Università degli Studi di Padova

UNIVERSITÀ DEGLI STUDI DI PADOVA

Department of Statistical Sciences

Ph.D. Course in Statistics

Cycle XXXIV

Bayesian infinite factorization methods with applications to tracking data in football

Course Coordinator:
Prof. Nicola Sartori

Supervisor:
Prof. Antonio Canale

Candidate:
Lorenzo Schiavon

2022

Supervisor: *Prof.* Antonio Canale

Abstract

Factorization models are a mathematical representation of multidimensional data objects as a collection of simpler components. For instance, a matrix can be characterized as a sum of latent rank one components, where the number of addends is generally much lower than the dimensions of the matrix. Factor models are commonly used across a variety of disciplines to deal with data sets whereby a large number of observed variables is thought to reflect a smaller number of latent variables. However, it can be challenging to infer the relative impact of the different components as well as the number of components. To address this issue, it has become popular to rely on overfitted factorization models that avoid strict constraints on either the number of factors and the ordering of the data. In the Bayesian framework, increasing shrinkage priors on latent elements have been proposed, allowing the introduction of infinitely many factors, albeit with impact decreasing with the component index, such that the unnecessary ones can be adaptively removed by increasingly shrinking their coefficients close to zero as the component index increases. These flexible approaches are usually named infinite factorization models.

This thesis aims to provide an overview on infinite factorization models, presenting the state of the art, discussing the limitations of the current models, and gradually composing a general Bayesian infinite factorization framework that includes novel methods to address such deficiencies. In particular, we consider the role of sparsity in the latent low-rank elements, as being crucial to improve the inference and facilitate interpretation. Firstly, we focus on the effect of the sparsity induced by the usual approximation of the infinite model through a truncated version to facilitate the posterior inference. In this regard, it is fundamental to carefully assess how the truncation criterion affects the inference performance and the factor model representation. We propose a novel truncation criterion that relates the level of truncation to the factor contribution to the global data variability, allowing one to easily calibrate the algorithm's parameters. Secondly, we

careful investigate the role of local sparsity within the low-rank latent elements by introducing a new general class of infinite factorization models. In this framework, we provide theoretical support to verify desirable shrinkage properties of the prior, including robustness to large signals and the sparsity behaviour to the increasing number of factors or dimension of the data. The main novelty of the proposed class of models lies on the dependence of the local sparse pattern of the latent elements on auxiliary information which is supposed to inform on the similarity among variables, that correspond to columns of the data matrix. This structure enables us to fill a key gap of the current infinite factor models that do not accommodate grouped variables and other nonexchangeable structures. We also propose extending this class to the more general class of matrix decomposition models. Symmetrically to the use of the exogenous information about variables, the matrix decomposition model also embeds auxiliary information about the row entities of the data matrix, enabling us to model the dependence through structured sparse latent elements with respect to both the matrix dimensions. A novel estimation algorithm inspired by boosting approaches is designed, overcoming the computational limits of the current Markov chain Monte Carlo approaches and the nonidentifiability issue which characterizes all the overfitted factorization models.

Practical gains with respect to the current state of art are demonstrated in simulation studies and discussed in real data applications, further illustrating benefits in terms of parameter estimations and model interpretation. Football player tracking data represent the common thread of the thesis. They motivate the introduction of the novel methodologies to address the challenges arising from the need of extracting valuable knowledge from a high dimensional dataset representing a complex phenomenon. The amount of information included is such that several aspects of interest can be explored. In this thesis, we focus on three of them: similarities among players, positional and technical predictors of the dangerousness of an action, and player run heatmaps. In all these cases, thoughtful insights and representations are provided, sheding ligth on the potential of our approach. However, the generality of the proposed framework is expected to impact many other application fields.

Sommario

I modelli fattoriali sono una rappresentazione matematica di dati multidimensionali tramite una collezione di oggetti più semplici Per esempio, una matrice di dati può essere descritta da una somma di componenti latenti a rango uno, dove il numero di componenti è solitamente molto più piccolo delle dimensioni della matrice. I modelli fattoriali vengono utilizzati frequentemente per l'analisi di dati in varie discipline, quando si suppone che un insieme di variabili osservate sia esprimibile con un numero più piccolo di variabili latenti. Ad ogni modo, può risultare molto difficile capire il numero e il peso delle diverse componenti latenti. Per rispondere a questo problema, si sta diffondendo l'utilizzo di modelli fattoriali sovra parametrizzati che evitano l'imposizione di vincoli sia sul numero di fattori che sull'ordinamento dei dati. Nel contesto bayesiano, si sono affermate delle distribuzioni a priori con compressione crescente che permettono di avere infiniti fattori, ma con impatto decrescente rispetto all'indice di componente, in modo tale che i fattori non necessari vengano rimossi comprimendone a zero i rispettivi coefficienti, in misura tanto maggiore al crescere dell'indice di componente. Questi modelli flessibili sono generalmente identificati con il nome di modelli infinito fattoriali.

Questa tesi si pone l'obbiettivo di fornire una panoramica sui modelli infinito fattoriali, presentandone lo stato dell'arte, discutendone i limiti e costruendo in modo incrementale una struttura generale per modelli bayesiani infinito fattoriali che includa nuovi metodi per sopperire a tali mancanze. In particolare, la tesi tratta il ruolo della sparsità negli elementi latenti di basso rango, in quanto cruciale per migliorare l'inferenza e facilitare l'interpretazione del modello. Inizialmente, focalizziamo la nostra attenzione sull'effetto della sparsità indotta dall'usuale approssimazione dei modelli a infiniti fattori dovuta a troncamento, effettuata per facilitare l'inferenza sulla distribuzione a posteriori. A tal proposito, è importante valutare attentamente come il criterio di troncamento influisca sulla capacità di inferenza e sulla rappresentazione del modello. Proponiamo

quindi un nuovo criterio di troncamento che pone in relazione il livello a cui viene troncato il modello con il contributo dei fattori alla spiegazione della variabilità totale dei dati, permettendo così di calibrare più facilmente i parametri dell'algoritmo. In secondo luogo, analizziamo il ruolo della sparsità locale negli elementi latenti a basso rango introducendo una nuova classe generale di modelli a infinito fattori. In questo scenario, forniamo gli strumenti teorici per verificare delle proprietà di compressione della distribuzione a priori, tra le quali la robustezza ai segnali evidenti e il comportamento della sparsità al crescere del numero di fattori o della dimensione dei dati. La maggior novità della classe di modelli proposta risiede nella specificazione della struttura di sparsità degli elementi latenti come dipendente da informazione ausiliaria che informi circa la similarità tra le variabili, che corrispondono alle colonne della matrice di dati. Questa specificazione permette di rispondere ad uno dei punti aperti degli attuali modelli a infiniti fattori che non permettevano la possibilità di indurre gruppi o altre strutture tra variabili. Proponiamo anche di estendere questa classe alla più generale classe di modelli per decomposizione di matrici. In modo simmetrico rispetto a quanto fatto con l'informazione esogena sulle variabili, il modello per la decomposizione di matrici include informazione aggiuntiva riguardante anche le righe della matrice di dati, consentendo di modellare la dipendenza lungo entrambe le dimensioni della matrice tramite elementi latenti con sparsità strutturata. Si definisce un nuovo algoritmo di stima ispirato dai metodi *boosting*, superando i limiti computazionali dei metodi basati su catene di Markov Monte Carlo e il problema di non-identificabilità che caratterizza tutti i modelli fattoriali sovra-parametrizzati.

I dati di tracciamento dei giocatori di calcio rappresentano il filo conduttore della tesi. Motivano l'introduzione dei nuovi metodi come risposta alle problematiche poste dal dover estrarre conoscenza da un insieme di dati ad alta dimensionalità che descrive un fenomeno complesso. La quantità di informazione in questi dati è tale che vi sono diversi aspetti di interesse da esplorare. In questa tesi, ci concentriamo su tre di essi: similarità tra giocatori, predittori posizionali e tattici della pericolosità di un'azione e *heatmaps* di corsa dei giocatori.In tutti questi casi, vengono fornite riflessioni approfondite e rappresentazioni che mettono in luce le potenzialità del nostro metodo. Ad ogni modo, data la generalità dell'approccio proposto, è lecito aspettarsi che vi sia un impatto su molti altri campi di applicazione.

Acknowledgements

First and foremost, I am extremely grateful to my supervisor Antonio Canale who has steered me through the world of research, by always supporting my ideas while providing feedbacks, insights and guidance.

I would like to express my gratitude to professors David Dunson and Bernardo Nipoti whose deep and thoughtful comments and edits has made possible for me to feel proud of this thesis.

I wish to acknowledge professor Otso Ovaskainen, Gleb Tikhonov, and Sirio Legramanti for their useful feedbacks on early versions of the chapters of this thesis.

I would also like to thank Daniele Durante and professor Anirban Bhattacharya for their precious reviews that have helped to improve this thesis.

Many thanks to Martino Tenconi, and Math&Sport for the kindness and confidence shown in supporting this project.

Furthermore, I want to thank the people of the department of Statistical Sciences in Padova, which has represented my second home in the last 9 years.

I would also like to thank my colleagues Nicolas, Cristian, Laura, Emanuele, Silvia, Jacopo, Dung, Anam, Jacopo, Mattia, Anna, and Fabio for having shared this trip.

To conclude, I cannot forget to thank Linda, my friends, and my family for all the unconditional support: they have been the real latent factors of this thesis.

Contents

List of Figures

List of Tables

1 Introduction

1 Overview

1.1 *Tracking data in football*

Data and statistics about football are commonly based on ball-related events, due to the manual system of data collection. Recently, advances in computer vision techniques have made it possible to automatically track every player on the pitch at discrete but very frequent time points. It should be emphasised that although these methods were introduced around ten years ago (Barros et al., 2007; Liu et al., 2009), it is only in the last couple of years that they have started to be routinely used. The time points frequency is such that the positions and the velocities of all the players and the ball are available at multiple times in a second, providing all necessary information to describe any game situation. For instance, we may be interested in estimating the probability that a certain shot is scored, given all the available information at the moment the shot is taken. Models to address this issue, usually named expected goals models, have been developed both in academia (e.g. Pollard & Reep, 1997) and by football analysts (Caley, 2015), but they have never relied on full tracking data, leading to well-known biases (Mackay, 2017). Data including the positions and velocities of all players on the pitch should now allow shot situations to be distinguished in terms of objective scoring probability. Despite recent remarkable attempts (Fernández et al., 2019), we are far from achieving such an objective because of the difficulty in handling such huge quantities of data.

In this thesis, we have access to new-generation tracking data provided by MathAndSport[1]. This motivates the development of statistical methods that aim to be able to extract and summarize valuable knowledge from the large amount of information available to provide useful and easily

[1] MathAndSport s.r.l. is a sport analytics company based in Milan: https://www.mathandsport.com/

accessible insights in such a new and complex context. Each player or action in the game can now be described by a large number of indicators. When the interest of the coaching staff or media is focused on a single aspect of football, the proliferation of data and indices represents an undoubted advantage, since it is likely that a suitable indicator addressing such a specific aspect exists. Nevertheless, football insiders are more often interested in detecting general traits and macro-trends to evaluate the impact of strategic decisions or assess and compare different players. Recognizing and isolating such traits might be challenging. To address this issue, we will focus on factorization models that are routinely used to express large statistical objects in terms of simpler components. Due to the broad applicability of the factorization models and to the generality of the methodology, in this thesis we will develop models and algorithms that are expected to impact many application fields beyond football player tracking data.

1.II *Factorization models*

Factorization models are a well known mathematical representation used across a variety of disciplines, and are based on the idea that one can more easily characterize structure in complex data by utilizing a collection of simple components. For example, a matrix or tensor can be characterized as a sum of rank one components. Suppose that a $n \times p$ matrix of data y is available, where $i = 1, \dots, n$ indicates the subjects, and $j = 1 \dots, p$ indicates the variables. The likelihood for y under a general class of factorization models can be expressed as $L(y; H, \Lambda, \Sigma)$, where $H = \{\eta_{ih},\, i = 1, \dots, n\, h = 1, \dots, k\}$ and $\Lambda = \{\lambda_{jh},\, j = 1, \dots, p,\, h = 1, \dots, k\}$ are matrices with rank k, and Σ is a matrix of additional parameters. There are many important special cases of this class, including Gaussian linear factor models (see Roweis & Ghahramani, 1999, for a discussion), exponential family factor models (Jun & Tao, 2013), Gaussian copula factor models (Murray et al., 2013), latent factor linear mixed models (An et al., 2013), probabilistic matrix factorization (Salakhutdinov & Mnih, 2008), underlying Gaussian factor models for mixed scale data (Reich & Bandyopadhyay, 2010), and functional data factor models (Montagna et al., 2012).

To be effective in dimensionality reduction, the number of simpler components k is generally much lower than $\min(n, p)$. In this case, the factor decomposition is condensing the available information into latent aggregates, which could be potentially meaningful. In view of this, factor models are commonly used in many applications to deal with data sets whereby a large number of observed variables p is thought to reflect a smaller number k of latent variables, especially if the latent constructs present an easy interpretation, as the individual skills and behaviours in

psychology (see, e.g. Fabrigar et al., 1999, for an extended review) or economics (Heckman et al., 2006). In studying biological activity (Carvalho et al., 2008) and diseases (West, 2003), factors analysis has been largely applied to summarize gene expression through latent aggregates that are possibly associated to variables of interest.

A small number of components k enables us to obtain a sparse representation of an object of interest, which is particularly attractive when n and p are huge, even when we are not interested in bringing hidden and meaningful relations to light. For instance, the envelope model (Cook et al., 2010) for multivariate regression is based on the key assumption that some aspects of the response vector are stochastically constant as the p predictors vary, meaning that a coefficient matrix can be represented in a k-dimensional subspace of its column space. This is equivalent to representing the coefficient matrix as a product between a smaller rank matrix and an orthonormal basis of the subspace. The choice of k is crucial in order to represent the minimum subspace that is relevant to the regression.

More generally, in factorization models, it can be challenging to infer the relative impact of the different components as well as the number of components k. Although there is a rich literature, selection of k is far from a solved problem. In unsupervised settings, it is common to fit the model for different choices of k and then choose the value with the best goodness-of-fit criteria. For likelihood models, the Bayesian information criteria is particularly popular. It is also common to use an informal elbow rule, selecting the smallest k such that the criteria improve only a small amount for $k+1$. In specific contexts, formal model selection methods have been developed. For example, taking a Bayesian approach, one can choose a prior for k and attempt to approximate the posterior distribution of k using Markov chain Monte Carlo; see Lopes & West (2004) for linear factor models, Miller & Harrison (2018) for mixture models, and Yang et al. (2018) for matrix factorization. Although such methods are conceptually appealing, their computation can be prohibitive outside of specialized settings. A popular idea to address this issue is by including infinitely many components having impact decreasing with the component index. Such flexible approaches were proposed by Rousseau & Mengersen (2011) for mixture models and Bhattacharya & Dunson (2011) for Gaussian linear factor models, and they are usually named infinite factorization models.

In this thesis, we will focus on Bayesian factorization models characterized by the underlying Gaussian structure

$$y = f(z), \quad z = H\Lambda^\top + \epsilon, \quad \text{vec}(\epsilon) \sim N_{np}(0, \Sigma), \tag{1.1}$$

with H an $n \times k$ factor matrix, Λ a $p \times k$ loadings matrix, ϵ an independent error matrix, and where f is a deterministic transformation and $N_q(\mu, \Sigma)$ denotes the multivariate Gaussian distribution with mean μ and variance matrix Σ.

Class (1.1) includes most of the cases mentioned above. For instance, when the row vectors $\epsilon_{i\cdot}$ and $\eta_{i\cdot}$ of ϵ and H are Gaussian random vectors and f is the identity function, model (1.1) is a Gaussian linear factor model. With similar assumptions for $\epsilon_{i\cdot}$ and $\eta_{i\cdot}$, and assuming the transformations $y_{ij} = F_j^{-1}\{F_N(z_{ij})\}$, with $F_N(z_{ij})$ the cumulative distribution function of the standard Gaussian distribution, model (1.1) is a Gaussian copula factor model (Murray et al., 2013). Exponential family factor models (Jun & Tao, 2013), probabilistic matrix factorization (Salakhutdinov & Mnih, 2008), and underlying Gaussian models for mixed-scale data (Reich & Bandyopadhyay, 2010) can be obtained by appropriately defining the elements in (1.1), whereas multivariate response regression models belong to this framework when H depends on a covariate matrix x.

Consistent with the literature on Bayesian factor models (see, e.g. Arminger & Muthén, 1998), we rely on error terms with diagonal covariance matrix Σ with inverse gamma priors on the diagonal elements and specify prior distributions on H and Λ. Recently, suitable increasing shrinkage priors on Λ have been proposed (Bhattacharya & Dunson, 2011; Legramanti et al., 2020), allowing the introduction of infinitely many factors, namely $k = \infty$, with the loadings elements increasingly shrunk towards zero as the component index increases. This allows us to specify an infinite factorization model that can be accurately approximated through a truncated version, which facilitates the inference computation.

II Main contributions of the thesis

Despite the spread of the infinite factorization models, existing methods present a lack of careful consideration of the structure induced by the prior, limiting the use of such models in practical applications. Firstly, the truncation criteria currently applied in the increasing shrinkage prior are heuristic and not invariant to the scale of the data. This leads to difficulties in calibrating the algorithm's parameters with negative consequences on the representation of large data through a small number of components. Secondly, the literature lacks investigation on how to handle sparsity and, in particular, how to induce shrinkage structure. A careful consideration of these aspects could help to detect hidden relation patterns among variables and to promote an easier

interpretation of the latent components. Finally, practical uses of the current increasing shrinkage priors are limited both by slow estimation methods and the fact that are thought to address model where the dependence structure among subjects $i = 1, \dots, n$ is carefully and strictly specified.

Motivated both by these limitations and the new challenges raised by the advances in football data collection technology, in this thesis we gradually compose a general Bayesian infinite factorization framework that includes novel methods to address such limitations and effectively apply factor models to football player tracking data as well as to other contexts. In each chapter, practical gains are demonstrated in simulation studies and discussed in real data applications, further illustrating benefits in terms of parameter estimations and model interpretation. Specifically, the thesis is organized as follows.

In Chapter 2, we present an overview of the current increasing shrinkage priors that are specifically designed to include infinitely many factors $k = \infty$ in the Gaussian linear factor model

$$y_i = \Lambda \eta_{i\cdot} + \epsilon_{i\cdot}, \quad \eta_{i\cdot} \sim N_k(0, \Psi), \quad \epsilon_{i\cdot} \sim N_p(0, \Sigma), \quad \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2), \quad (1.2)$$

with Ψ a $k \times k$ covariance matrix. Assuming the error term $\epsilon_{i\cdot}$ independent of $\eta_{i\cdot}$, the matrix $\Omega = \text{var}(y_{i\cdot})$ can be expressed as $\Omega = \Lambda \Psi \Lambda^\top + \Sigma$. These methods are designed to facilitate posterior computation via simple Gibbs sampling algorithms, approximating the infinite model through a truncated version. In this regard, it is important to carefully consider which truncation criterion should be adopted. A discussion about this is outlined in this chapter, as well as a new proposed truncation criterion. Such a criterion allows a more intuitive and general way to calibrate the algorithm's parameters relating the level of truncation to the factor contribution to the global data variability. Following this idea, it is easy to note the remarkable importance of using methods that are invariant to the scale of the data. Part of the results reported in the second chapter are presented in Schiavon & Canale (2020).

Chapter 3 presents a new general class of infinite factor models, named generalized infinite factorization models, which is designed for underlying Gaussian factor models $y_{ij} = f_j(z_{ij})$ with $z_{i\cdot} = \Lambda \eta_{i\cdot} + \epsilon_{i\cdot}$. In this framework, we report a careful discussion on the shrinkage properties of increasing shrinkage priors. Based on the rich literature about shrinkage priors outside the factorization context, we provide theoretical support to verify desirable properties in our general framework. These properties include robustness of the prior to large signals and the asymptotic behaviour of the prior to the increasing dimension of the data. Another key gap of the existing methods is the lack of accommodation for grouped variables and other nonexchangeable structures.

We address this issue in the new class of models defining the prior on Λ as dependent on a meta-covariate matrix w informing about the sparsity similarities among the different variables $j = 1, \ldots, p$. Simulation and application studies show the benefits of using our approach with respect to the current state of the art both in terms of variance inference and model interpretation. Part of the results reported in the third chapter are presented in Schiavon et al. (in press) and Schiavon & Canale (2021).

In Chapter 4, we propose extending the generalized infinite factorization models approach to the more general class of matrix decomposition models. Symmetrically to the prior on Λ, we define a shrinkage prior on the elements of H as depending on a covariate matrix x, promoting the identification of complex dependence structures related to similarities among the subjects. To estimate the model, we design a novel algorithm inspired by gradient boosting approaches, overcoming the computational limits of the current Markov chain Monte Carlo approaches with which we also come up against in Chapter 3. In addition, this algorithm naturally handles and solves the nonidentifiability issue which characterizes all the overfitted factorization models. The application of the algorithm to a dataset of player tracking heatmaps provides thoughtful insights and representations of a complex phenomenon, enlightening on the potential of this approach with high-dimensional data.

Finally, Section 5 provides some final remarks on the achievements of this thesis and on future possible developments.

2 Infinite factorization models

1 Background

In recent decades, the sparse structure of the variance matrix characterizing factor models has been particularly attractive for dimensionality reduction purposes, and several recent works have mainly concerned on estimating the variance and covariance matrix, rather than the factor loadings (see Kastner, 2019, for an example in econometrics). Because of this fact and to the aforementioned challenges from the choice of the number of components k, it has become popular to rely on overfitted factorization models that avoid strict constraints on either the number of factors and the ordering of the data (Frühwirth-Schnatter & Lopes, 2018), since a unique identification of the loadings matrix is not necessary. The Bayesian approach proposed by Bhattacharya & Dunson (2011) for Gaussian linear factor models considers more than enough factors, albeit with shrinkage priors that adaptively remove unnecessary ones by shrinking their coefficients close to zero. Given model (1.2), the authors specify the variance of each element λ_{jh} $(j = 1, \dots, p; h = 1, \dots, k)$ as the product between a local scale ϕ_{jh} and a factor-specific scale θ_h. An increasing shrinkage behaviour on the columns of Λ through a multiplicative gamma process prior on θ_h $(h = 1, \dots, \infty)$ is imposed, allowing the introduction of infinitely many factors, $k = \infty$. Specifically, the prior on Λ proposed by Bhattacharya & Dunson (2011) can be written as

$$\lambda_{jh} \mid \phi_{jh}, \theta_h \sim N(0, \phi_{jh}\theta_h), \quad \theta_h = \vartheta_h \rho_h, \tag{2.1}$$
$$\rho_h = \prod_{l=0}^{h-1} \vartheta_l, \quad \vartheta_0 = 1, \quad \vartheta_1^{-1} \sim \mathrm{Ga}(a_1, b_1), \quad \vartheta_m^{-1} \sim \mathrm{Ga}(a_2, b_2), \ m \geq 2.$$

Expression (2.1) induces a class of scale-mixture of Gaussian shrinkage priors (Polson & Scott, 2010) for the loadings. Under this construction, the columns of Λ progressively play a less important role in characterizing the covariance structure of the data and their value is increasingly shrunk towards zero, so that we can control and discard redundant factors. The authors specify an inverse gamma prior on the local scale, setting $\phi_{jh}^{-1} \sim \text{Ga}(\nu/2, \nu/2)$, where $\text{Ga}(a, b)$ denotes a gamma distribution with mean a/b and variance a/b^2.

Similar in spirit is the more recent cumulative stick-breaking process proposed by Legramanti et al. (2020), who introduced a spike-and-slab structure (Mitchell & Beauchamp, 1988) that increases the mass on the spike for later columns. Thanks to the convenient notation we introduced, the cumulative stick-breaking prior can be represented under the same setting (2.1), fixing local scales to one and defining factor-specific scale priors

$$\vartheta_h^{-1} \sim \text{Ga}(a_\theta, b_\theta), \quad \rho_h = \text{Ber}(1 - \pi_h), \quad h = 1, \ldots, \infty,$$

with $\text{Ber}(p)$ denoting a Bernoulli distribution with mean p. The process $\pi_h = \text{pr}(\theta_h = 0)$ $(h = 1, \ldots, \infty)$ follows a stick-breaking construction,

$$\pi_h = \sum_{l=1}^{h} u_l, \quad u_l = v_l \prod_{m=1}^{l-1} v_m, \quad v_m \sim (1, \alpha), \tag{2.2}$$

with $\text{Be}(a, b)$ the beta distribution with mean $a = a/(a + b)$, such that $\pi_{h+1} > \pi_h$ is guaranteed for any $h = 1, \ldots, \infty$ and $\lim_{h\to\infty} \pi_h = 1$ almost surely. The main difference with respect to the multiplicative gamma process lies on the separation between the parameters that control the rate of shrinkage of redundant factors from those regulating the magnitude of the nonneglectable factors. To induce a continuous shrinkage prior on every λ_{jh}, the authors suggested possibly adjusting the definition of θ_h as $\theta_h = \rho_h(\vartheta_h - \theta_\infty) + \theta_\infty$, where θ_∞ is positive but close to zero.

In both models, representing the current state of art, the increasing shrinkage allows one to accurately approximate the likelihood $L(y; \Lambda, \Psi, \Sigma)$ by $L(y; \Lambda_{k^*}, \Psi_{k^*}, \Sigma)$, with Λ_{k^*} containing the first k^* columns of the infinite matrix Λ and Ψ_{k^*} the first k^* rows and columns of Ψ, lessening the computational burden. Posterior inference for truncated infinite factor models can be conducted in different ways. For fixed truncation k^*, one could run separate Gibbs samplers for each value of k^* to choose the best k^*. Clearly, this approach would lead to computational hurdles, especially for large p, where a large grid of possible k^* values should be tested. An alternative would be to consider a varying k^* and implement a reversible jump Markov chain Monte Carlo approach (Lopes

& West, 2004), a formally valid solution that still remains not convenient from the computational viewpoint. A computationally efficient proposal introduced in Bhattacharya & Dunson (2011) and followed in Legramanti et al. (2020) consists in defining an adaptive Gibbs sampler that attempts to infer the best value of k^* while it runs. In this way, a single run of the algorithm is sufficient to choose the number of latent factors and to draw from the posterior distributions of the parameters. The value of k^* is adapted only at some Gibbs iterations by discarding redundant factors and, if no redundant factors are identified, by adding a new factor by sampling its parameters from the prior distribution. Convergence of the Markov chain is guaranteed by satisfying the diminishing adaptation condition in Theorem 5 of Roberts & Rosenthal (2007), by specifying the probability of occurrence of an adaptive iteration t as equal to $\text{pr}(t) = \exp(\alpha_0 + \alpha_1 t)$, where α_0 and α_1 are negative constants, such that frequency of adaptation decreases exponentially fast. In every adaptive iteration, redundant factors are discarded according to a specific criterion. If no redundant factors are identified, a new factor is added by sampling its parameters from the prior distribution. The truncation criteria clearly play a central role in determining the final number of factors and, consequently, the posterior inference. In the adaptive Gibbs sampler for the multiplicative gamma process factor model, for example, the authors considered as redundant those factors characterized by loadings having all their elements λ_{jh} within a neighbourhood of zero. In other terms, at each adaptive iteration, the quantity $m_h = \max_{1 \leq j \leq p}\{|\lambda_{jh}|\}$ of every column h of Λ is compared with a pre-determined threshold ζ, dropping the columns where $m_h < \zeta$. We will argue that the choice of the tolerance parameter ζ is a delicate issue and, motivated by a lack of clear guidelines for choosing such a value, we propose an alternative truncation criterion for the multiplicative gamma process model. We will show that our proposal is interpretable, robust with respect to the dimension and scale of the data, and able to identify all important factors. Our contribution further illustrates the importance of using criteria that are invariant to the scale of the data when we apply any infinite factor model, including the cumulative shrinkage process factor model.

11 A NOVEL TRUNCATION CRITERION FOR INFINITE FACTORIZATION MODELS

11.1 *A new proposal*

Our goal is to define a criterion that is interpretable, robust with respect to the dimension and scale of the data, and able to identify all important factors. Under the class of models (1.2), let Λ_{k^*} denote the matrix obtained by discarding the columns of Λ from k^*+1 onwards, or, equivalently, setting them to zero. Our idea consists of truncating Λ such that the induced truncated model $\Omega_{k^*} = \Lambda_{k^*}\Psi_{k^*}\Lambda_{k^*}^\top + \Sigma$ is able to explain at least a fraction $\tau \in (0,1)$ of the total variability of the data. To measure the induced truncation error of Ω_{k^*}, we use the trace of Ω. The trace is justified by the fact that the maximum error occurring in an element of Ω due to truncation always lies along the diagonal and by the relation between the difference of traces and the nuclear norm, routinely used to approximate low rank minimization problems (Liu & Vandenberghe, 2010).

Heuristically, we would like to have

$$\frac{\mathrm{tr}(\Lambda_{k^*}\Psi_{k^*}\Lambda_{k^*}^\top) + \mathrm{tr}(\Sigma)}{\mathrm{tr}(\Omega)} \geq \tau, \tag{2.3}$$

but we will make this rule concrete later. In this way, the number of columns in Λ is still affected by a subjective choice, the value of τ, but this can be decided according to a criterion that is consistent with what is commonly done in other similar contexts (e.g. principal component analysis) and independently from the value of p and the scale of the data.

As formal justification for this approach, we obtain an upper bound on the probability that condition (2.3) is not satisfied. The following proposition provides conditions on prior (2.1) so that the underestimation of it occurs by truncating decreases exponentially fast as k^* increases.

PROPOSITION 2.1: *Let $E(\phi_{jh})$ be finite for $j = 1,\dots,p$ and $h = 1,\dots,\infty$ and $E(\theta_h) = ab^{h-1}$ with $a > 0$ and $b \in (0,1)$ for all $h = 1,\dots,\infty$. Let $c > 0$ be a sufficiently large number such that $c \geq \max_{h=1,\dots,\infty}\psi_{hh}$. If*

$$m_\Omega = \min_{j=1,\dots,p}\left[E(\sigma_j^{-2}),\ E\left\{\left(\sum_{h=1}^{\infty}\psi_{hh}\lambda_{jh}^2\right)^{-1}\right\}\right] < \infty,$$

then for any $\tau \in (0,1)$,

$$\text{pr}\left\{\frac{\text{tr}(\Omega_{k^*})}{\text{tr}(\Omega)} \leq \tau\right\} \leq \left(\frac{1}{1-\tau}\right) ac\frac{b^H}{1-b}\, m_\Omega \sum_{j=1}^{p} E(\phi_{j1}).$$

Proof of Proposition 2.1. The trace of Ω is $\text{tr}(\Sigma)+\text{tr}(\Lambda_{k^*}\Psi_{k^*}\Lambda_{k^*}^{\text{T}})+\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})$, where $\Lambda_{\Delta_{k^*}} = \Lambda - \Lambda_H$ and $\Psi_{\Delta_{k^*}} = \Psi - \Psi_H$. Hence, it is equivalent to rewriting the probability of interest as

$$\text{pr}\left\{\frac{\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})}{\text{tr}(\Omega)} \geq 1-T\right\}.$$

By Markov's inequality

$$\text{pr}\left\{\frac{\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})}{\text{tr}(\Omega)} \geq 1-\tau\right\} \leq E\left\{\frac{\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})}{\text{tr}(\Omega)}\right\}/(1-\tau).$$

The expected ratio of two random variables u and v is $E(u/v) = \text{cov}(u, 1/v) + E(u)E(1/v)$, which allows us to write $E(u/v) \leq E(u)E(1/v)$ if $\text{cov}(u, 1/v) \leq 0$. Then, since the covariance between $\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})$ and $\text{tr}(\Omega)$ is nonnegative, the following inequality holds

$$E\left\{\frac{\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})}{\text{tr}(\Omega)}\right\} \leq E\{\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})\}E\left(\frac{1}{\text{tr}(\Omega)}\right).$$

The trace $\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})$ is equal to $\sum_{j=1}^{p}\sum_{h=H+1}^{\infty}\psi_{hh}\lambda_{jh}^2$. The variance of λ_{jh} is $E(\lambda_{jh}^2) = E(\phi_{j1})E(\theta_h)$. Let c satisfy $c \geq \max_{h=1,\ldots,\infty}\psi_{hh}$. Since $E(\phi_{j1})$ is finite and $E(\theta_h) = ab^{h-1}$ with a, b positive constants and $b < 1$, then

$$E\{\text{tr}(\Lambda_{\Delta_{k^*}}\Psi_{\Delta_{k^*}}\Lambda_{\Delta_{k^*}}^{\text{T}})\} \leq ca\frac{b^H}{1-b}\sum_{j=1}^{p}E(\phi_{j1}).$$

Since $\text{tr}(\Omega) = \text{tr}(\Lambda\Psi\Lambda^{\text{T}}) + \text{tr}(\Sigma)$, we know that $\text{tr}(\Omega) \geq \sum_{h=1}^{\infty}\psi_{hh}\lambda_{jh}^2 + \sigma_j^2$ for any j in $1,\ldots,p$, where σ_j^2 is the jth diagonal element of Σ. Then, for any j in $1,\ldots,p$, we obtain

$$\frac{1}{\text{tr}(\Omega)} \leq \frac{1}{\sum_{h=1}^{\infty}\psi_{hh}\lambda_{jh}^2 + \sigma_j^2},$$

and, consequently,

$$E\left\{\frac{1}{\text{tr}(\Omega)}\right\} \leq E\left(\sigma_j^{-2}\right), \qquad E\left\{\frac{1}{\text{tr}(\Omega)}\right\} \leq E\left\{\left(\sum_{h=1}^{\infty}\psi_{hh}\lambda_{jh}^2\right)^{-1}\right\}.$$

Therefore, since $m_\Omega = \min_{j=1,\ldots,p}\left[E(\sigma_j^{-2}), E\left\{\left(\sum_{h=1}^{\infty}\psi_{hh}\lambda_{jh}^2\right)^{-1}\right\}\right] < \infty$, then

$$\text{pr}\left\{\frac{\text{tr}(\Lambda_{k^*}\Psi_{k^*}\Lambda_{k^*}^{\text{T}}) + \text{tr}(\Sigma)}{\text{tr}(\Omega)} \leq \tau\right\} \leq \left(\frac{1}{1-\tau}\right) a\, c\, \frac{b^H}{1-b}\, m_\Omega \sum_{j=1}^{p} E(\phi_{j1}),$$

as stated by the proposition. □

Note that generally, in infinite factor models, the parameters σ_j^2 (for $j = 1 \ldots, p$) are identically distributed according to an inverse gamma distribution. This fact guarantees $E(\sigma_j^{-2})$ constant and finite for every $j = 1, \ldots, p$. In particular, all the conditions of Proposition 2.1 hold for the multiplicative gamma process under the condition of Theorem 1 of Bhattacharya & Dunson (2011) and for the cumulative shrinkage process of Legramanti et al. (2020).

To apply rule (2.3) through an adaptive Gibbs sampler, in theory, we should have a realization of the matrix Ω, which is in fact not observed. Hence, in practice, rule (2.3) translates to the following operative procedure for the multiplicative gamma process factor model, where $\Psi = I_k$. Consider the empirical estimator of $\text{tr}(\Omega)$ given by $\widehat{\text{tr}(\Omega)} = \sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}^2/n$, assuming that the data have been transformed to have mean zero. Let $\sigma_j^{2(t)}$ the draw of the parameter σ_j^2 at the t-th iteration of the Markov chain. The full conditional posterior expected value of $\sum_{j=1}^{p}\sigma_j^{(t)2}$ under the multiplicative gamma process specification is a linear transformation of $\sum_{i=1}^{n}(y_{ij} - \lambda_{j\cdot}^{(t)\top}\eta_{i\cdot}^{(t)})^2$, which is also the sample deviance estimate of the residuals $e_{ij}^{(t)} = y_{ij} - \lambda_{j\cdot}^{(t)\top}\eta_{i\cdot}^{(t)}$ at the t-th iteration. Thus, we can write

$$n\,\widehat{\text{tr}(\Omega)} = \sum_{j=1}^{p}\sum_{i=1}^{n} y_{ij}^2 = \sum_{j=1}^{p}\sum_{i=1}^{n} e_{ij}^{(t)2} + S^{(t)},$$

where we interpret $S^{(t)} = \sum_{j=1}^{p}\sum_{i=1}^{n}(y_{ij}^2 - e_{ij}^{(t)2})$ as an estimate of the variability due to the

factors at iteration t. In particular, $S^{(t)}$ can be decomposed as the sum of

$$s_h^{(t)} = \sum_{j=1}^{p} \sum_{i=1}^{n} \lambda_{jh}^{(t)} \eta_{hi}^{(t)} (y_{ij} + e_{ij}^{(t)}),$$

for $h = 1, \dots, k^{(t)}$ and where $k^{(t)}$ is the number of factors at iteration t. Considering each $s_h^{(t)}$ as the contribution of the h-th column to $S^{(t)}$, we discard the columns of Λ with lowest $s_h^{(t)}$ while

$$\frac{n^{-1} \left(\sum_{j=1}^{p} \sum_{i=1}^{n} e_{ij}^{(t)2} + S^{(t)} \right)}{\widehat{\mathrm{tr}(\Omega)}} > \tau, \tag{2.4}$$

is satisfied for a fixed $\tau \in (0, 1)$.

A remarkable property implied by this procedure, reported in the following lemma, is that it guarantees a finite and deterministic upper bound on the number of factors, for any $\tau \in (0, 1)$.

LEMMA 2.1: *Let $k^{(t+1)}$ denote the number of latent factors after the truncation at iteration t, determined as the minimum number of summands $s_h^{(t)}$ such that condition (2.4) is satisfied. Let t_κ the iteration corresponding to the κ-th adaptation and $\tilde{\kappa} = \lceil k^{(0)} - 1 - (1 - \tau)^{-1} \rceil$, where $k^{(0)}$ is the starting number of factors. Then, for any $t \geq t_{\tilde{\kappa}}$ and for any $\tau \in (0, 1)$,*

$$k^{(t+1)} \leq \frac{1}{1 - \tau} + 1.$$

Proof of Lemma 2.1. Consider the decomposition

$$n\widehat{\mathrm{tr}(\Omega)} = \sum_{j=1}^{p} \sum_{i=1}^{n} e_{ij}^{(t)2} + \sum_{h=1}^{k^{(t)}} s_h^{(t)}.$$

Let $s_m^{(t)}$ denote the minimum $\min_{1 \leq h \leq k^{(t)}} \{ s_h^{(t)} \}$ between the summands. Then,

$$s_m^{(t)} k^{(t)} \leq n\widehat{\mathrm{tr}(\Omega)} - \sum_{j=1}^{p} \sum_{i=1}^{n} e_{ij}^{(t)2} \leq n\widehat{\mathrm{tr}(\Omega)},$$

from which $s_m^{(t)}/n\widehat{\text{tr}(\Omega)} \leq 1/k^{(t)}$. If $k^{(t)} > 1/(1-\tau)$, then $s_m^{(t)}/(n\widehat{\text{tr}(\Omega)}) < 1-\tau$ and

$$\tau < 1 - \frac{s_m^{(t)}}{n\widehat{\text{tr}(\Omega)}} = \frac{n^{-1}(\sum_{j=1}^{p}\sum_{i=1}^{n} e_{ij}^{(t)2} + S^{(t)} - s_m^{(t)})}{\widehat{\text{tr}(\Omega)}},$$

which implies

$$\tau < \frac{n^{-1}(\sum_{j=1}^{p}\sum_{i=1}^{n} e_{ij}^{(t)2} + S^{(t)})}{\widehat{\text{tr}(\Omega)}},$$

which satisfies condition (2.4). Therefore, $k^{(t+1)} \leq k^{(t)} - 1$. By algorithm construction $k^{(t)} \leq k^{(t-1)} + 1$ and, given $t \geq t_{\tilde{\kappa}}$, with $\tilde{\kappa} = \lceil k^{(0)} - 1 - (1-\tau)^{-1} \rceil$, we can observe

$$k^{(t+1)} \leq \frac{1}{1-\tau} + 1.$$

□

This result, which at a glance seems trivial, is in fact very important as, for a given ζ, using in the adaptation step the criterion of Bhattacharya & Dunson (2011), k^* does not have a deterministic upper bound. See the next section for details.

11.11 *Relations with alternative criteria*

Consider the issue of choosing ζ in the algorithm proposed in Bhattacharya & Dunson (2011) and recalled in Section 1. Let $\Delta^{k^*} = (\Lambda - \Lambda_{k^*})(\Lambda - \Lambda_{k^*})^T$ such that we can decompose the matrix Ω as $\Omega = \Delta^{k^*} + \Lambda_{k^*}\Lambda_{k^*}^{\top} + \Sigma$, since Ψ is the identity matrix in the multiplicative gamma process. Then,

$$\max_{1\leq j\leq p}\{\Omega_{jj}\} \geq \max_{1\leq j\leq p}\{\Delta_{jj}^{k^*}\} = \max_{1\leq j\leq p}\left\{\sum_{h=k^*+1}^{+\infty}\lambda_{jh}^2\right\} \geq \max_{1\leq j\leq p,\, h>k^*}\{\lambda_{jh}^2\} = \max_{h>k^*} m_h^2.$$

Therefore, if ζ is big enough so that the posterior probability that $\sqrt{\max_{1\leq j\leq p}\{\Omega_{jj}\}}$ is smaller than ζ is close to one, then a fortiori also the posterior probability that $m_h < \zeta$ is close to one

for all $h > k^*$ and for any possible truncation level k^*. This will lead to the highly deprecable consequence of discarding almost all the columns of Λ. In particular, the relation presented in the next theorem shows that an unwisely large ζ could determine poor results in terms of explained variance.

THEOREM 2.1: *Consider the set of values* $m_h = \max_{1 \leq j \leq p}\{|\lambda_{jh}|\}$ *(*$h = 1, \ldots, +\infty$*), and the subset*

$$\mathcal{M}_\tau = \left\{ m_h \text{ for } h = 1, \ldots, +\infty : m_h > \sqrt{(1-\tau)tr(\Omega)} \right\}.$$

Let $\tau \in (0, 1)$ *such that the posterior probability* $pr(\mathcal{M}_\tau = \emptyset \mid y)$ *is close to zero. Let* Λ_ζ *denote the matrix defined discarding every column* h *of the random* Λ *such that* $m_h < \zeta$*. If* ζ *is big enough so that the posterior probability that* $\min_{m_h \in \{\mathcal{M}_\tau\}} m_h < \zeta$ *is close to one when* $\mathcal{M}_\tau$ *is not empty, then the posterior probability*

$$pr\left\{ \frac{tr(\Lambda_\zeta \Lambda_\zeta^\top) + tr(\Sigma)}{tr(\Omega)} < \tau \mid y \right\}$$

is close to one.

Proof of Theorem 2.1. Consider the case of $\mathcal{M}_\tau$ not empty, denoting $p_\emptyset = \text{pr}(\mathcal{M}_\tau = \emptyset \mid y)$ the posterior probability of the complementary event. Let h^* denote the index of the minimum element belonging to $\mathcal{M}_T$, and $p_{m_{h^*}}$ denote the posterior probability $p_{m_{h^*}} = \text{pr}(m_{h^*} < \zeta \mid y, \mathcal{M}_\tau \neq \emptyset)$, conditionally on $\mathcal{M}_\tau$ not empty. Consider the event $m_{h^*} < \zeta$, so that the index h^* belongs to the set of column indices $\mathcal{H}_\zeta$ that defines the column of the random matrix $\Lambda - \Lambda_\zeta$. Then,

$$\text{tr}\left\{ (\Lambda - \Lambda_\zeta)(\Lambda - \Lambda_\zeta)^\top \right\} = \sum_{j=1}^{p} \sum_{h_\zeta \in \mathcal{H}_\zeta} \lambda_{jh_\zeta}^2 > m_{h^*}^2 \geq (1-\tau)\text{tr}(\Omega),$$

from which

$$\frac{\text{tr}\left\{ (\Lambda - \Lambda_\zeta)(\Lambda - \Lambda_\zeta)^\top \right\}}{\text{tr}(\Omega)} > 1 - \tau$$
$$\frac{\text{tr}(\Lambda_\zeta \Lambda_\zeta^\top) + \text{tr}(\Sigma)}{\text{tr}(\Omega)} < \tau.$$

In other terms, $\mathcal{M}_\tau \neq \emptyset$ and $m_{h^*} < \zeta$ are sufficient conditions to guarantee $\{\text{tr}(\Lambda_\zeta \Lambda_\zeta^\top) + \text{tr}(\Sigma)\}/\text{tr}(\Omega) < \tau$.

Therefore,

$$\text{pr}\left\{\frac{\text{tr}(\Lambda_\zeta \Lambda_\zeta^\top) + \text{tr}(\Sigma)}{\text{tr}(\Omega)} < \tau \mid y\right\} > \text{pr}\left(m_{h^*}^2 \geq (1-\tau)\text{tr}(\Omega) \mid y\right) = p_{m_{h^*}}(1 - p_\emptyset).$$

When τ and ζ are chosen large enough that $p_\emptyset$ is close to 0 and $p_{m_{h^*}}$ is close to 1, we obtain

$$\text{pr}\left\{\frac{\text{tr}(\Lambda_\zeta \Lambda_\zeta^\top) + \text{tr}(\Sigma)}{\text{tr}(\Omega)} < \tau \mid y\right\}$$

close to 1. □

Theorem 2.1 further sheds light on the importance of carefully considering the scale of the data, represented here by the trace of Ω, to explain the variance of the data through the factor model $\Omega_\zeta = \Lambda_\zeta \Lambda_\zeta^\top + \Sigma$.

On the other side, ζ may be too small, leading to an unnecessarily large number of irrelevant factors. Specially, for large p, the probability of having all values of $|\lambda_{jh}|$ smaller than ζ goes to zero exponentially. The consequence is that an unnecessarily large set of columns is kept. In general, for a given ζ, the number of factors determined according to the procedure in Bhattacharya & Dunson (2011) is random and unbounded.

While our method is defined for the multiplicative gamma process model, its guidelines provide useful practical suggestions for the application of the cumulative shrinkage process of Legramanti et al. (2020). According to the aforementioned notation, the columns to discard are naturally identified as those modelled by the spike θ_∞, while, in the columns retained, λ_{jh} is modelled by the slab, and specifically, it is marginally distributed as a Student's t-distribution with $2a_\theta$ degrees of freedom. Consider the ratio r between the two posterior full conditional probabilities that a column is modelled by the spike and the slab, respectively. Then, r is affected by the trace of Ω. In particular, in those cases where the slab marginal distribution can be approximated by a normal distribution, i.e., when a_θ is large, if Ω is scaled by a factor c, the new ratio between probabilities r_c is given by $r_c = r^c$. For this reason, despite a standardization procedure not being included in the original algorithm (Legramanti et al., 2020), it appears crucial in applications.

III Simulation experiments

To illustrate the performance of the different criteria, a simulation study has been conducted. This also enables a comparison of the behaviour of the different methods when the scale of the problem changes, which has been ignored in previous studies.

Specifically and consistently with Legramanti et al. (2020), we simulate 20 independent data sets with $n = 100$ observations from the Gaussian linear factor model $y_i \sim N_p(0, \Lambda_0\Lambda_0^\top + I_p)$ by sampling the loadings λ_{jh0} independently from a Gaussian distribution with mean zero and variance ς^2 with $h \leq k$, k finite, and $\varsigma \in \{1, 50\}$. We consider three different scenarios based on different dimensions of Λ, and specifically we let $(p, k) \in \{(20, 5), (50, 10), (100, 15)\}$. We compare the performance of different models and criteria, namely the multiplicative gamma process with adaptation algorithm and $\zeta = 10^{-4}$ as in Bhattacharya & Dunson (2011), the multiplicative gamma process with adaptation based on the criterion described in Section II.i with $\tau = 0.999$, the cumulative shrinkage process as in Legramanti et al. (2020), and the cumulative shrinkage process fitted on the standardized data assuming $\theta_\infty = 0.01$ and $a_\theta = 15$. The hyperparameters of the four algorithms, unless otherwise specified, are assumed to be equal to those specified in Section 4 of Legramanti et al. (2020). Table 2.1 reports the median of the posterior mean of a Monte Carlo estimate of the number of factors $E(k^*|y)$ and its interquartile range. To limit the computational time, we fix, in the algorithms, the maximum value of k^* equal to p. We notice that the standard multiplicative gamma process is severely biased. This bias is due to the difficulty

Table 2.1: Median and interquartile range of the estimated $E(k^*|y)$ for the different models and truncation criteria.

ς	(p, k)	MGP_ζ		MGP_τ		CUSP_0		CUSP_{std}	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
1	(20, 5)	20.00	0.00	9.72	1.21	5.00	0.00	5.00	0.00
	(50, 10)	49.99	0.29	14.02	0.95	10.00	0.00	10.00	0.00
	(100, 15)	93.16	2.12	17.80	0.30	15.00	0.00	14.00	1.00
50	(20, 5)	20.00	0.00	7.01	0.81	11.15	2.23	5.14	0.41
	(50, 10)	50.00	0.12	10.77	0.37	21.35	4.45	10.00	0.00
	(100, 15)	94.82	2.31	15.97	0.27	15.00	0.07	14.50	1.00

CUSP, cumulative shrinkage process; MGP_ζ, multiplicative gamma process with truncation criterion as in Bhattacharya & Dunson (2011); MGP_τ, multiplicative gamma process with truncation criterion proposed in Section II.i; $Q_{0.5}$, median; IQR, interquartile range.

of defining a value for ζ without taking into consideration the scale or the dimension of the data. Our proposal based on the interpretable quantity τ instead consistently estimates the number of factors and is robust to the scale of the data. Moreover, the last two columns of Table 2.1 show that the standardization in the cumulative shrinkage process is fundamental for performing adequate inference on the parameters in different contexts.

In addition to the posterior mean of k^*, we also compute the computational time of each procedure, a Monte Carlo estimate of the mean squared error $\sum_{j=1}^{p}\sum_{l=j}^{p} E\{(\omega_{jl}-\omega_{jl0})^2 \mid y\}/\{p(p+1)/2\}$, where ω_{jl} and ω_{jl0} are the elements jl of Ω and $\Omega_0 = \Lambda_0\Lambda_0^\top + I_p$, respectively. Furthermore, we also compute the logarithm of the pseudo-marginal likelihood, a convenient index derived from predictive considerations (Gelfand & Dey, 1994). The median and the interquartile range of these quantities are reported in Tables 2.2--2.4.

TABLE 2.2: Median and interquartile range of the estimated mean squared error for the different models and truncation criteria.

ς	(p,k)	MGP_ζ		MGP_τ		CUSP_0		CUSP_{std}	
		$\text{Q}_{0.5}$	IQR	$\text{Q}_{0.5}$	IQR	$\text{Q}_{0.5}$	IQR	$\text{Q}_{0.5}$	IQR
1	$(20,5)$	0.76	0.21	0.77	0.20	0.73	0.24	0.70	0.22
	$(50,10)$	2.08	0.38	2.09	0.30	2.13	0.32	2.34	0.25
	$(100,15)$	3.69	0.37	3.65	0.37	3.70	0.44	5.64	0.62
50	$(20,5)$	13.38	4.00	12.57	2.74	11.98	3.78	3.19	1.71
	$(50,10)$	26.27	2.85	26.52	5.00	20.72	2.57	11.12	3.49
	$(100,15)$	99.64	53.56	72.19	47.08	287.38	140.13	28.29	3.11

CUSP, cumulative shrinkage process; MGP_ζ, multiplicative gamma process with truncation criterion as in Bhattacharya & Dunson (2011); MGP_τ, multiplicative gamma process with truncation criterion proposed in Section 11.1; $\text{Q}_{0.5}$, median; IQR, interquartile range.
To facilitate reading the results have been scaled by a factor of 10^{-6} when $\varsigma = 50$.

It is worth reporting that the performances of MGP_τ in terms of runtime, mean squared error, and logarithm of the pseudo-marginal likelihood are comparable to or even better than those of the original cumulative shrinkage process implementation. This fact suggests that the poor performance of the multiplicative gamma process discussed in the comparison in Legramanti et al. (2020) could be mainly related to a misleading method to select the relevant latent factors. This stresses the dramatic importance of an interpretable and general truncation criterion in infinite factor models.

TABLE 2.3: Median and interquartile range of the runtime in seconds for the different models and truncation criteria.

ς	(p, k)	MGP_ζ		MGP_τ		CUSP_0		CUSP_{std}	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
1	$(20, 5)$	88.44	15.93	67.08	8.49	69.56	5.74	69.41	6.26
	$(50, 10)$	308.4	5.97	156.04	7.98	160.77	5.09	165.23	3.93
	$(100, 15)$	1425.62	39.77	338.41	9.35	369.01	7.05	369.28	6.40
50	$(20, 5)$	80.66	4.79	65.28	4.98	80.13	4.10	73.60	5.23
	$(50, 10)$	304.93	10.62	151.00	10.59	201.37	11.18	169.61	2.87
	$(100, 15)$	1425.99	39.23	333.03	9.23	380.14	8.36	363.35	12.20

CUSP, cumulative shrinkage process; MGP_ζ, multiplicative gamma process with truncation criterion as in Bhattacharya & Dunson (2011); MGP_τ, multiplicative gamma process with truncation criterion proposed in Section II.i; $Q_{0.5}$, median; IQR, interquartile range.

IV Football player tracking data application

IV.1 *Dimensionality reduction of large indicators dataset*

Player tracking data is emerging as a good testing ground for methods that aim to reduce the dimensionality of large datasets to allow complex phenomena to be represented and visualized. For instance, each player can be described by a large number of key performance indicators, although it makes challenging to find an interpretable representation of the players and identify common traits and similarities. Factor models seem particularly suitable for addressing this issue. The large amount of data contained in the key performance indicators can be represented through a lower dimensional set of latent factors, reducing dimensionality to a few dimensions that can be visualized and interpreted through simple charts. Since there is no clue as to how many factors should be considered to isolate and sufficiently represent the underlying covariance structure among the indicators, we rely on the Bayesian infinite factor models framework, which allows a flexible specification without imposing strict constraints.

We consider a dataset of $n = 178$ players described by $p = 13$ key performance indicators, such as number of sprints, passes and conduction choices, pressure applied and received, and other physical and technical metrics measured in a professional European league match and scaled per 90 minutes. We exclude from the dataset goalkeepers and players who played less than 60 minutes. Due to data confidentiality agreements, both players and teams have been anonymized.

TABLE 2.4: Median and interquartile range of the log-pseudo-marginal likelihood for the different models and truncation criteria.

ς	(p, k)	MGP_ζ		MGP_τ		CUSP_0		CUSP_{std}	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
1	$(20, 5)$	-36.25	0.79	-36.24	0.78	-36.22	0.78	-36.22	0.78
	$(50, 10)$	-92.46	0.71	-92.43	0.74	-92.47	0.63	-92.55	0.82
	$(100, 15)$	-182.25	0.80	-182.53	0.85	-182.67	0.79	-188.23	7.77
50	$(20, 5)$	-56.66	1.04	-56.39	0.89	-56.54	1.03	-80.40	3.27
	$(50, 10)$	-133.53	1.09	-133.15	0.70	-135.39	1.21	-216.16	2.06
	$(100, 15)$	-243.80	1.55	-242.52	0.95	-243.25	1.43	-445.23	60.46

CUSP, cumulative shrinkage process; MGP_ζ, multiplicative gamma process with truncation criterion as in Bhattacharya & Dunson (2011); MGP_τ, multiplicative gamma process with truncation criterion proposed in Section 11.i; $Q_{0.5}$, median; IQR, interquartile range.

To represent the players in a low-dimensional space, we define a multiplicative gamma process factor model on the data. After standardizing the data, we set $\nu = 3$, $a_1 = 1$, $a_2 = 2$, and $b_1 = b_2 = 1$, consistent with the simulation experiments. Then, we run the Gibbs sampler assuming the current standard practice of adapting the number of factors as reported in Bhattacharya & Dunson (2011) for different values of $\zeta \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. We compare the dimensionality reduction capacity of this approach with the results obtained by running the algorithm with adaptation based on the criterion described in Section 11.i for different values of $\tau \in \{0.75, 0.9, 0.95\}$. We run the algorithms for $25,000$ iterations discarding the first $10,000$ iterations. Then, we thin the Markov chain, saving every 5-th sample, and adapt the number of active factors at iteration t with probability $p(t) = \exp(-1 - 5\,10^{-4}t)$. We set the maximum number of factors equal to $2p$.

The posterior mean of the number of factors for the two citeria is reported in Table 2.5. The

TABLE 2.5: Posterior mean of k^* for the different truncation criteria and thresholds.

	MGP_ζ			MGP_τ		
	$\zeta = 10^{-4}$	$\zeta = 10^{-3}$	$\zeta = 10^{-2}$	$\tau = 0.75$	$\tau = 0.9$	$\tau = 0.95$
$E(k^* \mid y)$	25.96	21.65	15.39	2.56	5.27	5.92

data dimension is not effectively reduced by applying the standard truncation criterion for any of the three values of ζ. Instead, the last column of the table shows that, by applying our proposed truncation criterion setting $\tau = 0.95$, fewer than six factors are sufficient to explain around the

95% of the model variance.

IV.II *Identifiability and two dimensional representation*

We are interested in visualizing the set of players in a two-dimensional space. Therefore we consider the space defined by the two first latent components under the multiplicative gamma process with $\tau = 0.75$, where two factors are able to represent a large part of the total variability of the key performance indicators. The player i can be represented by an estimated summary of the couple (η_{i1}, η_{i2}), as the posterior mean $\{E(\eta_{i1} \mid y), E(\eta_{i2} \mid y)\}$. Nevertheless, non-identifiability of the latent structure creates problems in representing the posterior mean of the Markov chain Monte Carlo samples. Indeed, both Λ and H are only identifiable up to an arbitrary rotation P with $PP^\top = I_k$, then each sample could be drawn from a different rotation, making difficult to summarize the elements of Λ and H through their posterior means. This is a well known problem in Bayesian factor models, and there is a rich literature proposing post-processing algorithms that align posterior samples of Λ and H so that one can then obtain interpretable posterior summaries. In particular, we follow the algorithm proposed and implemented by Poworoznek et al. (2021) in the R package `infinitefactor`. The algorithm firstly applies the varimax orthogonal rotation (Kaiser, 1958) to the samples to maximize the sum of the variances, then it switches the sign and the order of columns to align all the samples to a pivotal matrix. Refer to McParland et al. (2014), Aßmann et al. (2016), and Roy et al. (2019) for alternative post-processing algorithms in related contexts.

Looking at the posterior mean of the aligned matrix of the Λ samples, reported in Table 2.6, we can roughly attribute meanings to the factors according to the most loaded performance indicators in each column of the table. The first factor mainly concerns the space available for each player, whereby high values of $\eta_{\cdot 1}$ indicate players who are not pressed and have a high chance of receiving or conducting the ball. Consistently, the first factor also explains the players in terms of gain obtained through passes and increase in the dangerousness of the actions. The second factor, instead, explains the run performances of the players represented by applied pressure, the number of sprints, and sprint distance. Figure 2.1 represents the players by plotting the posterior mean of $\eta_{\cdot 1}$ and $\eta_{\cdot 2}$ estimated on the aligned Markov chain Monte Carlo samples. Colours highlight the association between the first factor and the role of the player. In fact, as we might expect, strikers are generally constrained to play in tight spaces and under considerable pressure. The second factor appears independent from the first, showing heterogeneity in physical performance over

TABLE 2.6: Posterior mean of the aligned samples of Λ truncated at $k^* = 2$.

	First factor	Second factor
Pass availability	0.90	−0.10
Pass risk	0.13	0.10
Pass gain	0.22	0.01
Dangerousness increase	0.64	0.01
Conduction time	0.42	0.10
Transfer time	0.43	0.18
Covered area	0.84	0.10
Received pressure	−0.95	−0.04
Applied pressure	−0.23	−0.25
Walk distance	0.22	0.38
Sprint distance	−0.03	0.87
Run distance	0.00	0.46
Number of sprints	−0.12	0.93

the entire distribution of the first factor with very low sample correlation $\text{cor}\{E(\eta_{\cdot 1} \mid y), E(\eta_{\cdot 2} \mid y)\} = 0.01$. However, looking at behaviour within the single groups defined by the role, we observe a slightly positive association between physical performance and space available, with sample correlation between factors equal to 0.19 and 0.38 when we consider only the midfielder or the striker group, respectively. This fact suggests an interesting perspective to underline the importance of physical performance during a match and the advantages that can be obtained in terms of dangerousness increase. More generally, Fig. 2.1 provides a useful tool for an immediate assessment of overall player performance during a match.

FIGURE 2.1: Posterior means of the aligned samples of the first two latent factors estimated by the multiplicative gamma process with $\tau = 0.75$. To facilitate the representation, axis measures are set such that 0 and 100 represent minimum and maximum observed values of the factors.

3 Generalized infinite factorization models

1 Structured shrinkage

Although overfitted factorizations and, specifically, infinite factor models are widely used in different contexts, as in the recent low-rank regression model of Chakraborty et al. (2020), there are two key gaps in the literature. The first one is a careful development of the shrinkage properties of increasing shrinkage priors (Durante, 2017). Outside the factorization context and mostly motivated by high-dimensional regression, there is a rich literature recommending specific desirable properties for shrinkage priors. These include high concentration at zero to favour shrinkage of small coefficients and heavy tails to avoid over shrinking large coefficients. Motivated by this principles, popular shrinkage priors have been developed including the Dirichlet--Laplace (Bhattacharya et al., 2015) and horseshoe (Carvalho et al., 2010). Current increasing shrinkage priors, such as those of Bhattacharya & Dunson (2011), were not designed to have the desirable shrinkage properties of these priors. For this reason, ad hoc truncation and use of the horseshoe or Dirichlet--Laplace can outperform increasing shrinkage priors in some contexts; for example, this was the case in Ferrari & Dunson (2021).

A second gap in the literature on overfitted factorization priors is the lack of structured shrinkage. The focus has been on priors for Λ that are exchangeable within columns, with the level of shrinkage increasing with the column index. However, it is common in practice to have meta-covariates encoding features of the rows of Λ. For example, the rows may correspond to different genes in genomics or species in ecology. There is a rich literature on incorporating gene ontology in statistical analyses of genomic data (see, for example, Thomas et al., 2009), while in ecology it is

common to include species traits in species distribution models (Ovaskainen & Abrego, 2020). Considering the football application discussed in Section 2.IV, we could be interested in including additional information about how the key performance indicators are measured to help with the identification of sparsity patterns on Λ according to our prior knowledge on similarities between indicators. Beyond the Bayesian literature, it is common to include structured penalties, with the group lasso (Yuan & Lin, 2006) being a notable example. The widespread use of the group lasso and the overlap group lasso (Jacob et al., 2009) for variable selection inspired the football application we present in this chapter. Indeed, factor models can be particularly suitable when we want to predict a variable of interested through the a large set of covariates, by performing a variable selection approach replacing the original very many predictors with the low-dimensional latent factors. In the football context we may be interested in modelling the dangerousness y_i^R of an action i through a regression on a set of p key performance indicators y_i^C describing the action. Then, we reduce the covariate dimensionality by considering a linear factor model for the $n \times p + 1$ matrix $y = c(y^R, y^C)$, with sparse pattern on Λ inducing structured penalty on regression coefficients.

Motivated by the aforementioned deficiencies of current factorizations priors, we propose in this chapter a broad class of generalized infinite factorization priors, along with corresponding theory and algorithms for routine Bayesian implementation. We will also present two applications of such models in two very different contexts, i.e., regularized regression in football and covariance modelling in ecology, showing the benefits provided by the structured shrinkage, regardless of the application field.

II Generalized infinite factor models

II.1 *Model specification*

Recalling the factor model notation previously introduced, we consider the following general class of factor models,

$$y_{ij} = f_j(z_{ij}), \quad z_i = \Lambda\eta_i + \epsilon_i, \quad \epsilon_i \sim f_\epsilon, \tag{3.1}$$

with the function $f_j : \Re \to \Re$, for $j = 1 \ldots, p$. In our motivating application, y_{ij} is the measure of the j-th ($j = 1, \ldots, p$) indicator of action i ($i = 1, \ldots, n$). We refer to (3.1) as the class of generalized factorization models.

Following common practice in infinite factor models (see Section 1 of Chapter 2), we avoid imposing identifiability constraints on Λ and assume that Ψ is prespecified. Our focus is on a new class of generalized infinite factor models where (2.1) holds and the novel class of priors of Λ allows infinitely many factors, $k = \infty$. In particular, the local ϕ_{jh} and the factor-specific θ_h scales are independent a priori and supported on $[0, \infty)$ with positive probability mass on $(0, \infty)$. We let $N(0, 0)$ denote a degenerate distribution with all its mass at zero. Although we allow infinitely many columns in Λ, (2.1) induces a prior for Ω supported on the set of $p \times p$ positive semi-definite matrices under mild conditions reported in the following proposition.

PROPOSITION 3.1: *Let $\Pi_\Lambda \otimes \Pi_\Sigma$ denote the prior on (Λ, Σ). Let Θ_Λ and Θ_Σ denote the sample spaces of the matrices Λ and Σ, respectively. If $E(\phi_{jh}) = E(\phi_{lh})$ for every $h, l \in \{1, \ldots, \infty\}$ and $\sum_{h=1}^{\infty} E(\theta_h) < \infty$, then, $\Pi_\Lambda \otimes \Pi_\Sigma(\Theta_\Lambda \times \Theta_\Sigma) = 1$.*

Proof of Proposition 3.1. Assume $\Sigma \in \Theta_\Sigma$ and $(\Psi, \Lambda) \in \Theta_\Psi \times \Theta_\Lambda$, with Θ_Σ the set of $p \times p$ positive semi-definite matrices with finite elements, and

$$\Theta_\Psi \times \Theta_\Lambda = \left\{ \Lambda = (\lambda_{jh}),\ \Psi = (\psi_{hh}) : \sum_{h=1}^{\infty} \lambda_{jh}\psi_{hh}\lambda_{sh} < \infty \ \forall\, j, s \in (1, \ldots, p) \right\}.$$

Due to independence, we can study the prior on Σ and Λ separately. The prior on Σ is defined on the set of positive semi-definite matrices. Therefore, it is sufficient to prove that the elements of $\Lambda\Psi\Lambda^{\mathrm{T}}$ are finite almost surely. Using Cauchy-Schwartz, it is straightforward to show that all the entries of $\Lambda\Psi\Lambda^{\mathrm{T}}$ are finite if and only if $\sum_{h=1}^{\infty} \psi_{hh}\lambda_{jh}^2 < \infty$ ($j = 1, \ldots, p$). Let c satisfy $c > \max_{h=1,\ldots,\infty} \psi_{hh}$. Since

$$E(\lambda_{jh}^2) = E\{E(\lambda_{jh}^2 \mid \phi_{jh}, \theta_h)\} = E(\phi_{jh})E(\theta_h),$$

and $E(\phi_{jh}) = E(\phi_{j1})$ ($j = 1, \ldots, p$; $h = 1, \ldots, \infty$), it is sufficient that $\sum_{h=1}^{\infty} E(\theta_h) < \infty$ to prove that $\sum_{h=1}^{\infty} E(\lambda_{jh}^2) = E(\phi_{j1}) \sum_{h=1}^{\infty} E(\theta_h) < \infty$ and then $\sum_{h=1}^{\infty} \psi_{hh}\lambda_{jh}^2 < c \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty$. □

In contrast to most of the existing literature on shrinkage priors, we want to define a nonexchangeable structure that includes meta-covariates w informing the sparsity structure of Λ. In our

context, meta-covariates provide information to distinguish the p different variables as opposed to traditional covariates that serve to distinguish the n subjects. Letting w denote a $p \times m$ matrix of such meta-covariates, we choose the density function of the local scale elements pr_{ϕ_j} not depending on the index h and such that

$$E(\phi_{jh} \mid \gamma_h) = g(w_j^\top \gamma_h), \quad \gamma_h = (\gamma_{1h}, \ldots, \gamma_{mh})^\top, \quad \gamma_{lh} \sim \text{pr}_\gamma \quad (l = 1, \ldots, m) \tag{3.2}$$

where $g : \Re \to \mathcal{A} \subset \Re_+$ is a known smooth one-to-one differentiable link function, $w_{.j} = (w_{j1}, \ldots, w_{jm})^T$ denotes the jth row vector of w, and γ_h is the hth column vector of the coefficient matrix Γ controlling the impact of the meta-covariates on shrinkage of the factor loadings in the hth column of Λ.

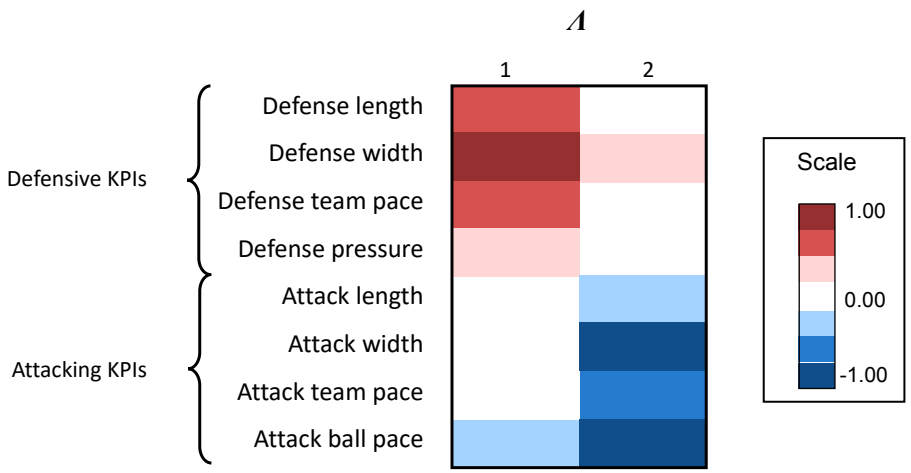

FIGURE 3.1: Illustrative loadings matrix of a football application, where the rows refer to eight action key performance indicators referring to the attacking or defending team. White cells represent the elements of Λ equal to zero, while blue and red cells represent negative and positive values, respectively.

To illustrate the usefulness of (3.2), consider the previously introduced study on action danger-

ousness and suppose $w_j = \{\mathbb{1}(\kappa_j = "a"), \mathbb{1}(\kappa_j = "d")\}^T$, where $\kappa_j \in \{"a", "d"\}$ denotes if the indicator j is referring to the attacking ("a") or defending ("d") team. Performance indices referring to the attacking team (or defending team) may tend to have similarities that can be expressed in terms of a shared pattern of high or low loadings on the same latent factors. To illustrate this situation, we simulate a loadings matrix, displayed in Fig. 3.1, sampling from the prior introduced in Section III.i where $\text{pr}(\lambda_{jh} = 0) > \text{pr}(\phi_{jh} = 0) > 0$. The loadings within each column are penalized based on the group structure identified by the $m = 2$ team roles, namely attacking and defending, of the $p = 10$ key performance indicators considered. Our proposed prior allows for the possibility of such structure while not imposing it. In the football application on tracking data indicators, w can be defined to include also other features of the indices, such as the type of measure the index considers: spatial measurements, physical performances, or possession choices. Related meta-covariates are widely available in several application fields as genomics (Thomas et al., 2009) or ecology. For instance, in species distribution modelling, it is common to consider in the model species traits as phylogenetic placement, size or diet (Miller et al., 2019; Tikhonov et al., 2020).

II.II *Properties*

In this section we present some properties motivating the shrinkage process defined in (2.1) and (3.2) and provide some insights into prior elicitation. Below we study key properties of our prior, including an increasing shrinkage property, the ability of the induced marginal prior to accommodate both sparse and large signals, and control of the multiplicity problem in sparse settings. This theory illuminates the role of hyperparameters; specific recommendations of hyperparameter choice in practice are illustrated under the model settings of Section III.i.

To formalize the increasing shrinkage property, we introduce the following definition.

DEFINITION 3.1: *Letting Π_Λ denote a shrinkage prior on Λ, Π_Λ is a weakly increasing shrinkage prior if $\text{var}(\lambda_{j(h-1)}) > \text{var}(\lambda_{jh})$ for j in $1, \ldots, p$ and $h = 2, \ldots, \infty$. Π_Λ is a strongly increasing shrinkage prior if $\text{var}(\lambda_{s(h-1)}) > \text{var}(\lambda_{jh})$, for j, s in $\{1, \ldots, p\}$ and $h = 2, \ldots, \infty$.*

Weakly increasing shrinkage corresponds to the prior variance increasing across columns within each row of Λ, while strongly increasing shrinkage implies that the prior variance of any loading element is larger than all elements with a higher column index. In the following theorem, we show that the process in (2.1) and (3.2) induces weakly increasing shrinkage under a simple sufficient

condition.

THEOREM 3.1: *Expression* (2.1) *is a weakly increasing shrinkage prior according to Definition 3.1 if* $E(\theta_h) > E(\theta_{h+1})$ *for any* h.

Proof of Theorem 3.1. The variance of λ_{jh} is

$$\text{var}(\lambda_{jh}) = E\{E(\lambda_{jh}^2 \mid \phi_{jh}, \theta_h)\} = E\{E(\theta_h \phi_{jh} \mid \phi_{jh}, \theta_h)\} = E(\theta_h \phi_{jh}).$$

Then,

$$\text{var}(\lambda_{jh}) = E(\phi_{jh}\theta_h) = E(\phi_{j1})E(\theta_h) > E(\phi_{j1})E(\theta_{h+1}) = \text{var}(\lambda_{jh+1}),$$

since the scale parameters are independent and the local scale ϕ_{jh} is equally distributed over the column index h. □

Increasing shrinkage priors favour a decreasing contribution of higher-indexed columns of Λ to the covariance Ω. In addition to inducing a flexible shrinkage structure that allows different factors to have a different sparsity structure in their loadings, this enables accurate approximation of the model by truncating the number of factors to k^*. Conditions on prior to control the induced truncation error of $\Omega_{k^*} = \Lambda_{k^*}\Psi_{k^*}\Lambda_{k^*}^\top + \Sigma$ has already been provided in Proposition 2.1 in Chapter 2. The above increasing shrinkage properties can be satisfied by naive priors that overshrink the elements of Λ. It is important to avoid such overshrinkage and allow not only many elements that are ≈ 0 but also a small proportion of large coefficients. A similar motivation applies in the literature on shrinkage priors in regression (Carvalho et al., 2010). Borrowing from that literature, the marginal prior for λ_{jh} should be concentrated at zero to reduce mean square error by shrinking small coefficients to zero, albeit with heavy tails to avoid overshrinking the signal.

To quantify the prior concentration of (2.1) in a ζ neighbourhood of zero, we can obtain

$$\text{pr}(|\lambda_{jh}| > \zeta) \leq \frac{E(\theta_h)\, E(\phi_{jh})}{\zeta^2} \tag{3.3}$$

as a consequence of Markov's inequality. It is common practice in working with local-global shrinkage priors to choose local or column scale small while assigning a heavy-tailed density to the other scale. In our case, (3.2) allows the bound in (3.3) to be regulated by meta-covariates w, while, under the condition in Theorem 3.1, decreasing $E(\theta_h)$ with column index causes an increasing concentration near zero, since $E(\phi_{jh}) = E(\phi_{jl})$ for every $h, l \in \{1, \ldots, \infty\}$. The means of

the factor and the local scales control prior concentration near zero, while overshrinkage can be ameliorated by choosing prior pr_{ϕ_j} or pr_{θ_h} $(h = 1, \ldots, \infty)$ heavy-tailed.

To show sufficient conditions to guarantee a heavy-tailed marginal distribution for λ_{jh}, we need to introduce the following Lemma. A random variable has power law tails if its cumulative distribution function F has $1 - F(t) \geq ct^{-\alpha}$ for constants $c > 0$, $\alpha > 0$, and for any $t > L$ for L sufficiently large.

LEMMA 3.1: *Let u, v denote two real positive random variables. If at least one among $(u \mid v)$ and $(v \mid u)$ is power law tail distributed, then the product uv is power law tail distributed.*

Proof of Lemma 3.1. For a positive value x, we can write

$$\text{pr}(uv > x) = \int_0^\infty \text{pr}(u > x/v \mid v)\,\text{pr}(v)\text{d}v = E\{F^C_{u|v}(x/v)\},$$

where $F^C_{u|v}(x) = \text{pr}(u > wx \mid v)$ and $\text{pr}(v)$ is the probability density function of v. If $F^C_{u|v}(x) \geq c\,x^{-\alpha}$ with c, α positive constants and x greater than a sufficiently large number L, then

$$\text{pr}(uv > x) \geq E\{c\,(x/v)^{-\alpha}\} = c\,x^{-\alpha}E(v^\alpha) \quad x > L \gg 0.$$

If $E(v^\alpha) = \infty$, then $\text{pr}(uv > x) > c\,x^{-\alpha} = O(x^{-\alpha})$, otherwise $\text{pr}(uv > x) \geq \nu(x)$ for $x > L$, with $\nu(x)$ a function of order $O(x^{-\alpha})$ as x goes to infinity. This shows that the right tail of the distribution of the random variable uv follows a power law behaviour. □

The following Proposition provides a condition on the prior to guarantee a heavy-tailed marginal distribution for λ_{jh}.

PROPOSITION 3.2: *If at least one scale parameter among θ_h or ϕ_{jh} is characterized by a power law tail prior distribution, then the prior marginal distribution of λ_{jh} has power law tails.*

Proof of Proposition 3.2. Consider the strictly positive random variables $\theta^*_h = (\theta_h \mid \theta_h > 0)$, and $\phi^*_{jh} = (\phi_{jh} \mid \phi_{jh} > 0)$. Since the positive part of the variance of the λ_{jh} is equal to the product of independent positive random variables $\theta^*_h\phi^*_{jh}$, Lemma 3.1 ensures that if at least one of those scale parameters follows a power law tail distribution, then the product is power law tail distributed, so that $\text{pr}(\theta^*_h\phi^*_{jh} > x) \geq c\,x^{-\alpha}$ for c, α positive constants and $x > L$. Without loss

of generality, we focus on the right tail of λ_{jh}. Let

$$\begin{aligned}\text{pr}(\lambda_{jh} > \lambda) = \text{pr}(\lambda_{jh} > \lambda \mid \theta_h\phi_{jh} > 0)\,\text{pr}(\theta_h\phi_{jh} > 0) \\ + \text{pr}(\lambda_{jh} > \lambda \mid \theta_h\phi_{jh} = 0)\,\text{pr}(\theta_h\phi_{jh} = 0).\end{aligned} \tag{3.4}$$

It is straightforward to observe that λ_{jh} marginally has a power law tail if and only if $(\lambda_{jh} \mid \theta_h\phi_{jh} > 0)$ is power law tail distributed and $\text{pr}(\theta_h\phi_{jh} > 0)$ is strictly positive. Since $\text{pr}(\theta_h > 0) > 0$, and $\text{pr}(\phi_{jh} > 0) > 0$, then $\text{pr}(\theta_h\phi_{jh} > 0) > 0$, given independence between the scale parameters. Focusing on $\theta_h\phi_{jh} > 0 > 0$ in the first term of the right hand side of (3.4), we have

$$\text{pr}(\lambda_{jh} > \lambda \mid \theta^*_h\phi^*_{jh}) = 1 - F_N(\lambda\,(\theta^*_h\phi^*_{jh})^{-0.5}),$$

with $F_N(x)$ indicating the cumulative distribution function of the standard Gaussian distribution. We want to prove that the marginal $F^c_{\lambda_{jh}}(\lambda) = \text{pr}(\lambda_{jh} > \lambda)$ is sub-exponential as $\lambda \to \infty$. Using the lower bound for the right tail of the standard Gaussian of Abramowitz & Stegun (1948),

$$1 - F_N(\lambda\,(\theta^*_h\phi^*_{jh})^{-0.5}) \geq \left(\frac{2}{\pi}\right)^{0.5} \frac{(\theta^*_h\phi^*_{jh})^{-0.5}}{\lambda + (\lambda^2 + 4\theta^*_{jh})^{0.5}} e^{-\lambda^2/(2\theta^*_{jh})}.$$

Marginalizing over the product $\theta^*_h\phi^*_{jh}$, we obtain

$$\text{pr}(\lambda_{jh} > \lambda \mid \theta^*_h\phi^*_{jh}) \geq E\left\{\left(\frac{2}{\pi}\right)^{0.5} \frac{\theta^{*\,0.5}_{jh}}{\lambda + (\lambda^2 + 4\theta^*_{jh})^{0.5}} e^{-\lambda^2/(2\theta^*_{jh})}\right\} = E\left\{t_\lambda(\theta^*_h\phi^*_{jh})\right\},$$

where $t_\lambda(\theta^*_h\phi^*_{jh})$ is a monotonically increasing nonnegative function defined on the positive real line. Applying Markov's inequality, we have $E\{t_\lambda(\theta^*_h\phi^*_{jh})\} > \text{pr}(\theta^*_h\phi^*_{jh} > \zeta)t_\lambda(\zeta)$, and letting $\zeta = \lambda^2$,

$$E\left\{t_\lambda(\theta^*_h\phi^*_{jh})\right\} > \text{pr}(\theta^*_h\phi^*_{jh} > \lambda^2)\frac{e^{-0.5}}{1 + 5^{0.5}}\left(\frac{2}{\pi}\right)^{0.5}.$$

If $\text{pr}(\theta^*_h\phi^*_{jh} > \lambda) \geq c\,\lambda^{-\alpha}$ for certain α, c positive constants and λ sufficiently large, then

$$\text{pr}(\lambda_{jh} > \lambda \mid \theta^*_h\phi^*_{jh}) \geq \frac{e^{-0.5}}{1 + 5^{0.5}}\left(\frac{2}{\pi}\right)^{0.5} c\,\lambda^{-2\alpha} = \tilde{c}\lambda^{-\tilde{\alpha}},$$

where $\tilde{c} = e^{-0.5}(1+5^{0.5})^{-1}(2/\pi)^{0.5}c > 0$ and $\tilde{\alpha} = \alpha/2 > 0$. By symmetry, $\text{pr}(\lambda_{jh} < -\lambda \mid \theta_{jh} > 0) \geq \tilde{c}\lambda^{-\tilde{\alpha}}$ for $\lambda > L$ sufficiently large. It is sufficient that the marginal distribution of θ_h^* or ϕ_{jh}^* has power law right tail to guarantee that $(\lambda_{jh} \mid \theta_h\phi_{jh} > 0)$ has power law tail and then that marginally λ_{jh} has power law tail. □

An important consequence of the heavy-tailed property is avoidance of overshrinkage of large signals. This is often formalized via a tail robustness property (Carvalho et al., 2010). As an initial result, key to showing sufficient conditions for a type of local tail robustness, we provide the following lemma on the derivative of the log prior in the limit as $\lambda_{jh} \to \infty$.

LEMMA 3.2: *If at least one scale parameter among θ_h or ϕ_{jh} has a prior with power law tails for any possible prior distribution of γ_h, then for any finite truncation level k^*,*

$$\lim_{\lambda\to\infty} \frac{\partial \log\{\text{pr}_{\lambda_{jh}|\Lambda_{-jh}}(\lambda)\}}{\partial\lambda} = 0$$

where $\text{pr}_{\lambda_{jh}|\Lambda_{-jh}}(\lambda)$ is the conditional distribution of λ_{jh} given the other elements of Λ_{k^*}.

Proof of Lemma 3.2. Consistent with Proposition 3.2, $(\lambda_{jh} \mid \Lambda_{-jh})$ has power law tail if $(\theta_h\phi_{jh} \mid \Lambda_{-jh})$ has power law tail. Furthermore, $\text{pr}(|\lambda_{jh}| > \lambda \mid \Lambda_{-jh})$ has power law tail for large λ if and only if $\text{pr}(|\lambda_{jh}| > \lambda \mid \Lambda_{-jh}, \theta_h\phi_{jh} > 0)$ has power law tail and $\text{pr}(\theta_h\phi_{jh} > 0 \mid \Lambda_{-jh}) > 0$. The latter inequality is always true when the marginal $\text{pr}(\theta_h\phi_{jh} > 0)$ is positive. To prove $(\theta_h\phi_{jh} \mid \Lambda_{-jh})$ has power law tail, we apply Lemma 3.1. We first focus on proving the lemma when ϕ_{jh} satisfies the power law tail condition. As the local scale ϕ_{jh} is independent from $(\Lambda_{-jh}, \theta_h)$ given γ_h, its conditional density is

$$\text{pr}_{\phi_{jh}|\theta_h,\Lambda_{-(jh)}}(\phi) = \int_{\Re} \text{pr}_{\phi_{jh}|\gamma_h}(\phi)\ \text{pr}_{\gamma_h|\theta_h,\Lambda_{-jh}}(\gamma)\,\text{d}\gamma.$$

As the tail conditions hold for any possible prior on γ, we have

$$\text{pr}_{\phi_{jh}}(\phi) = \int_{\Re} \text{pr}_{\phi_{jh}|\gamma_h}(x)\ \text{pr}(\gamma)\,\text{d}\gamma, \quad \text{pr}_{\phi_{jh}}(\tilde{\phi}) \propto \tilde{\phi}^{-\alpha}, \quad \tilde{\phi} = \{\phi : \phi > L\}, \quad L \gg 0,$$

for any prior density function pr defined on $\Re$. Hence, $(\phi_{jh} \mid \theta_h, \Lambda_{-jh})$ is power law tail distributed. We now focus on proving the lemma when θ_h is power law tail distributed. Let

$\theta_h^* = (\theta_h \mid \theta_h > 0)$ and $\phi_{jh}^* = (\phi_{jh} \mid \phi_{jh} > 0)$. By Bayes' Theorem

$$\mathrm{pr}_{\theta_h^*|\phi_{jh}^*,\Lambda_{-jh}}(\theta) = \frac{\mathrm{pr}_{\Lambda_{-jh}|\phi_{jh}^*,\theta_h^*}(\Lambda_{-jh}; r)\,\mathrm{pr}_{\theta_h^*|\phi_{jh}^*}(\theta)}{\mathrm{pr}_{\Lambda_{-jh}|\phi_{jh}^*}(\Lambda_{-jh})}.$$

Since θ_h^* is independent from ϕ_{jh}^*, it is sufficient to prove that the function $\mathrm{pr}_{\Lambda_{-jh}|\phi_{jh}^*,\theta_h^*}(\Lambda_{-jh};\theta)$ decreases slower than $c\,\theta^{-\alpha}$, for $c, \alpha >$ positive constants, when $r \to \infty$.

Denoting $F_{\phi_{11}\ldots\phi_{pk},\theta_1,\ldots,\theta_{h-1},\theta_{h+1},\ldots,\theta_{k^*}|\phi_{jh}^*,\theta_h^*}$ the probability measure for conditional density $\mathrm{pr}_{\phi_{11}\ldots\phi_{pk},\theta_1,\ldots,\theta_{h-1},\theta_{h+1},\ldots,\theta_{k^*}|\phi_{jh}^*,\theta_h^*}$, we can write

$$\begin{aligned}
\mathrm{pr}_{\Lambda_{-jh}|\phi_{jh}^*,\theta_h^*}(\Lambda_{-jh};\theta) &= \int \mathrm{pr}_{\Lambda_{-jh}|\phi_{11}\ldots\phi_{pk},\theta_1,\ldots,\theta_{k^*}}(\Lambda_{-jh};\theta)\,\mathrm{d}F_{\phi_{11}\ldots\phi_{pk},\theta_1,\ldots,\theta_{h-1},\theta_{h+1},\ldots,\theta_{k^*}|\phi_{jh}^*,\theta_h^*} \\
&= \int \prod_{(s,l)\neq(j,h)} \mathrm{pr}_{\lambda_{sl}|\phi_{sl},\theta_l}(\lambda_{sl};\theta)\,\mathrm{d}F_{\phi_{11},\ldots,\theta_{k^*}|\phi_{jh}^*,\theta_h^*} \\
&= E\left\{ \prod_{(s,l)\neq(j,h)} \mathrm{pr}_{\lambda_{sl}|\phi_{sl},\theta_l}(\lambda_{sl};\theta) \;\middle|\; \phi_{jh}^*, \theta_h^* \Lambda_{-jh} \right\}
\end{aligned}$$

The product inside the expectation is zero when there is a pair of indices (s,l) such that $\lambda_{sl} \neq 0$ and $\theta_{sl} = 0$. However, since the probability $\mathrm{pr}(\theta_{sl} = 0 \mid \lambda_{sl} \neq 0) = 0$, we know that the expected value of the product between the functions $\mathrm{pr}_{\lambda_{sl}|\phi_{sl},\theta_l}(\lambda_{sl};\theta)$, given $\phi_{jh}^*, \theta_h^*, \Lambda_{-jh}$, is strictly positive. We now prove that $\mathrm{pr}_{\Lambda_{-jh}|\phi_{jh}^*,\theta_h^*}(\Lambda_{-jh};\theta)$ decreases slower than $c\theta^{-\alpha}$ for $c, \alpha > 0$. We can write the above expectation as

$$E\left\{ \prod_{s=1,l\neq h}^{p} \mathrm{pr}_k(\lambda_{sl}) \prod_{s\neq j} \mathrm{pr}_{\lambda_{sh}|\theta_{sh}}(\lambda_{sh};\theta_h^*) \;\middle|\; \phi_{jh}^*, \theta_h^*, \Lambda_{-jh} \right\},$$

where $\prod_{s=1,l\neq h}^{p} \mathrm{pr}_k(\lambda_{sl})$ is a product between $(k-1)\times p$ strictly positive random variables that does not depend on ϕ_{jh}^* and θ_h^*, while $\prod_{s\neq j} \mathrm{pr}_{\lambda_{sh}|\phi_{sh}\theta_h}(\lambda_{sh};\theta_h^*)$ is a product between p strictly positive random variables. In particular, if $\phi_{sh} = 0$, then $f_{\lambda_{sl}|\phi_{sh}\theta_h}(\lambda_{sh};\theta_h^*) = \mathbb{1}(\lambda_{sh} = 0)$. If $\phi_{sh} > 0$, then

$$\mathrm{pr}_{\lambda_{sh}|\phi_{sh},\theta_h}(\lambda_{sh};\theta_h^*) = (2\pi\phi_{sh}^*\theta_h^*)^{-0.5}\exp\left(-\frac{\lambda_{sh}^2}{2\phi_{sh}^*\theta_h^*}\right) > (2\pi\phi_{sh}^*\theta_h^*)^{-0.5}\exp\left(-\frac{\lambda_{sh}^2}{2\phi_{sh}^*}\right).$$

Therefore, the upper bound

$$\mathrm{pr}_{\lambda_{sh}|\phi_{sh},\theta_h}(\lambda_{sh};\theta_h^*) \geq \begin{cases} \min\{1,\ (2\pi\phi_{sh}^*\theta_h^*)^{-0.5}\}, & \text{if } \lambda_{sh=0} \\ (2\pi\phi_{sh}^*\theta_h^*)^{-0.5}\exp\{-\lambda_{sh}^2/(2\phi_{sh}^*)\} & \text{if } \lambda_{sh\neq 0}, \end{cases}$$

holds with probability equal to 1. For $\theta > 1$, we note that $f_{\lambda_{sh}|\phi_{sh},\theta_h}(\lambda_{sh};\theta) \geq \theta^{-0.5}u_{\lambda_{sh}}$ with

$$u_{\lambda_{sh}} = \begin{cases} \min\{1,\ (2\pi\phi_{sh}^*)^{-0.5}\}, & \text{if } \lambda_{sh} = 0, \\ (2\pi\phi_{sh}^*)^{-0.5}\exp\{-\lambda_{sh}^2/(2\phi_{sh}^*)\} & \text{if } \lambda_{sh} \neq 0. \end{cases}$$

Then,

$$E\Big\{ \prod_{(s,l)\neq(j,h)} \mathrm{pr}_{\lambda_{sl}|\phi_{sl},\theta_l}(\lambda_{sl};\theta_h^*) \ \Big|\ \phi_{jh}^*, \theta_h^*, \Lambda_{-jh} \Big\} \geq$$

$$E\Big\{ \prod_{s=1,l\neq h}^{p} \mathrm{pr}_k(\lambda_{sl}) \prod_{s\neq j} \theta_h^{*-0.5} u_{\lambda_{sh}} \ \Big|\ \phi_{jh}^*, \theta_h^*, \Lambda_{-jh} \Big\} =$$

$$\theta_h^{*-0.5\,(p-1)} E\Big\{ \prod_{s=1,l\neq h}^{p} \mathrm{pr}_k(\lambda_{sl}) \prod_{s\neq j} u_{\lambda_{sh}} \ \Big|\ \phi_{jh}^*, \Lambda_{-jh} \Big\},$$

where the expectation is strictly positive and not depending on θ_h. Therefore, for θ sufficiently large, $\mathrm{pr}_{\Lambda_{-jh}|\phi_{jh}^*,\theta_h^*}(\Lambda_{-jh};\theta) \geq c\theta^{-\alpha}$ holds with $c, \alpha > 0$, so that $(\theta_h \mid \phi_{jh}, \Lambda_{-jh})$ is power law tail distributed.

Hence, if any of the scale parameters is power law tail distributed for any prior on γ, then its distribution, conditionally on Λ_{-jh} and on the product of the other scale parameter, is power law tail distributed and, as a consequence, $(\lambda_{jh} \mid \Lambda_{-jh})$ is power law tail distributed. Since $\mathrm{pr}_{\lambda_{jh}|\Lambda_{-jh}}(\lambda) \geq c|\lambda|^{-\alpha}$ for certain c, α positive constants and $|\lambda| > L$ sufficiently large, in the same settings, we can write

$$\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda) = c|\lambda|^{-\alpha}\{1 + f(|\lambda|)\},$$

where $f(|\lambda|)$ is a positive function. Then,

$$\frac{\partial[\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} = -\frac{\alpha}{\lambda} + \frac{\partial f(\lambda)}{\partial\lambda} \qquad \text{for } \lambda > L \quad \text{and } L \gg 0,$$
$$\frac{\partial[\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} = \frac{\alpha}{\lambda} + \frac{\partial\{-f(\lambda)\}}{\partial\lambda} \qquad \text{for } \lambda < -L \quad \text{and } L \gg 0,$$

We now consider the sign of the derivative of $f(|\lambda|)$. If $f(|\lambda|)$ is not decreasing,

$$\frac{\partial[\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \geq -\frac{\alpha}{\lambda}, \qquad \text{for } \lambda > L \quad \text{and } L \gg 0,$$
$$\frac{\partial[\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \leq \frac{\alpha}{\lambda}, \qquad \text{for } \lambda < -L \quad \text{and } L \gg 0,$$

whereas if $f(|\lambda|)$ is decreasing, its derivative goes to zero when $|\lambda|$ goes to infinity. Therefore,

$$\frac{\partial[\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \geq f'_{lb}(\lambda) \qquad \text{for } \lambda > L \quad \text{and } L \gg 0,$$
$$\frac{\partial[\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \leq -f'_{lb}(|\lambda|) \qquad \text{for } \lambda < -L \quad \text{and } L \gg 0,$$

where $f'_{lb}(\lambda) < 0 \ \forall\, \lambda > 0$ and $\lim_{\lambda\to\infty} f'_{lb}(|\lambda|) = 0$. The proof is concluded by using this result along with the fact that $f_{\lambda_{jh}|\Lambda_{-(jh)}}(|\lambda|)$ is decreasing when $\lambda \to \infty$,

$$\frac{\partial[\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \leq 0 \qquad \text{for } \lambda > L \quad \text{and } L \gg 0$$
$$\frac{\partial[\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \geq 0 \qquad \text{for } \lambda > -L \quad \text{and } L \gg 0,$$

showing that the limit of the derivative for $|\lambda| \to \infty$ is equal to zero. □

The following definition introduces a type of local tail robustness property.

DEFINITION 3.2: *Consider model* (1.1) *with factors* H *known. Let* $\mathrm{pr}_{\lambda_{jh}|y,H,\Lambda_{-jh}}(\lambda)$ *denote the posterior density of* λ_{jh}*, given the data, conditional on any possible value of the other elements of* Λ_{k^*} *for any finite* k^**, and let* $\hat{\lambda}_{jh}$ *denote the conditional maximum likelihood estimate of* λ_{jh}

for any possible value of the other elements of Λ_{k^*}. *We say that the prior on* λ_{jh} *is tail robust if*

$$\lim_{\hat{\lambda}_{jh}\to\infty} \left| \hat{\lambda}_{jh} - \arg\max_{\lambda} \mathrm{pr}_{\lambda_{jh}|y,H,\Lambda_{-jh}}(\lambda) \right| = 0.$$

For a given sample, $\hat{\lambda}_{jh}$ is a fixed quantity; the above limit should be interpreted as what happens as the data support a larger and larger maximum likelihood estimate. In order for tail robustness to hold, we need the data to be sufficiently informative about the parameter λ_{jh} and the likelihood to be sufficiently regular; this is formalized as follows.

ASSUMPTION 3.1: *Let* $L(y;\Lambda,H,\Sigma)$ *denote the likelihood for data* y *conditionally on latent variables* H, *let* $l_s(\lambda)$ *denote the derivative function of the loglikelihood with respect to* λ_{jh}, *and let* $\mathcal{J}(\hat{\lambda}_{jh})$ *denote the negative of the second derivative of the loglikelihood with respect to* λ_{jh}, *evaluated at the conditional maximum likelihood estimate* $\hat{\lambda}_{jh}$. *Then* $l_s(\lambda)$ *is a continuous function for every* $\lambda \in \Re$ *and* $\mathcal{J}(\hat{\lambda}_{jh}) \geq \nu(\hat{\lambda}_{jh})$, *where* $\nu(\hat{\lambda}_{jh})$ *is of order* $O(1)$ *as* $\hat{\lambda}_{jh} \to \infty$.

This assumption can be verified for most of the cases mentioned in Chapter 1; for example, for Gaussian linear factor models $\mathcal{J}(\hat{\lambda}_{jh})$ is of order $O(1)$ with respect to $\hat{\lambda}_{jh}$.

THEOREM 3.2: *Under Assumption 3.1, if at least one scale parameter among* θ_h *or* ϕ_{jh} *is power law tail distributed for any possible prior distribution of* γ_h, *then the prior on* λ_{jh} *is tail robust according to Definition 3.2.*

Proof of Theorem 3.2. The mode of the conditional posterior density of λ_{jh} is $\tilde{\lambda}_{jh}$ such that

$$l_s(\tilde{\lambda}_{jh};y,H) + \frac{\partial}{\partial\lambda}\log\{\mathrm{pr}_{\lambda_{jh}|\Lambda_{-jh}}(\lambda)\}\big|_{\lambda=\tilde{\lambda}_{jh}} = 0, \tag{3.5}$$

where $l_s(\tilde{\lambda}_{jh};y,H)$ is the jhth element of the score function of the likelihood for the data y conditionally on the latent variables H, and $\mathrm{pr}_{\lambda_{jh}|\Lambda_{-jh}}$ is the conditional prior density function of $(\lambda_{jh} \mid \Lambda_{-jh})$. Given prior symmetry, without loss of generality, we focus on $\hat{\lambda}_{jh} > 0$. In a neighbourhood $(\hat{\lambda}_{jh}-\varepsilon, \hat{\lambda}_{jh}+\varepsilon)$ of the conditional maximum likelihood estimate $\hat{\lambda}_{jh}$ of λ_{jh}, we can approximate the score function using a Taylor expansion: $l_s(\lambda;y) = -\mathcal{J}(\hat{\lambda}_{jh})\,(\lambda-\hat{\lambda}_{jh})+\zeta_\varepsilon$, where $\mathcal{J}(\hat{\lambda}_{jh}) > 0$ is the negative of the derivative of $l_s(\lambda;y)$ evaluated at $\lambda = \hat{\lambda}_{jh}$, and ζ_ε is an approximation error term such that $\lim_{\varepsilon\to 0}\zeta_\varepsilon/\varepsilon = 0$. For $\hat{\lambda}_{jh}$ large enough, such that $\hat{\lambda}_{jh} - \varepsilon > L$ with $L \gg 0$, Lemma 3.2 holds for every λ in $(\hat{\lambda}_{jh} - \varepsilon, \hat{\lambda}_{jh} + \varepsilon)$, leading to the

lower bound

$$-\mathcal{J}(\hat{\lambda}_{jh})\,(\lambda-\hat{\lambda}_{jh})+f'_{lb}(\lambda)+\zeta_\varepsilon \leq l_s(\lambda;y)+\frac{\partial}{\partial\lambda}\log\{f_{\lambda_{jh}|\Lambda_{-jh}}(\lambda)\},$$

where $f'_{lb}(\lambda)$ is a non positive continuous function for every $\lambda>0$, $\lim_{\lambda\to+\infty} f'_{lb}(\lambda)=0$. Let ε be a function of $\hat{\lambda}_{jh}$ such that $\lim_{\hat{\lambda}_{jh}\to\infty}\varepsilon=0$ and $\lim_{\hat{\lambda}_{jh}\to\infty} f'_{lb}(\hat{\lambda}_{jh})/\varepsilon=0$. The limit for $\hat{\lambda}_{jh}\to\infty$ of the lower bound evaluated in $\hat{\lambda}_{jh}-\varepsilon$ is

$$\lim_{\hat{\lambda}_{jh}\to\infty}\mathcal{J}(\hat{\lambda}_{jh})\,\varepsilon+f'_{lb}(\hat{\lambda}_{jh}-\varepsilon)+\zeta_\varepsilon=\lim_{\hat{\lambda}_{jh}\to\infty}|\varepsilon|\,\{\mathcal{J}(\hat{\lambda}_{jh})+f'_{lb}(\hat{\lambda}_{jh}-\varepsilon)/|\varepsilon|+\zeta_\varepsilon/|\varepsilon|\}.$$

Under Assumption 3.1, $\lim_{\hat{\lambda}_{jh}\to\infty}\mathcal{J}(\hat{\lambda}_{jh})+f'_{lb}(\hat{\lambda}_{jh}-\varepsilon)/|\varepsilon|+\zeta_\varepsilon/|\varepsilon|\geq 0$, which guarantees $\hat{\lambda}_{jh}-\varepsilon\leq\tilde{\lambda}_{jh}\leq\hat{\lambda}_{jh}$, and, hence $\lim_{\hat{\lambda}_{jh}\to\infty}|\tilde{\lambda}_{jh}-\hat{\lambda}_{jh}|=0$, which proves the theorem. □

As an additional desirable property, we would like to control for the multiplicity problem within each column λ_h of the loadings matrix, corresponding to increasing numbers of false signals as the dimension p increases. This can be accomplished by imposing an asymptotically increasingly sparse property on the prior, which is defined as follows.

DEFINITION 3.3: *Let* $|\text{supp}_\zeta(\lambda_h)|$ *denote the cardinality of* $\text{supp}_\zeta(\lambda_h)=(j:|\lambda_{jh}|>\zeta)$. *Let* $s_p=o(p)$ *such that* $s_p\geq c_s\log(p)/p$ *for some constant* $c_s>0$. *We say that the prior on* Λ *defined in* (2.1) *is an asymptotically increasingly sparse prior if*

$$\lim_{p\to\infty}\text{pr}\{|\text{supp}_\zeta(\lambda_h)|>a\,s_p\mid\theta_h\}=0,\qquad \textit{for some constant } a>0.$$

The quantity $|\text{supp}_\zeta(\lambda_h)|$ represents an approximate measure of model size for continuous shrinkage priors and, conditionally on γ_h and θ_h, it is a priori distributed as a sum of independent Bernoulli random variables $\text{Ber}(\varpi_{\zeta jh})$, where

$$\varpi_{\zeta jh}=\text{pr}(|\lambda_{jh}|>\zeta\mid\gamma_h,\theta_h)\leq\frac{\theta_h\,g(w_j^\top\gamma_h)}{\zeta^2}.$$

We now provide sufficient conditions for an asymptotically increasingly sparse prior, allowing regulation of the sparsity behaviour of the prior of the columns of Λ for increasing dimension p.

THEOREM 3.3: *Consider the prior defined in* (2.1) *and* (3.2), *with* ϕ_{jh} $(j=1,\dots,p)$ *a priori independent given* γ_h. *If* $\text{pr}\{g(w_j^\top\gamma_h)\leq\nu_j(p)\}=1$, *with* $\nu_j(p)=O\{\log(p)/p\}$,

($j = 1, \ldots, p$), then the prior on Λ is asymptotically increasingly sparse according to Definition 3.3.

Proof of Theorem 3.3. Since the local scales are independent, conditionally on γ, we can apply the Chernoff's method and obtain the following upper bound

$$\text{pr}\{|\text{supp}_\zeta(\lambda_h)| > a s_p \mid \gamma_h, \theta_h\} \leq \exp(-s_p a\, t)\, \exp\left\{(e^t - 1) \sum_{j=1}^{p} \varpi_{\zeta j h}\right\},$$

for every $t > 0$ and $\varpi_{\zeta j h} = \{\theta_h\, g(w_j^\top \gamma_h)\}/\zeta^2$ a function of γ_h. Since $g(w_j^\top \gamma_h)$ is of order $\leq O(\log(p)/p)$ by assumption and is limited above with respect to γ_h, we can deduce $g(w_j^\top \gamma_h) \leq c_j \log(p)/p$ for p sufficiently large and for some constant $c_j > 0$ that does not depend on γ_h and is asymptotically of order $O(1)$ with respect to p. Then, for $p \gg 0$,

$$\sum_{j=1}^{p} g(w_j^\top \gamma_h) \leq \sum_{j=1}^{p} c_j \log(p)/p \leq p\, \log(p)/p \max_{1 \leq j \leq p} c_j = M \log(p),$$

where $M = \max_{1 \leq j \leq p} c_j$ does not depend on γ_h. Then, the upper bound is

$$\text{pr}\{|\text{supp}_\zeta(\lambda_h)| > a s_p \mid \gamma_h, \theta_h\} \leq \exp\left\{ - s_p a\, t + (e^t - 1)\, \frac{\theta_h}{\zeta^2}\, M \log(p)\right\}.$$

Let us choose $t = \log\{\zeta^2/(\theta_h M) + 1\}$. Since $s_p \geq \log(p)\, c_s$ for a certain $c_s > 0$, then, for any $a > (c_s t)^{-1}$, we can write

$$\text{pr}\{|\text{supp}_\zeta(\lambda_h)| > a s_p \mid \gamma_h, \theta_h\} \leq \exp\left\{ - \log(p)\, \tilde{a}\right\},$$

where $\tilde{a}$ is a positive constant such that $a = (1 + \tilde{a})(c_s t)^{-1}$. The upper bound does not depend on γ_h, so

$$\text{pr}\{|\text{supp}_\zeta(\lambda_h)| > a s_p \mid \theta_h\} \leq \nu(p)$$

with $\nu(p)$ of order $O(p^{-1})$ that goes to zero. □

The condition of the theorem is easily satisfied, for example when g is the multiplication of a bounded function and a suitable offset depending on p as assumed in Section III.i. The multiplicative gamma process (Bhattacharya & Dunson, 2011) and the cumulative shrinkage process

(Legramanti et al., 2020) do not satisfy the sufficient conditions of Theorem 3.3, and, furthermore, the following lemma holds.

LEMMA 3.3: *The multiplicative gamma process prior (Bhattacharya & Dunson, 2011) and the cumulative shrinkage process prior (Legramanti et al., 2020) are not asymptotically increasing sparse according to Definition 3.3.*

Proof of Lemma 3.3. In both the multiplicative gamma process and cumulative shrinkage process, priors on Λ are exchangeable within columns, that is $\text{pr}(|\lambda_{jh}| > \zeta \mid \theta_h) = \varpi_{\zeta h}$ does not depend on j. Then, the prior density of $|\text{supp}_\zeta(\lambda_h)|$, conditionally on θ_h is a priori distributed as a sum of independent and identically distributed Bernoulli random variables $\text{Ber}(\varpi_{\zeta h})$. Furthermore, $\varpi_{\zeta h}$ does not depend on p. By applying the Chernoff's method, we obtain

$$\text{pr}\{|\text{supp}_\zeta(\lambda_h)| < as_p \mid \theta_h\} \leq \exp\left\{ats_p + p\varpi_{\zeta h}(e^{-t} - 1)\right\},$$

for any $t > 0$ and with $1 - e^{-t} > 0$. Hence,

$$\text{pr}\{|\text{supp}_\zeta(\lambda_h)| > as_p \mid \theta_h\} \geq 1 - \exp[-p\{(1 - e^{-t})\varpi_{\zeta h} - ats_p/p\}],$$

where the limit of the lower bound is $\lim_{p\to\infty} 1 - \exp[-p\{(1 - e^{-t})\varpi_{\zeta h} - ats_p/p\}] = 1$, which concludes the proof. □

Although this section has focused on properties of the prior, we find empirically that these properties tend to carry over to the posterior, as will be illustrated in the subsequent sections. For example, the posterior exhibits asymptotic increasing sparsity; see Table 3.2 of Section IV, which shows results for a novel process in our proposed class that is much more effective than current approaches at identifying the true sparsity structure, particularly when p is large.

III Structured increasing shrinkage prior

III.1 *Model specification*

In this section we propose a structured increasing shrinkage process prior for generalized infinite factor models satisfying all the sufficient conditions in Propositions 2.1, 3.2 and Theorems 3.2--3.3.

Following the notation previously introduced, we specify

$$\begin{aligned} &\theta_h = \vartheta_h \rho_h, \quad \phi_{jh} \mid \gamma_h \sim \text{Ber}\{\text{logit}^{-1}(w_j^\top \gamma_h)\, c_p\}, \\ &\vartheta_h^{-1} \sim \text{Ga}(a_\theta, b_\theta), \quad a_\theta > 1, \quad \rho_h = \text{Ber}\,(1-\pi_h)\,, \quad \gamma_h \sim N_m(0, \sigma_\gamma^2 I_m), \end{aligned} \tag{3.6}$$

where we assume the link $g(x) = \text{logit}^{-1}(x)c_p$, with $\text{logit}^{-1}(x) = e^x/(1+e^x)$ and $c_p \in (0,1)$ a possible offset. Inspired by the cumulative shrinkage process (Legramanti et al., 2020), the process $\{\pi_h\}$, with $\pi_h = \text{pr}(\theta_h = 0)$, is defined as in (2.2). The prior expected number of non degenerate Λ columns is $E(\sum_{h=1}^{\infty} \rho_h) = \alpha$, suggesting setting α equal to the expected number of active factors. The prior specification is completed assuming $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with $\sigma_j^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$ for $j = 1, \ldots, p$. The hyperparameters can be chosen based on one's prior expectation of the signal-to-noise ratio, as σ_j^2 is the contribution of the noise component to the total variance of the jth variable. Figure 3.2 displays the prior distribution, obtained simulating 10 000 samples from the prior, of the proportion of variance $\text{tr}(\Lambda\Lambda^{\text{T}})/\text{tr}(\Omega)$ explained by the structured increasing shrinkage factor model for varying α, $\{E(\sigma^{-2}), \text{var}(\sigma^{-2})\}$, and $\{E(\vartheta_h^{-1}), \text{var}(\vartheta_h^{-1})\}$. The role of $\{E(\sigma^{-2}), \text{var}(\sigma^{-2})\}$ is not obvious, but suggests that sufficiently large mean and variance can guarantee higher flexibility. A sensitivity study, however, shows that posterior distributions of the same variance proportion tend to be robust to the specification of a_σ, b_σ as can be seen in Fig. 3.3, where posterior distributions are estimated on synthetic data sets with $n = 100$ and $p = 50$ generated from the Gaussian linear factor model $y_i \sim N_p(0, \Lambda_0\Lambda_0^T + I_p)$, assuming $\text{tr}(\Lambda_0\Lambda_0^{\text{T}})/\text{tr}(\Omega_0) = 0.966$, with $\Omega_0 = \Lambda_0\Lambda_0^{\text{T}} + I_{50}$. If we have incorrect expectations on the number of factors, i.e., α set small, a sufficiently concentrated prior on a large value of $E(\sigma^{-2})$ seems more suitable to model such data. Regarding prior elicitation, we recommend setting $b_\theta \geq a_\theta$ to induce a high enough proportion of variance explained by the factor model. Figure 3.2 reports empirical evidence of the influence of the hyperparameters regulating the distribution of ϑ_h on this quantity, showing that concentrated prior on a large value of $E(\vartheta^{-1})$ induces a smaller proportion of variance explained by the factor model.

The above specification respects the general class of priors defined in the previous section and, consequently, the following corollary holds.

COROLLARY 3.1: *The structured increasing shrinkage process defined in* (3.6)

i. is a strongly increasing shrinkage prior according to Definition 3.1;

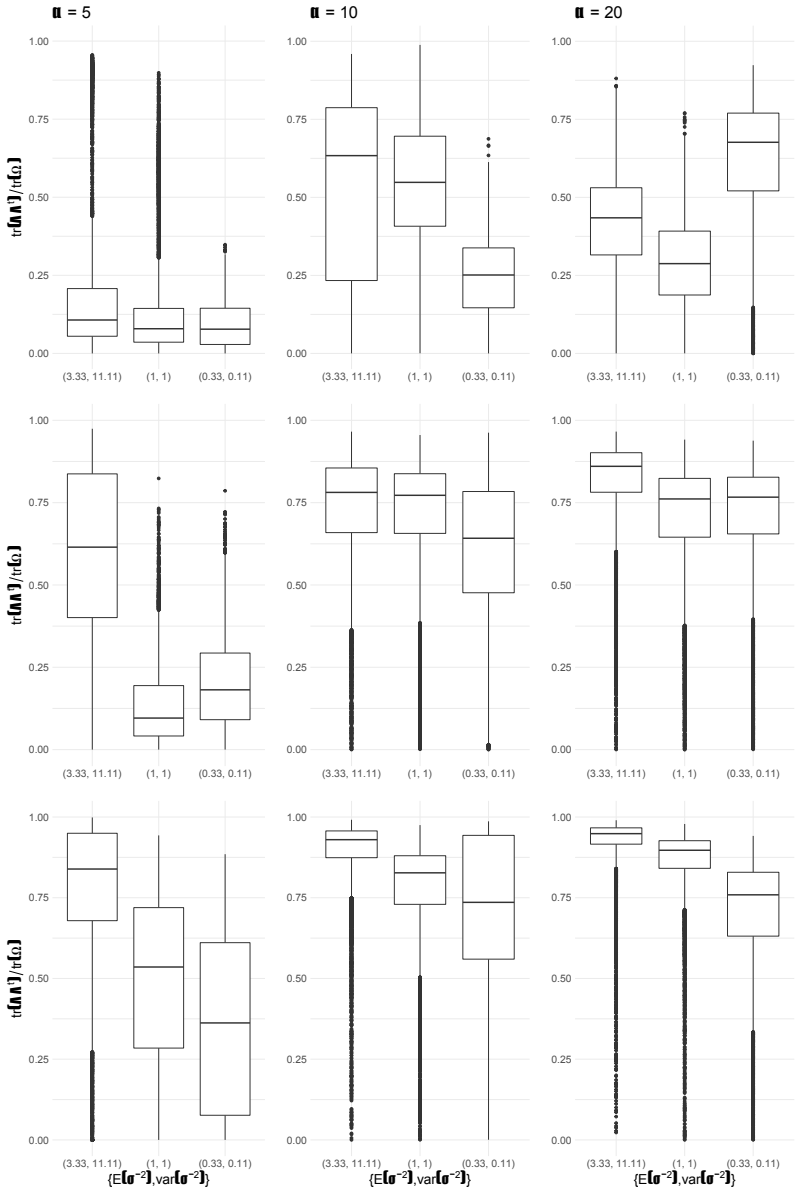

FIGURE 3.2: Boxplots of the prior distribution of the proportion of variance explained by the factor model $\text{tr}(\Lambda\Lambda^T)/\text{tr}(\Omega)$. The quantity is obtained simulating 10,000 samples from the prior distribution with varying values of the parameters. The horizontal axis characterize the effect of $\{E(\sigma^{-2}), \text{var}(\sigma^{-2})\}$; differences for $\alpha \in (5, 10, 20)$ are reported in each column; differences for $\{E(\vartheta^{-1}), \text{var}(\vartheta^{-1})\} \in \{(2, 2), (1, 0.5), (0.5, 0.125)\}$ are reported in each row.

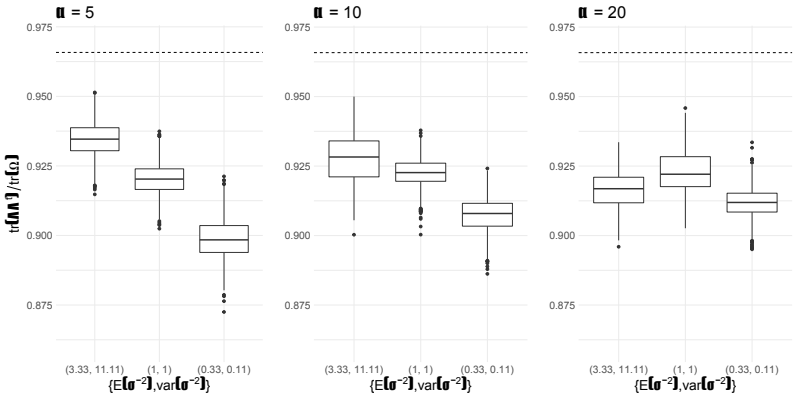

FIGURE 3.3: Boxplots representing the simulated posterior distribution of the proportion of variance explained by the factor model $\text{tr}(\Lambda\Lambda^{\text{T}})/\text{tr}(\Omega)$ for varying α and $\{E(\sigma^{-2}), \text{var}(\sigma^{-2})\}$. The dashed lines represent the proportion computed on the true value of Λ and Ω.

ii. *for any* $\tau \in (0, 1)$,

$$\text{pr}\left\{\frac{\text{tr}(\Omega_{k^*})}{\text{tr}(\Omega)} \leq \tau\right\} \leq \left(\frac{1}{1-\tau}\right) \frac{b^H}{1-b}\, \theta_0\, \frac{a_\sigma}{b_\sigma} \sum_{j=1}^{p} E(\phi_{j1}),$$

with $b = \{\alpha(1+\alpha)\}^{-1}$ *and* $\theta_0 = E(\vartheta_h)$.

Proof of Corollary 3.1. *i.* It is sufficient to prove the conditions required by Theorem 3.1. We have $E(\theta_h) = E(\vartheta_h)E(\rho_h) = E(\rho_h)\, b_\theta/(a_\theta - 1)$, where

$$E(\rho_h) = 1 - \sum_{l=1}^{h} E(u_l) = 1 - \sum_{l=1}^{h-1} E(u_l) - E(u_h) = E(\rho_{h-1}) - E(u_h).$$

Since the random variable u_l is obtained as a product of positive random variables, $E(u_l) > 0$ for every $l = 1, \ldots, h$. Therefore $E(\theta_h) < E(\theta_{h-1})$ for each $h = 2, \ldots, \infty$.

ii. It is sufficient to prove the conditions required by Proposition 2.1. It is straightforward to verify $E(\phi_{jh}) \leq 1$ for $j = 1, \ldots, p$ and $h = 1, \ldots, \infty$. The factor-specific scale expectation is

$$E(\theta_h) = E(\vartheta_h) \left(\frac{\alpha}{1+\alpha}\right) \left(\frac{\alpha}{1+\alpha}\right)^{h-1},$$

which can be written in a form ab^{h-1}. The elements σ_j^{-2} are gamma distributed guaranteeing finite expectation for all $j = 1, \ldots, p$.

□

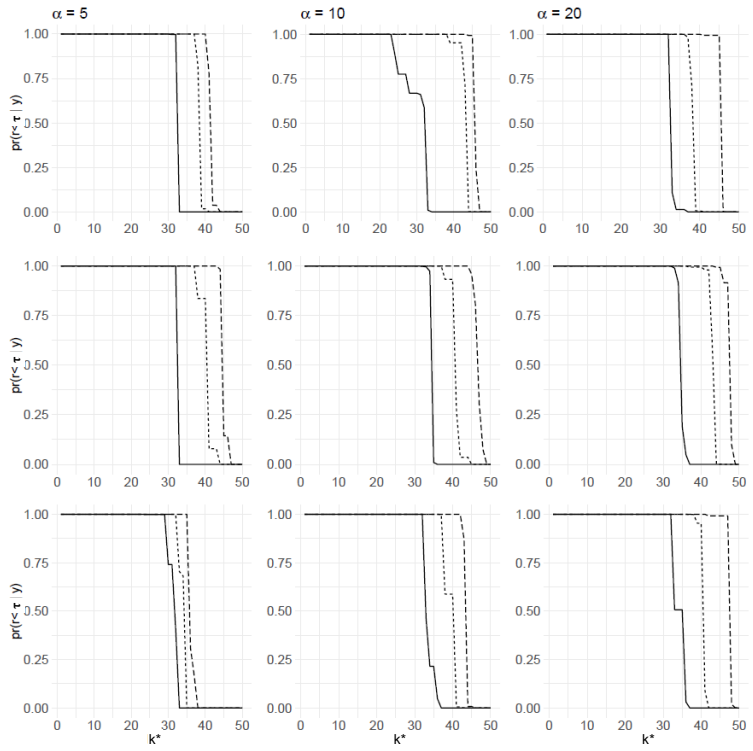

FIGURE 3.4: Monte Carlo approximation of the posterior probability of truncation error $\text{pr}(r < \tau \mid y)$, with $r = \text{tr}(\Omega_{k^*})/\text{tr}(\Omega)$, at varying of k^*. The quantity is computed for τ equal to 0.75 (—), 0.9 (- - -), and 0.95 (– – –) and varying $\alpha \in (5, 10, 20)$ over the columns and $\{E(\sigma^{-2}), \text{var}(\sigma^{-2})\} \in \{(3.33, 11.11), (1, 1), (0.33, 0.11)\}$ over the rows of the figure.

We conducted a simulation study on the posterior distribution of $\{\text{tr}(\Omega_{k^*})/\text{tr}(\Omega) \leq \tau\}$ for varying hyperparameters, and found that the results, summarised in Fig. 3.4, were quite consistent with our prior truncation error bounds. Considering the same synthetic data presented before, we found that if Λ_0 is sparse, a small value of α induces good approximations even with k^* smaller than the true number of factors. The inferred sparsity pattern in Λ is robust to the prior distribution for σ^{-2}.

The prior concentration of the structured increasing shrinkage process in (3.6) follows from

(3.3):

$$\text{pr}(|\lambda_{jh}| > \zeta) \leq \frac{E(\vartheta_h)\{1 - E(\pi_h)\}E(\phi_{jh})}{\zeta^2} = \frac{\theta_0 \{\alpha/(1+\alpha)\}^h}{\zeta^2}\frac{c_p}{2}.$$

In addition, the inverse gamma prior on ϑ_h implies a power law tail distribution on θ_h inducing robustness properties on λ_{jh} as formalized by the next corollary of Proposition 3.2 and Theorem 3.2.

COROLLARY 3.2: *Under the structured increasing shrinkage process defined in* (3.6)

i. the marginal prior distribution on λ_{jh} $(j = 1, \ldots, p; h = 1, 2, \ldots)$ *has power law tails;*

ii. under Assumption 3.1, the prior on λ_{jh} $(j = 1, \ldots, p; h = 1, 2, \ldots)$ *is tail robust according to Definition 3.2.*

Proof of Corollary 3.2. It is sufficient to prove the conditions required by Theorem 3.2. The probability density function of the column scale θ_h $(h = 1, \ldots, \infty)$ of model (3.6) evaluated at a certain $\theta > 0$ is

$$\text{pr}_{\theta_h}(\theta) = \text{pr}(\rho_h = 1)\,\text{pr}_{\vartheta_h}(\theta) \propto \theta^{-a_\theta - 1}\exp(-b_\theta/\theta),$$

where $\text{pr}_{\vartheta_h}(\theta)$ is the inverse gamma probability density function evaluated at θ. The function $\theta^{-a_\theta-1}\exp(-b_\theta/\theta)$ is of order $O(\theta^{-a_\theta-1})$ as θ goes to infinity. Since $a_\theta > 0$, we conclude that the factor-specific scale θ_h is power law tail distributed. The independence between θ_h and γ_h $(h = 1, \ldots, \infty)$ guarantees that the latter result holds for any possible prior distribution pr_γ on γ. □

Finally, it is important to assess the joint sparsity properties of the prior on each column of Λ. This is formalized in the following corollary of Theorem 3.3.

COROLLARY 3.3: *If* $c_p = O\{\log(p)/p\}$ *the structured increasing shrinkage process defined in* (3.6) *is asymptotically increasingly sparse according to Definition 3.3.*

Proof of Corollary 3.3. It is sufficient to prove the conditions required by Theorem 3.3. The structured increasing shrinkage prior is such that, for every $j = 1, \ldots, p$ and $h \geq 1$, we have $g(w_j^\top \gamma_h) \leq c_p < 1$. The proof is obtained under the assumption $c_p = O\{\log(p)/p\}$. □

III.II *Posterior computations*

Posterior inference is conducted via Markov chain Monte Carlo sampling. Following common practice in infinite factor models (Bhattacharya & Dunson, 2011; Legramanti et al., 2020) we use an adaptive Gibbs algorithm, which truncate the model to k^* factors, adapting the value of k^* only at some Gibbs iterations (see 1 in Chapter 2 for further details).

The decomposition of θ_h into two parameters ρ_h and ϑ_h allows one to identify the inactive columns of Λ, corresponding to the redundant and neglectable factors, as those with $\rho_h = 0$, while k_a indicates the number of active columns of Λ. Consequently, at the adaptive iteration $t+1$, the truncation level $k^{(t+1)}$ is set to $k^{(t+1)} = k_a^{(t)} + 1$ if $k_a^{(t)} < k^{(t)} - 1$, and $k^{(t+1)} = k^{(t)} + 1$ otherwise. Given $k^{(t+1)}$, the number of factors of the truncated model at iteration $t+1$, the sampler draws the model parameters from the corresponding posterior full conditional distributions. The detailed steps of the adaptive Gibbs sampler for the structured increasing shrinkage prior in case of Gaussian data $y_i = z_i \sim N_p(0, \Lambda\Lambda^\top + \Sigma)$ are reported below.

The notation $(x \mid -)$ denotes the full conditional distribution of x conditionally on everything else. Given k^* the number of factors of the truncated model, the sampler cycles through the following steps.

STEP 1 Update in parallel the p elements of Σ, by sampling

$$(\sigma_j^{-2} \mid -) \sim \text{Ga}\left\{a_\sigma + \frac{n}{2},\, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(z_{ij} - \lambda_j^\top \eta_i)^2\right\}.$$

STEP 2 Update, for $i = 1, \dots, n$, the factor η_i according to the posterior full conditional

$$(\eta_i \mid -) \sim N_{k^*}\big\{(I_{k^*} + \Lambda_{k^*}^\top \Sigma^{-1}\Lambda_{k^*})^{-1}\Lambda_{k^*}^\top \Sigma^{-1} z_i,\, (I_{k^*} + \Lambda_{k^*}^\top \Sigma^{-1}\Lambda_{k^*})^{-1}\big\}.$$

The distribution is conditional to Λ and ϵ, so we can update in parallel the n vectors η_i.

STEP 3 Let $\tilde{\lambda}_{jh}$ denote the continuous underlying loadings element such that $\lambda_{jh} = \phi_{jh}\rho_h\tilde{\lambda}_{jh}$ $(h = 1, \dots, k^*)$. Then, update the elements $\tilde{\lambda}_{jh}$ by sampling from the independent full conditional posterior distributions of the row vector $\tilde{\lambda}_j$ $(j = 1, \dots, p)$,

$$(\tilde{\lambda}_j \mid -) \sim \mathcal{N}_{n_f}\big\{(D_j^{-1} + F_j H^\top H F_j)^{-1} F_j H^\top z_j,\, (D_j^{-1} + F_j H^\top H F_j)^{-1}\big\},$$

where $D_j^{-1} = \text{diag}(\vartheta_1^{-1}, \dots, \vartheta_{k^*}^{-1})$, $F_j = \text{diag}(\rho_1\phi_{1j}, \dots, \rho_{k^*}\phi_{k^*})$, and z_j is the column

data vector. The distribution is conditional to H and the scale matrix and vector Φ and θ, thus we can update in parallel the p vectors $\tilde{\lambda}_j$. Finally, set $\lambda_{jh} = \phi_{jh}\rho_h\tilde{\lambda}_{jh}$ for any $j = 1, \dots, p$ and $h = 1, \dots, k^*$.

STEP 4 Update the local scale parameters proceeding as follows.

STEP 4.I Update the local scale by sampling from the full conditional distributions of ϕ_{jh}. If $h \in \{1, \dots, k^* : \rho_h = 0\}$, then sample from the Bernoulli defined as

$$\text{pr}(\phi_{jh} = \xi) = \begin{cases} \{1 - \text{logit}^{-1}(w_j^\top \gamma_h)\, c_p\} & \text{for } \xi = 0 \\ \text{logit}^{-1}(w_j^\top \gamma_h)\, c_p & \text{for } \xi = 1. \end{cases}$$

Given the linear predictor matrix $w\Gamma$, we can update in parallel ϕ_{jh} for $j = 1, \dots, p$ and $h \in \{1, \dots, k^* : \rho_h = 0\}$. If $h \notin \{1, \dots, k^* : \rho_h = 0\}$, sample from

$$\text{pr}(\phi_{jh} = \xi) \propto \begin{cases} \{1 - \text{logit}^{-1}(w_j^\top \gamma_h)\, c_p\}\ \text{pr}_N\left(z_j; H\lambda_j - \phi_{jh}^{(t-1)}\tilde{\lambda}_{jh}\eta_h, I_n\right) & \text{for } \xi = 0 \\ \text{logit}^{-1}(w_j^\top \gamma_h)\, c_p\ \text{pr}_N\left\{z_j; H\lambda_j + \left(1 - \phi_{jh}^{(t-1)}\right)\tilde{\lambda}_{jh}\eta_h, I_n\right\} & \text{for } \xi = 1. \end{cases}$$

where $\text{pr}_N(x; \mu, I_n)$ is the multivariate density function of the n-variate Gaussian distribution with mean μ, variance equal to the identiy matrix, and evaluated at x. We use $\phi_{jh}^{(t-1)}$ to denote the parameter ϕ_{jh} sampled at the previous iteration of the Gibbs and η_h indicating the hth column of H. The distribution of each ϕ_{jh} depends on $w\Gamma$ and on the elements ϕ_{lj} $(l = 1, \dots, h)$ via λ_j. Therefore, the update is sequential with respect to the index h and requires to set $\lambda_{jh} = \phi_{jh}\tilde{\lambda}_{jh}$ after having sampled ϕ_{jh} for any $h \notin \{1, \dots, k^* : \rho_h = 0\}$. On the other hand, we can update in parallel with respect to the index $j = 1, \dots, p$.

STEP 4.II Let $\phi_{jh} = \varphi_{jh}\tilde{\phi}_{jh}$, with φ_{jh} and $\tilde{\phi}_{jh}$ independent a priori and distributed as $\text{Ber}\{\text{logit}^{-1}(w_j^\top \gamma_h)\}$ and $\text{Ber}(c_p)$, respectively. Update φ_{jh}, for $j = 1, \dots, p$ and $h = 1, \dots, k^*$, setting $\varphi_{jh} = 1$ if $\phi_{jh} = 1$ and sampling from the full conditional distribution

$$\text{pr}(\varphi_{jh} = l) \propto \begin{cases} 1 - \text{logit}^{-1}(w_j^\top \gamma_h) & \text{for } l = 0, \\ \text{logit}^{-1}(w_j^\top \gamma_h)(1 - c_p) & \text{for } l = 1, \end{cases}$$

if $\phi_{jh} = 0$. Given $w\Gamma$, the elements φ_{jh} $(j = 1, \dots, p, h = 1, \dots, k^*)$ are indepen-

dently distributed and can be updated in parallel.

STEP 4.III Update each γ_h exploiting the Pólya-Gamma data-augmentation strategy (Polson et al., 2013). Let $\text{pr}(x) \propto \sum_{n=0}^{\infty}(-1)^n A_n(2\pi x^3)^{-0.5}\exp\{-(2n+b)^2(8x)^{-1} - 0.5c^2x\}$ indicate the probability density function of a Pólya-Gamma distributed random variable $x \sim \text{PG}(b,c)$. For each $h = 1,\dots,k^*$, generate p independent random variables $d_{j(h)}$ sampling from the full conditional distribution $(d_{j(h)} \mid -) \sim \text{PG}(1, w_j^\top \gamma_h^{(t-1)})$. Let $D_{(h)}$ denote the $p \times p$ diagonal matrix with entries $d_{j(h)}$ $(j = 1,\dots,p)$ and define the $m \times m$ diagonal matrix $S = \sigma_\gamma^2 I_m$. For each $h = 1,\dots,k^*$, update γ_h sampling from

$$(\gamma_h \mid -) \sim N_m\{(w^\top D_{(h)} w + S^{-1})^{-1}(w^\top \kappa_h),\ (w^\top D_{(h)} w + S^{-1})^{-1}\},$$

where κ_h is a p-dimensional vector with the j-th entry equal to $\varphi_{jh} - 0.5$. Given φ_{jh} $(j = 1,\dots,p;\ h = 1,\dots,k^*)$, we can update in parallel all the vectors γ_h $(h = 1,\dots,k^*)$.

STEP 5 Update the factor-specific scale parameters proceeding as follows.

STEP 5.I Update the parameter ϑ_h $(h = 1,\dots,k^*)$ by sampling ϑ_h^{-1} from the full conditional distribution $\text{Ga}(a_\theta + 0.5p, b_\theta + 0.5\sum_{j=1}^{p}\tilde{\lambda}_{jh}^2)$.

STEP 5.II Following Legramanti et al. (2020), define the independent indicators ξ_h $(h = 1,\dots,p)$ with prior $\text{pr}(\xi_h = l) = u_l$. Update the augmented data ξ_h by sampling from the full conditional distribution

$$\text{pr}(\xi_h = l) \propto \begin{cases} u_l\ \text{pr}_N\{\text{vec}(z); \text{vec}(H\Lambda) - \text{vec}(\eta_h\lambda_h^\top), I_{np}\} & \text{for} \quad l = 1,\dots,h \\ u_l\ \text{pr}_N\{\text{vec}(z); \text{vec}(H\Lambda) + (1-\rho_h^{(t-1)})\text{vec}(\eta_h\lambda_h^{*\top}), I_{np}\} & \text{for} \quad l = h+1,\dots,k^*, \end{cases}$$

where we define the row vector $\lambda_h^{*\top}$ such that $\lambda^\top = \rho_h\lambda_h^{*\top}$. Then, $\rho_h = 1$ if $\xi_h > h$, else $\rho_h = 0$. The full conditional distribution of ξ_h depends on the value of ρ_l $(l = 1,\dots,k^*)$, implying to immediately set $\lambda^\top = \rho_h\lambda_h^{*\top}$ and to update ρ_h sequentially with respect to the index $h = 1,\dots,k^*$.

STEP 5.III For $l = 1,\dots,k^*-1$, sample v_l from

$$(v_l \mid -) \sim \text{Be}\Big\{1 + \sum_{h=1}^{k^*}\mathbb{1}(\xi_h = l), \alpha + \mathbb{1}(\xi_h > l)\Big\},$$

while set $v_{k^*} = 1$. Since, given u, the distribution are conditionally independents, we can update the k^* elements of the vector v in parallel. Finally, update $u_l = v_l \prod_{m=1}^{l-1}(1 - v_m)$, for $l = 1, \dots, k^*$.

The computational complexity of the algorithm is of order $nk^3 + np^3 + nkp^2 + pk^3 + npk^2 + pkm + km^3$ per iteration, assuming a standard implementation for the inversion of a $n \times n$ matrix with complexity n^3. Generally $p > k$ and $p > m$, such that the leading term is np^3, which is the computational complexity of updating n rows of a factor matrix H in a Gibbs sampler of any Gaussian linear factor model. When multiple processors are available, the computational complexity is divided by the number of processors, since in each step it is possible to compute in parallel at least k elements.

III.III *Identifiability and posterior summaries*

As already mentioned in the previous chapter, the matrices H and Λ are only identifiable up to an arbitrary rotation P with $PP^\top = I_k$. This is a well known problem in Bayesian factor models since makes difficult to extract meaningful posterior summaries from Markov chain Monte Carlo samples. In Section 2.IV.ii we overcome this issue following the literature and applying a post-processing algorithms that aligns the posterior samples of Λ and H through orthogonal rotations and label and sign column switching.

Unfortunately, such post hoc alignment algorithms destroy the structure we have carefully imposed on the loadings in terms of sparsity and dependence on meta-covariates. Therefore, we propose in this chapter a different solution to obtain a point estimate of Λ based on finding a representative Monte Carlo draw $\Lambda^{(t)}$ consistently with the proposals of Dahl (2006) and Wade et al. (2018) in the context of Bayesian model-based clustering. Specifically, we summarize Λ and $\gamma = (\gamma_1, \gamma_2, \dots)$ through $\Lambda^{(t^*)}$ and $\gamma^{(t^*)}$ sampled at iteration t^*, characterized by the highest marginal posterior density function $\text{pr}(\Lambda, \gamma, \Sigma \mid y)$ obtained by integrating out the scale parameters θ_h, ϕ_{jh} $(j = 1, \dots, p, h = 1, \dots)$ and the latent factors H from the posterior density function. Formally, we select the iteration $t^* \in \{1, \dots, T\}$ such that

$$\text{pr}(\Lambda^{(t^*)}, \gamma^{(t^*)}, \Sigma^{(t^*)} \mid y) > \text{pr}(\Lambda^{(t)}, \gamma^{(t)}, \Sigma^{(t)} \mid y) \quad (t = 1, \dots, T),$$

where $t = 1, \dots, T$ indexes the posterior samples. Under the structured increasing shrinkage prior described in Section III.i, these computations are straightforward. The matrices $\Lambda^{(t^*)}, \gamma^{(t^*)}, \Sigma^{(t^*)}$

are Monte Carlo approximations of the maximum a posteriori estimator, which corresponds to the Bayes estimator under L_∞ loss. Although one can argue that L_∞ is not an ideal choice of loss philosophically in continuous parameter problems, it nonetheless is an appealing pragmatic choice in our context and is broadly used in other sparse estimation contexts, as in the algorithm proposed by Ročková & George (2016) that similarly aims to recover a strongly sparse posterior mode of an overparameterized factor model. In addition, while the likelihood is symmetric over all the infinite orthogonal rotations of Λ and H, the posterior is symmetric only with respect to the sign switching of each factor, as a consequence of the asymmetry caused by the increasings and local shrinkage induced by the prior. In other terms, given fixed the sign of an element in each column of Λ, there exists a unique global maximum a posteriori, which partially justifies our approach.

IV Simulation experiments

We assess the performance of our structured increasing shrinkage prior compared with existing approaches (Bhattacharya & Dunson, 2011; Legramanti et al., 2020) through a simulation study. We have a particular interest in inferring sparse and interpretable loadings matrices Λ, but also assessing performance in estimating the induced covariance matrix Ω and number of factors. We generate synthetic data from four scenarios based on different loadings structures. For each scenario we simulate $R = 25$ data sets with $n = 250$ observations from $y_i \sim N_p(0, \Lambda_0\Lambda_0^\top + I_p)$ $(i = 1, \dots, n)$. In Scenario a, we assume non sparse Λ_0, sampling the loadings λ_{jh} from a Gaussian distribution with mean zero, variance equal to $\sigma^2_\lambda = 1$ and ordering them to obtain decreasing variance over the columns. To ensure that each element λ_{jh} represents a signal, we shifted them away from zero by $\sigma^2_\lambda/3$. In Scenario b we remove the decreasing behaviour and introduce a random sparsity pattern characterized by an increasing number of zero entries over the column index. The loadings matrix for Scenario c is characterized by both the decreasing behaviour over the columns of Scenario a and the random sparsity structure of Scenario b. Finally, in Scenario d, while the decreasing behaviour is kept, we induce a sparsity pattern dependent on a meta-covariate matrix w_0 including variable with four balanced categories, a continuous variable sampled from a multivariate Gaussian distribution, and a continuous variable where the p elements are sampled from p gamma distributions.

For each scenario we consider four combinations of dimension and sparsity level of Λ_0. We let $(p, k, s) \in \{(16, 4, 0.6),\ (32, 8, 0.4),\ (64, 12, 0.3),\ (128, 16, 0.2)\}$, where s is the proportion

of nonzero entries of Λ, with the exception of Scenario a where $s = 1$. To estimate the structured increasing shrinkage model, we set w equal to the p-variate column vector of 1s, $\sigma_\gamma = 1$ and, consistent with Corollary 3.3, $c_p = 2e \log(p)/p$, which belongs to $(0, 1)$ for every $p \geq 15$. In Scenario d we also estimate and compare a correctly specified structured increasing model with $w = w_0$. In these settings, the algorithm reported in Section III.ii takes $0.07 - -0.73$ seconds of computational time per iteration, depending on the dimension p, using an R implementation on a laptop with Intel Core i5-6200U CPU and 15.8 GB of RAM. The computational cost of each iteration is notable when p is large and motivates further researches on possible alternative algorithms, as that we will largely discuss in Chapter 4.

For the method proposed by Ročková & George (2016), we set the hyperparameters they suggested; for the remaining approaches, we set $a_\sigma = 1$ and $b_\sigma = 0.3$ and follow the hyperparameter specification and factor selection guidelines in Section III of Chapter 2. Results are obtained by running the algorithms for 25000 iterations discarding the first 10000 iterations. Then, we thin the Markov chain, saving every 5-th sample. We adapt the number of active factors at iteration t with probability $p(t) = \exp(-1 - 5\,10^{-4}t)$.

TABLE 3.1: Median and interquartile range of LPML and $E(k_a \mid y)$ in 25 replications of Scenario a for different combinations of (p, k); Scenario a is a worst case for the proposed SIS method.

	(p, k)	MGP		CUSP		SIS	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
LPML	$(16, 4)$	-28.68	0.42	-28.68	0.43	-28.65	0.41
	$(32, 8)$	-60.08	0.45	-60.09	0.45	-60.07	0.49
	$(64, 12)$	-117.68	0.56	-117.75	0.53	-117.88	0.56
	$(128, 16)$	-225.04	1.04	-225.13	1.04	-228.76	1.47
$E(k_a \mid y)$	$(16, 4)$	8.17	1.44	4.00	0.00	4.00	0.00
	$(32, 8)$	10.68	0.33	8.00	0.00	8.00	0.00
	$(64, 12)$	14.16	1.09	12.00	0.00	12.00	0.00
	$(128, 16)$	17.03	0.47	16.00	0.00	18.00	0.02

LPML, logarithm of the pseudo-marginal likelihood; CUSP, cumulative shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process; $Q_{0.5}$, median; IQR, interquartile range.

Scenario a is a worst case for the proposed method since there is no sparsity, no structure, and the elements of the loadings matrix are similar in magnitude. However, even in this case, structured increasing shrinkage performs essentially identically to the best competitor, as illustrated by the

results in Table 3.1. We report the median and interquartile range over the R replicates of the logarithm of the pseudo-marginal likelihood (Gelfand & Dey, 1994) and of the estimated posterior mean of the number of factors $E(k_a \mid y)$.

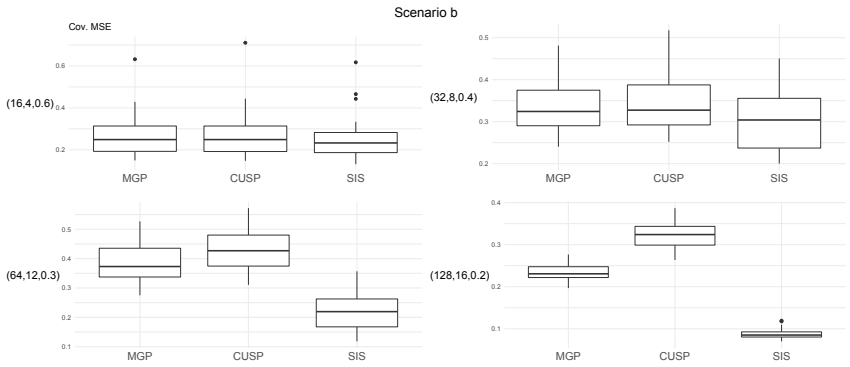

FIGURE 3.5: Boxplots of mean squared error of the covariance matrix of each model for different combinations of (p, k, s) in Scenario b. Cov. MSE, covariance mean squared error; CUSP, cumulative shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process.

Scenario b judges performance in detecting sparsity. The proposed approach shows better performance in terms of the logarithm of the pseudo-marginal likelihood and mean squared error of the covariance matrix, particularly as sparsity increases, as displayed in Fig. 3.5. Consistent with Legramanti et al. (2020), the covariance mean squared error is estimated in each simulation by $\sum_{j,l}^{p} \sum_{t=1}^{S} (\omega_{jl}^{(t)} - \omega_{jl0})^2 / \{p(p+1)/2\}$, where ω_{jl0} and $\omega_{jl}^{(t)}$ are the elements jl of $\Omega_0 = \Lambda_0 \Lambda_0^{\mathrm{T}} + I_p$ and $\Omega^{(t)} = \Lambda^{(t)} \Lambda^{(t)\mathrm{T}} + I_p$, respectively. The proposed approach allows exact zeros in the loadings, while the competitors require thresholding to infer sparsity. In particular, we set λ_{jh} to zero when $|\lambda_{jh}|$ $(j = 1, \ldots, p)$ is under a certain threshold. We choose the threshold equal to 0.05, which is consistent with the value of the hyperparameter θ_∞ used in the cumulative shrinkage process. We evaluate performance in inferring the sparsity pattern via the mean classification error:

$$MCE = \frac{1}{S} \sum_{t=1}^{S} \frac{\sum_{j=1}^{p} \sum_{h=1}^{k^{*(t)}} |\mathbb{1}(\lambda_{jh0} = 0) - \mathbb{1}(\lambda_{jh}^{(t)} = 0)|}{pk},$$

where $k^{*(t)}$ is the maximum between the true number of factors k and $k_a^{(t)}$, and λ_{jh0} and $\lambda_{jh}^{(t)}$ are the elements jh of Λ_0 and $\Lambda^{(t)}$, respectively. If $k_a^{(t)}$ or k are smaller than k^*, we fix the higher indexed columns at zero, possibly leading to a mean classification error bigger than one. To

TABLE 3.2: Median and interquartile range of the mean classification error computed in 25 replications assuming Scenario b and several combinations of (p, k, s)

MCE	(p, k, s)	MGP		CUSP		SIS	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
	$(16, 4, 0.6)$	1.06	0.16	0.38	0.01	0.24	0.09
	$(32, 8, 0.4)$	0.70	0.07	0.48	0.08	0.16	0.09
	$(64, 12, 0.3)$	0.61	0.07	0.58	0.01	0.09	0.06
	$(128, 16, 0.2)$	0.49	0.03	0.52	0.08	0.04	0.01

MCE, mean classification error; MGP, multiplicative gamma process; CUSP, cumulative shrinkage process; SIS, structured increasing shrinkage process; $Q_{0.5}$, median; IQR, interquartile range.

address column order ambiguity and label switching, we compute the mean classification error only after having ordered the columns of $\Lambda^{(t)}$ (for $t = 1, \ldots, T$), for each model, increasingly with respect to the number of zero entries identified. The results reported in Table 3.2 show that the proposed structured increasing shrinkage prior is much more effective at identifying sparsity in Λ, maintaining good performance even with large p and in strongly sparse contexts. Also, more accurate estimation of the number of factors is obtained, as reported in Table 3.3.

Similar comments apply in Scenarios c and d reported in Fig. 3.6. The superior performance of the structured increasing shrinkage model is only partially diminished in Scenario c for large p for the logarithm of the pseudo-marginal likelihood. In Scenario d, the use of meta-covariates has a mild benefit in identifying the sparsity pattern. In lower signal-to-noise settings, meta-covariates have a bigger impact, and they also aid interpretation, as illustrated in the next section.

TABLE 3.3: Median and interquartile range of the LPML, Cov. MSE and of $E(k_a \mid y)$ computed in 25 replications assuming Scenario b and several combinations of (p, k, s)

	(p, k, s)	MGP		CUSP		SIS	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
LPML	$(16, 4, 0.6)$	-28.20	0.33	-28.20	0.33	-28.17	0.32
	$(32, 8, 0.4)$	-56.95	0.53	-57.00	0.51	-56.80	0.49
	$(64, 12, 0.3)$	-111.35	0.70	-111.71	0.74	-110.76	0.89
	$(128, 16, 0.2)$	-211.65	0.74	-215.94	1.57	-210.19	0.86
Cov. MSE	$(16, 4, 0.6)$	0.25	0.12	0.25	0.12	0.23	0.10
	$(32, 8, 0.4)$	0.32	0.08	0.33	0.10	0.30	0.12
	$(64, 12, 0.3)$	0.37	0.10	0.43	0.11	0.22	0.09
	$(128, 16, 0.2)$	0.23	0.03	0.32	0.04	0.09	0.01
$E(k_a \mid y)$	$(16, 4, 0.6)$	8.91	1.52	4.00	0.00	4.00	0.00
	$(32, 8, 0.4)$	11.27	1.48	7.00	1.00	8.00	0.00
	$(64, 12, 0.3)$	14.72	1.49	11.00	0.00	12.00	0.00
	$(128, 16, 0.2)$	17.16	0.81	12.00	1.75	16.00	0.00

LPML, logarithm of the pseudo-marginal likelihood; Cov. MSE, covariance mean squared error; CUSP, cumulative increasing shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process; $Q_{0.5}$, median; IQR, interquartile range.

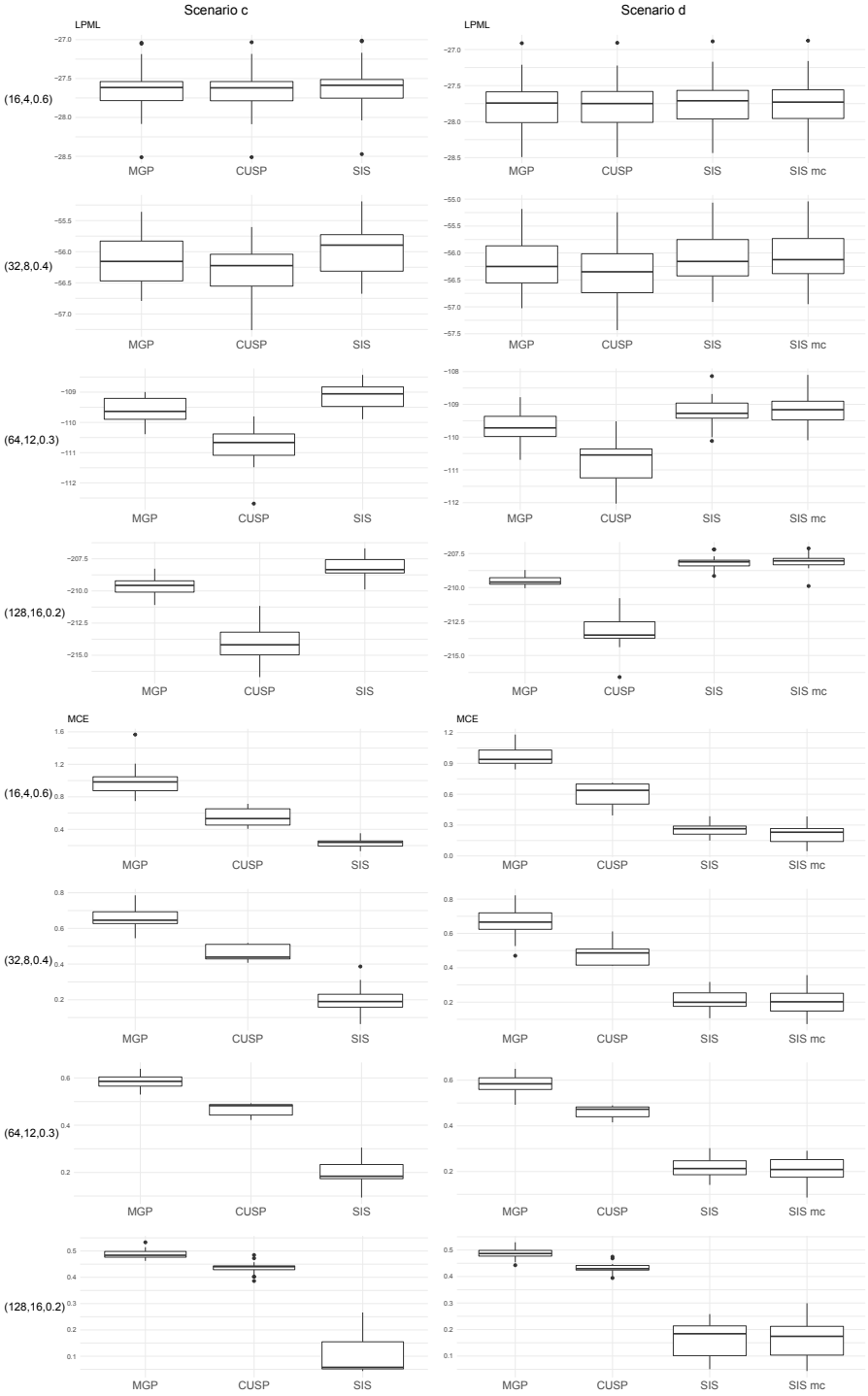

FIGURE 3.6: Boxplots of the LPML and MCE of each model for all combinations of (p, k, s) in Scenario c (left panel) and Scenario d (right panel). LPML, logarithm of the pseudo-marginal likelihood; MCE, mean classification error; CUSP, cumulative shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process; SIS mc, structured increasing shrinkage process with meta-covariates.

V Applications to real data

V.1 *Regularized regression for tracking data of football actions*

We illustrate our approach by implementing a factor based regularization of a linear regression model for football tracking data. We are interested in modelling the dangerousness y_i^R of the action i through a regression on a p-variate vector y_i^C including p key performance indicators of the action obtained by aggregating the tracking data of all the players. We consider a dataset composed by $n = 125$ independent actions of three matches of a professional European league. Due to data confidentiality agreements, the team names are not reported. The covariate matrix, opportunely standardized, include $p = 21$ indicators such as the length and width of the teams during the action, the distance run, the number of players involved and other spatial, temporal and tactical metrics. The response variable is the dangerous index computed by MathAndSport as a weighted estimate of the maximum probability to score during the action, assuming continuous values in $(0, 1)$. We expect redundant information in y^C, then we reduce covariates dimensionality by considering the Gaussian linear factor model for the $n \times p$ covariate matrix y^C

$$y_i^C = \Lambda\eta_i + \epsilon_i, \quad \epsilon_i^C \sim N(0, \Sigma^C), \quad i = 1, \dots, n,$$

where Λ is characterized by $k^* \ll p$ nonneglectable columns. The linear regression model for the n-variate response vector y^R on the latent covariates is

$$y_i^R = \beta_0 + \eta_i^\top \tilde{\beta} + \epsilon_i^R, \quad \epsilon_i^R \sim N(0, \sigma_R^2), \quad i = 1, \dots, n, \tag{3.7}$$

where β_0 is the intercept coefficient and β is the k^*-variate coefficient vector. Let β denote a p-variate coefficient vector such that $\beta = \Lambda\tilde{\beta}$, the model above can be re-written as

$$y_i^R = \beta_0 + y_i^{C\top}\beta + \nu_i, \quad \nu_i \sim N(0, \sigma_R^2 + \beta^T\Sigma^C\beta) \quad i = 1, \dots, n. \tag{3.8}$$

Model (3.7) is equivalent to model (3.8) where regularization on β is applied through k^* linear constraints determined by the columns of Λ. A sparsity pattern on Λ implies that the linear constraints act only on subsets, possibly overlapped, of elements of β and can be seen as a group penalty. Therefore, we induce the group penalty by factorizing the covariate matrix and applying the structured increasing shrinkage prior on the loadings Λ. Assuming Gaussian independent

latent factors $\eta_i \sim N(0, I_k)$ $i = 1, \ldots, p$, we can write

$$y_i = (0_p^\top, \beta_0)^\top + z_i \quad z_i \sim N(0, [\Lambda^\top, \beta]^\top [\Lambda^\top, \beta] + \Sigma)$$

where $y_i = (y_i^{C\top}, y_i^R)^\top$, $\Sigma = I_{p+1}(\sigma_1^2, \ldots . \sigma_p^2, \sigma_R^2)^\top$, and $0_p^\top$ is the p-variate null row vector. This specification presents several benefits. Firstly, the k^* latent factors can be interpreted as latent covariates that summarize the information of the observed covariates. In our case, this means the construction of a new set of k^* more informative performance indicators of the action. Secondly, the constraints definition is flexible, allowing to adaptively choose the number of linear constraints as well as the weights of such constraints without imposing any fixed structure. Furthermore, we can include the meta-covariate $p \times m$ matrix w informing on two indicator characteristics to help with the identification of sparsity pattern on Λ, implying a coherent group penalty on β. The first meta-covariate indicates if each key performance indicator is referred to the attacking, to the defending team or to both, while the second one classifies the indicators according to the type of measure they consider: spatial measurements, physical performances, or possession choices. This leads to $m = 5$. The prior specification is then completed assuming usual conjugate priors in regression models, namely $\sigma_R^{-2} \sim \text{Ga}(a_\sigma, b_\sigma^R)$, $\sigma_j^2 \sim \text{Ga}(a_\sigma, b_\sigma)$ $(j = 1, \ldots, p)$, and $\beta_h \sim N(0, \sigma_\beta^2)$ $(h = 0, 1, \ldots, \infty)$.

Consistent with simulation studies of Section IV, we fix the hyperparameters $a_\sigma = 1, b_\sigma = 0.3$, $\sigma_\gamma^2 = 1$, and $a_\theta = b_\theta = 2$. We set $b_\sigma^R = 2, \sigma_\beta^2 = 1$ and $\alpha = 4$. Then, we run the algorithm for 15000 iterations after a burnin of 10000 iterations and we thin the Markov Chain, discarding all but every third sampled parameters. We verified satisfying convergence and low autocorrelation in the sampled parameters. To evaluate the model in terms of predictions, we consider y^R as missing values in a subset of $n_v = 25$ randomly sampled actions. Their root mean squared error (RMSE) with respect to the predictive posterior mean is 0.1637.

In addition to the possible advantages in terms of prediction provided by regularized regression, the proposed structured prior helps with interpretation of the relations among the large set of key performance indicators and the response variable. The estimate of k^* strongly suggests six main factors, whose impact can be illustrated by the estimates of $[\Lambda^\top, \beta]^\top$ and meta-covariate coefficients Γ reported in Fig. 3.7. The loadings matrix is quite sparse, indicating that each latent factor impacts a small group of key performance indicators. Lower elements of $\Gamma^{(t^*)}$, represented with light cells on the right panel, induce higher shrinkage on the group of indicators described by the corresponding meta-covariate. The indicators influenced by the first factor are fairly homogeneous,

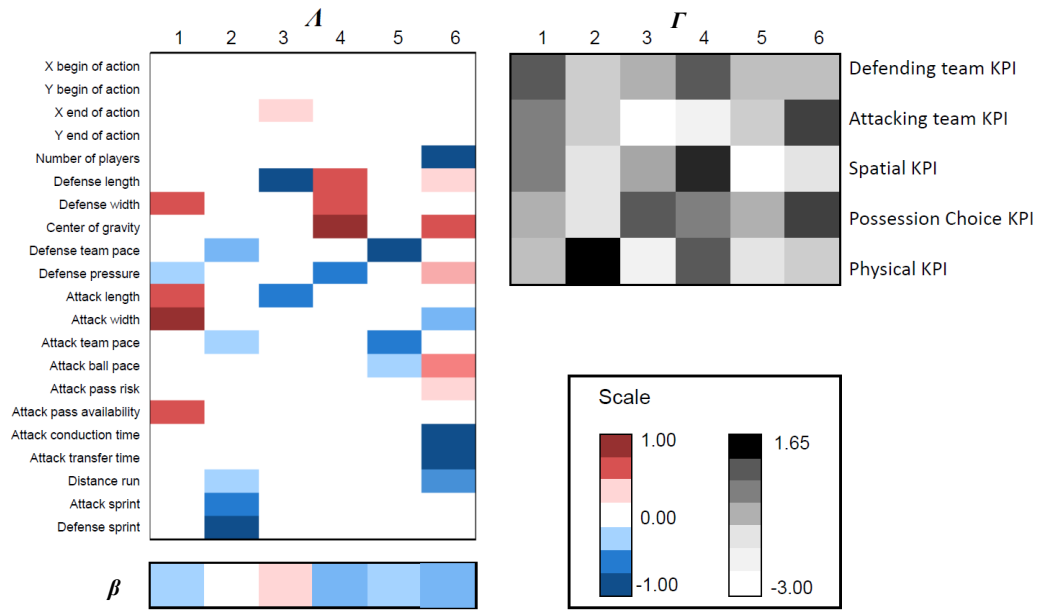

FIGURE 3.7: Posterior summaries $(\Lambda^\top, \beta)^{(t^*)}$ and $\Gamma^{(t^*)}$ of the structured regression model for football actions, where the rows of the left matrix refer to the 21 action indicators considered, while the rows of the right matrix refer to the five meta-covariates. Light coloured cells of $\Gamma^{(t^*)}$ induce shrinkage on corresponding cells of $\Lambda^{(t^*)}$.

all related to the amount of spaces between defenders. The more spaces there are during the action, the more is likely that the action has not entailed dangerous situations, which are usually characterized by the collapsing toward the box of the defence. The high level of $\gamma_{jh}^{(t^*)}$ with $j = 5$ and $h = 2$ suggests that indicators measuring physical performance tend to have loadings different from zero for the second factor and an overall very low influence on the dangerousness of the action. This fact suggests that increasing physical capacity of the players can impact the probability to score only when they produce differences and advantages in strategic and technical aspects. Focusing on the most important factors in terms of explaining dangerousness, we note that high levels of both the fourth and the sixth factors decrease the probability of scoring, although they represent distinct aspects of the action as we notice looking at the fourth and sixth columns of Γ. The indicators influenced by the fourth factor are mostly related to the defenders attitude during the action: as expected, we observe low, narrow, and high defensive pressure when the ball is in dangerous areas and close to the goal. The last factor describes the attacking strategy providing the most interesting insights. Long actions that involves a lot of players well distributed along the

width of the attacking pitch are generally more dangerous than other actions. Surprisingly, the loadings $\lambda_{jh}^{(t^*)}$ with $j = 14$ and $j = 15$ and $h = 6$ indicate that risky and fast passes are generally not worth, since they are not rewarded in terms of scoring probability.

V.11 *Effectiveness beyond football: bird species co-occurence*

As already mentioned, the generality and the effectiveness of the methodology are expected to have a dramatic impact also in other scenarios in terms of both application field and specification of the model. In particular, we illustrate our proposed method by modelling co-occurrence of the fifty most common bird species in Finland (Lindström et al., 2015), focusing on data in 2014. In ecology applications, it is quite common to have data on species traits, which are routinely used as meta-covariates informing on the different species effects of some environmental covariates. In our case, an $n \times c$ environmental covariate matrix x is available, including a five-level habitat type, `spring temperature' (mean temperature in April and May), and the square of `spring temperature', leading to $c = 7$. We also have a meta-covariate $p \times m$ matrix w of species traits: logarithm of typical body mass, migratory strategy, which is classified into short-distance migrant, resident species or long-distance migrant, and a seven-level superfamily index. Response y is an $n \times p$ binary matrix denoting occurrence of $p = 50$ species in $n = 137$ sampling areas. We model species presence or absence using a multivariate probit regression model:

$$y_{ij} = \mathbb{1}(\tilde{z}_{ij} > 0), \quad \tilde{z}_{ij} = x_i^\top \beta_j + z_{ij}, \quad z_i = (z_{i1}, \ldots, z_{ip})^T \sim N_p(0, \Lambda\Lambda^\top + I_p), \tag{3.9}$$

where β_j is the jth column of the coefficient matrix B characterizing impact of environmental covariates on species occurrence probabilities. Covariance of the latent $\tilde{z}_i$ vector is characterized through a linear Gaussian factor model. To borrow information across species while incorporating species traits, we let

$$\beta_j \sim N_c(b\, w_j, \sigma_\beta^2 I_c), \quad b = (b_1, \ldots, b_m), \quad b_l \sim N_c(0, \sigma_b^2 I_c), \tag{3.10}$$

where b is a $c \times m$ coefficient matrix with column vectors b_l given Gaussian priors.

Model (3.9)--(3.10) is consistent with popular joint species distribution models (Ovaskainen et al., 2016; Tikhonov et al., 2017; Ovaskainen & Abrego, 2020), with current standard practice using a multiplicative gamma process for Λ. We compare this approach to an analysis that instead

uses our proposed structured increasing shrinkage prior to allow the species traits w to impact Λ and hence the covariance structure across species. After standardizing x and w, we set $\alpha = 4$, $a_\theta = b_\theta = 2$ and $\sigma_\beta = \sigma_b = 1$. Posterior sampling is straightforward via the Gibbs sampler reported below, where the notation $(x \mid -)$ denote the full conditional distribution of x conditionally on everything else.

STEP 1 Update in parallel the p vectors β_j $(j = 1, \ldots, p)$ by sampling from the independent full conditional posterior distributions

$$(\beta_j \mid -) \sim N_c\big[(\sigma_\beta^{-2} I_c + x^{\mathrm{T}} x)^{-1}\{x^{\mathrm{T}}(\tilde{z}_j^{\mathrm{T}} - H\lambda_j) + b w_j\}, (\sigma_\beta^{-2} I_c + x^{\mathrm{T}} x)^{-1}\big],$$

where $\tilde{z}_j$ is the jth row vector of $\tilde{z}$.

STEP 2 Update b_l $(l = 1, \ldots, c)$ sampling from conditionally independent posteriors

$$(b_l \mid -) \sim N_m\big\{(\sigma_b^{-2} I_m + \sigma_\beta^{-2} w^{\mathrm{T}} w)^{-1} \sigma_\beta^{-2} (w^{\mathrm{T}} \beta_l^{\mathrm{T}}),\ (\sigma_b^{-2} I_m + \sigma_\beta^{-2} w^{\mathrm{T}} w)^{-1}\big\},$$

where β_l is the lth row vector of the coefficient matrix B.

STEP 3 Update the elements $\tilde{z}_{ij}$ $(i = 1, \ldots, n;\ j = 1 \ldots, p)$ sampling independently and in parallel from the truncated normal

$$(\tilde{z}_{ij} \mid -) \sim TN(\lambda_j^{\mathrm{T}} \eta_i + x_i^T \beta_j, 1, l_{ij}, u_{ij}),$$

where the lower bound l_{ij} is equal to 0 if $y_{ij} = 1$ and $-\infty$ otherwise and the upper bound is $u_{ij} = 0$ if $y_{ij} = 0$ and ∞ otherwise. Then, we set $z = \tilde{z} - xB$.

STEP 4 Given the sampled value of z, we follow steps 2--6 of the algorithm in Section III.ii to sample H,Λ,Φ, and θ where Σ is replaced by I_p.

We run the algorithm for 40000 iterations discarding the first 20000 iterations. Then, we thin the Markov Chain, saving every 5-th sample. We adapt the number of active factors at iteration t with probability $p(t) = \exp(-1 - 2.5\, 10^{(-4)} t)$ and, given the high value of p considered, we choose the offset constant $c_p = 2e \log(p)/p$.

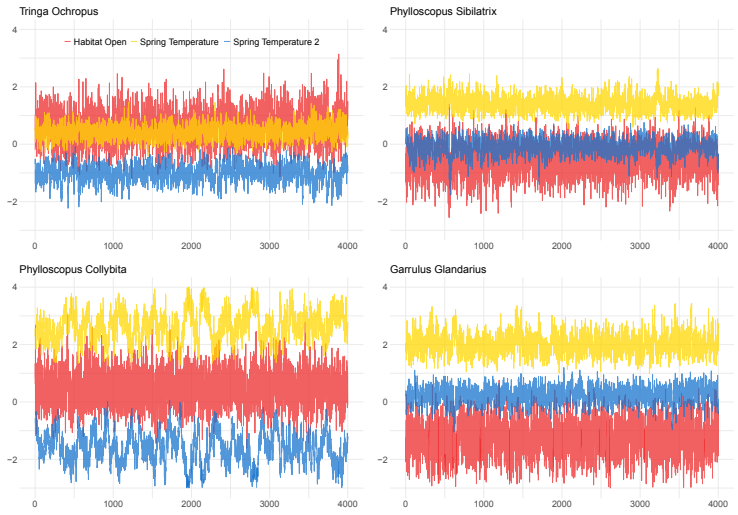

FIGURE 3.8: Chain plots of the marginal posterior samples of 12 mean coefficients of the matrix B obtained by the Gibbs sampler, discarding the first 20000 iterations and saving every 5-th sample.

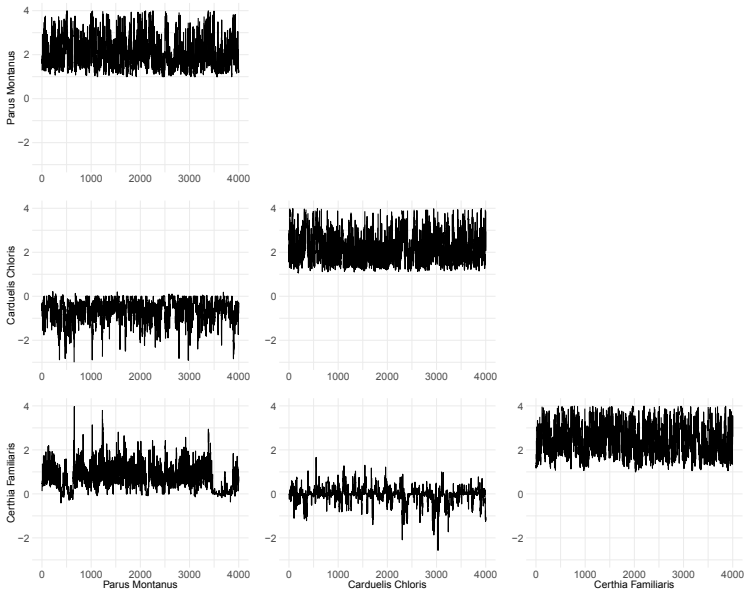

FIGURE 3.9: Chain plots of the marginal posterior samples of six elements of the covariance matrix obtained by the Gibbs sampler, discarding the first 20000 iterations and saving every 5-th sample.

Figures 3.8--3.9 report the trace plots of the posterior samples for some parameters of the model, which show a discrete mixing.

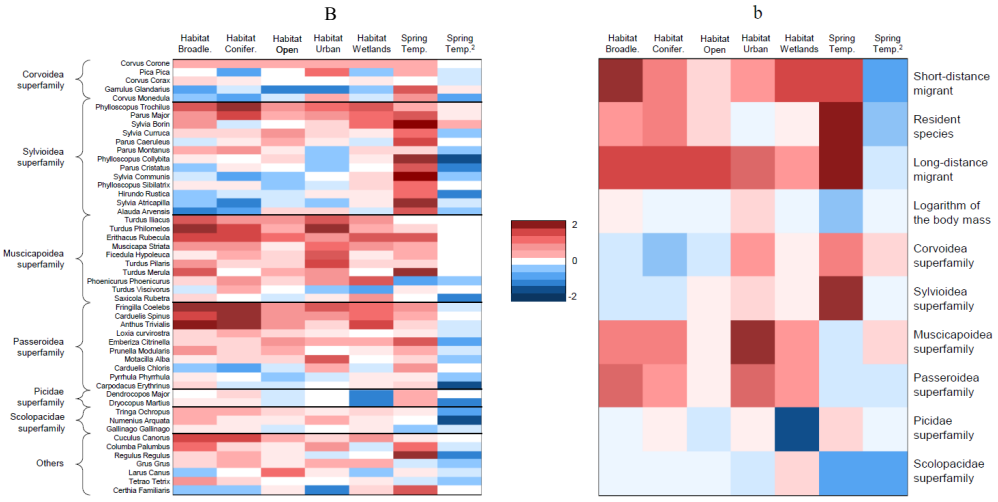

FIGURE 3.10: Posterior mean of B and b for the structured increasing shrinkage model, where the rows of the left matrix refer to the 50 birds species considered, while the rows of the right matrix refer to the ten species traits considered. Broadle: broadleaved forests; Conifer: coniferous forests; Temp: temperature.

Figure 3.10 displays the posterior means of B and b. A first investigation of the left panel shows large heterogeneity of the habitat-type effects across different species. For instance, superfamilies Corvoidea and Sylvoidea are rarely observed in forest habitats. This is also reflected in the negative posterior mean of the corresponding coefficients $b_{41}, b_{51}, b_{42}, b_{52}$ in the right panel of Fig. 3.10. This panel also shows that covariate effects tend to not depend on migratory strategy or body mass, with the exception of urban habitats tending to have more migratory birds.

The estimated Λ and meta-covariate coefficients Γ, following the guidelines of Section III.iii, are displayed in Fig. 3.11. The loadings matrix is quite sparse, indicating that each latent factor impacts a small group of species. Positive sign of the loadings means that high levels of the corresponding factors increase the probability of observing birds from those species. Lower elements of $\Gamma^{(t^*)}$, represented with light cells on the right panel, induce higher shrinkage on the corresponding group of birds. To facilitate interpretation, we rearrange the rows of $\Lambda^{(t^*)}$ according to the most relevant species traits in terms of shrinkage, which are migration strategy and body mass. The species influenced by the first factor are fairly homogeneous, characterized by short distance or resident

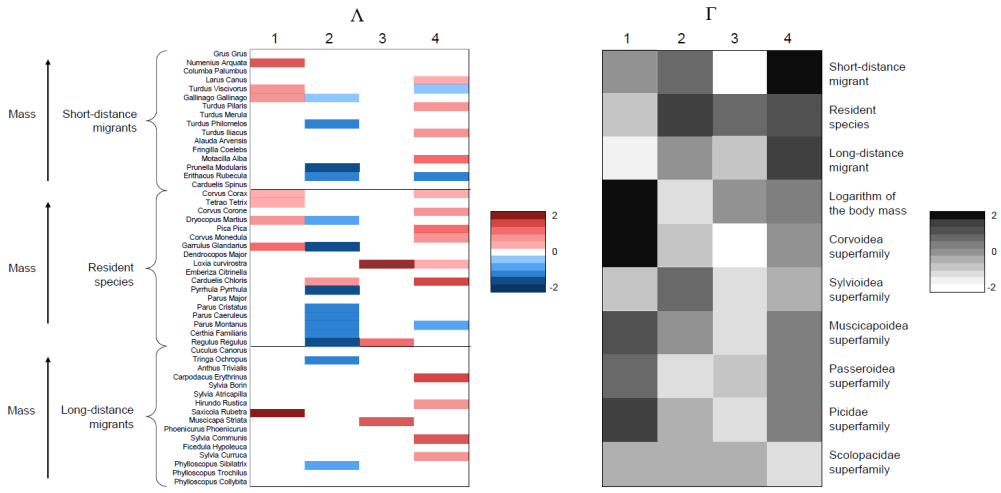

FIGURE 3.11: Posterior summaries $\Lambda^{(t^*)}$ and $\Gamma^{(t^*)}$ of the structured increasing shrinkage model; rows of left matrix refer to 50 birds species, and rows of right matrix to ten species traits. Light coloured cells of $\Gamma^{(t^*)}$ induce shrinkage on corresponding cells of $\Lambda^{(t^*)}$.

migratory strategies and larger body mass. The strongly negative value of the $\Gamma^{(t^*)}$ element $(4, 2)$ suggests heavier species of birds tend to have loadings close to zero for the second factor. This is also true for the third factor, which likewise does not impact short-distance migrants.

Figure 3.12 shows a spatial map of the sampling units coloured accordingly to the values of the first and the third latent factors. We can interpret these latent factors as unobserved environmental covariates. We find that the species traits included in our analysis only partially explain the loadings structure; this is as expected and provides motivation for the proposed approach. Sparsity in the loadings matrix helps with interpretation. Species may load on the same factor not just because they have similar traits but also because they tend to favour similar habitats for reasons not captured by the measured traits.

The induced covariance matrix $\Omega = \Lambda\Lambda^T + I_p$ across species is of particular interest. We compare estimates of Ω under the multiplicative gamma process, estimated using the R package `hmsc` (Tikhonov et al., 2020), and our proposed structured increasing shrinkage model. Figure 3.13 plots the posterior mean of the correlation matrices under the two competing models. The network graph based on the posterior mean of the partial correlation matrices, reported in Fig. 3.14, reveals several communities of species under the proposed structured increasing shrinkage prior that are not evident under the multiplicative gamma.

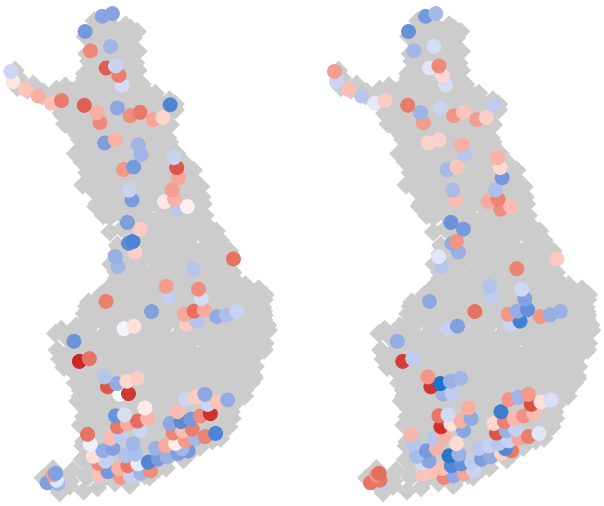

FIGURE 3.12: Maps of the sampling units in Finland coloured accordingly to the values of the first and the third latent factors sampled at iteration t^*. Red and blue spots represent the environments with positive and negative values of the factors, respectively.

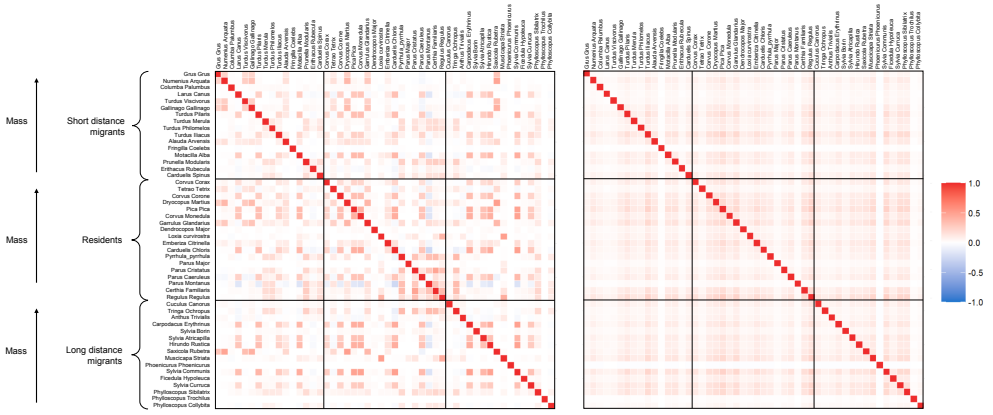

FIGURE 3.13: Posterior mean of the correlation matrices estimated by the structured increasing shrinkage model (on the left) and the multiplicative gamma process model (on the right).

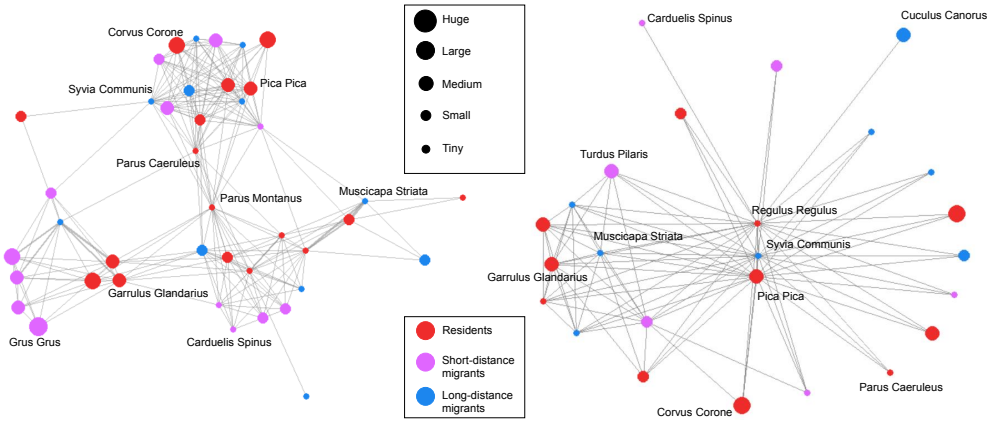

FIGURE 3.14: Graphical representation based on the inverse of the posterior mean of the correlation matrices estimated by the structured increasing shrinkage model (on the left) and the multiplicative gamma process model (on the right). Edge thicknesses are proportional to the latent partial correlations between species. Values below 0.025 are not reported. Nodes are positioned using a Fruchterman–Reingold force-direct algorithm.

We also find that the multiplicative gamma process provides a slightly worse fit to the data. The logarithm of the pseudo-marginal likelihood computed on the posterior samples of the structured increasing shrinkage model is equal to -21.06, higher than that achieved by the competing model, which is -21.36. Using four-fold cross-validation, we compared the loglikelihood evaluated in the held-out data, with μ and Ω estimated by the posterior mean in the training set. The mean of the loglikelihood was -22.62 under the structured increasing shrinkage and -23.22 under the multiplicative gamma process prior.

4 Structured matrix factorization

1 Motivation and matrix factorization notation

Despite their unquestionable charm, football player tracking data raise a noticeable number of statistical challenges that have not been considered in the football analysis framework until now. Most of them are related to the fact that those data are multivariate in several directions, needing careful and sufficiently general approaches to deal with their dependence structure. Consider, for instance, one of the most natural representations of football player tracking data: player heatmaps. A player heatmap is a graphical representation whereby the pitch areas involved in the player's action in a certain period of time are coloured depending on the intensity of the action. Figure 4.1 reports the heatmap of the metres run by a professional football player during a match when his team is in possession of the ball. Mathematically, we can describe a heatmap as a p-variate vector, where p is the number of cells in which the pitch is divided such that the jth vector component reports the intensity of the player action in the jth cell. In a football match, we have 22 player heatmaps available at the same time, generating a further dimension of the problem, and raising exponentially the complexity of analysis. More generally, we might be interested in analysing a collection of heatmaps that could be represented by a two dimensions array y, where the presence of a dependence structure cannot be excluded in any of the two dimensions.

Motivated by such a challenge, in this chapter we propose extending the Bayesian factor models to model the hidden dependence structure in a multidimensional array, imposing only weak and dimension-symmetric assumptions on the dependence. In the heatmaps example, we expect that the relation between two elements of the data matrix y_{ij} and y_{ls} depends on the similarity between the two players involved i and l, and it also depends on the spatial relation between the pitch cells j and s. Therefore, inspired by the general class of infinite factor models with structured shrinkage

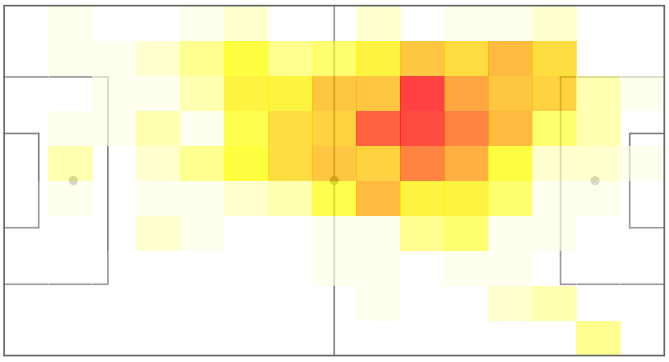

FIGURE 4.1: Illustrative heatmap representing the metres run by a professional football player during the possession time of his team in different areas of the pitch. Dark red areas are those where the player has run larger distances, white areas are those not touched by the player.

presented in the previous chapter, we introduce a new model that allows us to naturally embed exogenous information about both the subjects and the variables through local scales of latent elements depending on covariate and meta-covariate matrices x and w, when available. Beyond the Bayesian literature, successful approaches in matrix factorization (Agarwal & Chen, 2009; Rendle et al., 2011; Chen et al., 2013) included such auxiliary information along with the latent terms produced by the factorization.

We extend the notation presented in the previous chapters. Consider some transformations f and the general class of underlying Gaussian factor models with k latent factors for the $n \times p$ data matrix y. We assume that both H and Λ depend on an $n \times q$ covariates matrix x and a $p \times m$ meta-covariate matrix w, respectively, such that

$$
\begin{aligned}
y_{ij} &= f_{ij}(z_{ij}), \\
z_{ij} &= \sum_{h=1}^{k} \eta_{ih}(x_i)\lambda_{jh}(w_j) + \epsilon_{ij} \quad (i = 1, \ldots, n; j = 1, \ldots, p)
\end{aligned}
$$

where x_i and w_j are a q-variate and an m-variate vectors, respectively, and ϵ_{ij} is an independent

Gaussian error $\epsilon_{ij} \sim N(0, \sigma^2_{ij})$. We can re-write the model as

$$z = \sum_{h=1}^{k} F_h + \epsilon, \quad F_h = \eta_{\cdot h}(x)\lambda^{\top}_{\cdot h}(w).$$

Recalling the specification of Λ elements in infinite factor models reported in (2.1), we define the followig Gaussian hierarchical priors

$$\eta_{ih} \mid \psi_{ih} \sim N\{0, \psi_{ih}(x_i)\}, \qquad \lambda_{jh} \mid \theta_h, \phi_{jh} \sim N\{0, \theta_h \phi_{jh}(w_j)\},$$

where the local scale matrices Ψ and Φ depend on covariates and meta-covariates such that $E\{\psi_{ih}(x_i) \mid B\} \propto g_x(x_i\beta_h)$ and $E\{\phi_{jh}(w_j) \mid \Gamma\} \propto g_w(w_j\gamma_h)$, with β_h and γ_h columns of the coefficient matrices B and Γ, and g_x and g_w known smooth one-to-one differentiable link functions. We can re-write the model as

$$z_{ij} = \sum_{h=1}^{k} \psi^{1/2}_{ih}(x_i)\, \tilde{\eta}_{ih}\, \phi^{1/2}_{jh}(w_j)\, \tilde{\lambda}_{jh}\, \theta^{1/2}_h + \epsilon_{ij},$$

where $\eta_{ih} = \psi^{1/2}_{ih}(x_i)\tilde{\eta}_{ih}$ and $\lambda_{jh} = \theta^{1/2}_h \phi^{1/2}_{jh}(w_j)\tilde{\lambda}_{jh}$.

Considering a proper increasing shrinkage prior on θ_h $(h = 1, \ldots, k)$; we can extend the model to infinite factors $k = \infty$, similar to what is done in infinite factor models discussed in previous chapters.

The generality of such a model is expected to have impact on other several application fields. For instance, in economics and finance, the data of interest are routinely stored in a matrix of n-variate time series with p data points observed over a certain period (Arellano, 2003; Tsay, 2013). Relations among the different time series and time points can be extracted via matrix factorization algorithms (Alquier & Marie, 2019; Kastner, 2019). In particular, we could model the dependence among the n different time series (consider e.g the market returns of n financial assets, as in the work of Cappiello et al., 2006) by inducing shrinkage patterns on the latent elements through a set of covariates x; on the other hand, we could induce flexible dependence patterns over the time points by exploiting meta-covariates w that include functions of the time as trends, cycles, and seasonality.

To overcome the computational limits of the current approaches, in Section III.i, we also design a novel algorithm to allow fast factorization of huge data sets based on a forward stagewise additive procedure, which is the common ground of the *boosting* algorithms (Friedman et al., 2000; Chen

& Guestrin, 2016).

II Model specification

We specify multiplicative idiosyncratic randomness on ψ_{ih} and ϕ_{jh} $(i = 1, \dots, n;\ j = 1, \dots, p;\ h = 1, \dots, \infty)$ with respect to their mean given B and Γ matrices, such that we can write

$$\psi_{ih} = \tilde{\psi}_{ih} g_x(x_i \beta_h), \quad \phi_{jh} = \tilde{\phi}_{jh} g_w(w_j \gamma_h),$$

different from the structured increasing shrinkage prior presented in the previous chapter, where local scales given γ coefficients cannot be factorized in random and deterministic factors. Recalling the previous notation, the model for the ij element of the matrix z is

$$z_{ij} = \sum_{h=1}^{\infty} g_x(x_i^\top \beta_h) \tilde{\psi}_{ih}^{1/2} \, \tilde{\eta}_{ih} \, g_w(w_j^\top \gamma_h) \tilde{\phi}_{jh}^{1/2} \, \tilde{\lambda}_{jh} \, \theta_h^{1/2} + \epsilon_{ij},$$

for $i = 1, \dots, n$ and $j = 1, \dots, p$.

The square root of the factor scale θ_h multiplies all the parameters referring to the factor h, such that the prior on θ_h plays a key role in regulating the importance of the contribution of the factor h to the model. The specification of an increasing shrinkage prior also makes θ_h crucial in determining the truncation level of the model. Indeed, if θ_h decreases over the factors $h = 1, \dots, \infty$, then we would observe decreasing factor contributions up to a neglectable level. Nevertheless, although a steep decrease of factor contributions would certainly keep the number of relevant factors low, it would lead to difficult model interpretation, due to the large difference between the importance of first and last factors. More generally, it is preferable to separate the parameters that control the rate of shrinkage of redundant factors from those regulating the contribution magnitude of the nonneglectable factors. Consistent with such consideration, we define the factor scale again by exploiting Legramanti et al. (2020) and the structured increasing shrinkage prior introduced in Chapter 3 to manage the increase of the truncation probability while maintaining similar scale among factors. In particular, we assume $\theta_h = \rho_h \vartheta_h$, where ϑ_h $(h = 1, \dots, \infty)$ are identical and independent distributed random variables, while ρ_h is a Bernoulli random variable $\text{Ber}(1 - \pi_h)$

with increasing probability π_h of being zero over h according to the stick breaking construction

$$\pi_h = \sum_{l=1}^{h} u_l, \quad u_l = v_l \prod_{m=1}^{l-1} (1 - v_m), \quad v_m \sim \text{Be}(1, \alpha),\ \alpha > 0.$$

The parameters ρ_h provide an implicit manner to select the truncation level as the number of nonneglectable factors, i.e., every factor h such that $\rho_h = 1$. This formulation guarantees that the prior variance of any loadings element is larger than all elements with a higher column index, being a strongly increasing shrinkage prior under 3.1 in Section 3.II.ii.

We consider the following set of priors on the other parameters:

$$\begin{aligned} &\tilde{\eta}_{ih} \sim N(0,1), \quad \tilde{\lambda}_{jh} \sim N(0,1), \\ &\tilde{\psi}_{ih} \sim \text{Ber}(c_n), \quad \tilde{\phi}_{jh} \sim \text{Ber}(c_p), \quad \vartheta_h^{-1} \sim \text{Ga}(a_\theta, b_\theta), \end{aligned}$$

where c_n and c_p are fixed constants in $(0, 1)$. The inverse gamma prior on ϑ_h implies a power law tail distribution on θ_h inducing robustness properties on $\lambda_j h$ $(j = 1, \ldots, p)$. Indeed, as discussed in the literature on shrinkage priors in regression (Carvalho et al., 2010), it is crucial that the prior is concentrated at zero to reduce mean square error by shrinking small coefficients to zero, albeit with heavy tails, as power law tails, to avoid overshrinking the obvious and large signals. Recalling Proposition 3.2 and Theorem 3.2 in Section 3.II.ii, given the known matrix H, we can show that the impact of the prior on λ_{jh} posterior mode goes to zero when the data are sufficiently informative and support an increasingly larger maximum likelihood estimate.

Notably, this model is a generalization of Bayesian neural networks (Burden & Winkler, 2008; Gal et al., 2016), which are spreading through the literature due to their flexibility and implicit regularization induced by the prior. Indeed, in the simple case with $p = 1$, k fixed, and no meta-covariates w available, we have

$$y_i = f_i(z_i) = f_i\left\{ \sum_{h=1}^{k} g_x(x_i^\top \beta_h) \tilde{\lambda}_{1h} \tilde{\phi}_{1h} \theta_h{}^{1/2} + \epsilon_{ij} \right\}.$$

if $\tilde{\eta}_{ih}$ and $\tilde{\psi}_{ih}$ are Dirac δ distributions on 1 for $i = 1, \ldots, n$ and $h = 1, \ldots, k$. In this scenario, the coefficients in B and Λ are the weights of the two layers of the neural network, with g_x the activation function of the *neurons* in the hidden layer. In the literature (Leshno et al., 1993), activation functions are generally chosen in the rectified linear unit class of functions, or ReLu.

Consistent with these considerations, we define g_x and g_w as FReLu activation functions (Qiu et al., 2018), i.e.,

$$\text{FReLu}(x) = \max(x, 0) + \varepsilon,$$

with $\varepsilon \geq 0$ fixed. These functions are nondecreasing, nonnegative, and piecewise linear, helping the update and interpretation of the coefficients B and Γ.

Finally, the elements of β_h and γ_h $(h = 1, \ldots, \infty)$ are distributed a priori as

$$\begin{aligned} \beta_{1h} &\sim N(1 - \varepsilon, 1), \quad \beta_{dh} \sim N(0, 1) \ (d = 2, \ldots, q), \\ \gamma_{1h} &\sim N(1 - \varepsilon, 1), \quad \gamma_{lh} \sim N(0, 1) \ (l = 2, \ldots, m). \end{aligned}$$

Consistent with the literature on Bayesian factor models (e.g Arminger & Muthén, 1998), the prior elicitation is concluded considering the $n \times p$ error matrix ϵ distributed as $\text{vec}(\epsilon) \sim N_{np}(0, \Sigma)$, with Σ a $np \times np$ diagonal matrix with inverse gamma distributed diagonal elements $\sigma_{ij}^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$. If we integrate out the terms σ_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, p)$ according to the measure defined by their prior distribution, each error ϵ_{ij} is independently distributed as a Student-t distribution $t_{2a_\sigma}(0, b_\sigma/a_\sigma)$, with $2a_\sigma$ degrees of freedom, location equal to zero, and scale equal to b_σ/a_σ. However, the algorithm we present below is suitable to any log-posterior that can be effectively minorized by a quadratic function with respect to the single factor parameters $\tilde{\eta}_h$, and $\tilde{\lambda}_h$.

III Accelerated factorization via infinite latent elements

III.I *Estimation via forward stagewise additive maximization*

To estimate the model presented above, we propose XFILE, a novel algorithm to perform an Accelerated Factorization via Infinite Latent Elements. It aims to compute fast pointwise estimates of the model parameters, preventing the possibility of obtaining full Bayesian inference. Although sampling from the posterior distribution by extending the Gibbs sampler discussed in Section 3.III.ii is straightforward, the overparametrization of the model would lead to a very slow algorithm and so highly parametrized that a careful inference analysis would be actually performed only on a

small part of the set of parameters $\mathcal{P} = \{\tilde{H}, \tilde{\Psi}, \tilde{\Lambda}, \tilde{\Phi}, B, \Gamma, \vartheta\}$. Specifically, XFILE provides an approximation of the posterior mode, as is common in the machine learning literature (Marcel & Millán, 2007) and, in particular in the case of factorization models (Gao et al., 2013; Ročková & George, 2016). This estimation method is also partially justified from a theoretical perspective by the aforementioned robustness property formalized in Theorem 3.2 in Section 3.II.ii and focused on the posterior mode behaviour.

Our approach belongs to the wide class of machine learning methods that exploit the Bayesian model construction to obtain regularized estimates of the parameters by minimizing a loss function penalized by the parameter priors (see e.g. Fraley & Raftery, 2007; Kayri, 2016), with the probabilistic matrix factorization (Salakhutdinov & Mnih, 2008) providing a notable example for our purposes. In most cases, the loss function corresponds to the opposite of a loglikelihood. In other terms, these methods aim to find the set of pointwise estimates $\hat{\mathcal{P}}$ that minimizes

$$-\log\{L(y; \mathcal{P}, \Sigma, x, w)\} - \log\{f(\mathcal{P})\},$$

where $\log\{L(y; \mathcal{P}, \Sigma, x, w)\}$ is the data loglikelihood, and $\log(f(\mathcal{P}, \Sigma))$ is the logarithm of the prior density.

Focusing on the underlying Gaussian factor model for z, the loglikelihood is

$$\log\{L(z; \mathcal{P}, \Sigma, x, w)\} = \sum_{i=1,j=1}^{n,p} \left(z_{ij} - \sum_{h=1}^{k} \eta_{jh}\lambda_{ih}\right)^2 / \sigma_{ij}^2.$$

We can interpret the σ_{ij}^2 parameters as weights of the loss function contributions with respect to the prior penalisation. In particular, when we assume common variance σ^2 for every contribution i and j, σ^2 plays the same role of the usual regularization parameter that regulates the importance of the penalty function in machine learning methods, which is often estimated through grid search and out-of-sample cross validation. As we already mentioned, we propose integrating out the parameters σ_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, p)$ to avoid estimation troubles, while high flexibility of the model is maintained. Then, the integrated loglikelihood for z is the Student-t loglikelihood

$$\log\{L(z; \mathcal{P}, x, w)\} = \sum_{i=1,j=1}^{n,p} \left\{1 + \frac{1}{2a_\sigma} \frac{(z_{ij} - \sum_{h=1}^{k} \eta_{jh}\lambda_{ih})^2}{b_\sigma / a_\sigma}\right\}^{-a_\sigma - 0.5}. \tag{4.1}$$

In this framework, the hyperparameters (a_σ, b_σ) become the tuning parameters to regulate the

importance of the penalisation. High values of b_σ would increase the mode of σ^2_{ij}, entailing a higher importance of the regularization, which is reflected on higher expected variance of ϵ_{ij} on the marginal data model. On the contrary, a high value of a_σ would induce a lower mode, increasing the importance of the loss function to estimate the other parameters in $\mathcal{P}$. In the marginal likelihood, this translates to thinning the tails of the error distribution. Nevertheless, variations in the value of a_σ and b_σ have smaller consequences on the predictive ability of the model with respect to directly amending the value of Σ.

Our algorithm relies on both the forward stagewise additive interpretation of a factor model and the possibility of defining the log-prior as a sum of factor contributions when priors on factors are independent. Thus, the model is estimated by sequentially adding a new factor $F_h = \eta_h \lambda_h^\top$, with

$$(\eta_h, \lambda_h) = \underset{\{\eta,\lambda\}}{\operatorname{argmin}} \log\{L(z; \sum_{l=1}^{h-1} \eta_l \lambda_l^\top + \eta\lambda^\top, x, w)\} + \sum_{l=1}^{h-1} \log\{f(\eta_l, \lambda_l)\} + \log\{f(\eta, \lambda)\},$$

while the previous $h-1$ terms are fixed. In other words, at each iteration, we greedily add the factor that most improves the fit of our model to the data, under the constraints induced by the prior. A second order approximation with respect to F_h and evaluated at $\sum_{l=1}^{h-1} \hat{F}_l$ might be used to quickly optimize the objective function, strengthening the connection with the gradient boosting algorithms (Friedman et al., 2000).

This approach presents two main advantages with respect to the Markov chain Monte Carlo methods used previously in this thesis. First, the computations are substantially faster, even in the case of highly overparametrized models. In addition, the forward stagewise additive estimation also enable us to easily overcome the well known nonidentifiability issue characterizing the posterior sampling methods, largely discussed in previous chapters, avoiding the use of post-processing algorithms (McParland et al., 2014; Aßmann et al., 2016; Roy et al., 2019). In fact, given $\sum_{l=1}^{h-1} F_l$, both λ_h and η_h are only identifiable up to an arbitrary rotation P, such that $PP^\top = 1$. However, such a condition is satisfied only by two possible univariate matrices $P = 1$ and $P = -1$. This fact, jointly with unimodal and symmetric priors of η_h and λ_h with respect to zero, leads to only two equally high posterior modes on $(\hat{\eta}_h, \hat{\lambda}_h)$ and $(-\hat{\eta}_h, -\hat{\lambda}_h)$, whose interpretation is symmetric. Convergence of the algorithm is guaranteed when a nonnegative constraint on a single element of either η_h or λ_h is fixed.

The *boosting* approach defined above sheds new light on the interpretation and use of some parameters. For instance, the parameter ϑ_h $(h = 1, \dots, \infty)$ could be interpreted as a dynamic

learning rate of the algorithm that controls the impact of each step, as routinely done in boosting algorithms (Chen & Guestrin, 2016). Reducing the impact of each step through a low learning rate is fundamental to induce a finer search of the optimum, allowing a better fit. However, as the learning rate gets lower, many more steps are needed, making computation slower, and interpretation harder, since in our case each step corresponds to a further factor. The prior on ϑ_h guarantees sufficient flexibility to balance these two opposite aspects, but we generally suggest setting $b_\theta \leq a_\theta$ to ensure sufficient prior mass concentration in $(0, 1)$.

The algorithm stops when it is not possible to add a factor while increasing the log-posterior of the model. This condition can be simply verified for the factor h by looking at the value of ρ_h that maximizes the log-posterior, given the estimates of the other factor parameters. In other words, given $\sum_{l=1}^{h-1} F_l$ known, we add the factor h to the model if

$$\log\left\{\left(\frac{1+\alpha}{\alpha}\right)^h - 1\right\} < \sum_{i=1}^{n}\sum_{j=1}^{p}\left(l_{ij}^{(\rho_h=0)} - l_{ij}^{(\rho_h=1)}\right),$$

where $\text{pr}(\rho_h = 1) = \{\alpha/(1+\alpha)\}^h$ is the prior probability of factor h being not shrunk, $l_{ij}^{(\rho_h=0)}$ is the loglikelihood of z_{ij} under $\rho_l = 0$ $(l = h, \ldots, \infty)$, and $l_{ij}^{(\rho_h=1)}$ is the maximum of the loglikelihood of z_{ij} under $\rho_h = 1$ and $\rho_l = 0$ $(l = h+1, \ldots, \infty)$.

III.II *Coordinate ascent algorithm for the single factor estimation*

To estimate (η_h, λ_h), given the first $h-1$ factors and under $\rho_h = 1$, we rely on a coordinate ascent algorithm (see Wright, 2015, for a complete discussion on these algorithms) and hierarchical prior structure defined in Section II to exploit closed form updates for blocks of parameters. The loglikelihood of latent data z is

$$\sum_{i=1,j=1}^{n,p}(a_\sigma + 0.5)\log\left[1 + \frac{\left\{\tilde{z}_{ij} + \eta_{ih}\lambda_{jh}^\top\right\}^2}{2b_\sigma}\right],$$

where $\tilde{z}_{ij} = z_{ij} - \sum_{l=1}^{h-1} \eta_{il}\lambda_{jl}$ is known. Hence, extending the parameter hierarchical structure, we want to maximize the log-posterior

$$\sum_{i=1,j=1}^{n,p} -(a_\sigma + 0.5)\log\left[1 + \frac{\left\{\tilde{z}_{ij} - \text{FReLu}(x_i^\top \beta_h)\tilde{\psi}_{ih}^{1/2}\,\tilde{\eta}_{ih}\,\text{FReLu}(w_j^\top \gamma_h)\tilde{\phi}_{jh}^{1/2}\,\tilde{\lambda}_{jh}\,\theta_h^{1/2}\right\}^2}{2b_\sigma}\right]$$
$$+ \sum_{i=1}^{n}\log\{\text{pr}(\tilde{\eta}_{ih})\} + \sum_{j=1}^{p}\log\{\text{pr}(\tilde{\lambda})\} + \log\{\text{pr}(\beta_h)\} + \log\{\text{pr}(\gamma_h)\}$$
$$+ \sum_{i=1}^{n}\log\{\text{pr}(\tilde{\psi}_{ih})\} + \sum_{j=1}^{p}\log\{\text{pr}(\tilde{\phi}_{jh})\} + \log\{\text{pr}(\theta_h)\},$$

where $\text{pr}(\cdot)$ indicates the prior probability function.

We set $\rho_h = 1$ and, after initially sampling the other parameters from the prior, we cycle over the steps below up to a maximum number of iterations or convergence of the algorithm. In every step of iteration t, we update a block of parameters moving towards the maximum, conditional on the value of the other parameters.

STEP 1 *Parameter vector $\tilde{\eta}_h$ update.*

Set $\tilde{\psi}_{ih} = 1$ for $i = 1, \ldots, n$. Then, exploiting the minorize-maximize paradigm, update $\tilde{\eta}_h$ using the following quadratic minorant of the Student-t loglikelihood (see Wu & Lange, 2010, for a complete presentation), tangent to the current value $\tilde{\eta}_h^{(t-1)}$:

$$\sum_{i=1,j=1}^{n,p} -(a_\sigma+0.5)\left[\log\left\{1 + \frac{(\tilde{z}_{ij} - \tilde{\eta}_h^{(t-1)}\xi_{ij})^2}{2b_\sigma}\right\} + \frac{(\tilde{y}_{ij} - \tilde{\eta}_h\xi_{ij})^2 - (\tilde{z}_{ij} - \tilde{\eta}_h^{(t-1)}\xi_{ij})^2}{2b_\sigma + (\tilde{z}_{ij} - \tilde{\eta}_h^{(t-1)}\xi_{ij})^2}\right],$$

where $\xi_{ij} = \vartheta_h^{1/2}\,\text{FReLu}(x_i^\top\beta_h)\,\text{FReLu}(w_j^\top\gamma_h)\,\tilde{\lambda}_{jh}\tilde{\phi}_{jh}$. Consider $\tilde{z}_J$, the set of p_J columns of $\tilde{z}$ with index in $J_h = \{j = 1, \ldots, p : \phi_{jh} = 1\}$, and let $\bar{z}_{\tilde{\eta}} = \Xi_{\tilde{\eta}}^{-1}\text{vec}(\tilde{z}_J)$, where ξ_{ij} is the generic entry of the $np_J \times np_J$ diagonal matrix $\Xi_{\tilde{\eta}}$ if and only if j belongs to J_h.

Then, maximize

$$-(a_\sigma + 0.5)||D_{\tilde{\eta}}^{-1}(\bar{z}_{\tilde{\eta}} - \mathbb{1}_{np_J}\tilde{\eta}_h)||^2 - \frac{||\tilde{\eta}_h||^2}{2},$$

with respect to $\tilde{\eta}_h$, where $D_{\tilde{\eta}}^2$ is a $np_J \times np_J$ diagonal matrix with the ij entry equal to

$\xi_{ij}^{-2}\{2b_\sigma + (\bar{z}_{\tilde{\eta};ij} - \tilde{\eta}_h^{(t-1)})^2\}$ and $\mathbb{1}_{np_J} = I_n \otimes (1, \ldots, 1)^\top$ is a $np_J \times n$ matrix obtained as the Kronecker product between the identity matrix and a p_J-variate vector of ones. Then, update

$$\tilde{\eta}_h^{(t)} = \left\{ \mathbb{1}_{np_J}^\top D_{\tilde{\eta}}^{-2} \mathbb{1}_{np_J} + \frac{1}{2(a_\sigma + 0.5)} I_n \right\}^{-1} \mathbb{1}_{np_J}^\top D_{\tilde{\eta}}^{-2} \bar{z}_{\tilde{\eta}}.$$

Notice that $\mathbb{1}_{np_J}^\top D_{\tilde{\eta}}^{-2} \mathbb{1}_{np_J}$ is a diagonal matrix with element i equal to $\sum_{j \in J_h} D_{\tilde{\eta};ij}^{-2}$, such that a low computational effort is required to perform the inversion.

STEP 2 *Scale $\tilde{\psi}_{ih}$ update.*
For $i = 1, \ldots, n$, set $\tilde{\psi}_{ih}^{(t)} = 1$ if

$$\log\left(\frac{c_n}{1 - c_n}\right) > -(a_\sigma + 0.5) \sum_{j=1}^{p} \left[\log\{1 + \tilde{z}_{ij}^2/(2b_\sigma)\} - \log\{1 + \varepsilon_{ij}^2/(2b_\sigma)\} \right],$$
$$\text{with} \quad \varepsilon_{ij} = \tilde{z}_{ij} - \mathrm{FReLu}(x_i^\top \beta_h)\, \mathrm{FReLu}(w_j^\top \gamma_h)\, \tilde{\eta}_{ih} \tilde{\lambda}_{jh} \tilde{\phi}_{jh} \vartheta_h^{1/2}$$

and o otherwise.

STEP 3 *Vector β_h update.*
The vector β_h is updated by applying a Newton-Raphson step to maximize the minorant of the Student-t loglikelihood tangent to the current value $\tilde{\beta}_h^{(t-1)}$. At iteration t, we identify the set of indices $I_t = \{i \in \{1, \ldots, n\} : x_i^\top \beta_h^{(t-1)} > 0\}$, where the cardinality of the set is n_I, letting x_I denote the submatrix of x composed by the rows with index i belonging to I_t. Then, we define the $n_I p_J \times n_I p_J$ diagonal matrix Ξ_β, with the generic entry $\xi_{ij} = \mathrm{FReLu}(w_j^\top \gamma_h)\, \lambda_{jh} \eta_{ih}$ if i and j belong to I_t and J_h, respectively. Define the $n_I p_J$-variate vector $\bar{z}_\beta$ as $\bar{z}_\beta = \Xi_\beta^{-1} \mathrm{vec}(\tilde{z})_{IJ}$, where $\mathrm{vec}(\tilde{z})_{IJ}$ is a vector including the elements $\tilde{z}_{ij}$ for $i \in I_t$ and $j \in J_h$, and the $n_I p_J \times n_I p_J$ diagonal matrix D_β, where the ij entry of D_β^2 is equal to $\xi_{ij}^{-2}[2b_\sigma + \{\bar{z}_{\beta;ij} - \mathrm{FReLu}(x_i^\top \beta_h^{(t-1)})\}]$. Then, update $\beta_h^{(t)}$ setting

$$\beta_h^{(t)} = \left\{ x_I^\top \mathbb{1}_{n_I p_J}^\top D_\beta^{-2} \mathbb{1}_{n_I p_J} x_I + \frac{1}{2(a_\sigma + 0.5)} I_{q_x} \right\}^{-1} \left\{ x_I^\top \mathbb{1}_{n_I p_J}^\top D_\beta^{-2} \bar{z}_\beta + \frac{1}{2(a_\sigma + 0.5)} \mu_\beta \right\},$$

where $\mu_\beta = (1 - \epsilon, 0, \ldots, 0)^\top$ is the prior mean of β_h and $\mathbb{1}_{n_I p_J} = (1, \ldots, 1)^\top \otimes I_{n_I}$.

Because of the shape of the FReLu function around zero, the gradient with respect to β_h does not exist for some points of the domain. To overcome this issue, we assume $\mathrm{d}\,\mathrm{FReLu}(x_i^\top \beta_h)/\mathrm{d}\beta_h = 0$ if $\mathrm{FReLu}(x_i^\top \beta_h) = 0$, relying on the subgradient concept (Lange, 2013).

STEP 4 *Vector $\tilde{\lambda}_h$ update.*
Set $\tilde{\phi}_{jh} = 1$ for $i = 1, \ldots, n$ and define $\lambda_h^* = \tilde{\lambda}_h \vartheta_h^{1/2}$, such that the prior on $\lambda_h^* \mid \vartheta_h$ is the p-variate Gaussian $N_p(0, \vartheta_h)$. Consider $\tilde{z}_I$, the set of n_I rows of $\tilde{z}$ with index in $I_h = \{i = 1, \ldots, n : \psi_{ih} = 1\}$ and $\bar{z}_{\lambda^*} = \Xi_{\lambda^*}^{-1}\mathrm{vec}(\tilde{z}_I)$, where $\xi_{ij} = \mathrm{FReLu}(x_i^\top \beta_h)\,\mathrm{FReLu}(w_j^\top \gamma_h)\,\tilde{\eta}_{ih}\tilde{\psi}_{ih}$ is the generic entry of the $n_I p \times n_I p$ diagonal matrix Ξ_{λ^*} if and only if i belongs to I_h. At each iteration t, update λ_h^* with

$$\lambda_h^{*(t)} = \left\{ \mathbb{1}_{n_I p}^\top D_{\lambda^*}^{-2} \mathbb{1}_{n_I p} + \frac{1}{2(a_\sigma + 0.5)\vartheta_h} I_n \right\}^{-1} \mathbb{1}_{n_I p}^\top D_{\lambda^*}^{-2} \bar{z}_{\lambda^*},$$

where D_{λ^*} is the $n_I p \times n_I p$ diagonal matrix such that the ij entry of $D_{\lambda^*}^2$ is equal to $\xi_{ij}^{-2}\{2b_\sigma + (\bar{z}_{\lambda^*;ij} - \lambda_h^{*(t-1)})^2\}$. Finally, set $\tilde{\lambda_{jh}}^{(t)} = \lambda_{jh}^{*(t)}/\vartheta_h^{1/2}$.

STEP 5 *Scale $\tilde{\phi}_{jh}$ update.*
For $j = 1, \ldots, p$, set $\tilde{\phi}_{jh}^{(t)} = 1$ if

$$\log\left(\frac{c_p}{1-c_p}\right) > -(a_\sigma + 0.5)\sum_{j=1}^{p}\left[\log\{1 + \tilde{z}_{ij}^2/(2b_\sigma)\} - \log\{1 + \varepsilon_{ij}^2/(2b_\sigma)\}\right],$$

$$\text{with} \quad \varepsilon_{ij} = \tilde{z}_{ij} - \mathrm{FReLu}(x_i^\top \beta_h)\,\mathrm{FReLu}(w_j^\top \gamma_h)\,\tilde{\eta}_{ih}\tilde{\psi}_{ih}\tilde{\lambda}_{jh}\vartheta_h^{1/2}$$

and 0 otherwise.

STEP 6 *Vector γ_h update.*
At iteration t, identify the set of indices $J_t = \{j = 1, \ldots, p : w_j^\top \gamma_h^{(t-1)} > 0\}$, where the cardinality of the set is p_J, letting w_J denote the submatrix of w composed by the rows with index j belonging to J_t. Then, define the $n_I p_J \times n_I p_J$ diagonal matrix Ξ_γ, where the generic entry is $\xi_{ij} = \mathrm{FReLu}(x_i^\top \beta_h)\,\lambda_{jh}\eta_{ih}$ if i and j belong to I_h and J_t, respectively. Further define the $n_I p_J$-variate vector $\bar{z}_\gamma = \Xi_\gamma^{-1}\mathrm{vec}(\tilde{z})_{IJ}$, where $\mathrm{vec}(\tilde{z})_{IJ}$ is a vector including the elements $\tilde{z}_{ij}$ for $i \in I_h$ and $j \in J_t$, and the $n_I p_J \times n_I p_J$ diagonal matrix D_γ, with the ij entry of D_γ^2 is equal to $\xi_{ij}^{-2}[2b_\sigma + \{\bar{z}_{\gamma;ij} - \mathrm{FReLu}(w_j^\top \gamma_h^{(t-1)})\}]$.

Finally, update $\gamma_h^{(t)}$ with

$$\gamma_h^{(t)} = \left\{ w_J^\top \mathbb{1}_{p_J n_I}^\top D_\gamma^{-2} \mathbb{1}_{p_J n_I} w_J + \frac{1}{2(a_\sigma + 0.5)} I_m \right\}^{-1} \left\{ w_J^\top \mathbb{1}_{p_J n_I}^\top D_\gamma^{-2} \bar{z}_\gamma + \frac{1}{2(a_\sigma + 0.5)} \mu_\gamma \right\},$$

where μ_γ is the prior mean of γ_h and $\mathbb{1}_{p_J n_I} = (1, \ldots, 1)^\top \otimes I_{p_J}$.

STEP 7 *Scale ϑ_h update.*
The update of ϑ_h exploits the hierarchical specification of λ. In particular, given $\lambda_h^* = \tilde{\lambda}_h \vartheta_h^{1/2}$ with $\lambda_h^* \sim N_p(0, \vartheta_h)$, the full conditional distribution of ϑ_h^{-1} given the other parameters is $\text{Ga}(a_\theta + 0.5p, b_\theta + 0.5 \sum_{j=1}^p \lambda_{jh}^{*2})$. Then, the value of ϑ_h maximizing the objective function is the mode of the inverse gamma distribution, i.e.,

$$\vartheta_h^{(t)} = \frac{b_\theta + 0.5 \sum_{j=1}^p \lambda_{jh}^{*2}}{a_\theta + 0.5p + 1}.$$

The algorithm structure and Steps 1,2,4,5,7 are greedy, ensuring the algorithm ascends the objective function. Then, in order to guarantee the convergence, we suggest adjusting Steps 3 and 6, relying on the Newton approximation, as follows. Given the log-posterior $l^{(t-1)}$, we perform the update as described in the algorithm if and only if the log-posterior $l^{(t-1)}$ evaluated after the step is equal or greater than $l^{(t-1)}$; otherwise, we simply move along the gradient of a small step. We also recommend performing several random initializations of the first step of each element of $\tilde{\eta}_h$, $\tilde{\lambda}_h$, β_h, and γ_h to mitigate the risk of starting the algorithm very far from the maximum of the log-posterior, which would entail a huge number of steps to reach convergence, due to the nature of the minorize-maximize approach.

To speed up the algorithm, it is possible to adaptively reduce the number of updating parameters at each iteration. To accomplish this, Glasmachers & Dogan (2013) proposed adaptively changing the frequency of steps occurrence to promote the update of the parameters that allow a larger increase of the objective function. In our case, we can promote higher frequent updates of the elements of $\tilde{\eta}_h$ and $\tilde{\lambda}_h$ corresponding to the elements of $\tilde{\psi}_h$ and $\tilde{\phi}_h$, respectively, equal to 1 in the last iteration. This approach entails computational benefits, especially when n and p are very large and when high sparsity is expected, i.e., when the constants c_n and c_p are set close to zero.

IV Football heatmaps decomposition

IV.I *Nongaussian distance run heatmaps*

We are interested in modelling the n heatmaps generated by a set of players in some matches. In particular, we consider the heatmaps obtained by measuring the distance covered during a match in the p different areas in which the pitch is divided. By construction, we would observe several areas with zero distance covered and positive continuous values elsewhere, leading to an $n \times p$ heatmaps data matrix y with $y_{ij} \geq 0$ $(i = 1, \dots, n; j = 1, \dots, p)$. Recalling the notation in Section I, we model the data as a deterministic transformation of an underlying Gaussian model $y_{ij} = f_{ij}(z_{ij})\,(i = 1, \dots, n;\ j = 1, \dots, p)$, where $z = H(x)\Lambda(w) + \epsilon$ and independent Gaussian error term $\epsilon_{ij} \sim N(0, \sigma^2_{ij})$. In our case, we consider the transformation

$$y_{ij} = \begin{cases} z_{ij} & \text{if } z_{ij} > 0 \\ 0 & \text{if } z_{ij} \leq 0, \end{cases}$$

which is

$$y_{ij} = \begin{cases} \sum_{l=1}^{h-1} \eta_{il}\lambda_{jl} + \tilde{z}_{ij} & \text{if } \tilde{z}_{ij} > -\sum_{l=1}^{h-1} \eta_{il}\lambda_{jl} \\ 0 & \text{if } \tilde{z}_{ij} \leq -\sum_{l=1}^{h-1} \eta_{il}\lambda_{jl}, \end{cases}$$

when the first $h-1$ factors are known. We treat the parameter matrix $\tilde{z}$ as a further factor-specific parameter that has to be updated at each additive iteration h in order to maximize the joint posterior. Following the algorithm described in Section III.ii, we add a greedy step in the loop that iteratively updates the matrix $\tilde{z}_{ij}$ maximizing the posterior density function, given the other fixed factor-specific parameters. The distribution of $\tilde{z}_{ij}$, conditional to the first $h-1$ factors and to the value of η_{ih} and λ_{jh}, is a Student-t distribution $t_{2a_\sigma}(\eta_{ih}\lambda_{jh}, b_\sigma/a_\sigma)$, leading to the following step.

Step 8 We set $\tilde{z}_{ij} = y_{ij} - \sum_{l=1}^{h-1} \eta_{il}\lambda_{jl}$ if $y_{ij} > 0$, for $i = 1, \dots, n$ and $j = 1, \dots, p$. If $y_{ij} < 0$, we independently update $\tilde{z}_{ij}$, setting it equal to the value that maximizes the full conditional distribution

$$(\tilde{z}_{ij} \mid y_{ij} < 0, -) \sim Tt_{2a_\sigma}(\sum_{l=1}^{h} \eta_{il}\lambda_{jl}, b_\sigma/a_\sigma, -\infty, -\sum_{l=1}^{h-1} \eta_{il}\lambda_{jl}),$$

where Tt indicates the truncated Student-t distribution in the interval $(-\infty, -\sum_{l=1}^{h-1} \eta_{il}\lambda_{jl})$. Then, we set $\tilde{z}_{ij}^{(t)} = \sum_{l=1}^{h} \eta_{il}\lambda_{jl}$ if $\sum_{l=1}^{h} \eta_{il}\lambda_{jl} < -\sum_{l=1}^{h-1} \eta_{il}\lambda_{jl}$ and $\tilde{z}_{ij}^{(t)} = -\sum_{l=1}^{h-1} \eta_{il}\lambda_{jl}$ otherwise.

IV.II *Application and results*

We apply XFILE to a dataset y of $n = 106$ heatmaps of different players collected over five professional European league football matches. Each heatmap is described by a vector of $p = 150$ elements corresponding to the 150 cells in which we divide the pitch. Each element y_{ij} reports the distance covered by the player i within the cell j during his team's possession time in the match. Due to data confidentiality agreements, both players and teams have been anonymized.

Different from most of the existing literature, our model allows one to naturally embed exogenous information as covariates and meta-covariates, informing the sparsity structure of H and Λ and ultimately inducing a nonexchangeable block dependence structure. The $n \times q$ covariate matrix x provides information to distinguish the n different players, and in our case, it includes the expected role of each player during the match defined by three binary variables based on the line-up provided before the match. In addition, we exploit the possibility of considering a $p \times m$ meta-covariates matrix w including information on the p pitch cells to naturally promote consistent spatial dependence, without imposing any fixed structure. In particular, we consider the distance in polar coordinates from the centre of the pitch, two binary variables indicating which quadrant of the pitch the cell belongs to, and a further binary variable that is equal to one when the cell belongs to one of the two boxes and zero elsewhere.

After standardizing the data and the meta-covariates matrices, consistent with simulation experiments in 3.IV, we set $a_\sigma = 1$ and $b_\sigma = 0.3$, and $b_\theta = 2$ lower than $a_\theta = 4$, as suggested in Section III.i. The parameter α corresponds to the prior expected number of factors, while the offset constants c_n and c_p represent the probability of nonzero elements in η_h and λ_h, respectively. In view of this, we set $\alpha = 5$, $c_n = 0.1$, and $c_p = 0.2$. Parameter estimation is straightforward via the algorithm reported in Section III.ii. The algorithm stops after four iterations, indicating that any possible additional factor does not sufficiently improve the fit. The structured shrinkage induced by covariates and meta-covariates identifies groups of both players and pitch cells in each factor, allowing for an easy interpretation of the model by looking at the estimated H and Λ.

Figure 4.2 displays the estimates of the four columns of the element-wise product $\tilde{\Phi} \cdot \tilde{\Lambda}$ in the form of four heatmaps. According to our model, a suitable linear combination of such factors is

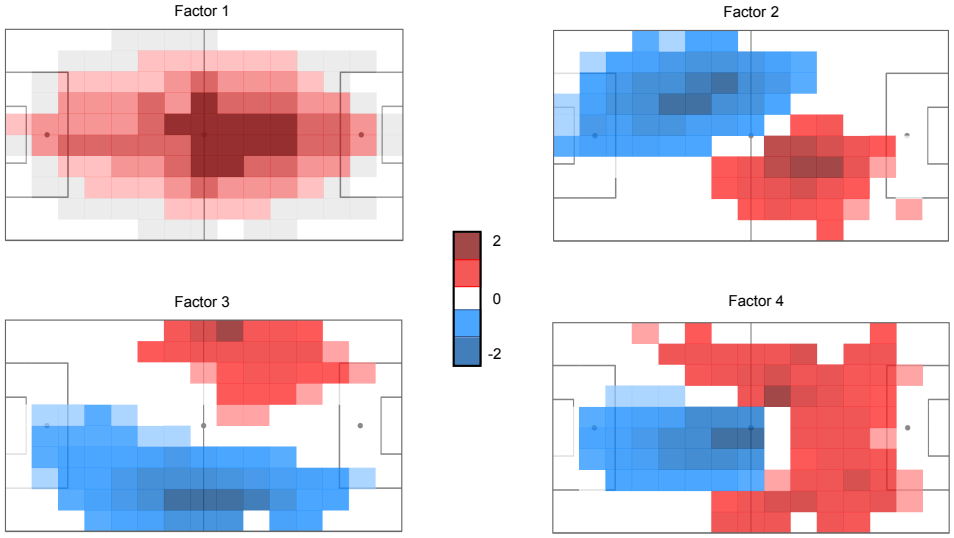

FIGURE 4.2: Heatmaps illustrating the estimates of the four columns of the element-wise product $\tilde{\Phi} \cdot \tilde{\Lambda}$. Players attack from left to right.

able to represent sufficiently well any player heatmap of the sample. The first factor in the top-left corner explains the areas of the pitch that are mostly involved in the heatmaps, acting as a sort of baseline heatmap. The top-right panel reports the second factor that helps to distinguish the players who mainly move in the right attacking area, characterized by positive values of η_2, from those that play in the left back, characterized by negative values of η_2. If $\eta_{i2} = 0$, player i does not move following one of these two patterns. Equivalent considerations can be applied to the third factor, shown in the bottom-left corner, while the fourth factor differentiates player behaviours according to a less obvious criterion. The blue cells highlight the areas of the pitch where players involved in the build-up of the action move, while the red areas characterize recurrent patterns of players who are involved forward in the action and who only provide wide pass lines during the build-up phase.

The estimate of the element-wise product $\Psi \cdot \text{sign}(H)$ is reported in Fig. 4.3, such that coloured cells in column h indicate the players influenced by the h-th factor. As previously mentioned, the first factor acts as a baseline heatmap and then affects almost all the players, with the sole exceptions being the goalkeepers (at the bottom of the figure) and a defender. Such exceptions are guaranteed by the flexible specification of our proposal. As expected, the second and third columns show clear blocks corresponding to the role of the players. This visualization can help one to immediately

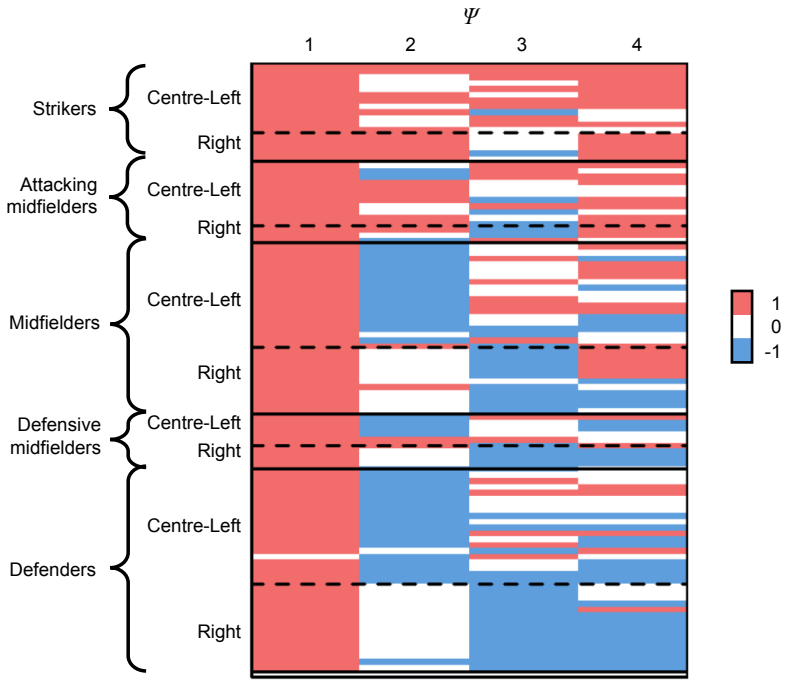

FIGURE 4.3: Estimate of the element-wise product $\tilde{\Psi} \cdot \text{sign}(H)$. The rows of the matrix refer to the 106 player heatmaps considered, and they are grouped according to the role indications provided before the match.

identify the players who played in a different role with respect to the line-up indications provided before the match. For instance, there are at least three left-side players with zero second factor and negative third factor, when the others players on the same side are generally characterized by the opposite behaviour. The influence of the last factor is heterogeneous within each role, specially among defenders and midfielders, meaning that our model is also able to identify clusters of players characterized by a similar play style, regardless of the expected role of the players. In particular, the last column of $\tilde{\Psi} \cdot \text{sign}(H)$ allows three groups to be defined according to their propensity for moving forward during attacking phases.

Each row vector of the estimated matrix H represents the play style of a single player during a match. Hence, closeness between different players can be measured by the similarity among the estimated row vectors η_i $(i = \ldots, \infty)$. In particular, we compute the Gaussian kernel similarity $-\exp\{-0.5(\eta_i - \eta_l)^{\mathrm{T}}\Theta^{-1}(\eta_i - \eta_l)\}$, where Θ is the diagonal matrix with the element h of the diagonal equal to ϑ_h. In Fig. 4.4, we report a net graph based on this similarity metric. Players

playing in similar roles tend to be clustered. Several groups of defenders are spread all over the graph, and this can be explained by the different style of play of defenders in different teams. *Player A*, *Player B*, and *Player C* lie close to centre strikers and attacking midfielders, indicating wrong expectations or representations before the match about their actual style of play. Comparing these facts with a qualitative football knowledge about these three professional players, we have gained confirmation about their ability to participate in attacking situations starting from a wider or more defensive nominal position. Instead, *Player D* is universally recognized among football insiders for his overall style of play; in fact he connects different clusters.

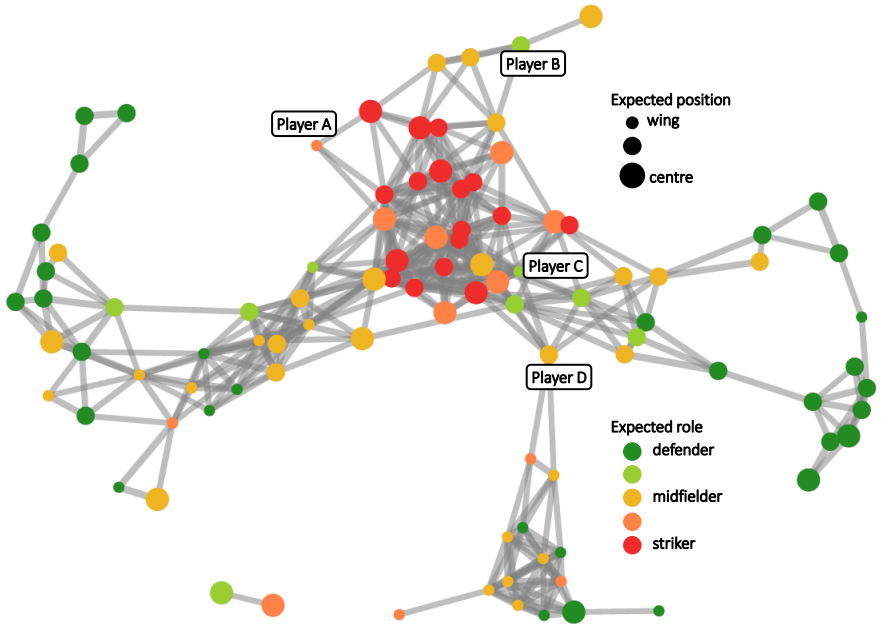

FIGURE 4.4: Graphical representation based on the Gaussian kernel similarity of estimated row vectors of H. Edge thicknesses are proportional to the similarities between players. Values below 0.5 are not reported. Nodes are positioned using a Fruchterman–Reingold force-direct algorithm.

5 Discussion

This thesis has aimed to provide an overview on infinite factorization models, presenting the state of the art, discussing the limitations of the current models, and proposing a general Bayesian infinite factorization framework including novel methods and algorithms to address such deficiencies. We have considered in particular the role of sparsity in the latent elements, to promote better inference by shrinking the noise and the redundant information and to facilitate an easier interpretation. We report below a brief discussion on the main achievements of this thesis and on the future related research topics that we think are worth exploring.

In Chapter 2, we have mainly focused on how the sparsity generated by truncating the model affects the inference performance and the factor model representation. We have noted that this effect is strongly influenced by using different truncation criteria. The novel truncation method we have proposed in Section II.i allows a more intuitive and general way to calibrate the algorithm's parameters, removing the factors that provide a negligible contribution to the global data variability. Following this idea, it is easy both to relate the threshold parameter to the interpretable model's quantities and to note the remarkable importance of using methods that are invariant to the scale of the data. Furthermore, by its construction, the algorithm guarantees a finite and deterministic upper bound on the number of factors, allowing for control of the maximum computational cost.

In Chapters 3--4, we have investigated the effects of inducing local sparsity within the factors and loadings matrices. The desired local sparsity patterns have been induced by using external auxiliary variables or exploiting unstructured prior information. The use of exogenous information to define the prior of the latent elements represents one of the more innovative contributions of the thesis. The generalized class for infinite factorization models discussed in Chapter 3 is characterized by the dependence of the loadings matrix on this additional information which is supposed to inform on the loadings sparsity structure. Theoretical support has been provided through the

definition of sparsity and robustness prior properties, while practical gains with respect to the current state of art have been demonstrated in simulation studies. The model presented in Chapter 4 embeds the auxiliary information in both the forms of covariates and meta-covariates, enabling us to model the dependence through structured sparse latent elements along both the matrix dimensions.

The available football player tracking data have represented the common thread of the thesis. They have motivated the introduction of the novel methodologies to address the challenges arising from the need of extracting valuable knowledge from such a huge amount of data, possibly framed in different shapes. The football heatmaps factorization in Section IV represents only one of the possible applications in which our methodology extracts useful insights and representations from a high dimensional dataset of a complex phenomenon. However, the generality of the framework proposed is expected to impact many other application fields, with first evidence of this fact on the improved performance in the cutting-edge ecological models discussed in Section V.ii, both in terms of variance inference and model interpretation.

However, some critical aspects need to be pointed out. The over-parametrization of the model uncovers some limitations of the current estimation methods. In generalize infinite factorization models, the adaptive Gibbs sampler generates highly autocorrelated samples, with negative impacts on the inference results. This is particularly true for the cumulative shrinkage process and structured shrinkage process that rely on a spike and slab prior on the factor scale θ_h to model the column of the loadings matrix. Indeed, as the dimension of the loadings column p increases, it becomes more difficult for θ_h to switch between the slab and spike components (Scheipl et al., 2012), inducing the draws of the number of latent factors k^* to be stuck on the same value along the Markov chain. A possible solution has recently been adopted by Kowal & Canale (2021). The authors suggested a redundant parameter expansion that introduces a further Gaussian distributed scale centred around 1 or -1 in the prior of the loadings element. Although this might lead to substantial improvements in terms of reliable posterior uncertainty quantification, Markov chain Monte Carlo methods for over-fitted factorization models are still affected by other limitations. Firstly, the algorithms could be very slow in case of big data problems. Secondly, if we are interested to interpret either the loadings or the factor, we need to overcome the problem of defining good posterior summaries, which is due to the nonidentifiability of the latent matrices. The methods adopted in this thesis to address this issue (see Section IV.ii and Section III.iii) are ad hoc solutions that cannot be taken as a complete answer for the generalized infinite factorization class of models. On the other hand, the XFILE algorithm proposed in Section III is not affected by the three mentioned

problems typical of the Markov chain Monte Carlo methods, albeit it prevents the possibility of obtaining full Bayesian inference. In addition, it still lacks of both theoretical support and comparisons with alternative algorithms. A careful investigation of the theoretical properties of the algorithm is needed, especially to explore the estimator's behaviour around the lower posterior modes. The algorithm should be largely tested on both synthetic and real data scenarios to verify the presence of practical gains with respect to the current state of the art.

Alternative scalable algorithms allowing full uncertainty quantification would be worth investigating. In this perspective, a future research path could be focused on studying the applicability of scalable algorithms that approximate the posterior distribution, as variational Bayes (Blei et al., 2017), expectation propagation (Minka, 2001), or integrated nested Laplace approximation (Rue et al., 2009). A first exploration shows that their application is not straightforward in case of over-fitted latent models as those discussed in this thesis. For instance, although it is possible in many cases to derive analytic full conditional posteriors that would encourage the use of simple coordinate ascent variational inference algorithms based on mean-field approximations (Blei et al., 2017), their application could lead to not interesting results. Indeed, the already discussed rotational symmetry of the parameter, induced by the non-identifiability of the latent structure, would be broken by the variational inference, which would approximate the symmetric modes of the posterior distribution of the latent elements with a new mode, corresponding to a degenerate solution, as a consequence of the independence assumption between H and Λ in the mean-field variational posterior distribution. The implicit regularization induced by mean-field variational approaches has been widely studied in case of Bayesian matrix factorization models (Nakajima et al., 2013) and could even represent a strength of the method when the focus of the inference is not on the low-rank latent matrices, differently from this thesis. To overcome this issue, Moore (2016) has recently proposed to model the symmetries directly in variational inference by using a symmetrized posterior as variational approximating distribution of the latent elements. Despite the promising initial results, the literature on this topic seems still immature for straightforward and immediate applications on the complex hierarchical models presented in this thesis.

Although the discussion about structured factorization is far from being completed, the encouraging results achieved in this thesis suggest that structured sparsity inducing prior should be seriously taken in consideration in future implementations of infinite factor models. Furthermore, the novel factorization algorithm inspired by gradient boosting approaches, introduced in Chapter 4, highlights the connections between the Bayesian nonparametric methods discussed in this thesis and machine learning, opening several future themes of research in this field. Bayesian nonpara-

metric modelling have already shown to be strongly effective in many machine learning contexts including variable selection (Kim et al., 2006), unsupervised learning (Broderick et al., 2013), and deep learning (Gal & Ghahramani, 2016; Polson & Sokolov, 2017). On the other hand, matrix factorization, embedding, data compression, and low-rank projections in general, are successful and widely used approaches for big data problems. For instance, in recommendation systems, the user's preference is modelled as the product of an item latent vector and a user-specific vector of latent factors. In this framework, context-aware recommendation systems (Rendle et al., 2011; Adomavicius & Tuzhilin, 2011) provide notable examples of the use of auxiliary context information in building supervised factorizations in machine learning, underline a further link between the machine learning literature and the structured priors for factorization model investigated in this thesis. The sparse pattern in the latent low-rank matrices induced by the prior might represent an appealing characteristics in big data analysis for the natural consequent regularization, which is crucial to avoid overfitting by reducing the variability in estimation, specifically if the data dimensions are large.

In this framework, we think that several directions are worth exploring. Firstly, matrix factorization in two lower-rank matrices is only one of the possible representations of matrix decomposition. It may be convenient defining a model for matrix factorization with further decomposition of the lower-rank matrices such that they can be represented as a product simpler objects. For instance, suppose we are interested in modelling the observed data collected for n subjects in p time points over a certain period of time. Either the trajectory of n players on the pitch or the evolution of n exchange rates in the forex exchange market are two possible motivating applications regarding high frequency multivariate time series. In these cases, classical approaches would suggest to induce a parametric dependence structure over time (Kastner, 2019) limiting the model flexibility, while the structured matrix decomposition proposed in Chapter 4 lacks in modelling the cycling mean behaviour of the process. An intuitive and natural alternative might be specifying a time dependence structure in terms of a function or combination of latent recurrent situations. For instance, the loadings matrix Λ may be further factorized through a $p \times s$ and an $s \times k$ matrices, such that one can characterize the jth loadings vector corresponding to time j as a linear combination of s latent states. Such states, in turn, can be represented as a linear combination of k latent factors. Application of online computational techniques, that are widely studied in machine learning, represents an additional goal. Investigations on this topic might provide interesting results in several application fields and possibly assuming different types of two-level matrix decompositions.

The second future research direction regards the generalization of the structured matrix decom-

position for tensors of order greater than 2. On this theme, tensor factorization relying on parallel factor analysis (Harshman, 1970; Bro, 1997), which expresses a rank k tensor as a sum of tensors, could represent an obvious extension of the matrix decomposition through two low rank matrices. To replicate the nonparametric framework discussed in this thesis, such extension should include the definition of a suitable notion of increasing shrinkage in high dimensions and a consistent method for the learning of the rank k^*. Both these aspects have been addressed by Dunson & Xing (2009) to model data in form of a d-dimensional contingency table. The table is modelled through an infinite sum of d-dimensional factors that are weighted by an infinite vector distributed a priori according to a Dirichlet process. This representation presents several similarities with infinite factor models. In case of location-scale family prior on the factor elements, factor weights are equivalent to the factor specific scales largely discussed in this thesis. In parallel factor analysis, the number of parameters increases linearly with the number of dimensions d, making crucial to induce a well specified sparsity structure. The approach proposed by Zhou et al. (2015) for multidimensional contingeny tables may recall the mixture structure of the local scale prior of the generalized infinite factorization models. However, to induce a strong dimensional reduction, the authors specified a unique mixture for all the entries within each dimension, which is in contrast with the main idea behind the structured factorization, where the sparsity structure within each dimension depends on the specific traits of the entries. Extension to tensor factorization should careful consider this aspect, having the necessity of accommodating both a locally induced sparsity and a low dimensional parameter space. In this perspective, it could be more promising to investigate tensor factorization with group sparse structure expected between dimensions and not within, following similar intuition to the group structure introduced in the collapsed Tucker model (see, e.g. Johndrow et al., 2017). Use cases of tensor decomposition models can be found in almost every application field. In football player tracking data, one can naturally store the data of a single match in a matrix, as mentioned above, and then store several matches along the tube of a tensor. Similar data characteristics are common in many other contexts such as neuroscience, finance, and genomics. Tensor decomposition is not in contrast with the two-level matrix decomposition research objective discussed above. In fact, a joint development could provide benefits to both directions and possible achievements would represent a major step ahead in the virgin field of Bayesian nonparametric prior for tensor factorization.

Bibliography

Abramowitz, M. & Stegun, I. A. (1948). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, vol. 55. US Government printing office.

Adomavicius, G. & Tuzhilin, A. (2011). Context-aware recommender systems. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira & P. B. Kantor, eds. Springer, pp. 217--253.

Agarwal, D. & Chen, B.-C. (2009). Regression-based latent factor models. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 19--28.

Alquier, P. & Marie, N. (2019). Matrix factorization for multivariate time series analysis. *Electron. J. Stat.* 13, 4346--4366.

An, X., Yang, Q. & Bentler, P. M. (2013). A latent factor linear mixed model for high-dimensional longitudinal data analysis. *Stat. Med.* 32, 4229--4239.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford university press.

Arminger, G. & Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika* 63, 271--300.

Assmann, C., Boysen-Hogrefe, J. & Pape, M. (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *J. Econom.* 192, 190--206.

Barros, R. M., Misuta, M. S., Menezes, R. P., Figueroa, P. J., Moura, F. A., Cunha, S. A., Anido, R. & Leite, N. J. (2007). Analysis of the distances covered by first division Brazilian soccer players obtained with an automatic tracking method. *J. Sports Sci. Med.* 6, 233.

Bhattacharya, A. & Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* 98, 291--306.

Bhattacharya, A., Pati, D., Pillai, N. S. & Dunson, D. B. (2015). Dirichlet--Laplace priors for optimal shrinkage. *J. Am. Statist. Assoc.* 110, 1479--1490.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *J. Am. Statist. Assoc.* 112, 859--877.

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst.* 38, 149--171.

Broderick, T., Jordan, M. I. & Pitman, J. (2013). Cluster and feature modeling from combinatorial stochastic processes. *Stat. Sci.* 28, 289--312.

Burden, F. & Winkler, D. (2008). Bayesian regularization of neural networks. In *Artificial Neural Networks*, D. J. Livingstone, ed. Springer, pp. 23--42.

Caley, M. (2015). Premier league projections and new expected goals. `http://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/-premier-league-projections-and-new-expected-goals` (visited on 2021-12-30).

Cappiello, L., Engle, R. F. & Sheppard, K. (2006). Asymmetric dynamics in the correlations of global equity and bond returns. *J. Financ. Econom.* 4, 537--572.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. & West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Am. Statist. Assoc.* 103, 1438--1456.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97, 465--480.

Chakraborty, A., Bhattacharya, A. & Mallick, B. K. (2020). Bayesian sparse multiple regression for simultaneous rank reduction and variable selection. *Biometrika* 107, 205--221.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 785--794.

Chen, T., Li, H., Yang, Q. & Yu, Y. (2013). General functional matrix factorization using gradient boosting. In *International Conference on Machine Learning*. PMLR. pp. 436--444.

Cook, R. D., Li, B. & Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Stat. Sin.* 20, 927--960.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics* 4, 201--218.

Dunson, D. B. & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Am. Statist. Assoc.* 104, 1042--1051.

Durante, D. (2017). A note on the multiplicative gamma process. *Stat. Probab. Lett.* 122, 198--204.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* 4, 272.

Fernández, J., Bornn, L. & Cervone, D. (2019). Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. In *13th MIT Sloan Sports Analytics Conference*.

Ferrari, F. & Dunson, D. B. (2021). Bayesian factor analysis for inference on interactions. *J. Am. Statist. Assoc.* 116, 1521--1532.

Fraley, C. & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classif.* 24, 155--181.

Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Ann. Stat.* 28, 337--407.

Frühwirth-Schnatter, S. & Lopes, H. F. (2018). Sparse Bayesian factor analysis when the number of factors is unknown. *arXiv preprint arXiv:1804.04231* .

Gal, Y. & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*. PMLR. pp. 1050--1059.

Gal, Y., McAllister, R. & Rasmussen, C. E. (2016). Improving PILCO with Bayesian neural network dynamics models. In *Data-Efficient Machine Learning workshop, ICML*. p. 25.

Gao, B., Woo, W. L. & Ling, B. W.-K. (2013). Machine learning source separation using maximum a posteriori nonnegative matrix factorization. *IEEE Trans. Cybern.* 44, 1169--1179.

Gelfand, A. E. & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. R. Statist. Soc. B* 56, 501--514.

Glasmachers, T. & Dogan, U. (2013). Accelerated coordinate descent with adaptive coordinate frequencies. In *Asian Conference on Machine Learning*. PMLR. pp. 72--86.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics* 16, 1--84.

Heckman, J. J., Stixrud, J. & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J. Labor Econ.* 24, 411--482.

Jacob, L., Obozinski, G. & Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *International Conference on Machine Learning*. pp. 433--440.

Johndrow, J. E., Bhattacharya, A. & Dunson, D. B. (2017). Tensor decompositions and sparse log-linear models. *Ann. Stat.* 45, 1--38.

Jun, L. & Tao, D. (2013). Exponential Family Factors for Bayesian Factor Analysis. *IEEE Trans. Neural Netw. Learn. Syst.* 24, 964--976.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 187--200.

Kastner, G. (2019). Sparse Bayesian time-varying covariance estimation in many dimensions. *J. Econom.* 210, 98--115.

Kayri, M. (2016). Predictive abilities of Bayesian regularization and Levenberg--Marquardt algorithms in artificial neural networks: A comparative empirical study on social data. *Math. Comput. Appl.* 21, 20.

Kim, S., Tadesse, M. G. & Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* 93, 877--893.

Kowal, D. R. & Canale, A. (2021). Semiparametric functional factor models with Bayesian rank selection. *arXiv preprint arXiv:2108.02151* .

Lange, K. (2013). *Optimization*. Springer.

Legramanti, S., Durante, D. & Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika* 107, 745--752.

Leshno, M., Lin, V. Y., Pinkus, A. & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* 6, 861--867.

Lindström, Å., Green, M., Husby, M., Kålås, J. A. & Lehikoinen, A. (2015). Large-scale monitoring of waders on their boreal and Arctic breeding grounds in northern Europe. *Ardea* 103, 3--15.

Liu, J., Tong, X., Li, W., Wang, T., Zhang, Y. & Wang, H. (2009). Automatic player detection, labeling and tracking in broadcast soccer video. *Pattern Recognit. Lett.* 30, 103--113.

Liu, Z. & Vandenberghe, L. (2010). Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.* 31, 1235--1256.

Lopes, H. F. & West, M. (2004). Bayesian model assessment in factor analysis. *Stat. Sin.* 14, 41--67.

Mackay, N. (2017). How accurate are xG models II: The 'Big Chance' dilemma. `http://mackayanalytics.nl/2017/06/19/how-accurate-are-xg-models-ii-the-big-chance-dilemma` (visited on 2021-12-30).

Marcel, S. & Millán, J. d. R. (2007). Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation. *IEEE PAMI* 29, 743--752.

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C. W. & Collinson, M. A. (2014). Clustering south African households based on their asset status using latent variable models. *Ann. Appl. Stat.* 8, 747.

Miller, J. E., Li, D., LaForgia, M. & Harrison, S. (2019). Functional diversity is a passenger but not driver of drought-related plant diversity losses in annual grasslands. *J. Ecol.* 107, 2033--2039.

Miller, J. W. & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *J. Am. Statist. Assoc.* 113, 340--356.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, Massachusetts Institute of Technology.

Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Am. Statist. Assoc.* 83, 1023--1036.

Montagna, S., Tokdar, S. T., Neelon, B. & Dunson, D. B. (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics* 68, 1064--1073.

Moore, D. A. (2016). Symmetrized variational inference. In *NIPS Workshop on Advances in Approximate Bayesian Inferece*, vol. 4. p. 31.

Murray, J. S., Dunson, D. B., Carin, L. & Lucas, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *J. Am. Statist. Assoc.* 108, 656--665.

Nakajima, S., Sugiyama, M., Babacan, S. D. & Tomioka, R. (2013). Global analytic solution of fully-observed variational Bayesian matrix factorization. *J. Mach. Learn. Res.* 14, 1--37.

Ovaskainen, O. & Abrego, N. (2020). *Joint Species Distribution Modelling: With Applications in R*. Cambridge University Press.

Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* 7, 549--555.

Pollard, R. & Reep, C. (1997). Measuring the effectiveness of playing strategies at soccer. *J. R. Statist. Soc. D* 46, 541--550.

Polson, N. G. & Scott, J. G. (2010). Shrink globally, act locally: Bayesian sparsity and regularization. *Bayesian Statistics* 9, 1--16.

Polson, N. G., Scott, J. G. & Windle, J. (2013). Bayesian inference for logistic models using pólya--gamma latent variables. *J. Am. Statist. Assoc.* 108, 1339--1349.

Polson, N. G. & Sokolov, V. (2017). Deep learning: A Bayesian perspective. *Bayesian Anal.* 12, 1275--1304.

Poworoznek, E., Ferrari, F. & Dunson, D. (2021). Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching. *arXiv preprint arXiv:2107.13783* .

Qiu, S., Xu, X. & Cai, B. (2018). FReLU: Flexible rectified linear units for improving convolutional neural networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE. pp. 1223--1228.

Reich, B. J. & Bandyopadhyay, D. (2010). A latent factor model for spatial data with informative missingness. *Ann. Appl. Stat.* 4, 439.

Rendle, S., Gantner, Z., Freudenthaler, C. & Schmidt-Thieme, L. (2011). Fast context-aware reccomendations with factorization machines. In *Proceedings of the 34th ACM SIGIR Conference on Research and Developement in Information Retrieval*. pp. 635--644.

Roberts, G. O. & Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Prob.* 44, 458--475.

Ročková, V. & George, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Am. Statist. Assoc.* 111, 1608--1622.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Statist. Soc. B* 73, 689--710.

Roweis, S. & Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neural Comput.* 11, 305--345.

Roy, A., Schaich-Borg, J. & Dunson, D. B. (2019). Bayesian time-aligned factor analysis of paired multivariate time series. *arXiv preprint arXiv:1904.12103* .

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc. B* 71, 319--392.

Salakhutdinov, R. R. & Mnih, A. (2008). Probabilistic matrix factorization. In *Proc. Adv. Neural Inf. Process. Syst. (NIPS 07)*. ACM Press. pp. 1257--1264.

Scheipl, F., Fahrmeir, L. & Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *J. Am. Statist. Assoc.* 107, 1518--1532.

Schiavon, L. & Canale, A. (2020). On the truncation criteria in infinite factor models. *Stat* 9, e298.

Schiavon, L. & Canale, A. (2021). Bayesian regularized regression of football tracking data through structured factor models. In *Book of Short Papers SIS 2021*, C. Perna, N. Salvati & F. Schirippa Spagnolo, eds.

Schiavon, L., Canale, A. & Dunson, D. B. (in press). Generalized infinite factorization models. *Biometrika* .

Thomas, D. C., Conti, D. V., Baurley, J., Nijhout, F., Reed, M. & Ulrich, C. M. (2009). Use of pathway information in molecular epidemiology. *Hum. Genomics* 4, 21.

Tikhonov, G., Abrego, N., Dunson, D. & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods Ecol. Evol.* 8, 443--452.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J. & Ovaskainen, O. (2020). Joint species distribution modelling with the R-package hmsc. *Methods Ecol. Evol.* 11, 442--447.

Tsay, R. S. (2013). *Multivariate Time Series Analysis: With R and Financial Applications*. John Wiley & Sons.

Wade, S., Ghahramani, Z. et al. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* 13, 559--626.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics* 7, 733--742.

Wright, S. J. (2015). Coordinate descent algorithms. *Math. Program.* 151, 3--34.

Wu, T. T. & Lange, K. (2010). The mm alternative to em. *Stat. Sci.* 25, 492--505.

Yang, L., Fang, J., Duan, H., Li, H. & Zeng, B. (2018). Fast low-rank Bayesian matrix completion with hierarchical Gaussian prior models. *IEEE Trans. Signal Process.* 66, 2804--2817.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* 68, 49--67.

Zhou, J., Bhattacharya, A., Herring, A. H. & Dunson, D. B. (2015). Bayesian factorizations of big sparse tensors. *J. Am. Statist. Assoc.* 110, 1562--1576.

Lorenzo Schiavon

Curriculum Vitae

Contact Information

Address	Department of Statistical Sciences(Università degli Studi di Padova) via Cesare Battisti, 241 35121 Padova, Italy
Mail	lorenzo.schiavon@phd.unipd.it
Website	lorenzo-schiavon.github.io

Current position

Apr 2022	*PhD Student* in Course, Università degli Studi di Padova
Oct 2018	Thesis Title: *Bayesian infinite factorization methods with applications to tracking data in football* Supervisor: *Prof.* Antonio Canale

Education

Sep 2018	*MSc in Statistical Sciences*	110/110 with honors
	Dissertation title: *Bias reduction in a fixed effects model for Expected Goals* Supervisor: *Prof.* Nicola Sartori	
Apr 2016	*MSc in Statistics, Economy and Finance*	110/110 with honors
	Dissertation title: *Human capital and economic growth: the role of heterogeneity among groups of countries* Supervisor: *Prof.* Stefano Galavotti	

Other educational experiences

Ongoing	*MBA fellow* at Collège des Ingénieurs Italia (Turin). Program: Science & Management

Other work experiences

Oct 2019	*External consultant* at Mercurius BI srl Statistical consultancy on research, definition and implementation of statistical models to predict football match outcomes.
Aug 2016	*Intern* at Sanmarco Informatica Spa (Grisignano di Zocco) Internship as junior programmer to complete a document management system.

Teaching Activities

Sep 2021	*Co-supervisor of MSc. thesis* at Department of Statistical Sciences(Università degli Studi di Padova) Dissertation title: Bayesian infinite factor models for count data
2017-2018	*Junior Tutor* at Department of Statistical Sciences(Università degli Studi di Padova)

Computer skills

Advanced knowledge of the R statistical software and basic knowledge of MATLAB, SAS and STATA.
Good knowledge of SQL, BigQuery and Google Analytics 360 frameworks.
Basic programming skills in Python Java and C. Advanced knowledge of the LaTeXtypesetting system for papers, reports and presentation.
Basic knowledge of markup languages as `html`, `xml` and `json`.

Language skills

Italian (mother language), English (C1), French (A2)

Publications

Schiavon, L., Canale, A., Dunson, D. B. (in press) Generalized infinite factorization models, *Biometrika*. doi: 10.1093/biomet/asab056.

Padovani, A., Canale, A., Schiavon, L., Masciocchi, S., Imarisio, A., Risi, B., Bonzi, G., De Giuli, V., Di Luca, M., Ashton, N.J., Blennow, K., Zetterberg, H., Pilotto, A. (in press) Is amyloid involved in acute neuroinflammation? A CSF analysis in encephalitis, *Alzheimer's & Dementia*. doi: 10.1002/alz.12554.

Schiavon, L., Canale, A. (2021). Bayesian regularized regression of football tracking data through structured factor models. *Book of Short Papers SIS 2021* (Editors: Perna, C., Salvati, N. and Schirippa Spagnolo, F.), ISBN: 9788891927361.

Schiavon, L., Canale, A. (2020) On the truncation criteria in infinite factor models, Stat, 9 (1), e298. doi: 10.1002/sta4.298.

Schiavon, L., Sartori, N. (2019). Bias reduced estimation of a fixed effects model for Expected Goals in association football. *Book of Short Papers SIS2019* (Editors: Arbia, G., Peluso, S., Pini, A., Rivellini, G.), ISBN: 9788891915108.

Petretta, M., Schiavon, L., Diquigiovanni, J. (2019). Betting on football: a model to predict match outcomes. *Book of Short Papers SIS2019* (Editors: Arbia, G., Peluso, S., Pini, A., Rivellini, G.), ISBN: 9788891915108.

F. Bortolon, C. Castiglione, L. Parolini, L. Schiavon (2017). A Markovian approach to darts. *Proceedings of MathSport International 2017* (Editors: De Francesco, C., De Giovanni, L., Ferrante, M., Fonseca, G., Lisi, F., Pontarollo, S.), ISBN: 9788869380587.

Conference presentations

Dec 2021	Generalized infinite factorization models with an application to Finnish bird co-occurence data. Invited presentation, *CMStatistics 2021*, virtual.
Sep 2021	Generalized infinite factorization models with an application to Finnish bird co-occurence data. Contributed presentation, *RSS International Conference 2021*, virtual.
Jun 2021	Generalized infinite factorization models with an application to Finnish bird co-occurence data. Contributed presentation, *2021 World Meeting of the International Society for Bayesian Analysis*, virtual.
Jun 2021	Bayesian regularized regression of football tracking data through structured factor models. Invited presentation, *SIS 2021 Intermediate meeting*, virtual.
Nov 2019	Predictions with Expected Goals: a model for the scoring process in a football match. Contributed presentation, *AUEB Sport Analytics Workshop 2019*, Athens, Greece
Jun 2019	Bias reduced estimation of a fixed effects model for Expected Goals in association football. Contributed presentation, *SIS 2019 Intermediate meeting*, Milan, Italy.

Awards and Grants

2021	Innovation 4 Change program (CDI Italia, Polytechnic University of Turin and CERN Ideasquare)
2020	Student funding grant ASA Statistics in Sports
2018	Special mention at Oliviero Lessi award (Società Italiana di Statistica)
2017	Best Report Prize at Stats Under the Stars3