UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Natália Caroline Costa de Oliveira

New extensions of the generalized mixed spatiotemporal modeling with random effect via factor analysis.

Belo Horizonte
2023

Natália Caroline Costa de Oliveira

New extensions of the generalized mixed spatiotemporal modeling with random effect via factor analysis.

Final Version

Thesis presented to the Graduate Program in Statistics at the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Statistics.

Advisor: Vinícius Diniz Mayrink

Belo Horizonte
2023

Oliveira, Natália Caroline Costa de.

O048n New extensions of the generalized mixed spatiotemporal modeling with random effect via factor analysis [recurso eletrônico] / Natália Caroline Costa de Oliveira – 2023.
1 recurso online 61 f. il., color.) : pdf.

Orientador: Vinícius Diniz Mayrink
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 55-57

1. Estatística – Teses. 2. Análise de regressão – Teses. 3. Processos gaussianos - Teses. 4. Análise fatorial – Teses. I. Mayrink, Vinícius Diniz. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

FOLHA DE APROVAÇÃO

New extensions of the generalized mixed spatio-temporal modeling with random effect via factor analysis

NATALIA CAROLINE COSTA DE OLIVEIRA

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Mestre em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.

Aprovada em 28 de fevereiro de 2023, pela banca constituída pelos membros:

Prof. Vinícius Diniz Mayrink - Orientador
DEST/UFMG

Prof. Marcos Oliveira Prates
DEST/UFMG

Profa. Thais Cristina Oliveira da Fonseca
IM/UFRJ

Belo Horizonte, 28 de fevereiro de 2023.

Acknowledgments

First, I would like to thank my advisor Vinícius Mayrink, for all the support I had during this period of training. The tips and guidance were essential for my learning.

My eternal thanks to my family, especially my parents, Naiara and Césio, and my brothers Lucas and Tiago, who are always willing to help me.

Thanks to all the professors and staff of the Department of Statistics. In addiction, I thank my friends who were by my side in the happy and difficult days, not letting me give up in the most challenging moments.

Finally, I would like to thank the Brazilian research agency CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for granting me the scholarship allowing me to dedicate to the studies and research. I also thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) for their support to the Department of Statistics.

"*Statistics is the grammar of science.*"
(Karl Pearson)

Resumo

Modelos Lineares Mistos Generalizados (GLMM) são utilizados quando há interesse em modelar efeitos fixos e aleatórios em uma regressão com variável resposta pertencente à família exponencial. Neste trabalho um GLMM com efeito aleatório é modelado via análise fatorial com interação não linear entre os escores fatoriais. Além disso, consideramos no estudo um efeito espaço-temporal. A dependência espacial é estabelecida através das colunas da matriz de cargas da abordagem fatorial usando a modelagem CAR (modelo Autoregressivo Condicional). A dependência temporal está relacionada às linhas da matriz de escores de fatores em uma estrutura AR(1)- modelo Autoregressivo. Um objetivo de se considerar um modelo com interação não linear entre fatores latentes é possibilitar a captura de níveis complexos de associações entre regiões e fatores. Um segundo objetivo é diversificar os tipos de clusters (grupos de regiões) estabelecidos por combinações de efeitos envolvendo fatores principais e a interação. Problemas de identificação podem ocorrer na modelagem de fatores, e aqui apresentamos algumas estratégias para tornar o modelo identificável. O foco deste trabalho é estabelecer uma proposta de modelagem para variável resposta contínua, como a Normal, Gama e Beta. Um estudo simulado é realizado para explorar a performance dos modelos. Um esquema Monte Carlo com replicações é também usado na análise. Finalmente,o modelo Beta é ilustrado em uma aplicação real usando dados (do INEP) sobre indivíduos que fizeram o Exame Nacional do Ensino Médio (período $2015 - 2021$) no Estado de Minas Gerais de modo a relacionar a combinação linear das notas das provas objetivas com outras covariáveis. Um dos objetivos é encontrar clusters que evidenciam padrões espaciais e temporais das notas dos indivíduos presentes no estudo.

Palavras-chave: Regressão. Interação. MCMC. Processo Gaussiano. Modelo CAR.

Abstract

Generalized Linear Mixed Models (GLMM) are used when there is an interest in modeling fixed and random effects in regression with response belonging to the exponential family. In this work, a GLMM with a random effect is modeled via factor analysis (FA) with nonlinear interaction between the factor scores. We consider in the framework a spatiotemporal effect. The spatial dependence is established via CAR model (Conditional Autoregressive model) for the columns of the loadings matrix defined FA part of the model. The temporal dependence is related to the rows of the factor score matrix in an AR(1) model - first-order Autoregressive model. One goal of considering a model with nonlinear interactions between latent factors is to enable capturing complex levels of associations between regions and factors. Another goal is to diversify the types of clusters (groups of regions) established by combinations of effects involving the main factors and the interaction. Identification problems can occur in factor modeling, and here we present some strategies to make the model identifiable. The focus of this work is to model continuous responses, such as those from Normal, Gamma, and Beta distributions. A comprehensive simulation study is developed to verify performance and explore the models. A Monte Carlo scheme with replications is also considered in the analysis. Finally, the Beta model is illustrated in a real application using data (from INEP) about individuals taking the Brazilian national exam ENEM (period $2015-2021$) in the State of Minas Gerais in order to relate the linear combination of the scores of the objective tests with other covariates. One of the goals is to find clusters that show spatial and temporal patterns of the scores of individuals present in the study.

Keywords: Regression. Interaction. MCMC. Gaussian Process. CAR Model.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Factor Analysis (FA) is a technique to summarize a large number of variables into a smaller number of factors. The main goal is to describe the original variability of q covariates through a number $K \leq q$ of latent factors. In multivariate statistics, there are other methodologies for dimensionality reduction such as Principal Component Analysis - PCA [see 28, 17] and Linear Discriminant Analysis - LDA [see 9, 25]. In FA, the design matrix is constructed from the product of the loadings matrix and the factor scores matrix plus an error matrix. The observations in each row of the latter are usually assumed normally distributed and independent. The information provided by the variables is summarized in the rows of the factor scores matrix, and it is possible to observe existing underlying patterns of the data set. Works such as [18] and [19] describe FA with more details. See also the book [17].

Over the years, new models based on FA were developed. In the standard factor analysis, assuming independence between observations is the usual choice. Nevertheless, one of the extensions, called the dynamic factor model considers the temporal correlation between observations and the formation of patterns in the factor scores over time. The references [11] and [20] have more information about this aspect. Another extension to be reported is the incorporation of spatial correlation in the FA through the columns of the loadings matrix and temporal correlation in the rows of the factor scores matrix. Including temporal and spatial correlations enables capturing capturing the behavior of factors over time, and a spatiotemporal factor model was obtained. In [20], the authors propose this type of spatiotemporal factor modeling structure by assuming a Gaussian distribution for the response variable. An extension was created in [19] where the distribution of the response belongs to the exponential family. This extension was possible due to the development of spatial statistics [2] and more efficient MCMC algorithms [10, 21]. Another extension for the FA was developed in [23], where a nonlinear interaction matrix between the rows of the factor scores matrix is inserted in an additive way. This framework considers a model using spike and slab priors, where scores with interaction are defined from a model with a Gaussian response. In [20], [19], and [23], the factor model is applied directly to the observed data.

Generalized Linear Models (GLM) have expanded the possibilities for linear mod-

els, where the response variable can have a probability distribution belonging to the exponential family [27]. Subsequently, the Generalized Linear Mixed Model (GLMM) was developed, incorporating correlation through random effects. The model with fixed and random components is usually called "mixed model" [24]. In [27], there are more details about the development of the Normal Bayesian GLMM modeling. In addition, works such as [8] and [5] have developed Bayesian GLMM models for cases in which the response variable has a Beta and Gamma distribution, respectively.

In [7], a generalized linear model was proposed to handle a database of individuals observed in time and space. A random effect is added to the model and the structure of the factor model with interactions [23] is incorporated to fit the data. The interaction effect is inserted through a Gaussian process with a covariance function depending on the columns of the factor scores matrix. Here, the Gaussian Process structure was built such that locations are affected by the same type of interaction between the main factors. The authors assume that an area may or may not be affected by the single interaction effect with probability defined through a spike and slab mixture [14, 15]. The main motivation of the methodology is to introduce a nonlinear interaction design to capture more complex associations between factors, enabling better identification of groups (clusters) of regions influenced by such associations. The spatial model for areal data considers a fixed region partitioned into a finite number of area units with well-defined boundaries. The spatial association is introduced by assuming a neighborhood structure defined by the map, which is incorporated via the so-called Conditional Autoregressive - CAR [4] modeling. The CAR model can also be considered for the temporal structure, but here, the time points are treated as regions, and neighbors are the past and future points; see [22].

In FA, there is an important identifiability issue that must be solved. To solve this issue, some constraints be added to the a prior distribution of the parameters related to this problem when the Bayesian approach is being considered. If no changes are made, the signs of the loadings and their respective factor scores may switch, and if two or more factors are involved, the rows of λ and the columns of α may change position. Some solutions have been presented in the literature, and it involves restricting the a prior distribution of the α. In [21], α is constructed as a block lower triangular matrix with strictly positive diagonal elements. In the work of [23], to constructed the loading matrix, some extra information was considered, that allowed creating groups of constraints in the prior distribution of this matrix. This modification made it possible to associate loads with only one set of factor scores. This last approach will be used in this work in the same way that it was considered in [7].

The authors in [7] evaluate two types of response variables by assuming the Bernoulli (for binary) and the Poisson (for count) distributions. They comment on the possibility of extending the analysis with other probability distributions for the response. In line with this idea, here we propose to explore the model under the following continuous response

variables: Normal, Gamma and Beta. A comprehensive simulation study is developed to explore the performance of the proposed models in terms of inference, i.e., evaluating how well the real values of parameters can be recovered. The authors in [7] analyze the predictive behavior of the Bernoulli model, which is standard practice for this type of GLM that can be used for classification problems. We emphasize that studying predictive behavior is not the focus of the present work. The simulation study was planned to contemplate different scenarios related to the number of regions, number of factors, number of time points, and number of regions affected by interaction. The present study shines a light to understand how the proposal in [7] behaves in situations where continuous distributions are assumed for the response. We assume the Bayesian approach to develop the whole study.

In order to motivate the analysis, a real data application is developed here assuming the Beta distribution for a bounded continuous response. The data refers to the *Exame Nacional do Ensino Médio* (ENEM; years 2015 to 2021) from candidates residing in the state of Minas Gerais (Brazil). The target response is built as a linear combination obtained through the first principal component (PCA) determined with respect to the grades, measuring performance in different fields-of-knowledge, obtained by the candidates taking the multiple-choice exam. Note that the response variable is indeed bounded, and the Beta model is a reasonable choice for this case. The goal of this real application is to evaluate how selected covariates and the spatiotemporal random effect impact the mean of the response. In addition, the analysis tries to identify regions in Minas Gerais showing similar associations with patterns displayed by the main latent factors and their interaction in the model. Note that each candidate has his/her location and year related to the exam. This is not a follow-up study, therefore, each student does not have repeated measurements throughout different years. Constraint groups are needed to handle the identification issue in the FA part of the model. We use the HDI (Human Development Index) to partition the municipalities of Minas Gerais into three groups (high, low, and moderate). The real data set can be accessed on the portal of the Brazilian National Institute of Educational Studies and Research (INEP).

The central contribution of this M.Sc. dissertation is to extend the model proposed in [7] by considering important continuous distributions for the response. The chosen options are (*i*) the Normal distribution widely used in regression settings, (*ii*) the Gamma distribution for positive continuous asymmetric data, and (*iii*) the Beta distribution defined for continuous bounded data. The FA with interactions structure for random effects, as in [7], is also considered here to determine the so-called clusters of regions. In addition to proposing new versions of the model, we also explore their performance and evaluate how they behave in terms of inference. A simulation study is developed to assess the three models under different scenarios and configurations. Finally, a database related to grades of students taking the Brazilian ENEM across distinct years is analyzed. This data set

has never been explored elsewhere in the literature, meaning that the analysis presented here can also be seen as a contribution.

The outline of this work is as follows. Section 2 presents the extensions of the model proposed in [7]. The formulations and notations for the Normal, Gamma, and Beta cases are introduced. In addition, all specifications and solutions to identification issues are discussed. Furthermore, Section 2 also indicates the hierarchical structures of the models and the details related to the Bayesian inference. Section 3 shows a simulation study exploring the models. We explain how to generate artificial data and summarize the central results. Section 4 presents the results from the real data application related to the ENEM exam. In this case, the Beta model is the reasonable choice to deal with the nature of the response variable. We show key results and discuss the main conclusions of the real analysis. Finally, Section 5 summarizes the main conclusions of the M.Sc. dissertation.

Chapter 2

Model specification

A Generalized Linear Mixed Model (GLMM) can be seen as an extension of a standard Generalized Linear Model [GLM; see 27]. In both cases, the goal is to describe the relationship between the response variable Y_i and a linear predictor. In the GLMM, this linear predictor includes the usual fixed effects and random effects to capture underlying structures not explained by covariates. Let Y_i, $i = 1, \ldots, n$, be the response variable related to i-th sample unit, whose distribution belongs to the exponential family, with outcomes observed independently. The corresponding density function is

$$f(Y_i|\theta, \psi) = \exp\left\{\sum_{j=1}^{J} T_j(Y_i)b_j(\theta_i) + c(\theta_i) + h(Y_i, \psi)\right\},$$

where θ_i is the mean (or it is proportional to the mean) and ψ is a dispersion parameter. The dispersion ψ can be embedded in $b_j(\theta_i)$ and $c(\theta_i)$ for some distributions, as the Bernoulli and Poisson. The parameter ψ can be 1, depending on the distribution. In addition, $h(Y_i, \psi)$ and $T_j(Y_i)$, $j = 1, \ldots, J$, are real-valued functions independent of θ_i. The terms $b_j(\theta_i)$ and $c(\theta_i)$ are real-valued functions of θ_i that do not contain Y_i in their formulations. The support of $f(Y_i|\theta_i, \psi)$ does not depend on θ_i.

Let X represent the $n \times q$ design matrix containing 1's in the first column and $q-1$ observed covariates in the remaining ones. Assume that β is a $q \times 1$ vector of unknown coefficients. The relationship between Y_i and the linear predictor is established by defining a connection between θ_i and $X_{\bullet i}^\top \beta + \delta_i$, for $i = 1, \ldots, n$. Consider $\theta_i = g(X_{\bullet i}^\top \beta + \delta_i)$, where $g(\bullet)$ is the link function. In particular, one may have $\theta_i = E[Y_i|X_{\bullet i}^\top \beta, \delta_i]$ for many important distributions such as Bernoulli, Poisson, Normal, Gamma, and Beta; the last two being parameterized here with respect to the mean.

In the context of GLMM, the model also includes random effects denoted by δ_i; see [24] for more details about GLMM's. In this study, we use a hierarchical structure assuming a spatial and temporal dependence between the random effects. This will be done as described in [7] using factor analysis with nonlinear interactions between latent factors. The analysis developed in [7], which is our main reference, was inspired by the model earlier proposed in [23]. The main difference between this M.Sc. dissertation and the study in [7] is the list of distributions being explored. Here, we consider Y_i with a

continuous distribution (Gaussian, Gamma, and Beta). The main reference evaluates two discrete distributions (Bernoulli and Poisson).

2.1 Likelihood function

The framework here is a multilevel (hierarchical) model developed based on the Gaussian, Gamma, and Beta distributions for the response Y_i. We assume that each sample unit i is obtained from a location and time. As a result, the random effect will be denoted by $\delta_{l_i^* t_i^*}$, for $i = 1, \ldots, n$. Let l_i^* be the location of the i-th observation, with $l^* = (l_1^*, \ldots, l_n^*)^\top$. One can say that $l_i^* = l$, if i belongs to region $l \in \{1, 2, \ldots, L\}$. Similarly, t_i^* is the time of the i-th observation, with $t^* = (t_1^*, \ldots, t_n^*)^\top$. We write $t_i^* = t$, if i belongs to time $t \in \{1, 2, \ldots, T\}$. As a consequence of this configuration, note that all observations from (l, t) contain information about δ_{lt}, therefore, we can organize all δ_{lt}'s in a $L \times T$ matrix called δ. The following descriptions indicate the link and likelihood functions related to the three continuous distributions explored in the present dissertation. It should be noted that, below, the link functions in equations 2.1, 2.3, and 2.5 can be replaced by others. However, for the Normal case, the Identity link function allows the conjugate analysis in the way shown in section 2.2. In the case of Gamma and Beta, the choice for these binding functions was due to the fact that they are the most used in the literature.

Gaussian distribution for Y_i:

Assume $Y_i|\theta_i, \psi \sim \text{Normal}(\theta_i, \psi)$, where θ_i is the mean and ψ is the variance. The connection between θ_i and the linear predictor is established through the identity link function:

$$\theta_i = g_N(X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}) = X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}, \quad i \in (1, \ldots, n). \tag{2.1}$$

Let $Y = (Y_1, Y_2, \ldots, Y_n)^\top$ be $n \times 1$ vector of independent random variables. In addition, denote $\delta_{l^* t^*} = (\delta_{l_1^* t_1^*}, \delta_{l_2^* t_2^*}, \ldots, \delta_{l_n^* t_n^*})^\top$. The likelihood function is then written as:

$$p(Y|X, \beta, \psi, \delta) = \left(\frac{1}{2\pi\psi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\psi}(Y - X\beta - \delta_{l^* t^*})^\top (Y - X\beta - \delta_{l^* t^*})\right\}, \tag{2.2}$$

where $Y_i \in \mathbb{R}$. The normal distribution for the response is widely used in a regression setting since it is the appropriate choice to deal with real-valued symmetric observations quite common in many applications.

Gamma distribution for Y_i:

Suppose that $Y_i|\theta_i, \psi \sim \text{Gamma}(\theta_i, \psi)$, with mean θ_i and shape parameter ψ. In this case, the log-linear modeling is determined through the following relationship:

$$\theta_i = g_G(X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}) = \exp\left\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\right\}, \quad i \in (1, \ldots, n). \tag{2.3}$$

When assuming that the Y_i's are all independent, the likelihood function is given by:

$$\begin{aligned} p(Y|X, \beta, \psi, \delta) &= \prod_{i=1}^{n} \exp\left\{\psi \log \psi - \psi X_{\bullet i}^\top \beta - \psi \delta_{l_i^* t_i^*} - \log \Gamma(\psi) + \right. \\ &\quad \left. + \psi \log(Y_i) - \log Y_i - \psi Y_i / \exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\}\right\}, \end{aligned} \tag{2.4}$$

where $Y_i > 0$. In GLM, the Gamma model is usually assumed for real-valued observations showing an asymmetric behavior determined by a heavier right tail. The indicated log link is widely used in this context since it provides an appropriate connection between the real-valued linear predictor and the strictly positive mean θ_i.

Beta distribution for Y_i:

Now assume that Y_i is a continuous random variable with bounded support given by the interval $(0, 1)$. In this case, the Beta distribution can be considered for $Y_i|\theta_i, \psi$. We follow the parameterization defined in [6], in which θ_i is the mean and the ψ can be seen as a dispersion parameter. The adopted distribution is specified by writing $Y_i|\theta_i, \psi \sim \text{Beta}(\theta_i, \psi)$. Note that the parametric space of θ_i is also the interval $(0, 1)$, therefore, the logit is a natural choice for the link function. In the present study, we write:

$$\theta_i = g_B(X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}) = \frac{\exp\left\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\right\}}{1 + \exp\left\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\right\}}, \quad i \in (1, \ldots, n). \tag{2.5}$$

When a random sample $Y_1, \ldots, Y_n$ is observed, all variables independent, the likelihood takes the form

$$\begin{aligned} p(Y|X, \beta, \psi, \delta) &= \prod_{i=1}^{n} p(Y_i|X_{\bullet i}^\top, \beta, \psi) \\ &= \prod_{i=1}^{n} \frac{\Gamma(\psi)}{\Gamma(\theta_i \psi)\Gamma((1-\theta_i)\psi)} Y_i^{\theta_i \psi - 1} (1 - Y_i)^{(1-\theta_i)\psi - 1}, \end{aligned} \tag{2.6}$$

where θ_i should be replaced by the expression in (2.5). Other distributions can be considered to handle the regression setting for continuous bounded response variables; see details in [3] and references therein. In the present dissertation, we choose to explore the well-known Beta model, which is used in most applications found in the literature.

Finally, the necessary details related to parameterization and likelihood were presented for the three target models that will be explored in this study. The next section will indicate the prior specifications required for the Bayesian inference. These prior specifications can be seen as elements representing levels below the distribution of Y_i in the hierarchical structure of the model.

2.2 Prior distributions

In the Bayesian approach for inference, it is necessary to obtain the posterior distribution by combining the prior and the likelihood through the Bayes rule. A Markov Chain Monte Carlo (MCMC) method will be applied for indirect sampling since the joint posterior distribution is only known up to a normalizing constant. The chosen MCMC is a Gibbs Sampling [13, 12] with some Metropolis-Hastings [26, 16] steps, which is also implemented in the main reference [7] and adapted here for the three continuous distributions. We highlight the fact that, for most parameters, the prior specification considered here is similar to the one defined in [7].

Assume that $\beta \sim N_q(\mu, \Sigma)$ with mean $\mu = (m_{\beta_0}, m_{\beta_1}, \ldots, m_{\beta_{q-1}})^\top$ and covariance matrix $\Sigma = \text{diag}\{s_{\beta_0}, s_{\beta_1}, \ldots, s_{\beta_{q-1}}\}$, both fixed. In terms of dispersion parameter, for the Gamma and Beta cases, consider $\psi \sim \text{Gamma}(a_\psi, b_\psi)$, with shape $a_\psi > 0$ and scale $b_\psi > 0$. Note that the parametrization based on the mean is not used in this Gamma prior for ψ. In the Normal case, we choose the prior for the variance ψ to allow a conjugate analysis simplifying the calculations to determine the corresponding full conditional posterior distribution. Let $\chi = 1/\psi$ be the precision parameter, then set $\chi \sim \text{Gamma}(a_\chi, b_\chi)$ with shape $a_\chi > 0$ and scale $b_\chi > 0$. Recall that if χ has a Gamma distribution, then ψ has an Inverse Gamma distribution with the same parameters.

The random effect matrix δ is decomposed according to the factor analytic structure proposed in [23]. We write

$$\delta \;=\; \alpha\,\lambda \;+\; \eta \;+\; \epsilon. \tag{2.7}$$

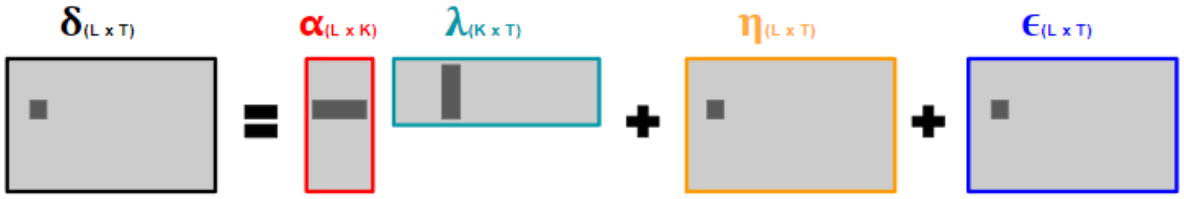
In this framework, $\alpha_{(L\times K)}$ is the factor loadings matrix, in which the spatial dependence will be defined. The term $\lambda_{(K\times T)}$, with $K << L$, is the factor scores matrix that will contain a temporal dependence. The element $\eta_{(L\times T)}$ is a matrix with zero and non-zero rows (non-zero rows represent the nonlinear interaction between the latent factors). Finally, $\epsilon_{(L\times T)}$ is the error matrix. The quantity K represents the number of factors, which must be fixed by an analyst in this study. In [23], the authors introduce the nonlinear interaction matrix η, and they discuss different ways of defining this element. In the present dissertation, we follow [7] to establish a simpler configuration for η, in which the non-zero rows represent only one type of nonlinear interaction. Figure 2.1 shows a scheme representing the terms of the factor model with a nonlinear interaction.

The spatial dependence between regions is defined for the columns of the loadings matrix α. This is achieved using the Conditional Autoregressive modeling - CAR [4], which is widely used in applications related to areal data. Let $\alpha_{\bullet k}$ represent the k-th column of α, the desired prior is based on the following multivariate Normal specification:

$$\alpha_{\bullet k}|\tau_\alpha \sim N_L(\mathbf{0}_{(L\times 1)}, \tau_\alpha[D_\alpha - \rho_\alpha W_\alpha]^{-1}), \quad k \in \{1, \ldots, K\}\,. \tag{2.8}$$

The scalar τ_α is a variance parameter for which we set the inverse gamma prior $\tau_\alpha \sim \text{IG}(a_{\tau_\alpha}, b_{\tau_\alpha})$, with $a_{\tau_\alpha} > 0$ and $b_{\tau_\alpha} > 0$ defined by the researcher. The term W_α is a $L \times L$ binary neighborhood matrix, where 1 indicates neighbor regions and 0 otherwise. Consider $D_\alpha = \text{diag}\{w_{\alpha_{1+}}, \cdots, w_{\alpha_{L+}}\}$, such that $w_{\alpha_{l+}}$ is the sum of the l-th row of W_α. In other words, $w_{\alpha_{l+}}$ represents the number of neighbors of the region l. The scalar ρ_α can be seen as a parameter controlling the strength of spatial dependence. Note that the matrix $D_\alpha - W_\alpha$ cannot be inverted, therefore, $\rho_\alpha = 1$ leads to an improper distribution. The parameter ρ is included to ensure a proper joint distribution whose covariance matrix can be obtained by inverting $D_\alpha - \rho_\alpha W_\alpha$. It can be shown that $\rho_\alpha \in (1/\lambda_{(1)}, 1/\lambda_{(L)})$, with $\lambda_{(1)}$ and $\lambda_{(L)}$ being, respectively, the smallest (negative) and largest (positive) eigenvalues of $D_\alpha^{-1/2} W_\alpha D_\alpha^{-1/2}$. In practice, ρ_α is usually chosen in the interval $(0, 1)$, where a value near 0 indicates weak spatial dependence, and a value close to 1 represents strong spatial dependence. Negative values of ρ_α are also possible, but this choice would lead to an unusual scenario difficult to interpret the relationship between each loading and its neighbors. The conditional distribution of each loading (given its neighbors) is basically the average of the loadings in the neighborhood weighted by ρ_α. The reader is referred to [2] for more details about the CAR modeling.

Figure 2.1: Structure of the factor model including the nonlinear interaction matrix.



One can assume, for example, the Uniform$(0, 1)$ prior for ρ_α and estimate the parameter in the Bayesian context. It is important to emphasize that making inferences about ρ_α is not an easy task, i.e., conclusions regarding the estimated magnitude and the strength of spatial dependence may not be too clear for the analyst. Our focus in the present dissertation is not testing the presence or absence of spatial dependence. In fact. we propose a study where including a spatial dependence is reasonable for the target application. Note that a simpler model without the spatial association can be easily defined as a particular case by setting a diagonal covariance matrix in (2.8). In the present M.Sc. dissertation, we choose to fix ρ_α in a value close to 1 to impose the presence of a spatial dependence assumed to exist in the data.

The temporal dependence is imposed on the rows of the scores matrix λ. This is done using again the first-order Autoregressive model AR(1). In this work, the CAR model defined can be seen as a AR(1) modeling structure [31] widely used in dynamic modeling and time series analysis. We choose to write the temporal dependence via the

CAR specification as a strategy to simplify the presentation using the same ideas from (2.8). Let $\lambda_{k\bullet}$ be the vector in the k-th row of λ. Then, the time dependence is inserted through the following multivariate normal distribution:

$$\lambda_{k\bullet}|\tau_\lambda \sim N_T(\mathbf{0}_{(T\times 1)}, \tau_\lambda[D_\lambda - \rho_\lambda W_\lambda]^{-1}). \tag{2.9}$$

Here, $\tau_\lambda = 1$ is fixed as a model simplification and also a strategy to avoid identifiability issues due to the product between loadings and latent factors in the model [7]; more details will be discussed in Chapter 3. The binary matrix W_λ is band diagonal to indicate the time neighbors. In other words, the main diagonal contains 0's and both the super- and sub-diagonal contain 1's [22] suggesting that time t has in general two neighbors ($t-1$ and $t+1$). The element D_λ is a diagonal matrix with the number of neighbors in the diagonal. If $T = 4$, for example, we have $D_\lambda = \text{diag}\{1, 2, 2, 1\}$. The previous discussion about choosing ρ_α is also valid for ρ_λ. Again, we will fix ρ_λ and allow a proper prior where $D_\lambda - \rho_\lambda W_\lambda$ can be inverted [2].

The presence of a nonlinear interaction effect in the rows of η is defined by a mixture distribution of the type mixture distribution (George and McCulloch, 1993, 1997). This prior involves a Gaussian Process (GP) component [23, 7] whose covariance function depends on the distance between columns of λ. Let $\eta_{l\bullet} = (\eta_{l1}, \ldots, \eta_{lT})^\top$ be the l-th row of η, then the mentioned prior specification is as follows

$$\eta_{l\bullet}|p_l, \eta^* \sim (1 - p_l)D_0 + p_l D_{\eta^*}. \tag{2.10}$$

The mixture prior configuration allows for sparsity in η. This choice is an interesting strategy to allow the model to decide, based on evidence from the data, the presence or absence of a non-null interaction affecting each region. We can see, D_0 is a degenerate distribution regarding the $(1 \times T)$ vector of zeros. The component D_{η^*} is another point mass distribution related to the $(1 \times T)$ vector η^* to be estimated. We set $\eta^*|\lambda \sim N_T(\mathbf{0}_{(T\times 1)}, \kappa(\lambda))$, with $\mathbf{0}_{(T\times 1)}$ being a vector of zeros and $\kappa(\lambda)$ is $T \times T$ covariance matrix depending on λ through a covariance function. Nonlinear interaction between latent factors (rows of λ) is introduced by assuming a covariance function to build the matrix $\kappa(\lambda)$. Let t_1 and t_2 be two-time points (columns of λ), the adopted covariance function is the squared exponential (or Gaussian) expressed by $\kappa(\lambda)_{t_1,t_2} = \gamma \exp\{-\phi^2||\lambda_{\bullet t_1} - \lambda_{\bullet t_2}||^2\}$. The symbol $||\bullet||$ indicates the Euclidean norm, and $\gamma > 0$ is the variance related to η^*. The Gaussian Process corresponding to this covariance function is stationary, isotropic, and infinitely differentiable [2].

The proposed GLMM is said to have a "nonlinear interaction" due to the fact that the surface generated by the GP may not be a plane. As an example, for $K = 2$ and for each time t, this surface will indicate the magnitude of interaction in the coordinate $(\lambda_{1t}, \lambda_{2t})$, where λ_{1t} is a score from Factor 1 and λ_{2t} is a score from Factor 2. As exposed by [23], if the coordinates $\lambda_{\bullet t_1}$ and $\lambda_{\bullet t_2}$ are close in the $\mathbb{R}^T$ space, then the scores in t_1 and

t_2 are similar. In this case, $\kappa(\lambda)_{t_1,t_2} \approx \gamma$. However, the greater the distance between these coordinates, the greater the dissimilarity between the scores of t_1 and t_2, then $\kappa(\lambda)_{t_1,t_2} \approx 0$. We can conclude that considering a large γ allows for an interaction away from zero. On the other hand, adopting a small γ determines an interaction near zero. Therefore, if the researcher desires to explore the model with nonlinear interactions, small γ is not a good choice. In addition, the fixed parameter $\phi > 0$ controls the strength of the distance $||\lambda_{\bullet t_1} - \lambda_{\bullet t_2}||$. The ϕ element is an adjustable scale parameter that controls how close λ_{t_1} and λ_{t_2} must be in order to allow a significant association between t_1 and t_2. If $\phi \approx 0$, an increase in this distance has a weak impact due to the product between ϕ^2 and the squared Euclidean norm. One can conclude that for a large ϕ, increasing the Euclidean norm will have a stronger impact forcing the covariance towards 0.

Moreover, the term p_l is the prior probability of $\eta_{l\bullet} = \eta^*$, i.e., the prior probability that the l-th region is affected by the non-null nonlinear interaction. When $p_l = 0\ \forall l$ in (2.10), the model is linear and can be written as a factor model without interactions. On the other hand, when $p_l \neq 0$ for at least one l, the matrix η will have $\eta_{l\bullet} = \eta^*$ determining the presence of the nonlinear interaction in the model. In this study, we set the prior $p_l \sim \text{Beta}(a_p, b_p)$, with $a_p > 0$ and $b_p > 0$, and parameterized such that the mean is $a_p/(a_p + b_p)$.

For computational reasons, the Equation (2.10) is restructured considering an indicator variable Z_l as follows:

$$\eta_{l\bullet}|Z_l, \eta^* \sim (1 - Z_l)D_0 + Z_l D^*_\eta, \quad \text{with} \quad Z_l \sim \text{Bernoulli}(p_l). \tag{2.11}$$

In terms of additional notation, consider: $p = (p_1, p_2, \cdots, p_L)^\top$ and $Z = (Z_1, Z_2, \cdots, Z_L)^\top$. For more details about the factor model with nonlinear interactions, the reader is referred to [23].

Finally, let $\epsilon_{l\bullet} = (\epsilon_{l1}, \ldots, \epsilon_{lT})^\top$ represent the l-th row of the error matrix ϵ. The model is configured by assuming that:

$$\epsilon^\top_{l\bullet}|\sigma^2 \sim N_T(\mathbf{0}, \sigma^2 I_T), \tag{2.12}$$

where I_T is a $T \times T$ identity matrix. The term σ^2 is the common variance for all regions. In a standard factor model, different variances would be assumed for each l. The proposed single variance version is a reasonable simplification justified by the fact that here the factor model is decomposing a random effect matrix instead of a data matrix. The list of prior specifications is completed by setting the prior $\sigma^2 \sim IG(a_{\sigma^2}, b_{\sigma^2})$, with $a_{\sigma^2} > 0$ and $b_{\sigma^2} > 0$ chosen by the practitioner. Again, the inverse gamma is adopted to allow a conjugate analysis.

As can be noted, the structure of the proposed factor model is quite complex involving the sum and product of terms to be estimated. As a result, this type of model has identifiability issues, which must be treated to allow correct estimation and conclusions

in the statistical inference. The next section is dedicated to reporting these issues and discussing the chosen solutions.

2.3 Identification issues

This section discusses the identifiability issues affecting the FA part of the proposed GLMM. The first issue is the fact that the FA can exchange signs between a column of α and the corresponding row of λ. This type of problem is due to the product $\alpha\lambda$ involving two unknown terms to be estimated. Another issue justified by this aspect is the possibility of having two columns of α, and the corresponding rows of λ, exchanging their positions. The authors in [19] provide a discussion about these particular problems. According to them, for any orthogonal matrix Q, we have $\alpha\lambda = \alpha Q\ Q^{\top}\lambda$. This means that many configurations of the loadings and scores matrices can be defined to determine the same result of the product. Note that the difficulty is not only related to the sign or column exchange, but it also involves the magnitudes of the loadings and scores. If we increase α, the scores in λ can be reduced accordingly to provide the same result of $\alpha\lambda$. As a solution for the sign exchange, one can simply require that a specific group of loadings should be positive, which will force the model to adapt the sign of λ to this restriction. Now, regarding the magnitude exchange, a widely-used strategy (previously mentioned in this dissertation) is to set $\tau_{\lambda} = 1$ to ensure that the factor scores are not free to increase their magnitude and then force a reduction in α.

The authors in [23] indicate another identification problem related to the sum $\alpha\lambda + \eta$, which involves two unknown matrices ($\alpha\lambda$ and η) to be estimated. Note that it is possible to represent the l-th row of the matrix resulting from this sum as $\alpha_{l\bullet}\lambda + \eta_{l\bullet} = C_1 + C_2$, where $C_1 = C\alpha_{l\bullet}\lambda$ and $C_2 = (1 - C)\alpha_{l\bullet}\lambda + \eta_{l\bullet}$. Here, C is any real value. Therefore, there are infinite choices of C that provide two matrices having the same result of $\alpha_{l\bullet}\lambda + \eta_{l\bullet}$. In the present work, we apply the same solution considered in [23]. The loadings matrix α is configured assuming a partition of the L regions in the space. We use some additional information (from the researcher or based on an auxiliary variable) to define $K + 1$ disjoints groups $G_1, G_2, \cdots, G_K$ and $G_{Extra} = G_E$, such that $G_1 \cup G_2 \cup \cdots \cup G_K \cup G_E = \{1, 2, \cdots, L\}$. Each group $G_{k \neq E}$ contains only the regions affected exclusively by the k-th factor and not affected by interactions or other factors. Furthermore, G_E includes the regions having unknown associations with the main factors and the interaction. According to this partition, we set $\alpha_{lk} = 0$ when $l \notin G_k$ and $l \notin G_E$. In addition, we assume the absence of interaction effect for the groups expected to be exclusively associated with one of the main factors, i.e, $\eta_{l\bullet} = 0$ for all $l \in G_{k \neq E}$. This

configuration will be inserted in the model through the prior distribution specified for α and η.

Figure 2.2: Configuration of α and η in an example with $K = 3$ factors. White indicates matrix elements defined to be zero. Grey indicates the elements allowed to differ from zero.

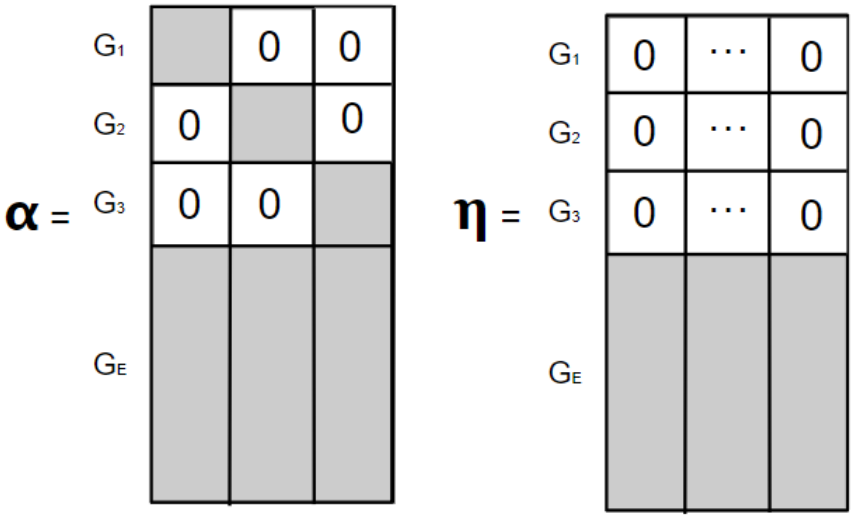


Figure 2.2 displays how α and η are configured to handle identifiability issues in a scenario with $K = 3$ factors. The sign exchange issue can be solved by requiring positive loadings in the grey parts related to $G_{k \neq E}$. Assuming the zeros (white parts of α) to ensure that the k-th column is strongly related to the group G_k is the solution to avoid the column exchange issue. The absence of interaction affecting the groups $G_{k \neq E}$ (white in the rows of η) is a restriction preventing a free communication between $\alpha_{l\bullet}\lambda$ and $\eta_{l\bullet}$. All regions affected by interactions are supposed to be a member of the group G_E. In order to impose that $\eta_{l\bullet} = \mathbf{0}$, one can simply set a Beta prior for p_l concentrating probability mass near 0.

Given the restrictions imposed to α, as represented in Figure 2.2, the prior in (2.8) must be adapted accordingly. The same notation defined in [7] is applied here. Let α_{0k} be a $L_{0_k} \times 1$ vector, related to the k-th column of α, containing the zero loadings; that is $\{\alpha_{lk} = 0;\ l \notin G_k \text{ and } l \notin G_E\}$. In addition, denote $\alpha_{\emptyset k}$ as the $L_{\emptyset_k} \times 1$ vector containing the non-null loadings from the k-th column; that is $\{\alpha_{lk} \neq 0;\ l \in G_k \text{ or } l \in G_E\}$. The multivariate normal specified in (2.8) can be rewritten assuming the mentioned partition $\alpha_{\bullet k} = (\alpha_{\emptyset k}, \alpha_{0k})^\top$ as follows:

$$(\alpha_{\emptyset k}, \alpha_{0k})^\top | \tau_\alpha \sim N_L(\mathbf{0}, \tau_\alpha B_k), \quad B_k = \begin{bmatrix} B_{k,11} & B_{k,12} \\ B_{k,21} & B_{k,22} \end{bmatrix}. \tag{2.13}$$

The covariance matrix B_k is basically the matrix $[D_\alpha - \rho_\alpha W_\alpha]^{-1}$ obtained by assuming the CAR model. Let $\partial_{\emptyset_k}$ be the set of indices in $\{1, 2, \cdots, L\}$ related to $\alpha_{\emptyset k}$. Similarly, ∂_{0_k} is the set of indices in $\{1, 2, \cdots, L\}$ corresponding to α_{0k}. We define $B_{k,11} = B[\partial_{\emptyset_k}, \partial_{\emptyset_k}]$ to represent the sub-matrix formed by the rows $\partial_{\emptyset_k}$ and columns $\partial_{\emptyset_k}$ of matrix B. In line with this notation, we can also write $B_{k,22} = B[\partial_{0_k}, \partial_{0_k}]$, $B_{k,12} = B[\partial_{\emptyset_k}, \partial_{0_k}]$, and $B_{k,21} = B[\partial_{0_k}, \partial_{\emptyset_k}]$. In terms of dimension we have: $B_{k,11}$ is $L_{\emptyset k} \times L_{\emptyset k}$, $B_{k,22}$ is $L_{0k} \times L_{0k}$, $B_{k,21}$ is $L_{0k} \times L_{\emptyset k}$, and $B_{k,12}$ is $L_{\emptyset k} \times L_{0k}$. Using well-known properties of the Multivariate Normal distribution, the following conditional distribution is determined:

$$\alpha_{\emptyset_k} | \alpha_{0_k}, \tau_\alpha \ \sim \ N_{L_{\emptyset_k}}(\mu_{\emptyset_k|0_k}, \tau_\alpha B_{\emptyset_k|0_k}), \tag{2.14}$$

where the mean is $\mu_{\emptyset_k|0_k} = \mathbf{0}_{L_{\emptyset_k} \times 1} + B_{k,12}(B_{k,22})^{-1}(\alpha_{0k} - \mathbf{0}_{L_{\emptyset_k} \times 1}) = \mathbf{0}_{L_{\emptyset_k} \times 1}$ and the covariance matrix is based on $B_{\emptyset_k|0_k} = B_{k,11} - B_{k,12}B_{k,22}^{-1}B_{k,21}$.

Now the discussion about the identifiability issues in the FA part of the proposed GLMM is complete. Note that solutions were presented to handle the problems, and the restrictions determined by these solutions were carefully considered in the adaptation of the prior with spatial dependence assumed for the columns of α. The next section discusses aspects related to the MCMC implemented to allow the Bayesian inference for the target models (Normal, Gamma, and Beta).

2.4 Bayesian inference

The joint posterior distribution $p(\beta, \delta, \psi, \alpha, \tau_\alpha, \lambda, \eta, \sigma^2, Z, p|X, Y)$ can only be determined up to an unknown normalizing constant. In order to estimate the parameters defined in the proposed model, and measure our posterior uncertainty about them, an MCMC algorithm is required to allow indirect sampling from this target distribution. As previously mentioned, the MCMC implemented here is a Gibbs Sampling with some Metropolis-Hastings (MH) steps [13, 12, 26, 16]. When using the MH, the candidate values are generated via Gaussian random walk in this dissertation. In addition, the MH algorithm is carefully tuned to determine levels of acceptance rates as recommended in the literature [30].

The algorithm used in this dissertation is in general similar to the one defined in [7]. The `R` programming language [29] is applied in the present study. Now, we list the key steps of the proposed MCMC.

1. Set initial values for all unknown parameters.

2. Generate $\eta^*|\bullet \sim N_T(M_{\eta^*}, V_{\eta^*})$, where
$M_{\eta^*} = V_{\eta^*} \sum_{l=1}^{L} (Z_l/\sigma^2)(\delta_{l\bullet}^\top - \lambda^\top \alpha_{l\bullet}^\top)$ and $V_{\eta^*} = \left[\sum_{l=1}^{L} (Z_l/\sigma^2)\mathbf{I}_T + \kappa(\lambda)^{-1}\right]^{-1}$.

3. Calculate the following normalization
$p^*(Z_l = 1|\bullet) = p(Z_l = 1|\bullet)/[p(Z_l = 1|\bullet) + p(Z_l = 0|\bullet)]$ such that
$p(Z_l = 1|\bullet) \propto \exp\{(-1/2\sigma^2)[(\eta^*)^\top \eta^* - 2\eta^*(\delta_{l\bullet} - (\alpha_{l\bullet}\lambda))^\top]\} \times$
$\times\, p(\eta_{l\bullet} = \eta^*|\lambda, Z_l = 1)\; p_l$, and
$p(Z_l = 0|\bullet) \propto \exp\{(-1/2\sigma^2)[(\mathbf{0})^\top \mathbf{0} - 2\mathbf{0}(\delta_{l\bullet} - (\alpha_{l\bullet}\lambda))^\top]\} \times$
$\times\, p(\eta_{l\bullet} = \mathbf{0}|\lambda, Z_l = 0)\; (1 - p_l) = (1 - p_l)$.
Generate $u \sim \mathrm{U}(0, 1)$. Set $(Z_l = 1,\ \eta_{l\bullet} = \eta^*)$, if $u < p^*(Z_l = 1|\bullet)$. Otherwise consider $(Z_l = 0,\ \eta_{l\bullet} = \mathbf{0})$.

4. Generate $p_l|\bullet \sim \mathrm{Beta}(a_p + Z_l, b_p - Z_l + 1)$.

5. Now, sample the coefficients in β. This step must be adapted for each probability distribution defined for Y_i. Here, the MH is necessary for the Gamma and Beta cases given that their complete conditional posterior distributions for β are not fully known. The Normal case is simpler and does not require the MH to generate β. We choose to generate each β_j separately, otherwise, it would be difficult to tune the MH related to the vector β.

6. Generate ψ accounting for the probability distribution defined for Y_i. Again, the Normal case is simpler and does not require the MH. Sampling from the complete conditional posterior requires the MH for the Gamma and Beta cases.

7. Generate δ_{lt} considering the chosen distribution for the response Y_i. The MH is necessary for all cases.

8. Sample $\sigma^2|\bullet \sim IG(a^*_{\sigma^2}, b^*_{\sigma^2})$ with $a^*_{\sigma^2} = LT/2 + a_{\sigma^2}$ and
$b^*_{\sigma^2} = b_{\sigma^2} + (1/2)\sum_{k=1}^{K} \alpha_{\emptyset_k}^\top B^{-1}_{\emptyset_k|0_k} \alpha_{\emptyset_k}$.

9. Generate $(\alpha_{\emptyset k}|\alpha_{0k} = \mathbf{0}, \bullet) \sim \mathrm{N}_{L_{\emptyset_k}}(M^*_{\emptyset_k|0_k}, V^*_{\emptyset_k|0_k})$, with
$M^*_{\emptyset_k|0_k} = (1/\sigma^2)V^*_{\emptyset_k|0_k} \sum_{t=1}^{T} \left(\delta_{\emptyset_k t} - \eta_{\emptyset_k} - \sum_{k' \neq k} \alpha_{\emptyset k'} \lambda_{k't}\right) \lambda_{kt}$, and
$V^*_{\emptyset_k|0_k} = \left[(1/\tau_\alpha)B^{-1}_{\emptyset_k|0_k} + (1/\sigma^2)\sum_{t=T}^{T} {}^{2}_{kt}\mathbf{I}_{L_{\emptyset_k}}\right]^{-1}$.

10. Generate $\tau_\alpha|\bullet \sim \mathrm{IG}(a^*_{\tau_\alpha}, b^*_{\tau_\alpha})$, with $a^*_{\tau_\alpha} = a_{\tau_\alpha} + \sum_{k=1}^{K} L_{\emptyset_k}/2$, and
$b^*_{\tau_\alpha} = b_{\tau_\alpha} + (1/2)\sum_{k=1}^{K} \alpha_{\emptyset_k}^\top B^{-1}_{\emptyset_k|0_k} \alpha_{\emptyset_k}$.

11. Sample $\lambda_{k\bullet}$. The complete conditional posterior has the kernel
$p(\lambda_{k\bullet}|\bullet) \propto \mathrm{N}_T[\lambda_{k\bullet}|M_\lambda, V_\lambda]\; |\kappa(\lambda)|^{-1/2}\; \exp\{-(1/2)\eta^{*\top}\kappa(\lambda)^{-1}\eta^*\}$, where
$\mathrm{N}_T[\lambda_{k\bullet}|M_\lambda, V_\lambda]$ is the density of $\mathrm{N}_T(M_\lambda, V_\lambda)$ evaluated at $\lambda_{k\lambda}$,
$V_\lambda = [(1/\tau_\lambda)(D_\lambda - \rho_\lambda W_\lambda) + \sum_{l=1}^{L} (\alpha_{lk}^2/\sigma^2)\mathbf{I}_T]^{-1}$, and

$M_\lambda = V_\lambda(1/\sigma^2)\sum_{l=1}^{L}\alpha_{lk}(\delta_{l\bullet}^\top - \eta_{l\bullet}^\top - \sum_{k'\neq k}\alpha_{lk'}\lambda_{k'\bullet}^\top)$.
Here, the MH step is required for all cases.

Now, we present the adaptations required, in Steps 5, 6, and 7 of the MCMC algorithm, for each distribution defined for the response Y_i. We begin with the simplest case assuming a Gaussian response.

Normal distribution for Y_i:

In this particular situation, the MH within the Gibbs Sampling is not necessary for Steps 5 and 6 due to the conjugate analysis. The likelihood function related to the Gaussian case can be found in (2.2). The Steps 5, 6, and 7 of the MCMC are as follows.

5. Sample $\beta|\bullet \sim N_q(M_\beta, V_\beta)$, with $V_\beta = [X^\top X + (S_\beta)^{-1}]^{-1}$, and $M_\beta = V_\beta(X^\top Y - X^\top\delta_{l^*t^*}) + S_\beta^{-1}m_\beta$.

6. Generate $\psi|\bullet \sim \text{Gamma}(a_\psi^*, b_\psi^*)$, with $a_\psi^* = \frac{n}{2} + a_\psi$, and $b_\psi^* = Y^\top Y - Y^\top\delta_{l^*t^*} - \delta_{l^*t^*}^\top Y + \delta_{l^*t^*}^\top\delta_{l^*t^*} + M_\beta(S_\beta m_\beta + 2b_\psi - M_\beta^\top V_\beta^{-1}M_\beta)$.

7. Sample $(\delta_{lt}|\bullet)$ via MH. The log kernel of the complete conditional posterior is
$\log p(\delta_{lt}|\bullet) = -[1/(2\psi)]\left[\sum_{i=1}^{n}(X_{\bullet i}^\top\beta + \delta_{l^*t^*})^2 1_{\{l_i^*=l\}\{t_i^*=t\}} + -2\sum_{i=1}^{n} y_i\delta_{l^*t^*}1_{\{l_i^*=l\}\{t_i^*=t\}}\right] - [1/(2\sigma^2)][\delta_{lt}^2 - 2\delta_{lt}(\alpha_{l\bullet}\lambda_{\bullet t} + \eta_{lt})] + C_{\delta_{Normal}}$,
where $C_{\delta_{Normal}}$ is an unknown constant.

Gamma distribution for Y_i:

When positive asymmetric continuous values are observed for the response, one may consider the Gamma version of the model with likelihood defined in (2.4). The Steps 5, 6, and 7 of the MCMC algorithm are adapted as follows for this case.

5. Generate $\beta_j|\bullet$ via MH. The log kernel of the complete conditional posterior is
$\log p(\beta_j|\bullet) = -\psi\beta_j\sum_{i=1}^{n}X_{ji} - \sum_{i=1}^{n}\psi y_i/\exp\{X_{\bullet i\beta} + \delta_{l_i^*t_i^*}\} + - [1/(2s_{\beta_j})][\beta_j^2 - 2\beta_j m_{\beta_j}] + C_{\beta_{Gamma}}$.

6. Sample $\psi|\bullet$ via MH. The complete conditional posterior has the following log kernel.
$\log p(\psi|\bullet) = n\psi\log(\psi) - \psi\sum_{i=1}^{n}(X_{\bullet i}\beta + \delta_{l^*t^*}) - n\log[\gamma(\psi)] + \psi\sum_{i=1}^{n}\log(y_i) + - \sum_{i=1}^{n}\psi y_i/\exp\{X_{\bullet i}^\top\delta_{l_i^*t_i^*}\} + a_{\psi-1}\log(\psi) - b_\psi\psi + C_{\psi_{Gamma}}$.

7. Generate $\delta_{lt}|\bullet$ via MH. The log kernel of the complete conditional posterior is
$\log p(\delta_{lt}|\bullet) = -\psi\sum_{i=1}^{n}\delta_{l^*t^*}1_{\{l_i^*=l\}\{t_i^*=t\}} - \sum_{i=1}^{n}1_{\{l_i^*=l\}\{t_i^*=t\}}\psi y_i/\exp\{X_{\bullet i}^\top\beta\delta_{l^*t^*}\} + - [1/(2\sigma^2)][\delta_{lt}^2 - 2\delta_{lt}(\alpha_{l\bullet}\lambda_{\bullet t} + \eta_{lt})] + C_{\delta_{Gamma}}$.

In the previous expressions, the terms $C_{\beta_{Gamma}}$, $C_{\psi_{Gamma}}$, and $C_{\delta_{Gamma}}$ represent the corresponding unknown normalizing constant in the log scale.

Beta distribution for Y_i:

In a scenario where the response variable is continuous and bounded, one can assume the Beta GLMM proposed in this dissertation. Details about parameterization and the corresponding likelihood function are given in (2.6). The Steps 5, 6, and 7 of the MCMC algorithm are adapted as follows in this situation.

5. Sample $\beta_j|\bullet$ via MH. The log kernel of the complete conditional posterior is
$$\begin{aligned}\log p(\beta_j|\bullet) \;=\; & -\textstyle\sum_{i=1}^{n}\log\Gamma(\psi E_i) - \sum_{i=1}^{n}\log\Gamma[\psi(1-E_i)] + \psi\sum_{i=1}^{n}[\log(y_i)E_i] + \\ & - \psi\textstyle\sum_{i=1}^{n}[(1-y_i)E_i] - [1/(2s_{\beta_j})](\beta_j^2 - 2\beta_j m_{\beta_j}) + C_{\beta_{Beta}},\end{aligned}$$
where $E_i = \exp\{X_{\bullet i}^{\top} + \delta_{l_i^* t_i^*}\}/(1+\exp\{X_{\bullet i}^{\top} + \delta_{l_i^* t_i^*}\})$.

6. Generate $\psi|\bullet$. The complete conditional posterior has the following log kernel.
$$\begin{aligned}\log p(\psi|\bullet) \;=\; & n\log\Gamma(\psi) - \textstyle\sum_{i=1}^{n}\log\Gamma(E_i) - \sum_{i-1}^{n}\log\Gamma[\psi E_i] + \\ & + \psi\textstyle\sum_{i=1}^{n}[\log(y_i)E_i]\;\; \psi\;\; \sum_{i=1}^{n}\log(1-y_i) - \psi\sum_{i=1}^{n}[\log(1-y_i)E_i] + \\ & + a_{\psi-1}\;\; \log(\psi) - b_\psi\psi + C_{\psi_{Beta}},\end{aligned}$$
where the term E_i is specified in the previous Step 5.

7. Sample $\delta_{lt}|\bullet$ via MH. The kernel of the complete conditional posterior is as follows.
$$\begin{aligned}\log p(\delta_{lt}|\bullet) \;=\; & -\textstyle\sum_{i=1}^{n} 1_{\{l_i^*=l\}\{t_i^*=t\}}\log\Gamma(\psi E_i) - \sum_{i=1}^{n} 1_{\{l_i^*=l\}\{t_i^*=t\}}\log\Gamma[\psi(1-E_i)] + \\ & + \psi\textstyle\sum_{i=1}^{n} 1_{\{l_i^*=l\}\{t_i^*=t\}}\log(y_i)E_i - \psi\sum_{i=1}^{n} 1_{\{l_i^*=l\}\{t_i^*=t\}}(1-y_i)E_i + \\ & - [1/(2\sigma^2)][\delta_{lt}^2 - 2\delta_{lt}(\alpha_{l\bullet}\lambda_{\bullet t} + \eta_{lt})] + C_{\delta_{Beta}}.\end{aligned}$$
The term E_i can be found in Step 5.

In the previous formulations, the elements $C_{\beta_{Beta}}$, $C_{\psi_{Beta}}$, and $C_{\delta_{Beta}}$ indicate the corresponding unknown normalizing constant in the log scale.

The reader can find in the present chapter all information regarding the description of the three versions of the GLMM with random effects structured via FA. Section 2.4 is now complete with all the necessary details about the MCMC algorithm developed for these models. We are ready to move to the next stage of the study, where results from applications involving distinct scenarios of artificial data are analyzed.

Chapter 3

Simulation study

In this chapter, a simulation study is developed to evaluate the three versions of the proposed model considering the Normal, Gamma, and Beta distributions for Y_i. Artificial data are generated under similar conditions for the three modeling versions. First, we explore results from a single database for each of the three cases. This analysis is important to validate the MCMC and verify aspects related to convergence. In addition, the behavior of the estimates is verified concerning the true values of the model parameters. After a single data set analysis, we move to a study based on Monte Carlo (MC) replications. In this case, the parameter estimation under the correct model specification is validated in an MC scheme with 50 replications for each scenario. Finally, a sensitivity study is performed for some parameters treated as fixed in the model.

3.1 Analysis based on a single data set

In this section, we analyze the results by fitting one artificial database for each modeling scenario. Details about the required MCMC algorithm to fit the data can be seen in Chapter 2. We emphasize that the chains generated from the algorithm are visually inspected to verify the required convergence to the target posterior distribution. The first step in the simulation study is to describe how the artificial data can be generated based on the structure of the proposed GLMM. This aspect is detailed in the next pages.

Procedure to generate data.

The chosen scenario for the present analysis assumes $L = 100$ regions and $T = 4$ time points. We set $K = 2$ factors and thus define the groups G_1, G_2, and G_E following the comments in Section 2.3. For each factor k, the loadings related to the regions in G_k are obtained from the Uniform$(1, 2)$. This uniform is chosen to ensure positive loadings away from 0, meaning that all regions in G_k are significantly impacted by Factor k with the same sign (direction) to interpret. Loadings related to the group G_E are generated according to the conditional specification given in (2.14). Consider group G_1 and G_2 with 10 regions each, which is inspired by the study in [7]. The remaining regions are allocated in G_E. Table 3.1 summarizes the configuration of the loadings matrix.

The spatial association considers a band diagonal binary neighborhood matrix W_α with 4 neighbors for most regions; see details in [22]. As a result, we obtained $D_\alpha = \text{diag}\{2, 3, 4, 4, \cdots, 4, 4, 3, 2\}$ with dimension 100×100. For the temporal structure, W_λ is such that $W_{\lambda,12} = W_{\lambda,21} = W_{\lambda,23} = W_{\lambda,32} = W_{\lambda,34} = W_{\lambda,43} = 1$, and the remaining elements are 0. This provides $D_\lambda = \text{diag}\{1, 2, 2, 1\}$. These configurations are motivated by the choices in [7].

Parameter	Index	Dimension	True value
α_{lk}	$l \in \{1, \ldots, 10\}$, $K = 1$	10×1	Uniform$(1, 2)$
	$l \in \{1, \ldots, 10\}$, $K = 2$	10×1	$\mathbf{0}$
α_{lk}	$l \in \{11, \ldots, 20\}$, $K = 1$	10×1	$\mathbf{0}$
	$l \in \{11, \ldots, 20\}$, $K = 2$	10×1	Uniform$(1, 2)$
$\alpha_{lk} \in G_E$	$l \in \{21, \ldots, L\}$, $K \in \{1, 2\}$	$(L - 20) \times 2$	See expression (2.14)

Table 3.1: Summary indicating the configuration of α to generate data with $K = 2$. The true values of the loadings related to the group G_E are generated conditional on the values obtained for the groups G_1 and G_2.

Two covariates are considered in this study. Let $X_{i1} = 1$ be the value in the first column of matrix X, for $i = 1, \cdots, n$. This is necessary to accommodate the intercept β_0. In addition, generate a binary covariate $X_{i2} \sim$ Bernoulli(0.5) for the second column of X, and a continuous covariate $X_{i3} \sim$ Uniform$(-1, 1)$ for the third column. In order to generate the errors in ϵ, consider $\sigma^2 = 0.8$. In addition, let $\psi = 2$ for the Normal, Gamma, and Beta distributions. Table 3.2 gives some extra information including the true values of the regression coefficients and the specifications of τ_α and ρ_α.

We set $\rho_\alpha = 0.9$ in the CAR structure defined within the model. One may argue that this choice is not close enough to 1 to impose a strong spatial association in the data. In order to verify whether 0.9 is an appropriate choice for the present study, a sensitivity analysis (results omitted here) was developed using other values in the interval (0.9, 1). The results were quite the same as those when assuming $\rho_\alpha = 0.9$. Consequently, we choose to generate data under this particular setting, which is also adopted in [7]. For

Covariate	Dimension	Value
X	$(N \times 3)$	
$X_{\bullet 1}$	$(N \times 1)$	1
$X_{\bullet 2}$	$(N \times 1)$	Bernoulli(0.5)
$X_{\bullet 3}$	$(N \times 1)$	Uniform(-1,1)
Parameter	**Dimension**	**True value**
ψ	scalar	2.00
β	(1×3)	$(0.50, -1.00, 1.00)$
σ^2	scalar	0.80
τ_α	scalar	4.00
ρ_α	scalar	0.90

Table 3.2: Summary to indicate how the covariates are generated and to show the true values of important parameters in the procedure to generate data.

more information, see Appendix A. When generating the data, the parameters τ_λ and ρ_λ do not need to be specified, as the values in λ are chosen so that $\lambda_{11} > \lambda_{12} > \lambda_{13} > \lambda_{14}$ (Factor 1 with a decreasing pattern) and $\lambda_{21} < \lambda_{22} < \lambda_{23} < \lambda_{24}$ (Factor 2 with an increasing pattern). In this case, consider the true values $\lambda_{1\bullet} = (2.0, 1.5, 0.5, -0.5)$ and $\lambda_{2\bullet} = (-1.0, 1.0, 1.5, 2.0)$.

The true nonlinear interaction is set to be $\eta^* = \lambda_{1\bullet}\lambda_{2\bullet} = (-2.0, 1.5, 0.75, -1.0)$. This is the same specification defined in [23]. We consider 50% of regions (randomly selected) in G_E to be affected by the non-null interaction effect η^*. Increasing this percentage implies that more information will contribute to estimating η^*, which means that the corresponding posterior uncertainty is reduced. We highlight the fact that 50% is a configuration explored in the main reference [7]. Note that if reducing this percentage, the model is basically moving towards the configuration of a standard FA structure without interactions.

Finally, we are ready to generate the response Y_i, for $i = 1, \cdots, n$. The magnitude of n will vary in different scenarios to be explored ahead. Recall that the proposed GLMM considers that each sample unit (or individual) i belongs to a region l and time t, therefore, we must specify the pair (l_i^*, t_i^*) for each i. This is done by randomly generating (uniformly with replacement) l_i^* from $\{1, 2, \cdots, 100\}$ and t_i^* from $\{1, 2, 3, 4\}$. Next, we calculate $\delta = \alpha\lambda + \eta + \epsilon$ and the linear predictor $X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}$. Given the linear predictor, one must consider the link function $g(\bullet)$ to obtain the mean parameter θ_i for each i; see the link functions in (2.1), (2.3), and (2.5). The response is then generated using the chosen distribution (Normal, Gamma or Beta) with parameters θ_i and ψ.

The description of the procedure to generate data is now complete. The next part of this section is dedicated to presenting some aspects related to the configuration of the MCMC algorithm that will be used to fit the proposed models to the artificial data sets.

MCMC configuration.

In terms of notation, a parameter name with the superscript "(0)" indicates the MCMC starting value for that parameter. We consider the following initial values. Set $\beta^{(0)} = (0, 0, 0)^\top$, $\sigma^{2(0)} = 1$, $\tau_\alpha^{(0)} = 1$, and $\psi^{(0)} = 3$. Generate values in $\alpha^{(0)}$ from the $U(-0.1, 0.1)$. Values in $\lambda^{(0)}$ are obtained from the $U(-2, 2)$ (Row 1 in ascending order, Row 2 in descending order); if $K > 2$, a random order will be used for any row $k > 2$. For $\eta^{(0)}$, the initial values is a matrix $(L \times T)$ with all entries with zero values. Consider $\delta^{(0)} = \alpha^{(0)}\lambda^{(0)} + \eta^{(0)}$. Let $Z^{(0)} = \mathbf{0}$ is a $L \times 1$ vector. Finally, set $p_l = 0.5$ for $l \in \partial_{\emptyset_k}$ (otherwise, $p_l = 0.01$).

In order to describe our initial uncertainty about the parameters, consider the following prior specifications. Set $\beta \sim \mathrm{N}_q(\mu, \Sigma)$, with $\mu = \mathbf{0}$ and $\Sigma = 10\ \mathbf{I}_q$. In addition, $\psi \sim \mathrm{Gamma}(0.1, 0.1)$, $\sigma^2 \sim \mathrm{IG}(2.1, 1.1)$, and $\tau_\alpha \sim \mathrm{IG}(2.1, 1.1)$. Note that the previous priors are defined with variance 10 to suggest a weakly informative scenario. We also consider $p_l \sim \mathrm{Beta}(1, 1)$ (the uniform) for $l \in \partial_{\emptyset_k}$. For $l \notin \partial_{\emptyset_k}$, let $p_l \sim \mathrm{Beta}(1, 100)$ concentrating the probability mass near 0. This highly informative option is used to induce a null interaction effect.

Assuming lag 1 for all MCMC chains, the Normal and Gamma models required 20,000 iterations; burn-in = 10,000 and posterior sample size = 10,000 iterations. In the Beta model, slow convergence was observed in the chains, thus this particular case required 60,000 interactions; burn-in = 50,000 iterations and posterior sample size 10,000. Again, we highlight the fact that the parameters involved in MH steps were properly tuned to establish acceptance rates as recommended in the literature [30]; Section 2.4 identifies these parameters and other details.

3.1.1 Results

The discussion presented here is related to the results from the analysis using a single artificial data set. Table 3.3 shows the posterior estimates for β, ψ, σ^2, and τ_α in all three models. Note that all parameters are within the corresponding 95% HPD (Highest Posterior Density) interval. In contrast with the study in [7], the dispersion parameter ψ is explored here due to its presence in the chosen continuous distribution adopted for the response.

Table 3.3 indicates, for all three models, that τ_α is the parameter having the largest posterior uncertainty (large SD). The intercept β_0 and the variance σ^2 also have larges SD when compared with β_1, β_2 and ψ. Among the three regression coefficients, and for all three models, the intercept β_0 is the one showing the largest distance to the

true value. The present analysis indicates that estimating β_0 is not an easy task in the context of the proposed models. This difficulty is perhaps explained by some sort of identification problem existing between β_0 and the random effects in δ. Results suggest that the model is able to increase (decrease) the estimate of β_0 and accordingly decrease (increase) the mean of the effects in δ. The authors in [7] detected this type of problem for the Poisson GLMM and they proposed a correction. In brief, the solution shown for them is obtained by setting the true value as the initial value of the MCMC to β_0 in the case of the simulation study. In an application to real data, it was recommended, first, the ordering of the estimated values of δ in a graph. Afterward, check the distance from the median δ_{ij} to 0 on the graph and apply this value to $\beta_0^{(0)}$. Fortunately, the mentioned problem is not too critical when comparing the three models in this dissertation and the Poisson case from the main reference. Some tests were performed by generating data with different values (negative and positive) of β_0. The result indicates that, regardless of the true β_0, the proposed GLMM tends to overestimate this intercept and underestimate the random effects in δ. The MC study developed ahead will show further details about this phenomenon. Our recommendation is that the analyst must be careful when interpreting β_0, since it may be overestimated. It is important to highlight the fact that our findings indicate that other coefficients (β_1 and β_2) and the parameter ψ are all well-estimated and not affected by the problem. For the parameters, the medians are close to the true values, indicating that the posterior distributions are symmetric.

		Normal			Gamma			Beta		
	True	Mean	SD	HPD	Mean	SD	HPD	Mean	SD	HPD
β_0	0.50	0.69	0.17	[0.35, 1.03]	0.74	0.15	[0.45, 1.01]	0.61	0.12	[0.42, 0.86]
β_1	-1.00	-1.05	0.05	[-1.15, -0.96]	-1.03	0.03	[-1.08, -0.98]	-0.99	0.01	[-1.01, -0.97]
β_2	1.00	0.97	0.04	[0.88, 1.06]	0.98	0.02	[0.94, 1.03]	0.99	0.01	[0.97, 1.00]
ψ	2.00	2.04	0.05	[1.94, 2.14]	1.98	0.05	[1.89, 2.07]	2.07	0.04	[1.98, 2.09]
σ^2	0.80	0.61	0.10	[0.44, 0.82]	0.97	0.11	[0.77, 1.21]	0.78	0.09	[0.62, 0.97]
τ_α	4.00	2.93	0.94	[1.25, 4.73]	2.82	1.18	[1.00, 5.13]	3.00	1.39	[1.01, 5.78]

Table 3.3: Posterior mean, standard deviation (SD) and 95% HPD interval for the regression coefficients in β, the dispersion parameter ψ, and the variances σ^2 and τ_α. The true values are reported in the first column.

Figure 3.1 shows the posterior estimates for all α_{lk}, λ_{kt}, and δ_{lt}. In particular, the loadings and random effects are ordered by true values to facilitate analysis and allow comparison. As can be seen, the Gamma and Beta cases have loadings in α slightly overestimated for $\alpha_{lk} < 0$ and underestimated for $\alpha_{lk} > 0$; see the misalignment between black and red points. In Panels $(a - c)$, short HPD centered around 0 can be found in the graphs. They are related to the locations in G_1 and G_2 for which we assume $\alpha_{lk} = 0$. Recall from Section 2.3 that this assumption is necessary to avoid a model identifiability issue. All graphs show that most parameters have the corresponding true value within the

95% credibility intervals. Panels $(g-i)$ indicate that all random effects seem to be well-estimated. If we take a close look comparing the posterior means (black points) and true values (red points), it is possible to detect a slight underestimation (not severe) of δ_{lt}'s for all three models. Another aspect to be emphasized is related to the amplitude observed for the HPD intervals of δ_{lt} in the Normal (largest amplitude), Gamma (intermediate), and Beta (shortest) models. One can conclude that the posterior uncertainty, about the random effects, differs between the models. Panels $(d-f)$ show that the increasing and decreasing patterns assumed for the true factor scores are well-captured by the models. In Appendix B it is possible to see more results referring to the matrices α, λ and δ.

Figure 3.1: HPD intervals (95%) for α_{lk} (Panels $a-c$), λ_{kt} (Panels $d-f$), δ_{lt} (Panels $g-i$). Each column of panels is related to one model. Black points indicate the posterior mean and the red points represent the true values. Graphs for α and δ_{lt} are ordered with respect to the true values to improve the visual analysis.

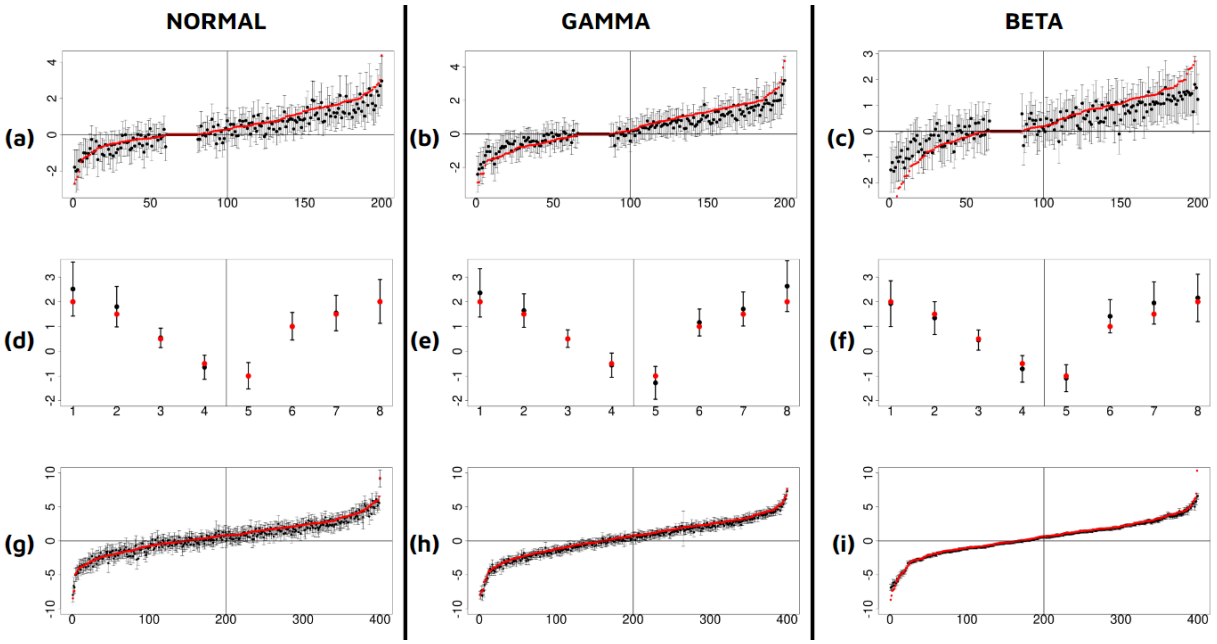
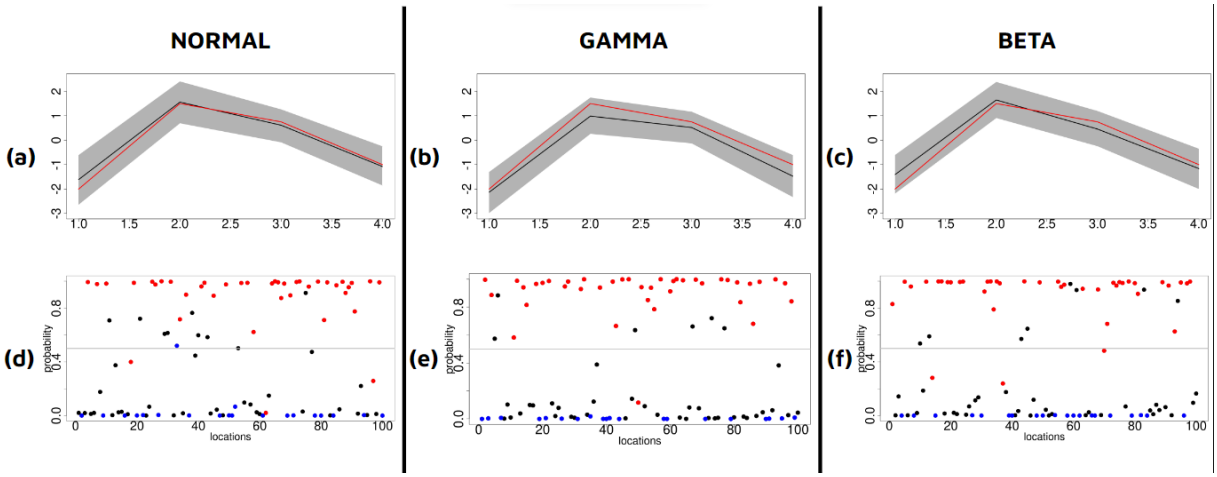


Figure 3.2 presents results related to the nonlinear interaction η^*. Panels $(a-c)$ indicate that the estimated η^* (posterior mean = black line) managed to capture the true pattern (red line) for all three models. The 95% HPD intervals (shaded areas) include all true values. Panels $(d-f)$ show the estimated probabilities that each location is affected by the non-null interaction η^*. Colors of the points are established in accordance with the data generation. The blue represents a region (in G_1 or G_2) known to be unaffected by the interaction. The red suggests a region (in G_E) generated with the interaction effect. Black indicates a region (in G_E) generated without interaction. Note that most red points are located above the level 0.5, which suggests that the models are able to detect the impact of the interaction effect for most regions. Similarly, many black points are located below 0.5 indicating that unaffected regions are also well-identified.

Figure 3.2: Panels $(a-c)$ show the posterior mean (black line), 95% HPD interval (shaded area), and true value (red line) of the interaction η^*. Panels $(d-f)$ present the probabilities (points) that the regions are affected by the nonlinear interaction. Blue indicates locations in G_1 and G_2 (without interaction), red represents locations in G_E affected by the interaction, and black denotes locations in G_E unaffected by the interaction.



The next stage of the analysis is to evaluate the model when increasing the sample size, but L and T are kept the same. The sample size in [7] is motivated by their real application, related to electrocardiogram data, where each region has 35 observations on average. In short, the authors assume $n = 35L$, with $L = 400$. We follow the steps of the main reference and set $n = P \times (35L)$ in our study. The term P is defined to calibrate the sample size. The case $P = 1$ indicates 35 observations per region in the artificial data, which is the configuration [7]. We increase the sample size by setting a second scenario with $P = 2$, meaning 70 observations per region. The goal of the analysis is to compare these two situations.

When fitting the generated data sets, with $P = 1$ and $P = 2$, the inference results show strong similarities between the point estimates from $P = 1$ and $P = 2$ for β_1, β_2, τ_α, σ^2, ψ, and η^*. As expected, increasing the sample size implies that shorter HPD intervals are obtained. When evaluating the results for β_0, a significant improvement is observed for $P = 2$, as shown in Table 3.4. Note that when for a larger n $(P = 2)$ the posterior mean is closer to the true $\beta_0 = 0.5$ and the posterior uncertainty is smaller (small SD and shorter HPD). The true values can be found within the 95% HPD for all cases. In addition to improving the estimation of β_0, we observed an improvement in the estimates of random effects in δ. It seems reasonable to conclude that increasing the sample size will alleviate the existing identification problem between β_0 and δ.

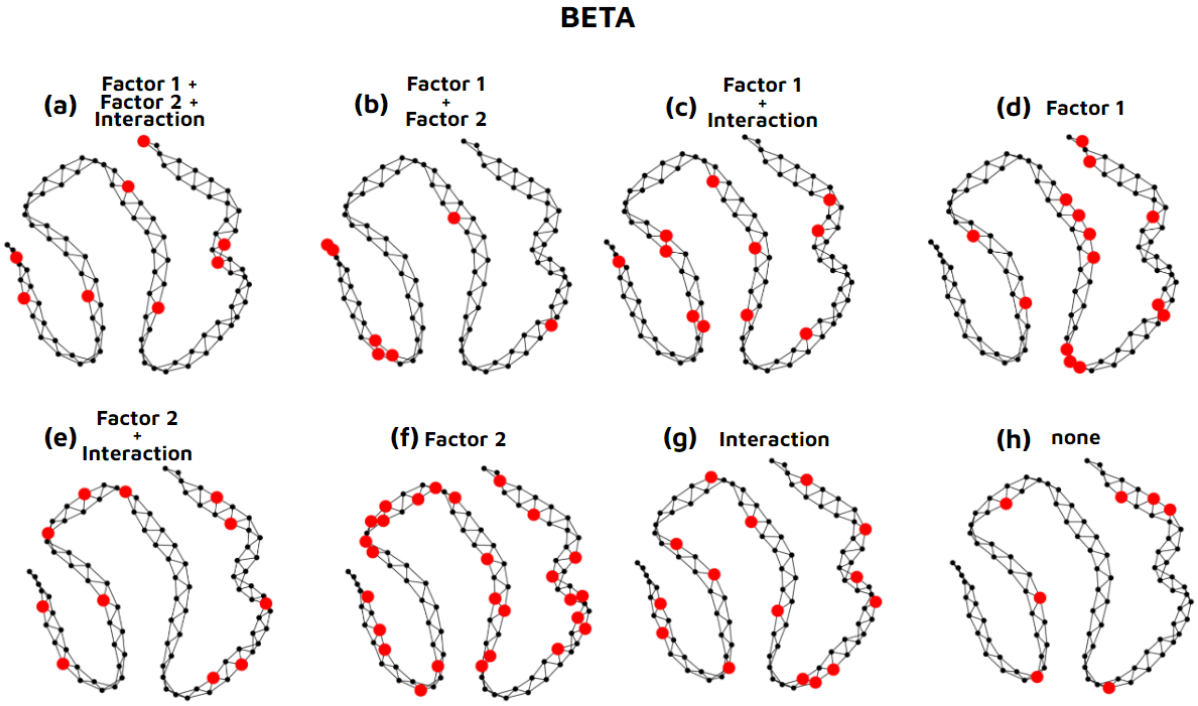
Figure 3.3 shows a generic graph scheme representing the spatial structure of the artificial data with 4 neighbors per region. The analysis is focused on the Beta model with n specified by using $P = 2$; similar conclusions are found for the Normal and Gamma cases. The points (vertices) indicate the regions and the lines connecting points (edges) identify

	P	Mean	SD	HPD
Normal	1	0.69	0.17	[0.35, 1.03]
	2	0.65	0.13	[0.40, 0.92]
Gamma	1	0.74	0.15	[0.45, 1.00]
	2	0.63	0.07	[0.47, 0.80]
Beta	1	0.60	0.11	[0.48, 0.86]
	2	0.58	0.07	[0.46, 0.77]

Table 3.4: Comparing scenarios with different sample sizes $n = P \times (35L)$, where $L = 100$, and P can be 1 or 2. Posterior estimates (mean, standard deviation, and 95% HPD interval) of β_0 for all three models. The true value is $\beta_0 = 0.5$.

the neighborhood. In any panel, the group of red points forms a cluster containing regions affected by the same combination of main factors and interaction. We reinforce to the reader that "cluster" refers to a group of regions, not necessarily adjacent or contiguous. As an example, Panel (a) indicates the regions affected by both main factors (1 and 2) and the non-null interaction. In addition, Panel (b) identifies the regions affected by both main factors and having a null interaction. Panel (h) shows the cluster of regions without main effects or interaction.

Figure 3.3: Generic graph scheme, displaying 4 neighbors per region, mimicking the spatial structure of the artificial data. The points (vertices) represent locations the lines (edges) indicate neighborhood. The red color identifies the so-called "clusters" formed by locations affected by the same combination of main effects and interaction effects (null or non-null). Here, assume the scenario $Y_i \sim \text{Beta}(\theta_i, \psi)$ and $P = 2$.



In this dissertation, a region l is said to be affected by η^* when $p^*(z_l = 1|\bullet) >$

0.5; see Step 3 in the algorithm presented in Section 2.4. Region l is considered to be significantly affected by Factor k, when the value 0 is outside the 90% HPD interval for α_{lk}. The level 90% is chosen here to allow the identification of cluster configurations with more than one region in Figure 3.3 $(a-g)$ and a few regions in Panel (h). Larger levels determine larger HPD intervals for the loadings, which tend to include 0 for almost all regions. The reader must note that a level $> 90\%$ can be chosen for larger n. Increasing the sample size will provide shorter HPD intervals that do not include 0.

The analyses involving a single data set to verify the performance of the proposed GLMMs are now complete. Recall that [7] shows similar results for the Bernoulli and Poisson regressions. In the next section, we develop a Monte Carlo simulation study, with replications, to provide broader conclusions about the behavior of the three proposed models.

3.2 Monte Carlo study

This section is dedicated to exploring results from an MC study based on 50 artificial data sets created under similar conditions. The MC scheme accounts for a similar scenario considered for the single data set analyzed in the previous section. We evaluate all three continuous distributions proposed for the response variable. The analyses consider the so-called Relative Bias (RB) to assess the quality of the estimates. This measurement is also adopted in the main reference [7]. In terms of formulation, we have $\text{RB}(\zeta) = 100(\hat{\zeta} - \zeta_{true})/|\zeta_{true}|$, where ζ is a generic parameter (scalar), $\hat{\zeta}$ is the posterior estimate, and ζ_{true} is the true value chosen to generate the data. Note that the RB indicates how large is the difference between the estimate and true value relative to the magnitude of the true value. If $RB(\zeta) > 0$, then the parameter is being overestimated. In contrast, $RB(\zeta) < 0$ indicates that the target parameter is underestimated.

For all replicates, we set $L = 100$ regions, $T = 4$ time points, $n = 35L$, 4 neighbors per region for most regions, $K = 2$ latent factors, 50% of the regions in G_E are affected by the interaction between the two factors. Consider the matrix λ specified in the previous section for all cases and replicas. The non-null interaction is given by $\eta^* = \lambda_{1\bullet}\lambda_{2\bullet}$. In addition, consider G_1 and G_2 with 10 regions each, and G_E has 80 regions. Again, we set ψ, β, σ^2, τ_α, and ρ_α as indicated in Table 3.2. The same distributions - Bernoulli(0.5) and Uniform$(-1, 1)$ - are used to generate the two covariates, but these covariates are not the same for each replica. We did not fix the covariates since the generation of covariates is controlled in each replica of the MCMC, making the results approximately similar for each database generated. A different error matrix is generated for each replication, assuming

$\epsilon_{lt} \sim N(0, \sigma^2)$ independently for all l and t. We also generate different loadings matrices for each MC replication (see Table 3.1). This is done to avoid exploring the same (but similar) spatial structure in the MC scheme. In order to explore the RB for the loadings in α, only the loadings related to the group G_E are considered. This strategy is used because the loads present in the restriction groups G_1 related to Factor 2 and G_2 related to Factor 1 are forced equal to zero, making the calculation of the relative bias inapplicable in these cases. The location l and time t is randomly generated for each individual establishing the $n \times 1$ vectors l^* and t^*. Finally, we calculate θ_i using (2.1), (2.3) or (2.5) and generate the target response Y_i for $i = 1, \cdots, n$.

The reason for not fixing ϵ, α, and the values of covariates in the MC scheme is that small changes in these elements cause, a few times, considerable changes in the quality of the estimates. Therefore, by fixing these quantities one could end up using specifications that promote only good results for all MC replications. Generating these elements (under the same conditions) for each MC sample is a strategy to avoid misleading conclusions.

Figure 3.4: Boxplots displaying the 50 Relative Biases obtained for each parameter in the MC study. Panels $(a - c)$ show RB(η_t^*), Panels $(d - f)$ present RB(β_j), Panels $(g - i)$ are related to ψ, σ^2 and τ_α.

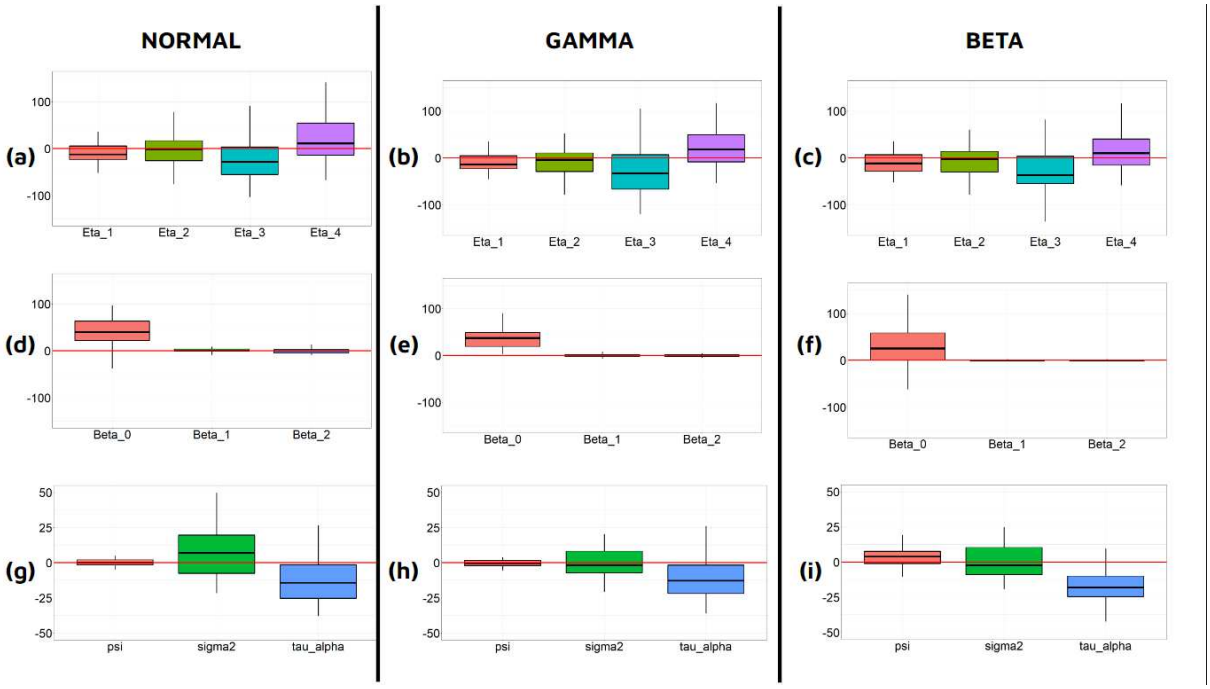
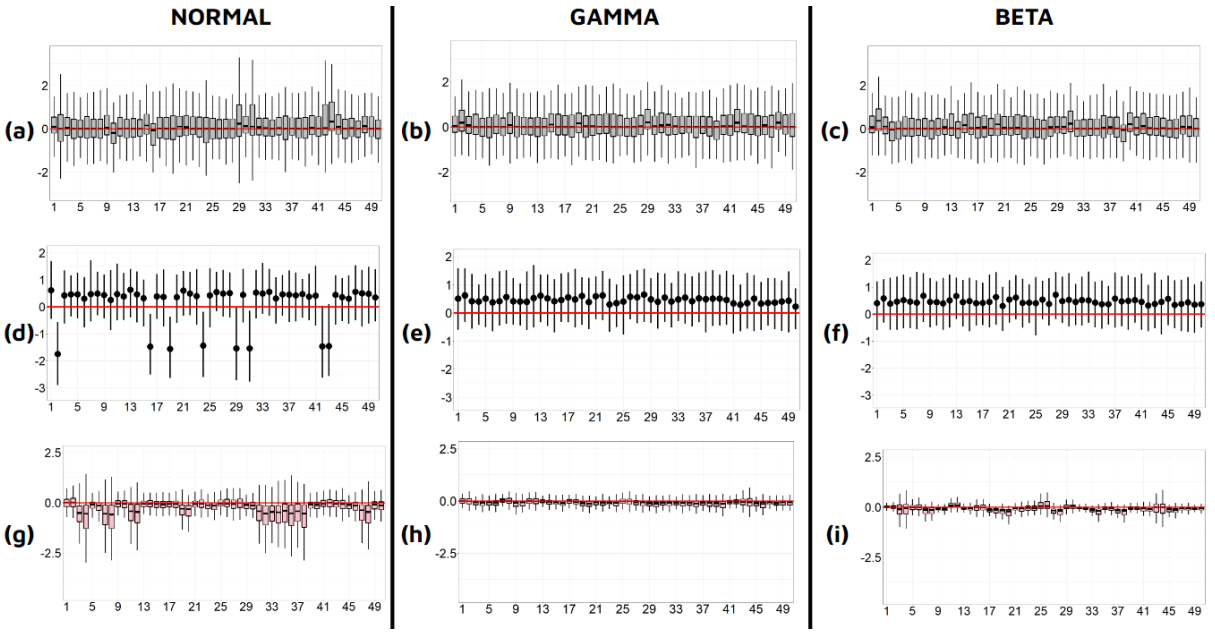


Figure 3.4 shows the RBs for elements for η_1^* to η_4^* (Panels $a - c$), coefficients in β (Panels $d - f$), and the parameters related to variability ψ, σ^2 and τ_α (Panels $g - i$). The boxplot allows us to see the median and the MC dispersion. For all models (Normal, Gamma, and Beta), the results can be considered similar. In the graphs related to η^*, we see that the median of η_4^* is above zero and with a greater interquartile range, leading to the conclusion of overestimation and greater posterior uncertainty. The boxplot of η_3^*

also has a large interquartile range, but with a median below zero (underestimation). For η_1^* and η_2^*, their interquartile ranges are smaller, indicating lower posterior variability. In addition, one can see medians closer to zero, suggesting good estimates.

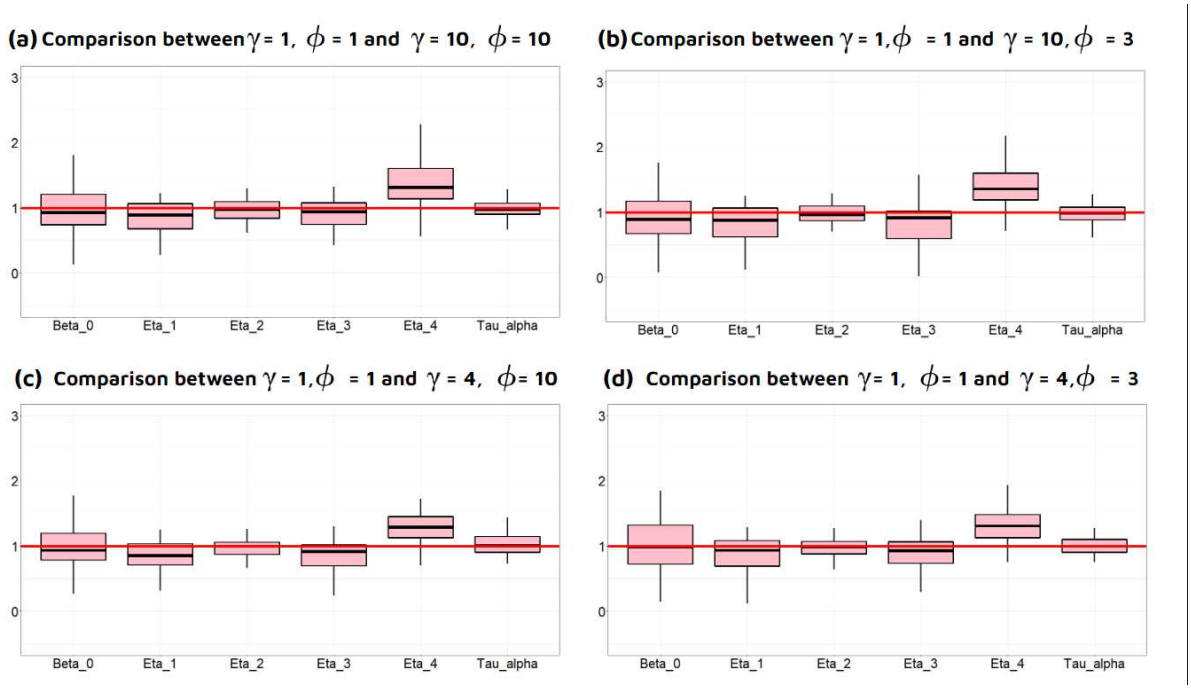
Figure 3.5: Graphs displaying the Relative Biases obtained for some parameters. Panels $(a-c)$ and $(g-i)$ show 50 boxplots (one for each MC replica) summarizing the RBs from the whole matrices α and δ, respectively. Panels $(d-f)$ show 50 bar plots summarizing the RBs from λ for each replica; λ is small 2×4, thus the bar plot provides better visualization. Black points represent the mean and the range of the bars indicates the minimum and maximum RBs.



Regarding the regression coefficients, Figure 3.4 $(d-f)$ indicates that the intercept is overestimated for all cases. The MC variability is also larger for β_0 than for β_1 and β_2. The coefficients β_1 and β_2 have a median quite close to zero and a small interquartile range. Panels $(g-i)$ present results regarding the variability parameters. As can be seen, τ_α is underestimated and has greater MC variability for all models. The parameter σ^2 is underestimated and has larger MC variability in the Normal model. In the Beta and Gamma cases, the median of σ^2 is closer to zero. Finally, ψ has low MC variability and a median near zero, especially for the Normal and Gamma models. Some slight overestimation is detected for ψ in the Beta case. Please note that ψ is usually seen as a nuisance parameter in GLMM, and the present dissertation is the first study fitting the proposed methodology in the context of continuous distributions indexed by ψ.

Figure 3.5 presents a boxplot for each MC replica. Each graph expresses the behavior of the RBs within α (group G_E) in Panels $(a-c)$, and the RBs from δ in Panels $(g-i)$. Given the small dimension (2×4) of λ, we choose to present its result in terms of a bar plot (black point = mean) ranging from min. to max. RB in λ; see Panels $(d-f)$. As can be seen, results for α show that most of the medians are near zero (small interquartile

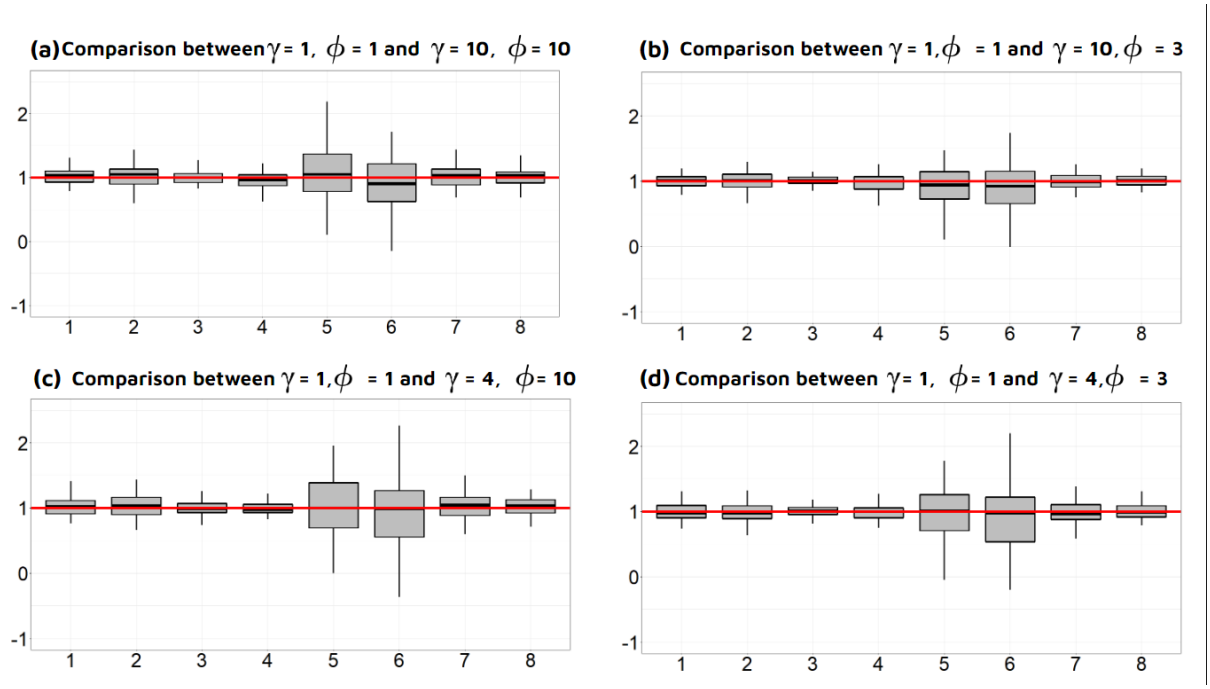
Figure 3.6: Sensitivity analysis. The boxplots summarize the 50 ratios of RBs obtained from two scenarios. For each MC replica, the numerator has the RB assuming $\gamma = 1$ and $\phi = 1$, and the denominator has the RB from the other configuration (see above each panel). These results are for β_0, η_t^*, and τ_α.



range) suggesting good estimates for the loadings. Now for the matrix λ, one can see that most averages (black points) are above zero for the Normal case, and all averages are above zero for the Gamma and Beta cases. We can conclude that the scores tend to be slightly overestimated, and the amplitudes of the bar plots are similar for all cases. Finally, the analysis for δ shows that the interquartile range is quite small (around 0) for the Gamma and Beta cases, indicating good estimates. Some MC samples have a median slightly below zero, especially for the Normal case, which indicates underestimation and provides evidence for the hypothesis discussed in Section 3.1 about identification issues involving δ and β_0.

As mentioned in Chapter 2, ϕ and γ are fixed. A sensitivity study is conducted to choose the best value for these parameters. Here, the case ($\gamma = 1$, $\phi = 1$) was compared with options: ($\gamma = 10$, $\phi = 10$), ($\gamma = 10$, $\phi = 3$), ($\gamma = 4$, $\phi = 10$), and ($\gamma = 4$, $\phi = 3$). These choices are also explored in [7]. Furthermore, values of ϕ less than 1 were not chosen since, according to [23], this would cause the model to estimate the effect of non-linear interaction between the factors in a flatter way, not this case being interesting for this work. The RBs of parameters whose true value is the same for all MC replicas are the focus of the analysis; these parameters are β, η^*, ψ, σ^2, τ_α and λ. We calculate the ratio between the RB from fitting ($\gamma = 1$, $\phi = 1$) divided by the RB from one of the other mentioned configurations. Note that a ratio below 1 indicates lower RB from ($\gamma = 1$,

Figure 3.7: Sensitivity analysis. The boxplots summarize the 50 ratios of RBs obtained from two scenarios. For each MC replica, the numerator has the RB assuming $\gamma = 1$ and $\phi = 1$, and the denominator has the RB from the other configuration (see above each panel). These results are for the factor scores (Boxplots $1-4$ = Factor 1, Boxplots $5-8$ = Factor 2).



$\phi = 1$) compared to the other configuration. The sensitivity analysis in Figure 3.6 shows most boxplots with a median near 1, suggesting a similarity between the RB from ($\gamma = 1$, $\phi = 1$) and the other case under investigation. Despite being close to 1, some medians are slightly below 1, which indicates that ($\phi = 1$, $\gamma = 1$) tend to provide lower RB in this MC study. Note that η_4 is the only parameter having a boxplot with a median above the level 1. We can conclude that most parameters give support to choose ($\phi = 1$, $\gamma = 1$). The graphs of RBs ratios for β_1, β_2, ψ, and σ^2 are not presented here due to the fact that their medians are quite close to 1 and a small interquartile range is observed (making their scale difficult to see in Figure 3.6).

Figure 3.7 $(a-d)$ presents the result from a sensitivity analysis comparing ($\phi = 1$, $\gamma = 1$) and other configurations of (γ, ϕ) for each score λ_{kt}. Recall that matrix λ is small with dimension 2×4. The scores of Factor 1 are related to the values $1-4$ on the horizontal axis. The scores of Factor 2 are related to the values $5-8$. Note that most medians of boxplots are close to 1, and the interquartile range is small between $(0.5, 1.5)$. The scores λ_{21} (fifth graph) and λ_{22} (sixth graph) tend to have a larger interquartile range in all panels (their median is close to 1). The results displayed in Figure 3.7 suggest a similarity between the compared configurations, thus the analysis here also supports the choice ($\phi = 1$, $\gamma = 1$). We highlight the fact that such sensitivity study was not developed

for α and δ because these matrices are not the same for each MC replica (real values are not the same to compute the RBs).

3.3 Conclusions of the simulation study

In the first part of Chapter 3, a single data set was analyzed assuming values of parameters and hyperparameters described under three different distributions for the response variable. Convergence was achieved in the MCMC for all inspected chains. The main conclusion is that all models provide good estimates when the correct model specification is chosen for the analysis. More specifically, the Beta model showed satisfactory results for all parameters, and the random effect matrix δ was well-estimated with short 95% HPD intervals. In addition, the parameter ψ, not considered in [7], is also well-estimated for the three models. As expected, an improvement in terms of estimation was observed for all parameters when increasing the sample n.

The second part of Chapter 3 is dedicated to an MC study with 50 replications. In general, the good inference results based on a single data set were confirmed here. In the comparison β_0 vs. δ, results seem to support the hypothesis of an identification problem between these parameters. An overestimation of β_0 can be connected with an underestimation of δ, and vice versa. Finally, a sensitivity analysis was carried out, where the objective was to choose (ϕ, γ) defined in the spatial structure of the model. The conclusion is that for the majority of the explored parameters, the configuration $(\phi = 1, \gamma = 1)$ is an appropriate choice in terms of Relative Bias (it also provides lower computational cost).

The Beta model will be explored in the next chapter in a real data application. The database refers to the ENEM (*Exame Nacional do Ensino Médio*), which is an exam taken by students in Brazil, and it is widely used by Brazilian public universities in their selection procedure of new students. This database was never explored using the GLMM methodology discussed in this work.

Chapter 4

Application

In this chapter, the goal is to show a real illustration of the proposed Beta model. The Beta case is chosen due to the bounded continuous nature of the dependent variable. The data set is related to the exam ENEM (*Exame Nacional do Ensino Médio*) being freely available for download from the website[1] of INEP (*Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira*). INEP is an agency, connected to the Brazilian Ministry of Education, in charge of evaluating educational systems and the quality of education in Brazil.

The ENEM was created in 1998 as a tool to evaluate the performance of students at the end of high school. In 2009, the exam increased its importance and became broader because it started to be used in the selection procedure to have access to Brazilian higher education institutions through the *Sistema de Seleção Unificada* (Sisu) and the *Programa Universidade para Todos* (ProUni). The ENEM is also accepted in more than 50 Portuguese institutions. ENEM participants can also apply for financial support from the *Fundo de Financiamento Estudantil* (Fies) to pay some percentage of the fee to study in a private university. The ENEM data allow researchers to develop studies accounting for educational indicators.

Anyone who has completed high school or is about to finish this stage can take the ENEM to access higher education. Participants who have not yet completed high school can only participate as "trainers", and their ENEM grades is only used for self-assessment of knowledge. In the present study, we consider only participants who are not in the category "trainers". The ENEM requires two days to be applied, and participants respond to 180 questions in four areas of knowledge: languages, codes, and technologies; human sciences and technologies; natural sciences and technologies; and mathematics and technologies. Participants are also evaluated through an essay, which requires the development of an argumentative dissertation text based on a proposed topic.

[1] `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos`

4.1 About the database

In this work, we consider data from individuals taking the ENEM in the Brazilian state of Minas Gerais. We also require the following features to select individuals: the student cannot be a "trainer", aged $17-22$ years, the participant indicated his/her type of high school (public or private), the student had the "regular" type of education, the student already completed high school or will finish it in the same year of the exam, the student was present in both examination days, and the student took ENEM between $2015-2021$. The reason for choosing these years is the fact that this period has more similarity in terms of organization and difficulty of the exam. As an example, tests applied between $2009-2014$ have more interpretative questions, while tests between $2015-2021$ have questions that require more knowledge of content [32]. The number of regions considered was $L = 188$ municipalities in Minas Gerais. The ENEM is not applied in all municipalities of the state, then some students must go to the nearest location where the exam can be taken. Two municipalities (Açucena and Extrema) do not have data for all years in the selected period, thus they are removed from the analysis. For all locations and time points, we have at least one student in the data set.

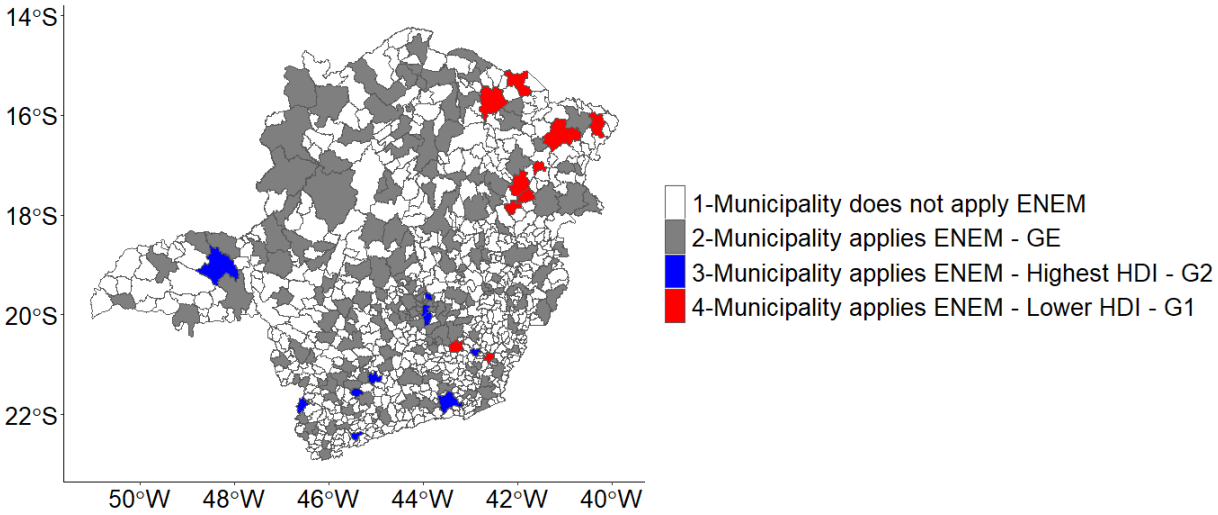
The ENEM data does not provide a longitudinal study where a subject has observations for each year in the period $2015-2021$. Note also that an individual may take the ENEM more than once in the period of study. However, according to the technical report released by INEP, these cases cannot be identified based on the available data. Each individual has an identification number in the database, but the same number is not allocated to the same individual when he/she takes the exam in different years. Given the impossibility to find such individuals, we basically do not impose any association between observations from the same individual in distinct years.

The covariates chosen for the study are: gender (X_{1i}: 0 = Male, 1 = Female), race (X_{2i}: 0 = white or Asiatic, 1 = non-white), and essay writing score (X_{3i}: scaled to the interval between $0-1$). After filtering the database with the required information, the final data set is composed of 701,930 observations. Fitting the proposed Beta GLMM to this large data set has a high computational cost justified by the large number of parameters to be estimated, especially in the FA structure of the model. Problems related to computer memory can appear when trying to run the corresponding MCMC in an ordinary computer, which makes it infeasible to obtain results for this dissertation. As an alternative to circumvent this difficulty, we randomly extract (with replacement) three smaller databases from the large one. The extracted data sets are as follows: Data set 1 has 15,000 observations, Data set 2 has 20,000 observations, and Data set 3 has 30,000 observations. The next results will be focused on the Data set 3. The Data sets 1 and 2 will be used in a comparison study with 3 to validate and verify whether conclusions are

similar. The response variable, considered in the present real application, is determined based on a linear combination of the scores obtained by the individual in the different areas of knowledge. This linear combination is given by the first principal component of the four scores; see Principal Component Analysis (PCA) in [17].

Regarding the binary neighborhood matrix W_α, note that its dimension is 186×186. In order to determine the neighborhood, we first obtain the Euclidean distance $d_{l_1 l_2}$ between all pairs of municipalities (l_1, l_2). Next, a threshold d^* is chosen and two locations (l_1, l_2) are considered neighbors if $d_{l_1 l_2} < d^*$. The value of d^* must ensure that every location has at least one neighbor. Subsequently, those municipalities with less than 6 neighbors remained with their number of neighbors, but locations with > 6 neighbors were truncated to have 6 neighbors. In this case, we select the nearest municipalities, ensure that W_α is symmetric, and verify that if the municipality l_1 is a neighbor of l_2, then l_2 is a neighbor of l_1 for this application. Note that the largest number in the diagonal of D_α is 6, and the smallest value is 1. Other choices (between 4 and 10) for the maximum number of neighbors were tested in the present study, but the results are quite similar to those reported here for the case 6. Values < 4 or > 10 led to different conclusions, however, these cases represent situations where the spatial structure has (i) few neighbors for all regions or (ii) regions with too many neighbors, including municipalities far from each other. The chosen configuration (max. 6 neighbors) has also the advantage of providing a reasonable computational cost to run the MCMC.

Figure 4.1: Spatial arrangement of the municipalities with respect to the auxiliary variable "HDI Income". The map identifies the groups G_1 in red, G_2 in blue, and G_E in gray.



As explained in Chapter 2, partitioning the set of regions into groups is a necessary strategy to guarantee the identifiability of the FA part of the model. Each group is supposed to have a strong association with one of the factors, except G_E, which is free to be affected or not by any factor. The researcher may choose different auxiliary variables

to build the restriction groups. In line with [7], the present analysis considers the HDI (Human Development Index) of the municipalities in Minas Gerais for this purpose. We used the 2010 HDI Income, which can be downloaded from the Brazilian Atlas of Human Development website[2]. Let $K = 2$ latent factors. The group G_1 is set to include the 10 regions with the lower HDI Income. In contrast, the group G_2 contains the 10 regions with the highest HDI Income. Figure 4.1 shows the positions of the municipalities considered in the analysis. The white color indicates the municipalities that do not apply the ENEM. The gray color represents the regions in G_E. The red color identifies G_1 (low HDI). Finally, the blue color indicates the members of G_2 (high HDI). As can be seen, most regions in G_1 are located in the Northeast part of Minas Gerais. On the other hand, regions in G_2 are found in the West, Central and South parts of the State.

This real application has a bounded continuous response variable for which we assume the Beta model defined in the present dissertation. The FA structure for the random effect matrix δ allows us to explore underlying information that cannot be captured by an ordinary Beta regression. The proposed model will provide the factor scores representing the patterns related to municipalities with low and high HDI. This will give an idea of how the average ENEM grade behaved across the years for these two groups. In addition, the method will establish clusters of municipalities affected by the same combination of factors and/or interaction. The next section shows the results.

4.2 Analyses via the Beta model

This section evaluates the results of fitting the Beta model to the real data. The MCMC setup (iterations, burn-in, and starting values) and prior distributions applied here were the same ones considered in Section 3.1, subsection "MCMC configuration". In line with the sensitivity analysis developed in Chapter 3, we set $\phi = 1$ and $\gamma = 1$ in the present real application. Table 4.1 shows the posterior means and standard deviations (HPD intervals) related to the coefficients in β, σ^2, τ_α and ψ for Data sets 1, 2, and 3. As can be seen, the posterior variance is close to zero for most parameters, indicating low posterior uncertainty. In addition, when comparing the three data sets, it is clear that the posterior uncertainty (see the HPD intervals) decreases when increasing the sample size. In addition, our findings confirmed that all parameters have very similar estimates for all three data sets by finding that most averages estimates are within the credibility intervals on the three datasets. Regardless of the sample size defined in the three data sets, the model provides the same conclusions. This result indicates that the extracted

[2] `http://www.atlasbrasil.org.br/2013/pt/ranking`

sub-samples are coherent in terms of inference.

In order to simplify the discussion and avoid repeating the conclusions, the analysis will be focused on Data set 3, which is the one having the largest sample size (30,000). All parameters indicated in Table 4.1 are significant (95% HPD intervals do not include 0). Note that $\beta_1 = -0.17$, therefore, the ENEM grade (1^{st} PCA) seems to be lower for women. In addition, $\beta_2 = -0.08$ indicates that non-white students tend to have lower ENEM grades. Finally, $\beta_3 = 1.90$ suggests, as expected, that the ENEM grade increases, when the essay score increases. Naturally, one must consider all other covariates and effects fixed in order to separately measure the magnitude of these impacts. Among all parameters listed in Table 4.1, the dispersion ψ is the one with the highest posterior uncertainty (large SD).

	Data set 1		Data set 2		Data set 3	
	Mean	HPD	Mean	HPD	Mean	HPD
β_0	-0.92	[-0.88, -0.97]	-0.77	[-0.80, -0.74]	-0.82	[-0.86, -0.79]
β_1	-0.18	[-0.19, -0.17]	-0.18	[-0.19, -0.17]	-0.17	[-0.17, -0.18]
β_2	-0.10	[-0.11, -0.08]	-0.09	[-0.10, -0.08]	-0.08	[-0.10, -0.08]
β_3	2.07	[1.91, 2.10]	1.97	[1.94, 2.08]	1.90	[1.89, 1.97]
ψ	29.07	[28.41, 34.74]	31.47	[28.64, 34.11]	33.55	[28.99, 34.06]
σ^2	0.01	[0.01, 0.01]	0.01	[0.01, 0.01]	0.01	[0.01, 0.01]
τ_α	0.20	[0.07, 0.34]	0.19	[0.08, 0.34]	0.20	[0.07, 0.33]

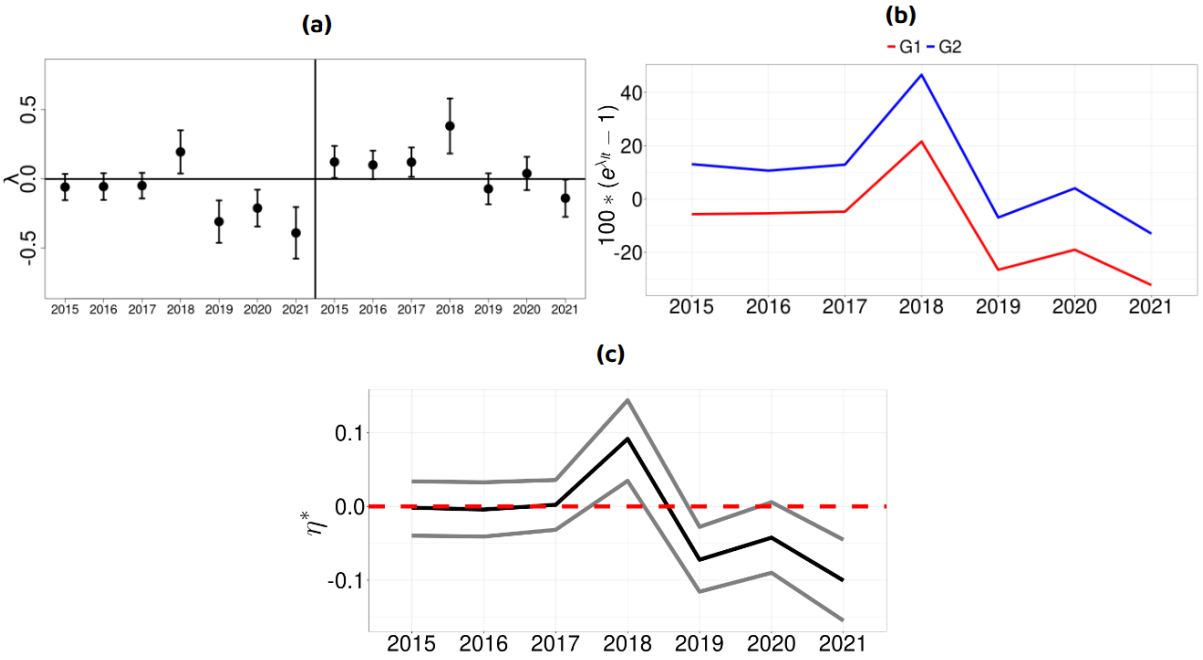
Table 4.1: Real application related to the ENEM data. Posterior estimates (fitting the Beta model) for β, ψ, σ^2, and τ_α. The values were configured with two decimal places, thus 0.00 is not exactly 0.

Recall that a sign exchange might occur between α and λ in the FA. In order to avoid this problem (previously mentioned in this dissertation), we simply require that the majority of non-null loadings related to the groups G_1 and G_2 are positive. If the column k of α is multiplied by -1 to meet this requirement, then the row k of λ must also have the sign change. This sign modification does not affect the covariance matrix defined in the prior for η^* [23].

Figure 4.2 $(a-b)$ shows results related to the factor scores in λ. As can be seen from Panel (a), the patterns of the scores for Factors 1 and 2 are similar. From 2015 to 2017, the scores remain stable, then an increase is observed between $2017-2018$. A fast decay is estimated between $2018-2019$, and small alterations are detected between $2019-2021$. The main difference between G_1 and G_2 is that the scores of Factor 1 are lower than those of Factor 2, for all years. Some 95% HPD intervals contain 0, suggesting that their corresponding scores may not be significant.

Figure 4.2 (b) presents the impact (in %) of the score λ_{kt} on the mean θ_i of the response variable (ENEM grade 1^{st} PCA). The mentioned impact is calculated as indicated in [7]. In brief, the relationship between θ_i and the linear predictor is given by $\theta_i/(1-\theta_i) =$

Figure 4.2: Analysis of factor scores and interaction. Panel (a) shows the 95% HPD intervals for the scores in λ. Panel (b) shows the values of $100(e^{\lambda_{kt}} - e^0)$, which can be seen as the impact (in %) of λ_{kt} over $\theta_i/(1-\theta_i)$. Panel (c) presents the impact of the interaction η^*_{lt} in red and the HPD interval in black.

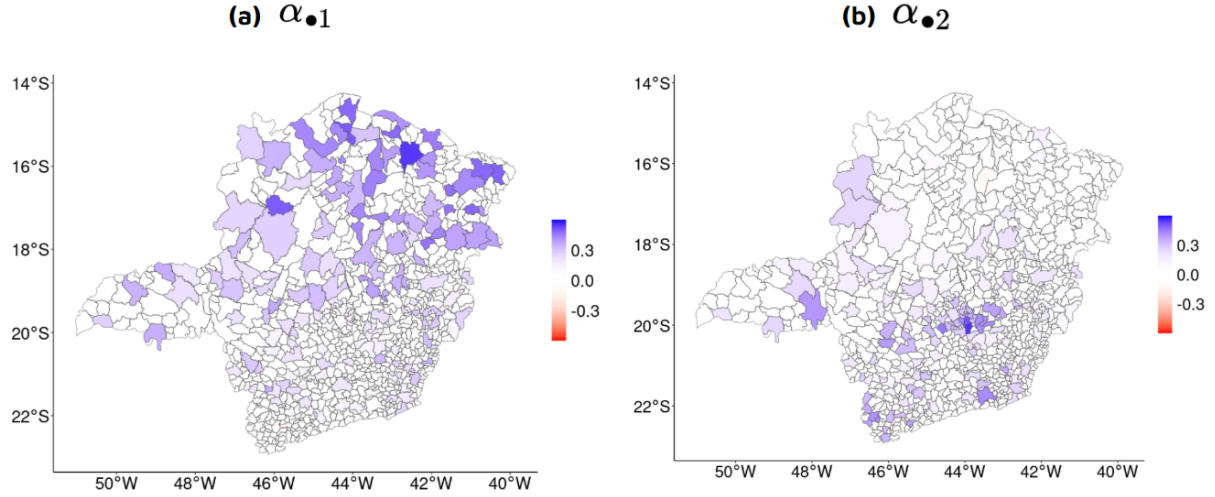


$\exp\{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}\}\exp\{\delta_{l_i^* t_i^*}\}$. The term $\exp\{\delta_{lt}\} = \exp\{\alpha_{l\bullet}\lambda_{\bullet t} + \eta_{lt} + \epsilon_{lt}\}$ can be seen as the impact of the FA structure in the ratio $\theta_i/(1-\theta_i)$. If $\alpha_{l\bullet} = (0,1)$ or $\alpha_{l\bullet} = (1,0)$, $\eta_{lt} = 0$ and $\epsilon_{lt} = 0$, then the expression of this impact simplifies to $\exp\{\lambda_{kt}\}$. The following measurement $100(e^{\lambda_{kt}} - e^0)$ provides the percentage in which the ratio $\theta_i/(1-\theta_i)$ increases or decreases (w.r.t. $\lambda_{kt} = 0$) given the score $\lambda_{kt} \neq 0$. The term e^0 refers to the impact of the null score. If $100(e^{\lambda_{kt}} - e^0) > 0$, then $\lambda_{kt} > 0$ and the estimated score increases the ratio $\theta_i/(1-\theta_i)$. If $100(e^{\lambda_{kt}} - e^0) < 0$, then the ratio decreases. In Panel (b), note that the trajectory of the line representing G_1 (low HDI regions) is mostly located below 0, suggesting a reduction in the ratio $\theta_i/(1-\theta_i)$ for every year (except 2018). In contrast, the trajectory related to G_2 (high HDI regions) is mostly located above 0, suggesting an increase in $\theta_i/(1-\theta_i)$ for every year (except 2021).

Figure 4.2 (c) shows the estimated $\eta^*_{l\bullet}$ in black line. As can be seen, the behavior observed for the scores in Panel (a) can also be detected for the nonlinear interaction. The 95% HPD intervals (in grey line) show that η^*_4 (2018), η^*_5 (2019) e η^*_7 (2021) are significantly different from zero. This conclusion is due to the fact that the red dashed line is not within the credibility interval. The other components of the interaction vector are not significant.

Figure 4.3 shows maps of Minas Gerais displaying the magnitude of the loadings for each municipality included in the study. Panel (a) indicates the loadings associated with Factor 1, and Panel (b) displays the loadings associated with Factor 2. As can be

Figure 4.3: Maps displaying the magnitude of the estimated loadings throughout the space. Loadings related to Factor 1 are presented in Panel (a). Loadings of Factor 2 are shown in Panel (b).



seen, blue is the main color in both maps, indicating that the loadings are mostly positive in this analysis. Therefore, when increasing the factor score, the random effects tend to increase as well. Note that Factor 1 tends to have higher loadings for municipalities in the north of the state. Factor 2 has larger loadings for a few municipalities in the center, west, and south parts of the map. In addition, among the regions considered in the analysis, many are mainly associated with one of the factors. In Appendix C it is possible to see extra results about the estimates and HPD intervals for α and δ.

The maps in Figure 4.4 are related to the columns (years) of δ. A pattern similar to that reported for η^* and λ is observed. Note that between $2015-2017$ the magnitudes of δ_{lt}'s are close to zero, and light colors (near white) are observed for most regions. In the North and Northeast parts, municipalities tend to have a light red color suggesting a small decrease in the average response. On the other hand, in the South, West, and Southeast, one can see light blue, indicating a small increase in the average response. This color perception is quite difficult to see in Panels ($a-c$), then for these years, it seems there is no association between location and ENEM performance. In 2018 (Panel d), it is clear that larger magnitudes (mostly positive) are detected for the random effects, which means that the average grade may have increased. Finally, note that between $2019-2021$ (Panel $e-g$) the red color (negative δ_{lt}) can be detected for most regions, which suggests that the average grade decreases for those years. Dark red (more significant negative effect) is especially related to municipalities in the North of the State. This visual analysis suggests that there may exist an important association between location and ENEM performance for the last three years of the study. No scientific studies were found to justify the upward trend in the magnitudes of δ_{lt}'s in 2018 and a decrease between $2019-2021$. However, some reports in the news media show that in the year 2018, there was an increase in the

average grade in three of the four objective tests (see the news here[3]). In addition, to justify the decrease between 2018−1029, the change of government, which generated many changes in the presidency of INEP (the institution that organizes ENEM), in addition to the pandemic, may be the main reason for this trend.

The final analysis of this dissertation is related to the identification of clusters including municipalities affected by the same combination of main factors and interaction. A significant effect of Factor 1 or Factor 2 is determined by verifying whether the HPD intervals for the loadings do not include zero. The credibility level of 95% is the first choice in practice, but this option provides (in the present application) a wide interval that incorporates 0 for most loadings in α. As an alternative to avoid the result where almost all loadings are not significant, we choose to reduce (to 70%) the credibility level of these HPD intervals considered in the criterion. Please note that this real application involves 186 municipalities, 7 years, and $n = 30{,}000$ individuals. If a larger n is adopted, then shorter 95% HPD intervals are obtained and lower credibility levels would not be necessary to evaluate significance. The interaction effect is considered significant for those locations having a posterior probability $p^*(Z_l = 1|\bullet) > 0.5$. The threshold 0.5 is the same used in [7].

Figure 4.5 shows the configurations of the clusters. Again, most municipalities with high HDI are located in the South and Southeast parts of Minas Gerais, whereas, low HDI regions are mainly found in the North and Northeast. As can be seen, regions affected by both factors are located in the central area of the map. Regions solely affected by Factor 1 are located in the North. In addition, regions solely affected by Factor 2 are found in the South. Panel (d) shows that a few regions are not affected by any of the two factors. The identification of these clusters can be useful for governments and education departments to assist in their decision-making regarding the municipalities showing worse performance in ENEM.

This section developed a real application involving the Beta model described in Chapter 2. The study has a continuous bounded response variable defined (via 1^{st} PCA) as an index representing the ENEM grade obtained by the individuals. The HDI Income was considered as an auxiliary variable to create groups of municipalities with contrasting HDIs. This strategy is intended to solve an identification issue in the FA part of the model. The estimated factor scores indicate that the year 2018 is particularly interesting displaying a higher score than the other years. The cluster analysis from Figure 4.5 aims to identify the locations associated with both, one of the factors, none of the factors, and the presence/absence of interaction. The trajectory of the factor scores across the years is quite similar for Factor 1, Factor 2, and also for η^*. The next chapter presents the main conclusions of the dissertation.

[3] `https://bit.ly/crise-no-inep`

Figure 4.4: Maps displaying the magnitude of the estimated random effects δ_{lt}'s throughout the space. Each panel represents one of the years considered in the study. This analysis involves 186 municipalities in Minas Gerais (Brazil).

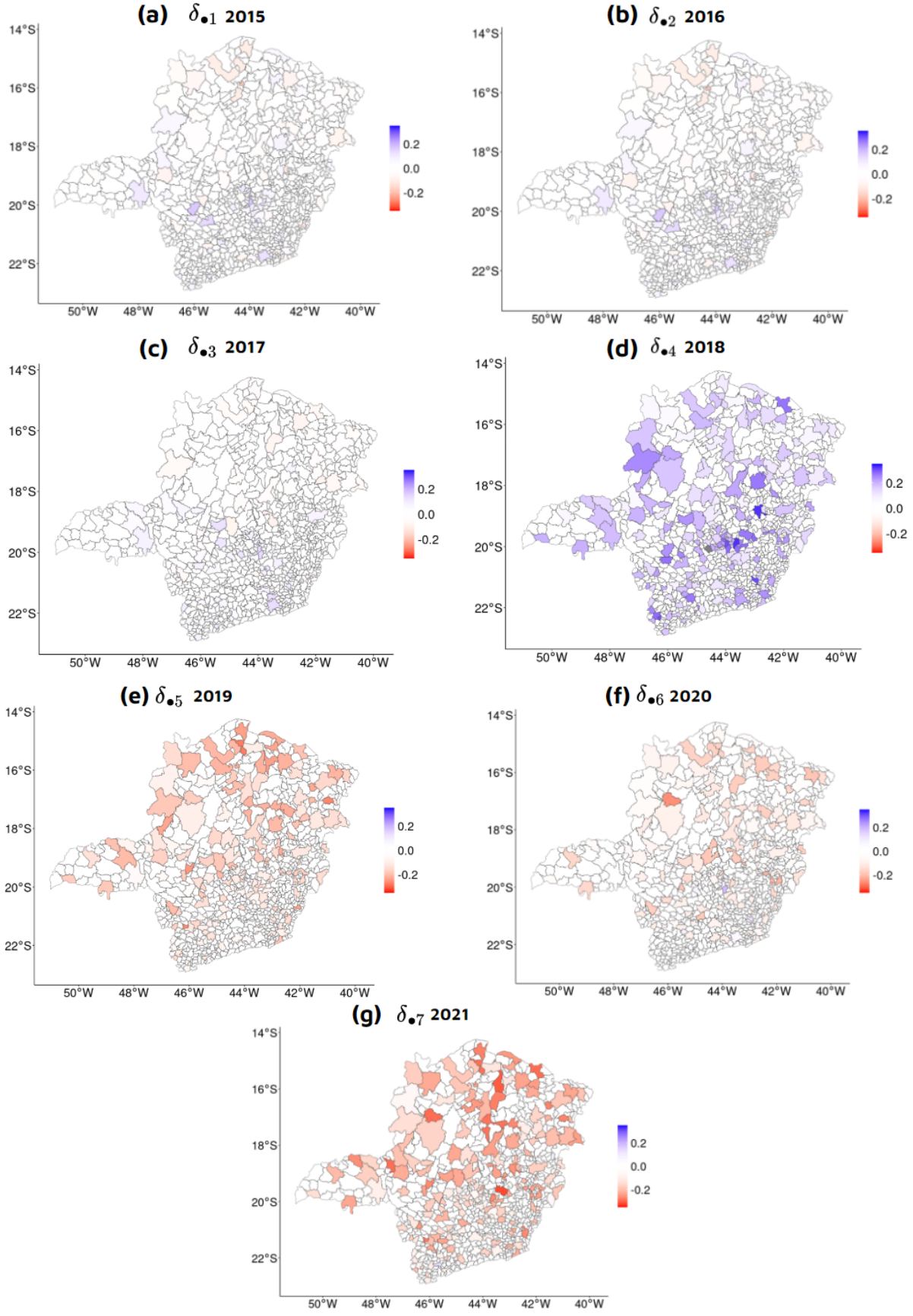
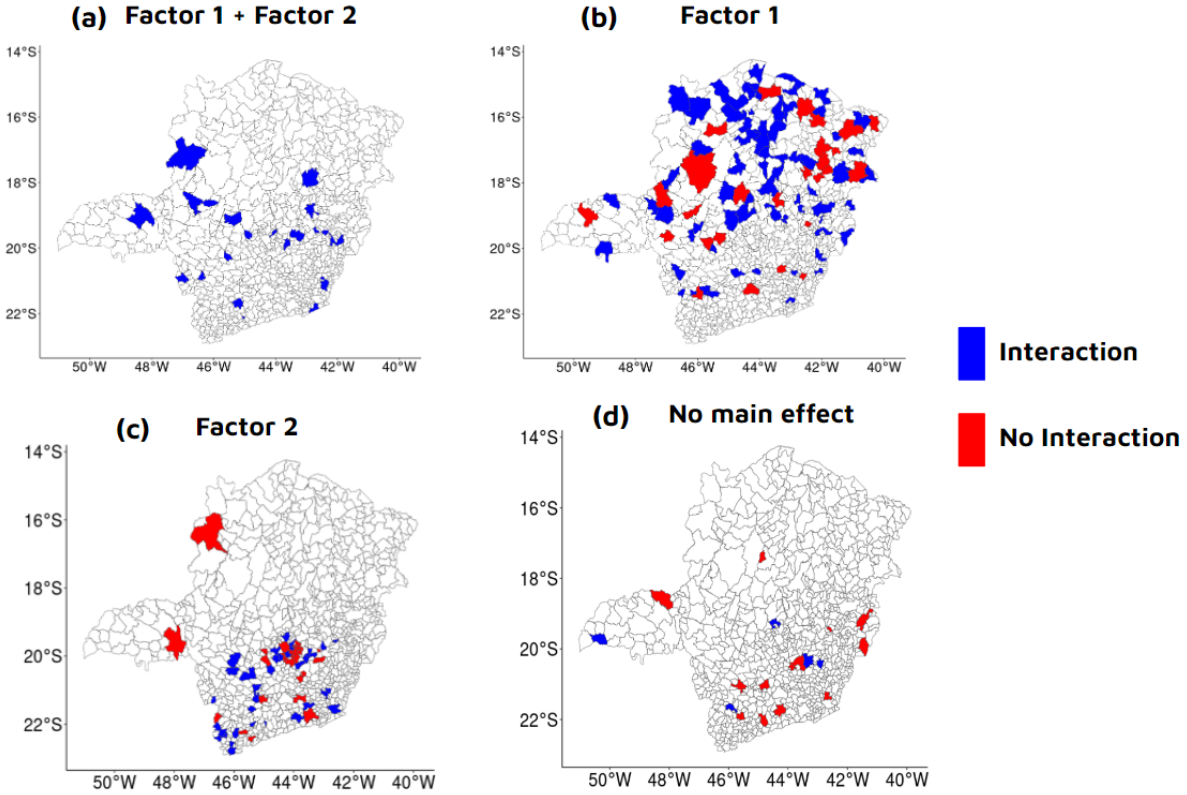


Figure 4.5: Partitioning the 186 municipalities in clusters according to the influence of factors and interaction. Blue = region affected by the interaction. Red = region without interaction effect. Panels $(a - d)$ contain the regions affected by both, at least one, or none of the main factors. The main factor effect is determined by looking at the loadings having 70% HPD interval without zero. The presence of an interaction effect is established based on the mixture posterior probability > 0.5 in Step 3 (Section 2.4).



Chapter 5

Conclusions

In this work, we extend the model developed in [7] where we fit a generalized mixed linear model (GLMM) considering continuous response variables (Normal, Gamma and Beta cases). The random effect is modeled using the structure of a factor model, including a nonlinear interaction between the factor scores. The proposal is a spatiotemporal model with spatial dependence defined for the columns of the factor loadings matrix, and temporal dependence established for the rows of the factor scores matrix. The spatial and temporal dependencies were defined via the CAR structure; the temporal dependence can also be seen through the AR(1) process. Identification problems between the matrices α, λ, and η can occur; we describe these issues and provide solutions. The nonlinear interaction added in the modeling aims to improve the identification of clusters (groups of regions), and it allows the detection of complex associations between factors that may influence the behavior of the data for each municipality.

One main aspect of this work is to extend the GLMM proposed in [7] for three scenarios of continuous response variables. Subsequently, there was also interest in analyzing and verifying the model's performance in simulation studies developed based on artificial data. A study with a single data set suggested that most parameters are well-estimated with low posterior uncertainty. The random effects in δ are well-estimated, especially for the Gamma and Beta cases. The estimation of these effects is satisfactory for the Normal case, despite the higher posterior variability. For the matrices α and λ, the estimates follow the pattern of the true values, and most of them fall within the 95% HPD credibility intervals.

Regarding the probability of locations being affected by the interaction, the model captures well the true and false indications of the areas affected by the interaction. However, some identification problems were identified between the intercept and the random effects in δ. An overestimation is observed for β_0 and, as a consequence, an underestimation is detected for δ. A Monte Carlo study was necessary to evaluate the impact of this problem on the model. The MC study suggested that such confusion between β_0 and δ does not seriously affect the model fit and the interpretation of the remaining parameters. In addition, a sensitivity study was performed to select key parameters defined in the covariance matrix specified via CAR modeling.

Finally, a real application was developed to illustrate in practice how one of the proposed GLMMs can be used. The chosen data set is related to individuals that were submitted, in Minas Gerais, to the Brazilian national exam ENEM. The period of the study involves the years between $2015 - 2021$, and the total number of municipalities is 186. The analysis is developed with $K = 2$ factors, 10 regions (high HDI Income) having high association with Factor 1, and 10 regions (low HDI Income) highly connected with Factor 2. The remaining regions in the study are assumed in the group G_E, for which the associations with the main factors are unknown and the presence of the interaction effect is also uncertain. It was verified that the scores of Factors 1 and 2 had similar patterns, and the Factor 1 scores are lower than those from Factor 2. The year 2018 is particularly different having a higher score than the other years in the study. The nonlinear interaction effect follows the same pattern throughout time when compared to the Factor scores. Finally, clusters of regions can be detected by using the proposed model with FA structure. These clusters are defined in terms of the combination of effects influencing their regions (not necessarily contiguous regions). A given cluster can be affected by: the interaction, both factors, one of the factors, or none of the factors. Identifying these groups can be useful to guide government public policies directed to high school students in regions highly associated with patterns observed in low HDI municipalities.

5.1 Future Work

This work and [7] focused on models whose probability distributions were discrete (Bernoulli and Poisson) and continuous (Normal, Gamma and Beta). Thus, one interesting future step is to develop the methodology for other families of distributions including, for example, the skew-normal and t-Student to handle atypical observations in the response. Another future work is to develop an adequate residual analysis for this type of model. Evaluating residuals is another way to determine whether the model fit is suitable for the data. We would also like to be able to run this model for the application considering the entire sample of 701,930. Finally, developing an `R` package to fit these models is certainly a critical step to disseminate the work and motivate researchers to use the methodology.

References

[1] R Assuncao and E Krainski. Neighborhood dependence in bayesian spatial models. *Biometrical Journal*, 51(5):851–869, 2009.

[2] S Banerjee, B P Carlin, and A E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Presss, New York, 1 edition, 2004.

[3] W Barreto-Souza, V D Mayrink, and A B Simas. Bessel regression and bbreg package to analyze bounded data. *Australian and New Zealand Journal of Statistics*, 63(4):685–706, 2021.

[4] J Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36(2):192–236, 1974.

[5] M L Corrales and E Cepeda-Cuervo. A bayesian approach to mixed gamma regression models. *Revista Colombiana de Estadística*, 42(1):81–99, 2019.

[6] S L P Ferrari and F Cribari-Neto. Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

[7] M P S Ferreira, V D Mayrink, and A L P Ribeiro. Generalized mixed spatio-temporal modeling: Random effect via factor analysis with nonlinear interaction for cluster detection. *Spatial Statistics*, 43:100515, 2021.

[8] J I Figueroa-Zúñiga, R B Arellano-Valle, and S L P Ferrari. Mixed beta regression: A bayesian perspective. *Computational Statistics and Data Analysis*, 61(7):137–147, 2013.

[9] R A Fisher. The use of multiple measures in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[10] D Gamerman and H F Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC, London, 1 edition, 2006.

[11] D Gamerman and E Salazar. *Hierarchical modeling in time series: the factor analytic approach*, pages 167–182. Oxford University Press, Oxford, 2013.

[12] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[13] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[14] E I George and E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[15] E I George and E McCulloch. Approaches for Bayesian variable selection. *Statistical Sinica*, 7:339–373, 1997.

[16] W K Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[17] R A Johnson and D W Wichern. *Applied Multivariate Statistical Analysis*. Pearson, Upper Saddle River, 6 edition, 2007.

[18] H F Lopes and C M Carvalho. Factor stochastic volatily with time varying loadings and Markov switching regimes. *Journal of Statistical Planning and Inference*, 137(10):3082–3091, 2007.

[19] H F Lopes, D Gamerman, and E Salazar. Generalized spatial dynamic factor models. *Computational Statistics and Data Analysis*, 55(3):1319–1330, 2011.

[20] H F Lopes, E Salazar, and D Gamerman. Spatial dynamic factor analysis. *Bayesian Analysis*, 3(4):759–792, 2008.

[21] H F Lopes and M West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 2004.

[22] V D Mayrink and D Gamerman. On computational aspects of Bayesian spatial models: influence of the neighboring structure in the efficiency of MCMC algorithms. *Computational Statistics*, 24(4):641–669, 2009.

[23] V D Mayrink and J E Lucas. Sparse latent factor models with interactions: analysis of gene expression data. *The Annals of Applied Statistics*, 7(2):799–822, 2013.

[24] C E McCulloch and J M Neuhaus. *Generalized Linear Mixed Models*, pages 845–852. Elsevier, Amsterdam, 2 edition, 2015.

[25] G J McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Hoboken, 1 edition, 2004.

[26] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[27] J A Nelder and R W M Wadderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384, 1972.

[28] K Pearson. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine and Journal of Science*, 11(423):559–572, 1901.

[29] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[30] G O Roberts and S K Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B*, 59(2):291–317, 1997.

[31] H Rue and L Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, London, 1 edition, 2005.

[32] R Da Silva, L C Lamb, and M C Barbosa. Universality, correlations, and rankings in the Brazilian universities national admission examinations. *Physica A: Statistical Mechanics and its Applications*, 457:295–306, 2021.

Appendix A

Simulated study to determine ρ_α.

Appendix A shows some results of the study carried out to determine the value for ρ_α. Some results in the literature report that, when the value of ρ_α approaches 1, the model produces greater spatial dependence in the covariance matrix defined in the CAR model, in addition to facilitating the interpretation [1]. We tested different values of this parameter (0.9, 0.999, and 0.99999). The goal is to verify whether changing the value of ρ_α determines different levels of Relative Bias (RB) for other key parameters of the model (see Section 3.2).

	Normal			Gamma			Beta		
ρ_α	0.9	0.999	0.99999	0.9	0.999	0.99999	0.9	0.999	0.99999
β_0	0.38	0.40	0.39	0.48	0.45	0.47	0.22	0.20	0.19
β_1	0.05	0.05	0.05	0.03	0.03	0.03	0.01	0.01	0.01
β_2	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01
ψ	0.02	0.02	0.02	0.01	0.01	0.01	0.04	0.04	0.04
σ^2	0.24	0.24	0.24	0.21	0.21	0.21	0.03	0.03	0.03
τ_α	0.27	0.31	0.21	0.29	0.25	0.26	0.25	0.31	0.28

Table A.1: Relative Bias of the parameters β_0, β_1, β_2, ψ, σ^2, and τ_α obtained by setting different specifications of ρ_α.

Table A.1 shows the results of the three cases specifications of ρ_α for the Normal, Gamma, and Beta cases. We do not include other results due to the fact that the conclusions are similar. The closer the RB is to 0, the closer the parameter's estimated value is to its true value. Also, the values have been rounded, and then equal results are not necessarily the same number. In this way, we can observe that, for all cases, the RB is close to 0, indicating that the estimates are close to the true values. Furthermore, the RB was very similar for all parameters, except for β_0 and τ_α. Therefore, we conclude that the estimates from each specification are close, indicating no difference between fitting the model with $\rho_\alpha = 0.9$ and fitting the model with larger ρ_α. When the value of ρ_α approaches 1, the computational time increases, therefore we prefer to keep $\rho_\alpha = 0.9$.

Appendix B

Extra results from the simulation study.

This Appendix shows some extra results for the matrices α and δ estimated in the simulation study for one replicate. Figure B.1 shows heat maps comparing the true and estimated values of α and figure B.2 of δ, which are the largest matrices in the factor model. Note that the real pattern of the matrices was well captured by the Bayesian model. Visually, for both matrices, it is noticed that the estimated values are slightly underestimated.

Figure B.1: Heat maps confronting estimated and true values for scenario Normal, Gamma, and Beta refers to α.

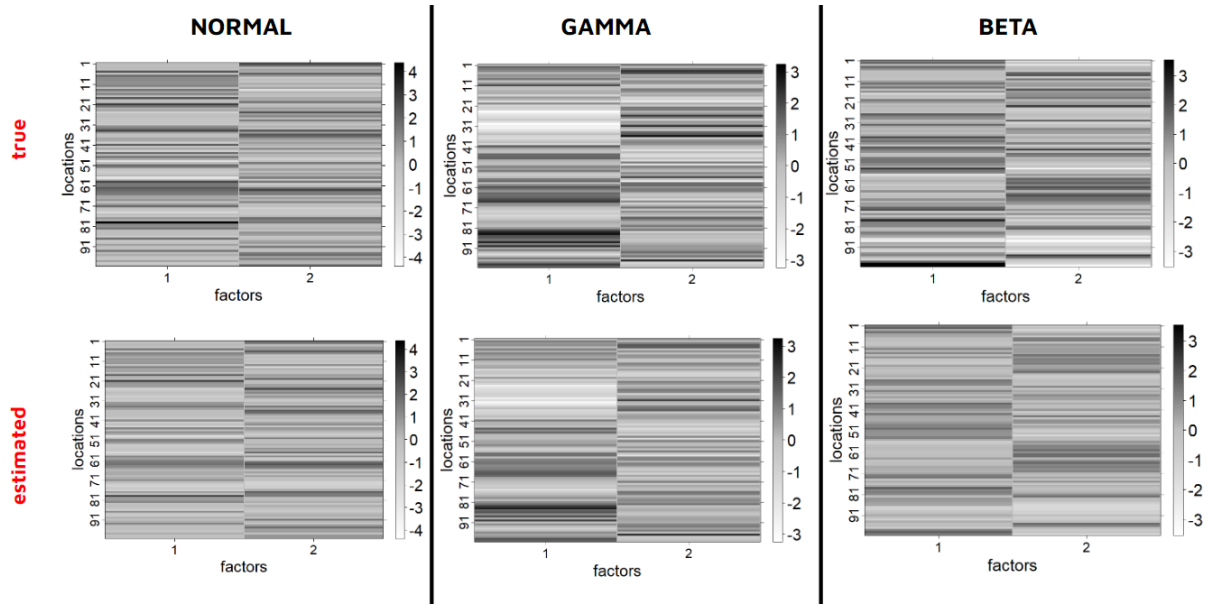
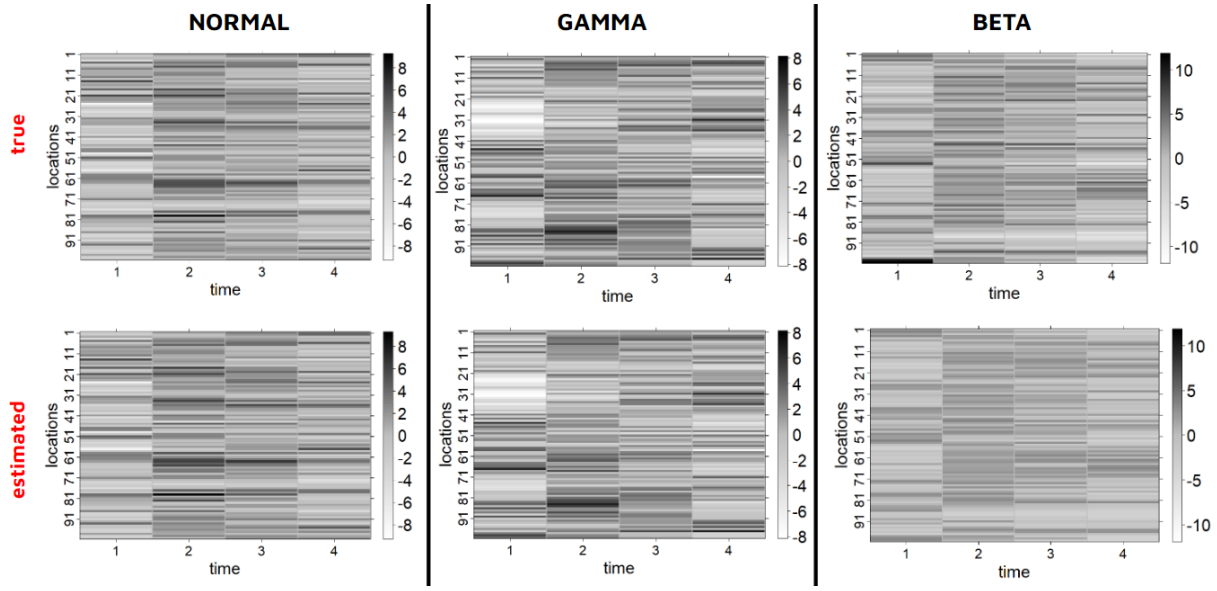


Figure B.2: Heat maps confronting estimated and true values for scenario Normal, Gamma, and Beta refers to δ.



Appendix C

Extra results from the real application.

This Appendix shows some extra results for the matrices α and δ estimated in the real application. In Figure C.1, the black point represents the posterior mean while the segments indicate the 95% HPD credibility intervals. As can be seen, Panel (a) shows that most posterior means have positive values (above the black horizontal line). Panel (b) presents the posterior mean and 95% HPD intervals for the random effects in δ. These results are in accordance with what was observed in the simulation study based on artificial data.

Figure C.1: Painel (a) and (c) present the 95% HPD interval for α and δ respectively.