,\mathrm{d}\tau \,. \tag{C.4c}$$

We define $n = n_1 + n_2 + n_3 + n_4 = I + J + K + Q$ and exploit the property $\det(kA) = k^n \det(A)$, for a square matrix A of size n and a scalar k. Finally, we assume:

$$\overline{a}^{\tau} = \overline{\alpha} R \tag{C.5}$$

which is allowed since the hyper-parameter $\overline{a}^{\tau}$ must be positive. We can thus obtain:

$$\propto \int_0^\infty \prod_{r=1}^{R} \phi_r^{\overline{\alpha}-1} \prod_{h=1}^{4} (\tau\phi_r w_{h,r,l} \mathbf{I}_{n_h})^{-1} \exp\left\{ -\frac{1}{2} \sum_{l=1}^{L} (\tau\phi_r w_{h,r,l})^{-1} \gamma_{h,l}^{(r)'} \gamma_{h,l}^{(r)} \right\}$$
$$\cdot\, \tau^{\overline{a}^{\tau} - 1} \exp\left\{ -\overline{b}^{\tau} \tau \right\} \,\mathrm{d}\tau \tag{C.6a}$$

$$\propto \int_0^\infty \prod_{r=1}^{R} (\tau\phi_r)^{\overline{\alpha}-1} (\tau\phi_r)^{-n} \exp\left\{ -\frac{1}{2} \left[2\overline{b}^{\tau}\tau + \sum_{h=1}^{4} \sum_{l=1}^{L} (\tau\phi_r w_{h,r,l})^{-1} \gamma_{h,l}^{(r)'} \gamma_{h,l}^{(r)} \right] \right\} \mathrm{d}\tau \tag{C.6b}$$

$$= \int_0^{\infty} \left(\prod_{r=1}^{R} (\tau\phi_r)^{\bar{\alpha}-n-1} \right) \exp\left\{ -\frac{1}{2} \sum_{r=1}^{R} \left[2\bar{b}^{\tau}\tau\phi_r + \frac{1}{\tau\phi_r} \sum_{h=1}^{4} \sum_{l=1}^{L} \frac{\gamma_{h,l}^{(r)\prime}\gamma_{h,l}^{(r)}}{w_{h,r,l}} \right] \right\} \mathrm{d}\tau \tag{C.6c}$$

where in the last line we used $\sum_{r=1}^{R} \phi_r = 1$. It can be seen that the integrand is the kernel of a GiG with respect to the random variable $\psi_r = \tau\phi_r$. Following Guhaniyogi et al. (2017), it is possible to sample from the posterior of ϕ_r, for each $r = 1, \ldots, R$ by first sampling ψ_r from a GiG with kernel given in eq. (C.6c), then normalising over r, as reported in eq. (C.2)-(C.1), respectively.

As an alternative, it is possible to sample from eq. (C.2) using a Hamiltonian Monte Carlo step (Neal (2011)).

C.3.3 Full conditional distribution of τ

The full conditional of the global component of the variance of the PARAFAC marginals is:

$$p(\tau|\mathcal{G}, \mathbf{W}, \boldsymbol{\phi}) \sim \mathrm{GiG}\left(2\bar{b}^{\tau}, \sum_{r=1}^{R} \sum_{h=1}^{4} \sum_{l=1}^{L} \frac{\gamma_{h,l}^{(r)\prime}\gamma_{h,l}^{(r)}}{\phi_r w_{h,r,l}}, (\bar{\alpha} - n)R \right), \tag{C.7}$$

The posterior full conditional distribution is derived from:

$$\begin{aligned} p(\tau|\mathcal{G}, \mathbf{W}, \boldsymbol{\phi}) &\propto \pi(\tau) p(\mathcal{G}|\mathbf{W}, \boldsymbol{\phi}, \tau) \\ &\propto \tau^{\bar{a}^{\tau}-1} \exp\left\{ -\bar{b}^{\tau}\tau \right\} \prod_{r=1}^{R} \prod_{h=1}^{4} \prod_{l=1}^{L} \left| \tau\phi_r w_{h,r,l} \mathbf{I}_{n_h} \right|^{-1/2} \exp\left\{ -\frac{1}{2} \gamma_{h,l}^{(r)\prime} (\tau\phi_r w_{h,r,l} \mathbf{I}_{n_h})^{-1} \gamma_{h,l}^{(r)} \right\} &&\text{(C.8a)} \\ &\propto \tau^{\bar{a}^{\tau}-nR-1} \exp\left\{ -\frac{1}{2} \left[2\bar{b}^{\tau}\tau + \frac{1}{\tau} \sum_{r=1}^{R} \sum_{h=1}^{4} \sum_{l=1}^{L} \frac{\gamma_{h,l}^{(r)\prime}\gamma_{h,l}^{(r)}}{\phi_r w_{h,r,l}} \right] \right\}, &&\text{(C.8b)} \end{aligned}$$

which is the kernel of the GiG in eq. (C.7), once the constraint in eq. (C.5) has been taken into account.

It is possible to sample from eq. (C.7) using a Hamiltonian Monte Carlo step (Neal (2011)).

C.3.4 Full conditional distribution of $w_{h,r,l}$

The full conditional distribution of the local component of the variance of each PARAFAC marginal, for $h = 1, \ldots, 4$, $r = 1, \ldots, R$ and $l = 1, \ldots, L$, is given by:

$$p(w_{h,r,l}|\gamma_{h,l}^{(r)}, \phi_r, \tau, \lambda_l) \sim \mathrm{GiG}\left(\lambda_l^2, \frac{\gamma_{h,l}^{(r)\prime}\gamma_{h,l}^{(r)}}{\tau\phi_r}, 1 - \frac{n_h}{2} \right), \tag{C.9}$$

which follows from:

$$\begin{aligned} p(w_{h,r,l}|\gamma_{h,l}^{(r)}, \phi_r, \tau, \lambda_l) &\propto \pi(w_{h,r,l}|\lambda_l) p(\gamma_{h,l}^{(r)}|w_{h,r,l}, \phi_r, \tau) &&\text{(C.10a)} \\ &\propto \exp\left\{ -\frac{\lambda_l^2}{2} w_{h,r,l} \right\} \left| \tau\phi_r w_{h,r,l} \mathbf{I}_{n_h} \right|^{-1/2} \exp\left\{ -\frac{1}{2} \gamma_{h,l}^{(r)\prime} (\tau\phi_r w_{h,r,l} \mathbf{I}_{n_h})^{-1} \gamma_{h,l}^{(r)} \right\} &&\text{(C.10b)} \end{aligned}$$

$$\propto \exp\left\{-\frac{\lambda_l^2}{2}w_{h,r,l}\right\} w_{h,r,l}^{-n_h/2} \exp\left\{-\frac{1}{2}\frac{\gamma_{h,l}^{(r)'}\gamma_{h,l}^{(r)}}{\tau\phi_r w_{h,r,l}}\right\} \tag{C.10c}$$

$$= w_{h,r,l}^{-n_h/2} \exp\left\{-\frac{1}{2}\left[\lambda_l^2 w_{h,r,l} + \frac{1}{w_{h,r,l}}\frac{\gamma_{h,l}^{(r)'}\gamma_{h,l}^{(r)}}{\tau\phi_r}\right]\right\} . \tag{C.10d}$$

It is possible to sample from eq. (C.9) using a Hamiltonian Monte Carlo step (Neal (2011)).

C.3.5 Full conditional distribution of λ_l

The full conditional distribution of λ_l, for $l = 1, \dots, L$, is given by:

$$p(\lambda_l|\mathbf{w}_l) \propto \lambda_l^{\bar{a}_l^\lambda + 8R - 1} \exp\left\{-\lambda_l \bar{b}_l^\lambda\right\} \cdot \exp\left\{-\frac{\lambda_l^2}{2}\sum_{r=1}^{R}\sum_{h=1}^{4} w_{h,r,l}\right\} . \tag{C.11}$$

It is obtained from:

$$p(\lambda_l|\mathbf{w}_l) \propto \pi(\lambda_l)p(\mathbf{w}_l|\lambda_l) \tag{C.12a}$$

$$\propto \lambda_l^{a_\lambda^l - 1} \exp\left\{-b_\lambda^l \lambda_l\right\} \prod_{r=1}^{R}\prod_{h=1}^{4} \frac{\lambda_l^2}{2} \exp\left\{-\frac{\lambda_l^2}{2} w_{h,r,l}\right\} \tag{C.12b}$$

$$\propto \lambda^{\bar{a}_l^\lambda + 8R - 1} \exp\left\{-\lambda_l \bar{b}_l^\lambda\right\} \cdot \exp\left\{-\frac{\lambda_l^2}{2}\sum_{r=1}^{R}\sum_{h=1}^{4} w_{h,r,l}\right\} . \tag{C.12c}$$

Since the second exponential is always smaller than one due to the positiveness of all the parameters $\lambda_l, w_{h,r,l}$, we can sample from this distribution by means of an accept/reject algorithm using as proposal density a Gamma distribution $\mathcal{G}a(\tilde{a}, \tilde{b})$ with parameters:

$$\tilde{a} = \bar{a}_l^\lambda + 8R \qquad \tilde{b} = \bar{b}_l^\lambda . \tag{C.13}$$

Since this sampling scheme has very low acceptance rate, it is possible to sample from eq. (C.11) using a Hamiltonian Monte Carlo step (Neal (2011)).

C.3.6 Full conditional distribution of $\gamma_{h,l}^{(r)}$

For deriving the full conditional distribution of each PARAFAC marginal, $\gamma_{h,l}^{(r)}$, of the tensor $\mathcal{G}_l$, $l = 1, \dots, L$, we start by manipulating the complete data likelihood in eq. (3.18) with the aim of singling out $\gamma_{h,l}^{(r)}$. From eq. (C.13), considering all the entries of $\mathcal{X}_t$ at a given $t \in \{1, \dots, T\}$ and denoting with $\pi(\mathcal{G}_l)$ the prior distribution induced on $\mathcal{G}_l$ by the hierarchical prior on the PARAFAC marginals in eq. (3.19), the following proportionality relation holds:

$$p(\mathcal{G}_l|\mathcal{X}_t, \mathcal{D}_t, \mathbf{\Omega}_t, s_t = l, \rho_l) \propto \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \exp\left\{-\frac{\omega_{ijk,t}}{2}(\mathbf{z}_t'\mathbf{g}_{ijk,l})^2 + \kappa_{ijk,t}(\mathbf{z}_t'\mathbf{g}_{ijk,l})\right\} p(\omega_{ijk,t})\pi(\mathcal{G}_l)$$

$$= \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \exp\left\{-\frac{1}{2\omega_{ijk,t}^{-1}}\left[(\mathbf{z}_t'\mathbf{g}_{ijk,l})^2 - 2\frac{\kappa_{ijk,t}}{\omega_{ijk,t}}(\mathbf{z}_t'\mathbf{g}_{ijk,l})\right]\right\} p(\omega_{ijk,t})\pi(\mathcal{G}_l)$$

$$= \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \exp \left\{ -\frac{1}{2\omega_{ijk,t}^{-1}} \left(\mathbf{z}_t' \mathbf{g}_{ijk,l} - \frac{\kappa_{ijk,t}}{\omega_{ijk,t}} \right)^2 \right\} p(\omega_{ijk,t}) \pi(\mathcal{G}_l). \tag{C.14}$$

Define $u_{ijk,t} = \kappa_{ijk,t}/\omega_{ijk,t}$, then we rewrite eq. (C.14) in more compact form as:

$$\begin{aligned} & p(\mathcal{G}_l | \mathcal{X}_t, \mathcal{D}_t, \mathbf{\Omega}_t, s_t = l, \rho_l) \propto \\ & \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{1}{\omega_{ijk,t}^{-1}} \left(\mathbf{z}_t' \mathbf{g}_{ijk,l} - u_{ijk,t} \right)^2 \right\} \cdot \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} p(\omega_{ijk,t}) \cdot \pi(\mathcal{G}_l) \\ & = \exp \left\{ -\frac{1}{2} \sum_{i=1}^{I} (\mathcal{G}_l \times_4 \mathbf{z}_t - \mathcal{U}_t)_i' \operatorname{diag}\left(\boldsymbol{\omega}_{i:,t} \right) (\mathcal{G}_l \times_4 \mathbf{z}_t - \mathcal{U}_t)_i \right\} \cdot \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} p(\omega_{ijk,t}) \cdot \pi(\mathcal{G}_l) \\ & = \exp \left\{ -\frac{1}{2} \left(\operatorname{vec}\left(\mathcal{G}_l \times_4 \mathbf{z}_t \right) - \operatorname{vec}\left(\mathcal{U}_t \right) \right)' \operatorname{diag}\left(\operatorname{vec}\left(\mathbf{\Omega}_t \right) \right) \left(\operatorname{vec}\left(\mathcal{G}_l \times_4 \mathbf{z}_t \right) - \operatorname{vec}\left(\mathcal{U}_t \right) \right) \right\} \\ & \cdot \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} p(\omega_{ijk,t}) \cdot \pi(\mathcal{G}_l) \\ & = f\left(\mathcal{G}_l, \mathbf{z}_t, \mathcal{U}_t, \mathbf{\Omega}_t \right) \cdot \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} p(\omega_{ijk,t}) \cdot \pi(\mathcal{G}_l), \end{aligned} \tag{C.15}$$

where $f(\cdot)$ is a function which contains the kernel of a multivariate normal distribution with respect to the variable $\operatorname{vec}\left(\mathcal{G}_l \times_4 \mathbf{z}_t \right)$.

Given the proportionality relation conditional on the latent variable s_t, the last step in the manipulation of the likelihood function consists in rewriting the complete data likelihood. Thus, considering eq. (3.18) and (C.15) we obtain the proportionality relation:

$$L(\boldsymbol{\mathcal{X}}, \mathcal{D}, \mathbf{\Omega}, \mathbf{s} | \boldsymbol{\theta}) = \prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} p(\mathcal{X}_t, \mathcal{D}_t, \mathbf{\Omega}_t, s_t | \boldsymbol{\theta}) \propto \prod_{l=1}^{L} \prod_{t \in \mathcal{T}_l} f\left(\mathcal{G}_{s_t}, \mathbf{z}_t, \mathcal{U}_t, \mathbf{\Omega}_t \right) . \tag{C.16}$$

We are now ready to compute the full conditional distributions of each vector $\gamma_{h,l}^{(r)}$, $h = 1, \ldots, 4$, $l = 1, \ldots, L$ and $r = 1, \ldots, R$. To this aim, notice that:

$$\mathcal{G}_l = \sum_{r=1}^{R} \gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)} \circ \gamma_{4,l}^{(r)} = \mathcal{G}_l^{(r)} + \mathcal{G}_l^{(-r)} , \tag{C.17}$$

where we have defined:

$$\mathcal{G}_l^{(r)} = \gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)} \circ \gamma_{4,l}^{(r)} \tag{C.18a}$$

$$\mathcal{G}_l^{(-r)} = \sum_{\substack{v=1 \\ v \neq r}}^{R} \mathcal{G}_l^{(v)} . \tag{C.18b}$$

By exploiting the definitions of mode-n product and PARAFAC decomposition, we obtain:

$$\mathcal{G}_{l,t} = \mathcal{G}_l \times_4 \mathbf{z}_t = \sum_{r=1}^{R} \left(\gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)} \right) \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle = \sum_{r=1}^{R} \mathcal{G}_{l,t}^{(r)} . \tag{C.19}$$

Here $\langle \cdot, \cdot \rangle$ denotes the standard inner product in the Euclidean space $\mathbb{R}^n$. Since the latter is a scalar, we have that:

$$\bar{\mathbf{g}}_{l,t} = \operatorname{vec}\left(\mathcal{G}_{l,t}\right) = \operatorname{vec}\left(\mathcal{G}_l \times_4 \mathbf{z}_t\right) = \sum_{r=1}^{R} \operatorname{vec}\left(\mathcal{G}_{l,t}^{(r)}\right) = \sum_{r=1}^{R} \bar{\mathbf{g}}_{l,t}^{(r)} . \tag{C.20}$$

The vectorisation of a tensor can be expressed in the following way, which is a generalisation of a well known property holding for matrices: it consists in stacking in a column vector all the vectorised slices of the tensor. For the sake of clarity, let $\boldsymbol{\alpha}_1 \in \mathbb{R}^I$, $\boldsymbol{\alpha}_2 \in \mathbb{R}^J$ and $\boldsymbol{\alpha}_3 \in \mathbb{R}^K$ and let the tensor $\mathcal{A} = \boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2 \circ \boldsymbol{\alpha}_3$. Denote $\mathcal{A}_{::k} \in \mathbb{R}^{I \times J}$ the k-th frontal slice of the tensor $\mathcal{A}$. Then, by applying the properties of Kronecker product, $\otimes$, and of the vectorization operator, vec, we obtain[3]:

$$\begin{aligned} \operatorname{vec}\left(\mathcal{A}\right) = \operatorname{vec}\left(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2 \circ \boldsymbol{\alpha}_3\right) &= \left[\operatorname{vec}\left(\mathcal{A}_{::1}\right)', \ldots, \operatorname{vec}\left(\mathcal{A}_{::K}\right)'\right]' \\ &= \left[\operatorname{vec}\left(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2\right)' \alpha_{3,1}, \ldots, \operatorname{vec}\left(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2\right)' \alpha_{3,K}\right]' \\ &= \boldsymbol{\alpha}_3 \otimes \operatorname{vec}\left(\boldsymbol{\alpha}_1 \circ \boldsymbol{\alpha}_2\right) = \boldsymbol{\alpha}_3 \otimes \operatorname{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right) . \end{aligned} \tag{C.21}$$

The use of the same property allows to rewrite eq. (C.21) in three equivalent ways, each one written as a product of a matrix and one of the vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3$, respectively. In fact, we have:

$$\operatorname{vec}\left(\mathcal{A}\right) = \boldsymbol{\alpha}_3 \otimes \operatorname{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right) = \boldsymbol{\alpha}_3 \otimes \left(\boldsymbol{\alpha}_2 \otimes \mathbf{I}_I\right) \operatorname{vec}\left(\boldsymbol{\alpha}_1\right) = \left(\boldsymbol{\alpha}_3 \otimes \boldsymbol{\alpha}_2 \otimes \mathbf{I}_I\right) \boldsymbol{\alpha}_1 \tag{C.22}$$

$$\operatorname{vec}\left(\mathcal{A}\right) = \boldsymbol{\alpha}_3 \otimes \operatorname{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right) = \boldsymbol{\alpha}_3 \otimes \left[\left(\mathbf{I}_J \otimes \boldsymbol{\alpha}_1\right) \operatorname{vec}\left(\boldsymbol{\alpha}_2'\right)\right] = \left(\boldsymbol{\alpha}_3 \otimes \mathbf{I}_J \otimes \boldsymbol{\alpha}_1\right) \boldsymbol{\alpha}_2 \tag{C.23}$$

$$\begin{aligned} \operatorname{vec}\left(\mathcal{A}\right) = \boldsymbol{\alpha}_3 \otimes \operatorname{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right) &= \operatorname{vec}\left(\operatorname{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right) \boldsymbol{\alpha}_3'\right) = \left(\mathbf{I}_K \otimes \operatorname{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right)\right) \operatorname{vec}\left(\boldsymbol{\alpha}_3'\right) \\ &= \left(\mathbf{I}_K \otimes \operatorname{vec}\left(\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2'\right)\right) \boldsymbol{\alpha}_3 = \left(\mathbf{I}_K \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_1\right) \boldsymbol{\alpha}_3 . \end{aligned} \tag{C.24}$$

The first line represents a product between the matrix $\boldsymbol{\alpha}_3 \otimes \boldsymbol{\alpha}_2 \otimes \mathbf{I}_I \in \mathbb{R}^{IJK \times I}$ and the vector $\boldsymbol{\alpha}_1$, the second is a product between the matrix $\boldsymbol{\alpha}_3 \otimes \mathbf{I}_J \otimes \boldsymbol{\alpha}_1 \in \mathbb{R}^{IJK \times J}$ and the vector $\boldsymbol{\alpha}_2$. Finally, the last row is a product between the matrix $\mathbf{I}_K \otimes \boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_1 \in \mathbb{R}^{IJK \times K}$ and the vector $\boldsymbol{\alpha}_3$.

Starting from eq. (C.20), we can apply for $\gamma_{1,l}^{(r)}, \ldots, \gamma_{3,l}^{(r)}$ the same argument as for $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_3$, with the aim of manipulating the likelihood function and obtain three different expressions

[3] The outer product and Kronecker products are two operators acting on:

$$\begin{aligned} \circ : \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_K} &\to \mathbb{R}^{n_1 \times \ldots \times n_K} \\ \otimes : \mathbb{R}^{n_1 \times m_1} \times \mathbb{R}^{n_2 \times m_2} &\to \mathbb{R}^{n_1 n_2 \times m_1 m_2} . \end{aligned}$$

Notice that the Kronecker product is defined on the space of matrices (and vectors, as a particular case), while the outer product is defined on arrays of possible different number of dimensions (e.g. it is defined between two vectors, and returns a matrix, as well as between a vector and a matrix, yielding a third order tensor). In practice, in the particular case arising when dealing with two vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^m$, their outer product and Kronecker product are related and given by, respectively:

$$\mathbf{u} \circ \mathbf{v} = \mathbf{u}\mathbf{v}' \in \mathbb{R}^{n \times m}$$

$$\mathbf{u} \otimes \mathbf{v} = \operatorname{vec}\left(\mathbf{v}\mathbf{u}'\right) = \operatorname{vec}\left(\mathbf{v} \circ \mathbf{u}\right) \in \mathbb{R}^{nm} .$$

For two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times k}$ it holds:

$$\operatorname{vec}\left(\mathbf{AB}\right) = \left(\mathbf{I}_k \otimes \mathbf{A}\right) \operatorname{vec}\left(\mathbf{B}\right) = \left(\mathbf{B}' \otimes \mathbf{I}_m\right) \operatorname{vec}\left(\mathbf{A}\right) \in \mathbb{R}^{mk \times 1} .$$

Moreover, if $n = 1$ then $\mathbf{B}$ is a row vector of length k, as a consequence $\mathbf{B}' = \operatorname{vec}\left(\mathbf{B}\right) \in \mathbb{R}^{k \times 1}$. See (Cichocki et al., 2009, p.31).

where the dependence on $\gamma_{1,l}^{(r)}, \gamma_{2,l}^{(r)}, \gamma_{3,l}^{(r)}$, respectively, is made explicit. This will then be used later on for deriving the posterior full conditional distributions of the PARAFAC marginals. Thus, from eq. (C.20) we have:

$$\bar{\mathbf{g}}_{l,t}^{(r)} = \operatorname{vec}\left(\mathcal{G}_{l,t}^{(r)}\right) = \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \operatorname{vec}\left(\gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)}\right) = \operatorname{vec}\left(\gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)}\right) \mathbf{z}_t' \gamma_{4,l}^{(r)} = \mathbf{A}_4 \gamma_{4,l}^{(r)}, \tag{C.25}$$

where:

$$\mathbf{A}_4 = \operatorname{vec}\left(\gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)}\right) \mathbf{z}_t'. \tag{C.26}$$

Exploiting eq. (C.22) we have:

$$\bar{\mathbf{g}}_{l,t}^{(r)} = \operatorname{vec}\left(\mathcal{G}_{l,t}^{(r)}\right) = \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \left(\gamma_{3,l}^{(r)} \otimes \gamma_{2,l}^{(r)} \otimes \mathbf{I}_I\right) \gamma_{1,l}^{(r)} = \mathbf{A}_1 \gamma_{1,l}^{(r)}, \tag{C.27}$$

with:

$$\mathbf{A}_1 = \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \left(\gamma_{3,l}^{(r)} \otimes \gamma_{2,l}^{(r)} \otimes \mathbf{I}_I\right). \tag{C.28}$$

Exploiting eq. (C.23) we have:

$$\bar{\mathbf{g}}_{l,t}^{(r)} = \operatorname{vec}\left(\mathcal{G}_{l,t}^{(r)}\right) = \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \left(\gamma_{3,l}^{(r)} \otimes \mathbf{I}_J \otimes \gamma_{1,l}^{(r)}\right) \gamma_{2,l}^{(r)} = \mathbf{A}_2 \gamma_{2,l}^{(r)}, \tag{C.29}$$

with:

$$\mathbf{A}_2 = \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \left(\gamma_{3,l}^{(r)} \otimes \mathbf{I}_J \otimes \gamma_{1,l}^{(r)}\right). \tag{C.30}$$

Finally, using eq. (C.24) we obtain:

$$\bar{\mathbf{g}}_{l,t}^{(r)} = \operatorname{vec}\left(\mathcal{G}_{l,t}^{(r)}\right) = \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \left(\mathbf{I}_K \otimes \gamma_{2,l}^{(r)} \otimes \gamma_{1,l}^{(r)}\right) \gamma_{3,l}^{(r)} = \mathbf{A}_3 \gamma_{3,l}^{(r)}, \tag{C.31}$$

with:

$$\mathbf{A}_3 = \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \left(\mathbf{I}_K \otimes \gamma_{2,l}^{(r)} \otimes \gamma_{1,l}^{(r)}\right). \tag{C.32}$$

By using the definition of $f(\mathcal{G}_l, \mathbf{z}_t, \mathcal{U}_t^{(l)}, \mathbf{\Omega}_t)$, eq. (C.20) and the notation of eq. (C.17) we can thus write:

$$\operatorname{vec}\left(\mathcal{G}_l \times_4 \mathbf{z}_t\right) = \bar{\mathbf{g}}_{l,t}^{(r)} + \sum_{\substack{v=1 \\ v \neq r}}^{R} \bar{\mathbf{g}}_{l,t}^{(v)} = \bar{\mathbf{g}}_{l,t}^{(r)} + \bar{\mathbf{g}}_{l,t}^{(-r)}. \tag{C.33}$$

From eq. (C.16), by focusing on regime $l \in \{1, \ldots, L\}$, we get:

$$\begin{aligned} & L(\boldsymbol{\mathcal{X}}, \mathcal{D}, \mathbf{\Omega}, \mathbf{s} | \boldsymbol{\theta}) \propto \\ & \propto \exp\left\{-\frac{1}{2}\left(\operatorname{vec}\left(\mathcal{G}_l \times_4 \mathbf{z}_t\right) - \operatorname{vec}\left(\mathcal{U}_t\right)\right)' \operatorname{diag}\left(\operatorname{vec}\left(\mathbf{\Omega}_t\right)\right) \left(\operatorname{vec}\left(\mathcal{G}_l \times_4 \mathbf{z}_t\right) - \operatorname{vec}\left(\mathcal{U}_t\right)\right)\right\} \\ & = \exp\left\{-\frac{1}{2}\left(\bar{\mathbf{g}}_{l,t}^{(r)} + \bar{\mathbf{g}}_{l,t}^{(-r)} - \mathbf{u}_t\right)' \overline{\overline{\mathbf{\Omega}}}_t \left(\bar{\mathbf{g}}_{l,t}^{(r)} + \bar{\mathbf{g}}_{l,t}^{(-r)} - \mathbf{u}_t\right)\right\} \end{aligned} \tag{C.34}$$

where, for reducing the burden of notation, we have defined:

$$\mathbf{u}_t = \operatorname{vec}\left(\mathcal{U}_t\right) \tag{C.35}$$

$$\overline{\overline{\mathbf{\Omega}}}_t = \operatorname{diag}\left(\operatorname{vec}\left(\mathbf{\Omega}_t\right)\right). \tag{C.36}$$

We can now single out a specific component $\mathcal{G}_l^{(r)}$ of the PARAFAC decomposition of the tensor $\mathcal{G}$, which is incorporated in $\bar{\mathbf{g}}_{l,t}^{(r)}$. In fact, we can manipulate the function in eq. (C.34) with the aim of finding a proportionality relation, as follows:

$$\begin{aligned} L(\boldsymbol{\mathcal{X}}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s}|\boldsymbol{\theta}) &\propto \prod_{t\in\mathcal{T}_l} \exp\left\{ -\frac{1}{2} \left[\bar{\mathbf{g}}_{l,t}^{(r)\prime} \overline{\overline{\boldsymbol{\Omega}}}_t \bar{\mathbf{g}}_{l,t}^{(r)} + \mathbf{g}_{l,t}^{(r)\prime} \overline{\overline{\boldsymbol{\Omega}}}_t (\bar{\mathbf{g}}_{l,t}^{(-r)} - \mathbf{u}_t) \right.\right. \\ &\qquad \left.\left. + (\bar{\mathbf{g}}_{l,t}^{(-r)} - \mathbf{u}_t)' \overline{\overline{\boldsymbol{\Omega}}}_t \bar{\mathbf{g}}_{l,t}^{(r)} + (\bar{\mathbf{g}}_{l,t}^{(-r)} - \mathbf{u}_t)' \overline{\overline{\boldsymbol{\Omega}}}_t (\bar{\mathbf{g}}_{l,t}^{(-r)} - \mathbf{u}_t) \right]\right\} \\ &\propto \prod_{t\in\mathcal{T}_l} \exp\left\{ -\frac{1}{2} \left[\bar{\mathbf{g}}_{l,t}^{(r)\prime} \overline{\overline{\boldsymbol{\Omega}}}_t \bar{\mathbf{g}}_{l,t}^{(r)} - 2(\mathbf{u}_t - \bar{\mathbf{g}}_{l,t}^{(-r)})' \overline{\overline{\boldsymbol{\Omega}}}_t \bar{\mathbf{g}}_{l,t}^{(r)} \right] \right\} . \end{aligned} \tag{C.37}$$

Full conditional distribution of $\gamma_{1,l}^{(r)}$

The full conditional distribution of $\gamma_{1,l}^{(r)}$ is given by:

$$p(\gamma_{1,l}^{(r)} | \boldsymbol{\mathcal{X}}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s}, \gamma_{2,l}^{(r)}, \gamma_{3,l}^{(r)}, \gamma_{4,l}^{(r)}, \mathcal{G}_l^{(-r)}, w_{1,r}, \phi_r, \tau) \sim \mathcal{N}_I(\tilde{\zeta}_{1,l}^r, \tilde{\boldsymbol{\Lambda}}_{1,l}^r) \tag{C.38}$$

where:

$$\tilde{\boldsymbol{\Lambda}}_{1,l}^r = \left[(\tau\phi_r w_{1,r} \mathbf{I}_I)^{-1} + \sum_{t\in\mathcal{T}_l} \left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)} \right)^{-1} \right]^{-1} \tag{C.39a}$$

$$\tilde{\zeta}_{1,l}^r = \tilde{\boldsymbol{\Lambda}}_{1,l}^{r\prime} \left[\bar{\zeta}_{1,l}^{r\prime} \left(\tau\phi_r w_{1,r} \mathbf{I}_I \right)^{-1} + \sum_{t\in\mathcal{T}_l} \bar{\boldsymbol{\mu}}_{1,l,t}^{(r)\prime} \left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)} \right)^{-1} \right]' . \tag{C.39b}$$

By exploiting the rightmost term in the equality chain in eq. (C.27), we can simplify the two addenda in eq. (C.37) as:

$$\begin{aligned} \bar{\mathbf{g}}_{l,t}^{(r)\prime} \overline{\overline{\boldsymbol{\Omega}}}_t \bar{\mathbf{g}}_{l,t}^{(r)} &= \left(\mathbf{A}_1 \gamma_{1,l}^{(r)} \right)' \overline{\overline{\boldsymbol{\Omega}}}_t \left(\mathbf{A}_1 \gamma_{1,l}^{(r)} \right) \\ &= \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \gamma_{1,l}^{(r)\prime} \left(\gamma_{3,l}^{(r)} \otimes \gamma_{2,l}^{(r)} \otimes \mathbf{I}_I \right)' \overline{\overline{\boldsymbol{\Omega}}}_t \left(\gamma_{3,l}^{(r)} \otimes \gamma_{2,l}^{(r)} \otimes \mathbf{I}_I \right) \gamma_{1,l}^{(r)} \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \\ &= \gamma_{1,l}^{(r)\prime} \left[\left(\gamma_{3,l}^{(r)\prime} \otimes \gamma_{2,l}^{(r)\prime} \otimes \mathbf{I}_I' \right) \overline{\overline{\boldsymbol{\Omega}}}_t \left(\gamma_{3,l}^{(r)} \otimes \gamma_{2,l}^{(r)} \otimes \mathbf{I}_I \right) (\langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle)^2 \right] \gamma_{1,l}^{(r)} \\ &= \gamma_{1,l}^{(r)\prime} \left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)} \right)^{-1} \gamma_{1,l}^{(r)} . \end{aligned} \tag{C.40}$$

and

$$\begin{aligned} -2(\mathbf{u}_t - \bar{\mathbf{g}}_{l,t}^{(-r)})' \overline{\overline{\boldsymbol{\Omega}}}_t \bar{\mathbf{g}}_{l,t}^{(r)} &= -2(\mathbf{u}_t - \bar{\mathbf{g}}_{l,t}^{(-r)})' \overline{\overline{\boldsymbol{\Omega}}}_t \left(\gamma_{3,l}^{(r)} \otimes \gamma_{2,l}^{(r)} \otimes \mathbf{I}_I \right) \gamma_{1,l}^{(r)} \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle \\ &= -2 \langle \gamma_{4,l}^{(r)}, \mathbf{z}_t \rangle (\mathbf{u}_t - \bar{\mathbf{g}}_{l,t}^{(-r)})' \overline{\overline{\boldsymbol{\Omega}}}_t \left(\gamma_{3,l}^{(r)} \otimes \gamma_{2,l}^{(r)} \otimes \mathbf{I}_I \right) \gamma_{1,l}^{(r)} \\ &= -2 \bar{\boldsymbol{\mu}}_{1,l,t}^{(r)\prime} \left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)} \right)^{-1} \gamma_{1,l}^{(r)} . \end{aligned} \tag{C.41}$$

Now, by applying Bayes' rule and plugging eq. (C.40) and eq. (C.41) into eq. (C.37) we get:

$$p(\gamma_{1,l}^{(r)} | -) \propto L(\boldsymbol{\mathcal{X}}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s}|\boldsymbol{\theta}) \pi(\gamma_{1,l}^{(r)} | \mathbf{w}_{1,:}, \boldsymbol{\phi}, \tau)$$

$$
\begin{aligned}
&\propto \prod_{t\in\mathcal{T}_l} \exp\left\{-\frac{1}{2}\left[\gamma_{1,l}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\gamma_{1,l}^{(r)} - 2\overline{\boldsymbol{\mu}}_{1,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\gamma_{1,l}^{(r)}\right]\right\}\\
&\cdot \exp\left\{-\frac{1}{2}\left[\gamma_{1,l}^{(r)'}\left(\overline{\boldsymbol{\Lambda}}_{1,l}^{r}\right)^{-1}\gamma_{1,l}^{(r)} - 2\overline{\zeta}_{1,l}^{r'}\left(\overline{\boldsymbol{\Lambda}}_{1,l}^{r}\right)^{-1}\gamma_{1,l}^{(r)}\right]\right\}\\
&= \exp\left\{-\frac{1}{2}\left[\sum_{t\in\mathcal{T}_l}\left(\gamma_{1,l}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\gamma_{1,l}^{(r)} - 2\overline{\boldsymbol{\mu}}_{1,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\gamma_{1,l}^{(r)}\right)\right.\right.\\
&\left.\left.+\left(\gamma_{1,l}^{(r)'}\left(\overline{\boldsymbol{\Lambda}}_{1,l}^{r}\right)^{-1}\gamma_{1,l}^{(r)} - 2\overline{\zeta}_{1,l}^{r'}\left(\overline{\boldsymbol{\Lambda}}_{1,l}^{r}\right)^{-1}\gamma_{1,l}^{(r)}\right)\right]\right\}\\
&= \exp\left\{-\frac{1}{2}\left[\gamma_{1,l}^{(r)'}\left(\sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\right)\gamma_{1,l}^{(r)} - 2\left(\sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}_{1,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\right)\gamma_{1,l}^{(r)}\right.\right.\\
&\left.\left.+\gamma_{1,l}^{(r)'}\left(\overline{\boldsymbol{\Lambda}}_{1,l}^{r}\right)^{-1}\gamma_{1,l}^{(r)} - 2\overline{\zeta}_{1,l}^{r'}\left(\overline{\boldsymbol{\Lambda}}_{1,l}^{r}\right)^{-1}\gamma_{1,l}^{(r)}\right]\right\}\\
&= \exp\left\{-\frac{1}{2}\left[\gamma_{1,l}^{(r)'}\left(\left(\overline{\boldsymbol{\Lambda}}_{1,l}^{r}\right)^{-1} + \sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\right)\gamma_{1,l}^{(r)}\right.\right.\\
&\left.\left.-2\left(\overline{\zeta}_{1,l}^{r'}\left(\overline{\boldsymbol{\Lambda}}_{1,l}^{r}\right)^{-1} + \sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}_{1,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\right)\gamma_{1,l}^{(r)}\right]\right\}. \qquad \text{(C.42)}
\end{aligned}
$$

This is the kernel of a multivariate normal distribution with parameters:

$$
\tilde{\boldsymbol{\Lambda}}_{1,l}^{r} = \left[\left(\tau\phi_r w_{1,r}\mathbf{I}_I\right)^{-1} + \sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\right]^{-1} \qquad \text{(C.43a)}
$$

$$
\tilde{\zeta}_{1,l}^{r} = \tilde{\boldsymbol{\Lambda}}_{1,l}^{r'}\left[\overline{\zeta}_{1,l}^{r'}\left(\tau\phi_r w_{1,r}\mathbf{I}_I\right)^{-1} + \sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}_{1,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{1,l,t}^{(r)}\right)^{-1}\right]'. \qquad \text{(C.43b)}
$$

Full conditional distribution of $\gamma_{2,l}^{(r)}$

The full conditional distribution of $\gamma_{2,l}^{(r)}$ is given by:

$$
p(\gamma_{2,l}^{(r)}|\boldsymbol{\mathcal{X}}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s}, \gamma_{1,l}^{(r)}, \gamma_{3,l}^{(r)}, \gamma_{4,l}^{(r)}, \mathcal{G}_l^{(-r)}, w_{2,r}, \phi_r, \tau) \sim \mathcal{N}_J(\tilde{\zeta}_{2,l}^{r}, \tilde{\boldsymbol{\Lambda}}_{2,l}^{r}) \qquad \text{(C.44)}
$$

where:

$$
\tilde{\boldsymbol{\Lambda}}_{2,l}^{r} = \left[\left(\tau\phi_r w_{2,r}\mathbf{I}_J\right)^{-1} + \sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\right]^{-1} \qquad \text{(C.45a)}
$$

$$
\tilde{\zeta}_{2,l}^{r} = \tilde{\boldsymbol{\Lambda}}_{2,l}^{r'}\left[\overline{\zeta}_{2,l}^{r'}\left(\tau\phi_r w_{2,r}\mathbf{I}_J\right)^{-1} + \sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}_{2,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\right]'. \qquad \text{(C.45b)}
$$

By exploiting the central term in the equality chain in eq. (C.29), we can simplify the two addenda in eq. (C.37) as:

$$
\begin{aligned}
\overline{\mathbf{g}}_{l,t}^{(r)'}\overline{\overline{\boldsymbol{\Omega}}}_t\overline{\mathbf{g}}_{l,t}^{(r)} &= \left(\mathbf{A}_2\gamma_{2,l}^{(r)}\right)'\overline{\overline{\boldsymbol{\Omega}}}_t\left(\mathbf{A}_2\gamma_{2,l}^{(r)}\right) \\
&= \langle\gamma_{4,l}^{(r)},\mathbf{z}_t\rangle\gamma_{2,l}^{(r)'}\left(\gamma_{3,l}^{(r)}\otimes\mathbf{I}_J\otimes\gamma_{1,l}^{(r)}\right)'\overline{\overline{\boldsymbol{\Omega}}}_t\left(\gamma_{3,l}^{(r)}\otimes\mathbf{I}_J\otimes\gamma_{1,l}^{(r)}\right)\gamma_{2,l}^{(r)}\langle\gamma_{4,l}^{(r)},\mathbf{z}_t\rangle \\
&= \gamma_{2,l}^{(r)'}\left[\left(\gamma_{3,l}^{(r)'}\otimes\mathbf{I}_J'\otimes\gamma_{1,l}^{(r)'}\right)\overline{\overline{\boldsymbol{\Omega}}}_t\left(\gamma_{3,l}^{(r)}\otimes\mathbf{I}_J\otimes\gamma_{1,l}^{(r)}\right)(\langle\gamma_{4,l}^{(r)},\mathbf{z}_t\rangle)^2\right]\gamma_{2,l}^{(r)} \\
&= \gamma_{2,l}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\gamma_{2,l}^{(r)}\,. \qquad \text{(C.46)}
\end{aligned}
$$

and

$$
\begin{aligned}
-2(\mathbf{u}_t-\overline{\mathbf{g}}_{l,t}^{(-r)})'\overline{\overline{\boldsymbol{\Omega}}}_t\overline{\mathbf{g}}_{l,t}^{(r)} &= -2(\mathbf{u}_t-\mathbf{g}_{l,t}^{(-r)})'\overline{\overline{\boldsymbol{\Omega}}}_t\langle\gamma_{4,l}^{(r)},\mathbf{z}_t\rangle\left(\gamma_{3,l}^{(r)}\otimes\mathbf{I}_J\otimes\gamma_{1,l}^{(r)}\right)\gamma_{2,l}^{(r)} \\
&= -2\langle\gamma_{4,l}^{(r)},\mathbf{z}_t\rangle(\mathbf{u}_t-\overline{\mathbf{g}}_{l,t}^{(-r)})'\overline{\overline{\boldsymbol{\Omega}}}_t\left(\gamma_{3,l}^{(r)}\otimes\mathbf{I}_J\otimes\gamma_{1,l}^{(r)}\right)\gamma_{2,l}^{(r)} \\
&= -2\overline{\boldsymbol{\mu}}_{2,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\gamma_{2,l}^{(r)}\,. \qquad \text{(C.47)}
\end{aligned}
$$

Now, by applying Bayes' rule and plugging eq. (C.46) and eq. (C.47) into eq. (C.37) we get:

$$
\begin{aligned}
p(\gamma_{2,l}^{(r)}|-) &\propto L(\boldsymbol{\mathcal{X}},\mathcal{D},\boldsymbol{\Omega},\mathbf{s}|\boldsymbol{\theta})\pi(\gamma_{2,l}^{(r)}|\mathbf{w}_{2,:},\boldsymbol{\phi},\tau) \\
&\propto \prod_{t\in\mathcal{T}_l}\exp\left\{-\frac{1}{2}\left[\gamma_{2,l}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\gamma_{2,l}^{(r)}-2\overline{\boldsymbol{\mu}}_{2,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\gamma_{2,l}^{(r)}\right]\right\} \\
&\cdot\exp\left\{-\frac{1}{2}\left[\gamma_{2,l}^{(r)'}\left(\overline{\boldsymbol{\Lambda}}_{2,l}^{r}\right)^{-1}\gamma_{2,l}^{(r)}-2\overline{\zeta}_{2,l}^{r'}\left(\overline{\boldsymbol{\Lambda}}_{2,l}^{r}\right)^{-1}\gamma_{2,l}^{(r)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\sum_{t\in\mathcal{T}_l}\left(\gamma_{2,l}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\gamma_{2,l}^{(r)}-2\overline{\boldsymbol{\mu}}_{2,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\gamma_{2,l}^{(r)}\right)\right.\right. \\
&+\left.\left.\left(\gamma_{2,l}^{(r)'}\left(\overline{\boldsymbol{\Lambda}}_{2,l}^{r}\right)^{-1}\gamma_{2,l}^{(r)}-2\overline{\zeta}_{2,l}^{r'}\left(\overline{\boldsymbol{\Lambda}}_{2,l}^{r}\right)^{-1}\gamma_{2,l}^{(r)}\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\gamma_{2,l}^{(r)'}\left(\sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\right)\gamma_{2,l}^{(r)}-2\left(\sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}_{2,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\right)\gamma_{2,l}^{(r)}\right.\right. \\
&+\left.\left.\gamma_{2,l}^{(r)'}\left(\overline{\boldsymbol{\Lambda}}_{2,l}^{r}\right)^{-1}\gamma_{2,l}^{(r)}-2\overline{\zeta}_{2,l}^{r'}\left(\overline{\boldsymbol{\Lambda}}_{2,l}^{r}\right)^{-1}\gamma_{2,l}^{(r)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\gamma_{2,l}^{(r)'}\left(\left(\overline{\boldsymbol{\Lambda}}_{2,l}^{r}\right)^{-1}+\sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\right)\gamma_{2,l}^{(r)}\right.\right. \\
&-\left.\left.2\left(\overline{\zeta}_{2,l}^{r'}\left(\overline{\boldsymbol{\Lambda}}_{2,l}^{r}\right)^{-1}+\sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}_{2,l,t}^{(r)'}\left(\overline{\overline{\boldsymbol{\Sigma}}}_{2,l,t}^{(r)}\right)^{-1}\right)\gamma_{2,l}^{(r)}\right]\right\}. \qquad \text{(C.48)}
\end{aligned}
$$

This is the kernel of a multivariate normal distribution with parameters:

$$\tilde{\boldsymbol{\Lambda}}^{r}_{2,l} = \left[\left(\tau\phi_r w_{2,r}\mathbf{I}_J\right)^{-1} + \sum_{t\in\mathcal{T}_l} \left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{2,l,t}\right)^{-1} \right]^{-1} \tag{C.49a}$$

$$\tilde{\zeta}^{r}_{2,l} = \tilde{\boldsymbol{\Lambda}}^{r\prime}_{2,l} \left[\overline{\zeta}^{r\prime}_{2,l} \left(\tau\phi_r w_{2,r}\mathbf{I}_J\right)^{-1} + \sum_{t\in\mathcal{T}_l} \overline{\boldsymbol{\mu}}^{(r)\prime}_{2,l,t} \left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{2,l,t}\right)^{-1} \right]^{\prime} . \tag{C.49b}$$

Full conditional distribution of $\gamma^{(r)}_{3,l}$

The full conditional distribution of $\gamma^{(r)}_{3,l}$ is given by:

$$p(\gamma^{(r)}_{3,l} | \boldsymbol{\mathcal{X}}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s}, \gamma^{(r)}_{1,l}, \gamma^{(r)}_{2,l}, \gamma^{(r)}_{4,l}, \mathcal{G}^{(-r)}_{l}, w_{3,r}, \phi_r, \tau) \sim \mathcal{N}_K(\tilde{\zeta}^{r}_{3,l}, \tilde{\boldsymbol{\Lambda}}^{r}_{3,l}) \tag{C.50}$$

where:

$$\tilde{\boldsymbol{\Lambda}}^{r}_{3,l} = \left[\left(\tau\phi_r w_{3,r}\mathbf{I}_K\right)^{-1} + \sum_{t\in\mathcal{T}_l} \left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1} \right]^{-1} \tag{C.51a}$$

$$\tilde{\zeta}^{r}_{3,l} = \tilde{\boldsymbol{\Lambda}}^{r\prime}_{3,l} \left[\overline{\zeta}^{r\prime}_{3,l} \left(\tau\phi_r w_{3,r}\mathbf{I}_K\right)^{-1} + \sum_{t\in\mathcal{T}_l} \overline{\boldsymbol{\mu}}^{(r)\prime}_{3,l,t} \left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1} \right]^{\prime} . \tag{C.51b}$$

By exploiting the rightmost term in the equality chain in eq. (C.31), we can simplify the two addenda in eq. (C.37) as:

$$\begin{aligned}
\overline{\mathbf{g}}^{(r)\prime}_{l,t} \overline{\overline{\boldsymbol{\Omega}}}_t \overline{\mathbf{g}}^{(r)}_{l,t} &= \left(\mathbf{A}_3\gamma^{(r)}_{3,l}\right)^{\prime} \overline{\overline{\boldsymbol{\Omega}}}_t \left(\mathbf{A}_3\gamma^{(r)}_{3,l}\right) \\
&= \langle\gamma^{(r)}_{4,l}, \mathbf{z}_t\rangle \gamma^{(r)\prime}_{3,l} \left(\mathbf{I}_K \otimes \gamma^{(r)}_{2,l} \otimes \gamma^{(r)}_{1,l}\right)^{\prime} \overline{\overline{\boldsymbol{\Omega}}}_t \left(\mathbf{I}_K \otimes \gamma^{(r)}_{2,l} \otimes \gamma^{(r)}_{1,l}\right) \gamma^{(r)}_{3,l} \langle\gamma^{(r)}_{4,l}, \mathbf{z}_t\rangle \\
&= \gamma^{(r)\prime}_{3,l} \left[\left(\mathbf{I}^{\prime}_K \otimes \gamma^{(r)\prime}_{2,l} \otimes \gamma^{(r)\prime}_{1,l}\right) \overline{\overline{\boldsymbol{\Omega}}}_t \left(\mathbf{I}_K \otimes \gamma^{(r)}_{2,l} \otimes \gamma^{(r)}_{1,l}\right) (\langle\gamma^{(r)}_{4,l}, \mathbf{z}_t\rangle)^2 \right] \gamma^{(r)}_{3,l} \\
&= \gamma^{(r)\prime}_{3,l} \left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1} \gamma^{(r)}_{3,l} .
\end{aligned} \tag{C.52}$$

and

$$\begin{aligned}
-2(\mathbf{u}_t - \overline{\mathbf{g}}^{(-r)}_{l,t})^{\prime} \overline{\overline{\boldsymbol{\Omega}}}_t \overline{\mathbf{g}}^{(r)}_{l,t} &= -2(\mathbf{u}_t - \overline{\mathbf{g}}^{(-r)}_{l,t})^{\prime} \overline{\overline{\boldsymbol{\Omega}}}_t \left(\mathbf{I}_K \otimes \gamma^{(r)}_{2,l} \otimes \gamma^{(r)}_{1,l}\right) \gamma^{(r)}_{3,l} \langle\gamma^{(r)}_{4,l}, \mathbf{z}_t\rangle \\
&= -2\langle\gamma^{(r)}_{4,l}, \mathbf{z}_t\rangle (\mathbf{u}_t - \overline{\mathbf{g}}^{(-r)}_{l,t})^{\prime} \overline{\overline{\boldsymbol{\Omega}}}_t \left(\mathbf{I}_K \otimes \gamma^{(r)}_{2,l} \otimes \gamma^{(r)}_{1,l}\right) \gamma^{(r)}_{3,l} \\
&= -2\overline{\boldsymbol{\mu}}^{(r)\prime}_{3,l,t} \left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1} \gamma^{(r)}_{3,l} .
\end{aligned} \tag{C.53}$$

Now, by applying Bayes' rule and plugging eq. (C.52) and eq. (C.53) into eq. (C.37) we get:

$$\begin{aligned}
p(\gamma^{(r)}_{3,l} | -) &\propto L(\boldsymbol{\mathcal{X}}, \mathcal{D}, \boldsymbol{\Omega}, \mathbf{s} | \boldsymbol{\theta}) \pi(\gamma^{(r)}_{3,l} | \mathbf{w}_{3,:}, \boldsymbol{\phi}, \tau) \\
&\propto \prod_{t\in\mathcal{T}_l} \exp\left\{ -\frac{1}{2} \left[\gamma^{(r)\prime}_{3,l} \left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1} \gamma^{(r)}_{3,l} - 2\overline{\boldsymbol{\mu}}^{(r)\prime}_{3,l,t} \left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1} \gamma^{(r)}_{3,l} \right] \right\}
\end{aligned}$$

$$
\begin{aligned}
&\cdot \exp\left\{-\frac{1}{2}\left[\gamma^{(r)\prime}_{3,l}\left(\overline{\boldsymbol{\Lambda}}^{r}_{3,l}\right)^{-1}\gamma^{(r)}_{3,l} - 2\overline{\zeta}^{r\prime}_{3,l}\left(\overline{\boldsymbol{\Lambda}}^{r}_{3,l}\right)^{-1}\gamma^{(r)}_{3,l}\right]\right\} \\
&= \exp\Bigg\{-\frac{1}{2}\Bigg[\sum_{t\in\mathcal{T}_l}\left(\gamma^{(r)\prime}_{3,l}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1}\gamma^{(r)}_{3,l} - 2\overline{\boldsymbol{\mu}}^{(r)\prime}_{3,l,t}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1}\gamma^{(r)}_{3,l}\right) \\
&+ \left(\gamma^{(r)\prime}_{3,l}\left(\overline{\boldsymbol{\Lambda}}^{r}_{3,l}\right)^{-1}\gamma^{(r)}_{3,l} - 2\overline{\zeta}^{r\prime}_{3,l}\left(\overline{\boldsymbol{\Lambda}}^{r}_{3,l}\right)^{-1}\gamma^{(r)}_{3,l}\right)\Bigg]\Bigg\} \\
&= \exp\Bigg\{-\frac{1}{2}\Bigg[\gamma^{(r)\prime}_{3,l}\left(\sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1}\right)\gamma^{(r)}_{3,l} - 2\left(\sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}^{(r)\prime}_{3,l,t}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1}\right)\gamma^{(r)}_{3,l} \\
&+ \gamma^{(r)\prime}_{3,l}\left(\overline{\boldsymbol{\Lambda}}^{r}_{3,l}\right)^{-1}\gamma^{(r)}_{3,l} - 2\overline{\zeta}^{r\prime}_{3,l}\left(\overline{\boldsymbol{\Lambda}}^{r}_{3,l}\right)^{-1}\gamma^{(r)}_{3,l}\Bigg]\Bigg\} \\
&= \exp\Bigg\{-\frac{1}{2}\Bigg[\gamma^{(r)\prime}_{3,l}\left(\left(\overline{\boldsymbol{\Lambda}}^{r}_{3,l}\right)^{-1} + \sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1}\right)\gamma^{(r)}_{3,l} \\
&- 2\left(\overline{\zeta}^{r\prime}_{3,l}\left(\overline{\boldsymbol{\Lambda}}^{r}_{3,l}\right)^{-1} + \sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}^{(r)\prime}_{3,l,t}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1}\right)\gamma^{(r)}_{3,l}\Bigg]\Bigg\}. \qquad \text{(C.54)}
\end{aligned}
$$

This is the kernel of a multivariate normal distribution with parameters:

$$
\tilde{\boldsymbol{\Lambda}}^{r}_{3,l} = \left[\left(\tau\phi_r w_{3,r}\mathbf{I}_K\right)^{-1} + \sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1}\right]^{-1} \qquad \text{(C.55a)}
$$

$$
\tilde{\zeta}^{r}_{3,l} = \tilde{\boldsymbol{\Lambda}}^{r\prime}_{3,l}\left[\overline{\zeta}^{r\prime}_{3,l}\left(\tau\phi_r w_{3,r}\mathbf{I}_K\right)^{-1} + \sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}^{(r)\prime}_{3,l,t}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{3,l,t}\right)^{-1}\right]^{\prime}. \qquad \text{(C.55b)}
$$

Full conditional distribution of $\gamma^{(r)}_{4,l}$

The full conditional distribution of $\gamma^{(r)}_{4,l}$ is given by:

$$
p(\gamma^{(r)}_{4,l}|\boldsymbol{\mathcal{X}},\mathcal{D},\boldsymbol{\Omega},\mathbf{s},\gamma^{(r)}_{1,l},\gamma^{(r)}_{2,l},\gamma^{(r)}_{3,l},\mathcal{G}^{(-r)}_{l},w_{4,r},\phi_r,\tau) \sim \mathcal{N}_Q(\tilde{\zeta}^{r}_{4,l},\tilde{\boldsymbol{\Lambda}}^{r}_{4,l}) \qquad \text{(C.56)}
$$

where:

$$
\tilde{\boldsymbol{\Lambda}}^{r}_{4,l} = \left[\left(\tau\phi_r w_{4,r}\mathbf{I}_Q\right)^{-1} + \sum_{t\in\mathcal{T}_l}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{4,l,t}\right)^{-1}\right]^{-1} \qquad \text{(C.57a)}
$$

$$
\tilde{\zeta}^{r}_{4,l} = \tilde{\boldsymbol{\Lambda}}^{r\prime}_{4,l}\left[\overline{\zeta}^{r\prime}_{4,l}\left(\tau\phi_r w_{4,r}\mathbf{I}_Q\right)^{-1} + \sum_{t\in\mathcal{T}_l}\overline{\boldsymbol{\mu}}^{(r)\prime}_{4,l,t}\left(\overline{\overline{\boldsymbol{\Sigma}}}^{(r)}_{4,l,t}\right)^{-1}\right]^{\prime}. \qquad \text{(C.57b)}
$$

By exploiting the central term in the equality chain in eq. (C.25), we can simplify the two addenda in eq. (C.37) as:

$$
\overline{\mathbf{g}}^{(r)\prime}_{l,t}\overline{\overline{\boldsymbol{\Omega}}}_t\overline{\mathbf{g}}^{(r)}_{l,t} = \left(\mathbf{A}_4\gamma^{(r)}_{4,l}\right)^{\prime}\overline{\overline{\boldsymbol{\Omega}}}_t\left(\mathbf{A}_4\gamma^{(r)}_{4,l}\right)
$$

$$
\begin{aligned}
&= \gamma_{4,l}^{(r)'} \mathbf{z}_t \operatorname{vec}\left(\gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)}\right) \overline{\overline{\mathbf{\Omega}}}_t \operatorname{vec}\left(\gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)}\right) \mathbf{z}_t' \gamma_{4,l}^{(r)} \\
&= \gamma_{4,l}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \gamma_{4,l}^{(r)} .
\end{aligned}
\tag{C.58}
$$

and

$$
\begin{aligned}
-2(\mathbf{u}_t - \overline{\mathbf{g}}_{l,t}^{(-r)})'\overline{\overline{\mathbf{\Omega}}}_t \overline{\mathbf{g}}_{l,t}^{(r)} &= -2(\mathbf{u}_t - \overline{\mathbf{g}}_{l,t}^{(-r)})'\overline{\overline{\mathbf{\Omega}}}_t \operatorname{vec}\left(\gamma_{1,l}^{(r)} \circ \gamma_{2,l}^{(r)} \circ \gamma_{3,l}^{(r)}\right) \mathbf{z}_t' \gamma_{4,l}^{(r)} \\
&= -2\overline{\boldsymbol{\mu}}_{4,l,t}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \gamma_{4,l}^{(r)} .
\end{aligned}
\tag{C.59}
$$

Now, by applying Bayes' rule and plugging eq. (C.58) and eq. (C.59) into eq. (C.37) we get:

$$
\begin{aligned}
p(\gamma_{4,l}^{(r)}|-) &\propto L(\boldsymbol{\mathcal{X}}, \mathcal{D}, \mathbf{\Omega}, \mathbf{s}|\boldsymbol{\theta})\pi(\gamma_{4,l}^{(r)}|\mathbf{w}_{4,:}, \boldsymbol{\phi}, \tau) \\
&\propto \prod_{t\in\mathcal{T}_l} \exp\left\{ -\frac{1}{2}\left[\gamma_{4,l}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \gamma_{4,l}^{(r)} - 2\overline{\boldsymbol{\mu}}_{4,l,t}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \gamma_{4,l}^{(r)} \right] \right\} \\
&\cdot \exp\left\{ -\frac{1}{2}\left[\gamma_{4,l}^{(r)'} \left(\overline{\mathbf{\Lambda}}_{4,l}^{r}\right)^{-1} \gamma_{4,l}^{(r)} - 2\overline{\zeta}_{4,l}^{r'} \left(\overline{\mathbf{\Lambda}}_{4,l}^{r}\right)^{-1} \gamma_{4,l}^{(r)} \right] \right\} \\
&= \exp\Bigg\{ -\frac{1}{2}\Bigg[\sum_{t\in\mathcal{T}_l} \left(\gamma_{3,l}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \gamma_{4,l}^{(r)} - 2\overline{\boldsymbol{\mu}}_{4,l,t}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \gamma_{4,l}^{(r)} \right) \\
&+ \left(\gamma_{4,l}^{(r)'} \left(\overline{\mathbf{\Lambda}}_{4,l}^{r}\right)^{-1} \gamma_{4,l}^{(r)} - 2\overline{\zeta}_{4,l}^{r'} \left(\overline{\mathbf{\Lambda}}_{4,l}^{r}\right)^{-1} \gamma_{4,l}^{(r)} \right) \Bigg] \Bigg\} \\
&= \exp\Bigg\{ -\frac{1}{2}\Bigg[\gamma_{4,l}^{(r)'} \left(\sum_{t\in\mathcal{T}_l} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \right) \gamma_{4,l}^{(r)} - 2\left(\sum_{t\in\mathcal{T}_l} \overline{\boldsymbol{\mu}}_{4,l,t}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \right) \gamma_{4,l}^{(r)} \\
&+ \gamma_{4,l}^{(r)'} \left(\overline{\mathbf{\Lambda}}_{4,l}^{r}\right)^{-1} \gamma_{4,l}^{(r)} - 2\overline{\zeta}_{4,l}^{r'} \left(\overline{\mathbf{\Lambda}}_{4,l}^{r}\right)^{-1} \gamma_{4,l}^{(r)} \Bigg] \Bigg\} \\
&= \exp\Bigg\{ -\frac{1}{2}\Bigg[\gamma_{4,l}^{(r)'} \left(\left(\overline{\mathbf{\Lambda}}_{4,l}^{r}\right)^{-1} + \sum_{t\in\mathcal{T}_l} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \right) \gamma_{4,l}^{(r)} \\
&- 2\left(\overline{\zeta}_{4,l}^{r'} \left(\overline{\mathbf{\Lambda}}_{4,l}^{r}\right)^{-1} + \sum_{t\in\mathcal{T}_l} \overline{\boldsymbol{\mu}}_{4,l,t}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \right) \gamma_{4,l}^{(r)} \Bigg] \Bigg\} .
\end{aligned}
\tag{C.60}
$$

This is the kernel of a multivariate normal distribution with parameters:

$$
\tilde{\mathbf{\Lambda}}_{4,l}^{r} = \left[\left(\tau\phi_r w_{4,r} \mathbf{I}_Q\right)^{-1} + \sum_{t\in\mathcal{T}_l} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \right]^{-1}
\tag{C.61a}
$$

$$
\tilde{\zeta}_{4,l}^{r} = \tilde{\mathbf{\Lambda}}_{4,l}^{r'} \left[\overline{\zeta}_{4,l}^{r'} \left(\tau\phi_r w_{4,r} \mathbf{I}_Q\right)^{-1} + \sum_{t\in\mathcal{T}_l} \overline{\boldsymbol{\mu}}_{4,l,t}^{(r)'} \left(\overline{\overline{\mathbf{\Sigma}}}_{4,l,t}^{(r)}\right)^{-1} \right]' .
\tag{C.61b}
$$

C.3.7 Full conditional distribution of $\omega_{ijk,t}$

The full conditional distribution for the latent variable $\omega_{ijk,t}$ for every $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, K$ and $t = 1, \dots, T$:

$$p(\omega_{ijk,t}|x_{ijk,t}, s_t, \mathcal{G}_{s_t}) \sim PG(1, \, \mathbf{z}_t' \mathbf{g}_{ijk,s_t}) \, . \tag{C.62}$$

To shorten the notation, define $\psi_{ijk,t} = \mathbf{z}_t' \mathbf{g}_{ijk,s_t}$. The full conditional is derived by integrating out the latent allocation variable $d_{ijk,t}$, as follows:

$$\begin{aligned}
&p(\omega_{ijk,t}|x_{ijk,t}, s_t, \mathcal{G}_{s_t}) \\
&= \int_D \int_\rho p(\omega_{ijk,t}, d_{ijk,t}|x_{ijk,t}, s_t, \mathcal{G}_{s_t}, \rho_{s_t}) p(\rho_{s_t}) \, \mathrm{d}\rho_{s_t} \mathrm{d}d_{ijk,t} \\
&= \int_D \int_\rho \frac{p(x_{ijk,t}, d_{ij,t}|\omega_{ijk,t}, s_t, \mathcal{G}_{s_t}, \rho_{s_t}) p(\omega_{ijk,t}) p(\rho_{s_t})}{\int_\Omega p(x_{ijk,t}, \omega_{ijk,t}, d_{ijk,t}|s_t, \mathcal{G}_{s_t}, \rho_{s_t}) \, \mathrm{d}\omega_{ijk,t}} \, \mathrm{d}\rho_{s_t} \mathrm{d}d_{ijk,t} \\
&= \int_D \int_\rho \frac{p(x_{ijk,t}, \omega_{ijk,t}, d_{ijk,t}|s_t, \mathcal{G}_{s_t}, \rho_{s_t})}{p(x_{ijk,t}, d_{ijk,t}|s_t, \mathcal{G}_{s_t}, \rho_{s_t})} p(\rho_{s_t}) \, \mathrm{d}\rho_{s_t} \mathrm{d}d_{ijk,t} \\
&= \int_D \int_\rho \frac{\left(\rho_{s_t} \delta_{\{0\}}(x_{ijk,t})\right)^{d_{ijk,t}} \left(\frac{1-\rho_{s_t}}{2}\right)^{d_{ijk,t}} \exp\left\{-\frac{\omega_{ijk,t}}{2} \psi_{ijk,t}^2 + \kappa_{ijk,t} \psi_{ijk,t}\right\} p(\omega_{ijk,t}) p(\rho_{s_t})}{\left(\rho_{s_t} \delta_{\{0\}}(x_{ijk,t})\right)^{d_{ijk,t}} \left(\frac{1-\rho_{s_t}}{2}\right)^{d_{ijk,t}} \left(\exp\{\psi_{ijk,t} x_{ijk,t}\} / (1 + \exp\{\psi_{ijk,t}\})\right)^{1-d_{ijk,t}}} \, \mathrm{d}\rho_{s_t} \mathrm{d}d_{ijk,t} \\
&= \int_D \int_\rho \exp\{\kappa_{ijk,t}^{(s_t)} \psi_{ijk,t}\} \frac{\exp\{\psi_{ijk,t} x_{ijk,t} (1 - d_{ijk,t})\}}{(1 + \exp\{\psi_{ijk,t}\})^{1-d_{ijk,t}}} \exp\left\{-\frac{\omega_{ijk,t}}{2} \psi_{ijk,t}^2\right\} p(\omega_{ijk,t}) p(\rho_{s_t}) \, \mathrm{d}\rho_{s_t} \mathrm{d}d_{ijk,t} \\
&= \int_D \int_\rho \left[\frac{1 + \exp\{\psi_{ijk,t}\}}{\exp\{\psi_{ijk,t} x_{ijk,t}\}} \cdot \frac{\exp\{\psi_{ijk,t} x_{ijk,t}\}}{\exp\{\psi_{ijk,t}/2\}}\right]^{1-d_{ijk,t}} \left[\exp\{-\psi_{ijk,t}^2 \omega_{ijk,t}/2\} p(\omega_{ijk,t})\right] p(\rho_{s_t}) \, \mathrm{d}\rho_{s_t} \mathrm{d}d_{ijk,t} \\
&= \left(1 + \frac{1 + \exp\{\psi_{ijk,t}\}}{\exp\{\psi_{ijk,t}/2\}}\right) \left[\exp\{-\psi_{ijk,t}^2 \omega_{ijk,t}/2\} p(\omega_{ijk,t})\right] \\
&\propto \exp\{-\psi_{ijk,t}^2 \omega_{ijk,t}/2\} p(\omega_{ijk,t}) \, .
\end{aligned} \tag{C.63}$$

Since $p(\omega_{ijk,t}) \sim PG(1, 0)$, by Theorem 1 in Polson et al. (2013) the result follows.

C.3.8 Full conditional distribution of $d_{ijk,t}$

The full conditional posterior probabilities for the latent allocation variables $d_{ijk,t}$, which select the component of the mixture in eq. (3.3), for each $t = 1, \dots, T$ and for every $i = 1, \dots, I$, $j = 1, \dots, J$ and $k = 1, \dots, K$, are given by:

$$p(d_{ijk,t} = 1|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t}) = \frac{\tilde{p}(d_{ijk,t} = 1|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t})}{\tilde{p}(d_{ijk,t} = 1|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t}) + \tilde{p}(d_{ijk,t} = 0|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t})} \tag{C.64a}$$

$$p(d_{ij,t} = 0|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t}) = \frac{\tilde{p}(d_{ijk,t} = 0|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t})}{\tilde{p}(d_{ijk,t} = 1|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t}) + \tilde{p}(d_{ijk,t} = 0|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t})} \, . \tag{C.64b}$$

The un-normalised posterior probabilities are given by:

$$\tilde{p}(d_{ijk,t} = 1|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t}) = \rho_{s_t} \delta_{\{0\}}(x_{ijk,t}) \tag{C.65a}$$

$$\tilde{p}(d_{ijk,t} = 0|\boldsymbol{\mathcal{X}}, \mathbf{s}, \mathcal{G}_{s_t}, \boldsymbol{\rho}_{s_t}) = (1 - \rho_{s_t}) \frac{\exp\left\{(\mathbf{z}_t' \mathbf{g}_{ijk,s_t}) x_{ijk,t}\right\}}{1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,s_t}\}} \, . \tag{C.65b}$$

We have obtained the result starting from eq. (3.18) after having integrated out the latent variables $\boldsymbol{\Omega}$, as follows:

$$\begin{aligned}
\tilde{p}(d_{ijk,t}|\boldsymbol{\mathcal{X}},\mathbf{s},\mathcal{G}_{s_t},\boldsymbol{\rho}_{s_t}) &\propto p(\boldsymbol{\mathcal{X}},\mathbf{s}|\mathcal{G}_{s_t},\boldsymbol{\rho}_{s_t},d_{ijk,t})\pi(d_{ijk,t}) \\
&= \rho_{s_t}^{d_{ijk,t}}\delta_{\{0\}}(x_{ijk,t})^{d_{ijk,t}}(1-\rho_{s_t})^{1-d_{ijk,t}}\frac{(\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,s_t}\})^{x_{ijk,t}(1-d_{ijk,t})}}{(1+\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,s_t}\})^{(1-d_{ijk,t})}} \\
&= \left[\rho_{s_t}\delta_{\{0\}}(x_{ijk,t})\right]^{d_{ijk,t}}\left[(1-\rho_{s_t})\frac{(\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,s_t}\})^{x_{ijk,t}}}{1+\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,s_t}\}}\right]^{1-d_{ijk,t}} .
\end{aligned} \tag{C.66}$$

C.3.9 Full conditional distribution of ρ_l

For each regime $l = 1, \dots, L$, the full conditional distribution for the mixing probability ρ_l of the observation model in eq (3.2) is given by:

$$p(\rho_l|\boldsymbol{\mathcal{X}},\mathcal{D},\mathbf{s}) = p(\rho_l|\mathcal{D},\mathbf{s}) \sim \mathcal{B}e(\tilde{a}_l,\tilde{b}_l)\,, \tag{C.67}$$

with:

$$\tilde{a}_l = N_1^l + \overline{a}_l^{\rho} \tag{C.68a}$$
$$\tilde{b}_l = N_0^l + \overline{b}_l^{\rho}\,. \tag{C.68b}$$

We get this result starting from eq. (3.18) and integrating out the latent variables $\boldsymbol{\Omega}$, as follows:

$$\begin{aligned}
p(\rho_l|\boldsymbol{\mathcal{X}},\mathcal{D},\mathbf{s}) &\propto \pi(\rho_l)\int_G L(\boldsymbol{\mathcal{X}},\mathcal{D},\mathbf{s}|\rho_l,\mathcal{G}_l)p(\mathcal{G}_l)\,\mathrm{d}\mathcal{G}_l \\
&\propto \Bigg[\int_G \prod_{t\in\mathcal{T}_l}\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\rho_l^{d_{ijk,t}}\cdot\left[\delta_{\{0\}}(x_{ijk,t})\right]^{d_{ijk,t}}\cdot(1-\rho_l)^{1-d_{ijk,t}} \\
&\quad\cdot\frac{\left(\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,l}\}\right)^{x_{ijk,t}(1-d_{ijk,t})}}{\left(1+\exp\{\mathbf{z}_t'\mathbf{g}_{ijk,l}\}\right)^{(1-d_{ijk,t})}}\,\mathrm{d}\mathcal{G}_l\Bigg]\cdot\rho_l^{\overline{a}_l^{\rho}-1}(1-\rho_l)^{\overline{b}_l^{\rho}-1} \\
&\propto \left[\prod_{t\in\mathcal{T}_l}\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\rho_l^{d_{ijk,t}}(1-\rho_l)^{1-d_{ijk,t}}\right]\rho_l^{\overline{a}_l^{\rho}-1}(1-\rho_l)^{\overline{b}_l^{\rho}-1} \\
&= \rho_l^{N_1^l}(1-\rho_l)^{N_0^l}\rho_l^{\overline{a}_l^{\rho}-1}(1-\rho_l)^{\overline{b}_l^{\rho}-1} \\
&= \rho_l^{N_1^l+\overline{a}_l^{\rho}-1}(1-\rho_l)^{N_0^l+\overline{b}_l^{\rho}-1}\,,
\end{aligned} \tag{C.69}$$

where we have defined the counting variables, for every $l = 1, \dots, L$:

$$N_1^l = \sum_{t\in\mathcal{T}_l}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbb{1}\{d_{ijk,t}=1\} \tag{C.70a}$$

$$N_0^l = \sum_{t\in\mathcal{T}_l}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbb{1}\{d_{ijk,t}=0\}\,. \tag{C.70b}$$

C.3.10 Full conditional distribution of $\boldsymbol{\mu}_l$

The full conditional distribution for the intercept term of the second equation of the model is given by:

$$p(\boldsymbol{\mu}_l|\mathbf{y},\mathbf{s},\boldsymbol{\Sigma}_l) \sim \mathcal{N}_M(\tilde{\boldsymbol{\mu}}_l,\tilde{\boldsymbol{\Upsilon}}_l)\,, \tag{C.71}$$

with:

$$\tilde{\boldsymbol{\mu}}_l = \tilde{\boldsymbol{\Upsilon}}_l' \left(\bar{\boldsymbol{\mu}}_l'\bar{\boldsymbol{\Upsilon}}_l^{-1} + \sum_{t\in\mathcal{T}_l} y_t'\boldsymbol{\Sigma}_l^{-1}\right)'\,, \tag{C.72a}$$

$$\tilde{\boldsymbol{\Upsilon}}_l = \left[T_l\boldsymbol{\Sigma}_l^{-1} + \bar{\boldsymbol{\Upsilon}}_l^{-1}\right]^{-1}\,, \tag{C.72b}$$

for each regime $l = 1,\dots,L$. We have derived the updated hyper-parameters from:

$$\begin{aligned}
p(\boldsymbol{\mu}_l|\mathbf{y},\mathbf{s},\boldsymbol{\Sigma}_l) &\propto \pi(\boldsymbol{\mu}_l)p(\mathbf{y}|\mathbf{s},\boldsymbol{\Sigma}_l,\boldsymbol{\mu}_l)\\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_l-\bar{\boldsymbol{\mu}}_l)'\bar{\boldsymbol{\Upsilon}}_l^{-1}(\boldsymbol{\mu}_l-\bar{\boldsymbol{\mu}}_l)\right\}\prod_{t\in\mathcal{T}_l}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}(\mathbf{y}_t-\boldsymbol{\mu}_l)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_l'\bar{\boldsymbol{\Upsilon}}_l^{-1}\boldsymbol{\mu}_l - 2\bar{\boldsymbol{\mu}}_l'\bar{\boldsymbol{\Upsilon}}_l^{-1}\boldsymbol{\mu}_l + \sum_{t\in\mathcal{T}_l}\boldsymbol{\mu}_l'\boldsymbol{\Sigma}_l^{-1}\boldsymbol{\mu}_l - 2\mathbf{y}_t'\boldsymbol{\Sigma}_l^{-1}\boldsymbol{\mu}_l\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_l'\left(T_l\boldsymbol{\Sigma}_l^{-1}+\bar{\boldsymbol{\Upsilon}}_l^{-1}\right)\boldsymbol{\mu}_l - 2\left(\sum_{t\in\mathcal{T}_l}\mathbf{y}_t'\boldsymbol{\Sigma}_l^{-1}+\bar{\boldsymbol{\mu}}_l'\bar{\boldsymbol{\Upsilon}}_l^{-1}\right)\boldsymbol{\mu}_l\right]\right\}\,.
\end{aligned} \tag{C.73}$$

C.3.11 Full conditional distribution of $\boldsymbol{\Sigma}_l$

The full conditional distribution for the covariance of the error term of the second equation of the model is given by:

$$p(\boldsymbol{\Sigma}_l|\mathbf{y},\mathbf{s},\boldsymbol{\mu}_l) \sim \mathcal{IW}_M(\tilde{\nu}_l,\tilde{\boldsymbol{\Psi}}_l)\,, \tag{C.74}$$

with:

$$\tilde{\nu}_l = \bar{\nu}_l + T_l\,, \tag{C.75a}$$

$$\tilde{\boldsymbol{\Psi}}_l = \bar{\boldsymbol{\Psi}}_l + \sum_{t\in\mathcal{T}_l}(\mathbf{y}_t-\boldsymbol{\mu}_l)(\mathbf{y}_t-\boldsymbol{\mu}_l)'\,, \tag{C.75b}$$

for each regime $l = 1,\dots,L$. We have derived the updated hyper-parameters from:

$$\begin{aligned}
p(\boldsymbol{\Sigma}_l|\mathbf{y},\mathbf{s},\boldsymbol{\mu}_l) &\propto \pi(\boldsymbol{\Sigma}_l)p(\mathbf{y}|\mathbf{s},\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)\\
&\propto |\boldsymbol{\Sigma}_l|^{-\frac{\bar{\nu}_l+m-1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\bar{\boldsymbol{\Psi}}_l\boldsymbol{\Sigma}_l^{-1}\right)\right\}\prod_{t\in\mathcal{T}_l}|\boldsymbol{\Sigma}_l|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}(\mathbf{y}_t-\boldsymbol{\mu}_l)\right\}\\
&= |\boldsymbol{\Sigma}_l|^{-\frac{\bar{\nu}_l+m-1+T_l}{2}}\exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\bar{\boldsymbol{\Psi}}_l\boldsymbol{\Sigma}_l^{-1}\right)+\sum_{t\in\mathcal{T}_l}(\mathbf{y}_t-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}(\mathbf{y}_t-\boldsymbol{\mu}_l)\right]\right\}\\
&= |\boldsymbol{\Sigma}_l|^{-\frac{\bar{\nu}_l+m-1+T_l}{2}}\exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(\bar{\boldsymbol{\Psi}}_l\boldsymbol{\Sigma}_l^{-1}\right)+\operatorname{tr}\left(\sum_{t\in\mathcal{T}_l}(\mathbf{y}_t-\boldsymbol{\mu}_l)(\mathbf{y}_t-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}\right)\right]\right\}
\end{aligned}$$

$$= |\mathbf{\Sigma}_l|^{-\frac{\bar{v}_l+m-1+T_l}{2}} \exp\left\{-\frac{1}{2}\left[\mathrm{tr}\left(\left(\overline{\mathbf{\Psi}}_l + \sum_{t\in\mathcal{T}_l}(\mathbf{y}_t-\boldsymbol{\mu}_l)(\mathbf{y}_t-\boldsymbol{\mu}_l)'\right)\mathbf{\Sigma}_l^{-1}\right)\right]\right\}, \tag{C.76}$$

where we have used the property of linearity of the trace operator.

C.3.12 Full conditional distribution of $\xi_{l,:}$

The full conditional distribution of each row $l = 1,\dots,L$ of the transition matrix of the hidden Markov chain, under the assumption that the initial distribution of the state s_t is independent from the transition matrix Ξ, that is $p(s_0|\Xi) = p(s_0)$, is given by:

$$p(\boldsymbol{\xi}_{l,:}|\mathbf{s}) \sim \mathcal{D}ir(\tilde{\mathbf{c}}), \tag{C.77}$$

where:

$$\tilde{\mathbf{c}} = \left(\bar{c}_1 + N_{l,1}(\mathbf{s}),\dots,\bar{c}_L + N_{l,L}(\mathbf{s})\right). \tag{C.78}$$

It can be derived from:

$$\begin{aligned} p(\boldsymbol{\xi}_{l,:}|\mathbf{s}) &\propto \pi(\boldsymbol{\xi}_{l,:})p(\mathbf{s}|\boldsymbol{\xi}_{l,:}) \\ &\propto \prod_{k=1}^{L}\xi_{l,k}^{\bar{c}_k-1}\prod_{g=1}^{L}\prod_{k=1}^{L}\xi_{g,k}^{N_{g,k}(\mathbf{s})}p(s_0|\Xi) \\ &\propto \prod_{k=1}^{L}\xi_{l,k}^{\bar{c}_k-1}\prod_{k=1}^{L}\xi_{l,k}^{N_{l,k}(\mathbf{s})}p(s_0|\Xi) \\ &= \prod_{k=1}^{L}\xi_{l,k}^{\bar{c}_k+N_{l,k}(\mathbf{s})-1}p(s_0|\Xi). \end{aligned} \tag{C.79}$$

Concerning the notation, we denoted the collection of hidden states up to time t by $\mathbf{s}^t = (s_0,\dots,s_t)$ and we used $N_{i,j}(\mathbf{s}) = \sum_t \mathbb{1}(s_{t-1}=i)\mathbb{1}(s_t=j)$ for counting the number of transitions from state i to state j up to time T. Under the assumption $p(s_0|\Xi) = p(s_0)$, we obtain the full conditional posterior in eq. (C.77). By contrast, if the initial distribution of s_0 depends on the transition matrix (for example, when it coincides with the ergodic distribution $\boldsymbol{\eta}^*(\Xi)$), we have:

$$p(\boldsymbol{\xi}_{l,:}|\mathbf{s}) \propto g_l(\boldsymbol{\xi}_{l,:})\boldsymbol{\eta}^*(\Xi), \tag{C.80}$$

where $g_l(\xi_{l,:})$ is the kernel of the Dirichlet distribution in eq. (C.79). We can sample from it via a Metropolis Hastings step, either for a single or for multiple rows of the transition matrix, using $g_l(\boldsymbol{\xi}_{l,:})$ as proposal for row l. See Frühwirth-Schnatter (2006) for further details.

C.3.13 Full conditional distribution of s_t

For sampling the trajectory $\mathbf{s} = (s_1,\dots,s_T)$, we can adopt two approaches: (i) update s_t for each $t = 1,\dots,T$ using a single-move Gibbs sampler step. This implies sampling each state s_t from its posterior distribution conditioning on all the other states. (ii) update the whole path $\mathbf{s}$ from the full joint conditional distribution in a multi-move Gibbs sampler step, also called the Forward-Filtering-Backward-Sampling (FFBS) algorithm (see Frühwirth-Schnatter (2006)).

Define $\mathbf{s}_{-t} = (s_0,\dots,s_{t-1},s_{t+1},\dots,s_T)'$. Since the hidden chain is assumed to be first order Markov, we can derive the full conditional distribution for the each state s_t:

$$p(s_t|\mathbf{s}_{-t},\boldsymbol{\mathcal{X}},\mathbf{y},\mathcal{D},\mathbf{\Omega},\mathcal{G},\boldsymbol{\mu},\mathbf{\Sigma},\Xi,\boldsymbol{\rho},\mathbf{W},\boldsymbol{\phi},\tau) = p(s_t|s_{t-1},s_{t+1},\mathcal{X}_t,\mathbf{y}_t,\mathcal{D},\mathcal{G},\boldsymbol{\mu},\mathbf{\Sigma},\Xi,\boldsymbol{\rho}). \tag{C.81}$$

Staring from the complete data likelihood in eq. (3.18) for a given time t and integrating out the latent variables $(\mathcal{D}, \boldsymbol{\Omega}_t)$, we obtain the un-normalised posterior probability of state l at time t for $l \in \{1, \ldots, L\}$:

$$p(s_t = l | s_{t-1} = u,\, s_{t+1} = v,\, \mathcal{X}_t, \mathbf{y}_t, \boldsymbol{\rho}, \mathcal{G}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Xi) \propto q^t_{l,uv}\,, \tag{C.82}$$

where $q^t_{l,uv}$ is given by:

$$\begin{aligned} q^t_{l,uv} = & \prod_{i=1}^{I} \prod_{j=1}^{J} \left[(1-\rho_l) \frac{\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}}{1+\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}} \right]^{x_{ij,t}} \left[\rho_l + (1-\rho_l) \frac{1}{1+\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}} \right]^{1-x_{ij,t}} \\ & \cdot |\boldsymbol{\Sigma}_l|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y}_t - \boldsymbol{\mu}_l)' \boldsymbol{\Sigma}_l^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_l) \right\} \\ & \cdot \left(\prod_{g=1}^{L} \xi_{g,l}^{\mathbb{1}\{s_{t-1}=u\}} \right) \left(\prod_{k=1}^{L} \xi_{l,k}^{\mathbb{1}\{s_{t+1}=v\}} \right). \end{aligned} \tag{C.83}$$

By normalizing one gets:

$$p(s_t = l | s_{t-1} = u,\, s_{t+1} = v,\, \mathcal{X}_t, \mathbf{y}_t, \boldsymbol{\rho}, \mathcal{G}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Xi) = \frac{q^t_{l,uv}}{\sum_{k=1}^{L} q^t_{k,uv}} \qquad \forall\, l\,. \tag{C.84}$$

Combining together all possible L values of the state variable, we can recognise that the posterior distribution of the state latent variable at time t follows a categorical distribution with probability vector $\tilde{\mathbf{p}}^t_{uv} = (\tilde{p}^t_{1,uv}, \ldots, \tilde{p}^t_{L,uv})'$:

$$p(s_t | s_{t-1} = u,\, s_{t+1} = v,\, \mathcal{X}_t, \mathbf{y}_t, \boldsymbol{\rho}, \mathcal{G}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Xi) \propto \prod_{l=1}^{L} (q^t_{l,uv})^{\mathbb{1}\{s_t=l\}}\,. \tag{C.85}$$

If we consider conditioning on (s_{t-1}, s_{t+1}) instead of on the specific couple $(s_{t-1} = u,\, s_{t+1} = v)$, we get an un-normalised posterior probability (denoted q^t_l) similar to eq. (C.84), but without the indicator variables. The result in eq. (C.85) thus translates in:

$$p(s_t = l | s_{t-1}, s_{t+1}, \mathcal{X}_t, \mathbf{y}_t, \mathcal{G}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Xi) = \frac{q^t_l}{\sum_{k=1}^{L} q^t_k} \propto q^t_l \qquad \forall\, l \tag{C.86}$$

$$p(s_t | s_{t-1}, s_{t+1}, \mathcal{X}_t, \mathbf{y}_t, \mathcal{G}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Xi) \propto \prod_{l=1}^{L} (q^t_l)^{\mathbb{1}\{s_t=l\}}\,. \tag{C.87}$$

By contrast, the multi-move Gibbs sampler consists in sampling the path from the joint full conditional distribution $p(\mathbf{s}|-)$. It is based on the factorisation of the full joint conditional distribution as the product of the entries of the transition matrix Ξ and the filtered probabilities. Since the observations $(\mathcal{X}_t, \mathbf{y}_t)$ depend only on the contemporaneous value of the hidden chain s_t, filtering the state probabilities is feasible. Staring from the complete data likelihood in eq. (3.18), we integrate the latent variables $(\mathcal{D}, \boldsymbol{\Omega})$ and sample the trajectory from the full joint conditional distribution:

$$p(\mathbf{s} | \boldsymbol{\mathcal{X}}, \mathbf{y}, \mathcal{G}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \Xi) \propto p(\boldsymbol{\mathcal{X}}, \mathbf{y}, \mathbf{s} | \mathcal{G}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{\mathcal{X}}, \mathbf{y} | \mathbf{s}, \mathcal{G}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\mathbf{s} | \Xi)\,. \tag{C.88}$$

Consequently, at each iteration of the Gibbs sampler we firstly compute the filtered state probabilities using $p(\boldsymbol{\mathcal{X}}, \mathbf{y} | \mathbf{s}, \mathcal{G}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ as likelihood function. Define $\mathcal{X}^{t-1} = \{\mathcal{X}_1, \ldots, \mathcal{X}_{t-1}\}$ and $\mathbf{y}^{t-1} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}\}$. Since the two observation processes are independent from each other as well as from their own past conditionally on the current state, the predictive probability correspond to the conditional distribution of the observations given the state:

$$\begin{aligned} p(\mathcal{X}_t, \mathbf{y}_t | s_t = l, \mathcal{X}^{t-1}, \mathbf{y}^{t-1}, \mathcal{G}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= p(\mathcal{X}_t, \mathbf{y}_t | s_t = l, \mathcal{G}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= p(\mathbf{y}_t | s_t = l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \cdot p(\mathcal{X}_t | s_t = l, \rho_l, \mathcal{G}_l) \qquad \text{(C.89a)} \\ &= p(\mathbf{y}_t | s_t = l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \cdot \prod_{i=1}^{I} \prod_{j=i}^{J} \prod_{k=1}^{K} p(x_{ijk,t} | s_t = l, \rho_l, \mathcal{G}_l) \, . \qquad \text{(C.89b)} \end{aligned}$$

From eq. (C.5), we have that the logarithm of the predictive probability is:

$$\begin{aligned} &\log p(\mathcal{X}_t, \mathbf{y}_t | s_t = l, \mathcal{X}^{t-1}, \mathbf{y}^{t-1}, \mathcal{G}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \\ &= \log p(\mathbf{y}_t | s_t = l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \log p(x_{ijk,t} | s_t = l, \rho_l, \mathcal{G}_l) \end{aligned} \tag{C.90}$$

where:

$$p(\mathbf{y}_t | s_t = l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) = (2\pi)^{-m/2} \left| \boldsymbol{\Sigma}_l \right|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y}_y - \boldsymbol{\mu}_l)' \boldsymbol{\Sigma}_l^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_l) \right\} \tag{C.91}$$

$$p(x_{ijk,t} = 1 | s_t = l, \rho_l, \mathcal{G}_l) = (1 - \rho_l) \frac{\exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}}{1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}} \tag{C.92}$$

$$p(x_{ijk,t} = 0 | s_t = l, \rho_l, \mathcal{G}_l) = \rho_l + (1 - \rho_l) \frac{1}{1 + \exp\{\mathbf{z}_t' \mathbf{g}_{ijk,l}\}} \, . \tag{C.93}$$

C.4 Computation for Pooled case

The complete data likelihood from (C.13) reads:

$$\begin{aligned} L(\boldsymbol{\mathcal{X}} | \boldsymbol{\theta}) = &\prod_{t \in \mathcal{T}_l} \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \rho_l^{d_{ijk,t}} \cdot \left(0^{x_{ijk,t}} 1^{1 - x_{ijk,t}} \right)^{d_{ijk,t}} \cdot \left(\frac{1 - \rho_l}{2} \right)^{1 - d_{ijk,t}} \\ &\cdot \exp \left\{ -\frac{\omega_{ijk,t}}{2} (\mathbf{z}_t' \mathbf{g}_{ijk,l})^2 + \kappa_{ijk,t} (\mathbf{z}_t' \mathbf{g}_{ijk,l}) \right\} p(\omega_{ijk,t}) \, . \end{aligned} \tag{C.1}$$

In the pooling case, we are assuming that the tensor of coefficients in each regime $l = 1, \ldots, L$ is given by:

$$\mathcal{G}_l = g_l \cdot \mathcal{I} \, , \tag{C.2}$$

where $\mathcal{I}$ is a $I \times J \times Q \times K$ tensor made of ones and $g_l \in \mathbb{R}$, for each $l = 1, \ldots, L$. Therefore $\mathbf{g}_{ijk,l} = g_l \boldsymbol{\iota}_Q$, where $\boldsymbol{\iota}_Q$ is a column vector of ones of length Q. We can rewrite the complete data likelihood as:

$$\begin{aligned} L(\boldsymbol{\mathcal{X}} | \boldsymbol{\theta}) &\propto \prod_{t \in \mathcal{T}_l} \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \exp \left\{ -\frac{\omega_{ijk,t}}{2} (\mathbf{z}_t' g_l \boldsymbol{\iota}_Q)^2 + \kappa_{ijk,t} (\mathbf{z}_t' g_l \boldsymbol{\iota}_Q) \right\} \\ &= \prod_{t \in \mathcal{T}_l} \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \exp \left\{ -\frac{\omega_{ijk,t}}{2} (g_l S_t^z)^2 + \kappa_{ijk,t} (g_l S_t^z) \right\} \, , \end{aligned} \tag{C.3}$$

where $S_t^z = \mathbf{z}_t' \boldsymbol{\iota}_Q = \sum_{q=1}^{Q} z_{q,t}$. Then:

$$L(\boldsymbol{\mathcal{X}}|\boldsymbol{\theta}) \propto \exp\left\{ -\frac{1}{2} \sum_{t\in\mathcal{T}_l} \sum_{i,j,k} g_l^2 (S_t^z)^2 \omega_{ijk,t} - 2 g_l S_t^z \kappa_{ijk,t} \right\}, \tag{C.4}$$

It is assumed that g_l, for each $l = 1, \dots, L$, has prior distribution:

$$g_l|\tau, w_l \sim \mathcal{N}(0, \tau w_l). \tag{C.5}$$

This yields the posterior distribution:

$$\begin{aligned} p(g_l|\boldsymbol{\Omega}_t, \tau, w_l) &\propto \exp\left\{ -\frac{1}{2} \sum_{t\in\mathcal{T}_l} \sum_{i,j,k} g_l^2 (S_t^z)^2 \omega_{ijk,t} - 2 g_l S_t^z \kappa_{ijk,t} \right\} \exp\left\{ -\frac{1}{2} \frac{g_l^2}{\tau w_l} \right\} \\ &= \exp\left\{ -\frac{1}{2} \left[\frac{g_l^2}{\tau w_l} + \sum_{t\in\mathcal{T}_l} \sum_{i,j,k} g_l^2 (S_t^z)^2 \omega_{ijk,t} - 2 g_l S_t^z \kappa_{ijk,t} \right] \right\} \\ &= \exp\left\{ -\frac{1}{2} \left[g_l^2 \left(\frac{1}{\tau w_l} + \sum_{t\in\mathcal{T}_l} \sum_{i,j,k} (S_t^z)^2 \omega_{ijk,t} \right) - 2 g_l \left(\sum_{t\in\mathcal{T}_l} \sum_{i,j,k} S_t^z \kappa_{ijk,t} \right) \right] \right\}. \end{aligned} \tag{C.6}$$

Therefore, for each $l = 1, \dots, L$:

$$\pi(g_l|\boldsymbol{\Omega}_t, \tau, w_l) \sim \mathcal{N}(m_l, s_l^2), \tag{C.7}$$

with:

$$s_l^2 = \left(\frac{1}{\tau w_l} + \sum_{t\in\mathcal{T}_l} \sum_{i,j,k} (S_t^z)^2 \omega_{ijk,t} \right)^{-1} \tag{C.8}$$

$$m_l = \left(\sum_{t\in\mathcal{T}_l} \sum_{i,j,k} S_t^z \kappa_{ijk,t} \right) \cdot s_l^{-2}. \tag{C.9}$$

Assume the prior distributions:

$$\pi(\tau) \sim \mathcal{G}a(\overline{a}^\tau, \overline{b}^\tau) \tag{C.10}$$

$$\pi(w_l|\lambda_l) \sim \mathcal{E}xp(\lambda_l^2/2) \tag{C.11}$$

$$\pi(\lambda_l) \sim \mathcal{G}a(a_\lambda^l, b_\lambda^l), \tag{C.12}$$

then the posterior distributions of the variance hyper-parameters τ, w_l, λ_l are obtained as follows.

The posterior distribution of τ is given by:

$$\begin{aligned} p(\tau|\mathbf{g}, \mathbf{w}) &\propto \pi(\tau) p(\mathbf{g}|\mathbf{w}, \tau) \\ &\propto \tau^{\overline{a}^\tau - 1} \exp\left\{ -\overline{b}^\tau \tau \right\} \prod_{l=1}^{L} \exp\left\{ -\frac{g_l^2}{2\tau w_l} \right\} \end{aligned}$$

$$
\begin{aligned}
&= \tau^{\bar{a}^\tau - 1} \exp\left\{ -\frac{1}{2}\left[2\bar{b}^\tau \tau + \sum_{l=1}^{L} \frac{g_l^2}{w_l}\frac{1}{\tau} \right] \right\} \\
&\sim \text{GiG}\left(\bar{a}^\tau - 1, 2\bar{b}^\tau, \sum_{l=1}^{L} \frac{g_l^2}{w_l} \right) . \qquad \text{(C.13)}
\end{aligned}
$$

The posterior distribution of w_l, for each $l = 1, \dots, L$, is given by:

$$
\begin{aligned}
p(w_l|g_l, \tau, \lambda_l) &\propto \pi(w_l|\lambda_l) p(g_l|w_l, \tau) \\
&\propto \frac{\lambda_l^2}{2} \exp\left\{ -\frac{\lambda_l^2}{2} w_l \right\} \exp\left\{ -\frac{g_l^2}{2\tau w_l} \right\} \\
&= \exp\left\{ -\frac{1}{2}\left[\lambda_l^2 w_l + \frac{g_l^2}{\tau}\frac{1}{w_l} \right] \right\} \\
&\sim \text{GiG}\left(1, \lambda_l^2, \frac{g_l^2}{\tau} \right) . \qquad \text{(C.14)}
\end{aligned}
$$

The posterior distribution of λ_l (integrating out w_l), for each $l = 1, \dots, L$, is given by:

$$
\begin{aligned}
p(\lambda_l|\tau, g_l) &\propto \pi(\lambda_l) \int p(g_l|\tau, w_l) p(w_l|\lambda_l) \, \mathrm{d}w_l \\
&\propto \pi(\lambda_l) p(g_l|\tau, \lambda_l) \\
&\propto \lambda_l^{a_\lambda^l - 1} \exp\left\{ -b_\lambda^l \lambda_l \right\} \frac{\sqrt{\tau}}{2\lambda_l} \exp\left\{ -\frac{|g_l|\sqrt{\tau}}{\lambda_l} \right\} \\
&\propto \lambda_l^{a_\lambda^l - 2} \exp\left\{ -\frac{1}{2}\left[2b_\lambda^l \lambda_l + |g_l|\sqrt{\tau}\frac{1}{\lambda_l} \right] \right\} \\
&\sim \text{GiG}\left(a_\lambda^l - 1, 2b_\lambda^l, |g_l|\sqrt{\tau} \right) . \qquad \text{(C.15)}
\end{aligned}
$$

C.5 Additional Simulations' Output

C.5.1 Size 100,100,3,2

Setup: $I = J = 100$, $Q = 3$, $M = 2$.
We run the Gibbs sampler for $N = 500$ iterations and the outcome is plotted from Fig. C.5 to C.11(b).

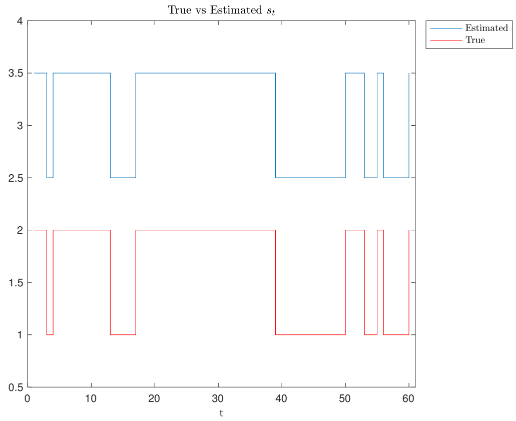
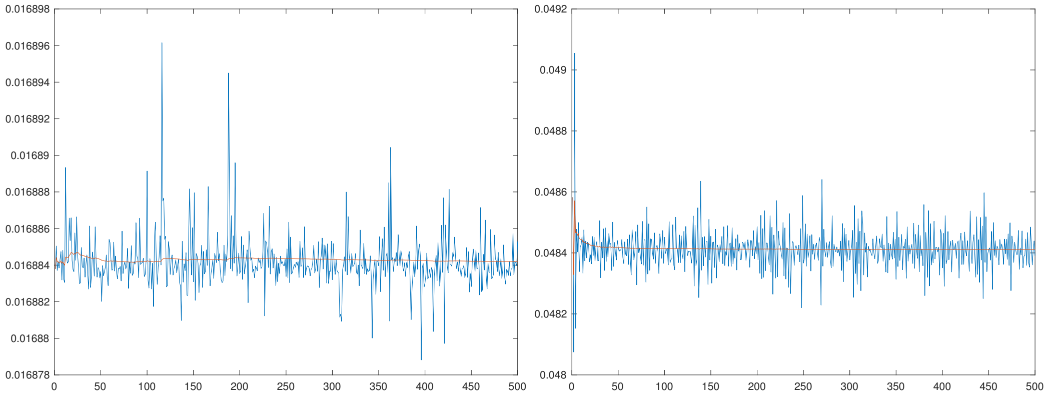


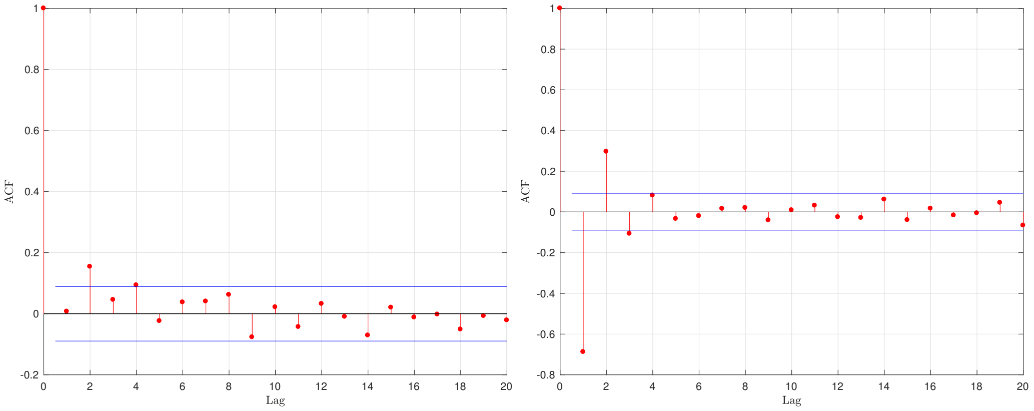
FIGURE C.5: Hidden Markov chain: true (blue) versus estimated (red).



(a) Regime 1.

(b) Regime 2.

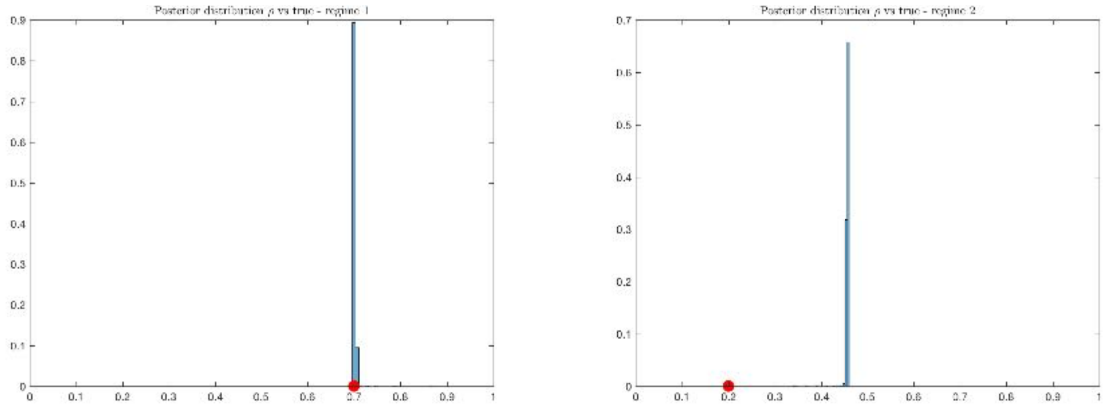
FIGURE C.6: Frobenious norm (blue line) and its progressive mean (red line) of the difference between the true tensor $\mathcal{G}_l^*$ and the MCMC samples of the tensor $\hat{\mathcal{G}}_l$.



(a) Regime 1.

(b) Regime 2.

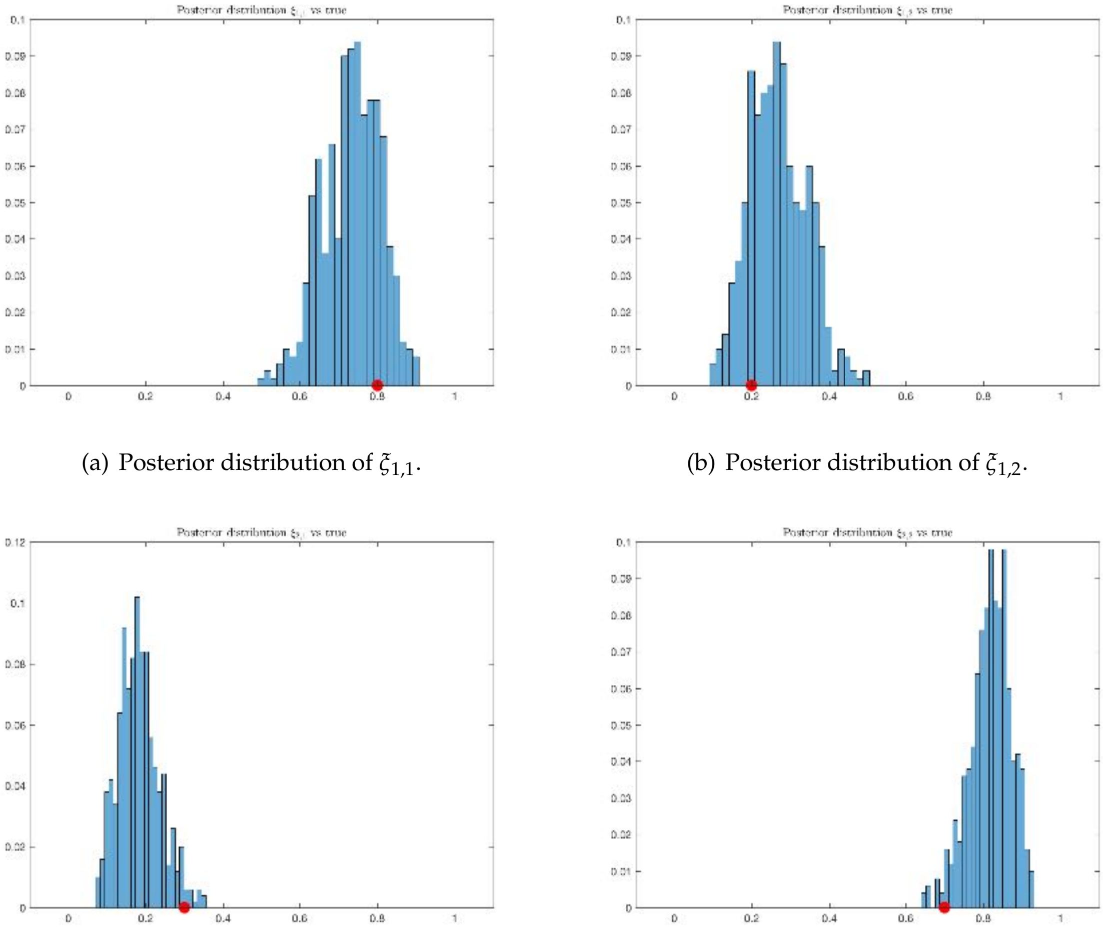
FIGURE C.7: ACF of Frobenious norm of the difference between the true tensor $\mathcal{G}_l^*$ and the MCMC samples of the tensor $\hat{\mathcal{G}}_l$.



(a) Regime 1.

(b) Regime 2.

FIGURE C.8: Posterior distribution of the mixing probability parameter ρ_l (blue) and the true value of the parameter (red).



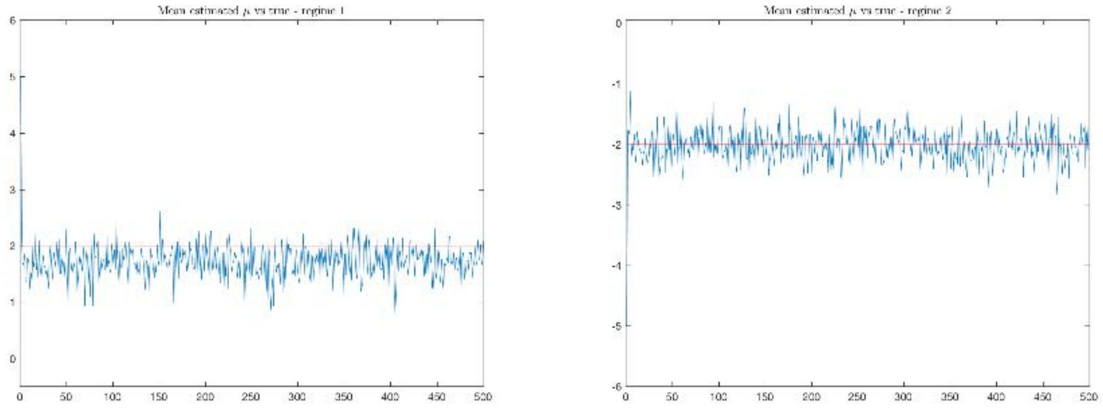
(a) Posterior distribution of $\xi_{1,1}$.

(b) Posterior distribution of $\xi_{1,2}$.

(c) Posterior distribution of $\xi_{2,1}$.

(d) Posterior distribution of $\xi_{2,2}$.

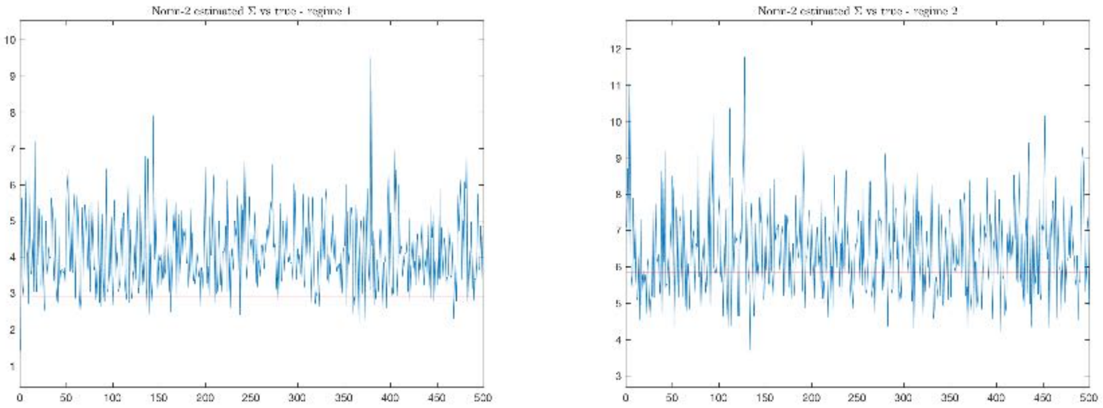
FIGURE C.9: Posterior distribution (blue) and true value (red) of the transition probabilities.



(a) Regime 1.

(b) Regime 2.

FIGURE C.10: Frobenious norm of the MCMC samples (blue line) and true value (red line) of the parameter $\hat{\boldsymbol{\mu}}_l$.



(a) Regime 1.

(b) Regime 2.

FIGURE C.11: Frobenious norm of the MCMC samples (blue line) and true value (red line) of the parameter $\hat{\boldsymbol{\Sigma}}_l$.

C.5.2 Size 150,150,3,2

Setup: $I = J = 150$, $Q = 3$, $M = 2$.
We run the Gibbs sampler for $N = 500$ iterations and the outcome is plotted in the following figures.

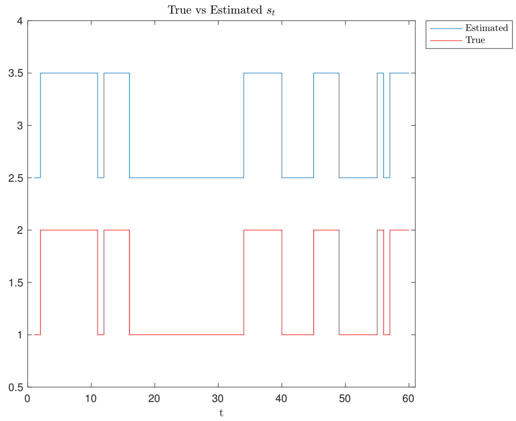
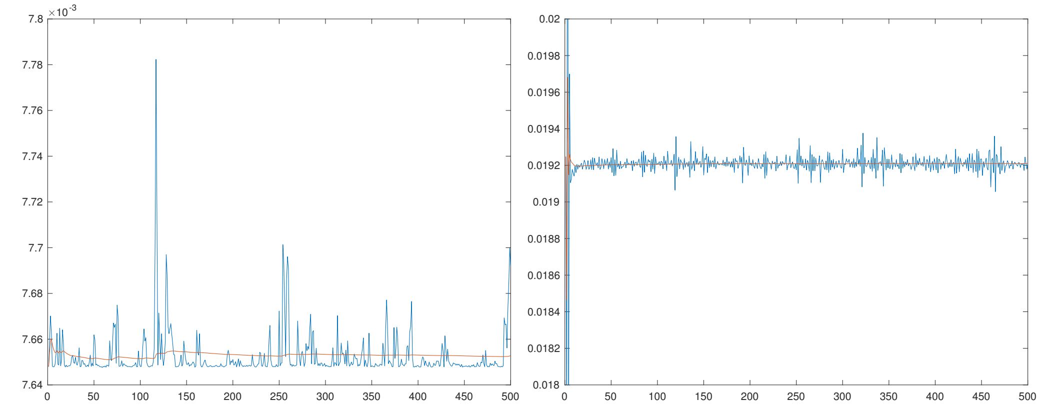


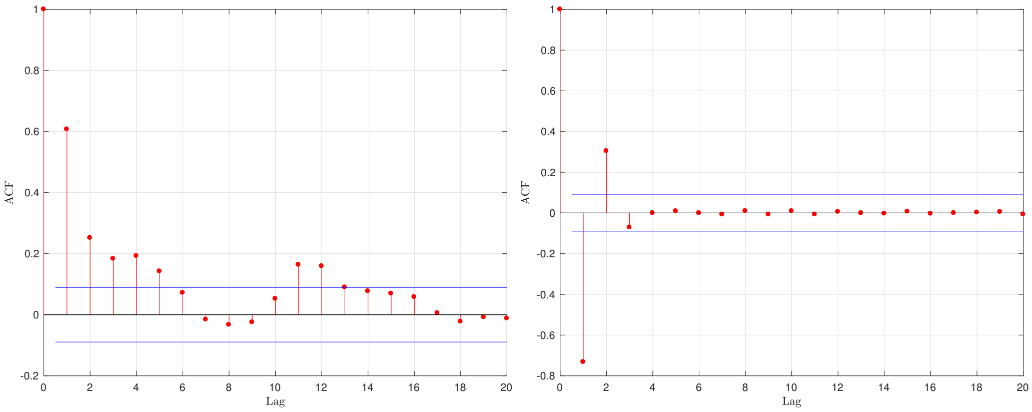
FIGURE C.12: Hidden Markov chain: true (blue) versus estimated (red).



(a) Regime $l = 1$.

(b) Regime $l = 2$.

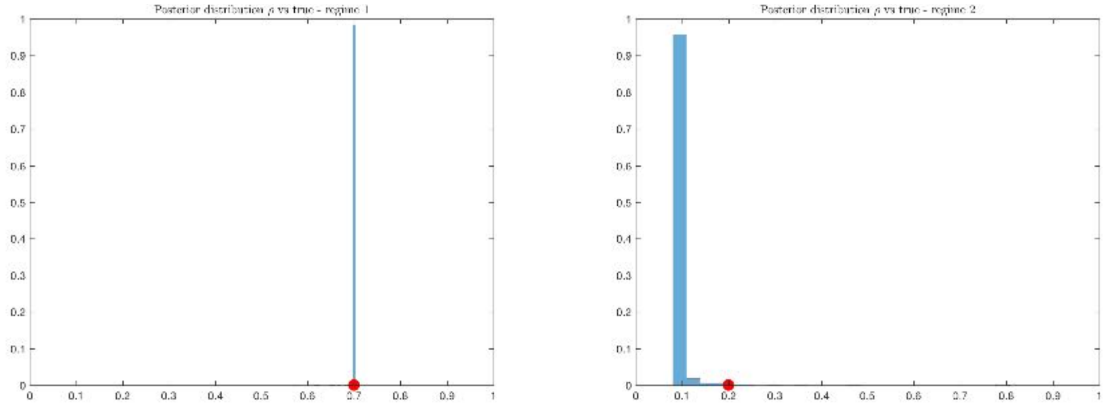
FIGURE C.13: Frobenious norm (blue line) and its progressive mean (red line) of the difference between the true tensor $\mathcal{G}_l^*$ and the MCMC samples of the tensor $\hat{\mathcal{G}}_l$.



(a) Regime 1.

(b) Regime 2.

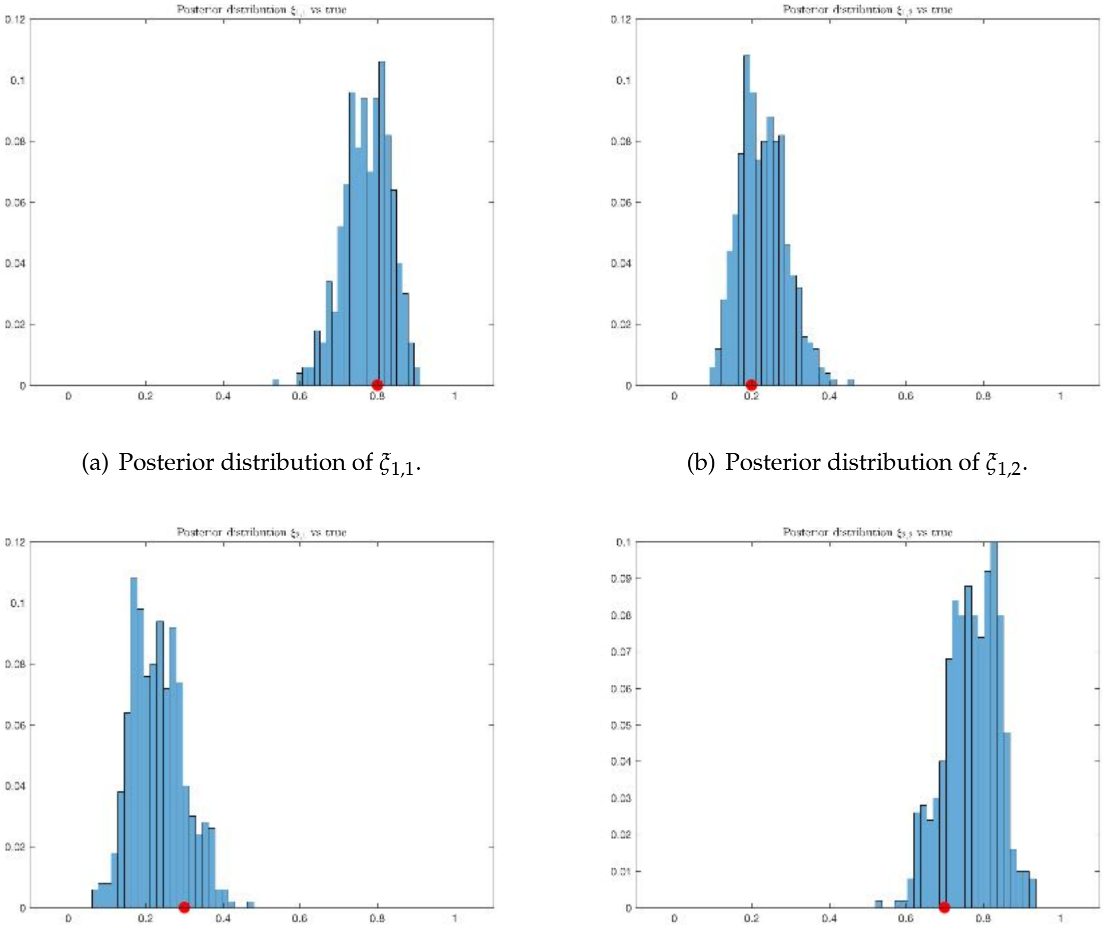
FIGURE C.14: ACF of Frobenious norm of the difference between the true tensor $\mathcal{G}_l^*$ and the MCMC samples of the tensor $\hat{\mathcal{G}}_l$.



(a) Regime 1.

(b) Regime 2.

FIGURE C.15: Posterior distribution of the mixing probability parameter ρ_l (blue) and the true value of the parameter (red).



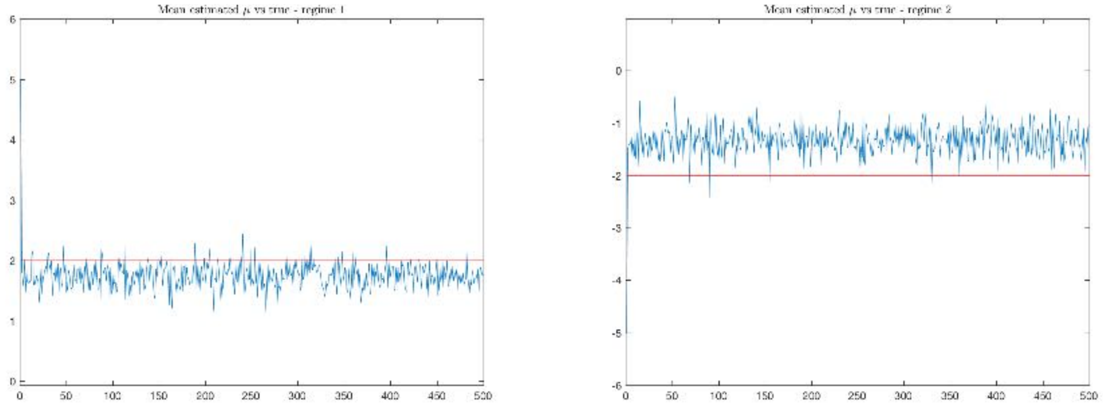
(a) Posterior distribution of $\xi_{1,1}$.

(b) Posterior distribution of $\xi_{1,2}$.

(c) Posterior distribution of $\xi_{2,1}$.

(d) Posterior distribution of $\xi_{2,2}$.

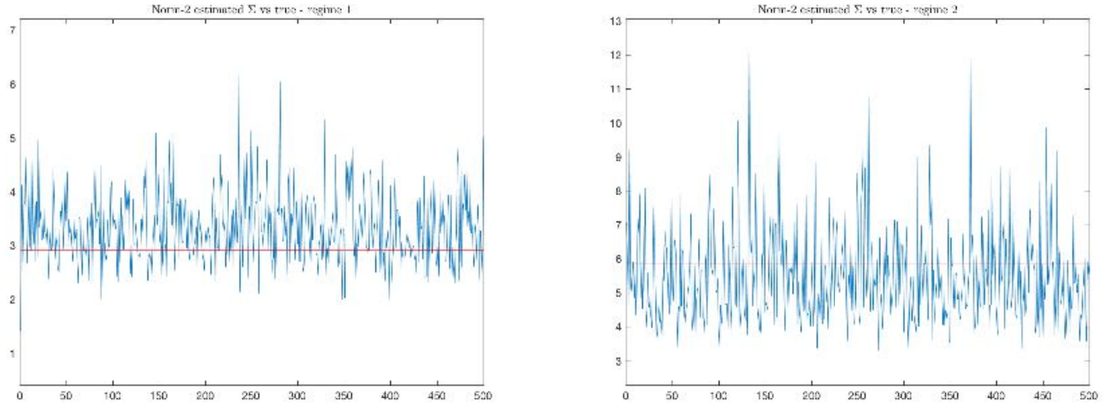
FIGURE C.16: Posterior distribution (blue) and true value (red) of the transition probabilities.



(a) Regime 1.

(b) Regime 2.

FIGURE C.17: Frobenious norm of the MCMC samples (blue line) and true value (red line) of the parameter $\hat{\boldsymbol{\mu}}_l$.



(a) Regime 1.

(b) Regime 2.

FIGURE C.18: Frobenious norm of the MCMC samples (blue line) and true value (red line) of the parameter $\hat{\boldsymbol{\Sigma}}_l$.

C.5.3 Size 200,200,3,2

Setup: $I = J = 200$, $Q = 3$, $M = 2$.
We run the Gibbs sampler for $N = 200$ iterations and the outcome is plotted in the following figures.

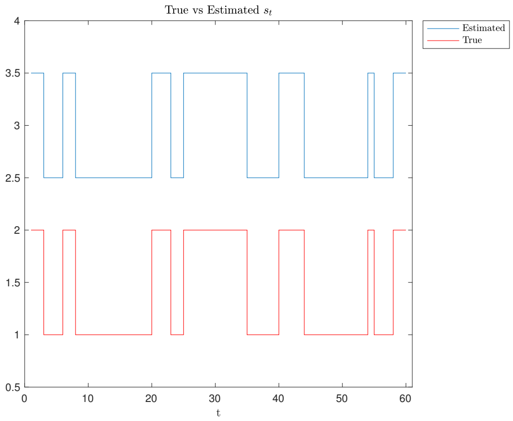
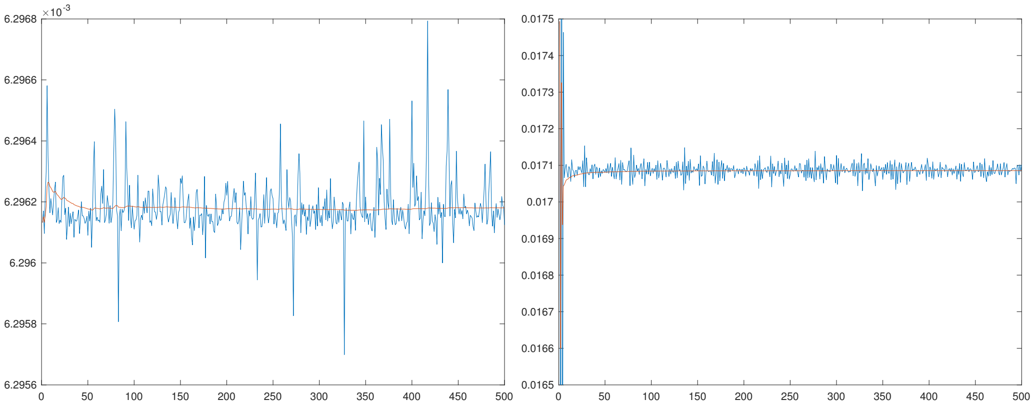


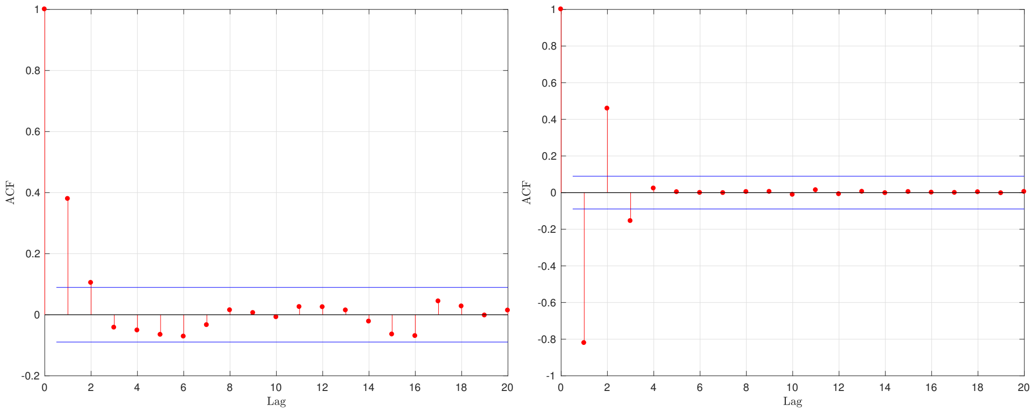
FIGURE C.19: Hidden Markov chain: true (blue) versus estimated (red).



(a) Regime 1.

(b) Regime 2.

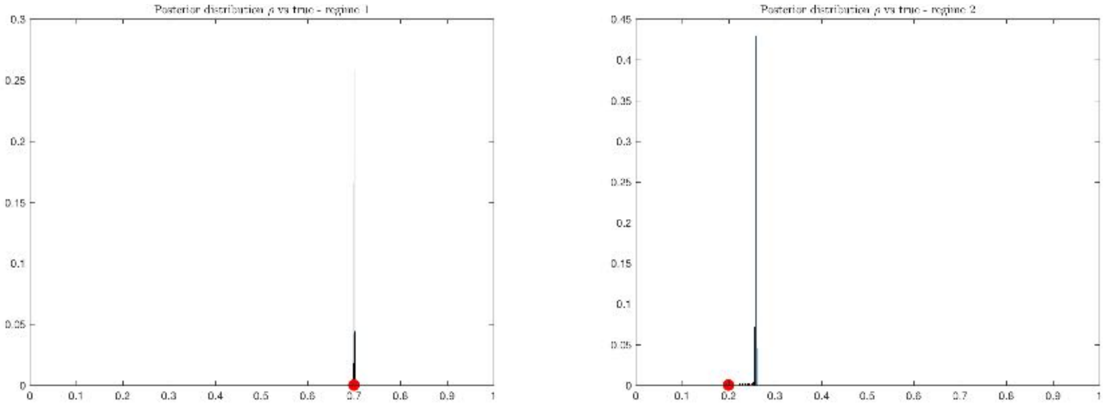
FIGURE C.20: Frobenious norm (blue line) and its progressive mean (red line) of the difference between the true tensor $\mathcal{G}_l^*$ and the MCMC samples of the tensor $\hat{\mathcal{G}}_l$.



(a) Regime 1.

(b) Regime 2.

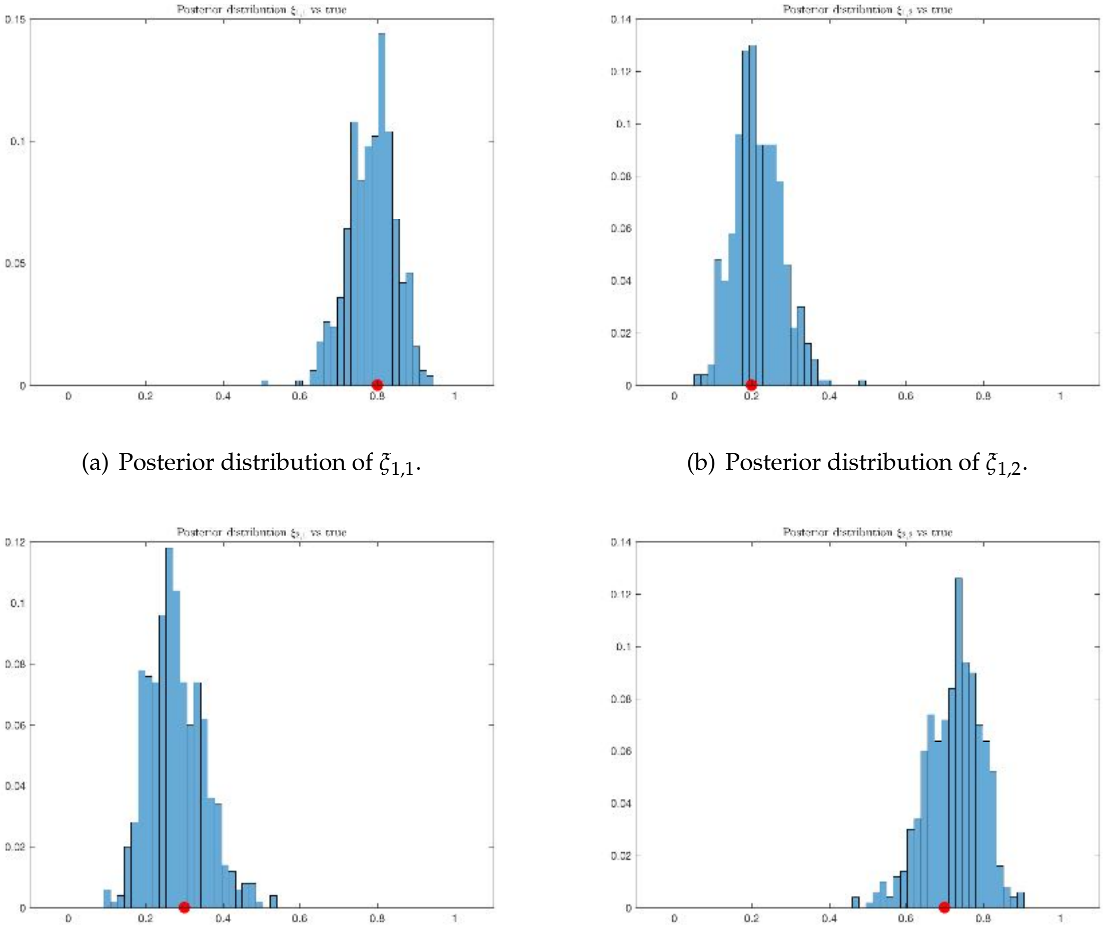
FIGURE C.21: ACF of Frobenious norm of the difference between the true tensor $\mathcal{G}_l^*$ and the MCMC samples of the tensor $\hat{\mathcal{G}}_l$.



(a) Regime 1.

(b) Regime 2.

FIGURE C.22: Posterior distribution of the mixing probability parameter ρ_l (blue) and the true value of the parameter (red).



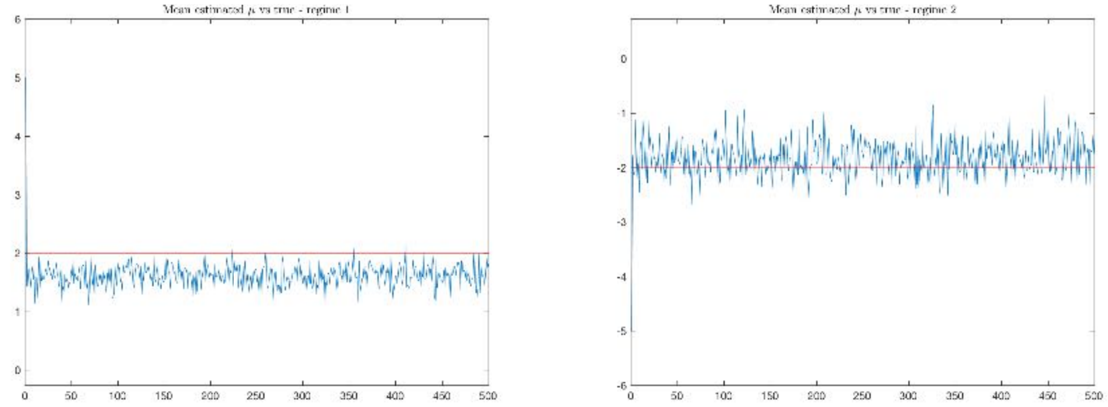
(a) Posterior distribution of $\xi_{1,1}$.

(b) Posterior distribution of $\xi_{1,2}$.

(c) Posterior distribution of $\xi_{2,1}$.

(d) Posterior distribution of $\xi_{2,2}$.

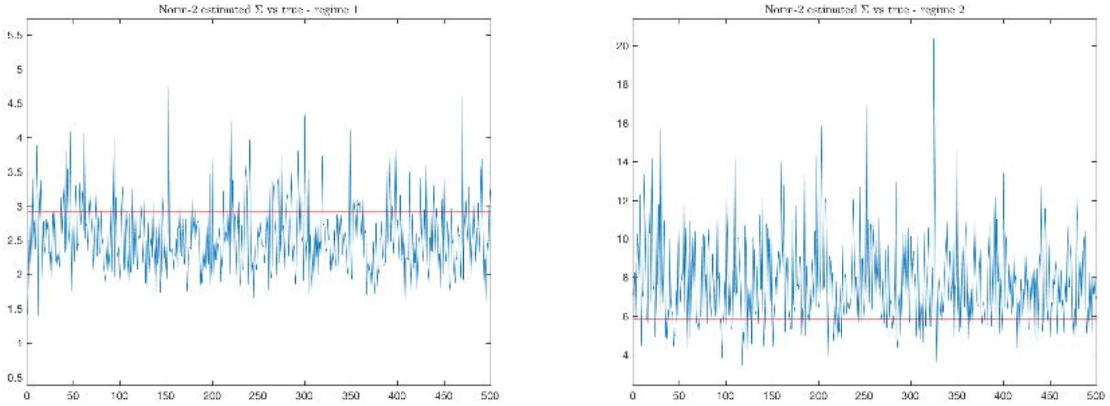
FIGURE C.23: Posterior distribution (blue) and true value (red) of the transition probabilities.



(a) Regime 1.

(b) Regime 2.

FIGURE C.24: Frobenious norm of the MCMC samples (blue line) and true value (red line) of the parameter $\hat{\boldsymbol{\mu}}_l$.



(a) Regime 1.

(b) Regime 2.

FIGURE C.25: Frobenious norm of the MCMC samples (blue line) and true value (red line) of the parameter $\hat{\boldsymbol{\Sigma}}_l$.

C.6 Additional Application's Output

In this section we report some additional plots concerning the Gibbs sampler's output for the estimation of the hyper-parameters in the application described in Section 3.6.

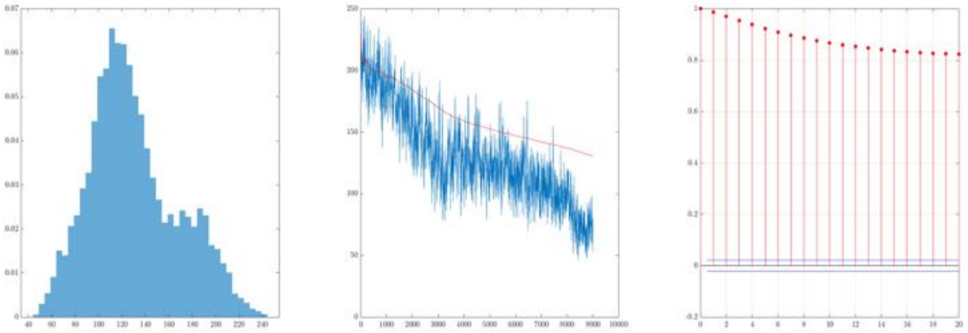


FIGURE C.26: Posterior distribution (*left*), MCMC output (*middle*) and autocorrelation function (*right*) of the global variance parameter τ.

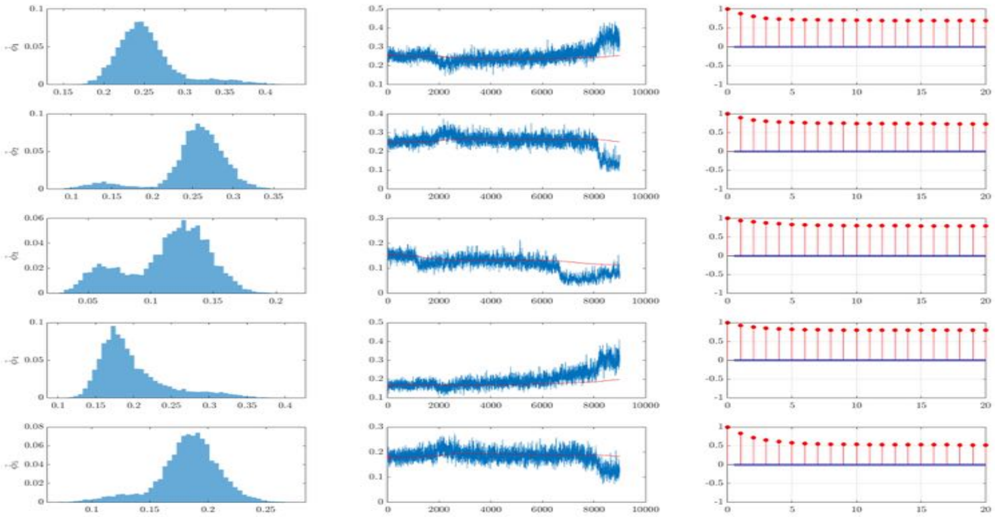


FIGURE C.27: Posterior distribution (*left* plots), MCMC output (*middle* plots) and autocorrelation functions (*right* plots) of the level-specific variance parameters $\boldsymbol{\phi}$. Each row corresponds to a different value of $r = 1, \ldots, R$.

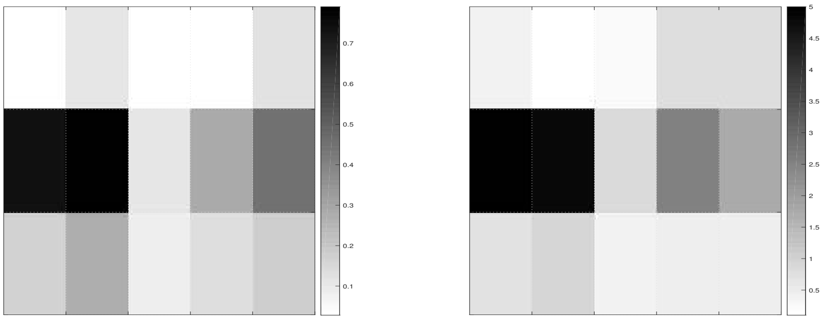


FIGURE C.28: Posterior mean of the variance each marginal of the tensor of coefficients, in state 1 (*left*) and state 2 (*right*). The cell (h, r) of each matrix, for $h = 1, \ldots, 3$ and $r = 1, \ldots, R$, corresponds to the estimated variance $\hat{\tau}\hat{\phi}_r\hat{w}_{h,r,l}$ of the marginal $\gamma_{h,l}^{(r)}$.

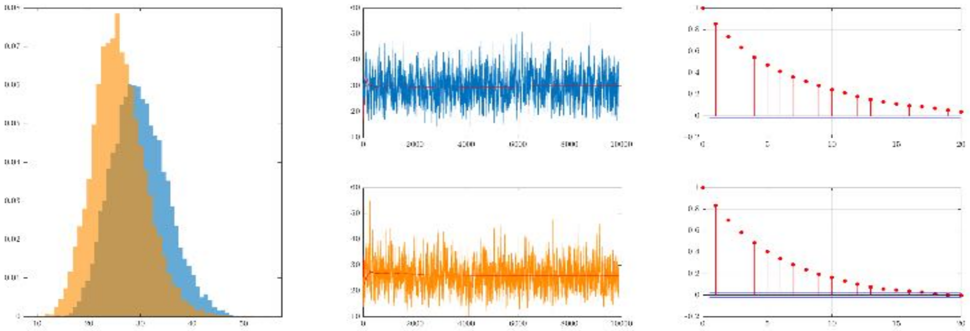


FIGURE C.29: Posterior distribution (*left* plot), MCMC output (*middle* plots) and autocorrelation functions (*right* plots) of the local variance hyper-parameters $\boldsymbol{\lambda}$. Regime 1 (*blue*) and regime 2 (*orange*).

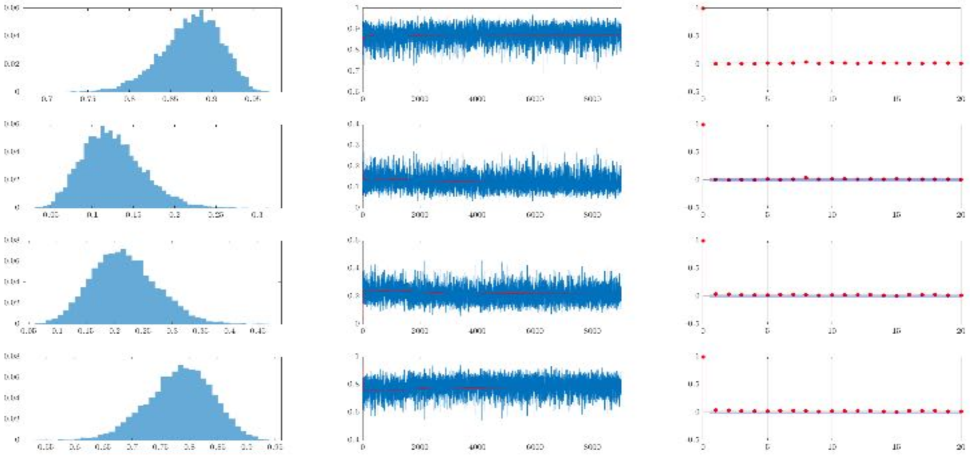


FIGURE C.30: Posterior distribution (*left* plots), MCMC output (*middle* plots) and autocorrelation functions (*right* plots) of the transition probabilities of the hidden Markov chain Ξ, in the order (*top to bottom*): $\xi_{1,1}, \xi_{1,2}, \xi_{2,1}, \xi_{2,2}$.

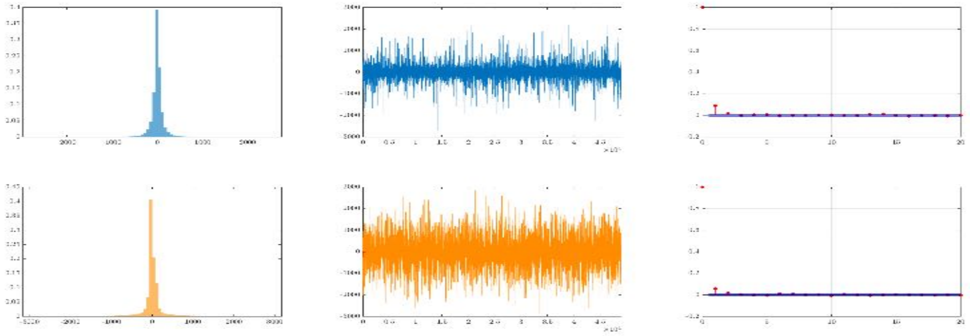


FIGURE C.31: Posterior distribution (*left* plots), MCMC output (*middle* plots) and autocorrelation functions (*right* plots) of the common coefficient g_l in the pooled model. Regime 1 (*blue*) and regime 2 (*orange*).

Appendix D

Appendix D

D.1 Functional PCA

In this section, denote $\mathbf{f}(\cdot) = (f_1(\cdot), \ldots, f_T(\cdot))'$ a sequence of T random functions $f_t : \mathbb{R}^n \to \mathbb{R}$ and let $V : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the covariance operator defined as:

$$V(f)(\cdot) = \int_{\mathbb{R}^n} v(\cdot, \mathbf{y}) f(\mathbf{y}) \, \mathrm{d}\mathbf{y} \,, \tag{D.1}$$

where the kernel $v : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, expressed as $v(\mathbf{x}, \mathbf{y})$, is the covariance function.

Functional principal component analysis (fPCA) is the infinite-dimensional analogue of multivariate principal component analysis (PCA), from which it borrows the terminology and interpretation (see (Ramsay and Silverman, 2005, ch.8) and Ferraty and Vieu (2006)). It is possible to interpret fPCA as a truncated the Karhunen-Loéve decomposition (Karhunen (1947), Loève (1945)). The latter is used to represent a function $f : \mathbb{R}^n \to \mathbb{R}$ via an infinite linear combination of basis functions $\xi_j(\cdot)$ with coefficients β_j given by:

$$f(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \xi_j(\mathbf{x}) \,. \tag{D.2}$$

In fPCA, the infinite sum is truncated by keeping only J components, thus reducing the infinite-dimensional problem into a finite-dimensional one, given by $(\xi_j(\cdot), \beta_j)$, $j = 1, \ldots, J$. In fact, the purpose of fPCA is to find out the linear combination of principal component functions (or factors) $\boldsymbol{\xi}(\cdot) = (\xi_1(\cdot), \ldots, \xi_J(\cdot))'$ and principal component scores (or loadings) $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$, which best approximates a given function (or series of functions). The factors represent the main modes of variability and the scores specify the weight of each principal component function in the approximation of the observed function. Let $\breve{\mathbf{f}} = (\breve{f}_1(\cdot), \ldots, \breve{f}_T(\cdot))'$ the set of functions approximating the series $\mathbf{f} = (f_1(\cdot), \ldots, f_T(\cdot))'$. Then each $\breve{f}_t(\cdot)$ is obtained as:

$$\breve{f}_t(\cdot) = \boldsymbol{\beta}_t' \boldsymbol{\xi}(\cdot) = \sum_{j=1}^{J} \beta_{t,j} \xi_j(\cdot) \,. \tag{D.3}$$

For identification purposes, the principal component functions are often constrained to be orthonormal, that is $||\xi_j(\cdot)||_2 = 1$, $j = 1, \ldots, J$ and $\langle \xi_k(\cdot), \xi_j(\cdot) \rangle = 0$, for $k \neq j$.

Different criteria are available for the choice of the number J of principal components to take in the approximation[1] of eq. (D.7). We interpret the estimation of the factors as an eigenproblem (see next paragraph) and, after having sorted the estimated eigenvalues in decreasing order, we keep the first J eigenfunctions (corresponding to the factors) such that the proportion of variability explained is above a threshold $\bar{d}$. In the empirical analysis,

[1] Notice that the number and shape of the factors necessary to approximate a function provide information about its complexity.

regardless of the criterion used, the value of J is generally very small, thus allowing to interpret and use fPCA as a dimensionality reduction technique for the original series $\mathbf{f}(\cdot)$.

There are several ways to estimate the principal component functions, according to the interpretation of the problem (see Ramsay and Silverman (2005) and Ferraty and Vieu (2006) for a review). By interpreting them as the eigenfunctions of the covariance operator of the functions $(f_1(\cdot), \ldots, f_T(\cdot))$, we can estimate each pair (ξ_j, ρ_j) of principal component function and score by solving the eigenproblem:

$$\begin{cases} \int_{\mathbb{R}^n} V(\mathbf{x}, \mathbf{y}) \xi_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \rho_j \xi_j(\mathbf{y}) & \text{(D.4a)} \\ \langle \xi_j(\cdot), \xi_j(\cdot) \rangle = 1 & \text{(D.4b)} \end{cases}$$

subject to the additional constraint $\langle \xi_k(\cdot), \xi_j(\cdot) \rangle = 0$, for $k \neq j$. For $t = 1, \ldots, T$, $j = 1, \ldots, J$, the principal component scores, in the case of orthonormal eigenfunctions, satisfy:

$$\beta_{j,t} = \int_{\mathbb{R}^n} f_t(\mathbf{x}) \xi_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \langle f_t(\cdot), \xi_j(\cdot) \rangle . \tag{D.5}$$

Following an alternative approach, each function $\xi_k(\cdot)$ is obtained by solving the optimization problem:

$$\begin{cases} \max_{\xi_k} \frac{1}{T} \sum_{t=1}^{T} \left(\int_{\mathbb{R}^n} f_t(\mathbf{x}) \xi_k(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right)^2 & \text{(D.6a)} \\ \text{s.t. } \left\| \xi_k(\mathbf{x}) \right\|_2 = 1 & \text{(D.6b)} \end{cases}$$

with the additional constraint that $\langle \xi_k, \xi_j \rangle = 0$, for $k \neq j$. For $t = 1, \ldots, T$, $j = 1, \ldots, J$, the scores are obtained again from eq. (D.5). In both cases, the output is a sequence of estimated factors $\widehat{\boldsymbol{\xi}}(\cdot) = (\widehat{\xi}_1(\cdot), \ldots, \widehat{\xi}_J(\cdot))'$ and scores $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_T)$, with $\widehat{\boldsymbol{\beta}}_t = (\widehat{\beta}_{t,1}, \ldots, \widehat{\beta}_{t,J})'$ for $t = 1, \ldots, T$. Then, we obtain:

$$\mathbf{f}(\cdot) \approx \check{\mathbf{f}}(\cdot) = \widehat{\mathbf{B}}' \widehat{\boldsymbol{\xi}}(\cdot) , \qquad \check{f}_t(\cdot) = \widehat{\boldsymbol{\beta}}_t' \widehat{\boldsymbol{\xi}}(\cdot) . \tag{D.7}$$

In the paper we follow the first interpretation and estimate the principal component functions and scores by solving an eigenproblem. This poses the preliminary problem of estimating the covariance of the observed sample of functions $\mathbf{f}(\cdot)$. The standard sample covariance function estimator is given by:

$$\widehat{V}(\mathbf{x}, \mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{x}) f_t(\mathbf{y}) . \tag{D.8}$$

Alternative non-parametric estimators have been developed in the earlier contributions of Hall et al. (2006), Li and Hsing (2010), Yao et al. (2005) and Staniswalis and Lee (1998). In matrix notation, eq. (D.8) is written as $v(\mathbf{x}, \mathbf{y}) = T^{-1} \mathbf{f}'(\cdot) \mathbf{f}(\cdot)$. By exploiting eqs. (D.8) and (D.7) we get:

$$T^{-1} \mathbf{f}'(\cdot) \mathbf{f}(\cdot) = T^{-1} \boldsymbol{\xi}(\cdot) \mathbf{B}' \mathbf{B} \boldsymbol{\xi}(\cdot) . \tag{D.9}$$

Therefore, the k-th principal component function $\widehat{\xi}_k(\cdot)$ and the score $\widehat{\boldsymbol{\beta}}_t = (\widehat{\beta}_{t,1}, \ldots, \widehat{\beta}_{t,J})' = (\langle f_t(\cdot), \xi_1(\cdot) \rangle, \ldots, \langle f_t(\cdot), \xi_J(\cdot) \rangle)'$, for $t = 1, \ldots, T$, are obtained by solving the eigenproblem[2]:

$$V \xi_k(\cdot) = \rho_k \xi_k(\cdot) , \tag{D.10}$$

[2] This can also be interpreted as a n-dimensional Fredhölm integral equation of the second type, see Atkinson (2009), Atkinson and Han (2005)

under the constraints $\langle \xi_k(\cdot), \xi_j(\cdot) \rangle = 0$ for $k \neq j$ and $||\xi_k(\cdot)||_2 = 1$. One way of solving the eigenproblem requires to discretize the functions on a specified grid of points $\{\mathbf{x}_i\}_{i=1}^N$, which permits to re-state eq. (D.10) as a finite-dimensional eigenproblem in matrix form. Then, standard methods used in multivariate PCA are applied for obtaining the solution.

Alternatively, one may assume that both the original functions $f_t(\cdot)$ and the eigenfunctions $\xi_k(\cdot)$ can be expressed as a finite linear combination of some chosen basis functions $\boldsymbol{\psi}(\cdot) = (\psi_1(\cdot), \ldots, \psi_K(\cdot))'$, with different coefficients:

$$f_t(\cdot) = \sum_{k=1}^{K} d_{t,k}\psi_k(\cdot) = \mathbf{d}_t'\boldsymbol{\psi}(\cdot), \qquad \xi_j(\cdot) = \sum_{k=1}^{K} a_{j,k}\psi_k(\cdot) = \mathbf{a}_j'\boldsymbol{\psi}(\cdot), \tag{D.11}$$

for $t = 1, \ldots, T$ and $j = 1, \ldots, J$. Given the choice of the basis functions, this reduces the infinite-dimensional problem for $\xi_j(\cdot)$ to a finite-dimensional one for the vector $\mathbf{a}_j = (a_{j,1}, \ldots, a_{j,K})'$. From eqs. (D.1), (D.8) and (D.10) we obtain:

$$T^{-1}\boldsymbol{\psi}(\cdot)'\mathbf{D}'\mathbf{D}\mathbf{M}\mathbf{a}_j = \rho_j\boldsymbol{\psi}(\cdot)'\mathbf{a}_j \tag{D.12}$$

$$T^{-1}\mathbf{D}'\mathbf{D}\mathbf{M}\mathbf{a}_j = \rho_j\mathbf{a}_j, \tag{D.13}$$

with $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_T)$ and $\mathbf{M} = (\langle \psi_k(\cdot), \psi_j(\cdot) \rangle)_{k,j}$, which is the identity matrix if the basis functions form an orthonormal system.

As common practice in multivariate PCA, the estimated eigenvalues $\widehat{\rho}_1, \widehat{\rho}_2, \ldots$ are then sorted in decreasing order and the number J of principal component functions to take is decided on the basis of the proportion of total variation explained $J = \arg\min_j\{\sum_j \widehat{\rho}_j > \bar{d}\}$.

D.2 Computations

In this section we provide the details of the computations needed in Section 4.3. We start by recalling a result from Lyche and Morken (2008) stating some useful properties of B-spline functions. A comprehensive discussion of spline functions and their properties can be found in De Boor (2001) and Schumaker (2007).

D.2.1 Proof of Lemma 4.3.0.1

In the following we show the procedure for solving the constrained optimal smoothing problem in eq. (4.23). Let $\bar{\boldsymbol{\lambda}}^{x,y} = \bar{\boldsymbol{\lambda}}^x \otimes \bar{\boldsymbol{\lambda}}^y$ denote an extended knot sequence (see Section 4.2.1 for the notation). We define the difference $\bar{\lambda}_{i,k}^{x,y} - \bar{\lambda}_{j,k}^{x,y}$ as the difference between the first coordinate, that is $\bar{\lambda}_{i,k}^{x,y} - \bar{\lambda}_{j,k}^{x,y} = \bar{\lambda}_i^x - \bar{\lambda}_j^x$ and $\bar{\lambda}_{k,i}^{x,y} - \bar{\lambda}_{k,j}^{x,y} = \bar{\lambda}_i^y - \bar{\lambda}_j^y$. In this section, for ease of notation we omit the bar and the superscripts and we implicitly refer to augmented knot sequences, that is we use $\lambda_{i,j}$ instead of $\bar{\lambda}_{i,j}^{x,y}$.

The integral constraint in eq. (4.23) yields:

$$\begin{aligned}\int_{a_1}^{b_1}\int_{a_2}^{b_2} s_m(u,v)\,\mathrm{d}v\,\mathrm{d}u &= \int_{a_1}^{b_1} \tilde{s}_m(u,b_2) - \tilde{s}_m(u,a_2)\,\mathrm{d}u \\ &= s_{m+1}(b_1,b_2) - s_{m+1}(b_1,a_2) - s_{m+1}(a_1,b_2) + s_{m+1}(a_1,a_2) \\ &= s_{m+1}(\lambda_{g+1,g+1}) - s_{m+1}(\lambda_{g+1,0}) - s_{m+1}(\lambda_{0,g+1}) + s_{m+1}(\lambda_{0,0}) = 0\end{aligned} \tag{D.14}$$

Starting from this result, we look for an equation allowing us to express the coefficient of a bivariate spline of order k with those of a spline obtained after differentiating it with respect to both arguments (that is, we look for an analogue of eq. (4.14)).

Now, we should derive the implication that the previous solution has on the coefficients $c_{i,j}$ of the spline $s_{m+1}(u,v)$ (e.g.: in the univariate case we end up with $0 = s_{m+1}(\lambda_{g+1}) - s_{m+1}(\lambda_0) = c_g - c_{-m-1}$ thus implying $c_{-m-1} = c_g$)[3].

By exploiting known properties of splines (see Lyche and Morken (2008)) we obtain:

$$0 = s_{m+1}(\lambda_{g+1,g+1}) - s_{k+1}(\lambda_{g+1,0}) - s_{m+1}(\lambda_{0,g+1}) + s_{m+1}(\lambda_{0,0})\,, \tag{D.15}$$

$$0 = \sum_i \sum_j c_{ij} \left[B_i(\lambda_{g+1})B_j(\lambda_{g+1}) - B_i(\lambda_{g+1})B_j(\lambda_0) - B_i(\lambda_0)B_j(\lambda_{g+1}) + B_i(\lambda_0)B_j(\lambda_0) \right] . \tag{D.16}$$

By property (i) and (iii):

- for $i = j = g$ it holds:

$$B_g(\lambda_{g+1})B_g(\lambda_{g+1}) - B_g(\lambda_{g+1})B_g(\lambda_0) - B_g(\lambda_0)B_g(\lambda_{g+1}) + B_g(\lambda_0)B_g(\lambda_0) = 1\,, \tag{D.17}$$

- for $i = j = -m-1$ it holds:

$$\begin{aligned} &B_{-m-1}(\lambda_{g+1})B_{-m-1}(\lambda_{g+1}) - B_{-m-1}(\lambda_{g+1})B_{-m-1}(\lambda_0) \\ &- B_{-m-1}(\lambda_0)B_{-m-1}(\lambda_{g+1}) + B_{-m-1}(\lambda_0)B_{-m-1}(\lambda_0) = 1\,, \end{aligned} \tag{D.18}$$

- for $i,j \notin \{-m-1, g\}$ the previous equation is always 0 since at least one of the terms of each product is 0.

Therefore we obtain:

$$s_{m+1}(\lambda_{g+1,g+1}) - s_{m+1}(\lambda_{g+1,0}) - s_{m+1}(\lambda_{0,g+1}) + s_{m+1}(\lambda_{0,0}) = 0\,, \tag{D.19}$$

which implies:

$$c_{g,g} + c_{-m-1,-m-1} = 0 \iff c_{g,g} = -c_{-m-1,-m-1}\,. \tag{D.20}$$

Now, by applying sequentially the recursion linking spline function with its partial derivatives and using a knot sequence (or an extended knot sequence) with equal number of knots along both directions:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}u}\frac{\mathrm{d}}{\mathrm{d}v} s_{m+1}(u,v) &= \frac{\mathrm{d}}{\mathrm{d}u}\frac{\mathrm{d}}{\mathrm{d}v} \sum_{i=-m-1}^{g} \sum_{j=-m-1}^{g} c_{ij} B_i^{m+2}(u) B_j^{m+2}(v) \\ &= \frac{\mathrm{d}}{\mathrm{d}v} \sum_{j=-m-1}^{g} B_j^{m+2}(v) \cdot \left(\frac{\mathrm{d}}{\mathrm{d}u} \sum_{i=-m-1}^{g} c_{ij} B_i^{m+2}(u) \right) \\ &= \frac{\mathrm{d}}{\mathrm{d}v} \sum_{j=-m-1}^{g} B_j^{m+2}(v) \cdot \left(\sum_{i=-m}^{g} c_{ij}^{u} B_i^{m+1}(u) \right) && \text{(D.21)} \\ &= \sum_{i=-m}^{g} B_i^{m+1}(u) \left(\frac{\mathrm{d}}{\mathrm{d}v} \sum_{j=-m-1}^{g} c_{ij}^{u} B_j^{m+2}(v) \right) && \text{(D.22)} \\ &= \sum_{i=-m}^{g} \sum_{j=-m}^{g} c_{ij}^{uy} B_i^{m+1}(u) B_j^{m+1}(v)\,. \end{aligned}$$

[3] They obtain the result by using the properties of B-splines in Lyche and Morken (2008). Some bases are exactly 1, others 0, reducing the spline function to the coefficient of the unique basis equal to 1.

Since by definition:

$$s_m(u,v) = \sum_{i=-m}^{g} \sum_{j=-m}^{g} b_{ij} B_i^{m+1}(u) B_j^{m+1}(v)\,, \tag{D.23}$$

by assuming $b_{ij} = c_{ij}^{uy}$ we get the equality:

$$\frac{\mathrm{d}}{\mathrm{d}u}\frac{\mathrm{d}}{\mathrm{d}v} s_{m+1}(u,v) = s_m(u,v)\,. \tag{D.24}$$

It is now necessary to develop the above expression for the constraint on the coefficients in order to find out the precise relationship between the c_{ij} (coefficients of s_{m+1}) and b_{ij} (coefficients of s_m). This is required in order to derive the solution of eq. (4.23) by minimizing the first equation, thus solving an unconstrained optimization problem. First, recall that the coefficients of a univariate spline are related to those of its first order derivative via the relation:

$$\frac{\mathrm{d}}{\mathrm{d}u} s_m(u) = \sum_i \check{c}_i B_i^m(u) = s_{m-1}(u) \qquad \check{c}_i = m\frac{c_i - c_{i-1}}{\lambda_{i+m} - \lambda_i}\,. \tag{D.25}$$

In the bivariate case, first define c_{ij}^u, for fixed j and $i = -m, \ldots, g$, as:

$$c_{ij}^u = (m+1)\frac{c_{i,j} - c_{i-1,j}}{\lambda_{i+m+1,j} - \lambda_{i,j}}\,. \tag{D.26}$$

Then, iterated application eq. (D.25) along each dimension gives, for $j = -m, \ldots, g$:

$$\begin{aligned} b_{i,j} = c_{i,j}^{uy} = (m+1)\frac{c_{i,j}^u - c_{i,j-1}^u}{\lambda_{i,j+m+1} - \lambda_{i,j}} &= \frac{(m+1)^2}{\lambda_{i,j+m+1} - \lambda_{i,j}}\left(\frac{c_{i,j} - c_{i-1,j}}{\lambda_{i+m+1,j} - \lambda_{i,j}} - \frac{c_{i,j-1} - c_{i-1,j-1}}{\lambda_{i+m+1,j-1} - \lambda_{i,j-1}}\right) \\ &= \frac{(m+1)^2}{\lambda_{i,j+m+1} - \lambda_{i,j}}\left(\frac{c_{i,j} - c_{i-1,j}}{\lambda_{i+m+1,j} - \lambda_{i,j}} - \frac{c_{i,j-1} - c_{i-1,j-1}}{\lambda_{i+m+1,j-1} - \lambda_{i,j-1}}\right)\,. \end{aligned} \tag{D.27}$$

This implies that the matrix **B** has the following top-left (i.e. $b_{-m,-m}$) and bottom-right (i.e. $b_{g,g}$) entries:

$$b_{-m,-m} = \frac{(m+1)^2}{\lambda_{-m,1} - \lambda_{-m,-m}}\left(\frac{c_{-m,-m} - c_{-m-1,-m}}{\lambda_{1,-m} - \lambda_{-m,-m}} - \frac{c_{-m,-m-1} - c_{-m-1,-m-1}}{\lambda_{1,-m-1} - \lambda_{-m,-m-1}}\right)\,, \tag{D.28}$$

$$b_{g,g} = \frac{(m+1)^2}{\lambda_{g,g+m+1} - \lambda_{g,g}}\left(\frac{c_{g,g} - c_{g-1,g}}{\lambda_{g+m+1,g} - \lambda_{g,g}} - \frac{c_{g,g-1} - c_{g-1,g-1}}{\lambda_{g+m+1,g-1} - \lambda_{g,g-1}}\right)\,. \tag{D.29}$$

We need conditions for linking the $(g+m+1) \times (g+m+1)$ coefficient matrix $\mathbf{B} = (b_{i,j})_{i,j}$ of the spline function $s_m(u,v)$ and the $(g+m+2) \times (g+m+2)$ coefficient matrix $\mathbf{C} = (c_{i,j})_{i,j}$ of the spline function $s_{m+1}(u,v)$. In the univariate case they are two vectors whose lengths differ by one, and the condition to be imposed consists in the equality of the first and last entry of the coefficient vector of the spline with higher degree. In the bivariate case, instead, $2(g+m+1)+1$ constraints are required:

$$\mathbf{C} = \left[\begin{array}{c|ccc} c_{-m-1,-m-1} & c_{-m-1,-m} & \cdots & c_{-m-1,g} \\ \hline c_{-m,-m-1} & & & \\ \vdots & & \bar{\mathbf{C}} & \\ c_{g,-m-1} & & & \end{array}\right] \tag{D.30}$$

From the previous computations, we obtain the constraint:

$$c_{-m-1,-m-1} = -c_{g,g}\,. \tag{D.31}$$

We need to incorporate this result obtained from the integral constraint. From eq. (D.27) we have that[4]:

$$\begin{aligned} b_{i,j} &= \frac{(m+1)^2}{\lambda_{i,j+(m+1)} - \lambda_{i,j}} \left[\frac{1}{\lambda_{i+(m+1),j} - \lambda_{i,j}}(c_{i,j} - c_{i-1,j}) - \frac{1}{\lambda_{i+(m+1),j-1} - \lambda_{i,j-1}}(c_{i,j-1} - c_{i-1,j-1}) \right] \\ &= \mathbf{D}^i_{j,j} \left[\mathbf{E}^j_{i,i}(\mathbf{K}\mathbf{c}_{:,j})_i - \mathbf{E}^{j-1}_{i,i}(\mathbf{K}\mathbf{c}_{:,j-1})_i \right] \,, \end{aligned} \tag{D.32}$$

which, using the shorthand $N = (g+m+1)$ and letting $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ be the canonical basis of the space of $N \times N$ matrices, gives the following expression for the column vector $\mathbf{b}_{:,j}$:

$$\begin{aligned} \mathbf{b}_{:,j} &= \left[\sum_{k=1}^{N} \mathbf{e}_k \mathbf{e}'_k \otimes (\mathbf{D}^j_{k,k}\mathbf{E}^j_{k,k}) \right] \mathbf{K}\mathbf{c}_{:,j} - \left[\sum_{k=1}^{N} \mathbf{e}_k \mathbf{e}'_k \otimes (\mathbf{D}^{j-1}_{k,k}\mathbf{E}^{j-1}_{k,k}) \right] \mathbf{K}\mathbf{c}_{:,j-1} \\ &= \begin{bmatrix} \mathbf{D}^j_{1,1}\mathbf{E}^j_{1,1} & & & & \\ & \mathbf{D}^j_{2,2}\mathbf{E}^j_{2,2} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \mathbf{D}^j_{N,N}\mathbf{E}^j_{N,N} \end{bmatrix} \begin{bmatrix} (\mathbf{K}\mathbf{c}_{:,j})_1 \\ (\mathbf{K}\mathbf{c}_{:,j})_2 \\ \vdots \\ (\mathbf{K}\mathbf{c}_{:,j})_N \end{bmatrix} \\ &\quad - \begin{bmatrix} \mathbf{D}^j_{1,1}\mathbf{E}^{j-1}_{1,1} & & & & \\ & \mathbf{D}^j_{2,2}\mathbf{E}^{j-1}_{2,2} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \mathbf{D}^j_{N,N}\mathbf{E}^{j-1}_{N,N} \end{bmatrix} \begin{bmatrix} (\mathbf{K}\mathbf{c}_{:,j-1})_1 \\ (\mathbf{K}\mathbf{c}_{:,j-1})_2 \\ \vdots \\ (\mathbf{K}\mathbf{c}_{:,j-1})_N \end{bmatrix} \,. \end{aligned} \tag{D.33}$$

For $j = -m, \ldots, g$, we define the $(g+m+1) \times (g+m+1)$ diagonal matrix $\mathbf{D}^i$ by:

$$\mathbf{D}^j = \operatorname{diag}\left(\frac{(m+1)^2}{\lambda_{-m,j+m+1} - \lambda_{-m,j}}, \frac{(m+1)^2}{\lambda_{-m+1,j+m+1} - \lambda_{-m+1,j}}, \ldots, \frac{(m+1)^2}{\lambda_{g,j+m+1} - \lambda_{g,j}} \right) \,. \tag{D.34}$$

and, for $j = -m-1, -m, \ldots, g$, we define the $(g+m+1) \times (g+m+1)$ diagonal matrix $\mathbf{E}^j$ by:

$$\mathbf{E}^j = \operatorname{diag}\left(\frac{1}{\lambda_{1,j} - \lambda_{-m,j}}, \frac{1}{\lambda_{2,j} - \lambda_{-m+1,j}}, \ldots, \frac{1}{\lambda_{g+m+1,j} - \lambda_{g,j}} \right) \,. \tag{D.35}$$

The matrix $\mathbf{K}$ coincides with the matrix representation of the linear operator L which performs first differences $\mathbf{L}_{g+m+1}$ given by the $(g+m+1) \times (g+m+2)$:

$$\mathbf{K} = \mathbf{L}_{g+m+1} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \,. \tag{D.36}$$

[4] We used the notation $\mathbf{A}^i_{j,j}$ to mean the (j,j)-th entry of the diagonal matrix $\mathbf{A}^i$.

Therefore, for $j = -m, \dots, g$, the $(g+m+1) \times 1$ vector of first differences $\mathbf{Kc}_{:,j}$ is given by:

$$\mathbf{Kc}_{:,j} = \begin{bmatrix} c_{-m,j} - c_{-m-1,j} \\ c_{-m+1,j} - c_{-m,j} \\ \vdots \\ c_{g,j} - c_{g-1,j} \end{bmatrix} . \tag{D.37}$$

In order to rewrite eq. (D.32) in more compact form, we introduce the following $N \times (N+1)$ matrices, which allow to select the first (or last, respectively) N columns from another one by post-multiplication. Let $\{\mathbf{e}_1^N, \dots, \mathbf{e}_N^N\}$ be the canonical basis of the space of square matrices of size $N \times N$ and define the $N \times (N+1)$ matrix $P_{N,m} = [\mathbf{e}_1^N, \dots, \mathbf{e}_{m-1}^N, \mathbf{0}_N, \mathbf{e}_m^N, \dots, \mathbf{e}_N^N]$ with $m \in \{1, \dots, N+1\}$, which, by pre-multiplying a vector of length $(N+1)$, selects all but the m-th entry. The $(N+1) \times N$ matrices $\mathbf{S}_c^f, \mathbf{S}_c^\ell$ defined as follows, instead, when pre-multiplied by a $(N+1) \times (N+1)$ matrix $\mathbf{A}$ select the sub-matrix made with the first (last, respectively) N columns and rows $\mathbf{A}$:

$$\mathbf{S}_c^f = P'_{N,N+1} = \left[\begin{array}{ccc} 1 & & \\ & \ddots & \\ & & 1 \\ \hline 0 & \dots & 0 \end{array}\right] , \qquad \mathbf{S}_c^\ell = P'_{N,1} = \left[\begin{array}{ccc} 0 & \dots & 0 \\ \hline 1 & & \\ & \ddots & \\ & & 1 \end{array}\right] . \tag{D.38}$$

For example, if $\mathbf{A} = [\mathbf{A}^f | \mathbf{a}_{N+1}] = [\mathbf{a}_1 | \mathbf{A}^\ell]$ then $\mathbf{AS}_c^f = \mathbf{A}^f$ and $\mathbf{AS}_c^\ell = \mathbf{A}^\ell$. Notice also that the transposed versions, that is $\mathbf{S}_r^f = (\mathbf{S}_c^f)'$ and $\mathbf{S}_r^\ell = (\mathbf{S}_c^\ell)'$, allow to select rows instead of columns. It is now possible to rewrite eq. (D.32) as:

$$\begin{aligned} \operatorname{vec}(\mathbf{B}) &= \begin{bmatrix} \mathbf{D}^{-m}\mathbf{E}^{-m} & & \\ & \ddots & \\ & & \mathbf{D}^g\mathbf{E}^g \end{bmatrix} \operatorname{vec}\left(\mathbf{S}_r^\ell \tilde{\mathbf{C}} \mathbf{S}_c^\ell\right) - \begin{bmatrix} \mathbf{D}^{-m}\mathbf{E}^{-m-1} & & \\ & \ddots & \\ & & \mathbf{D}^g\mathbf{E}^{g-1} \end{bmatrix} \operatorname{vec}\left(\mathbf{S}_r^f \tilde{\mathbf{C}} \mathbf{S}_c^f\right) \\ &= \mathbf{DE} \operatorname{vec}\left(\mathbf{S}_r^\ell \tilde{\mathbf{C}} \mathbf{S}_c^\ell\right) - \mathbf{DF} \operatorname{vec}\left(\mathbf{S}_r^f \tilde{\mathbf{C}} \mathbf{S}_c^f\right) , \end{aligned} \tag{D.39}$$

where, letting $\{\mathbf{e}_{-m}, \dots, \mathbf{e}_g\}$ be the canonical basis for the space of square matrices of size $(g+m+1) \times (g+m+1)$, we defined:

$$\mathbf{D} = \sum_{i=-m}^{g} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{D}^i \quad \text{size } (g+m+1)^2 \times (g+m+1)^2$$

$$\mathbf{E} = \sum_{i=-m}^{g} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{E}^i \quad \text{size } (g+m+1)^2 \times (g+m+1)^2$$

$$\mathbf{F} = \sum_{i=-m-1}^{g-1} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{E}^i \quad \text{size } (g+m+1)^2 \times (g+m+1)^2$$

$$\tilde{\mathbf{C}} = \left[\mathbf{Kc}_{:,-m} | \dots | \mathbf{Kc}_{:,g+1}\right] \quad \text{size } (g+m+1+1) \times (g+m+1+1) ,$$

and $\mathbf{S}_r^\ell \tilde{\mathbf{C}} \mathbf{S}_c^\ell$ (respectively, $\mathbf{S}_r^f \tilde{\mathbf{C}} \mathbf{S}_c^f$) select the bottom-right (respectively, top-left) square sub-matrix of $\tilde{\mathbf{C}}$ of size $(g+m+1) \times (g+m+1)$, denoted $(\tilde{\mathbf{C}})_{-m}^{g}$ (respectively, $(\tilde{\mathbf{C}})_{-m-1}^{g-1}$).

Now, we need to include in eq. (D.39) the information obtained in eq. (D.31) from the integral constraint in eq. (4.23), that is:

$$c_{-m-1,-m-1} = -c_{g,g} \,. \tag{D.40}$$

To this end, notice that there is no possibility of defining a unique matrix $\mathbf{K}$ which is able to give the desired result by a suitable choice of its entries (as opposed to the univariate case). This is mainly due to the fact that in the linear system representation of $\tilde{\mathbf{C}}$ the terms $c_{-m-1,-m-1}$ and $c_{g,g}$ are never in the same equation and the constraints on the other coefficients of $\mathbf{K}$ prevent to obtain the result. We propose to solve this issue as follows. Instead of transforming the matrix $\mathbf{C}$ and then vectorize it, reverse the order, that is, vectorize $\mathbf{C}$ then apply a suitable transformation. Then, define the $((g+m+1+1)^2-1) \times 1$ vector:

$$\tilde{\mathbf{c}} = P_{(g+m+1+1)^2-1,1} \cdot \operatorname{vec}\left(\mathbf{C}\right) \,, \tag{D.41}$$

which corresponds to the vectorization of the matrix $\mathbf{C}$ without the element[5] $c_{-m-1,-m-1}$. Consider the $N \times (N+1)$ matrix $\mathbf{L}_N$ representing the first difference operator L:

$$\mathbf{L}_N = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \,. \tag{D.42}$$

Since we are dealing with the vectorisation of a matrix, vec $(\mathbf{C})$, we must keep in mind that taking first differences of the whole vector implies taking differences also between the first entry of a column and the last of the previous one, which is undesired. Therefore we need to modify the structure of the difference operator matrix accordingly: we can do it by "shifting" to the right the blocks of non-zero entries every $N-1$ rows, where N is the number of rows of the original matrix $\mathbf{C}$. Moreover, taking into account the integral constraint[6] in eq. (D.31) we get in top right corner 1 instead of 0 whereas the first column of the difference operator matrix is removed. Consequently, we define $\mathbf{K}^*$ to be the matrix with number of columns equal to the size of $\tilde{\mathbf{c}}$ (that is, $(g+m+1+1)^2-1$) and number of rows equal to the number of entries of $\mathbf{C}$ minus a row (that is, $(g+m+1)(g+m+2)$), which is lost by taking differences. We obtain the $(g+m+1)(g+m+2) \times ((g+m+2)^2-1)$ block diagonal matrix:

$$\mathbf{K}^* = \left[\begin{array}{c|c|c|c|c} \mathbf{K}^{11} & & & & \mathbf{K}^{1N} \\ \hline & \mathbf{K}^{-m} & & & \\ \hline & & \mathbf{K}^{-m+1} & & \\ \hline & & & \ddots & \\ \hline & & & & \mathbf{K}^{g} \end{array}\right] \,, \tag{D.43}$$

where each $(g+m+1) \times (g+m+2)$ matrix $\mathbf{K}^i = \mathbf{L}_{g+m+1}$, for $i = -m, \ldots, g$, whereas $\mathbf{K}^{1N}$ is a $(g+m+1) \times (g+m+2)$ with all zeros but the top-right entry and the $(g+m+1) \times$

[5] We remove from vec $(\mathbf{C})$ all the entries equal to $c_{-m-1,-m-1}$, thus obtaining a vector $\mathbf{c}^*$ of length equal to the length of vec $(\mathbf{C})$ minus the number of occurrences of $c_{-m-1,-m-1}$.

[6] The constraint here has the opposite sign as compared to the univariate case, but it is coherent. In fact, it stems from the integral constraint and in the univariate case the integral is obtained by taking the difference between the value at the extrema of integration, while in the bivariate case the values at the top-right and bottom-left corners are added while those at the other two vertices of the rectangle are subtracted.

$(g+m+1)$ square matrix $\mathbf{K}^{11}$ is given by:

$$\mathbf{K}^{11} = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}. \tag{D.44}$$

For example, let I_R, I_C be the number of rows and columns in $\mathbf{C}$, respectively. Then the size of the matrix $\mathbf{K}^*$ is $(I_C(I_R-1)) \times (I_C I_R - 1)$.

By exploiting the previously defined diagonal matrices $\mathbf{D}^j, \mathbf{E}^j$, we define the following $(g+m+1)^2 \times (g+m+1)^2$ block diagonal matrices:

$$\mathbf{D} = \sum_{i=-m}^{g} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{D}^i = \left[\begin{array}{c|c|c|c} \mathbf{D}^{-m} & & & \\ \hline & \mathbf{D}^{-m-1} & & \\ \hline & & \ddots & \\ \hline & & & \mathbf{D}^{g} \end{array}\right], \tag{D.45}$$

$$\mathbf{E} = \sum_{i=-m}^{g} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{E}^i = \left[\begin{array}{c|c|c|c} \mathbf{E}^{-m} & & & \\ \hline & \mathbf{E}^{-m-1} & & \\ \hline & & \ddots & \\ \hline & & & \mathbf{E}^{g} \end{array}\right], \tag{D.46}$$

$$\mathbf{F} = \sum_{i=-m}^{g} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{E}^{i-1} = \left[\begin{array}{c|c|c|c} \mathbf{E}^{-m-1} & & & \\ \hline & \mathbf{E}^{-m} & & \\ \hline & & \ddots & \\ \hline & & & \mathbf{E}^{g-1} \end{array}\right]. \tag{D.47}$$

Finally, we define $\mathbf{T}^f, \mathbf{T}^l$ to be two selection matrices of size $(g+m+1)^2 \times (g+m+1)(g+m+2)$ which select entries from a vector of length $(g+m+1)(g+m+2)$ by pre-multiplication:

$$\mathbf{T}^f = \left[\begin{array}{ccc|ccc} 1 & & & 0 & \dots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & 1 & 0 & \dots & 0 \end{array}\right], \qquad \mathbf{T}^l = \left[\begin{array}{ccc|ccc} 0 & \dots & 0 & 1 & & \\ \vdots & & \vdots & & \ddots & \\ 0 & \dots & 0 & & & 1 \end{array}\right]. \tag{D.48}$$

Finally, define $\mathbf{A} = [\mathbf{E}\mathbf{T}^f - \mathbf{F}\mathbf{T}^l]\mathbf{K}^*$. We obtain the following equation relating the vectorized matrices of spline coefficients $\mathbf{B}$ and $\mathbf{C}$:

$$\bar{\mathbf{b}} = \operatorname{vec}(\mathbf{B}) = \mathbf{D}\left[\mathbf{E}\mathbf{T}^f\mathbf{K}^*\tilde{\mathbf{c}} - \mathbf{F}\mathbf{T}^l\mathbf{K}^*\tilde{\mathbf{c}}\right] = \mathbf{D}\left[\mathbf{E}\mathbf{T}^f - \mathbf{F}\mathbf{T}^l\right]\mathbf{K}^*\tilde{\mathbf{c}} = \mathbf{D}\mathbf{A}\tilde{\mathbf{c}}, \tag{D.49}$$

The next step consists in re-writing the objective function of the optimization problem (4.23) using matrix notation. First of all, since the B-spline basis for the bivariate spline is the product of two univariate B-splines, given a sample $(\mathbf{z}, \mathbf{u}, \mathbf{v}) = \{z_i, (u_i, v_i)\}_{i=1}^n$ we define the modified version of the matrix $\mathbf{C}^{m+1}(\mathbf{u})$ used in the univariate case as follows:

$$\mathbf{C}^{m+1}(\mathbf{u},\mathbf{v}) = \begin{bmatrix} B_{-m}^{m+1}(u_1)B_{-m}^{m+1}(v_1) & \dots & B_{g}^{m+1}(u_1)B_{-m}^{m+1}(v_1) & \dots & B_{g}^{m+1}(u_1)B_{g}^{m+1}(v_1) \\ \vdots & & & & \vdots \\ B_{-m}^{m+1}(u_n)B_{-m}^{m+1}(v_n) & \dots & B_{g}^{m+1}(u_n)B_{-m}^{m+1}(v_n) & \dots & B_{g}^{m+1}(u_n)B_{g}^{m+1}(v_n) \end{bmatrix}, \tag{D.50}$$

whose size is $n \times (g+m+1)^2$ and generic entry $\mathbf{C}_{i,j}^{m+1}(\mathbf{u},\mathbf{v}) = B_{j_1}^{m+1}(u_i)B_{j_2}^{m+1}(v_i)$, with j_1, j_2

obtained by inverting[7] the linear indexing $j = j_1 + (j_2 - 1)(g + m + 1)$. Each row is constructed by first fixing the index for the B-spline along the direction of v, that is $B^{m+1}_j(v_i)$, then considering all the combinations with the B-spline along the direction of u, that is $B^{m+1}_h(u_i)$. Then the index j is incremented and the process is iterated until the exhaustion of the basis. Notice that each row of the matrix corresponds to the same observation point. This construction is necessary to rewrite a bivariate spline function given the observation points $\{(u_i, v_i)\}_{i=1}^n$ in matrix form, obtaining a vector of size $n \times 1$:

$$s_m(\mathbf{u}, \mathbf{v}) = \mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v})\overline{\mathbf{b}}\,. \tag{D.51}$$

In order to write the matrix form of the integral of the squared derivative (of order $\ell \leq m - 1$) of the bivariate spline function, start by defining the $(g + m + 1 - \ell)^2 \times (g + m + 1 - \ell)^2$ matrix of inner products of the B-spline basis functions of order $m - \ell$ as follows:

$$\mathbf{M}_{m,\ell} = \begin{bmatrix} \langle B^{m+1-\ell}_{-m+\ell} B^{m+1-\ell}_{-m+\ell},\, B^{m+1-\ell}_{-m+\ell} B^{m+1-\ell}_{-m+\ell} \rangle & \dots & \langle B^{m+1-\ell}_{g} B^{m+1-\ell}_{g},\, B^{m+1-\ell}_{-m+\ell} B^{m+1-\ell}_{-m+\ell} \rangle \\ \vdots & & \vdots \\ \langle B^{m+1-\ell}_{-m+\ell} B^{m+1-\ell}_{-m+\ell},\, B^{m+1-\ell}_{g} B^{m+1-\ell}_{g} \rangle & \dots & \langle B^{m+1-\ell}_{g} B^{m+1-\ell}_{g},\, B^{m+1-\ell}_{g} B^{m+1-\ell}_{g} \rangle \end{bmatrix}, \tag{D.52}$$

where the generic entry is $\mathbf{M}_{m,\ell;i,j} = \langle B^{m+1-\ell}_{j_1} B^{m+1-\ell}_{j_2}, B^{m+1-\ell}_{i_1} B^{m+1-\ell}_{i_2} \rangle$, where i_1, i_2, j_1, j_2 are obtained, as for $\mathbf{C}^{m+1}_{i,j}(\mathbf{u}, \mathbf{v})$, by inverting the linear indexing $i = i_1 + (i_2 - 1)(g + m + 1 - \ell)$ and $j = j_1 + (j_2 - 1)(g + m + 1 - \ell)$. The inner product is defined in the usual way (see Algorithm 5.22 in Schumaker (2007) for numerical computation) as:

$$\begin{aligned} &\langle B^{m+1-\ell}_i B^{m+1-\ell}_j,\, B^{m+1-\ell}_h B^{m+1-\ell}_l \rangle \\ &= \int_{\lambda_0}^{\lambda_g} \int_{\lambda_0}^{\lambda_g} B^{m+1-\ell}_i(u)\, B^{m+1-\ell}_j(v)\, B^{m+1-\ell}_h(u)\, B^{m+1-\ell}_l(v)\;\; \mathrm{d}u\, \mathrm{d}v \geq 0\,. \end{aligned} \tag{D.53}$$

Under the assumptions made in the text, that is $(a_1, b_1) = (\lambda_{0,0}, \lambda_{g,g})$, $(a_2, b_2) = (\lambda_{u,0}, \lambda_{u,g})$, $\ell_1 = \ell_2 = \ell$, $n_1 = n_2 = n$, the objective function in eq. (4.23) can be re-written as the sum of two terms:

$$\begin{aligned} J_\ell(s_m) &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \left[s_m^{(\ell_1,\ell_2)}(u, v) \right]^2 \mathrm{d}v\, \mathrm{d}u + \alpha \left[\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{ij} \left(z_{ij} - s_m(u_i, v_j) \right)^2 \right] \\ &= \int_{\lambda_{0,0}}^{\lambda_{g,g}} \int_{\lambda_{u,0}}^{\lambda_{u,g}} \left[s_m^{(\ell,\ell)}(u, v) \right]^2 \mathrm{d}v\, \mathrm{d}u + \alpha \left[\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(z_{ij} - s_m(u_i, v_j) \right)^2 \right] \\ &= \int_{\lambda_{0,0}}^{\lambda_{g,g}} \int_{\lambda_{u,0}}^{\lambda_{u,g}} \left[s_m^{(\ell,\ell)}(u, v) \right]^2 \mathrm{d}v\, \mathrm{d}u + \alpha \left[\sum_{i'=1}^{n'} w_{i'} \left(z_{i'} - s_m(u_{i'}, v_{i'}) \right)^2 \right] \qquad \text{(D.54)} \\ &= J^1_\ell(s_m) + J^2_\ell(s_m)\,. \qquad \text{(D.55)} \end{aligned}$$

The double sum has been reduced to a single sum under the hypothesis that the sample consists of a value z_i and a point (u_i, v_i). The third line has been obtained after vectorization. The extrema of integration are the same as in the univariate case, but now it is necessary to

[7] The inversion is obtained by solving a linear system with two equations and two unknowns, j_1, j_2, which has a unique solution: $j_1 = j - (j_2 - 1)(g + m + 1)$ and $j_2 = \lfloor j/(g + m + 1) \rfloor$, where $\lfloor x \rfloor$ denote the integer part of x.

stress formally that when integrating with respect to v, the extrema of integration in principle may depend on u. The idea is nonetheless simple: the area of integration consists of the points included in the square with vertices $(\lambda_{0,0}, \lambda_{0,g}, \lambda_{g,0}, \lambda_{g,g})$. Concerning the derivative, by choosing $\ell_1 = \ell_2 = \ell = 2$ we are performing second-order derivative, thus obtaining a solution in the class of cubic spline functions.

Let $\mathbf{Z} = (z_{ij})_{ij}$, $\mathbf{W} = (w_{ij})_{ij}$ be $n \times n$ matrices and $\mathbf{z} = \mathrm{vec}\,(\mathbf{Z})$, $\mathbf{w} = \mathrm{vec}\,(\mathbf{W})$ be their vectorization. By exploiting eq. (D.51), the second component of eq. (D.55) can be written in matrix form as:

$$J^2_\ell(\overline{\mathbf{b}}) = \alpha \left(\mathbf{z} - \mathbf{C}^{m+1}(\mathbf{u},\mathbf{v})\overline{\mathbf{b}}\right)' \mathbf{W} \left(\mathbf{z} - \mathbf{C}^{m+1}(\mathbf{u},\mathbf{v})\overline{\mathbf{b}}\right) . \tag{D.56}$$

As for the first addendum, since the derivative of spline is another spline of lower degree, it can be represented in matrix form. In particular, it is given by the product of the vectorised coefficient matrix $\overline{\mathbf{b}}$ and a vector of B-spline basis functions $\mathbf{g}(u,v)$ defined as:

$$\mathbf{g}(u,v) = \left[B^{m+1}_{-m} B^{m+1}_{-m}(v), \dots, B^{m+1}_{g}(u) B^{m+1}_{-m}(v), \dots, B^{m+1}_{g}(u) B^{m+1}_{g}(v)\right] . \tag{D.57}$$

This allows to write:

$$\begin{aligned}
J^1_\ell(s_m) &= \int_{\lambda_{0,0}}^{\lambda_{g,g}} \int_{\lambda_{u,0}}^{\lambda_{u,g}} [s_m^{(\ell,\ell)}(u,v)]^2 \,\mathrm{d}v\,\mathrm{d}u = \int_{\lambda_{0,0}}^{\lambda_{g,g}} \int_{\lambda_{u,0}}^{\lambda_{u,g}} \overline{\mathbf{b}}^{(\ell)'} \mathbf{g}(u,v)'\mathbf{g}(u,v)\overline{\mathbf{b}}^{(\ell)} \,\mathrm{d}v\,\mathrm{d}u \\
&= \overline{\mathbf{b}}^{(\ell)'} \left[\int_{\lambda_{0,0}}^{\lambda_{g,g}} \int_{\lambda_{u,0}}^{\lambda_{u,g}} \mathbf{g}(u,v)'\mathbf{g}(u,v) \,\mathrm{d}v\,\mathrm{d}u\right] \overline{\mathbf{b}}^{(\ell)} \\
&= \overline{\mathbf{b}}^{(\ell)'} \mathbf{M}_{m,\ell} \overline{\mathbf{b}}^{(\ell)} = J^1_\ell(\overline{\mathbf{b}}) ,
\end{aligned} \tag{D.58}$$

where the last line follows from the definition of the matrix $\mathbf{M}_{m,\ell}$. We are left to find an explicit form for the vectorised coefficient matrix of the original spline of degree m, $\overline{\mathbf{b}}$, and that of its ℓ-th derivative, $\overline{\mathbf{b}}^{(\ell)}$. Recall the previous manipulation of the integral constraint gave a linear relation between $\overline{\mathbf{b}}$ and $\tilde{\mathbf{c}}$, with a restriction was accounted for in the construction of the matrix $\mathbf{K}^*$. In the problem at hand there no constraints, therefore we use the matrix $\mathbf{K}$ defined in eq. (D.36). Finally, notice that we can compute the ℓ-th derivative of $s_m(u,v)$ by simply iterating ℓ times the procedure previous used for the first order derivative, with a slight modification of the matrices involved. In fact, the size of the matrices need to shrink at each derivation step (in fact, the degree of a spline determines the length of its coefficient vector). Therefore, by indexing each matrix with a subscript corresponding to the order of the derivative, we obtain for the ℓ-th order derivative (similar to eq. (D.49)):

$$\overline{\mathbf{b}}^{(\ell)} = \mathbf{S}_\ell \overline{\mathbf{b}} = \left[\prod_{h=1}^{\ell} \mathbf{D}_h \left[\mathbf{E}_h \mathbf{T}^f_h - \mathbf{F}_h \mathbf{T}^l_h\right] \mathbf{K}_h\right] \overline{\mathbf{b}} . \tag{D.59}$$

All the definitions are provided below in eq. (D.60), (D.62), (D.63), (D.64), (D.65). Notice however that they are just simple generalisations of the matrices used when we dealing with the integral constraint: in fact in that case we were considering a first order derivative, while here we are considering a ℓ-th order derivative. The only significant difference consists in the substitution of the matrix $\mathbf{K}$ with the difference operator matrix $\mathbf{K}_h$ defined in eq. (D.62). Let $\{\mathbf{e}_{-m+h}, \dots, \mathbf{e}_g\}$ be the canonical basis of the space of $(g+m+1-h) \times (g+m+1-h)$ matrices, with $h = 1, \dots, \ell$. The block diagonal $(g+m+1-h)^2 \times (g+m+1-h)^2$ matrix

$\mathbf{D}_h$ is given by:

$$\mathbf{D}_h = \sum_{i=-m+h}^{g} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{D}_h^i = \left[\begin{array}{c|c|c|c} \mathbf{D}_h^{-m+h} & & & \\ \hline & \mathbf{D}_h^{-m+h+1} & & \\ \hline & & \ddots & \\ \hline & & & \mathbf{D}_h^g \end{array}\right], \tag{D.60}$$

where each block $\mathbf{D}_h^j$, for $j = -m+h, \ldots, g$, $h = 1, \ldots, \ell$, is a $(g+m+1-h) \times (g+m+1-h)$ diagonal matrix:

$$\mathbf{D}_h^j = \operatorname{diag}\left(\frac{(m+1-h)^2}{\lambda_{-m+h,j+m+1-h} - \lambda_{-m+h,j}}, \frac{(m+1-h)^2}{\lambda_{-m+h+1,j+m+1-h} - \lambda_{-m+h+1,j}}, \ldots, \frac{(m+1-h)^2}{\lambda_{g,j+m+1-h} - \lambda_{g,j}} \right). \tag{D.61}$$

Notice that when $\ell = 1$ and the original spline has degree $m+1$ we are back in the previous case. As previously noted, in dealing with derivatives without constraints, the matrix $\mathbf{K}$ can be substituted by the difference operator matrix defined in eq. (D.42). It has the same structure and entries for each $h = 1, \ldots, \ell$, but with different size. Since we are dealing with the iterative vectorisation of a matrix, we must account that at each derivative the last row of the original matrix is lost due to differentiation. This is reflected in a reduction of the number of rows $\mathbf{K}_h$ by a factor of hN at each step, where N is the number of columns of the original matrix. Finally, since differencing is performed iteratively, the vector to be differenced is the outcome of the previous iteration, hence its length (which is equal to the number of columns of $\mathbf{K}_h$) corresponds to the number of rows of $\mathbf{K}_h$ plus hN. Summarizing, for $h = 1, \ldots, \ell$ we define the $(g+m+1)(g+m+1-h) \times (g+m+1)(g+m+2-h)$ matrix:

$$\mathbf{K}_h = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}. \tag{D.62}$$

Similarly, she selection matrices $\mathbf{T}_h^f, \mathbf{T}_h^l$, with $h = 1, \ldots, \ell$, have the structure as $\mathbf{T}^f, \mathbf{T}^l$, but their size is $(g+m+1-h)^2 \times (g+m+1)(g+m+1-h)$. They are defined as follows:

$$\mathbf{T}_h^f = \left[\begin{array}{ccc|ccc} 1 & & & 0 & \ldots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & 1 & 0 & \ldots & 0 \end{array}\right] \qquad \mathbf{T}_h^l = \left[\begin{array}{ccc|ccc} 0 & \ldots & 0 & 1 & & \\ \vdots & & \vdots & & \ddots & \\ 0 & \ldots & 0 & & & 1 \end{array}\right]. \tag{D.63}$$

Finally, the matrices $\mathbf{E}$ and $\mathbf{F}$ are generalized to obtain the $(g+m+1-h)^2 \times (g+m+1-h)^2$ block diagonal matrices:

$$\mathbf{E}_h = \sum_{i=-m+h}^{g} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{E}_h^i = \left[\begin{array}{c|c|c|c} \mathbf{E}_h^{-m} & & & \\ \hline & \mathbf{E}_h^{-m-1} & & \\ \hline & & \ddots & \\ \hline & & & \mathbf{E}_h^g \end{array}\right], \tag{D.64}$$

$$\mathbf{F}_h = \sum_{i=-m+h}^{g} \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{E}_h^{i-1} = \left[\begin{array}{c|c|c|c} \mathbf{E}_h^{-m-1} & & & \\ \hline & \mathbf{E}_h^{-m} & & \\ \hline & & \ddots & \\ \hline & & & \mathbf{E}_h^{g-1} \end{array}\right], \tag{D.65}$$

where each $\mathbf{E}_h^j$, with $j = -m-1+h, \dots, g$, $h = 1, \dots, \ell$, is the $(g+m+1-h) \times (g+m+1-h)$ diagonal matrix:

$$\mathbf{E}_h^j = \operatorname{diag}\left(\frac{1}{\lambda_{1,j} - \lambda_{-m+h,j}}, \frac{1}{\lambda_{2,j} - \lambda_{-m+h+1,j}}, \dots, \frac{1}{\lambda_{g+m+1-h,j} - \lambda_{g,j}} \right) . \tag{D.66}$$

To sum up, we can re-write the first addendum of the objective function of the optimization problem in eq. (D.55) in compact form as follows:

$$J_\ell^1(\overline{\mathbf{b}}) = \overline{\mathbf{b}}^{(\ell)'} \mathbf{M}_{m,\ell} \overline{\mathbf{b}}^{(\ell)} = \overline{\mathbf{b}}' \mathbf{S}_\ell' \mathbf{M}_{m,\ell} \mathbf{S}_\ell \overline{\mathbf{b}} \, . \tag{D.67}$$

Putting together eq. (D.67) and (D.56) we obtain the following matrix representation of the objective function of the optimisation problem in eq. (4.23):

$$J_\ell(\overline{\mathbf{b}}) = \overline{\mathbf{b}}' \mathbf{S}_\ell' \mathbf{M}_{m,\ell} \mathbf{S}_\ell \overline{\mathbf{b}} + \alpha \left(\mathbf{z} - \mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v}) \overline{\mathbf{b}} \right)' \mathbf{W} \left(\mathbf{z} - \mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v}) \overline{\mathbf{b}} \right) . \tag{D.68}$$

We can now exploit the linear relation obtained in eq. (D.49) by working out the integral constraint and substitute it in eq. (D.68). This transforms the constrained optimization problem in eq. (4.23) into an unconstrained optimisation problem for $\tilde{\mathbf{c}}$, with objective function:

$$J_\ell(\tilde{\mathbf{c}}) = \tilde{\mathbf{c}}' \mathbf{A}' \mathbf{D}' \mathbf{S}_\ell' \mathbf{M}_{m,\ell} \mathbf{S}_\ell \mathbf{D} \mathbf{A} \tilde{\mathbf{c}} + \alpha \left(\mathbf{z} - \mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v}) \mathbf{D} \mathbf{A} \tilde{\mathbf{c}} \right)' \mathbf{W} \left(\mathbf{z} - \mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v}) \mathbf{D} \mathbf{A} \tilde{\mathbf{c}} \right) . \tag{D.69}$$

The system of first order necessary conditions for an optimum is obtained from:

$$\begin{aligned} \frac{\mathrm{d}J_\ell(\tilde{\mathbf{c}})}{\mathrm{d}\tilde{\mathbf{c}}'} &= 2\mathbf{A}'\mathbf{D}'\mathbf{S}_\ell'\mathbf{M}_{m,\ell}\mathbf{S}_\ell\mathbf{D}\mathbf{A}\tilde{\mathbf{c}} - 2\alpha\mathbf{A}'\mathbf{D}'\mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{z} \\ &+ 2\alpha\mathbf{A}'\mathbf{D}'\mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v})\mathbf{D}\mathbf{A}\tilde{\mathbf{c}} = 0 \, . \end{aligned} \tag{D.70}$$

Define the following variables for easing the notation:

$$\begin{aligned} \mathbf{N}_{m,\ell} &= \mathbf{A}'\mathbf{D}'\mathbf{S}_\ell'\mathbf{M}_{m,\ell}\mathbf{S}_\ell\mathbf{D}\mathbf{A} \, , \\ \mathbf{H}(\mathbf{u}, \mathbf{v}) &= \mathbf{C}^{m+1}(\mathbf{u}, \mathbf{v})\mathbf{D}\mathbf{A} \, . \end{aligned}$$

Therefore, one gets:

$$\begin{aligned} \mathbf{N}_{m,\ell}\tilde{\mathbf{c}} + \alpha\mathbf{H}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{H}(\mathbf{u}, \mathbf{v})\tilde{\mathbf{c}} &= \alpha\mathbf{H}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{z} \\ \left[\mathbf{N}_{m,\ell} + \alpha\mathbf{H}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{H}(\mathbf{u}, \mathbf{v})\right] \tilde{\mathbf{c}} &= \alpha\mathbf{H}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{z} \, . \end{aligned} \tag{D.71}$$

If the condition of the Rouché-Capelli theorem for the system to admit solution is satisfied and the matrix $\left[\mathbf{N}_{m,\ell} + \alpha\mathbf{H}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{H}(\mathbf{u}, \mathbf{v})\right]$ has full rank, then the system has a unique solution $\tilde{\mathbf{c}}^*$. By contrast, if the matrix is singular the problem admits an infinite number of solutions which can be obtained by computing the Moore-Penrose pseudo-inverse (denoted by †):

$$\tilde{\mathbf{c}}^* = \alpha \left[\mathbf{N}_{m,\ell} + \alpha\mathbf{H}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{H}(\mathbf{u}, \mathbf{v})\right]^\dagger \mathbf{H}(\mathbf{u}, \mathbf{v})'\mathbf{W}\mathbf{z} \, , \tag{D.72}$$

with a slight abuse of notation. Among this set of solutions, we choose the one with smallest norm. As a final step, we use eq. (D.49) and plug-in the optimal value of $\tilde{\mathbf{c}}$ for obtaining the optimal value of the coefficients $\overline{\mathbf{b}}^*$:

$$\overline{\mathbf{b}}^* = \mathbf{D}\mathbf{A}\tilde{\mathbf{c}}^* \, . \tag{D.73}$$

D.2.2 Eigenproblem

In the following we show the computations required for obtaining eq. (4.29). In this section, differently from the previous ones, we explicitly denote all the arguments of a function for making notation clearer. Thus, for example, we have that $\check{f}_t(\cdot) = \check{f}_t(\cdot,\cdot)$, where the first is follows the notation of the previous sections, while the second is according to the notation of this section. We use the standard estimator for the sample covariance (alternative non-parametric estimators have been proposed by Hall et al. (2006), Li and Hsing (2010), Yao et al. (2005) and Staniswalis and Lee (1998)), that is:

$$v(l_1, m_1, l_2, m_2) = \frac{1}{T}\sum_{t=1}^{T} \check{f}_t(l_1, m_1)\check{f}_t(l_2, m_2)\,. \tag{D.74}$$

Then, the eigenproblem can be formulated as follows, for $j = 1, 2, \ldots$:

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} v(\cdot,\cdot,l_2,m_2)\xi_j(l_2,m_2)\ \mathrm{d}l_2\,\mathrm{d}m_2 = \rho_j\xi_j(\cdot,\cdot)\,. \tag{D.75}$$

In order to solve this problem, we choose to express the eigenfunctions as finite linear combinations of the same set of basis functions used for the functions $\check{f}_t(\cdot,\cdot)$, that is the basis B-spline functions $\boldsymbol{\psi}(\cdot,\cdot) = (\psi_1(\cdot,\cdot),\ldots,\psi_K(\cdot,\cdot))'$ in eq. (4.26). Define the coefficient vectors $\mathbf{a}_j = (a_{j,1},\ldots,a_{j,K})'$. To summarize, we have:

$$\check{f}_t(\cdot,\cdot) = \mathbf{d}_t'\boldsymbol{\psi}(\cdot,\cdot) = \sum_{k=1}^{K} d_{t,k}\psi_k(\cdot,\cdot)\,, \tag{D.76}$$

$$\xi_j(\cdot,\cdot) = \mathbf{a}_j'\boldsymbol{\psi}(\cdot,\cdot) = \sum_{k=1}^{K} a_{j,k}\psi_k(\cdot,\cdot)\,. \tag{D.77}$$

By stacking all data together in $\check{\mathbf{f}}(\cdot,\cdot) = (\check{f}_1(\cdot,\cdot),\ldots,\check{f}_T(\cdot,\cdot))'$ and $\mathbf{D} = (\mathbf{d}_1,\ldots,\mathbf{d}_T)$, we obtain:

$$\mathbf{f}(\cdot,\cdot) = \mathbf{D}\boldsymbol{\psi}(\cdot,\cdot)\,. \tag{D.78}$$

We can thus rewrite eq. (D.75) in matrix notation:

$$\frac{1}{T}\int_a^b\int_a^b \boldsymbol{\psi}(\cdot,\cdot)'\mathbf{D}'\mathbf{D}\boldsymbol{\psi}(l_2,m_2)\boldsymbol{\psi}(l_2,m_2)'\mathbf{a}_j\ \mathrm{d}l_2\,\mathrm{d}m_2 = \rho_j\boldsymbol{\psi}(\cdot,\cdot)'\mathbf{a}_j \tag{D.79}$$

$$\frac{1}{T}\boldsymbol{\psi}(\cdot,\cdot)'\mathbf{D}'\mathbf{D}\left[\int_a^b\int_a^b \boldsymbol{\psi}(l_2,m_2)\boldsymbol{\psi}(l_2,m_2)'\ \mathrm{d}l_2\,\mathrm{d}m_2\right]\mathbf{a}_j = \rho_j\boldsymbol{\psi}(\cdot,\cdot)'\mathbf{a}_j\,, \tag{D.80}$$

then define the matrix of inner products:

$$\mathbf{M} = \int_a^b\int_a^b \boldsymbol{\psi}(l_2,m_2)\boldsymbol{\psi}(l_2,m_2)'\ \mathrm{d}l_2\,\mathrm{d}m_2 = \begin{bmatrix} \langle\psi_1(\cdot,\cdot),\psi_1(\cdot,\cdot)\rangle & \ldots & \langle\psi_1(\cdot,\cdot),\psi_K(\cdot,\cdot)\rangle \\ \vdots & & \vdots \\ \langle\psi_K(\cdot,\cdot),\psi_1(\cdot,\cdot)\rangle & \ldots & \langle\psi_K(\cdot,\cdot),\psi_K(\cdot,\cdot)\rangle \end{bmatrix} \tag{D.81}$$

$$\langle\psi_i(\cdot,\cdot),\psi_j(\cdot,\cdot)\rangle = \int_a^b\int_a^b \psi_i(l,m)\psi_j(l,m)\ \mathrm{d}l\,\mathrm{d}m\,, \tag{D.82}$$

thus obtaining:

$$T^{-1}\mathbf{D}'\mathbf{D}\mathbf{M}\mathbf{a}_j = \rho_j\mathbf{a}_j\,. \tag{D.83}$$

In order to obtain a positive semi-definite matrix, we apply the linear transformation $\mathbf{u}_j = \mathbf{M}^{1/2}\mathbf{a}_j$, where $\mathbf{A}^{1/2}$ is the principal square root of the positive definite matrix $\mathbf{A}$. Then, re-write the previous equation as an eigenproblem for $\mathbf{u}_j$ as follows:

$$T^{-1}\mathbf{M}^{1/2}\mathbf{D}'\mathbf{D}\mathbf{M}^{1/2}\mathbf{u}_j = \rho_j \mathbf{u}_j\,, \tag{D.84}$$

which is a standard multivariate eigenproblem for the matrix $T^{-1}\mathbf{M}^{1/2}\mathbf{D}'\mathbf{D}\mathbf{M}^{1/2}$. The number of components to take, J, is determined by the fraction of variability explained: we sort the estimated eigenvalues $\widehat{\rho}_j$, for $j = 1, 2, \ldots$, in decreasing order. Then, we fix a threshold $\bar{d}$ and retain all the pairs of eigenvalues and eigenvectors until the corresponding cumulated proportion of explained variability reaches, that is $J = \arg\min_j \{\sum_j \widehat{\rho}_j \geq \bar{d}\}$.

D.3 Additional plots

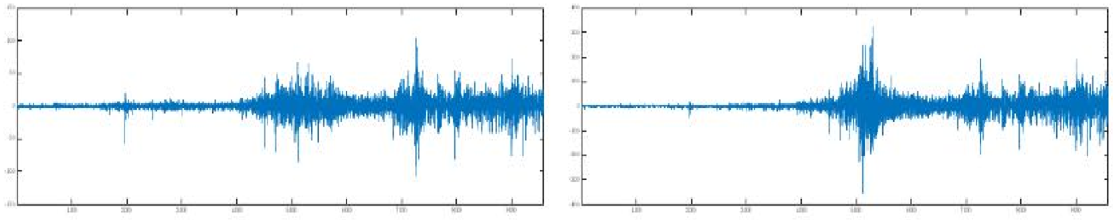


FIGURE D.1: First differenced series of S&P500 (*left*) and NASDAQ (*right*).

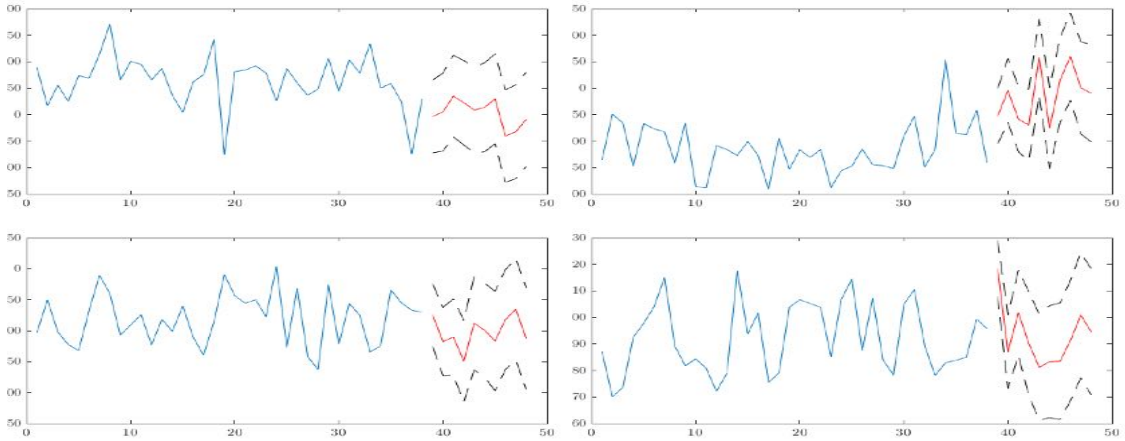


FIGURE D.2: Estimated time series (*solid, blue*) and forecast (*solid, red*) with 95% confidence intervals (*dashed, black*) of each entry of the vector of fPCA scores $\{\widehat{\boldsymbol{\beta}}_t\}_t$,from $j = 1$ (*top left*) to $j = J$ (*bottom*).

D.4 Bandwidth selection

In this section we present the estimation results under a different specification of the bandwidth parameter m. We defined two equally spaced discrete grids, $\mathcal{M}_1^n$ between 0.01 and 0.1 and $\mathcal{M}_2^n$ between 0.01 and 0.3, and for each grid we specify a length of $n = 6$ or $n = 12$ points. Therefore, we have in total four different cases $\mathcal{M}_1^6, \mathcal{M}_1^{12}, \mathcal{M}_2^6, \mathcal{M}_2^{12}$. We estimated the copula pdf in each period $t = 1, \ldots, T$ for each value of m in each grid and, then we compute the optimal bandwidth based on least squares cross validation method (LSCV).

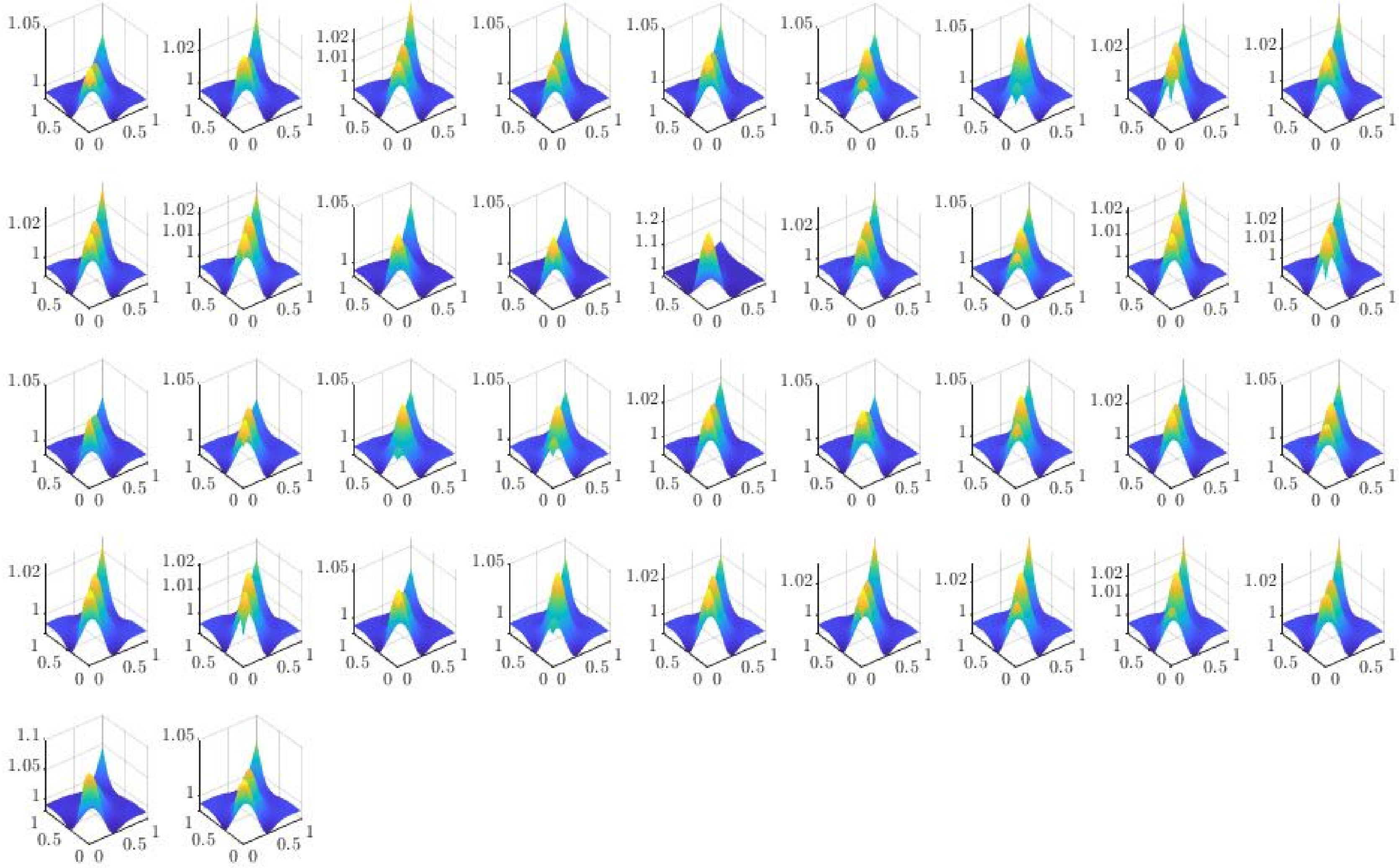


FIGURE D.3: 3D density plot of time series of bivariate copula pdfs, approximated via fPCA, for each year $t = 1, \dots, T$, starting from $t = 1$ in the top-left panel.

The least squares cross validation criterion (see Silverman (1986), Wand and Jones (1994)) to be minimized is given by:

$$LSCV(m) = \int \widehat{f}_m(\mathbf{y})\,\mathrm{d}\mathbf{y} - \frac{2}{N}\sum_{i=1}^{N}\widehat{f}_{m;-i}(\mathbf{x}_i)\,, \tag{D.85}$$

where

$$\widehat{f}_{m;-i}(\mathbf{x}) = \frac{1}{N-1}\sum_{j\neq i}^{N}\mathcal{K}_m(\mathbf{x}-\mathbf{y}_j) \tag{D.86}$$

is the leave-one-out estimate of the pdf. By minimizing the $LSCV(m)$ criterion for each period, we obtain $T = 38$ bandwidths. Finally, we choose their mean as the value of m to be used in the application (the results are similar for all the four grids).

We checked also the results of the procedure for $m = 0.05$ and Figs. D.4-D.10 report the corresponding outcome. We considered also cases for values higher values of the bandwidth above 0.05, by using $m = 0.07$ and $m = 0.11$. The results highlighted that such choices lead to clear oversmoothing, thus we not report here the plots.

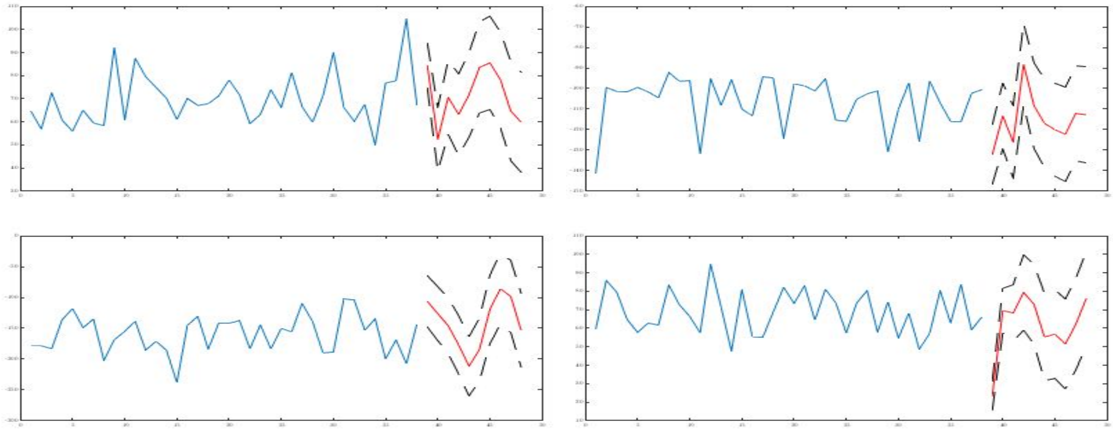


FIGURE D.6: Estimated time series (*solid, blue*) and forecast (*solid, red*) with 95% confidence intervals (*dashed, black*) of each entry of the vector of fPCA scores $\{\widehat{\boldsymbol{\beta}}_t\}_t$,from $j = 1$ (*top left*) to $j = 4$ (*bottom*).

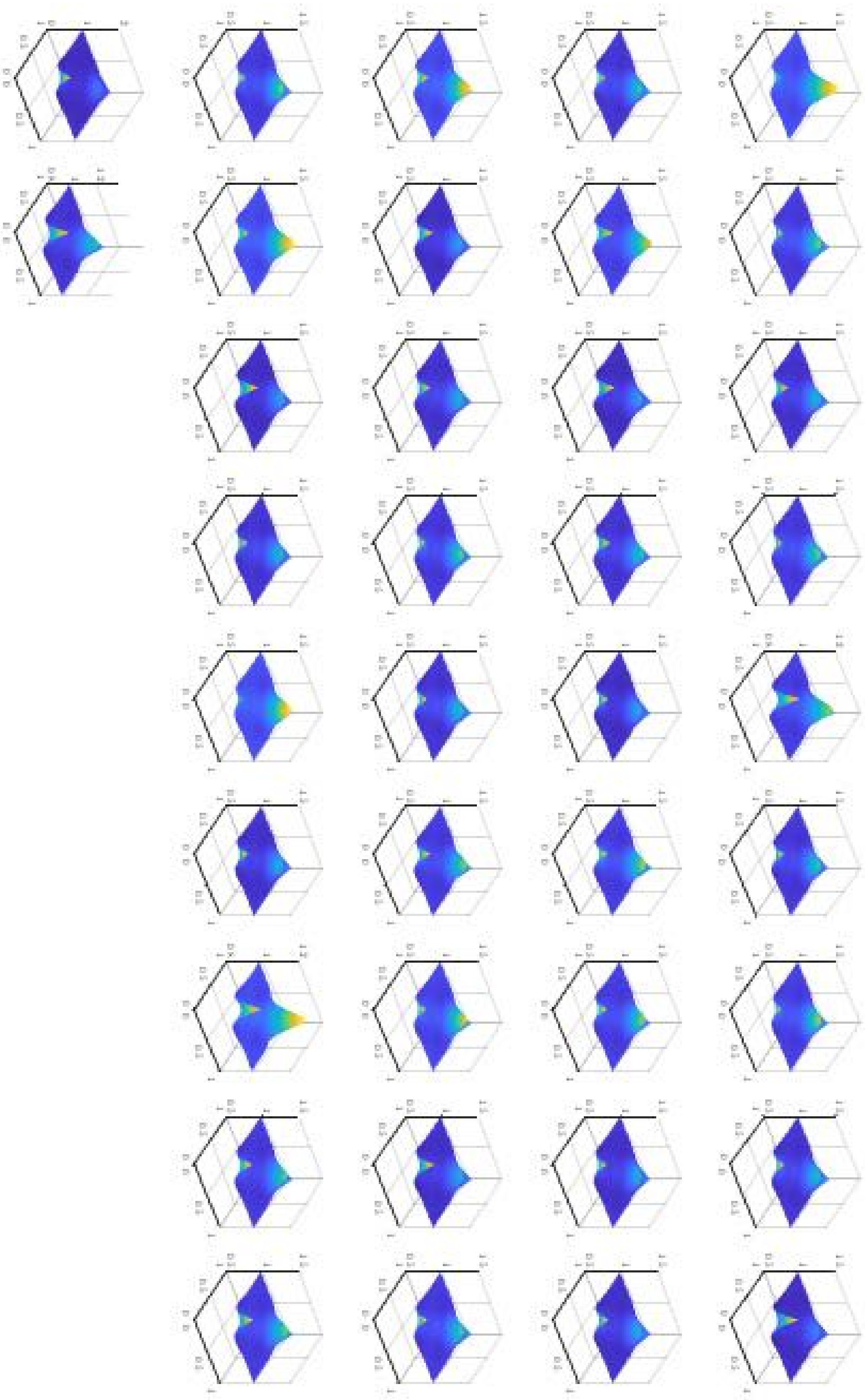


FIGURE D.4: 3D density plot of time series of bivariate copula pdfs, approximated via fPCA, for each year $t = 1, \dots, T$, starting from $t = 1$ in the top-left panel.

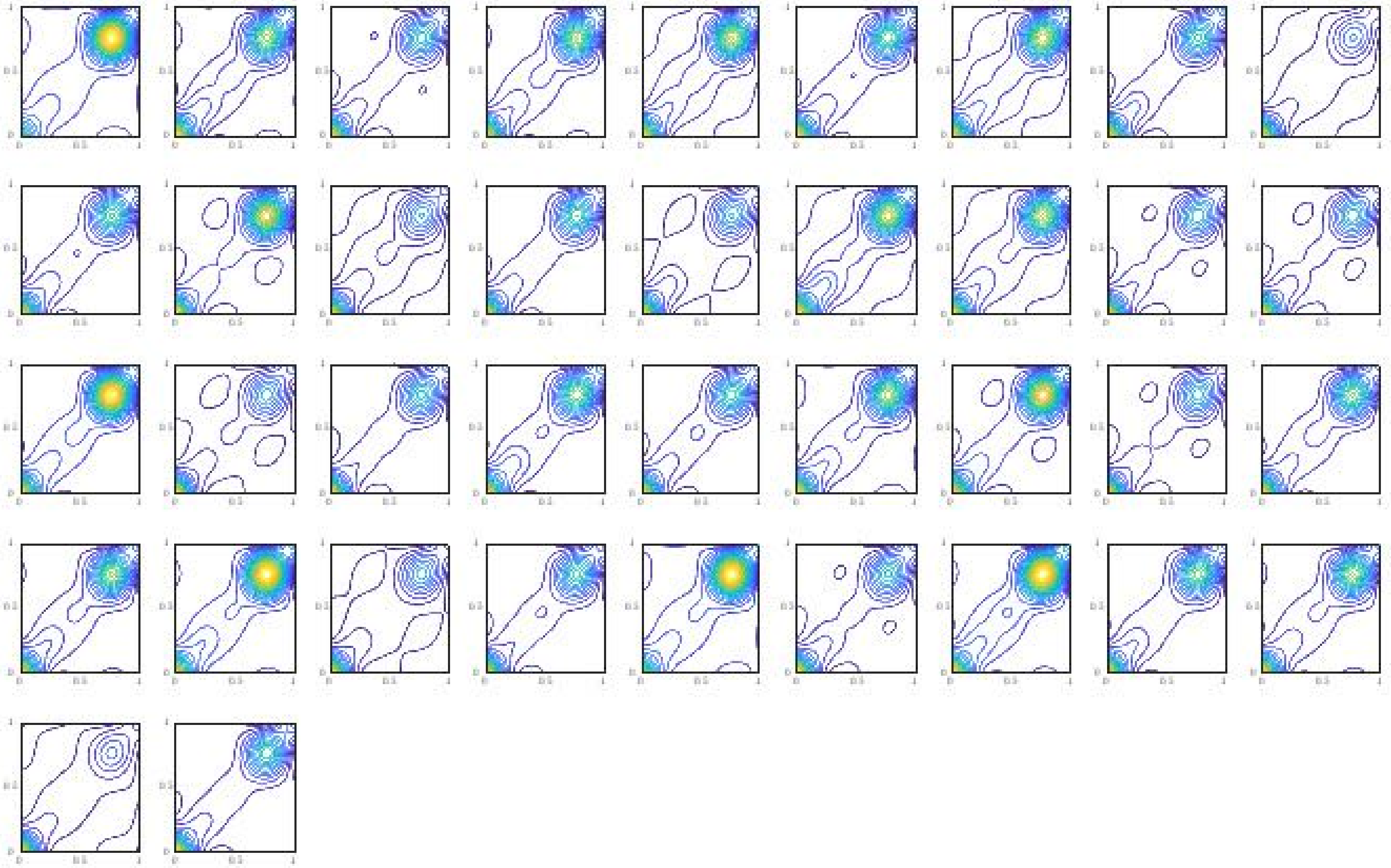


FIGURE D.5: Contour plot of time series of bivariate copula pdfs, approximated via fPCA, for each year $t = 1, \dots, T$, starting from $t = 1$ in the top-left panel.

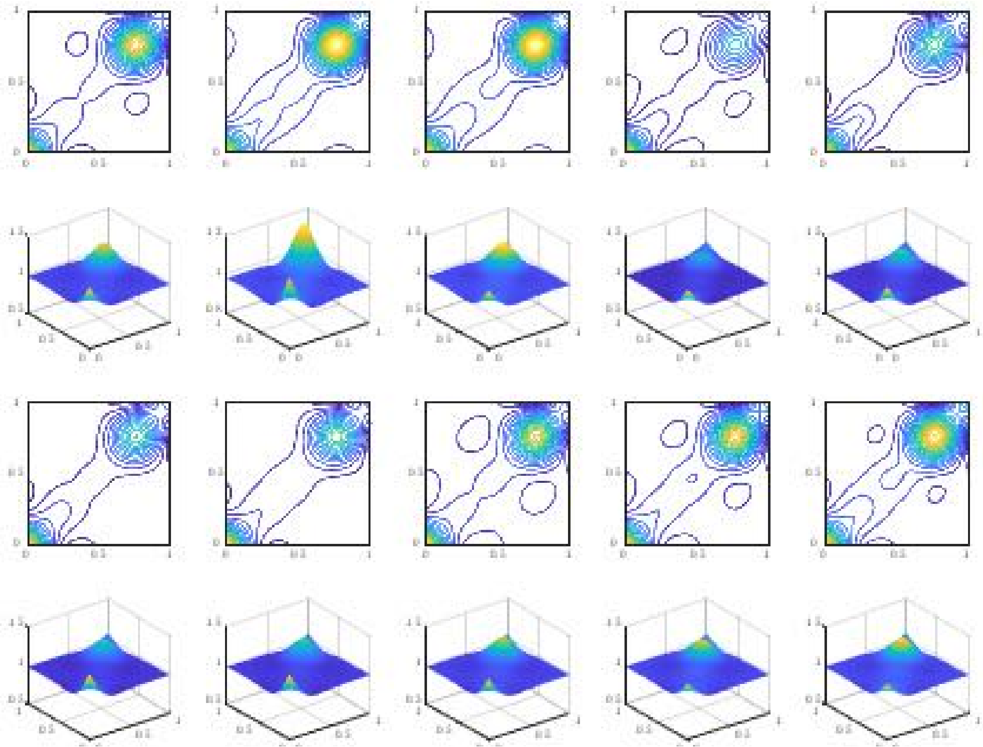


FIGURE D.7: Contour plots (*first* and *third* row) and the corresponding 3D density plot (*second* and *fourth* row) of the forecasted bivariate copula pdfs, approximated via fPCA, for each horizon $h = 1, \dots, 5$ (first and second rows) and $h = 6, \dots, 10$ (third and fourth rows), starting from the top-left panel.

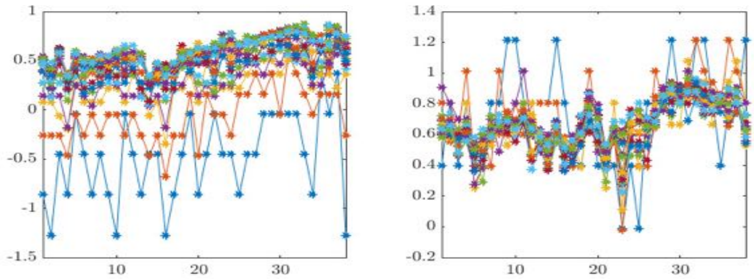


FIGURE D.8: Upper (*left*) and lower (*right*) tail dependence coefficients of the bivariate time series $(\mathbf{x}_t, \mathbf{y}_t)$, for $t = 1, \dots, 38$ (*x-axis*). Each curve corresponds to a different threshold $u = 0.01, 0.02, \dots, 0.20$.

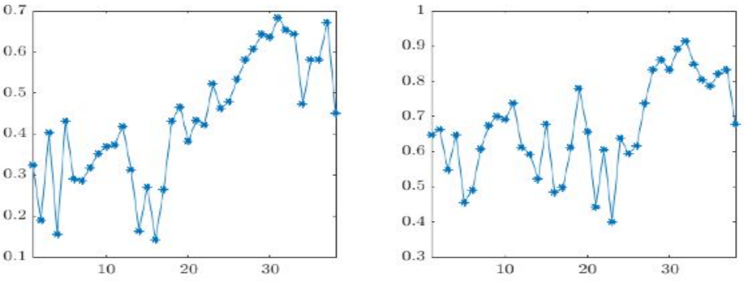


FIGURE D.9: Upper (*left*) and lower (*right*) tail dependence coefficients of the bivariate time series $(\mathbf{x}_t, \mathbf{y}_t)$, for $t = 1, \dots, 38$ (*x-axis*), threshold $u = 0.10$.

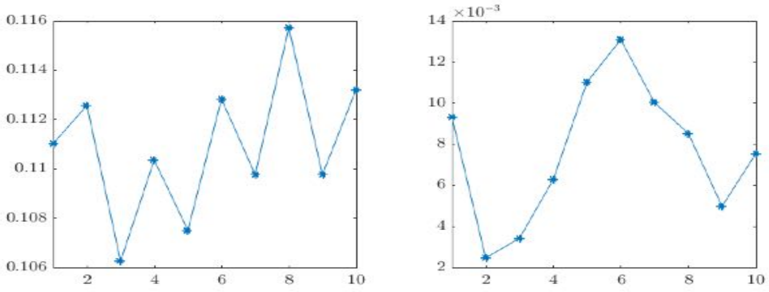


FIGURE D.10: Upper (*left*) and lower (*right*) tail dependence coefficients of the forecasted bivariate copula pdf $c_{T+h}(\cdot)$, for $h = 1, \ldots, 10$ (*x-axis*), threshold $u = 0.10$.

Bibliography

ABDI, H. (2003): "Factor rotations in factor analyses," *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, 792–795.

ABRAHAM, R., J. E. MARSDEN, AND T. RATIU (2012): *Manifolds, tensor analysis, and applications*, Springer Science & Business Media.

ACAR, E., S. A. ÇAMTEPE, M. S. KRISHNAMOORTHY, AND B. YENER (2005): "Modeling and multiway analysis of chatroom tensors," in *International conference on Intelligence and Security Informatics*, Springer, 256–268.

ACAR, E., S. A. ÇAMTEPE, AND B. YENER (2006): "Collective sampling and analysis of high order tensors for chatroom communications," in *International conference on Intelligence and Security Informatics*, Springer, 213–224.

ACEMOGLU, D., V. M. CARVALHO, A. OZDAGLAR, AND A. TAHBAZ-SALEHI (2012): "The network origins of aggregate fluctuations," *Econometrica*, 80, 1977–2016.

ACEMOGLU, D., A. MALEKIAN, AND A. OZDAGLAR (2016): "Network security and contagion," *Journal of Economic Theory*, 166, 536–585.

ACEMOGLU, D., A. OZDAGLAR, AND A. TAHBAZ-SALEHI (2015): "Systemic risk and stability in financial networks," *The American Economic Review*, 105, 564–608.

ADLER, R., M. BAZIN, AND M. SCHIFFER (1975): *Introduction to general relativity*, McGraw-Hill New York.

AHELEGBEY, D. F., M. BILLIO, AND R. CASARIN (2016a): "Bayesian graphical models for structural vector autoregressive processes," *Journal of Applied Econometrics*, 31, 357–386.

——— (2016b): "Sparse graphical vector autoregression: a Bayesian approach," *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, 333–361.

AIROLDI, E. M., D. M. BLEI, S. E. FIENBERG, AND E. P. XING (2008): "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014.

AITCHISON, J. (1986): *The statistical analysis of compositional data*, Chapman & Hall.

ALDASORO, I. AND I. ALVES (2016): "Multiplex interbank networks and systemic importance: an application to European data," *Journal of Financial Stability*.

ALMEIDA, C., C. CZADO, AND H. MANNER (2016): "Modeling high-dimensional time-varying dependence using dynamic D-vine models," *Applied Stochastic Models in Business and Industry*, 32, 621–638.

ANACLETO, O. AND C. QUEEN (2017): "Dynamic chain graph models for time series network data," *Bayesian Analysis*, 12, 491–509.

ARASHI, M. (2017): "Some theoretical results on tensor elliptical distribution," *arXiv preprint arXiv:1709.00801*.

ARIS, R. (2012): *Vectors, tensors and the basic equations of fluid mechanics*, Courier Corporation.

ATKINSON, K. (2009): *The numerical solution of integral equations of the second kind*, Cambridge Monographs on Applied and Computational Mathematics 4, Cambridge University Press.

ATKINSON, K. AND W. HAN (2005): *Theoretical numerical analysis*, Springer.

AUE, A., D. D. NORINHO, AND S. HÖRMANN (2015): "On the prediction of stationary functional time series," *Journal of the American Statistical Association*, 110, 378–392.

BALABANIS, G. AND A. DIAMANTOPOULOS (2004): "Domestic country bias, country-of-origin effects, and consumer ethnocentrism: a multidimensional unfolding approach," *Journal of the Academy of Marketing Science*, 32, 80.

BALAZSI, L., L. MATYAS, AND T. WANSBEEK (2015): "The estimation of multidimensional fixed effects panel data models," *Econometric Reviews*, 1–23.

BANERJEE, O., L. E. GHAOUI, AND A. D'ASPREMONT (2008): "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *Journal of Machine learning research*, 9, 485–516.

BARABÁSI, A.-L. AND R. ALBERT (1999): "Emergence of scaling in random networks," *Science*, 286, 509–512.

BARGIGLI, L., G. DI IASIO, L. INFANTE, F. LILLO, AND F. PIEROBON (2015): "The multiplex structure of interbank networks," *Quantitative Finance*, 15, 673–691.

BARIGOZZI, M. AND C. BROWNLEES (2016): "NETS: network estimation for time series," Tech. rep., Department of Economics and Business, Universitat Pompeu Fabra.

BARIGOZZI, M., G. FAGIOLO, AND G. MANGIONI (2011): "Identifying the community structure of the international-trade multi-network," *Physica A: statistical mechanics and its applications*, 390, 2051–2066.

BARRAT, A., B. FERNANDEZ, K. K. LIN, AND L.-S. YOUNG (2013): "Modeling temporal networks using random itineraries," *Physical Review Letters*, 110.

BARTRAM, S. M., S. J. TAYLOR, AND Y.-H. WANG (2007): "The Euro and European financial market dependence," *Journal of Banking & Finance*, 31, 1461–1481.

BATTISTON, F., V. NICOSIA, M. CHAVEZ, AND V. LATORA (2017): "Multilayer motif analysis of brain networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27, 047404.

BATTISTON, S., M. PULIGA, R. KAUSHIK, P. TASCA, AND G. CALDARELLI (2012): "Debtrank: too central to fail? Financial networks, the FED and systemic risk," *Scientific reports*, 2, 541.

BATTISTON, S., J. F. RODRIGUES, AND H. ZEYTINOGLU (2007): "The network of interregional direct investment stocks across Europe," *Advances in Complex Systems*, 10, 29–51.

BAZZI, M., M. A. PORTER, S. WILLIAMS, M. MCDONALD, D. J. FENN, AND S. D. HOWISON (2016): "Community detection in temporal multilayer networks, with an application to correlation networks," *Multiscale Modeling & Simulation*, 14, 1–41.

BECKMANN, C. F. AND S. M. SMITH (2005): "Tensorial extensions of independent component analysis for multisubject fMRI analysis," *Neuroimage*, 25, 294–311.

BEDFORD, T. AND R. M. COOKE (2002): "Vines: a new graphical model for dependent random variables," *The Annals of Statistics*, 30, 1031–1068.

BETANCOURT, B., A. RODRÍGUEZ, AND N. BOYD (2017): "Bayesian fused lasso regression for dynamic binary networks," *Journal of Computational and Graphical Statistics*, 26, 840–850.

BHATTACHARYA, A., D. PATI, N. S. PILLAI, AND D. B. DUNSON (2015): "Dirichlet-Laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, 110, 1479–1490.

BIANCHI, D., M. BILLIO, R. CASARIN, AND M. GUIDOLIN (2018): "Modeling systemic risk with Markov switching graphical SUR models," *Journal of Econometrics*.

BILLIO, M., M. CAPORIN, R. PANZICA, AND L. PELIZZON (2015a): "Network connectivity and systematic risk," Tech. rep., European Financial Management Association.

BILLIO, M., R. CASARIN, F. RAVAZZOLO, AND H. K. VAN DIJK (2016): "Interactions between Eurozone and US booms and busts: a Bayesian panel Markov-switching VAR model," *Journal of Applied Econometrics*, 31, 1352–1370.

BILLIO, M., M. GETMANSKY, D. GRAY, A. LO, R. MERTON, AND L. PELIZZON (2015b): "Sovereign, bank, and insurance credit spreads: connectedness and system networks," *MIT Working paper*.

BILLIO, M., M. GETMANSKY, A. W. LO, AND L. PELIZZON (2012): "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *Journal of Financial Economics*, 104, 535–559.

BLISS, R. R. AND N. PANIGIRTZOGLOU (2002): "Testing the stability of implied probability density functions," *Journal of Banking & Finance*, 26, 381–422.

BOCCALETTI, S., G. BIANCONI, R. CRIADO, C. I. DEL GENIO, J. GÓMEZ-GARDENES, M. ROMANCE, I. SENDINA-NADAL, Z. WANG, AND M. ZANIN (2014): "The structure and dynamics of multilayer networks," *Physics Reports*, 544, 1–122.

BOLLEN, K. A., S. RAY, J. ZAVISCA, AND J. J. HARDEN (2012): "A comparison of Bayes factor approximation methods including two new methods," *Sociological Methods & Research*, 41, 294–324.

BOLLOBÁS, B. (2012): *Graph theory: an introductory course*, vol. 63, Springer Science & Business Media.

——— (2013): *Modern graph theory*, vol. 184, Springer Science & Business Media.

BONANNO, G., G. CALDARELLI, F. LILLO, S. MICCICHE, N. VANDEWALLE, AND R. N. MANTEGNA (2004): "Networks of equities in financial markets," *The European Physical Journal B*, 38, 363–371.

BORGATTI, S. P., A. MEHRA, D. J. BRASS, AND G. LABIANCA (2009): "Network analysis in the social sciences," *Science*, 323, 892–895.

BORGS, C., J. T. CHAYES, H. COHN, AND N. HOLDEN (2016): "Sparse exchangeable graphs and their limits via graphon processes," *arXiv preprint arXiv:1601.07134*.

BOSQ, D. (2000): *Linear processes in function spaces: theory and applications*, vol. 149, Springer Science & Business Media.

BOSSAERTS, P. AND C. MURAWSKI (2015): "From behavioural economics to neuroeconomics to decision neuroscience: the ascent of biology in research on human decision making," *Current Opinion in Behavioral Sciences*, 5, 37–42.

BRÄUNING, F. AND S. J. KOOPMAN (2016): "The dynamic factor network model with an application to global credit risk," Tech. rep., Federal Reserve Bank of Boston.

BRILLINGER, D. R. (2001): *Time series: data analysis and theory*, vol. 36, SIAM.

BROWNLEES, C., E. NUALART, AND Y. SUN (2017): "Realized networks," *Available at SSRN: https://ssrn.com/abstract=2506703*.

CAI, D., T. CAMPBELL, AND T. BRODERICK (2016): "Edge-exchangeable graphs and sparsity," in *Neural Information Processing Systems*, 4242–4250.

CAIMO, A. AND N. FRIEL (2011): "Bayesian inference for exponential random graph models," *Social Networks*, 33, 41–55.

CALVO-ARMENGOL, A. AND M. O. JACKSON (2004): "The effects of social networks on employment and inequality," *The American Economic Review*, 94, 426–454.

CAMASTRA, F. (2003): "Data dimensionality estimation methods: a survey," *Pattern recognition*, 36, 2945–2954.

CAMERER, C., G. LOEWENSTEIN, AND D. PRELEC (2005): "Neuroeconomics: How neuroscience can inform economics," *Journal of Economic Literature*, 43, 9–64.

CANALE, A. AND M. RUGGIERO (2016): "Bayesian nonparametric forecasting of monotonic functional time series," *arXiv preprint arXiv:1608.08056*.

CANALE, A. AND S. VANTINI (2016): "Constrained functional time series: applications to the Italian gas market," *International Journal of Forecasting*, 32, 1340–1351.

CANOVA, F. AND M. CICCARELLI (2004): "Forecasting and turning point predictions in a Bayesian panel VAR model," *Journal of Econometrics*, 120, 327–359.

——— (2009): "Estimating multicountry VAR models," *International Economic Review*, 50, 929–959.

——— (2013): *Panel vector autoregressive models: a survey*, Emerald Group Publishing Limited, vol. 32, chap. 12, 205–246, VAR models in macroeconomics – new developments and applications: essays in honor of Christopher A. Sims ed.

CANOVA, F., M. CICCARELLI, AND E. ORTEGA (2007): "Similarities and convergence in G-7 cycles," *Journal of Monetary Economics*, 54, 850–878.

——— (2012): "Do institutional changes affect business cycles? Evidence from Europe," *Journal of Economic Dynamics and Control*, 36, 1520–1533.

CAPORIN, M., L. PELIZZON, F. RAVAZZOLO, AND R. RIGOBON (2017): "Measuring sovereign contagion in Europe," *Journal of Financial Stability*.

CARON, F. AND E. B. FOX (2017): "Sparse graphs using exchangeable random measures," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 1295–1366.

CARRIERO, A., G. KAPETANIOS, AND M. MARCELLINO (2016): "Structural analysis with multivariate autoregressive index models," *Journal of Econometrics*, 192, 332–348.

CARROLL, J. D. AND J.-J. CHANG (1970): "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," *Psychometrica*, 35, 283–319.

CARVALHO, C. M., H. MASSAM, AND M. WEST (2007): "Simulation of hyper-inverse Wishart distributions in graphical models," *Biometrika*, 94, 647–659.

CARVALHO, C. M., N. G. POLSON, AND J. G. SCOTT (2010): "The horseshoe estimator for sparse signals," *Biometrika*, 97, 465–480.

CARVALHO, C. M. AND M. WEST (2007): "Dynamic matrix-variate graphical models," *Bayesian Analysis*, 2, 69–97.

CASARIN, R., F. LEISEN, G. MOLINA, E. TER HORST, ET AL. (2015): "A Bayesian beta Markov random field calibration of the term structure of implied risk neutral densities," *Bayesian Analysis*, 10, 791–819.

CASARIN, R., D. SARTORE, AND M. TRONZANO (2018): "A Bayesian Markov-switching correlation model for contagion analysis on exchange rate markets," *Journal of Business & Economic Statistics*, 36, 101–114.

CASTEIGTS, A., P. FLOCCHINI, W. QUATTROCIOCCHI, AND N. SANTORO (2012): "Time-varying graphs and dynamic networks," *International Journal of Parallel, Emergent and Distributed Systems*, 27, 387–408.

CELEUX, G. (1998): "Bayesian inference for mixture: the label switching problem," in *Compstat*, Springer, 227–232.

CERCHIELLO, P. AND P. GIUDICI (2016): "Conditional graphical models for systemic risk estimation," *Expert systems with applications*, 43, 165–174.

CHANEY, T. (2014): "The network structure of international trade," *The American Economic Review*, 104, 3600–3634.

CHARPENTIER, A., J.-D. FERMANIAN, AND O. SCAILLET (2007): *Copulas: from theory to application in finance*, Risk Books, chap. The estimation of copulas: Theory and practice.

CHEN, S. X. (1999): "Beta kernel estimators for density functions," *Computational Statistics & Data Analysis*, 31, 131–145.

CHERUBINI, U., E. LUCIANO, AND W. VECCHIATO (2004): *Copula methods in finance*, John Wiley & Sons.

CHERUBINI, U., S. MULINACCI, F. GOBBI, AND S. ROMAGNOLI (2011): *Dynamic copula methods in finance*, John Wiley & Sons.

CHIB, S., F. NARDARI, AND N. SHEPHARD (2002): "Markov Chain Monte Carlo methods for stochastic volatility models," *Journal of Econometrics*, 108, 281–316.

CHINAZZI, M., G. FAGIOLO, J. A. REYES, AND S. SCHIAVO (2013): "Post-mortem examination of the international financial network," *Journal of Economic Dynamics and Control*, 37, 1692–1713.

CHRISTAKIS, N. A. AND J. H. FOWLER (2008): "The collective dynamics of smoking in a large social network," *New England journal of medicine*, 358, 2249–2258.

CHUDIK, A., V. GROSSMAN, AND M. H. PESARAN (2016): "A multi-country approach to forecasting output growth using PMIs," *Journal of Econometrics*, 192, 349–365.

CICHOCKI, A. (2014): "Era of Big data processing: a new approach via tensor networks and tensor decompositions," *arXiv preprint arXiv:1403.2048*.

CICHOCKI, A., N. LEE, I. OSELEDETS, A. PHAN, Q. ZHAO, AND D. MANDIC (2016): "Low-rank tensor networks for dimensionality reduction and large-scale optimization problems: perspectives and challenges PART 1," *arXiv preprint arXiv:1609.00893*.

CICHOCKI, A., D. MANDIC, L. DE LATHAUWER, G. ZHOU, Q. ZHAO, C. CAIAFA, AND H. A. PHAN (2015): "Tensor Decompositions for Signal Processing Applications: From two-way to Multiway Component Analysis," *IEEE Signal Processing Magazine*, 32, 145–163.

CICHOCKI, A., R. ZDUNEK, A. H. PHAN, AND S.-I. AMARI (2009): *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons.

CONT, R. AND J. DA FONSECA (2002): "Dynamics of implied volatility surfaces," *Quantitative finance*, 2, 45–60.

COOK, R. D., B. LI, AND F. CHIAROMONTE (2010): "Envelope models for parsimonious and efficient multivariate linear regression," *Statistica Sinica*, 20, 927–960.

CORSI, F., N. FUSARI, AND D. LA VECCHIA (2013): "Realizing smiles: pptions pricing with realized volatility," *Journal of Financial Economics*, 107, 284–304.

CORSI, F., F. LILLO, AND D. PIRINO (2015): "Measuring flight-to-quality with Granger-causality tail risk networks," *Available at SSRN: https://ssrn.com/abstract=2576078*.

COZZO, E., G. F. DE ARRUDA, F. A. RODRIGUES, AND Y. MORENO (2016): "Multilayer networks: metrics and spectral properties," in *Interconnected Networks*, Springer, 17–35.

DAGLISH, T., J. HULL, AND W. SUO (2007): "Volatility surfaces: theory, rules of thumb, and empirical evidence," *Quantitative Finance*, 7, 507–524.

DAMOISEAUX, J., S. ROMBOUTS, F. BARKHOF, P. SCHELTENS, C. STAM, S. M. SMITH, AND C. BECKMANN (2006): "Consistent resting-state networks across healthy subjects," *Proceedings of the national academy of sciences*, 103, 13848–13853.

DAVIDSON, I., S. GILPIN, O. CARMICHAEL, AND P. WALKER (2013): "Network discovery via constrained tensor analysis of fMRI data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 194–202.

DAWID, A. P. AND S. L. LAURITZEN (1993): "Hyper Markov laws in the statistical analysis of decomposable graphical models," *The Annals of Statistics*, 21, 1272–1317.

DE BOOR, C. (2001): *A practical guide to splines*, Springer-Verlag New York.

DE DOMENICO, M., C. GRANELL, M. A. PORTER, AND A. ARENAS (2016): "The physics of spreading processes in multilayer networks," *Nature Physics*, 12, 901.

DE PAULA, À. (2017): *Advances in economics: the eleventh World Congress of the Econometric Society*, Cambridge University Press, chap. Econometrics of network models, 268–323.

DEHEUVELS, P. (1978): "Caractérisation complète des lois extrêmes multivariées et de la convergence des types extrêmes," *Publ. Inst. Statist. Univ. Paris*, 23, 1–36.

——— (1979): "La fonction de dépendance empirique et ses Proprétés. Un test non paramétrique d'indépendance," *Académie Royale de Belgique. Bulletin de la Classe des Sciences (5)*, 65, 274–292.

DELLAPORTAS, P., J. J. FORSTER, AND I. NTZOUFRAS (2002): "On Bayesian model and variable selection using MCMC," *Statistics and Computing*, 12, 27–36.

DELLAPORTAS, P., P. GIUDICI, AND G. ROBERTS (2003): "Bayesian inference for nondecomposable graphical Gaussian models," *Sankhyā: The Indian Journal of Statistics*, 43–55.

DELPINI, D. AND G. BORMETTI (2015): "Stochastic volatility with heterogeneous time scales," *Quantitative Finance*, 15, 1597–1608.

DESSART, L., C. VELOUTSOU, AND A. MORGAN-THOMAS (2016): "Capturing consumer engagement: duality, dimensionality and measurement," *Journal of Marketing Management*, 32, 399–426.

DI GIOVANNI, J., A. A. LEVCHENKO, AND I. MÉJEAN (2014): "Firms, destinations, and aggregate fluctuations," *Econometrica*, 82, 1303–1340.

DIAS, A. AND P. EMBRECHTS (2004): "Dynamic copula models for multivariate high-frequency data in finance," *Manuscript, ETH Zurich*, 81.

DICKISON, M. E., M. MAGNANI, AND L. ROSSI (2016): *Multilayer social networks*, Cambridge University Press.

DIEBOLD, F. X. AND K. YILMAZ (2014): "On the network topology of variance decompositions: measuring the connectedness of financial firms," *Journal of Econometrics*, 182, 119–134.

——— (2015): *Financial and macroeconomic connectedness: a network approach to measurement and monitoring*, Oxford University Press, USA.

DIESTEL, R. (2012): *Graph theory*, vol. 173 of *Graduate Texts in Mathematics*, Springer.

DIETZENBACHER, E., B. LOS, R. STEHRER, M. TIMMER, AND G. DE VRIES (2013): "The construction of world input–output tables in the WIOD project," *Economic Systems Research*, 25, 71–98.

DING, S. AND R. D. COOK (2016): "Matrix-variate regressions and envelope models," *arXiv preprint arXiv:1605.01485*.

DOBRA, A. (2015): *Handbook of Spatial Epidemiology*, Chapman & Hall /CRC, chap. Graphical Modeling of Spatial Health Data, first ed.

DUBEY, S. D. (1970): "Compound gamma, beta and F distributions," *Metrika*, 16, 27–31.

DUEKER, M. J. (1997): "Markov switching in GARCH processes and mean-reverting stock-market volatility," *Journal of Business & Economic Statistics*, 15, 26–34.

——— (2005): "Dynamic forecasts of qualitative variables: a Qual VAR model of US recessions," *Journal of Business & Economic Statistics*, 23, 96–104.

DURANTE, D. AND D. B. DUNSON (2014a): "Bayesian logistic Gaussian process models for dynamic networks," *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 33.

——— (2014b): "Nonparametric Bayesian dynamic modelling of relational data," *Biometrika*, 101, 883–898.

DURANTE, F. AND C. SEMPI (2015): *Principles of copula theory*, CRC Press.

ECKART, C. AND G. YOUNG (1936): "The approximation of one matrix by another of lower rank," *Psychometrika*, 1, 211–218.

EGOZCUE, J., J. DÍAZ-BARRERO, AND V. PAWLOWSKY-GLAHN (2006): "Hilbert space of probability density function based on Aitchison geometry," *Acta Mathematica Sinica*, 22, 1175–1182.

EGOZCUE, J. AND V. PAWLOWSKY-GLAHN (2015): "Changing the reference measure in the simplex and its weighting effects," in *Welcome to CoDawork*.

ENGLE, R. F. AND C. W. GRANGER (1987): "Co-integration and error correction: representation, estimation, and testing," *Econometrica*, 251–276.

ERDÖS, P. AND A. RÉNYI (1959): "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, 6, 290–297.

EREVELLES, S., N. FUKAWA, AND L. SWAYNE (2016): "Big Data consumer analytics and the transformation of marketing," *Journal of Business Research*, 69, 897–904.

ESTIENNE, F., N. MATTHIJS, D. MASSART, P. RICOUX, AND D. LEIBOVICI (2001): "Multi-way modelling of high-dimensionality electroencephalographic data," *Chemometrics and Intelligent Laboratory Systems*, 58, 59–72.

FAGIOLO, G. (2010): "The international-trade network: gravity equations and topological properties," *Journal of Economic Interaction and Coordination*, 5, 1–25.

FAGIOLO, G., J. REYES, AND S. SCHIAVO (2008): "On the topological properties of the world trade web: a weighted network analysis," *Physica A: Statistical Mechanics and its Applications*, 387, 3868–3873.

——— (2009): "World-trade web: topological properties, dynamics, and evolution," *Physical Review E*, 79, 036115.

——— (2010): "The evolution of the world trade web: a weighted-network analysis," *Journal of Evolutionary Economics*, 20, 479–514.

FENGLER, M. R. (2012): "Option data and modeling BSM implied volatility," in *Handbook of computational finance*, Springer, 117–142.

FENGLER, M. R., W. K. HÄRDLE, AND E. MAMMEN (2007): "A semiparametric factor model for implied volatility surface dynamics," *Journal of Financial Econometrics*, 5, 189–218.

FERMANIAN, J.-D. AND O. SCAILLET (2003): "Nonparametric estimation of copulas for time series," *Journal of Risk*, 5, 25–54.

——— (2004): "Some statistical pitfalls in copula modeling for financial applications," Tech. rep., International Center for Financial Asset Management and Engineering.

FERMANIAN, J.-D. AND M. H. WEGKAMP (2012): "Time-dependent copulas," *Journal of Multivariate Analysis*, 110, 19–29.

FERRATY, F. AND P. VIEU (2006): *Nonparametric functional data analysis: theory and practice*, Springer Science & Business Media.

FODOR, I. K. (2002): "A survey of dimension reduction techniques," Tech. rep., Lawrence Livermore National Lab., CA (US).

FORBES, K. AND R. RIGOBON (2001): "Measuring contagion: conceptual and empirical issues," in *International financial contagion*, Springer, 43–66.

FORBES, K. J. AND R. RIGOBON (2002): "No contagion, only interdependence: measuring stock market comovements," *The journal of Finance*, 57, 2223–2261.

FRAHM, G., M. JUNKER, AND R. SCHMIDT (2005): "Estimating the tail-dependence coefficient: properties and pitfalls," *Insurance: mathematics and economics*, 37, 80–100.

FRANK, O. AND D. STRAUSS (1986): "Markov graphs," *Journal of the american Statistical association*, 81, 832–842.

FRIEDMAN, J., T. HASTIE, AND R. TIBSHIRANI (2008): "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

FRIEL, N., R. RASTELLI, J. WYSE, AND A. E. RAFTERY (2016): "Interlocking directorates in Irish companies using a latent space model for bipartite networks," *Proceedings of the National Academy of Sciences*, 113, 6629–6634.

FRÜHWIRTH-SCHNATTER, S. (2001): "Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models," *Journal of the American Statistical Association*, 96, 194–209.

——— (2006): *Finite mixture and Markov switching models*, Springer.

GABAIX, X. (2011): "The granular origins of aggregate fluctuations," *Econometrica*, 79, 733–772.

Gai, P., A. Haldane, and S. Kapadia (2011): "Complexity, concentration and contagion," *Journal of Monetary Economics*, 58, 453–470.

Gallotti, R. and M. Barthelemy (2015): "The multilayer temporal network of public transport in Great Britain," *Scientific data*, 2, 140056.

Gatheral, J. (2011): *The volatility surface: a practitioner's guide*, vol. 357, John Wiley & Sons.

Gefang, D. (2014): "Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage," *International Journal of Forecasting*, 30, 1–11.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014): *Bayesian data analysis*, vol. 2, CRC Press.

Gelman, A. and D. B. Rubin (1992): "Inference from iterative simulation using multiple sequences," *Statistical science*, 457–472.

George, E. I. and R. E. McCulloch (1993): "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

——— (1997): "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–373.

George, E. I., D. Sun, and S. Ni (2008): "Bayesian stochastic search for VAR model restrictions," *Journal of Econometrics*, 142, 553–580.

Geweke, J. (1991): *Bayesian Statistics 4*, Clarendon Press, chap. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, 169–193.

Giraitis, L., G. Kapetanios, A. Wetherilt, and F. Žikeš (2016): "Estimating the dynamics and persistence of financial networks, with an application to the Sterling money market," *Journal of Applied Econometrics*, 31, 58–84.

Giudici, P. and A. Spelta (2016): "Graphical network models for international financial flows," *Journal of Business & Economic Statistics*, 34, 128–138.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, E. M. Airoldi, et al. (2010): "A survey of statistical network models," *Foundations and Trends in Machine Learning*, 2, 129–233.

Graham, B. S. (2017): "An econometric model of network formation with degree heterogeneity," *Econometrica*, 85, 1033–1063.

Granger, C. W. (1988): "Some recent development in a concept of causality," *Journal of econometrics*, 39, 199–211.

Grossmann, A., I. Love, and A. G. Orlov (2014): "The dynamics of exchange rate volatility: a panel VAR approach," *Journal of International Financial Markets, Institutions and Money*, 33, 1–27.

Guégan, D. and J. Zhang (2010): "Change analysis of a dynamic copula for measuring dependence in multivariate financial data," *Quantitative Finance*, 10, 421–430.

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2017): "Bayesian tensor regression," *Journal of Machine Learning Research*, 18, 1–31.

Guidolin, M. and A. Timmermann (2006): "An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns," *Journal of Applied Econometrics*, 21, 1–22.

Gupta, A. K. and D. K. Nagar (1999): *Matrix variate distributions*, CRC Press.

GURTNER, G., S. VITALI, M. CIPOLLA, F. LILLO, R. N. MANTEGNA, S. MICCICHE, AND S. POZZI (2014): "Multi-scale analysis of the European airspace using network community detection," *PloS one*, 9, e94414.

HAAS, M., S. MITTNIK, AND M. S. PAOLELLA (2004): "A new approach to Markov-switching GARCH models," *Journal of Financial Econometrics*, 2, 493–530.

HACKBUSCH, W. (2012): *Tensor spaces and numerical tensor calculus*, Springer Science & Business Media.

HAFNER, C. M. AND H. MANNER (2012): "Dynamic stochastic copula models: estimation, inference and applications," *Journal of Applied Econometrics*, 27, 269–295.

HAFNER, C. M. AND O. REZNIKOVA (2010): "Efficient estimation of a semiparametric dynamic copula model," *Computational Statistics & Data Analysis*, 54, 2609–2627.

HALL, P., H.-G. MÜLLER, AND J.-L. WANG (2006): "Properties of principal component methods for functional and longitudinal data analysis," *The Annals of Statistics*, 34, 1493–1517.

HAMILTON, J. D. (1989): "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, 57, 357–384.

HAMILTON, J. D. AND R. SUSMEL (1994): "Autoregressive conditional heteroskedasticity and changes in regime," *Journal of Econometrics*, 64, 307–333.

HANDCOCK, M. S., A. E. RAFTERY, AND J. M. TANTRUM (2007): "Model-based clustering for social networks," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 301–354.

HANNEKE, S., W. FU, AND E. P. XING (2010): "Discrete temporal models of social networks," *Electronic Journal of Statistics*, 4, 585–605.

HARRIS, M. N. AND X. ZHAO (2007): "A zero-inflated ordered probit model, with an application to modelling tobacco consumption," *Journal of Econometrics*, 141, 1073–1099.

HARRISON, J. AND M. WEST (1999): *Bayesian forecasting & dynamic models*, Springer.

HARSHMAN, R. A. (1970): "Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics, 16, 1- 84*.

HAUTSCH, N., J. SCHAUMBURG, AND M. SCHIENLE (2014): "Financial network systemic risk contributions," *Review of Finance*, 19, 685–738.

HAYS, S., H. SHEN, AND J. Z. HUANG (2012): "Functional dynamic factor models with application to yield curve forecasting," *The Annals of Applied Statistics*, 6, 870–894.

HIDALGO, C. A. AND R. HAUSMANN (2009): "The building blocks of economic complexity," *Proceedings of the national academy of sciences*, 106, 10570–10575.

HOFF, P. D. (2011): "Separable covariance arrays via the Tucker product, with applications to multivariate relational data," *Bayesian Analysis*, 6, 179–196.

——— (2015): "Multilinear tensor regression for longitudinal relational data," *The Annals of Applied Statistics*, 9, 1169–1193.

HOFF, P. D., A. E. RAFTERY, AND M. S. HANDCOCK (2002): "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

HOLLAND, P. W. AND S. LEINHARDT (1981): "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, 76, 33–50.

HOLME, P. (2005): "Network reachability of real-world contact sequences," *Physical Review E*, 71, 046119.

HOLME, P. AND J. SARAMÄKI (2012): "Temporal networks," *Physics Reports*, 519, 97–125.

——— (2013): *Temporal networks*, Springer.

HOLSCLAW, T., A. M. GREENE, A. W. ROBERTSON, AND P. SMYTH (2017): "Bayesian nonhomogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling," *arXiv preprint arXiv:1701.02856*.

HOMESCU, C. (2011): "Implied volatility surface: Construction methodologies and characteristics," *arXiv preprint arXiv:1107.1834*.

HÖRMANN, S., L. KIDZIŃSKI, AND M. HALLIN (2015): "Dynamic functional principal components," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 319–348.

HÖRMANN, S. AND P. KOKOSZKA (2012): *Functional time series*, Elsevier, vol. 30.

HORVÁTH, L., M. HUŠKOVÁ, AND P. KOKOSZKA (2010): "Testing the stability of the functional autoregressive process," *Journal of Multivariate Analysis*, 101, 352–367.

HORVÁTH, L., P. KOKOSZKA, AND G. RICE (2014): "Testing stationarity of functional time series," *Journal of Econometrics*, 179, 66–82.

HRON, K., A. MENAFOGLIO, M. TEMPL, K. HRUZOVÁ, AND P. FILZMOSER (2016): "Simplicial principal component analysis for density functions in Bayes spaces," *Computational Statistics & Data Analysis*, 94, 330–350.

HU, J. (2010): "Dependence structures in Chinese and US financial markets: a time-varying conditional copula approach," *Applied Financial Economics*, 20, 561–583.

HUNG, H. AND C.-C. WANG (2013): "Matrix variate logistic regression model with application to EEG data," *Biostatistics*, 14, 189–202.

IMAIZUMI, M. AND K. HAYASHI (2016): "Doubly decomposing nonparametric tensor regression," in *International Conference on Machine Learning*, 727–736.

ISHWARAN, H. AND J. S. RAO (2005): "Spike and slab variable selection: frequentist and Bayesian strategies," *Annals of Statistics*, 730–773.

JACKSON, M. O. (2010): *Social and economic networks*, Princeton University Press.

JACKSON, M. O. AND A. WATTS (2002): "The evolution of social and economic networks," *Journal of Economic Theory*, 106, 265–295.

JACKSON, M. O. AND A. WOLINSKY (1996): "A strategic model of social and economic networks," *Journal of Economic Theory*, 71, 44–74.

JOCHMANN, M., G. KOOP, AND R. W. STRACHAN (2010): "Bayesian forecasting using stochastic search variable selection in a VAR subject to breaks," *International Journal of Forecasting*, 26, 326–347.

JOE, H. (1997): *Multivariate models and multivariate dependence concepts*, CRC Press.

JOE, H. AND D. KUROWICKA (2011): *Dependence modeling: vine copula handbook*, World Scientific.

JOHNSON, N. L., S. KOTZ, AND N. BALAKRISHNAN (1995): *Continuous univariate distributions*, vol. 2 of *Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics*, Wiley, New York.

JONDEAU, E. AND M. ROCKINGER (2006): "The copula-GARCH model of conditional dependencies: an international stock market application," *Journal of International Money and Finance*, 25, 827–853.

JONES, B. AND M. WEST (2005): "Covariance decomposition in undirected Gaussian graphical models," *Biometrika*, 92, 779–786.

KAISER, H. F. (1958): "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, 23, 187–200.

KALI, R. AND J. REYES (2010): "Financial contagion on the international trade network," *Economic Inquiry*, 48, 1072–1101.

KARGIN, V. AND A. ONATSKI (2008): "Curve forecasting by functional autoregression," *Journal of Multivariate Analysis*, 99, 2508–2526.

KARHUNEN, K. (1947): "Über lineare methoden in der wahrscheinlichkeitsrechnung," *Annales Academiae Scientiarum Fennicae*, 1–79.

KASS, R. E. AND A. E. RAFTERY (1995): "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

KAUFMANN, S. (2000): "Measuring business cycles with a dynamic Markov switching factor model: An assessment using Bayesian simulation methods," *The Econometrics Journal*, 3, 39–65.

——— (2010): "Dating and forecasting turning points by Bayesian clustering with dynamic structure: A suggestion with an application to Austrian data," *Journal of Applied Econometrics*, 25, 309–344.

——— (2015): "K-state switching models with time-varying transition distributions—Does loan growth signal stronger effects of variables on inflation?" *Journal of Econometrics*, 187, 82–94.

KHARRAZI, A., E. ROVENSKAYA, AND B. D. FATH (2017): "Network structure impacts global commodity trade growth and resilience," *PloS one*, 12, e0171184.

KIDZIŃSKI, L. (2015): "Functional time series," *arXiv preprint arXiv:1502.07113*.

KIDZIŃSKI, L., P. KOKOSZKA, AND N. M. JOUZDANI (2016): "Principal component analysis of periodically correlated functional time series," *arXiv preprint arXiv:1612.00040*.

KIERS, H. A. (2000): "Towards a standardized notation and terminology in multiway analysis," *Journal of Chameometrics*, 14, 105–122.

KIM, C.-J. AND C. R. NELSON (1998): "Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching," *Review of Economics and Statistics*, 80, 188–201.

KIVELÄ, M., A. ARENAS, M. BARTHELEMY, J. P. GLEESON, Y. MORENO, AND M. A. PORTER (2014): "Multilayer networks," *Journal of Complex Networks*, 2, 203–271.

KIVELÄ, M., R. K. PAN, K. KASKI, J. KERTÉSZ, J. SARAMÄKI, AND M. KARSAI (2012): "Multiscale analysis of spreading in a large communication network," *Journal of Statistical Mechanics: Theory and Experiment*, 2012, P03005.

KLAASSEN, F. (2002): "Improving GARCH volatility forecasts with regime-switching GARCH," in *Advances in Markov-Switching Models*, Springer, 223–254.

KLEPSCH, J., C. KLÜPPELBERG, AND T. WEI (2017): "Prediction of functional ARMA processes with an application to traffic data," *Econometrics and Statistics*, 1, 128–149.

KOH, K., S.-J. KIM, AND S. BOYD (2007): "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine learning research*, 8, 1519–1555.

KOKOSZKA, P. (2012): "Dependent functional data," *ISRN Probability and Statistics*.

KOLACZYK, E. D. (2009): *Statistical analysis of network data: methods and models*, Springer Science & Business Media.

KOLAR, M., L. SONG, A. AHMED, AND E. P. XING (2010): "Estimating time-varying networks," *The Annals of Applied Statistics*, 4, 94–123.

KOLDA, T. G. (2006): "Multilinear operators for higher-order decompositions." Tech. rep., Sandia National Laboratories.

KOLDA, T. G. AND B. W. BADER (2009): "Tensor decompositions and applications," *SIAM Review*, 51, 455–500.

KOLDA, T. G., B. W. BADER, AND J. P. KENNY (2005): "Higher-order web link analysis using multilinear algebra," in *Fifth IEEE International Conference on Data Mining*, IEEE Computer Society, 242–249.

KÖNIG, M. D., D. ROHNER, M. THOENIG, AND F. ZILIBOTTI (2017): "Networks in conflict: theory and evidence from the great war of Africa," *Econometrica*, 85, 1093–1132.

KOOP, G. AND D. KOROBILIS (2016): "Model uncertainty in panel vector autoregressive models," *European Economic Review*, 81, 115–131.

KOOP, G., D. KOROBILIS, AND D. PETTENUZZO (2018): "Bayesian compressed vector autoregressions," *Journal of Econometrics*.

KOROBILIS, D. (2013a): "Hierarchical shrinkage priors for dynamic regressions with many predictors," *International Journal of Forecasting*, 29, 43–59.

——— (2013b): "VAR forecasting using Bayesian variable selection," *Journal of Applied Econometrics*, 28, 204–230.

——— (2016): "Prior selection for panel vector autoregressions," *Computational Statistics & Data Analysis*, 101, 110–120.

KOSTAKOS, V. (2009): "Temporal graphs," *Physica A: Statistical Mechanics and its Applications*, 388, 1007–1023.

KRIVITSKY, P. N. AND M. S. HANDCOCK (2014): "A separable model for dynamic networks," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 29–46.

KROONENBERG, P. M. (2008): *Applied multiway data analysis*, John Wiley & Sons.

KRUIJER, W., J. ROUSSEAU, AND A. VAN DER VAART (2010): "Adaptive Bayesian density estimation with location-scale mixtures," *Electronic Journal of Statistics*, 4, 1225–1257.

LAMBERT, D. (1992): "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, 34, 1–14.

LAURITZEN, S. L. (1996): *Graphical Models*, Clarendon Press.

LEE, N. AND A. CICHOCKI (2016): "Fundamental tensor operations for large-scale data analysis in tensor train formats," *arXiv preprint arXiv:1405.7786*.

LENZEN, M., L.-L. PADE, AND J. MUNKSGAARD (2004): "CO2 multipliers in multi-region input-output models," *Economic Systems Research*, 16, 391–412.

LENZEN, M., R. WOOD, AND T. WIEDMANN (2010): "Uncertainty analysis for multi-region input–output models–a case study of the UK's carbon footprint," *Economic Systems Research*, 22, 43–63.

LI, L. AND X. ZHANG (2017): "Parsimonious tensor response regression," *Journal of the American Statistical Association*, 112, 1131–1146.

LI, Y. AND T. HSING (2010): "Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data," *The Annals of Statistics*, 38, 3321–3351.

LIEBL, D. (2010): "Modeling hourly electricity spot market prices as non stationary functional times series," Tech. rep., University Library of Munich, Germany.

——— (2013): "Modeling and forecasting electricity spot prices: a functional data perspective," *The Annals of Applied Statistics*, 7, 1562–1592.

LINDQUIST, M. A. (2008): "The statistical analysis of fMRI data," *Statistical science*, 23, 439–464.

LOÈVE, M. (1945): "Fonctions aléatoires de second ordre," Tech. Rep. 220.

LOF, M. AND T. MALINEN (2014): "Does sovereign debt weaken economic growth? A panel VAR analysis," *Economics Letters*, 122, 403–407.

LOPES, H. F. (2014): "A tutorial on computation of Bayes factors," *INSPER working paper*, 134.

LOVE, I. AND L. ZICCHINO (2006): "Financial development and dynamic investment behavior: evidence from panel VAR," *The Quarterly Review of Economics and Finance*, 46, 190–210.

LOVELOCK, D. AND H. RUND (1989): *Tensors, differential forms, and variational principles*, Courier Corporation.

LYCHE, T. AND K. MORKEN (2008): *Spline methods draft*, Department of Informatics, Center of Mathematics for Applications, University of Oslo.

MACHALOVÀ, J. (2002a): "Optimal interpolating and optimal smoothing spline," *Journal of Electrical Engineering*, 53, 79–82.

——— (2002b): "Optimal interpolatory splines using B-spline representation," *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica*, 41, 105–118.

MACHALOVÀ, J., K. HRON, AND G. S. MONTI (2016): "Preprocessing of centred logratio transformed density functions using smoothing splines," *Journal of Applied Statistics*, 43, 1419–1435.

MAGNUS, J. R. AND H. NEUDECKER (1999): *Matrix differential calculus with applications in statistics and econometrics*, Wiley, New York.

MAJEWSKI, A. A., G. BORMETTI, AND F. CORSI (2015): "Smile from the past: a general option pricing framework with multiple volatility and leverage components," *Journal of Econometrics*, 187, 521–531.

MALVERN, L. E. (1986): *Introduction to the mechanics of a continuous medium*, Englewood.

MANCEUR, A. M. AND P. DUTILLEUL (2013): "Maximum likelihood estimation for the tensor normal distribution: algorithm, minimum sample size, and empirical bias and dispersion," *Journal of Computational and Applied Mathematics*, 239, 37–49.

MANNER, H. AND O. REZNIKOVA (2012): "A survey on time-varying copulas: specification, simulations, and application," *Econometric Reviews*, 31, 654–687.

MAZZARISI, P., P. BARUCCA, F. LILLO, AND D. TANTARI (2017): "A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market," *arXiv preprint arXiv:1801.00185*.

MEINSHAUSEN, N. AND P. BÜHLMANN (2006): "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, 1436–1462.

MELE, A. (2017): "A structural model of Dense Network Formation," *Econometrica*, 85, 825–850.

MENAFOGLIO, A., A. GUADAGNINI, AND P. SECCHI (2014): "A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers," *Stochastic Environmental Research and Risk Assessment*, 28, 1835–1851.

MEYFROIDT, P., T. K. RUDEL, AND E. F. LAMBIN (2010): "Forest transitions, trade, and the global displacement of land use," *Proceedings of the National Academy of Sciences*, 107, 20917–20922.

MISTRULLI, P. E. (2011): "Assessing financial contagion in the interbank market: maximum entropy versus observed interbank lending patterns," *Journal of Banking & Finance*, 35, 1114–1127.

MITCHELL, T. J. AND J. J. BEAUCHAMP (1988a): "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1032.

——— (1988b): "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1032.

MIWAKEICHI, F., E. MARTÍNEZ-MONTES, P. A. VALDÉS-SOSA, N. NISHIYAMA, H. MIZUHARA, AND Y. YAMAGUCHI (2004): "Decomposing EEG data into space–time–frequency components using parallel factor analysis," *NeuroImage*, 22, 1035–1045.

MONTAGNA, M. AND C. KOK (2016): "Multi-layered interbank model for assessing systemic risk," Tech. rep., ECB Working Paper.

MURASE, Y., J. TÖRÖK, H.-H. JO, K. KASKI, AND J. KERTÉSZ (2014): "Multilayer weighted social network model," *Physical Review E*, 90, 052810.

NAIK, P., M. WEDEL, L. BACON, A. BODAPATI, E. BRADLOW, W. KAMAKURA, J. KREULEN, P. LENK, D. M. MADIGAN, AND A. MONTGOMERY (2008): "Challenges and opportunities in high-dimensional choice data analyses," *Marketing Letters*, 19, 201.

NAKAJIMA, J. AND M. WEST (2015): "Dynamic network signal processing using latent threshold models," *Digital Signal Processing*, 47, 5–16.

NEAL, R. M. (1994): "Contribution to the discussion of "Approximate Bayesian inference with the weighted likelihood bootstrap" by Newton MA, Raftery AE," *Journal of the Royal Statistical Society: Series A (Methodological)*, 56, 41–42.

——— (2011): "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, ed. by S. Brooks, A. Gelman, J. L. Galin, and X.-L. Meng, Chapman & Hall /CRC, chap. 5.

NELSEN, R. B. (2013): *An introduction to copulas*, vol. 139, Springer Science & Business Media.

NICHOLSON, W., J. BIEN, AND D. MATTESON (2016): "High dimensional forecasting via interpretable vector autoregression," *ArXiv e-prints*.

NICHOLSON, W. B., D. S. MATTESON, AND J. BIEN (2017): "VARX-L: structured regularization for large vector autoregressions with exogenous variables," *International Journal of Forecasting*, 33, 627–651.

NOWICKI, K. AND T. A. B. SNIJDERS (2001): "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087.

OH, D. H. AND A. J. PATTON (2017): "Time-varying systemic risk: evidence from a dynamic copula model of CDS spreads," *Journal of Business & Economic Statistics*, 1–15.

OHLSON, M., M. R. AHMAD, AND D. VON ROSEN (2013): "The multilinear normal distribution: introduction and some basic properties," *Journal of Multivariate Analysis*, 113, 37–47.

OSELIO, B., A. KULESZA, AND A. O. HERO (2014): "Multi-layer graph analysis for dynamic social networks," *IEEE Journal of Selected Topics in Signal Processing*, 8, 514–523.

PALLA, K., F. CARON, AND Y. W. TEH (2016): "Bayesian nonparametrics for sparse dynamic networks," *arXiv preprint arXiv:1607.01624*.

PAN, R. (2014): "Tensor transpose and its properties," *arXiv preprint arXiv:1411.1503*.

PARK, T. AND G. CASELLA (2008): "The Bayesian lasso," *Journal of the American Statistical Association*, 103, 681–686.

PATTON, A. J. (2006a): "Estimation of multivariate models for time series of possibly different lengths," *Journal of Applied Econometrics*, 21, 147–173.

——— (2006b): "Modelling asymmetric exchange rate dependence," *International Economic Review*, 47, 527–556.

——— (2012): "A review of copula models for economic time series," *Journal of Multivariate Analysis*, 110, 4–18.

PENSKY, M. (2016): "Dynamic network models and graphon estimation," *arXiv preprint arXiv:1607.00673*.

PETERSEN, A. AND H.-G. MÜLLER (2016): "Functional data analysis for density functions by transformation to a Hilbert Space," *The Annals of Statistics*, 44, 183–218.

PETRIS, G. (2013): "A Bayesian framework for functional time series analysis," *arXiv preprint arXiv:1311.0098*.

POLEDNA, S., J. L. MOLINA-BORBOA, S. MARTÍNEZ-JARAMILLO, M. VAN DER LEIJ, AND S. THURNER (2015): "The multi-layer network nature of systemic risk and its implications for the costs of financial crises," *Journal of Financial Stability*, 20, 70–81.

POLSON, N. G., J. G. SCOTT, AND J. WINDLE (2013): "Bayesian inference for logistic models using Pólya–Gamma latent variables," *Journal of the American Statistical Association*, 108, 1339–1349.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING, AND B. P. FLANNERY (2007): *Numerical recipes: the art of scientic computing*, Cambridge University Press.

RAJARATNAM, B., H. MASSAM, AND C. M. CARVALHO (2008): "Flexible covariance estimation in graphical Gaussian models," *The Annals of Statistics*, 36, 2818–2849.

RAMSAY, J. O. AND B. W. SILVERMAN (2005): *Functional data analysis*, Springer Series in Statistics, Springer, second edition ed.

RASTELLI, R., N. FRIEL, AND A. E. RAFTERY (2016): "Properties of latent variable network models," *Network Science*, 4, 407–432.

RAVIKUMAR, P., M. J. WAINWRIGHT, AND J. D. LAFFERTY (2010): "High-dimensional Ising model selection using l1-regularized logistic regression," *The Annals of Statistics*, 38, 1287–1319.

ROBERT, C. P. AND G. CASELLA (2004): *Monte Carlo statistical methods*, Springer.

ROBERT, C. P. AND D. WRAITH (2009): "Computational methods for Bayesian model choice," in *AIP Conference Proceedings*, vol. 1193, 251–262.

ROBINS, G. AND P. PATTISON (2001): "Random graph models for temporal processes in social networks," *Journal of Mathematical Sociology*, 25, 5–41.

ROBINS, G., P. PATTISON, Y. KALISH, AND D. LUSHER (2007): "An introduction to exponential random graph (p*) models for social networks," *Social networks*, 29, 173–191.

ROČKOVÁ, V. AND E. I. GEORGE (2014): "EMVS: The EM approach to Bayesian variable selection," *Journal of the American Statistical Association*, 109, 828–846.

RUBINOV, M. AND O. SPORNS (2010): "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, 52, 1059–1069.

SALAZAR, E., R. GIRALDO, AND E. PORCU (2015): "Spatial prediction for infinite-dimensional compositional data," *Stochastic Environmental Research and Risk Assessment*, 29, 1737–1749.

SANZ DÍAZ, M. T., R. YÑÍGUEZ OVANDO, AND J. M. RUEDA CANTUCHE (2015): "The relevance of multicountry input-output tables in measuring emissions trade balance of countries: the case of Spain," *Sort*, 40, 3–30.

SARANTIS, N. AND C. STEWART (2001): "Saving behaviour in OECD countries: evidence from panel cointegration tests," *The Manchester School*, 69, 22–41.

SCHIAVO, S., J. REYES, AND G. FAGIOLO (2010): "International trade and financial integration: a weighted network analysis," *Quantitative Finance*, 10, 389–399.

SCHILDCROUT, J. S. AND P. J. HEAGERTY (2005): "Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency," *Biostatistics*, 6, 633–652.

SCHOTMAN, P. AND H. K. VAN DIJK (1991): "A Bayesian analysis of the unit root in real exchange rates," *Journal of Econometrics*, 49, 195–238.

SCHUMAKER, L. (2007): *Spline functions: basic theory*, Cambridge University Press.

SCHWEITZER, F., G. FAGIOLO, D. SORNETTE, F. VEGA-REDONDO, A. VESPIGNANI, AND D. R. WHITE (2009): "Economic networks: the new challenges," *Science*, 325, 422–425.

SCOTT, J. (2017): *Social network analysis*, Sage.

SCOTT, J. G. AND C. M. CARVALHO (2008): "Feature-inclusion stochastic search for Gaussian graphical models," *Journal of Computational and Graphical Statistics*, 17, 790–808.

SEN, R. AND C. KLÜPPELBERG (2015): "Time series of functional data," *Reports of the Indian Statistical Institute.*

SEN, R. AND C. MA (2015): "Forecasting density function: application in finance," *Journal of Mathematical Finance*, 5, 433–447.

SEWELL, D. K. AND Y. CHEN (2015): "Latent space models for dynamic networks," *Journal of the American Statistical Association*, 110, 1646–1657.

SHERMAN, M., T. V. APANASOVICH, AND R. J. CARROLL (2006): "On estimation in binary autologistic spatial models," *Journal of Statistical Computation and Simulation*, 76, 167–179.

SILVERMAN, B. W. (1986): *Density estimation for statistics and data analysis*, Springer-Science & Business Media.

SIMS, C. A., D. F. WAGGONER, AND T. ZHA (2008): "Methods for inference in large multiple-equation Markov-switching models," *Journal of Econometrics*, 146, 255–274.

SIMS, C. A. AND T. ZHA (1998): "Bayesian methods for dynamic multivariate models," *International Economic Review*, 39, 949–968.

——— (2006): "Were there regime switches in US monetary policy?" *The American Economic Review*, 96, 54–81.

SKIADOPOULOS, G., S. HODGES, AND L. CLEWLOW (2000): "The dynamics of the S&P 500 implied volatility surface," *Review of derivatives research*, 3, 263–282.

SKLAR, A. (1959): *Fonctions de répartition à n dimensions et leurs marges*, Université Paris 8.

SMILDE, A., R. BRO, AND P. GELADI (2005): *Multi-way analysis: applications in the chemical sciences*, John Wiley & Sons.

SMITH, D. R. (2002): "Markov-switching and stochastic volatility diffusion models of short-term interest rates," *Journal of Business & Economic Statistics*, 20, 183–197.

SO, M. K. AND C. Y. YEUNG (2014): "Vine-copula GARCH model with dynamic conditional dependence," *Computational Statistics & Data Analysis*, 76, 655–671.

SORZANO, C. O. S., J. VARGAS, AND A. P. MONTANO (2014): "A survey of dimensionality reduction techniques," *arXiv preprint arXiv:1403.2877*.

SPRINGER, M. AND W. THOMPSON (1970): "The distribution of products of beta, gamma and Gaussian random variables," *SIAM Journal on Applied Mathematics*, 18, 721–737.

SQUARTINI, T., G. FAGIOLO, AND D. GARLASCHELLI (2011): "Randomizing world trade. I. A binary network analysis," *Physical Review E*, 84, 046117.

STANISWALIS, J. G. AND J. J. LEE (1998): "Nonparametric regression analysis of longitudinal data," *Journal of the American Statistical Association*, 93, 1403–1418.

STANLEY, N., S. SHAI, D. TAYLOR, AND P. J. MUCHA (2016): "Clustering network layers with the strata multilayer stochastic block model," *IEEE transactions on network science and engineering*, 3, 95–105.

SYNGE, J. L. AND A. SCHILD (1969): *Tensor calculus*, Courier Corporation.

TADDY, M. (2013): "Multinomial inverse regression for text analysis," *Journal of the American Statistical Association*, 108, 755–770.

TADDY, M. A. (2010): "Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime," *Journal of the American Statistical Association*, 105, 1403–1417.

TANNER, M. A. AND W. H. WONG (1987): "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–540.

THIEMICHEN, S., N. FRIEL, A. CAIMO, AND G. KAUERMANN (2016): "Bayesian exponential random graph models with nodal random effects," *Social Networks*, 46, 11–28.

TIMMER, M. P., E. DIETZENBACHER, B. LOS, R. STEHRER, AND G. J. VRIES (2015): "An illustrated user guide to the world input–output database: the case of global automotive production," *Review of International Economics*, 23, 575–605.

TODESCHINI, A. AND F. CARON (2016): "Exchangeable random measures for sparse and modular graphs with overlapping communities," *arXiv preprint arXiv:1602.02114*.

TUMMINELLO, M., S. MICCICHE, F. LILLO, J. PIILO, AND R. N. MANTEGNA (2011): "Statistically validated networks in bipartite complex systems," *PloS one*, 6, e17994.

TURNEY, P. D. (2002): "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 417–424.

VAN DEN GOORBERGH, R. W., C. GENEST, AND B. J. WERKER (2005): "Bivariate option pricing using dynamic copula models," *Insurance: Mathematics and Economics*, 37, 101–114.

VAN DER BOOGAART, K. G., J. J. EGOZCUE, AND V. PAWLOWSKY-GLAHN (2010): "Bayes linear spaces," *SORT: Statistics and Operations Research Transactions*, 34, 201–222.

——— (2014): "Bayes Hilbert spaces," *Australian & New Zealand Journal of Statistics*, 56, 171–194.

VAN DYK, D. A. AND T. PARK (2008): "Partially collapsed Gibbs samplers: theory and methods," *Journal of the American Statistical Association*, 103, 790–796.

VEITCH, V. AND D. M. ROY (2015): "The class of random graphs arising from exchangeable random measures," *arXiv preprint arXiv:1512.03099*.

VERGOTE, O. AND J. M. P. GUTIÉRREZ (2012): "Interest rate expectations and uncertainty during ECB governing council days: evidence from intraday implied densities of 3-month Euribor," *Journal of Banking & Finance*, 36, 2804–2823.

VIROLI, C. (2011): "Finite mixtures of matrix normal distributions for classifying three-way data," *Statistics and Computing*, 21, 511–522.

——— (2012): "On matrix-variate regression analysis," *Journal of Multivariate Analysis*, 111, 296–309.

VIROLI, C. AND L. ANDERLUCCI (2013): "Modelling longitudinal data through matrix-variate normal mixtures," in *Advances in Latent Variables - Methods, Models and Applications*.

VISAYA, M. V., D. SHERWELL, B. SARTORIUS, AND F. CROMIERES (2015): "Analysis of binary multivariate longitudinal data via 2-dimensional orbits: an application to the Agincourt health and socio-demographic surveillance system in South Africa," *PloS one*, 10, e0123812.

VITALI, S., J. B. GLATTFELDER, AND S. BATTISTON (2011): "The network of global corporate control," *PloS one*, 6, e25995.

WALKER, S. G. (2014): "Computing Marginal Likelihoods via Posterior Sampling," *Communications in Statistics-Simulation and Computation*, 43, 520–527.

WAND, M. P. AND M. C. JONES (1994): *Kernel smoothing*, CRC Press.

WANG, H. (2010): "Sparse seemingly unrelated regression modelling: Applications in finance and econometrics," *Computational Statistics & Data Analysis*, 54, 2866–2877.

——— (2012): "Bayesian graphical lasso models and efficient posterior computation," *Bayesian Analysis*, 7, 867–886.

Wang, H., C. Reeson, C. M. Carvalho, et al. (2011): "Dynamic financial index models: modeling conditional dependencies via graphs," *Bayesian Analysis*, 6, 639–664.

Wang, H. and M. West (2009): "Bayesian analysis of matrix normal graphical models," *Biometrika*, 96, 821–834.

Wang, L., D. Durante, R. E. Jung, and D. B. Dunson (2017): "Bayesian network–response regression," *Bioinformatics*, 33, 1859–1866.

Wang, Y. J. and G. Y. Wong (1987): "Stochastic blockmodels for directed graphs," *Journal of the American Statistical Association*, 82, 8–19.

Wasserman, S. and K. Faust (1994): *Social network analysis: methods and applications*, Cambridge University Press.

Watts, D. J. and S. H. Strogatz (1998): "Collective dynamics of "small-world" networks," *Nature*, 393, 440.

Weber, C. L. and H. S. Matthews (2007): "Embodied environmental emissions in US international trade, 1997- 2004," *Envoronmental Science & Technology*, 41, 4875–4881.

Wehmuth, K., A. Ziviani, and E. Fleury (2015): "A unifying model for representing time-varying graphs," in *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, IEEE, 1–10.

Weiss, G. N. and H. Supper (2013): "Forecasting liquidity-adjusted intraday value-at-risk with vine copulas," *Journal of Banking & Finance*, 37, 3334–3350.

White, H. and X. Lu (2010): "Granger causality and dynamic structural systems," *Journal of Financial Econometrics*, 8, 193–243.

Whittaker, J. (2009): *Graphical models in applied multivariate statistics*, Wiley Publishing.

Wilbur, J., J. Ghosh, C. Nakatsu, S. Brouder, and R. Doerge (2002): "Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial community DNA fingerprints," *Biometrics*, 58, 378–386.

Williamson, S. (2016): "Nonparametric network models for link prediction," *Journal of Machine Learning Research*, 17, 1–21.

Wixted, B., N. Yamano, and C. Webb (2006): "Input-output analysis in an increasingly globalised world: applications of OECD's harmonised international tables," Tech. rep., OECD Publishing.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT press.

Xu, T., Z. Yin, T. Siliang, S. Jian, W. Fei, and Z. Yueting (2013): *Logistic tensor regression for classification*, Springer, vol. 7751, chap. Intelligent science and intelligent data engineering, 573–581.

Yao, F., H.-G. Müller, and J.-L. Wang (2005): "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, 100, 577–590.

Yoshida, R. and M. West (2010): "Bayesian learning in sparse graphical factor models via variational mean-field annealing," *Journal of Machine Learning Research*, 11, 1771–1798.

Yuan, M. and Y. Lin (2007): "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94, 19–35.

YUE, Y. R., M. A. LINDQUIST, AND J. M. LOH (2012): "Meta-analysis of functional neuroimaging data using Bayesian nonparametric binary regression," *The Annals of Applied Statistics*, 697–718.

ZELLNER, A. (1962): "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," *Journal of the American Statistical Association*, 57, 348–368.

ZHANG, X., L. LI, H. ZHOU, AND D. SHEN (2014): "Tensor generalized estimating equations for longitudinal imaging analysis," *arXiv preprint arXiv:1412.6592*.

ZHANG, X., C. MOORE, AND M. E. NEWMAN (2017): "Random graph models for dynamic networks," *The European Physical Journal B*, 90, 200.

ZHAO, Q., L. ZHANG, AND A. CICHOCKI (2013): "A tensor-variate Gaussian process for classification of multidimensional structured data," in *Twenty-seventh AAAI conference on artificial intelligence*.

ZHAO, Q., G. ZHOU, L. ZHANG, AND A. CICHOCKI (2014): "Tensor-variate Gaussian processes regression and its application to video surveillance," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, 1265–1269.

ZHOU, H., L. LI, AND H. ZHU (2013): "Tensor regression with applications in neuroimaging data analysis," *Journal of the American Statistical Association*, 108, 540–552.

ZHOU, J., A. BHATTACHARYA, A. H. HERRING, AND D. B. DUNSON (2015): "Bayesian factorizations of big sparse tensors," *Journal of the American Statistical Association*, 110, 1562–1576.

ZHOU, S., J. LAFFERTY, AND L. WASSERMAN (2010): "Time varying undirected graphs," *Machine Learning*, 80, 295–319.

ZHU, Z., F. CERINA, A. CHESSA, G. CALDARELLI, AND M. RICCABONI (2014): "The rise of China in the international trade network: a community core detection approach," *PloS one*, 9, e105496.