UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Bayesian variable selection for logistic mixture models with Pólya-Gamma data augmentation

Mariella Ananias Bogoni

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Mariella Ananias Bogoni

Bayesian variable selection for logistic mixture models with Pólya-Gamma data augmentation

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Graduate Program in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
February 2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Bogoni, Mariella Ananias
B674b Bayesian variable selection for logistic mixture models with Pólya-Gamma data augmentation / Mariella Ananias Bogoni; orientadora Daiane Aparecida Zuanetti. -- São Carlos, 2022.
99 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Variable selection. 2. g-prior. 3. spike and slab prior. 4. Pólya-Gamma-sampling. I. Zuanetti, Daiane Aparecida, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Mariella Ananias Bogoni

Seleção Bayesiana de variáveis para modelos de mistura de regressão logística com variáveis latentes Pólya-Gamma

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
Fevereiro de 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Mariella Ananias Bogoni, realizada em 15/02/2022.

Comissão Julgadora:

Profa. Dra. Daiane Aparecida Zuanetti (UFSCar)

Prof. Dr. Carlos Tadeu Pagani Zanini (UFRJ)

Profa. Dra. Rosineide Fernando da Paz (UFC)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

I dedicate this work to my mother Francismara, my brother Otávio and my grandparents Selma and Nelson.

ACKNOWLEDGEMENTS

Firstly, I would like to thank God for giving me the opportunity to have the Masters training I dreamt of. I thank God for His protection, love and faithfulness during all this time away from home and for giving me strength and intellect to complete this work.

I would like to thank my mother Francismara, for all love and encouragement. To my brother, Otávio, for the support. And to my grandparents, Selma and Nelson, for keeping me in prayer.

During my Master's degree, I had the opportunity to meet with nice people along the way who became my friends. I thank my friend and partner Osafu Egbon for walking by my side all this time. I thank my dear friend Ritha Huaysara who has been walking with me since the summer course in January 2020. I also thank my friends Danillo Assunção and Fabiano Rodrigues, for all the moments of play and joy, and my friend Asrat Mekonnen, for the daily coffee support in the laboratory. I have to thank my friends from Volta Redonda, Daiana de Menezes and Jeihcio Francis for all the emotional support.

I also would like to thank my advisor, Daiane Zuanetti for all the patience and attention given to me during this period. She is excellent at what she does and undoubtedly an inspiration to all her students. I could not have asked for a better advisor. Thank you! A special thank to the professor Gustavo Pereira for all the knowledge provided in the regression course and also the encouragement to go for PHD.

Finally, I would like to thank CAPES, as this work was carried out with the financial support of the Coordenação de Aperfeiçoamento Pessoal de Nível Superior - Brazil (CAPES) - Code of Financing 001.

RESUMO

BOGONI, M.A. **Seleção Bayesiana de variáveis para modelos de mistura de regressão logística com variáveis latentes Pólya-Gamma** . 2022. 99 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Neste trabalho, métodos Bayesianos para estimação e seleção de variáveis em um modelo de mistura de regressão logística são apresentados. Com o objetivo de simplificar a inferência Bayesiana e ganhar eficiência computacional, a abordagem de aumento de dados com variáveis latentes Pólya-Gama é estendida para modelos de mistura de regressão logística. Através dela, o algoritmo amostrador de Gibbs é aplicado para a estimação do modelo completo, com a estimação do número de componentes da mistura sendo feita através de critérios Bayesianos de seleção de modelos. Para a seleção de variáveis, duas distribuições a priori para os coeficientes de regressão são investigadas, adicionando um segundo conjunto de variáveis latentes para indicar a presença e ausência das variáveis preditoras em cada componente da mistura. De modo análogo ao modelo completo, o algoritmo amostrador de Gibbs é aplicado no modelo com a seleção de variáveis e a conjugação obtida para a distribuição dos coeficientes de regressão, com a inclusão das variáveis Pólya-Gama, nos permite calcular analiticamente a verossimilhança marginal e ganhar eficiência computacional no processo de seleção de variáveis. Para analisar a performance dos métodos, as metodologias apresentadas são aplicadas em dados simulados e reais.

Palavras-chave: Seleção de variáveis, *g*-priori, priori spike e slab, Pólya-Gamma-sampling.

ABSTRACT

BOGONI, M.A. **Bayesian variable selection for logistic mixture models with Pólya-Gamma data augmentation**. 2022. 99 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

In this work, Bayesian methods for estimating and selecting variables in a mixture of logistic regressions model are presented. In order to simplify its Bayesian estimation, we extend the data augmentation approach with Pólya-Gamma random variables to the mixture of logistic regression models. Through the data augmentation approach, we present a Gibbs sampling algorithm for estimating the full model, and the number of components in the mixture is identified by Bayesian model selection criteria. In the model with variable selection, we investigate the performance of two prior distributions for the regression coefficients, adding a second set of latent variables to indicate the presence and non-presence of the predictor variables at each component of the mixture. Analogously to the full model, a Gibbs sampling algorithm is applied to the model with variable selection and the conjugation obtained for the distribution of the regression coefficients, through the inclusion of Pólya-Gamma variables, allows us to analytically calculate the marginal likelihood and gain computational efficiency in the variable selection process. To analyse the performance, the presented methodologies are applied in simulated and real data.

Keywords: Variable selection, *g*-prior, spike and slab prior, Pólya-Gamma-sampling.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

Finite mixture models can be applied to model data in many contexts (MELNYKOV; MAITRA, 2010), however, it is very common to be applied to model heterogeneous data. When this is the case, we say that the population is made up of K subpopulations and within each subpopulation i, the random variable Y of interest is modelled by a distribution $f(y|\theta_i)$. The subpopulations are called components and they are weighted by the proportion π_i of observations that belong to the component i. The mixture model of logistic regressions allows us to model the relationship between a binary or count outcome Y and a set of predictor variables $\mathbf{x}^\intercal = (x_1, \cdots, x_p)$ when there is presence of heterogeneity in the data, that is, when the outcome Y may be differently affected by the predictor variables across the population. Some examples are Li (2018), that used a finite mixture of logistic regression models to analyze the heterogeneity of the merging behavior of the driver population, and Deng, Chen and Li (2006), that applied a finite mixture of logistic regression to model the heterogeneity in the binary trait locus (BTL) mapping.

From the frequentist perspective, the estimation of a mixture model is based on the maximization of the likelihood function, considering a fixed number of components K, through the iterative algorithm Expectation-Maximization (EM) (DEMPSTER; LAIRD; RUBIN, 1977). When the number K of components is unknown, model selection criteria such as Akaike Information Criterion (AIC) (AKAIKE, 1998) and Bayesian Information Criterion (BIC) (SCHWARZ, 1978) are the most common ways to select the best value of K. One of the challenges in estimating the model with the EM algorithm is the dependency on the initial values. The EM algorithm is usually initial values sensitive to and, in some situations, it can also presents slow convergence. Some examples of frequentist estimation of a mixture of logistic regressions are Deng, Chen and Li (2006) and Li (2018).

The estimation of the model from the Bayesian perspective, however, can be done through MCMC algorithms, assigning prior distributions to the parameters. When the number of components is unknown, model selection criteria can also be applied to choose the best value. Deviance Information Criterion (DIC) (SPIEGELHALTER *et al.*, 2014) and Extended Bayesian

Information Criterion (EBIC) (CHEN; CHEN, 2008) criteria, for example, are common choices for Bayesian estimation of the number of components. For a pre-specified K, MCMC algorithms are usually used for simulating samples of the joint posterior distribution and estimate parameters of each component, such as Gibbs sampling (CASELLA; GEORGE, 1992) or Metropolis Hastings (CHIB; GREENBERG, 1995). Another option for the case with unknown number of components is the Reversible Jump (RJ) algorithm (GREEN, 1995), that performs estimation and model selection simultaneously.

A very common issue when dealing with Bayesian logistic models, specifically, is its intractable likelihood, which prevents us from applying simpler Bayesian algorithm, as Gibbs sampling, since there is no conjugation. Wan and Griffin (2021) describe some recent approaches to deal with the intractable likelihood of the logistic model, including Laplace approximation, Metropolis Hastings based sampler and data augmentation.

The data augmentation technique has been widely employed in binary models. Tüchler (2012), for example, introduced a two data-augmentation technique into the binary logit regression model to obtain a normal linear regression model. A new and recent data augmentation approach to the logistic regression model was proposed by Polson, Scott and Windle (2013), which introduces latent variables with Pólya-Gamma distribution, leading to a tractable likelihood and obtaining a simple and effective method for posterior inference. In this work, we explore this approach to solve the intractable likelihood of the mixture of logistic regression models.

Another crucial problem in fitting good regression models is the selection of predictor variables, especially when the population is made up of K subpopulations. The non-Bayesian variable selection methods for mixture models include the information criteria such as AIC and BIC (NAIK; SHI; TSAI, 2007) and methods based on the penalized log-likelihood function. The methodologies based on information criteria, however, take into account all the 2^p possible models, where p is the number of available predictor variables. And this number gets larger when considering the number of components in the mixture, inducing a high computational cost. For penalization methods, Khalili and Chen (2007) provided a penalized likelihood approach by introducing a new class of penalty functions, which solves the computation limitations of the information criterion approaches. Later, Städler, Bühlmann and Geer (2010) proposed a l_1-penalization approach with a specific parametrization, which leads to a better computational performance. Following the same idea, Khalili and Lin (2013) proposed a penalization approach for high-dimension data by changing the penalization functions proposed by Khalili and Chen (2007) and Städler, Bühlmann and Geer (2010). More recent works of penalization methods for variable selection in mixture models can be found on Devijver (2015) and Lloyd-Jones, Nguyen and McLachlan (2018).

The Bayesian variable selection methods, on the other hand, include prior distribution to the regression coefficients and latent variables to identify the presence and absence of pre-

dictors in the fitted model. The most common Bayesian methods are the Stochastic Search Variable Selection (GEORGE; MCCULLOCH, 1993), with the spike and slab prior, and the g-prior (ZELLNER, 1986). Both of them have been increasingly applied to variable selection in traditional regression models and also in mixture models. Chen and Ye (2015) studied the performance of both methods to select predictors in a mixture of linear regressions through the Gibbs sampling algorithm. The g-prior was also investigated by Lee, Chen and Wu (2016), that proposed a Gibbs sampling to variable selection in cases where $p > n$, where n is the sample size, or the predictor variables are correlated. Recently, Lee, Feldkircher and Chen (2021) proposed a RJ algorithm to fit each component as a sparse regression model in a mixture of linear regressions with spike and slab prior.

The main motivation for this work is the possibility of joining the data augmentation technique proposed by Polson, Scott and Windle (2013), which solves the problem of the intractable likelihood in logistic regressions, with the Bayesian variable selection methods using prior distributions that induce sparsity. Once the data augmentation is extended and applied to the mixture case, becomes possible to apply the Gibbs sampling algorithm, usually applied for fitting mixture of linear regression, to estimate a mixture of logistic regression and select relevant predictors. Such method is straightforward, efficient and easy to implement since the conjugation obtained for the distribution of the regression coefficients, with the inclusion of Pólya-Gamma variables, allows us to analytically calculate the marginal likelihood used in the variable selection process.

In this work, we extend the data augmentation approach presented by Polson, Scott and Windle (2013) to a mixture of logistic regression models in order to facilitate its Bayesian estimation. For the Bayesian variable selection, we investigate the performance of two prior distributions for the regression coefficients: the g-prior and the spike and slab prior with simulated data. We also analyze the estimation of the full model assuming a traditional normal prior distribution for the regression coefficients.

This work is organized as follows: in the Chapter 2, mixture models are formally presented, starting with mixture of distributions, extending to mixture of generalized linear models and to mixture of logistic regression models. In the Chapter 3, we present the extension of the Pólya-Gamma data augmentation to the mixture case, discussing its properties and benefits to the Bayesian inference of the model. The Chapter 4 is intended for the Bayesian estimation method of the full model and the model with the variable selection included. In the Chapter 5, we present the simulation results of the full model and variable selection. In the Chapter 6, we apply the methodologies to select variables in a real data set. Finally, in Chapter 7 we present the final considerations of this work.

CHAPTER

2

MIXTURE MODELS

Finite mixture models are known for modeling data that is not well described by a single unimodal distribution. In other words, it means that there are subpopulations in the data where the random variable of interest behaves differently. The presence of heterogeneity in the data leads to the search for models that can accommodate these changes across the subpopulations.

Additionally, mixture models are also used to deal with more complex models. In the binary and count data cases, for example, mixture models can be used to account for overdispersion or zero inflation (WANG; PUTERMAN, 1998), typical in these type of data, and also for clustering when the number and the members of the subpopulations are unknown. Because of this flexibility, mixture model are increasingly being used in statistical modelling.

There exist an extensive literature about mixture models and its Bayesian inference, in this chapter the theory is based on Frühwirth-Schnatter (2006), Mclachlan and Peel (2000) and Frühwirth-Schnatter and Celeux (2018).

2.1 Finite Mixture of Distributions

Suppose that the observed data belongs to a population composed by K subpopulations with proportions $\pi_1, \cdots, \pi_K$. Let Y be a random variable of interest and consider the random sample $\mathbf{Y} = (Y_1, \cdots, Y_n)$ of Y. Each Y_j will be heterogeneous in relation to the population and homogeneous in relation to the subpopulation from which Y_j belongs. Due to this heterogeneity, it is reasonable to affirm that the probability distribution modelling each subpopulation is different. These distributions are called components of the mixture and, if the components belong to the same parametric family, their parameters will differ across the subpopulations. However, they can also belong to different parametric families. Below, mixture model of distributions is defined. Throughout this chapter the definitions are given considering continuous random variables, however the discrete case is analogous.

Definition 2.1.1. A random variable Y, assuming values in $\Omega \in \mathbb{R}$, follows a finite mixture distribution if its probability density function is given as

$$f(y|\boldsymbol{\theta},\boldsymbol{\pi}) = \sum_{i=1}^{K} \pi_i f(y|\boldsymbol{\theta}_i), \quad \forall y \in \Omega \tag{2.1}$$

where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ is the vector of weights (mixing probabilities) of the mixture with $0 \leq \pi_i \leq 1$, $\sum_{i=1}^{K} \pi_i = 1$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$ represents the parameters vector of all components.

Under the model in Definition 2.1.1, we assume that there are K subpopulations in the data and each Y in the subpopulation i follows a distribution corresponding to the density $f(y|\boldsymbol{\theta}_i)$. Thus the density of the model is written as a convex linear combination of the densities for the K components.

An example of a mixture of two-normal distributions is shown in Figure 1. The distribution of the component 1 is the standard normal distribution with parameters $\mu_1 = 0$ and $\sigma_1 = 1$, and the distribution of the component 2 is a normal distribution with parameters $\mu_2 = 5$ and $\sigma_2 = 2$. The mixing probabilities are $\pi_1 = \pi_2 = 0.5$, so that the mixture density is given by $f(y|\boldsymbol{\mu},\boldsymbol{\sigma^2}) = 0.5N(0,1) + 0.5N(5,4)$, where $N(\cdot,\cdot)$ represents the density function of a normal distribution, and is represented by the red curve in Figure 1.

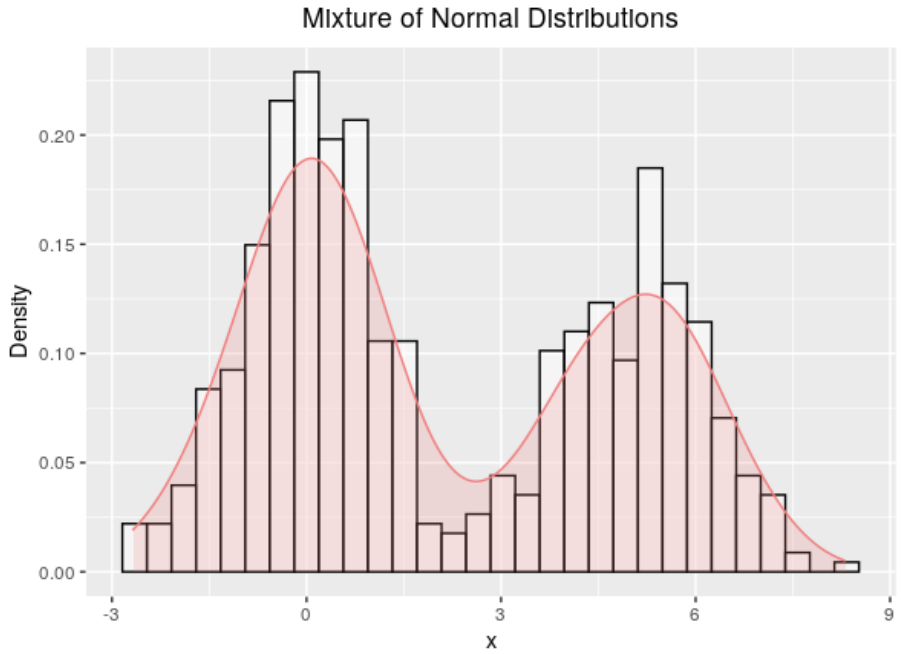

Figure 1 – Simulated data set of a mixture of two normal distributions, $f(y|\boldsymbol{\mu},\boldsymbol{\sigma^2}) = 0.5N(0,1) + 0.5N(5,4)$.

In many applications the variable of interest in the data can depend on other observable factors, which we call as predictor variables or covariates. Mixture models can also be extended to those applications and the model's interpretation remains the same: in each component of the

mixture there is a regression model to explain the response variables belonging to it as a function of relevant covariates and regression coefficients, that can differ among components. In the next section we formalize this class of models considering a more general case of regression models: the generalized linear models.

2.2 Finite Mixture of Generalized Linear Models

From the generalized linear models theory seen in Mclachlan and Nelder (1989) and Dobson and Barnett (2008), given the independent variables $Y_1, \cdots, Y_n$, their dependency on covariates $\mathbf{x}_1, \cdots, \mathbf{x}_n$ is expressed by

$$g(E[Y_j|\mathbf{x}_j]) = \mathbf{x}_j^\intercal \boldsymbol{\beta} = \eta_j, \tag{2.2}$$

where $\boldsymbol{\beta}^\intercal = (\beta_0, \beta_1, \cdots, \beta_p)$ is the unknown parameters vector and $g(\cdot)$ is a monotone and differentiable function, called link function.

In a case where the population is composed by K subpopulations, the presence of heterogeneity implies that in each subpopulation there is a distinct generalized linear model. This means that if Y_j belongs to subpopulation i, there is a subpopulation-specific parameter vector $\boldsymbol{\beta}_i$ so that the relation in (2.2) is modified to

$$g(E[Y_j|\mathbf{x}_j]) = \mathbf{x}_j^\intercal \boldsymbol{\beta}_i = \eta_{ij}, \tag{2.3}$$

where $\boldsymbol{\beta}_i^\intercal = (\beta_{i0}, \beta_{i1}, \cdots, \beta_{ip})$ is the vector of regression coefficients in the subpopulation i, for $i = 1, \cdots, K$ and $j = 1, \cdots, n$. In these conditions, Y follows a generalized linear mixture model.

Definition 2.2.1. A random variable Y, assuming values in $\Omega \in \mathbb{R}$, follows a generalized linear mixture model if its probability density function is given by

$$f(y|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{i=1}^{K} \pi_i f(y|\boldsymbol{\beta}_i, \boldsymbol{\theta}_i), \quad \forall y \in \Omega \tag{2.4}$$

where $\boldsymbol{\beta}_i^\intercal = (\beta_{i0}, \beta_{i1}, \cdots, \beta_{ip})$ is the regression coefficients vector in the component i, $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ is the weights vector and each $f(\cdot|\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$, $i = 1, \cdots, K$, belongs to a parametric exponential family, with mean

$$E[Y|\mathbf{x}] = \sum_{i=1}^{K} \pi_i g^{-1}(\mathbf{x}^\intercal \boldsymbol{\beta}_i), \tag{2.5}$$

where $g(\cdot)$ is the link function and $\mathbf{x}$ is the covariates vector.

Figure 2 shows a particular case of the generalized linear mixture model, where the response variable is normally distributed and the link function is the identity function. The plot shows the fit of a linear mixture model in the GNP and CO2 Data Set provided by the R package `mixtools` (BENAGLIA *et al.*, 2009). This data set provides the gross national product (GNP)

per capita in 1996 for various countries as well as their per capita estimated carbon dioxide emission (CO2) for the same year. In this example, looking only at the dispersion of the points we can clearly see that this data is generated by a mixture of two components. Covariate's effect is different in each component, since the values of the regression coefficients are different for each component, and this change of effects within the subpopulations is exactly what characterizes the mixture.

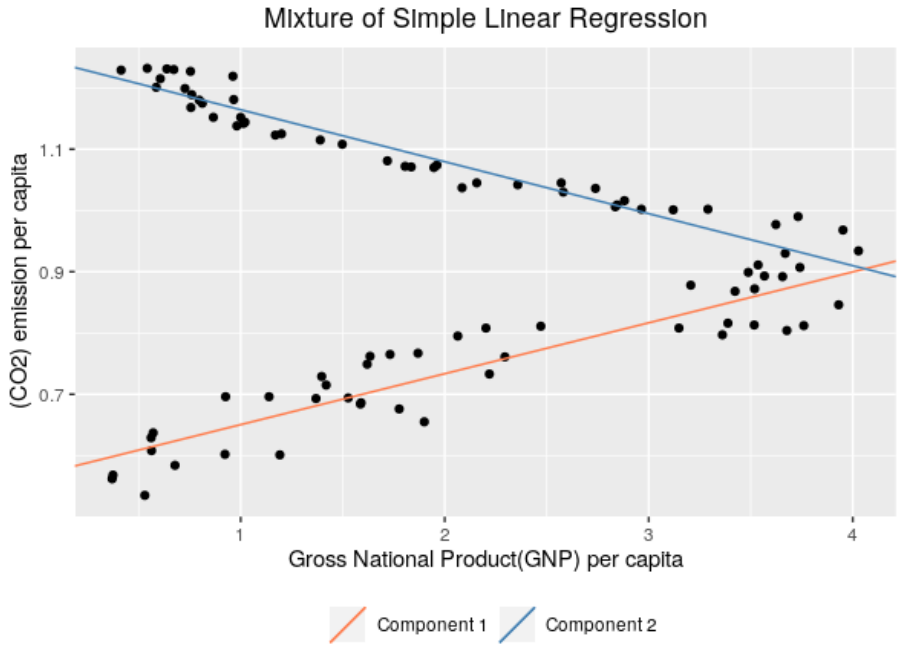

Figure 2 – The linear regression mixture model fit in a real data set provided by the R package `mixtools`.

Mixture of generalized linear models is well known in the literature (GRÜN; LEISCH, 2008), (YANG; MUTHÉN; YANG, 1999) with applications in agriculture (WANG; PUTERMAN, 1998), cognitive development (DAUVIER; CHEVALIER; BLAYE, 2012), economy (KONISHI; NAKAMURA; KIYOKI, 2019), biology (BELL; ZHANG; NIU, 2011) and more.

In the next section we restrict this class of models considering the response as a binary or count random variable, which characterizes a mixture of logistic models.

2.2.1 Finite Mixture of Logistic Models

Mixture of logistic regressions arise as a particular case of generalized linear mixture model, when the response variable Y_j is either binary or a count and the link function is the logit function. Following, a formal definition of mixture of logistic regressions is presented.

Definition 2.2.2. A random variable Y_j assuming values in $\Omega = \{0, 1, \cdots, N\}$ follows a logistic regression mixture model if its probability density function is given by

$$f(y_j|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{i=1}^{K} \pi_i \left[\binom{N}{y_j} \theta_{ij}^{y_j} (1-\theta_{ij})^{N-y_j} \right], \quad \forall y_j \in \Omega \tag{2.6}$$

where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ is the weights vector, N is the number of Bernoulli trials and the success probability θ_{ij} is defined as

$$\theta_{ij} = \frac{\exp(\mathbf{x}_j^\intercal \boldsymbol{\beta}_i)}{1 + \exp(\mathbf{x}_j^\intercal \boldsymbol{\beta}_i)}, \tag{2.7}$$

where $\boldsymbol{\beta}_i^\intercal = (\beta_{i0}, \beta_{i1}, \cdots, \beta_{ip})$ is the vector of regression coefficients in the component i for $i = 1, \cdots, K$, and $\mathbf{x}_j^\intercal$ is the covariates vector associated to y_j.

The Definition 2.2.2 says that the population is made up of K subpopulations, whose success probability varies across them. The most common distribution assigned to the response variable in logistic regression models is the Bernoulli distribution. However, in mixture model framework this choice leads to a generic non-identifiability problem (FRüHWIRTH-SCHNATTER, 2006; FRüHWIRTH-SCHNATTER; CELEUX, 2018).

Generic non-identifiability implies that the likelihood of the observed data is the same for any pair $(\boldsymbol{\pi}, \boldsymbol{\theta}) \neq (\boldsymbol{\pi}^*, \boldsymbol{\theta}^*)$ of parameter vectors that are not obtained by permuting each other. For instance, let Y be a random variable that follows a mixture distribution of $K = 2$ Binomial distributions with $N = 2$ Bernoulli trials with success probabilities θ_1, θ_2 as in Definition 2.2.2. We have that

$$P(Y = 0|\pi, \boldsymbol{\theta}) = \pi(1 - \theta_1)^2 + (1 - \pi)(1 - \theta_2)^2, \tag{2.8}$$

$$P(Y = 1|\pi, \boldsymbol{\theta}) = 2\pi\theta_1(1 - \theta_1) + 2(1 - \pi)\theta_2(1 - \theta_2), \tag{2.9}$$

$$P(Y = 2|\pi, \boldsymbol{\theta}) = \pi\theta_1^2 + (1 - \pi)\theta_2^2. \tag{2.10}$$

Because the probabilities in Equations (2.8), (2.9) and (2.10) have to sum one, we have only two linearly independent equations to identify the three parameters $(\pi, \theta_1, \theta_2)$ of the model. In this case, there will be parameters vectors $(\pi, \theta_1, \theta_2) \neq (\pi^*, \theta_1^*, \theta_2^*)$ that satisfy the Equations (2.8), (2.9) and (2.10) and are not written as a permutation of each other, which implies non-identifiability.

From this example it is possible to see that the number of Bernoulli trials and the number of components are directly related to identifiability of this model. According to Teicher (1963), a necessary and sufficient condition to identifiability in mixture of Binomial distributions is that $N \geq 2K - 1$, where N is the number of Bernoulli trials and K is the number of components of the mixture. For more details and references about mixture of non-normal regression models see Mclachlan and Peel (2000) and Frühwirth-Schnatter (2006).

2.2.1.1 The Likelihood Function

Considering independence among $Y_1, \cdots, Y_n$ and the fact that the random sample $Y_1, \cdots, Y_n$ is not identically distributed, the likelihood function can be written as

$$f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{j=1}^{n} \left[\sum_{i=1}^{K} \pi_i f(y_j|\theta_{ij}) \right]. \tag{2.11}$$

As seen in Ribeiro, Saraiva and Suzuki (2019), the product-of-sums form in (2.11) is not analytically tractable from a maximum likelihood estimation point of view. Similarly, in the Bayesian estimation framework, Dempster, Laird and Rubin (1977) observed that a finite mixture model can always be written as an incomplete-data problem by introducing latent variables in the model, favoring the use of the MCMC algorithms. Moreover, from the computational point of view the likelihood in (2.11) is not feasible for n or K large, since for each observation y_j there is a sum of K elements. Thus, latent variables are introduced in the model so that it is possible to rewrite the likelihood function in such way that we can facilitate the estimation process and classify the data into the components.

Let the discrete random variables $S_1, \cdots, S_n$ such that $P(S_j = i) = \pi_i$ for all $j = 1, \cdots, n$ and $i = 1, \cdots, K$. Conditioning in $S_j = i$, Y_j has density $f(y_j|\theta_{ij})$. In other words, $S_j = i$ indicates that the observation y_j comes from the component i and that occurs with probability π_i. Given the probability distribution function of each S_j,

$$P(S_j = i|\boldsymbol{\pi}) = \prod_{i=1}^{K} \pi_i^{\mathbb{1}_{S_j}(i)}, \tag{2.12}$$

where $\mathbb{1}_{S_j}(i) = 1$ if $S_j = i$ or 0 otherwise, and rewriting the conditional probability distribution function of Y_j as

$$f(y_j|S_j = i, \boldsymbol{\theta}) = \prod_{i=1}^{K} \left[f(y_j|\theta_{ij})\right]^{\mathbb{1}_{S_j}(i)}, \tag{2.13}$$

where $\mathbb{1}_{S_j}(i) = 1$ if $S_j = i$ or 0 otherwise. We rewrite the likelihood function as follows

$$\begin{aligned} f(\mathbf{y}, \mathbf{S}|\boldsymbol{\theta}, \boldsymbol{\pi}) &= \prod_{j=1}^{n} f(y_j, S_j|\boldsymbol{\theta}, \boldsymbol{\pi}) \\ &= \prod_{j=1}^{n} f(y_j|S_j = i, \boldsymbol{\theta}) P(S_j = i|\boldsymbol{\pi}) \\ &= \prod_{j=1}^{n}\prod_{i=1}^{K} \left[f(y_j|\theta_{ij})\pi_i\right]^{\mathbb{1}_{S_j}(i)} \\ &= \prod_{i=1}^{K} \pi_i^{n_i} \prod_{j:S_j=1} f(y_j|\theta_{1j}) \times \cdots \times \prod_{j:S_j=K} f(y_j|\theta_{Kj}), \end{aligned} \tag{2.14}$$

where $n_i = \sum_{j=1}^{n} \mathbb{1}_{S_j}(i)$ is the size of component i. This likelihood is frequently mentioned as the augmented likelihood, since it considers the non-observed data.

The likelihood function written as a product-of-product form in (2.14) makes it easier to identify the posterior distribution and decreases the computational burden in the simulation. However, for the logistic mixture model case, the likelihood is still not analytically convenient. To see that, consider $Y_1, \cdots, Y_n$ a random sample from a mixture of logistic model as in Definition 2.2.2. According to (2.14), the likelihood function is written as

$$
\begin{aligned}
f(\mathbf{y},\mathbf{S}|\boldsymbol{\beta},\boldsymbol{\pi}) &= \prod_{i=1}^{K}\pi_i^{n_i}\prod_{j:S_j=1}\left[\binom{N}{y_j}\theta_{1j}^{y_j}(1-\theta_{1j})^{N-y_j}\right]\times\cdots\times\prod_{j:S_j=K}\left[\binom{N}{y_j}\theta_{Kj}^{y_j}(1-\theta_{Kj})^{N-y_j}\right] \\
&= \left[\prod_{i=1}^{K}\pi_i^{n_i}\right]\prod_{j:S_j=1}\binom{N}{y_j}\frac{(\exp\{\mathbf{x}_j^{\intercal}\boldsymbol{\beta}_1\})^{y_j}}{(1+\exp\{\mathbf{x}_j^{\intercal}\boldsymbol{\beta}_1\})^{N}}\times\cdots \qquad (2.15) \\
&\times \prod_{j:S_j=K}\binom{N}{y_j}\frac{(\exp\{\mathbf{x}_j^{\intercal}\boldsymbol{\beta}_K\})^{y_j}}{(1+\exp\{\mathbf{x}_j^{\intercal}\boldsymbol{\beta}_K\})^{N}}.
\end{aligned}
$$

The exponential factors in (2.15) represents the likelihood function of each mixture component. Note that, considering this likelihood, there is no conjugate prior distribution available for $\boldsymbol{\beta}_i$ and it may be hard to sample from its conditional posterior distribution, since it will not belong to a well-known distribution family. To simplify the sampling process, a data augmentation strategy is proposed in the next section. When implementing this strategy, we can rewrite the likelihood in a way that we obtain a well-known conditional posterior for the regression coefficients, which is a interesting result for the Bayesian estimation since it can be carried out via Gibbs sampling.

CHAPTER

3

DATA AUGMENTATION WITH PÓLYA-GAMMA DISTRIBUTION

In this section, we propose an extension of the data augmentation method introduced by Polson, Scott and Windle (2013) to a mixture of logistic regressions. As we will see later, this method has a differential advantage of allowing to rewrite the likelihood in a way that we can obtain a well-known conditional posterior distribution for regression coefficients and then apply classical MCMC methods such as the Gibbs Sampling for estimating the model. Moreover the method is exact, that is, rather than sampling from an approximation of the posterior obtained from an approximation of the logistic function, we sample from the correct posterior. This approach is only possible due the main result presented in this section.

3.1 The Pólya-Gamma Distribution

In this section we define the Pólya-Gamma distribution and present some properties. More details can be seen in Polson, Scott and Windle (2013).

Definition 3.1.1. A random variable X assuming values in $\Omega = \mathbb{R}^+$ follows a Pólya-Gamma distribution with parameters $b > 0$ and $c \in \mathbb{R}$, denoted by $X \sim PG(b,c)$, if

$$X \overset{D}{=} \frac{1}{2\pi^2} \sum_{m=1}^{\infty} \frac{g_m}{(m-1/2)^2 + c^2/4\pi^2}, \tag{3.1}$$

where $g_m \sim Gamma(b,1)$, $m = 1,2,\cdots$, are independent random variables.

In other words, a random variable X has Pólya-Gamma distribution if its distribution is the same as the distribution of the sum in the left side in (3.1). The family of Pólya-Gamma distribution can be seen as a class of infinite convolutions of Gamma distributions. A particular case is when $b = 1$ and $c = 0$, where we obtain an infinite convolution of exponential distributions, known as Pólya distribution and reported by Barndorff-Nielsen, Kent and Sørensen (1982). In

case $b > 0$ we obtain an infinite convolution of Gamma distributions, giving rise to the name Pólya-Gamma distribution.

Before presenting the main result involving Pólya-Gamma distribution, we present an important object to characterize the Pólya-Gamma distribution, its Laplace transform.

Proposition 3.1.1. Let X be a random variable distributed as the Pólya-Gamma distribution with parameters $b > 0$ and $c = 0$. Its Laplace transform is given by

$$E\left[\exp(-Xt)\right] = \cosh^{-b}\left(\sqrt{t/2}\right), \tag{3.2}$$

for $t > 0$ and the function $\cosh(\cdot)$ denotes the hyperbolic cosine function.

Proof. Let $X \sim PG(b,0)$. By Definition 3.1.1 we have that

$$X \overset{D}{=} \frac{1}{2\pi^2} \sum_{m=1}^{\infty} \frac{g_m}{(m-1/2)^2}, \tag{3.3}$$

where $g_m \sim Gamma(b,1)$ for $m = 1,2,\ldots$. Defining $c_m = \dfrac{1}{2\pi^2(m-1/2)^2}$ and considering that g_m , $m = 1,2,\ldots$ are independent we have that

$$E\left[\exp(-Xt)\right] = E\left[\exp\left\{\sum_{m=1}^{\infty} -c_m g_m t\right\}\right] = E\left[\prod_{m=1}^{\infty} \exp\left\{-c_m g_m t\right\}\right] = \prod_{m=1}^{\infty} E\left[\exp\left\{-c_m g_m t\right\}\right] \tag{3.4}$$

which is the product of Laplace transforms of a Gamma distributions.

From the Laplace transform of Gamma distribution and the Weierstrass Factorization theorem we rewrite (3.4) as

$$\begin{aligned} E\left[\exp(-Xt)\right] &= \prod_{m=1}^{\infty} E\left[\exp\left\{-c_m g_m t\right\}\right] = \prod_{m=1}^{\infty} (1+c_m t)^{-b} = \prod_{m=1}^{\infty} \left(1 + \frac{t/2}{\pi^2(m-1/2)^2}\right)^{-b} \\ &= \cosh^{-b}\left(\sqrt{t/2}\right). \end{aligned}$$

□

The derivation of the Laplace transform for the general case where $b > 0$ and $c \in \mathbb{R}$ is analogous. Polson, Scott and Windle (2013) and Windle (2013) provide a probability density function for $PG(b,c)$ that arises through the exponential tilting of the $PG(b,0)$ density, obtaining

$$f(x|b,c) = \frac{\exp\left\{-xc^2/2\right\} f(x|b,0)}{\int_0^{\infty} \exp\left\{-xc^2/2\right\} f(x|b,0)dx} = \frac{\exp\left\{-xc^2/2\right\} f(x|b,0)}{E\left[\exp\left\{-Xc^2/2\right\}\right]}, \tag{3.5}$$

where the expected value is taken with respect to $X \sim PG(b,0)$ and $f(x|b,0)$ denotes its density probability function.

Polson, Scott and Windle (2013) also derived some good properties of the Pólya-Gamma distribution. The most important one is that all finite moments of the Pólya-Gamma random

variable are available in a closed form, making it possible to calculate the expectation and variance directly. Next, we will present the main result that will enable to rewrite the likelihood of the model.

Theorem 3.1.1. Let $f(x|b,0)$ be the probability density function of the Pólya-Gamma distribution with parameters $b > 0$ and $c = 0$. The following identity holds for all $a \in \mathbb{R}$

$$\frac{(\exp\{\eta\})^a}{(1+\exp\{\eta\})^b} = 2^{-b}\exp\{k\eta\}\int_0^\infty \exp\left\{-x\eta^2/2\right\} f(x|b,0)dx \tag{3.6}$$

where $k = a - b/2$ and $\eta \in \mathbb{R}$.

Proof. To see the identity in (3.6), consider $a = k + b/2$. Replacing a in the left side of the (3.6) we obtain

$$\frac{(\exp\{\eta\})^a}{(1+\exp\{\eta\})^b} = \frac{(\exp\{\eta\})^{k+b/2}}{(1+\exp\{\eta\})^b} = \frac{(\exp\{\eta\})^k(\exp\{\eta\})^{b/2}}{(1+\exp\{\eta\})^b} = \frac{(\exp\{\eta\})^k}{\left(\frac{1+\exp\{\eta\}}{\exp\{\eta/2\}}\right)^b}.$$

Note however that,

$$\left(\frac{1+\exp\{\eta\}}{\exp\{\eta/2\}}\right)^b = \left(\frac{\exp\{\eta\}}{exp\{\eta/2\}} + \frac{1}{exp\{\eta/2\}}\right)^b = (\exp\{\eta/2\}+\exp\{-\eta/2\})^b = (2\cosh(\eta/2))^b.$$

Thus,

$$\frac{(\exp\{\eta\})^a}{(1+\exp\{\eta\})^b} = \frac{(\exp\{\eta\})^k}{(2\cosh(\eta/2))^b} = 2^{-b}(\exp\{\eta\})^k.\cosh^{-b}(\eta/2).$$

Applying the Proposition 3.1.1, we conclude that

$$\frac{(\exp\{\eta\})^a}{(1+\exp\{\eta\})^b} = 2^{-b}(\exp\{\eta\})^k E\left[\exp\left\{-X\eta^2/2\right\}\right] = 2^{-b}\exp\{k\eta\}\int_0^\infty \exp\left\{-x\eta^2/2\right\} f(x|b,0)dx.$$

□

Theorem 3.1.1 provide us a different way to express the likelihood function of a logistic regression model, which is exactly the left side of Equation (3.6). In this next section we apply Theorem 3.1.1 in the likelihood of a logistic regression mixture model.

3.2 The Data Augmentation Strategy

In order to rewrite the likelihood function, and thus the conditional posterior distribution, we will extend and apply the data augmentation strategy presented in Polson, Scott and Windle (2013) to mixture of logistic regressions.

Consider $Y_1, \cdots, Y_n$ a random sample from a mixture of K Binomial distributions with N Bernoulli trials and success probability θ_{ij} for $i = 1, \cdots, K$ and $j = 1, \cdots, n$. In the regression context, we assume that for each response variable Y_j there is a vector of covariates $\mathbf{x}_j$ and a vector of parameters $\boldsymbol{\beta}_i$ so that the success probability θ_{ij} can be written through the inverse of the logistic function as

$$\theta_{ij} = \frac{\exp\{\mathbf{x}_j^\intercal \boldsymbol{\beta}_i\}}{1+\exp\{\mathbf{x}_j^\intercal \boldsymbol{\beta}_i\}} \tag{3.7}$$

for each component i, $i = 1, \cdots, K$. Then, we say that each Y_j is distributed as a mixture of logistic regression models whose the likelihood, according to the Section 2.2.1.1, is given by

$$f(\mathbf{y}|\mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\pi}) \propto \left[\prod_{i=1}^{K} \pi_i^{n_i}\right] \prod_{j:S_j=1} \frac{(\exp\{\mathbf{x}_j^\intercal \boldsymbol{\beta}_1\})^{y_j}}{(1+\exp\{\mathbf{x}_j^\intercal \boldsymbol{\beta}_1\})^N} \times \cdots \times \prod_{j:S_j=K} \frac{(\exp\{\mathbf{x}_j^\intercal \boldsymbol{\beta}_k\})^{y_j}}{(1+\exp\{\mathbf{x}_j^\intercal \boldsymbol{\beta}_K\})^N} \tag{3.8}$$

where n_i is the number of response variables in the component i or, in other words, the size of the component i, for $i = 1, \cdots, K$. Then we associate, for each Y_j in the component i, an auxiliary variable W_j that follows a Pólya-Gamma distribution with parameters $b_j = N$ and $c_j = 0$ so that we rewrite the likelihood applying the Theorem 3.1.1 to each product in (3.8) considering $b = N$, $a = y_j$ and $\eta = \mathbf{x}_j^\intercal \boldsymbol{\beta}_i$ for $i = 1, \cdots, K$ and $j = 1, \cdots, n$.

Considering the Theorem 3.1.1 , for a fixed component i, we have that

$$\begin{aligned}
f(\mathbf{y}_i, \mathbf{w}_i|\mathbf{S}, \boldsymbol{\beta}_i, \pi_i) &\propto \prod_{j:S_j=i} \left[\frac{(\exp\{\mathbf{x}_j^\intercal \boldsymbol{\beta}_i\})^{y_j}}{(1+\exp\{\mathbf{x}_j^\intercal \boldsymbol{\beta}_i\})^N}\right] \\
&= 2^{-Nn_i} \prod_{j:S_j=i} \left[\exp\{(y_j - N/2)\,\mathbf{x}_j^\intercal \boldsymbol{\beta}_i\} \exp\{-w_j(\mathbf{x}_j^\intercal \boldsymbol{\beta}_i)^2/2\} f(w_j|N,0)\right] \\
&\propto \prod_{j:S_j=i} \exp\{(y_j - N/2)\,\mathbf{x}_j^\intercal \boldsymbol{\beta}_i\} \exp\{-w_j(\mathbf{x}_j^\intercal \boldsymbol{\beta}_i)^2/2\} f(w_j|N,0) \\
&= \prod_{j:S_j=i} \exp\{(y_j - N/2)\,\mathbf{x}_j^\intercal \boldsymbol{\beta}_i - w_j(\mathbf{x}_j^\intercal \boldsymbol{\beta}_i)^2/2\} f(w_j|N,0) \\
&= \prod_{j:S_j=i} \exp\left\{-\frac{w_j}{2}\left((\mathbf{x}_j^\intercal \boldsymbol{\beta}_i)^2 - 2\left(\frac{y_j - N/2}{w_j}\right)\mathbf{x}_j^\intercal \boldsymbol{\beta}_i\right)\right\} f(w_j|N,0) \\
&= \prod_{j:S_j=i} \exp\left\{-\frac{w_j}{2}\left(\frac{y_j - N/2}{w_j} - \mathbf{x}_j^\intercal \boldsymbol{\beta}_i\right)^2\right\} \exp\left\{\frac{w_j}{2}\left(\frac{y_j - N/2}{w_j}\right)^2\right\} f(w_j|N,0) \\
&= \exp\left\{-\frac{1}{2}(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}_i)^\intercal \mathbf{W}_i(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}_i)\right\} \exp\left\{\frac{1}{2}\mathbf{z_i}^\intercal \mathbf{W}_i \mathbf{z}_i\right\} \prod_{j:S_j=i} f(w_j|N,0),
\end{aligned} \tag{3.9}$$

where $\mathbf{z}_i^\intercal = \left(\frac{(y_{i1} - N/2)}{w_{i1}}, \cdots, \frac{(y_{in_i} - N/2)}{w_{in_i}}\right)$, the matrix $\mathbf{X}_i$ is the design matrix for the component i and the matrix $\mathbf{W}_i$ is a diagonal matrix containing the Pólya-Gamma random variables associated to the response variables in the component i. Note that, when applying the Theorem 3.1.1, we did not integrate out the Pólya-Gamma variables as in (3.6), since the goal is to rewrite the likelihood as a function of $\mathbf{w}$ as well.

Through Equations (3.8) and (3.9) we rewrite the likelihood of the mixture as

$$f(\mathbf{y},\mathbf{w}|\mathbf{S},\boldsymbol{\beta},\boldsymbol{\pi}) \propto \prod_{i=1}^{K} \pi_i^{n_i} \left[\exp\left\{ -\frac{1}{2}(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}_i)^\intercal \mathbf{W}_i (\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}_i) \right\} \right.$$
$$\left. \times \exp\left\{ \frac{1}{2}\mathbf{z_i}^\intercal \mathbf{W}_i \mathbf{z}_i \right\} \prod_{j:S_j=i} f(w_j|N,0) \right]. \tag{3.10}$$

The main consequence is that when joining the likelihood in (3.10) with the Normal prior of the regression coefficients, there is conjugation, since the second term in (3.10) does not depend on $\boldsymbol{\beta}_i$. It means that we can apply the classical MCMC algorithm Gibbs Sampling for sampling from the posterior distribution of the regression coefficients. This conjugation also allows us to calculate analytically the marginal likelihood which simplifies and make the variable selection efficient, as we will see later.

The derivation of the Equation (3.9) reveals us one way of updating the Pólya-Gamma variables at the MCMC iterations. The conditional posterior distribution of W_j, for $j = 1, \cdots, n$, can be found using the Equation (3.9) by considering the terms that do not depends on w_j as a constant, obtaining the conditional posterior distribution of W_j as

$$\begin{aligned} p(w_j|\cdot) &\propto \exp\{\left(y_j - {}^{N}/_{2}\right)\mathbf{x}_j^\intercal\boldsymbol{\beta}_i\}\exp\{-w_j(\mathbf{x}_j^\intercal\boldsymbol{\beta}_i)^2/2\}f(w_j|N,0) \\ &\propto \exp\{-w_j(\mathbf{x}_j^\intercal\boldsymbol{\beta}_i)^2/2\}f(w_j|N,0), \end{aligned} \tag{3.11}$$

which is proportional to the density of $PG(b,c)$ in Equation (3.5) with $c = \mathbf{x}_j^\intercal\boldsymbol{\beta}_i$ and $b = N$. Thus, we can consider that

$$W_j|\cdot \sim PG(N, \mathbf{x}_j^\intercal\boldsymbol{\beta}_i). \tag{3.12}$$

The estimation and variable selection for a mixture of logistic regressions is described in the next section, as well as the choice of prior distributions of the regression coefficients, which plays an important role to the variable selection.

CHAPTER

4

BAYESIAN ESTIMATION AND VARIABLE SELECTION

Mixture models are sometimes referred as the model where "the number of things you do not know is one of the things you do not know", since in most applications, as well as the parameters, the number of components K is unknown. The simplest case is when both number of components K and the allocation vector $\mathbf{S} = (S_1, \cdots, S_n)$ is known. In this case, the only concern is the estimation of the regression coefficients of each component and the weights. This can done by simply allocating the observations according to $\mathbf{S}$ and then applying the Bayesian estimation for a logistic regression in each component individually.

For the case where the number K of components is known but the allocation vector $\mathbf{S} = (S_1, \cdots, S_n)$ is unknown, the estimation process is no longer so straightforward. In this chapter we describe the Bayesian estimation process of a mixture model of logistic regressions as well as the Bayesian variable selection approach considered in this work. The Bayesian estimation process is discussed considering the full model, without selecting variables. The variable selection will be discussed further, with respect to the prior distributions of the regression coefficients that will be considered for the selection of the variables.

4.1 Bayesian Estimation of the Full Model

The estimation of a mixture model from the Bayesian point of view requires specification of the prior distributions to the parameters. The prior distribution usually assigned to the mixing probabilities is the Dirichlet distribution, that is, we assume

$$\boldsymbol{\pi} \sim Dirichlet(\alpha_1, \cdots, \alpha_K) \tag{4.1}$$

so that the posterior distribution, according to Appendix A.1, is given by

$$\boldsymbol{\pi}|\mathbf{S} \sim Dirichlet(n_1 + \alpha_1, \cdots, n_K + \alpha_K) \tag{4.2}$$

where n_i is the size of the component i for $i = 1, \cdots, K$, and $\alpha_1, \cdots, \alpha_k$ are known parameters.

The posterior distribution for the allocation variables $S_1, \cdots, S_n$ is derived through the Bayes theorem, computing the probability of $S_j = i$ given that we observed the event $Y_j = y_j$ and the regression coefficients $\boldsymbol{\beta}_i$ as

$$\begin{aligned} P(S_j = i|y_j, \boldsymbol{\pi}, \boldsymbol{\beta}_i) &= \frac{f(y_j|\boldsymbol{\pi}, \boldsymbol{\beta}_i)P(S_j = i|\boldsymbol{\pi})}{\sum_{h=1}^{K} f(y_j|\boldsymbol{\pi}, \boldsymbol{\beta}_h)P(S_j = h|\boldsymbol{\pi})} \\ &= \frac{f(y_j|\boldsymbol{\pi}, \boldsymbol{\beta}_i)\pi_i}{\sum_{h=1}^{K} f(y_j|\boldsymbol{\pi}, \boldsymbol{\beta}_h)\pi_h}, \end{aligned} \tag{4.3}$$

for $j = 1, \cdots, n$ and $i = 1, \cdots, K$.

For the regression coefficients in each component we consider a normal prior distribution as

$$\boldsymbol{\beta}_i \sim N(\boldsymbol{\mu}_i, \Sigma_i), \tag{4.4}$$

where $\boldsymbol{\mu}_i$ and Σ_i are the vector of means and variance-covariance matrix of the regression coefficients in the component i. The posterior distribution, according to the Appendix A.2, is also a normal distribution, that is,

$$\boldsymbol{\beta}_i|\cdot \sim N(\mathbf{m}_i, \mathbf{V}_i) \tag{4.5}$$

where $\mathbf{V}_i = \left(\Sigma_i^{-1} + \mathbf{X}_i^{\intercal}\mathbf{W}_i\mathbf{X}_i\right)^{-1}$ and $\mathbf{m}_i = \mathbf{V}\left(\mathbf{X}_i^{\intercal}\mathbf{W}_i\mathbf{z}_i\right)$.

Note that the update of the regression coefficients in each component depends on the update of the Pólya-Gamma latent variables added in the likelihood. This update is done in two steps, first sampling from Pólya-Gamma distribution and then sampling from the distribution of the regression coefficients, as in the following scheme

$$\begin{aligned} W_j|\cdot &\sim PG(N, \mathbf{x}_j^{\intercal}\boldsymbol{\beta}_i), \\ \boldsymbol{\beta}_i|\cdot &\sim N(\mathbf{m}_i, \mathbf{V}_i). \end{aligned} \tag{4.6}$$

It is important to make it clear that the data augmentation strategy with Pólya-Gamma distribution is applied only to the sampling process of the regression coefficients of each component. So that for sampling from the posterior distributions in (4.6) a Gibbs sampling can be applied. Details are presented in Section 4.1.1.

The estimation of the number of components for a mixture model can be done either through model selection criteria or simultaneously with the other parameters, assigning a prior distribution to K. In this work, the estimation of the number of components will be done through the model selection criteria presented in the Section 4.1.2.

4.1.1 Gibbs sampling to Estimate the Full Model

In this section we describe the Gibbs sampling algorithm for Bayesian estimation of the full model. The Gibbs sampling algorithm works by successively sampling the parameters from

their posterior distribution in (4.2), (4.3) and (4.6) a large enough number of times, obtaining a posterior sample that will contain all the relevant information about the regression coefficients, the weights and the allocation variables. For sampling from the Pólya-Gamma distribution, Polson, Scott and Windle (2013) provides a sampling method that is implemented in the R package `BayesLogit`.

The Gibbs sampling applied in this work is described in the following steps,

Step 1: Initialize the weights by sampling from the prior distribution $\boldsymbol{\pi}^{(0)} \sim Dirichlet(\alpha_1, \cdots, \alpha_K)$;

Step 2: Initialize the allocation variables $\mathbf{S}^{(0)}$ by sampling from its prior distribution $S_j^{(0)} \sim Discrete(\boldsymbol{\pi}^{(0)})$ for $j = 1, \cdots, n$;

Step 3: Initialize each $\boldsymbol{\beta}_i^{(0)}$ by sampling $\boldsymbol{\beta}_i^{(0)} \sim N(\boldsymbol{\mu}_i, \Sigma_i)$ for $i = 1, \cdots, K$;

Step 4: At each iteration l, after allocating each observation according to $\mathbf{S}^{(l-1)}$, we update the Pólya-Gamma latent variables and the regression coefficients of each component i by sampling

$$W_j^{(l)}|\cdot \sim PG(N, \mathbf{x}_j^\intercal \boldsymbol{\beta}_i^{(l-1)}), \tag{4.7}$$

$$\boldsymbol{\beta}_i^{(l)}|\cdot \sim N(\mathbf{m}_i, \mathbf{V}_i); \tag{4.8}$$

Step 5: Update the weights $\boldsymbol{\pi}^{(l)}$ and the allocation variables $\mathbf{S}^{(l)}$ by sampling from

$$\boldsymbol{\pi}^{(l)}|\mathbf{S}^{(l-1)} \sim Dirichlet(n_1 + \alpha_1, \cdots, n_K + \alpha_K) \tag{4.9}$$

and from

$$P(S_j^{(l)} = i|y_j, \boldsymbol{\pi}^{(l)}, \boldsymbol{\beta}_i^{(l)}) = \frac{f(y_j|\boldsymbol{\pi}^{(l)}, \boldsymbol{\beta}_i^{(l)})\pi_i^{(l)}}{\sum_{h=1}^K f(y_j|\boldsymbol{\pi}^{(l)}, \boldsymbol{\beta}_h^{(l)})\pi_h^{(l)}} \tag{4.10}$$

and return to the Step 4.

After obtaining the posterior sample of $\boldsymbol{\pi}$ and each $\boldsymbol{\beta}_i$, the point estimates can be obtained by taking the mean of the posterior sample, which is the optimal Bayesian estimator with respect to the quadratic loss function. From the I iterations, we discard the first B iterations as burn-in period and consider J jumps between two recorded iterations to obtain a non-correlated sample. So that the final size of each posterior sample is $I_{final} = \dfrac{(I-B)}{J}$. Thus, the point estimates of $\pi_1, \cdots, \pi_K$ are computed as

$$\hat{\pi}_i = \frac{1}{I_{final}} \sum_{h=1}^{I_{final}} \pi_i^{(h)}. \tag{4.11}$$

The same idea is applied to compute the point estimates of each $\boldsymbol{\beta}_i$. For $i = 1, \cdots, K$ and $t = 1, \cdots, p$, the point estimate of β_{it} is given by

$$\hat{\beta}_{it} = \frac{1}{I_{final}} \sum_{h=1}^{I_{final}} \beta_{it}^{(h)}. \tag{4.12}$$

From the posterior sample of the allocation variables $S_1, \cdots, S_n$ we can compute a point estimate of the probability that an observation y_j belongs to each component and use them to classification. Let N_{ij} be the number of times that y_j was allocated to the component i for $i = 1, \cdots, K$. For each y_j, the point estimate of the probability that y_j belongs to component i is given by

$$\hat{P}(S_j = i|\cdot) = \frac{N_{ij}}{I_{final}}. \tag{4.13}$$

After obtaining the point estimates $\hat{P}(S_j = i|\cdot)$ for each y_j, we use them to randomly classify y_j into the component i.

This summarization only works if we do not observe label switching problem in the samples or if it was already corrected as we discuss below in Section 4.1.3.

4.1.2 Estimating the Number of Components

The Gibbs sampling algorithm previously presented consider that the number K of components in the mixture is known. However, this is not a realistic situation. In this work, the estimation of the number of components in the mixture will be done through Bayesian model selection criteria, namely Deviance Information Criterion (DIC) proposed by Spiegelhalter *et al.* (2014) and Extended Deviance Information Criterion (EBIC) proposed by Chen and Chen (2008). Both of them are composed by two terms, the first accounts for goodness-of-fit and the second penalizes the complexity of the model, through its number of parameters.

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}})$ be the Bayesian point estimates of the parameters, $\hat{\mathbf{S}}$ the predicted allocation variables and d_K the number of parameters of the model. Considering the Deviance function given by $D(\boldsymbol{\theta}) = -2\log(f(\mathbf{y}|\mathbf{S}, \boldsymbol{\theta}))$, the DIC and EBIC criterion are calculated as

$$\text{DIC} = D(\hat{\boldsymbol{\theta}}) + 2p_D \tag{4.14}$$

and

$$\text{EBIC} = \bar{D}(\boldsymbol{\theta}) + d_K \log(n), \tag{4.15}$$

where $\bar{D}(\boldsymbol{\theta})$ is the average of the Deviance function calculated in the parameters $\boldsymbol{\theta}$ and $\mathbf{S}$ of each iteration of the final chain and p_D is a measure of the effective number of parameters, that can be estimated from the data by

$$\hat{p}_D = \bar{D}(\boldsymbol{\theta}) - D(\hat{\boldsymbol{\theta}}). \tag{4.16}$$

For selecting the appropriate model, we fit the mixture model for $K \in \{1, 2, \cdots, K_{max}\}$ and compute the criteria previously presented for each one of them. The model chosen is the one that minimize the criteria.

4.1.3 Label Switching Problem

One of the challenges when dealing with Bayesian estimation of mixture models is the label switching problem. To understand the root of this problem, consider the incomplete

likelihood of the model, given by

$$f(\mathbf{y}|\boldsymbol{\pi},\boldsymbol{\beta}) = \prod_{j=1}^{n} \pi_1 f(y_j|\boldsymbol{\beta}_1) + \cdots + \pi_K f(y_j|\boldsymbol{\beta}_K)\,, \tag{4.17}$$

where $\boldsymbol{\beta}^\intercal = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K)$ and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$. For each permutation $\rho = (\rho_1, \cdots, \rho_K)$ of $\{1, \cdots, K\}$, it is possible to obtain new vectors $\boldsymbol{\beta}_\rho^\intercal = (\boldsymbol{\beta}_{\rho_1}, \cdots, \boldsymbol{\beta}_{\rho_K})$ and $\boldsymbol{\pi}_\rho = (\pi_{\rho_1}, \cdots, \pi_{\rho_K})$, obtained by permuting the original vector $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ through $\boldsymbol{\rho}$, so that the likelihood in (4.17) written under the permutation in $\boldsymbol{\beta}_\rho$ and in $\boldsymbol{\pi}_\rho$ will be the same as under $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$. To see that, consider $K = 3$ and $\rho = (1,3,2)$. In this case, $\boldsymbol{\beta}_\rho^\intercal = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_3, \boldsymbol{\beta}_2)$ and $\boldsymbol{\pi}_\rho = (\pi_1, \pi_3, \pi_2)$, then

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\pi}_\rho, \boldsymbol{\beta}_\rho) &= \prod_{j=1}^{n} \left[\pi_1 f(y_j|\boldsymbol{\beta}_1) + \pi_3 f(y_j|\boldsymbol{\beta}_3) + \pi_2 f(y_j|\boldsymbol{\beta}_2)\right] \\ &= \prod_{j=1}^{n} \left[\pi_1 f(y_j|\boldsymbol{\beta}_1) + \pi_2 f(y_j|\boldsymbol{\beta}_2) + \pi_3 f(y_j|\boldsymbol{\beta}_3)\right] \\ &= f(\mathbf{y}|\boldsymbol{\pi},\boldsymbol{\beta}). \end{aligned} \tag{4.18}$$

Moreover, given the allocation $\mathbf{S} = (S_1, \cdots, S_n)$ of observations, the complete likelihood is invariant when permuting the components as well. In the same way, if there is no prior knowledge about the difference among the components, the prior assigned to the components parameters will be the same and consequently, the prior distribution will also be invariant when permuting the label of the components. Hence, the posterior distribution will inherit the invariance of the likelihood function and prior distribution.

During the MCMC draws, the labels of the components can permute many times over iterations and due to the invariance discussed above, the final MCMC sample obtained will not be useful to make inference about the components. This characterizes the label switching problem. It is worth to note that the label switching problem is also an identifiability problem, since different vectors of parameters (permutations of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$) may leads to the same model.

In order to correct the whole MCMC sample from the label switching problem, an appropriate permutation must to be applied to each MCMC draw. Many algorithms have been proposed in the literature to correct label switching. The R package `label.switching` introduced by Papastamoulis (2016) provides many of them. In this work, the Equivalence Classes Representatives (ECR) algorithm, originally proposed by Papastamoulis and Iliopoulos (2010), is applied. The ECR algorithm search for the permutation, for each MCMC iteration, that makes its prediction of $\mathbf{S}$ as close as possible to the prediction of the first iteration excluding the burn in period and jumps.

4.2 Bayesian Variable Selection

Consider the observations $(y_j, \mathbf{x}_j)$, for $j = 1, \cdots, n$, from a population where each Y_j follows a mixture of logistic regressions given in Definition 2.2.2, where $\mathbf{x}_j$ is the vector of

covariates associated to y_j. Fixing the number of components K, let $\mathbf{y}_i$ and $\mathbf{X}_i$ be the vector of response variables and design matrix for the observations allocated in component i. The variable selection problem is to select, among the covariates $x_1, \cdots, x_p$, a subset $\{x_1^*, \cdots, x_d^*\}$ of relevant covariates to explain the success probability in each component of the mixture.

We start by introducing indicator latent variables $\boldsymbol{\gamma}_i^\intercal = (1, \gamma_{i1}, \cdots, \gamma_{ip})$ associated to the parameter vector $\boldsymbol{\beta}_i$ in the component i, so that $\gamma_{it} = 1$ if $\beta_{it} \neq 0$, and $\gamma_{it} = 0$ if $\beta_{it} = 0$ for $t = 1, \cdots, p$. In other words, $\gamma_{it} = 1$ indicates that the covariate x_t is relevant for observations in component i.

For each component, a natural choice of prior distribution for γ_{it} is the Bernoulli distribution. Considering independence, the prior distribution of the vector $\boldsymbol{\gamma}_i$ is given by

$$p(\boldsymbol{\gamma}_i) = \prod_{t=1}^{p} p_{it}^{\gamma_{it}} (1 - p_{it})^{1-\gamma_{it}}. \tag{4.19}$$

The main goal is to obtain the marginal posterior of each $\boldsymbol{\gamma}_i$ that will contain all the relevant information to select the best covariates.

The update of the indicator latent variable of vector $\boldsymbol{\gamma}_i$ is done through the posterior probability of accepting a covariate as relevant. This probability is calculated considering the marginalized likelihood function, integrating out $\boldsymbol{\beta}_{\gamma_i}$, that represents the vector of regression coefficients with $\gamma_{it} = 1$ or non-zero coefficients (see Appendix A.3). Then, the posterior probability of accepting a covariate as relevant does not depend on the value of its coefficient. For each γ_{it} we calculate the posterior probability of $\gamma_{it} = 1$ and $\gamma_{it} = 0$ as

$$P(\gamma_{it} = 1 | \mathbf{y}, \mathbf{S}, \boldsymbol{\pi}, \mathbf{w}, \boldsymbol{\gamma}_{i(-t)}) \propto f(\mathbf{y}_i | \mathbf{S}_i, \mathbf{w}_i, \boldsymbol{\pi}, \boldsymbol{\gamma}_i) p_{it} \tag{4.20}$$

and

$$P(\gamma_{it} = 0 | \mathbf{y}, \mathbf{S}, \boldsymbol{\pi}, \mathbf{w}, \boldsymbol{\gamma}_{i(-t)}) \propto f(\mathbf{y}_i | \mathbf{S}_i, \mathbf{w}_i, \boldsymbol{\pi}, \boldsymbol{\gamma}_i)(1 - p_{it}), \tag{4.21}$$

where $\boldsymbol{\gamma}_{i(-t)}^\intercal = (1, \gamma_{i1}, \cdots, \gamma_{it-1}, \gamma_{it+1}, \cdots, \gamma_{ip})$. Under this specification, we update only those regression coefficients β_{it} for which $\gamma_{it} = 1$ at each MCMC iteration, because $\beta_{it} = 0$ by definition when $\gamma_{it} = 0$. This method has been widely applied in Bayesian variable selection (GEORGE; MCCULLOCH, 1997; LEE; CHEN; WU, 2016; CAO; LEE; HUANG, 2020). Figure 3 shows an example of how we update the regression coefficients at each iteration considering their indicators. The blue balls represent those regression coefficients that have indicator 1 and thus are updated. The white balls are those regression coefficients with null indicator and thus they are set to zero.

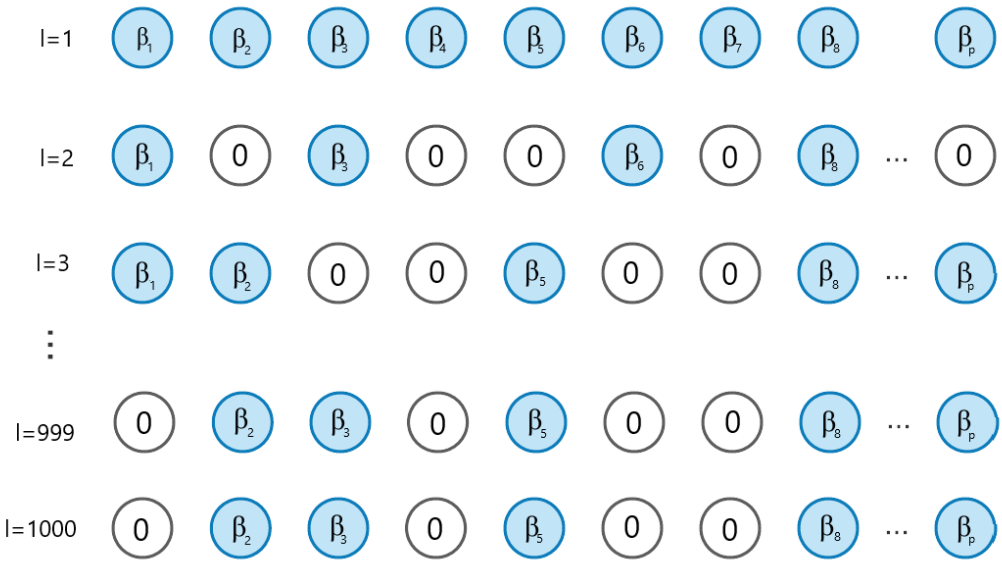

Figure 3 – Updating scheme of the regression coefficients. The coefficients that have indicator equals to 0 are set to be zero in the white balls.

To complete the specification of the variable selection process, it remains to specify the prior distribution of $\boldsymbol{\beta}_{\gamma_i}|\boldsymbol{\gamma}_i$. This choice also plays an important role in the selection. The prior distributions considered in this work are discussed in the following sections.

So far, we have not yet discussed the inclusion of each intercept term in the variable selection, which is usually "excluded" from the model by centralizing the covariates. Here, without loss of generality, we will treat the intercept term as a regression coefficient of a covariate x_0 that is always presents in the model, and the selection will be done only for the covariates x_t with $t = 1, \cdots, p$.

4.2.1 Spike and Slab Prior

The first variable selection approach considered in this work is the Stochastic Search Variable Selection (SSVS) method. First introduced by Mitchell and Beauchamp (1988) and further improved by George and McCulloch (1993), George_ and McCulloch (1996) and Ishwaran and Rao (2005), the SSVS method aims to stochastically search for the best set of covariates through a mixture prior with a spike and slab components. The spike component aims to shrink those coefficients β_{it} with small effect in the model, and has its mass concentrated at zero. The slab component, on the other hand, aims to sample plausible values for the coefficients with significant effect to the model, having its mass spread over a wide range of value.

There are basically two types of spike and slab prior proposed in the literature. The one presented by George and McCulloch (1993), in which the spike component follows a normal distribution centred at zero, that is

$$p(\beta_{it}|\gamma_{it}) = (1-\gamma_{it})N(0,\tau_i^2) + \gamma_{it}N(0,\sigma_i^2), \tag{4.22}$$

where τ_i^2 must be small and σ_i^2 must be large, since when $\gamma_{it} = 0$ then β_{it} is likely to be close to zero and if $\gamma_{it} = 1$, a non-close to zero estimate would be more appropriate to the coefficient β_{it}. And the one presented by Kuo and Mallick (1998), in which the spike component is a point of mass at zero, that is,

$$p(\beta_{it}|\gamma_{it}) = (1-\gamma_{it})\mathbf{1}(\beta_{it} = 0) + \gamma_{it}N(0, \sigma_i^2), \tag{4.23}$$

where σ_i^2 is usually large.

In this work we consider the spike and slab prior with point of mass at zero, given by Equation (4.23). However, when sampling only the regression coefficients with $\gamma_{it} = 1$, as previously commented, and assuming the spike and slab prior distribution in Equation (4.23) for the regression coefficients, it is equivalent to assuming that

$$\boldsymbol{\beta}_{\gamma_i}|\boldsymbol{\gamma}_i \sim N(\mathbf{0}, \sigma_i^2\mathbf{I}), \tag{4.24}$$

where $\boldsymbol{\beta}_{\gamma_i}$ represents the vector of regression coefficients with $\gamma_{it} = 1$ and σ_i^2 is the variance of the regression coefficients in the component i.

Thus, considering the likelihood in (3.10), the conditional posterior distribution of $\boldsymbol{\beta}_{\gamma_i}$ is calculated as in Appendix A.2 with $\Sigma_h = \sigma_i^2\mathbf{I}$ and $\boldsymbol{\mu}_h = \mathbf{0}$, obtaining

$$\begin{aligned} p(\boldsymbol{\beta}_{\gamma_i}|\cdot) &\propto \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\beta}_{\gamma_i} - \mathbf{m}\right)^{\mathsf{T}}\mathbf{V}^{-1}\left(\boldsymbol{\beta}_{\gamma_i} - \mathbf{m}\right)\right]\right\} \\ &\propto Normal\left(\mathbf{m}, \mathbf{V}\right), \end{aligned} \tag{4.25}$$

where $\mathbf{X}_{\gamma_i}$ contains only the covariates with $\gamma_{it} = 1$, $\mathbf{V} = \left(\left(\sigma_i^2\mathbf{I}\right)^{-1} + \mathbf{X}_{\gamma_i}^{\mathsf{T}}\mathbf{W}_i\mathbf{X}_{\gamma_i}\right)^{-1}$ and $\mathbf{m} = \mathbf{V}\left(\mathbf{X}_{\gamma_i}^{\mathsf{T}}\mathbf{W}_i\mathbf{z}_i\right)$.

In general, the hyperparameter σ_i^2 is chosen as a large value in order to obtain a vague prior. Moreover, it is also common to assign a prior distribution to σ_i^2, which is usually the Gamma or Inverse Gamma distribution. For the spike and slab prior in Equation (4.22), the choice of the hyperparameters τ_i and σ_i is discussed with details in George and McCulloch (1993).

4.2.2 g-Prior

The last choice of prior distribution for the regression coefficients considered in this work is the g-prior. First introduced by Zellner (1986), the g-prior for each $\boldsymbol{\beta}_{\gamma_i}$ is given by

$$\boldsymbol{\beta}_{\gamma_i}|\boldsymbol{\gamma}_i \sim N(\mathbf{0}, g_i\sigma_i^2\left(\mathbf{X}_{\gamma_i}^{\mathsf{T}}\mathbf{X}_{\gamma_i}\right)^{-1}) \tag{4.26}$$

where g_i is a constant and $\mathbf{X}_{\gamma_i}$ is the design matrix containing only those covariates x_t in which $\gamma_{it} = 1$ for $t = 1, \cdots, p$, that is, the relevant covariates for the model in the component i.

A very common problem when using the Zellner's g-prior is the singularity of the matrix $(\mathbf{X}_{\gamma_i}^{\mathsf{T}}\mathbf{X}_{\gamma_i})$. In models where either the number of observations is lower than the number of

covariates or the covariates are correlated, the matrix $(\mathbf{X}_{\gamma_i}^\intercal \mathbf{X}_{\gamma_i})$ is singular. Following Baragatti and Pommeret (2012), one way to avoid this problem is to add a ridge hyperparameter $\lambda > 0$ so that the matrix $(\mathbf{X}_{\gamma_i}^\intercal \mathbf{X}_{\gamma_i})$ is replaced by $(\mathbf{X}_{\gamma_i}^\intercal \mathbf{X}_{\gamma_i} + \lambda \mathbf{I})$, and the prior considered to $\boldsymbol{\beta}_{\gamma_i}$ is given by

$$\boldsymbol{\beta}_{\gamma_i} | \boldsymbol{\gamma}_i \sim N(\mathbf{0}, g_i \sigma_i^2 \left(\mathbf{X}_{\gamma_i}^\intercal \mathbf{X}_{\gamma_i} + \lambda \mathbf{I}\right)^{-1}). \tag{4.27}$$

Considering the likelihood given by (3.10) and the g-prior in (4.27), the conditional posterior distribution of $\boldsymbol{\beta}_{\gamma_i}$ is calculated as in Appendix A.2 considering $\boldsymbol{\Sigma}_h = \left(g_i \sigma_i^2 \left(\mathbf{X}_{\gamma_i}^\intercal \mathbf{X}_{\gamma_i} + \lambda \mathbf{I}\right)^{-1}\right)$ and $\boldsymbol{\mu}_h = \mathbf{0}$, obtaining

$$\begin{aligned} p(\boldsymbol{\beta}_{\gamma_i} | \cdot) &\propto \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\beta}_{\gamma_i} - \mathbf{m}\right)^\intercal \mathbf{V}^{-1} \left(\boldsymbol{\beta}_{\gamma_i} - \mathbf{m}\right)\right]\right\} \\ &\propto Normal\left(\mathbf{m}, \mathbf{V}\right) \end{aligned} \tag{4.28}$$

where $\mathbf{V} = \left(\left(g_i \sigma_i^2 \left(\mathbf{X}_{\gamma_i}^\intercal \mathbf{X}_{\gamma_i} + \lambda \mathbf{I}\right)^{-1}\right)^{-1} + \mathbf{X}_{\gamma_i}^\intercal \mathbf{W}_i \mathbf{X}_{\gamma_i}\right)^{-1}$ and $\mathbf{m} = \mathbf{V}\left(\mathbf{X}_{\gamma_i}^\intercal \mathbf{W}_i \mathbf{z}_i\right)$. To sample from the posterior in (4.28) the Gibbs sampling algorithm is applied.

The choice of the hyperparameter g_i in the g-prior plays an important role in the variable selection. Liang *et al.* (2008) provides a review of the choices of g_i in Bayesian variable selection. These choices include $g_i \in \{p^2, n_i, max(n_i, p^2)\}$. For the mixture model case, simulations studies in Lee, Chen and Wu (2016) suggest to take $g_i = n_i$ when p/n is less than 3 and $g_i = 100 \cdot p \cdot K/n$ otherwise. In the other hand, Gupta and Ibrahim (2007) suggested taking g_i above 100 to ensure a vague prior. In both works, a Inverse Gamma prior was assign to the hyperparameter σ_i^2. The possible choices to the ridge hyperparameter λ can be found in Baragatti and Pommeret (2012), that suggests $\lambda = 1/p$.

4.2.3 Gibbs sampling to Variable Selection

After rewriting the likelihood as in (3.10) and combining with the priors presented in the previous sections, a Gibbs sampling can be applied to sample from the posterior distributions in (4.25) and (4.28). The Gibbs sampling for variable selection follows the same idea of the algorithm presented in the Section 4.1.1, now including the sampling of the indicator latent variables γ_{it}. The Gibbs sampling algorithm to variable selection is described below.

Step 1: Initialize the weights by sampling from the prior distribution $\boldsymbol{\pi}^{(0)} \sim Dirichlet(\alpha_1, \cdots, \alpha_K)$;

Step 2: Initialize the allocation variables $\mathbf{S}^{(0)}$ by sampling from its prior distribution $S_j^{(0)} \sim Discrete(\boldsymbol{\pi}^{(0)})$ for $j = 1, \cdots, n$;

Step 3: Initialize the indicator variables $\boldsymbol{\gamma}_i^{(0)}$ and then each $\boldsymbol{\beta}_{\gamma_i}^{(0)}$ by sampling $\boldsymbol{\beta}_{\gamma_i}^{(0)} \sim N(\boldsymbol{\mu}_i, \Sigma_i)$ and Pólya-gamma variables $\mathbf{W}_i$ for $i = 1, \cdots, K$;

Step 4: At each iteration l, after allocating each observation according to $\mathbf{S}^{(l-1)}$, for $t = 1, \cdots, p$, we update the Pólya-Gamma latent variables by sampling

$$W_j^{(l)}|\cdot \sim PG(N, \mathbf{x}_j^\intercal \boldsymbol{\beta}_{\gamma_i}^{(l-1)}); \tag{4.29}$$

Step 5: Update the indicator variables γ_{it} sampling from a Bernoulli distribution with success probability given by the posterior probability $P(\gamma_{it}^{(l)} = 1|\mathbf{y}, \mathbf{S}^{(l-1)}, \boldsymbol{\pi}^{(l-1)}, \mathbf{w}^{(l)}, \boldsymbol{\gamma}_{i(-t)}^{(l-1)})$ of keeping the covariate x_t in the model of component i;

Step 6 Update the regression coefficients with $\gamma_{it}^{(l)} = 1$ of each component i by sampling

$$\boldsymbol{\beta}_{\gamma_i}^{(l)}|\cdot \sim N(\mathbf{m}_i, \mathbf{V}_i); \tag{4.30}$$

Step 7: Update the weights $\boldsymbol{\pi}^{(l)}$ and the allocation variables $\mathbf{S}^{(l)}$ by sampling from

$$\boldsymbol{\pi}^{(l)}|\mathbf{S}^{(l-1)} \sim Dirichlet(n_1 + \alpha_1, \cdots, n_K + \alpha_K) \tag{4.31}$$

and from

$$P(S_j^{(l)} = i|y_j, \boldsymbol{\pi}^{(l)}, \boldsymbol{\beta}_{\gamma_i}^{(l)}) = \frac{f(y_j|\boldsymbol{\pi}^{(l)}, \boldsymbol{\beta}_{\gamma_i}^{(l)})\pi_i^{(l)}}{\sum_{h=1}^{K} f(y_j|\boldsymbol{\pi}^{(l)}, \boldsymbol{\beta}_{\gamma_h}^{(l)})\pi_h^{(l)}}, \tag{4.32}$$

and then return to the Step 4.

After obtaining the posterior sample of the indicator variables $\boldsymbol{\gamma}_i$ of each component i, $i = 1, \cdots, K$, we select the relevant covariates based on their posterior inclusion probability. From the I iterations, we discard the first B iterations as burn-in period and consider J jumps between two recorded iterations to obtain a non-correlated sample. So that the final size of each posterior sample is $I_{final} = \frac{(I-B)}{J}$. The posterior inclusion probability is calculate as

$$\hat{P}(\gamma_{it} = 1|\cdot) = \frac{1}{I_{final}} \sum_{l=1}^{I_{final}} \mathbf{1}(\gamma_{it}^{(l)} = 1), \tag{4.33}$$

for $i = 1, \cdots, K$ and $t = 1, \cdots, p$. To finally select the important covariates, we adopt the Median Probability Criterion (BARBIERI; BERGER, 2004), that classifies a covariate as relevant if $\hat{P}(\gamma_{it} = 1|\cdot) \geq 0.5$.

Once the important covariates were selected, the point estimates of the regression coefficients associated to each selected covariate is calculated as in the Equation (4.12). In the same way, the point estimates of the weights are calculated as in the Equation (4.11) and the classification of the observations follows the same idea as in the Equation (4.13). As mentioned before, this summarization, including the selection of relevant variables, only works if we do not observe label switching problem in the samples or if it was already corrected.

CHAPTER

5

SIMULATION STUDY

This section illustrates the performance of the methods for selecting covariates in simulated data. The Section 5.1 is intended for the simulations of the methodology presented in Section 4.1 for the full model without variable selection. However, we can discuss the relevance of each covariate if the zero value is present or not in the associated regression coefficient's credibility interval. The second section brings simulation results of the variable selection methodology presented in the Section 4.2. The measures considered to assess the performance in each case are discussed within each section.

5.1 Estimation of the Full Model

In the simulation of the full model, the goal is to assess the estimation of the number of components, the goodness of fit through the obtained estimates, the classification rate and convergence. This assessment will be done assuming different scenarios to the generated data, in order to compare the performance of the methodology in each case.

For every scenario of simulation, we run 30 replications of data. In each replication, we select the number of components fitting the model for $K = 1, 2, \cdots, 4$ and apply the model selection criteria DIC and EBIC. For assess the goodness of fit, we calculate the bias of the estimates at each replication. The Highest Posterior Density credibility interval (HPD) is also considered. In this work, the HPD interval is calculated as the mean of the lower bound and the upper bound of the HPD intervals obtained at each replication, except for the TCO defined below.

Following Lee, Chen and Wu (2016), to measure the capacity of classifying observations, the True Classification of Observations (TCO) rate was computed at each replication as

$$\text{TCO} = \frac{\text{number of correct classification of observations}}{\text{the number of observations}}.$$

To verify convergence we analyse the log-likelihood of the model through the Geweke's convergence diagnostic (COWLES; CARLIN, 1996).

5.1.1 Scenario 1

In this first scenario, the data of the 30 replications was generated from a mixture of Binomial distributions with $N = 50$ Bernoulli trials, $K = 3$ components and $p = 5$ covariates, simulated from a standard normal distribution, with regression coefficients given by

$$\begin{aligned} \boldsymbol{\beta}_1^\top &= (1,-1,0,1,0), \\ \boldsymbol{\beta}_2^\top &= (-1,0,1,0,1) \quad \text{and} \\ \boldsymbol{\beta}_3^\top &= (-0.5,0,-0.5,0,-0.5). \end{aligned} \tag{5.1}$$

The sample size considered was $n = 200$ and the weights were fixed as $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$. Figure 4 shows how the smoothed histogram of a simulated data looks like under these definitions. In this figure we see that the components are evident and reasonably separated from each other. For each replicate we ran 65000 iterations with 5000 of burn-in period and jumps of 10.

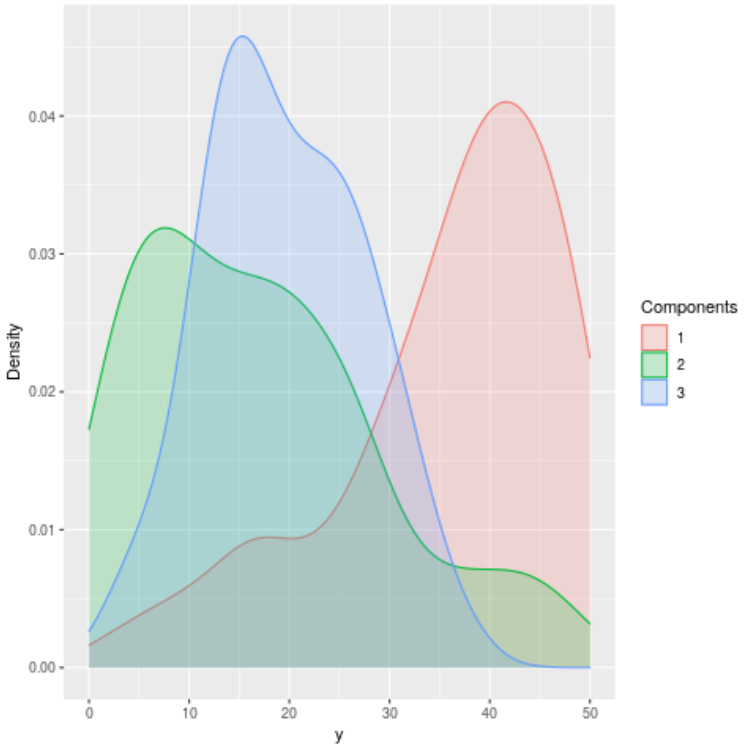

Figure 4 – Smoothed histogram for a data simulated in scenario 1.

Figure 5 presents the boxplot of the Geweke's diagnostic of convergence of the fitted models with different number of components. The Geweke's diagnostic indicates convergence when its value is in the interval $(-1.96, 1.96)$. According to this convergence diagnostic we see that some replications did not show convergence. However, this number is small, for $K = 3$ only

one presents an outlier value. These replications were not considered to summarize the results that will be presented posteriorly.

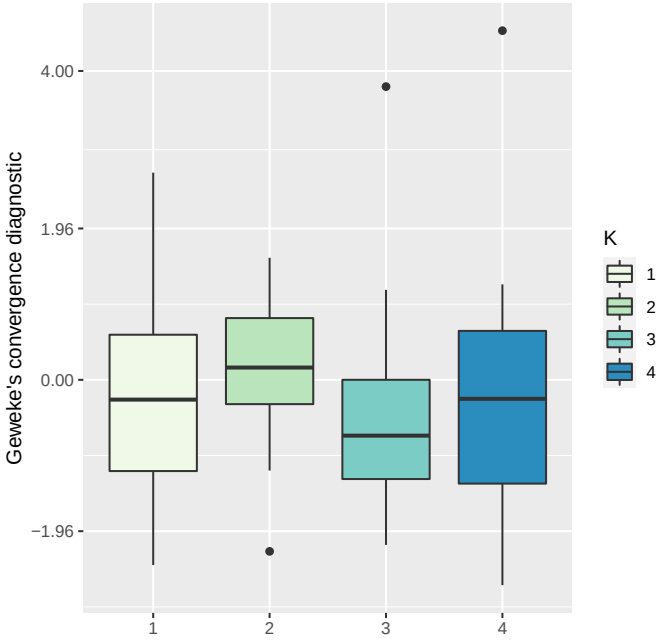

Figure 5 – Scenario 1: Geweke's convergence diagnostic of the 30 models.

For the estimation of the number of components, Table 1 presents the average of the criterion DIC and EBIC for each value of K. The criterion EBIC selected the correct model in all replications, whereas the DIC selected the correct model only in 28% of replications, tending to select the model with $K = 4$. However, the models with $K = 4$ had a $\hat{\pi}_4$ very close to zero with few observations allocated to it. Besides that, the estimates of the regression coefficients of the model with $K = 4$ components reveal that probably the algorithm created a fourth component only to allocate possible outliers. Based on these results, we selected the model with $K = 3$ components.

Table 1 – Scenario 1: Average of criteria to estimate K and their correct estimation percentage.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	2727.75	1621.03	-833.83	-3683.03	28
EBIC	2754.53	1678.34	1501.14	1533.72	100

The estimates of the regression coefficients and the weights of the model with $K = 3$ are shown in Table 2 as well as its Highest Posterior Density credibility interval (HPD). These estimates are the average of the obtained estimates at each replication. It is possible to see that the proposed algorithm could estimates well the regression coefficient and the weights. Besides that, the credibility intervals of the regression coefficients associated to covariates that had no effect on the response variable contains the zero value, expect for β_4 of the component 1.

Table 2 – Scenario 1: Estimate, true value and credibility interval for the parameters of the full model.

	Component 1	Component 2	Component 3
β_0	0.78 (1) (0.52,1.04)	-0.99 (-1) (-1.11,-0.88)	-0.39 (-0.5) (-0.59,-0.19)
β_1	-1.03 (-1) (-1.17,-0.89)	0.02 (0) (-0.09,0.14)	-0.01 (0) (-0.13,0.10)
β_2	0.01 (0) (-0.11,0.14)	1.01 (1) (0.90,1.13)	-0.49 (-0.5) (-0.60, 0.38)
β_3	1.03 (1) (0.90,1.17)	0.00 (0) (-0.11,0.11)	0.00 (0) (-0.12,0.11)
β_4	-0.23 (0) (-0.46,-0.01)	0.66 (1) (0.25,1.09)	-0.62 (-0.5) (-0.80,-0.43)
π	0.33 (0.33) (0.25, 0.40)	0.33 (0.33) (0.24, 0.40)	0.34 (0.33) (0.27, 0.44)

To better assess the goodness of fit we also analyze the bias of the estimates of the regression coefficients. Figure 6 shows the boxplot of the biases obtained at each replication. It is evident that the bias of the estimates are distributed very close to zero.

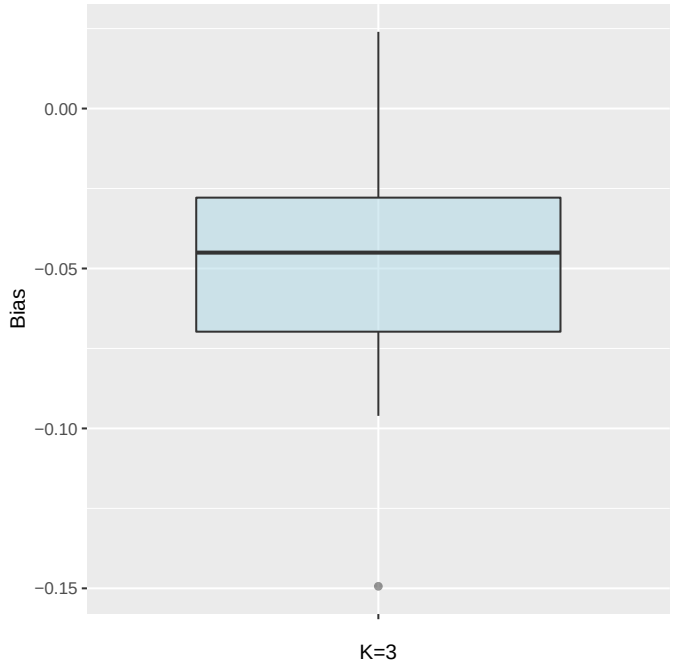

Figure 6 – Scenario 1: Bias of the estimates of the model with $K = 3$ components in each replication.

The ability of the algorithm to cluster and classify observations was also investigated. The median of the TCO was 74.5% with HPD interval of $(10, 82)\%$. It is not so close to 100% because there is a large intersection between components 2 and 3 and smaller intersection between components 2 and 1.

From the results we conclude that the proposed algorithm had a good performance in both estimating the regression coefficients, weights and classifying observations. Besides that, only the model selection criterion EBIC identified the correct number of components.

5.1.2 Scenario 2

In this second scenario the goal is to investigate the performance of the algorithm to estimate a mixture of logistic regression with similar regression coefficients. The data of the 30 replications was generated from a mixture of Binomial distributions with $N = 50$ Bernoulli trials, $K = 3$ components and $p = 5$ covariates, simulated from a standard normal distribution, with regression coefficients given by

$$\begin{aligned} \boldsymbol{\beta}_1^\top &= (1,-4,0,2,0), \\ \boldsymbol{\beta}_2^\top &= (-1,-3,0,1,0) \quad \text{and} \\ \boldsymbol{\beta}_3^\top &= (-1,4,0,-2,0). \end{aligned} \tag{5.2}$$

The sample size considered was $n = 200$ and the weights were fixed as $\boldsymbol{\pi} = (0.3, 0.4, 0.3)$. Figure 7 shows how the smoothed histogram of a simulated data looks like under these definitions. Note that the smoothed histogram of the data for each component has a very similar shape. For each replicate we kept the number of iteration as 65000, with 5000 of burn-in period and jumps of 10.

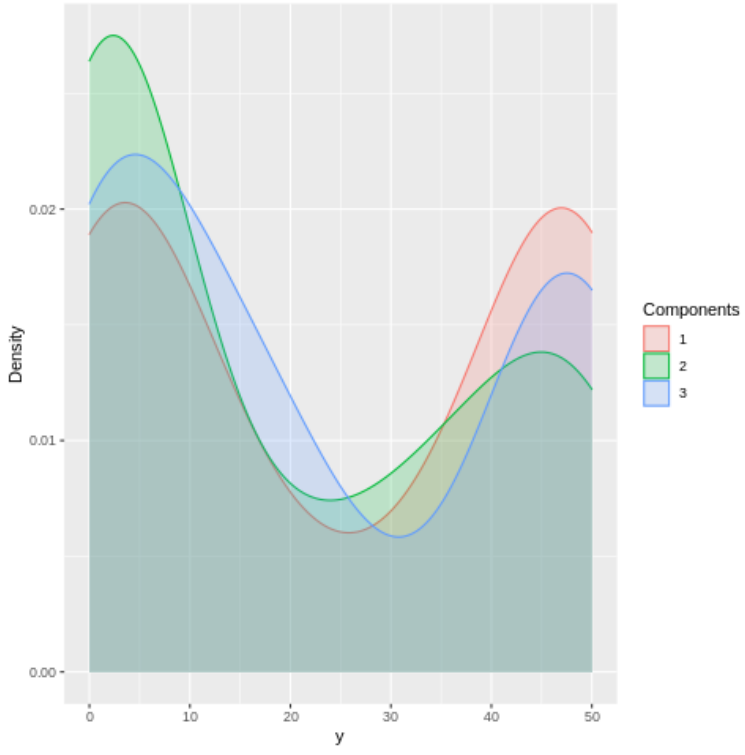

Figure 7 – Smoothed histogram for a data simulated in scenario 2.

Figure 8 show the Geweke's diagnostic of convergence of the models for each value

of K. Again, the chain of few replications did not show convergence, and thus, they were not considered to the results that will be shown later.

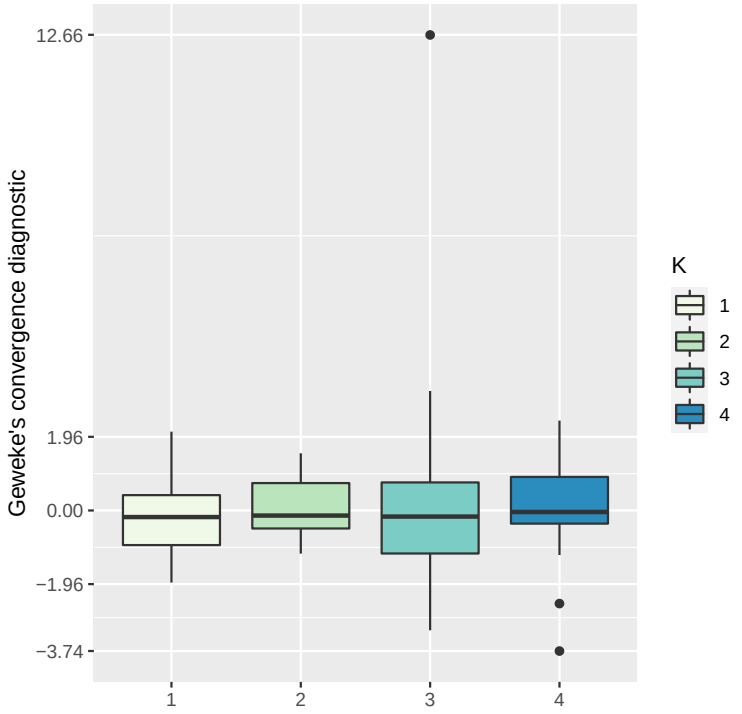

Figure 8 – Scenario 2: Geweke's convergence diagnostic of the replications of the models.

In the estimation of the number of components, the EBIC criterion selected the correct model in 60% of the replications. In remaining replications, the criterion selected the model with $K = 2$. The DIC criterion selected the correct model only in 30% of replications, selecting the model with $K = 4$ components in 65% of the replications. But the models with $K = 4$ had a $\hat{\pi}_4$ very close to zero with few observations allocated to it. And also, the estimates of the regression coefficients of the model with $K = 4$ components reveal that probably the algorithm created a fourth component only to allocated possible outliers. The average of the criterion and their correct estimation percentage are presented in Table 3. Based on these results, we selected the model with $K = 3$.

Table 3 – Scenario 2: Average of criteria to estimate K and their correct estimation percentage.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	7510.14	750.16	-12092.27	-19740.86	30
EBIC	7536.92	1296.19	1199.44	1241.84	60

The estimates of the regression coefficients and the weights considering $K = 3$ are shown in Table 4. The algorithm had a good performance in estimating the regression coefficients, except for a few estimates. In this case, some estimates of the regression coefficients associated

to covariates with no effect in the model was not so close to zero as in the scenario previously presented. However, their credibility intervals contains zero, expect for β_4 of the component 2.

Table 4 – Scenario 2: Estimate, true value and credibility interval for the parameters of the full model.

	Component 1	Component 2	Component 3
β_0	0.49 (1) (-0.27,1.25)	-0.99 (-1) (-1.16, -0.82)	-0.98 (-1) (-1.47,-0.50)
β_1	-3.99 (-4) (-5.09, -2.90)	-3.03 (-3) (-3.30, -2.76)	4.02 (4) (3.58,4.47)
β_2	-0.01 (0) (-0.3,0.29)	-0.01 (0) (-0.14, 0.12)	-0.03 (0) (-0.23, 0.17)
β_3	2.12 (2) (1.71,2.53)	0.99 (1) (0.81,1.18)	-2.02 (-2) (-2.28,-1.76)
β_4	-0.55 (0) (-1.35,0.25)	-0.63 (0) (-0.96,-0.3)	0.00 (0) (-0.36, 0.36)
π	0.26 (0.3) (0.17, 0.36)	0.44 (0.4) (0.34, 0.54)	0.29 (0.3) (0.26, 0.36)

The boxplot of the biases associated with fitted models models with $K = 3$ in Figure 9 shows that the bias is distributed close to zero, indicating a good fitting of the model.

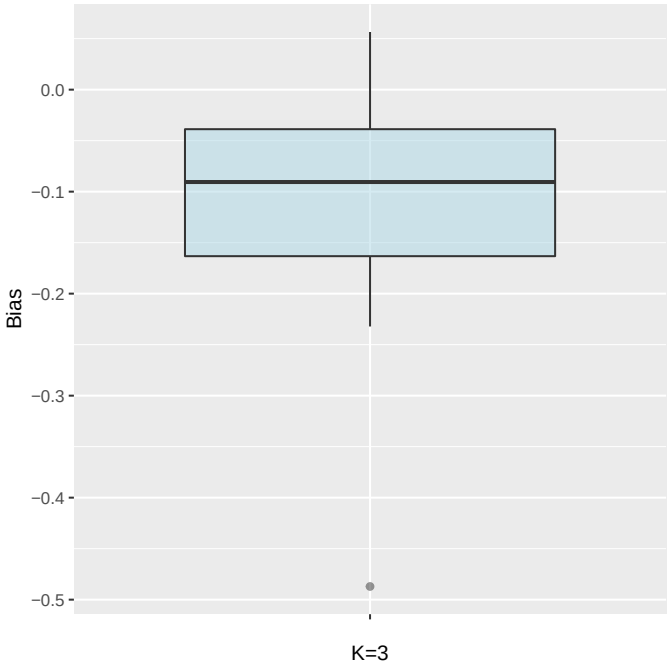

Figure 9 – Scenario 2: Bias of the estimates of the model with $K = 3$ components in each replication.

Finally, we analyze the capacity of classifying observations in this scenario. The median of the TCO was 73.5% with HPD interval of $(4, 88)\%$.

Estimating mixture of similar components can be challenging, since similar components can make it difficult for the algorithm to identify the components and estimate their parameters. The results presented reveal that the algorithm had a good performance in this case.

5.1.3 Scenario 3

In this third scenario the goal was to investigate the performance of the algorithm in the case where $p > n_i$, that is, the number of covariates is greater than the number of observations. To do this, the data of 30 replications was generated as in the Scenario 1, from a mixture of $K = 3$ Binomial distributions with $N = 50$, increasing the number of covariates to $p = 10$ in each component, simulated from a standard normal distribution, with their respective coefficients given by

$$\begin{aligned} \boldsymbol{\beta}_1^\top &= (1,-1,0,1,0,-1,0,1,0,-1), \\ \boldsymbol{\beta}_2^\top &= (-1,0,1,0,1,0,1,0,1,0) \quad \text{and} \\ \boldsymbol{\beta}_3^\top &= (-0.5,0,-0.5,0,-0.5,0,0.5,0,-0.5,0). \end{aligned} \tag{5.3}$$

The sample size was fixed as $n = 30$ and the weights fixed as $\boldsymbol{\pi} = (0.3, 0.3, 0.4)$. In Figure 10 we can see how the estimated densities are very similar and close with the fixed parameters. We also kept 6500 iterations with 5000 iterations as a burn-in period and jumps of 10.

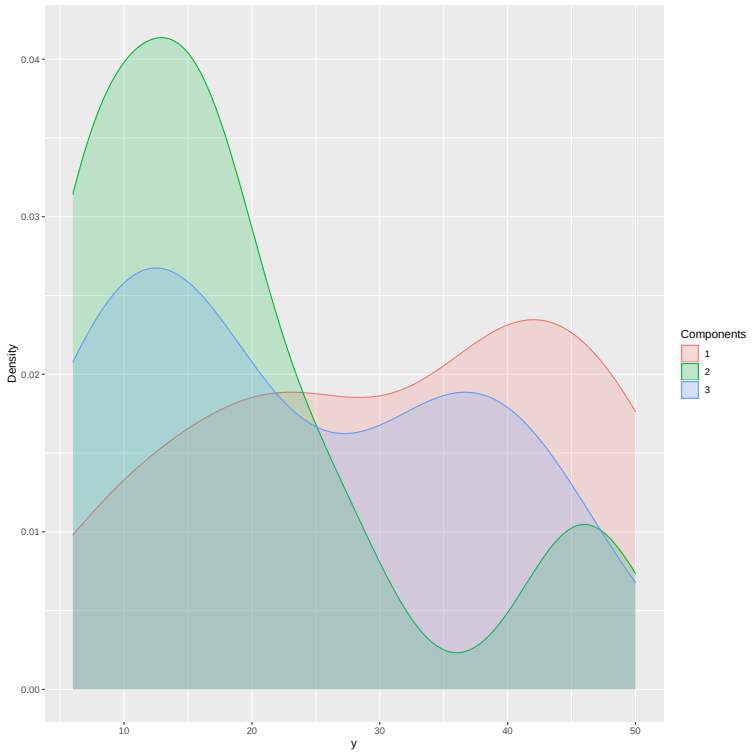

Figure 10 – Smoothed histogram for a data simulated in scenario 3.

The Geweke's convergence diagnostic is shown in Figure 11, where is possible to see that some replications did not converge, with some outliers for $K = 2$.

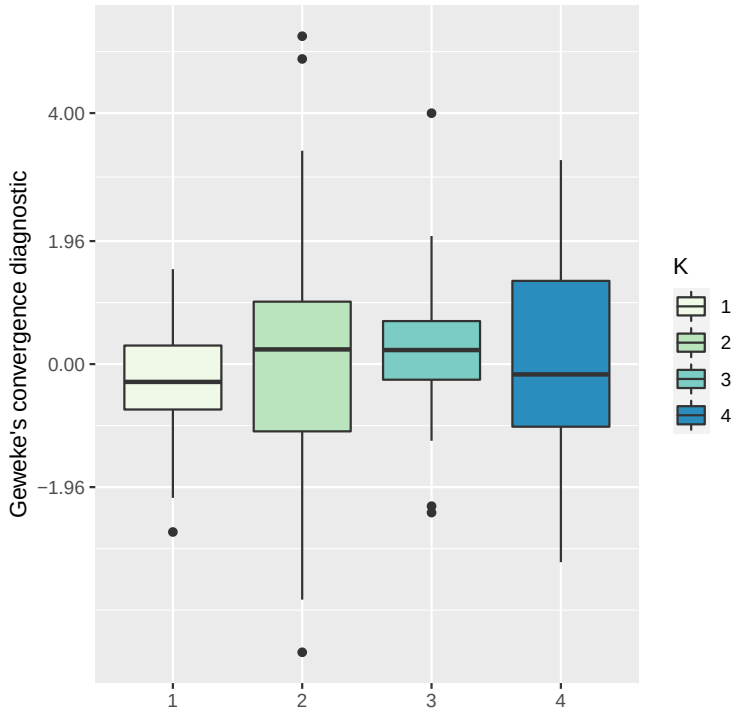

Figure 11 – Scenario 3: Geweke's convergence diagnostic of the replications of the models.

In the estimation of the number of components, the DIC selected the correct model only in 30.78% of replications, whereas both models with $K = 2$ and $K = 4$ were selected 34.61% of replications. The criterion EBIC selected the model with $K = 2$ components in all replications. The results are shown in Table 5. Note that, considering the results of both criteria, the selected model is the one with $K = 2$ components.

Table 5 – Scenario 3: Average of criteria to estimate K and their correct estimation percentage.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	425.04	-777.03	-971.34	-1173.45	30.78
EBIC	452.44	249.47	294.80	340.20	0

The results of the estimation of the regression coefficients and the weights of the model with $K = 2$ components are summarized in Table 6. From the estimates, it seems that the method could not identify the component 3, merging the components 1 and 3 into one. Besides that, the credibility interval obtained for the regression coefficients and weights are large. Although the sample size is small, the main consequence of the large credibility interval is high variance, which leads to low precision.

Figure 12 shows the boxplot of the obtained biases, which are still distributed close zero. The bias was calculated considering the true regression coefficients of components 1 and 2.

Table 6 – Scenario 3: Estimates, true value and credibility interval for the parameters of the full model.

	Component 1	Component 2
β_0	0.51 (1) (-4.81, 5.48)	0.12 (-1) (-6.77, 6.75)
β_1	-0.90 (-1) (-5.99, 4.45)	-0.23 (0) (-7.21, 6.00)
β_2	0.43 (0) (-4.85, 5.78)	0.72 (1) (-6.21, 7.54)
β_3	0.14 (1) (-5.21, 6.24)	-0.08 (0) (-6.32, 6.31)
β_4	0.29(0) (-5.59, 5.82)	0.29 (1) (-6.08, 6.58)
β_5	-0.57 (-1) (-6.45, 4.66)	0.12 (0) (-7.51, 6.81)
β_6	0.49 (0) (-5.05, 6.52)	0.95 (1) (-5.50, 7.58)
β_7	0.69 (1) (-4.31, 5.94)	0.47 (0) (-5.88, 7.62)
β_8	-0.22 (0) (-5.97, 5.43)	0.82 (1) (-5.70, 7.27)
β_9	-1.96 (-1) (-6.91, 3.96)	-1.36 (0) (-8.80, 5.21)
π	0.52 (0.3) (0.25, 0.80)	0.48 (0.3) (0.20, 0.75)

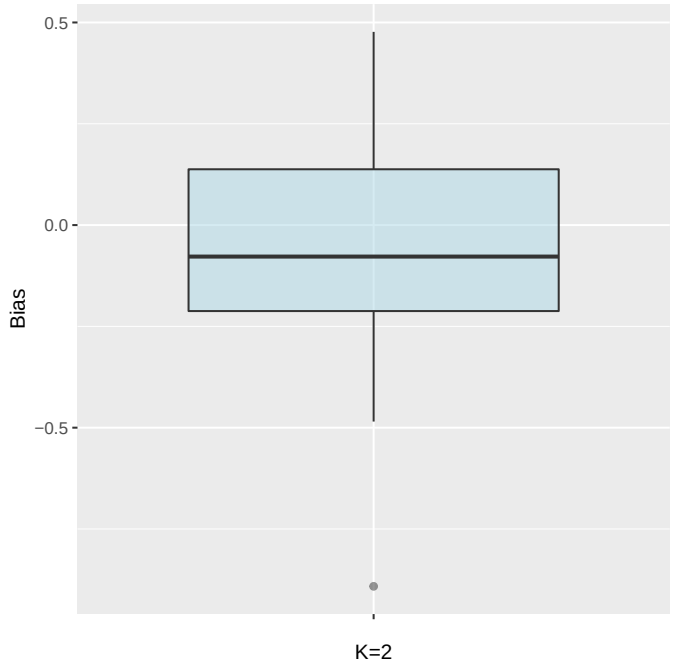

Figure 12 – Scenario 3: Bias of the estimates of the model with $K = 2$ components in each replication.

The classification results also reveal the merging of components 1 and 3 since around 63% of the observations from component 3 was allocated into component 1. The median of the TCO was 33.33% with HPD interval of $(20,60)\%$.

Through these results we realize that the algorithm did not fit the data so well. The model selection criteria could not identify the correct model and consequently the correct classification rate was not so high. Due the small number of observations, the algorithm could not capture enough information to identify the 3 components and their parameters, resulting in low precision in the estimates and favoring the model with less components.

5.2 Estimation of the Model with Variable Selection

Following Lee, Chen and Wu (2016), the variable selection performance in each component will be measured through the True Positive Rate (TPR) and the False Positive Rate (FPR), that are calculated as

$$\text{TPR} = \frac{\text{number of correctly selected variables}}{\text{the number of active variables}}$$

and

$$\text{FPR} = \frac{\text{number of incorrectly selected variables}}{\text{the number of inactive variables}}.$$

Thus, values of TPR close to one and FPR close to zero indicate a good performance of variable selection. The measures used in the simulation of the full model in the Section 5.1 will also be applied to the simulation with the variable selection. For every method of variable selection, we run 30 replications of data and for each replication we kept the number of iterations as 65000 with burn-in period of 5000 iterations and jumps of 10. The relevant covariates will be selected by applying the Median Probability Criterion to the average of the posterior inclusion probability obtained in the replications. To select the number of component we fit the model for $K = 1,2,\cdots,4$, and apply the model selection criteria DIC and EBIC for selecting the best value of K. The goal of this simulation is to assess and compare the performance of the two methods presented in Section 4.2 to select variable in a mixture of logistic regressions.

5.2.1 Scenario 1:

In this first scenario, the data was generated as in Scenario 5.1.1, from a mixture of $K = 3$ logistic regression model where the response variable follows a Binomial distribution with $N = 50$ Bernoulli trials and $p = 5$ covariates. Their respective regression coefficients are given by

$$\begin{aligned} \boldsymbol{\beta}_1^\intercal &= (1,-1,0,1,0), \\ \boldsymbol{\beta}_2^\intercal &= (-1,0,1,0,1) \quad \text{and} \\ \boldsymbol{\beta}_3^\intercal &= (-0.5,0,-0.5,0,-0.5). \end{aligned} \tag{5.4}$$

The sample size was fixed as $n = 200$ and the weights fixed as $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$. The hyperparameters of the prior distributions was chosen as follows. For the g-prior, we set $g = n_i$ as commented in the Section 4.2.2, and $\sigma_i^2 = 1$ for $i = 1, \cdots, K$. The ridge parameter was fixed as $\lambda = 1/p$. For the spike and slab prior distribution we fixed $\sigma_i^2 = 100$ for $i = 1, \cdots, K$. A summary of the hyperparameters of this simulation is shown in Table 7. The prior probability p_{it} of keeping the covariate x_t in the model for every component was fixed as $p_{it} = 0.5$.

Table 7 – Scenario 1: Summary of the hyperparameters of the prior distributions.

Prior	Hyperparameters
Spike and Slab	$\sigma_i^2 = 100$
g-prior	$g = n_i$, $\sigma_i^2 = 1$, $\lambda = 1/p$

The results of the variable selection with each method are discussed in the next subsections, followed by a final discussion of their performance in this scenario.

5.2.1.1 Spike and Slab prior

After running the 30 replications we assess their convergence through the Geweke's convergence diagnostic shown in Figure 13. We observe that, in general, the chains converged, with some outliers in the models with $K = 2$.

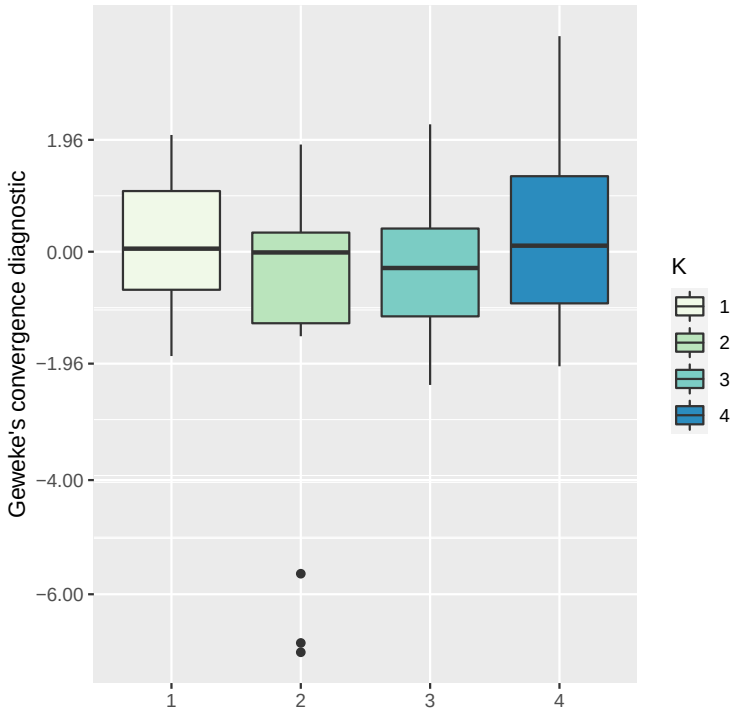

Figure 13 – Scenario 1: Geweke's convergence diagnostic of the replications of the models with the spike and slab prior.

In the estimation of the number of components, the criterion DIC selected the correct model only in 31.25% of replications and the model with $K = 4$ in 62.5% of replications. However, the models with $K = 4$ components had a estimate of π_4 very close to zero and did not select any covariates. By contrast with the DIC, the criterion EBIC selected the correct model in 100% of replications. The results of the criteria is summarized in Table 8. Based on them, we select the model with $K = 3$ components.

Table 8 – Scenario 1: Average of criteria to estimate K and their correct estimation percentage obtained with spike and slab prior.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	2716.55	1555.05	-1855.27	-2973.14	31.25
EBIC	2743.26	1688.30	1505.34	1540.51	100

Analysing posterior inclusion probability, that is, $\hat{P}(\gamma_{it} = 1)$ for $i = 1,\ldots,K$ and $t = 1,\ldots,p$, obtained in the replications we observed that the covariate associated to the regression coefficient β_4 of component 2 was excluded of the model in 29% of the replications. Similarly, in the component 1, the covariate associated to the coefficient β_4 was included in the model in 22% of the replications. Figure 14 presents the boxplots of TPR and FPR obtained in the replications.

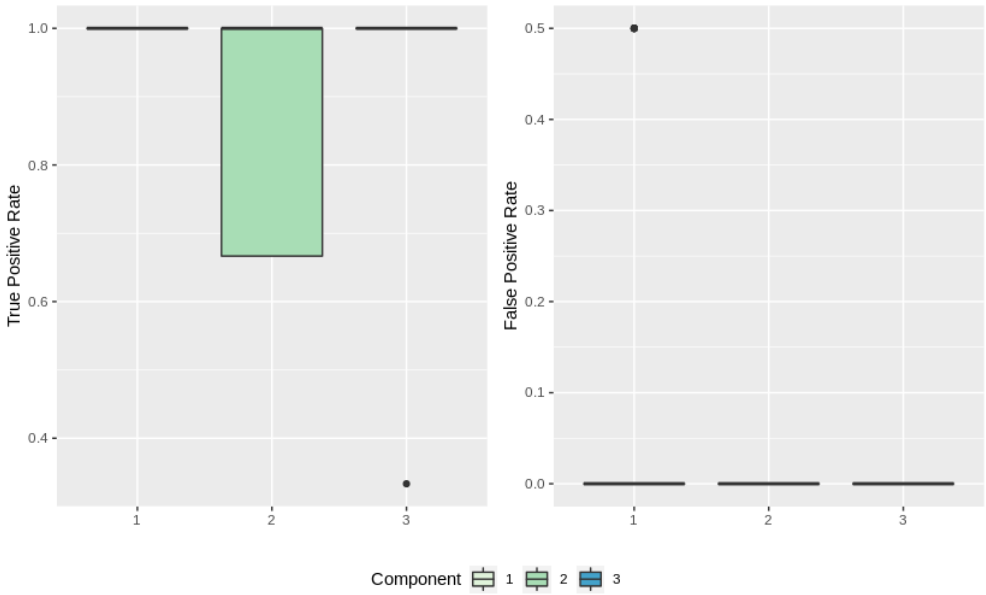

Figure 14 – Scenario 1: False and True Positive Rate obtained from the selection with the spike and slab prior.

The estimates of the regression coefficients associated to the selected covariates are shown in Table 9. In it we see that the covariates were selected correctly in each component, with good estimates. The estimate of β_4 of component 2 is not so close to the true value because its posterior inclusion probability was lower than 50% in 29% of the replications, occasioning its exclusion of the model, as previously commented.

Table 9 – Scenario 1: Estimates, true value and credibility interval for the parameters of the model with spike and slab prior.

	Component 1	Component 2	Component 3
β_0	0.92 (1) (0.71,1.12)	-0.99(-1) (-1.10,-0.88)	-0.37 (-0.5) (-0.57,-0.18)
β_1	-1.04 (-1) (-1.17,-0.92)	-	-
β_2	-	1.01 (1) (0.89,1.13)	-0.5 (-0.5) (-0.6, -0.39)
β_3	1.03 (1) (0.90,1.16)	-	-
β_4	-	0.59 (1) (0.19,1)	-0.64 (-0.5) (-0.81,-0.46)
π	0.32 (0.33) (0.24, 0.39)	0.32 (0.33) (0.25, 0.40)	0.36 (0.33) (0.28, 0.44)

To confirm the goodness of fit, the boxplot of the biases associated to the fitted models with $K = 3$ is shown in Figure 15, where it is possible to see that the bias is distributed close to zero.

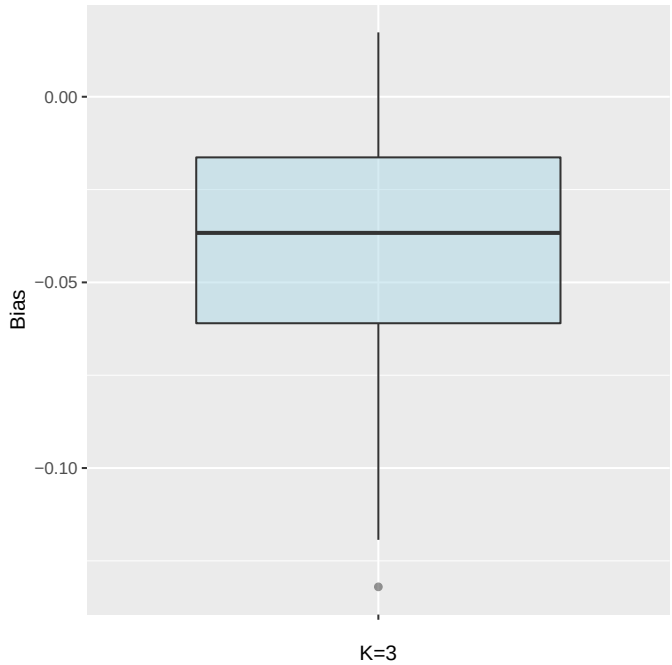

Figure 15 – Scenario 1: Bias of the estimates of the model with $K = 3$ components in each replication with the spike and slab prior.

We finally analyse the performance of the method for classifying observations through the TCO. The median of the TCO was 75.5% with HPD interval $(10, 83)\%$, which is also a good

rate of correct classification. Note that these results are very similar to the Simulation 5.1.1 of the full model, as expected.

5.2.1.2 g-Prior

For the simulation with the g-prior, the Geweke's convergence diagnostic is shown in Figure 16. Some replications did not converge, however, this number is still small. In the model with $K = 3$ components, only one replication did not converge.

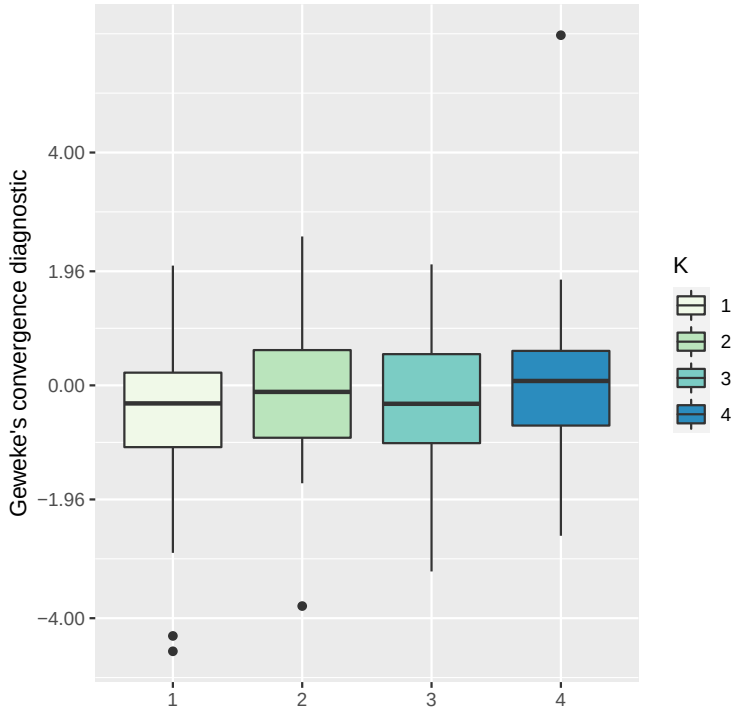

Figure 16 – Scenario 1: Geweke's convergence diagnostic of the replications of the models with the g-prior.

In the estimation of the number of components, the DIC criterion selected the correct model in 26% of the replications and selected the model with $K = 4$ in the most of replications. It is worth to point out that in the model with $K = 4$ components, the fourth component had only two selected covariates and their respective regression coefficients had estimates close to zero. Beside that, the estimate of π_4 was also close to zero. The EBIC criterion had a great performance in the model's selection, choosing the correct model 100% of replications. Table 10 presents the results of each criterion. Based on these results we choose the model with $K = 3$ components.

Table 10 – Scenario 1: Average of criteria to estimate K and their correct estimation percentage obtained with g-prior.

Criterion	$K=1$	$K=2$	$K=3$	$K=4$	%
DIC	2680.62	1579.72	-738.97	-1759.71	0.26
EBIC	2707.49	1672.46	1504.69	1568.81	100

In the selection of the predictor variables, the g-prior had a good performance. In Figure 17 we see the boxplots of the TPR and FPR obtained in each component. When checking the posterior inclusion probability of the covariates in the replications, it was observed that the covariate associated to the regression coefficient β_4 of component 2 was excluded of the model in 17% of replications. In the component 1, the covariate associated to the regression coefficient β_4 was included in the model of some replications. More precisely, it was included in 41% of the replications.

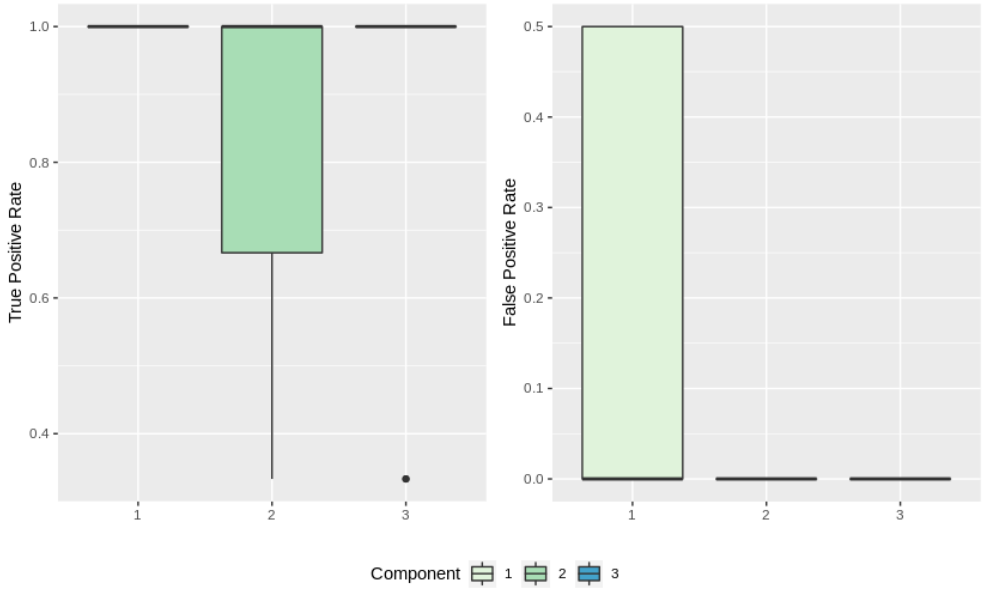

Figure 17 – Scenario 1: False and True Positive Rate obtained from the selection with the g-prior.

In Table 11 we see the estimates of coefficients for selected covariates of each component and the weights, with their respective credibility interval. Note that the estimates are close to the true value of the both regression coefficients and the weights, except for the regression coefficient β_4 in the component 2, that had posterior inclusion probability lower than 50% in 17% of the replications as previously commented.

The boxplot of the observed biases is shown in Figure 18, which are still distributed close to zero.

Table 11 – Scenario 1: Estimates, true value and credibility interval for the parameters of the model with by g-prior.

	Component 1	Component 2	Component 3
β_0	0.82 (1) (0.59,1.05)	-0.98(-1) (-1.09,-0.86)	-0.39 (-0.5) (-0.60,-0.17)
β_1	-1.03 (-1) (-1.17,-0.90)	-	-
β_2	-	1.0 (1) (0.88,1.12)	-0.49 (-0.5) (-0.60, -0.38)
β_3	1.02 (1) (0.89,1.15)	-	-
β_4	-	0.41 (1) (0.07,0.74)	-0.62 (-0.5) (-0.82,-0.42)
π	0.32 (0.33) (0.24, 0.40)	0.33 (0.33) (0.25, 0.41)	0.35 (0.33) (0.27, 0.43)

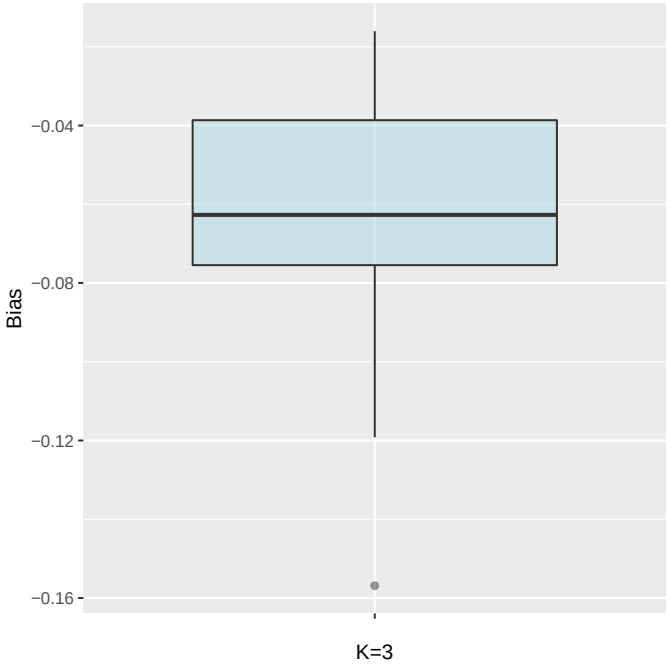

Figure 18 – Scenario 1: Bias of the estimates of the model with $K = 3$ components in each replication with the g-prior.

The performance of the method in classifying observations with the g-prior was also analysed. The median of TCO was 75%, which indicates a good performance. Its credibility interval is given by $(9,82)\%$.

To summarize and compare the results of the variable selection in the scenario 1 with the different prior distributions, Table 12 presents the average of the TPR and FPR obtained from the replications of each model with variable selection.

Table 12 – Scenario 1: Average of TPR and FPR.

Prior	Component 1		Component 2		Component 3	
	TPR	FPR	TPR	FPR	TPR	FPR
Spike and Slab	1.0	0.11	0.90	0.0	0.97	0.0
g-prior	1.0	0.21	0.90	0.0	0.95	0.0

Note that, based on the TPR and FPR values, the performance of the two prior distributions to the variable selection in the scenario 1 was very similar. Both of them presented some issue in selecting variables in the component 1. However, the prior distributions had a very good performance to select the correct covariates in all components.

5.2.2 Scenario 2:

This second scenario is generated as the Simulation 5.1.2, with the goal of investigating the performance of the algorithm to select the variables in the case where the logistic regressions of the components have similar regression coefficients. The data of the 30 replications was generated from a mixture of Binomial distributions with $N = 50$ Bernoulli trials, $K = 3$ components and $p = 5$ covariates, simulated from a standard normal distribution, with regression coefficients given by

$$\begin{aligned} \boldsymbol{\beta}_1^\intercal &= (1,-4,0,2,0), \\ \boldsymbol{\beta}_2^\intercal &= (-1,-3,0,1,0) \quad \text{and} \\ \boldsymbol{\beta}_3^\intercal &= (-1,4,0,-2,0). \end{aligned} \tag{5.5}$$

The sample size considered was $n = 200$ and the weights were fixed as $\pi = (0.3, 0.4, 0.3)$. The hyperparameters of the prior distributions were kept as in the previous scenario and are summarized in Table 13.

Table 13 – Scenario 2: Summary of the hyperparameters of the prior distributions.

Prior	Hyperparameters
Spike and Slab	$\sigma_i^2 = 100$
g-prior	$g = n_i$, $\sigma_i^2 = 1$, $\lambda = 1/p$

5.2.2.1 Spike and Slab Prior

After running the 30 replications of data for each method, we assess the convergence by the the Geweke's convergence diagnostic. Figure 19 shows the boxplot of the Geweke's diagnostic obtained from the replications, where it is possible to see that the most of chains has converged.

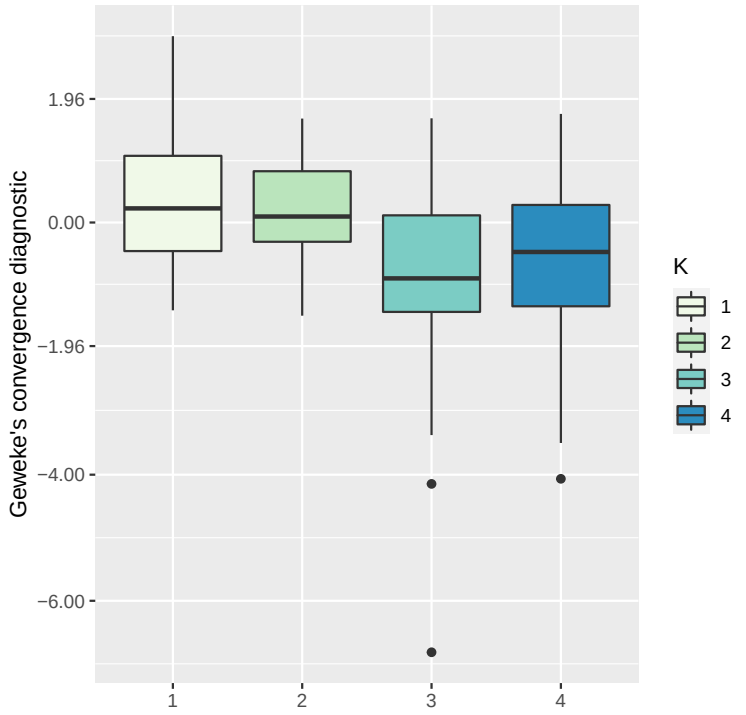

Figure 19 – Scenario 2: Geweke's convergence diagnostic of the replications of the models with the spike and slab prior.

The performance of the model selection criteria to estimate the number of components are shown in Table 14. The DIC criterion has a percentage of correct estimation of 24%, tending to select the model with $K = 4$ component in 71% of replications. The EBIC criterion, on the other hand, had a percentage of correct estimation of 86%, selecting the model with $K = 2$ in 14% of replications. Taking into to account the results of both criteria, we selected the model with $K = 3$ components.

Table 14 – Scenario 2: Average of criteria to estimate K and their correct estimation percentage obtained with spike and slab prior.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	7491.26	1200.08	-7261.53	-7428.55	24
EBIC	7518.67	1298.34	1189.94	1231.01	86

After the variable selection process, the posterior inclusion probability of the covariates in the replications revealed that all important covariates of each component were selected in every replication. However in the component 1, the covariate associated to the regression coefficient β_4 was included in the model in 2 replications. The same happened in the component 2, the covariate associated to the regression coefficient β_4 was included in 43% of the replications. The boxplots of the TPR and FPR are shown in Figure 20.

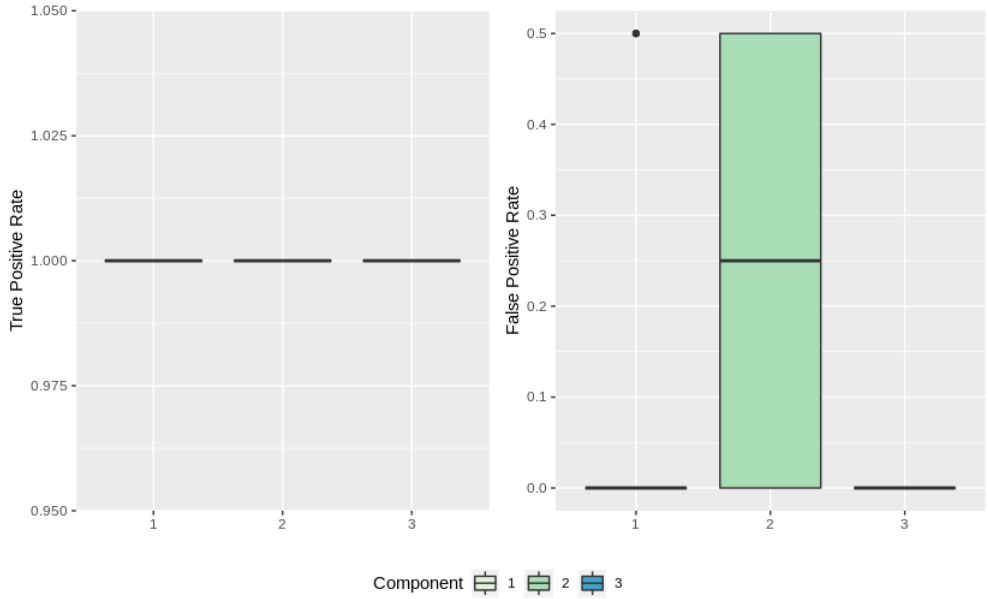

Figure 20 – Scenario 2: False and True Positive Rate obtained from the selection with the spike and slab prior.

Table 15 shows the estimates of the regression coefficient associated to the selected covariates and the weights of the mixture. Note that, in general, the method with the spike and slab prior could select the correct covariates in all components. Some regression coefficients were overestimated, however, the estimates are still close to the true value.

Table 15 – Scenario 2: Estimates, true value and credibility interval for the parameters in the model with spike and slab prior.

	Component 1	Component 2	Component 3
β_0	0.80 (1) (0.33,1.28)	-1.00(-1) (-1.14,-0.87)	-0.98 (-1) (-1.20,-0.75)
β_1	-4.12 (-4) (-4.68,-3.55)	-3.04 (-3) (-3.27,-2.82)	4.01 (4) (3.66,4.37)
β_2	-	-	-
β_3	2.10 (2) (1.74,2.46)	0.99 (1) (0.85,1.13)	-2.00 (-2) (-2.23,-1.78)
β_4	-	-	-
π	0.28 (0.3) (0.19, 0.36)	0.44 (0.4) (0.34, 0.53)	0.29 (0.3) (0.22, 0.35)

The boxplot of the obtained biases in the models in each replication are shown in Figure 21. This distribution is more spread out than of other scenarios presented, but it is still distributed close to zero.

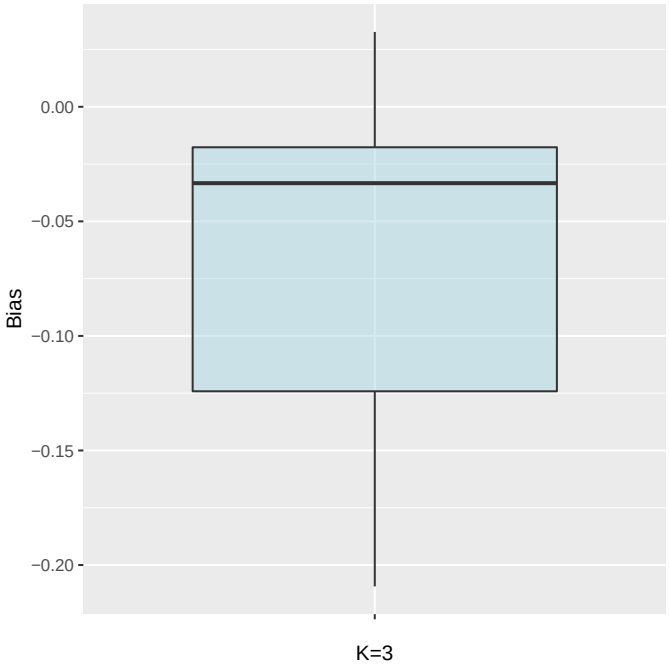

Figure 21 – Scenario 2: Bias of the estimates of the model with $K = 3$ components in each replication with the spike and slab prior.

Regarding the classification of the observations, the method with the spike and slab prior had a median of TCO of 81%, with credibility interval given by $(7, 87)\%$. This result is very similar to the one in the Scenario 5.1.2.

5.2.2.2 *g*-Prior

The Geweke's convergence diagnostic of the replications with g-prior is shown in Figure 22. The replications that did show convergence were excluded from the results.

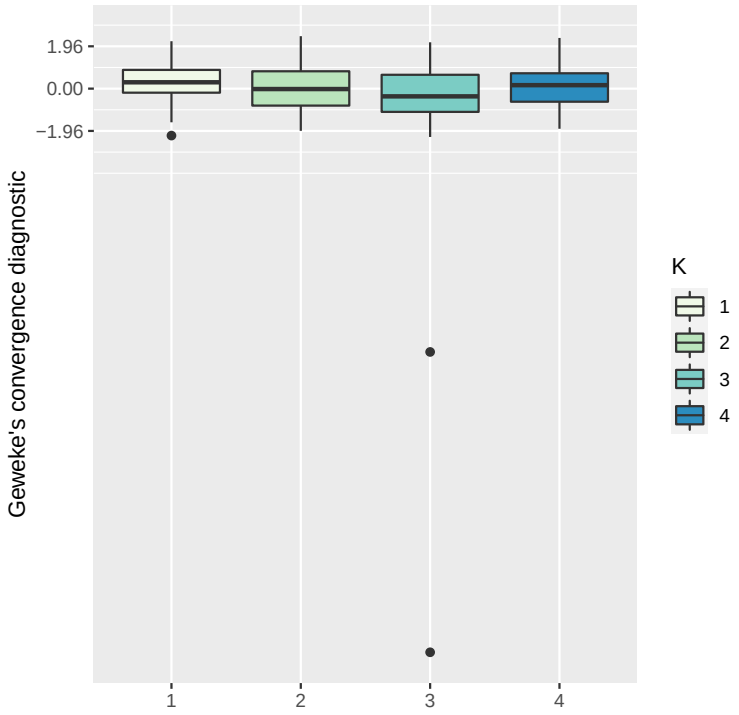

Figure 22 – Scenario 2: Geweke's convergence diagnostic of the replications with the g-prior.

In the estimation of the number of components, the DIC criterion selected the correct model in 53% of the replications, while EBIC selected the correct model in 65% of the replications. A summary of the results of each criterion is shown in Table 16. Based on these results, both criteria was minimized in the correct model and thus we choose the model with $K = 3$ components.

Table 16 – Scenario 2: Average of criteria to estimate K and their correct estimation percentage obtained with g-prior.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	7713.38	1335.85	-13831.45	-13296.58	53
EBIC	7740.88	1389.68	1197.73	1222.61	65

Regarding the variable selection, when checking the posterior inclusion probability of the covariates obtained in each replication we see that, in the component 1, the covariate associated to the regression coefficient β_1 was excluded from the model in 28% of the replications. In the same way, the covariate associated to the regression coefficient β_3 was excluded from the model in 19% of the replications. The covariate associated to the regression coefficient β_2 was included in the model in 33.33% of the replications. Similarly, the covariate associated to the regression coefficient β_4 in the component 1 was included in the model in 66.67% of the replications. In the component 2, the covariate x_4 was included in the model in 90% of the replications.

The covariates x_2 and x_4 of component 3 were included in 9% and 23% of the replications, respectively. The boxplots of TPR and the FPR are presented in Figure 23.

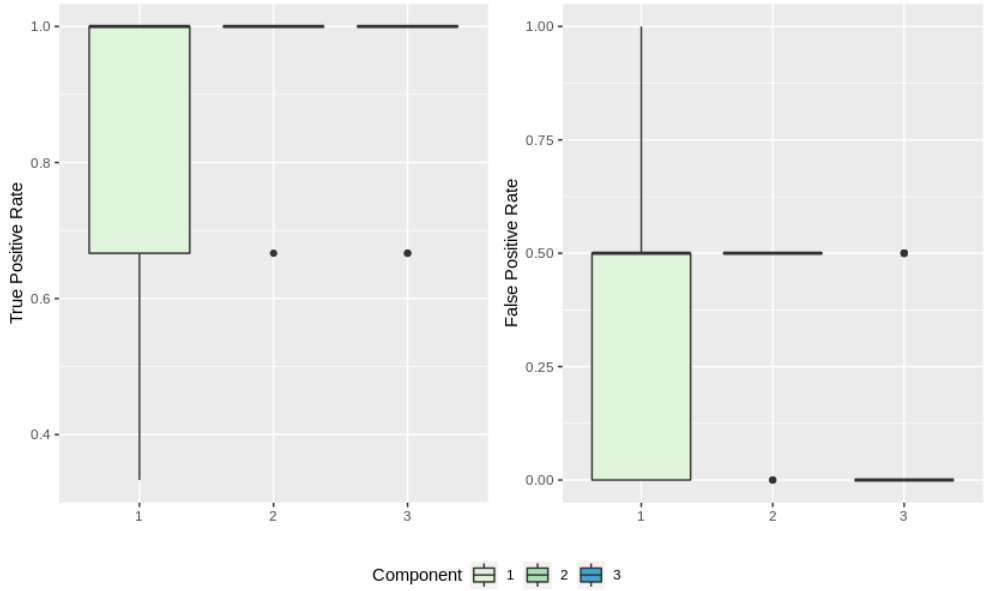

Figure 23 – Scenario 2: False and True Positive Rate obtained from the selection with the g-prior.

Table 17 shows the estimates of the regression coefficients associated to the relevant covariates and the estimates of the weights of the mixture. As discussed previously, the variable selection with the g-prior in this scenario selected some covariates that should not have been selected. As a consequence, the covariate x_4 was mistakenly included in the model of component 1 and 2. Table 17 presents the estimates of the regression coefficient associated to the selected covariates.

Table 17 – Scenario 2: Estimates, true value and credibility interval for the parameters of the model with g-prior.

	Component 1	Component 2	Component 3
β_0	0.57 (1) (0.14, 1.01)	-0.96 (-1) (-1.09, -0.82)	-0.82 (-1) (-1.09, -0.55)
β_1	-1.76 (-4) (-2.14, -1.38)	-3.14 (-3) (-3.34, -2.93)	3.88 (4) (3.54, 4.23)
β_2	-	-	-
β_3	1.01 (2) (0.66, 1.36)	1.15 (1) (1.01, 1.29)	-1.92 (-2) (-2.14, -1.71)
β_4	-0.14 (0) (-0.76, 0.08)	-0.84 (0) (-1.12 -0.56)	-
π	0.15 (0.3) (-0.56, 0.28)	0.55 (0.4) (0.47, 0.63)	0.29 (0.3) (0.22, 0.35)

In general, the estimates of the regression coefficients are close to the true value, except the estimate of β_1 of component 1. As a consequence, the bias is distributed close to zero as shown in Figure 24.

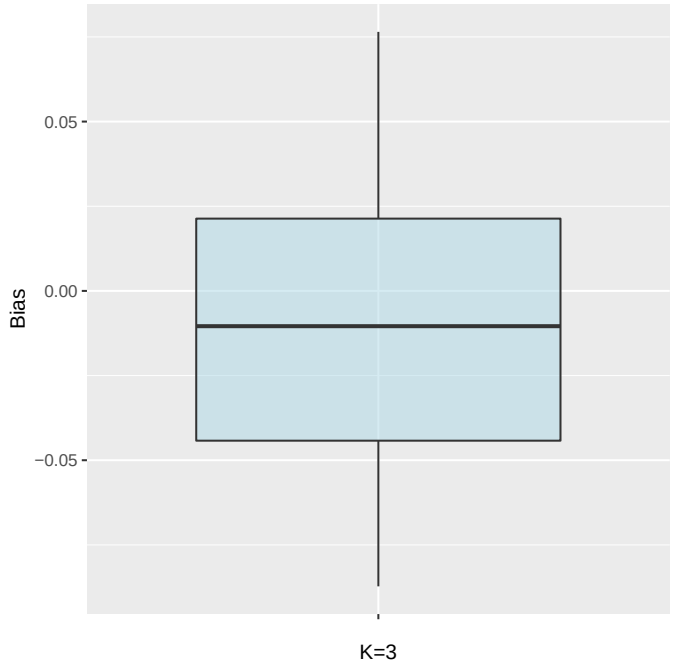

Figure 24 – Scenario 2: Bias of the estimates of the model with $K = 3$ components in each replication with the g-prior.

Finally, to assess the classification rate we analyze the median of TCO obtained from the replications. The median obtained was 64%, and its HPD interval is given by $(4, 88)\%$. Note that this performance is similar to the one obtained in the Scenario 5.1.2, with no variable selection.

For an overview of the variable selection in this scenario, Table 18 presents the average of TPR and FPR in the replications of each method.

Table 18 – Scenario 2: Average of TPR and FPR.

Prior	Component 1		Component 2		Component 3	
	TPR	FPR	TPR	FPR	TPR	FPR
Spike and Slab	1.0	0.06	1.0	0.25	1.0	0.00
g-prior	0.81	0.36	0.98	0.45	0.97	0.12

In general, the performance of the variable selection methods in a scenario where the components are very similar was good. The methods could select correctly most of the important covariates. However, comparing the two methods, we conclude that the variable selection with the spike and slab prior outperformed the one with the g-prior.

5.2.3 Scenario 3

In this scenario, the goal is to explore the performance of the variable selection method when there is a large number of inactive covariates and few active covariates. The data of the 30 replications was generated from a mixture of Binomial distributions with $N = 50$ Bernoulli trials, $K = 3$ components and $p = 100$ covariates where only 2 of them are non-zero, simulated from a standard normal distribution. The vectors of regression coefficients are given by

$$\begin{aligned} \boldsymbol{\beta}_1^\intercal &= (1,-1,0,1,0,0,\ldots,0), \\ \boldsymbol{\beta}_2^\intercal &= (-1,0,1,0,1,0,\ldots,0) \quad \text{and} \\ \boldsymbol{\beta}_3^\intercal &= (-0.5,0,-0.5,0,-0.5,0,\ldots,0). \end{aligned} \tag{5.6}$$

The sample size considered was $n = 300$ and the weights were fixed as $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$. Note that, with the sample size considered, we may fall in the case where $p > n$. Table 19 shows the settings of the hyperparameters for the simulation of this scenario.

Table 19 – Scenario 3: Summary of the hyperparameters of the prior distributions.

Prior	Hyperparameters
Spike and Slab	$\sigma_i^2 = 10$
g-prior	$g = n_i$, $\sigma_i^2 = 1$, $\lambda = 1/p$

In the next sections we present the results of the variable selection with each one of prior distributions considered in this work.

5.2.3.1 Spike and Slab prior

Figure 25 shows the Geweke's convergence diagnostic of the replications with the spike and slab prior. It it possible to see that most of replications showed convergence according to this criterion.

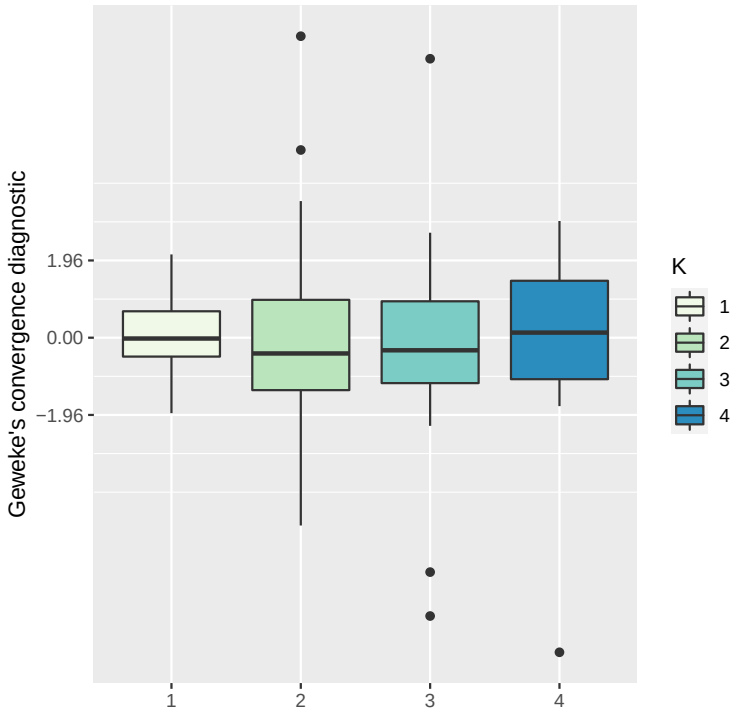

Figure 25 – Scenario 3: Geweke's convergence diagnostic of the replications with the spike and slab prior.

In the estimation of the number of components, the DIC criterion selected the correct model in 36% of the replications, selecting the model with $K = 4$ components in 43% of the replications. The same behavior seen in the other scenarios was observed. The model with $K = 4$ components had a estimated of π_4 very close to zero with few observations allocated to it. Moreover, no covariate had frequency greater or equal to 50% in the model of component 4, which means that no covariate was selected. The EBIC, on the other side, selected the model with $K = 2$ components 100% of the replications. This similar behavior was observed in Scenario 5.1.3, where $n < p$. When checking the model with $K = 2$ components we observed that some covariates was wrongly selected in the two components, and in both of them, the estimates of the regression coefficients associated to the selected covariates were close to zero. A summary of these results are given in Table 20. Based on these results we considered the model with $K = 3$ components.

Table 20 – Scenario 3: Average of criteria to estimate K and their correct estimation percentage obtained with spike and slab prior.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	4546.856	265.155	-2969.638	-3583.928	36
EBIC	5089.818	3193.564	3841.516	4414.648	0

Through the posterior inclusion probability of the covariates in the replications we see that, in the component 2, both covariates associated to the regression coefficients β_2 and β_4 was excluded from the model in 41.66% of the replications. In same way, the covariates associated

to the regression coefficients β_2 and β_4 of the model in component 3 was excluded from the model in 41.66% and 54.16% of the replications, respectively. In the model of component 1, the covariates associated to the regression coefficients β_2 and β_4 was included in the model 12.6% and 16.66% of the replications, respectively. In the model of component 2, the covariates associated to the regression coefficients β_1 and β_3 was included in the model 30% and 16.66% of the replications, respectively. Finally, in the model of component 3, the covariates associated to the regression coefficients β_1 and β_3 was included in the model 8.3% and 12.5% of the replications, respectively. The boxplots of TPR and FPR are shown in Figure 26.

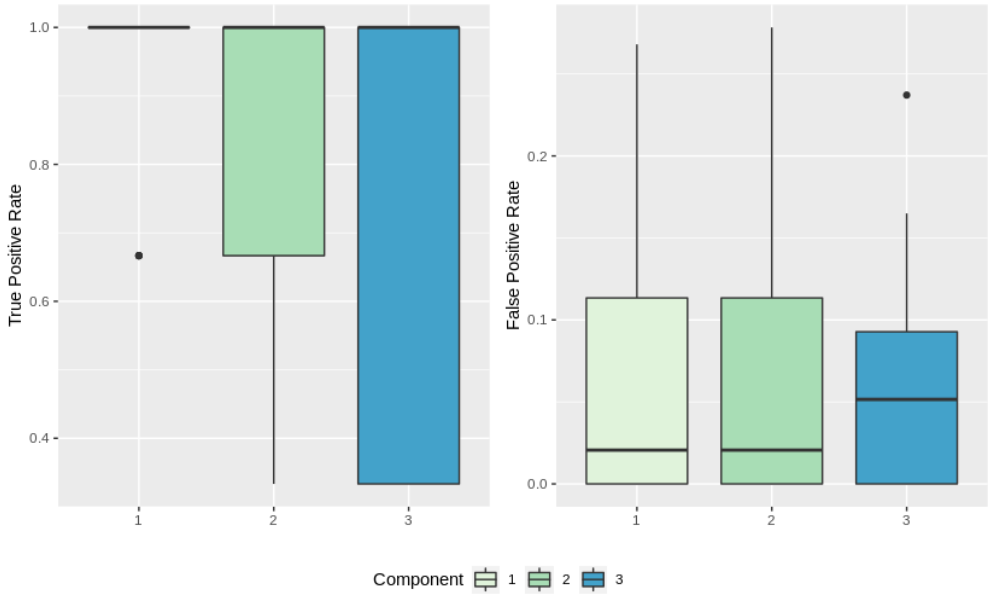

Figure 26 – Scenario 3: False and True Positive Rate obtained from the selection with the spike and slab prior.

In Table 21, we present the estimates of the regression coefficients of the selected covariates. The credibility interval of the regression coefficients was obtained through the 2.5% and 97.5% quantiles of the values obtained in the replications. In Table 21, we observe that all covariates were selected correctly in each component. However, the variation of TPR previously discussed affected the estimates of some regression coefficients.

Table 21 – Scenario 3: Estimates, true value and credibility interval for the parameters of the model with spike and slab prior.

	Component 1	Component 2	Component 3
β_0	0.6 (1) (-0.20, 1.05)	-0.7 (-1) (-1.10, 0.05)	-0.43 (-0.5) (-0.66, -0.08)
β_1	-0.83 (-1) (-1.08, 0.00)	-	-
β_2	-	0.55 (1) (0.00, 1.09)	-0.18 (-0.5) (-0.55, 0.52)
β_3	0.80 (1) (0.00, 1.63)	-	-
β_4	-	0.63 (1) (0.00, 1.09)	-0.21 (-0.5) (-0.55, 0.10)
π	0.33 (0.3) (0.25, 0.40)	0.32 (0.4) (0.24, 0.41)	0.35 (0.3) (0.27, 0.43)

For better assessing the estimates obtained after the selection, we analyse the bias them. In Figure 27, we see that the bias is distributed close to zero.

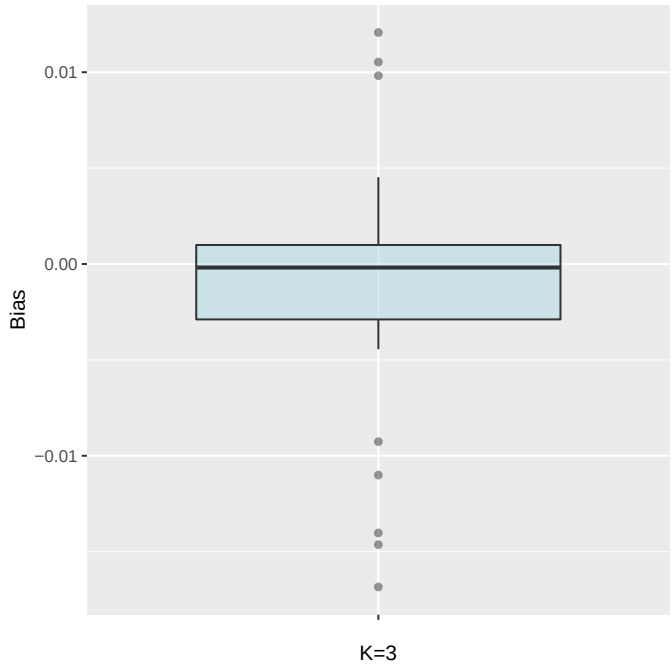

Figure 27 – Scenario 3: Bias of the estimates of the model with $K = 3$ components in each replication with the spike and slab prior.

To assess the classification rate we analyze the median of TCO obtained from the replications. The median obtained was 40%, and its HPD interval is given by $(8, 80)\%$.

5.2.3.2 g-Prior

For the convergence assessment, we analyse the Geweke's convergence diagnostic of the replications, shown in Figure 28. The replications that did not show convergence were not considered in the final results.

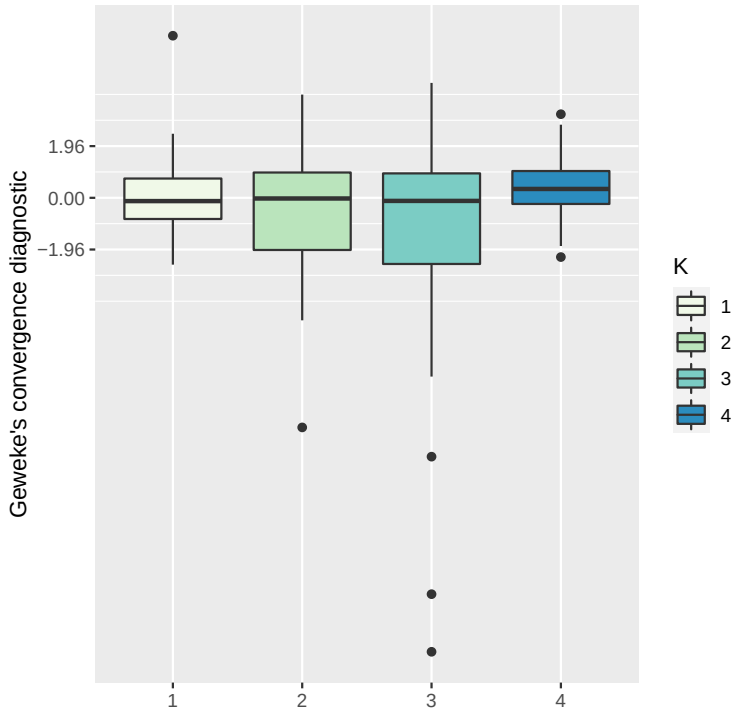

Figure 28 – Scenario 3: Geweke's convergence diagnostic of the replications with the g-prior.

The estimation of the number of components in the model with g-prior had the same behavior of the one with the spike slab prior. The DIC criterion selected the correct model in 40% of the replications and the model with $K = 4$ components in the remaining replications. However, the model with $K = 4$ components had a estimated of π_4 very close to zero with few observations allocated to it. Moreover, no covariate was selected in the model of component 4. The EBIC selected the model with $K = 2$ components 100% of the replications. Nonetheless, when checking the results of the selection in the model with $K = 2$ components we observed that some covariates was mistakenly selected in the two components and, in both of them, the estimates of the regression coefficients associated to the selected covariates were close to zero. Table 22 shows the mean of the criteria in the replications and the percentage of success in selecting K. Based on all these results we considered the model with $K = 3$ components.

Through the posterior inclusion probability of the covariates in the replications we see that the covariate associated to the regression coefficient β_1 was mistakenly excluded from the model of one of the replications. The covariates associated to the regression coefficients β_2 and β_4 was included in the model 11.76% of the replications. Some other null covariates were also wrongly selected in one of replications. In the model of component 2, the covariates associated

Table 22 – Scenario 3: Average of criteria to estimate K and their correct estimation percentage obtained with g-prior.

Criterion	$K=1$	$K=2$	$K=3$	$K=4$	%
DIC	4507.32	1206.74	-3820.58	-8372.72	40
EBIC	5032.27	3148.64	3827.80	4434.36	0

to the regression coefficients β_1 and β_3 was included in the model 11.76% and 17.64% of the replications, respectively. Again, some other null covariates were wrongly selected in some replications. However, this number was small. Finally, in the model of component 3, some null covariates were wrongly selected as well. Figure 29 presents the boxplots of TPR and FPR obtained in the replications of the model.

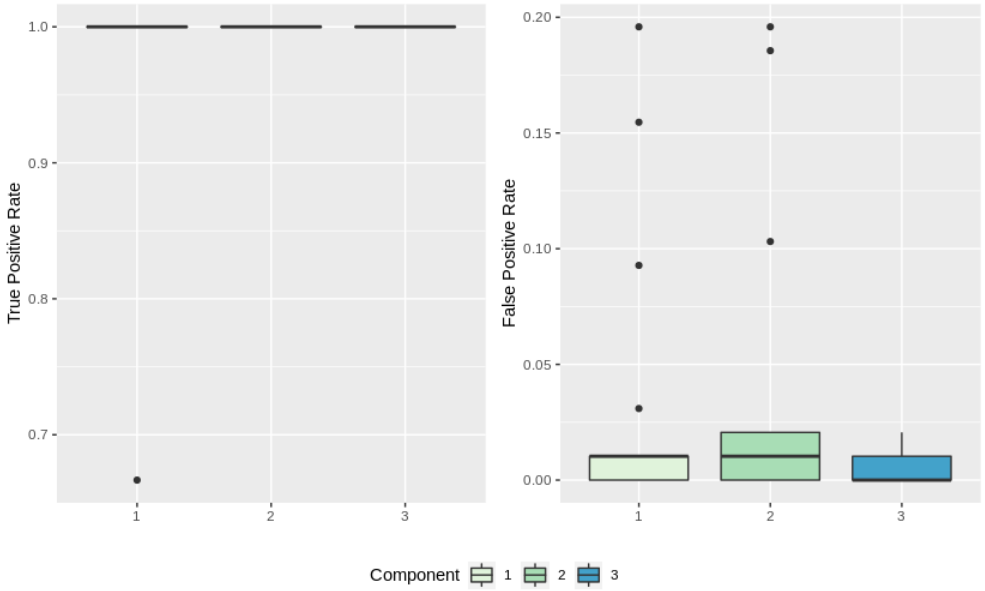

Figure 29 – Scenario 3: False and True Positive Rate obtained from the selection with the g-prior.

In Table 23 we present the estimates of the regression coefficients of the selected covariates. Despite of the variation of FPR, all covariates were selected correctly in each component with good estimates. At the bottom of this table, the estimates of the weights are presented.

Table 23 – Scenario 3: Estimates, true value and credibility interval for the parameters of the model with g-prior.

	Component 1	Component 2	Component 3
β_0	0.86 (1) (0.11, 1.06)	-0.9 (-1) (-1.09, 0.04)	-0.46 (-0.5) (-0.60, -0.04)
β_1	-0.90 (-1) (-1.05, 0)	-	-
β_2	-	0.85 (1) (0, 1.07)	-0.43 (-0.5) (-0.59, 0.04)
β_3	0.94 (1) (0.63, 1.08)	-	-
β_4	-	0.84 (1) (0.19, 1.12)	-0.43 (-0.5) (-0.57, 0)
π	0.32 (0.3) (0.26, 0.39)	0.34 (0.4) (0.28, 0.41)	0.33 (0.3) (0.27, 0.40)

Analysing the bias of the estimates, shown in Figure 30, we see that the bias is distributed very close to zero, which indicates a good fit.

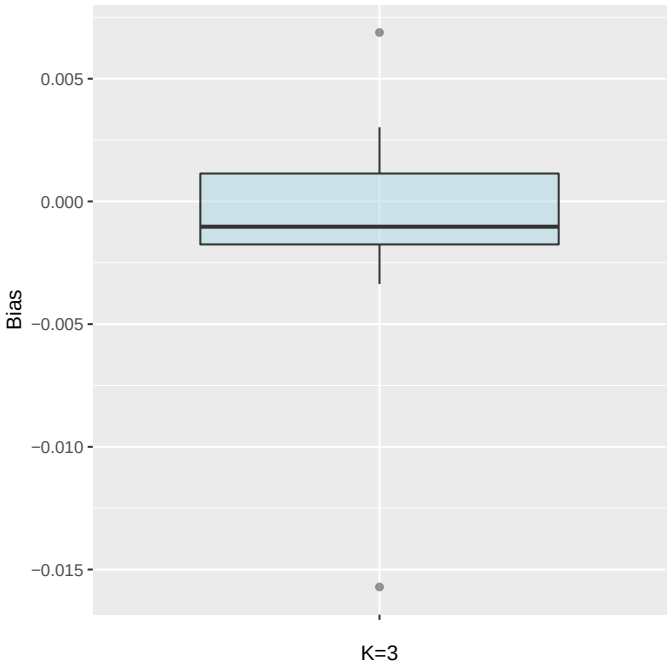

Figure 30 – Scenario 3: Bias of the estimates of the model with $K = 3$ components in each replication with the g-prior.

Regarding the final classification of the observations, the median of the TCO obtained was 76%, and its HPD interval is given by $(9, 82)\%$.

For an overview of the variable selection in this scenario, Table 24 presents the average of TPR and FPR in the replications of each method.

Table 24 – Scenario 3: Average of TPR and FPR.

Prior	Component 1		Component 2		Component 3	
	TPR	FPR	TPR	FPR	TPR	FPR
Spike and Slab	0.90	0.09	0.72	0.07	0.68	0.07
g-prior	0.98	0.03	1.0	0.04	1.0	0.01

Through Table 24 we observe that, in general, the performance of the variable selection in both model with spike and slab prior and g-prior was good. But it is evident that in the model with the g-prior the variable selection was better.

5.3 Estimation of the Model with Binary Response

As commented in Section 2.2.1, mixture model of logistic regressions is not identifiable when the number of Bernoulli trials is equal to one, that is, when the response variable is Bernoulli. The identifiability condition for this model was then proposed by Teicher (1963), in which it says that to ensure identifiability, the condition $N \geq 2K-1$ needs to be satisfied, where K is the number of components and N the number of Bernoulli trials. A further work by Follmann and Lambert (1991) provides a sufficient condition to identifiability in a mixture of logistic regressions with binary response when only the intercept is random, that is, when only the intercept varies across the components. According to Follmann and Lambert (1991), this model is identifiable if $K = \sqrt{N_{11}+2}-1$, where N_{11} is the maximum number of observations in the sample that differ, at most, in the value of only one covariate.

These simulations analyse the behavior of estimation and selection methods when the response variable is binary and check the identifiability condition presented by Follmann and Lambert (1991).

The data was generated as in Scenario 5.2.1, from a mixture of $K = 3$ logistic regression models where the response variable follows a Bernoulli distribution ($N = 1$) and $p = 5$ covariates. First, their respective regression coefficients are given by

$$\begin{aligned} \boldsymbol{\beta}_1^\intercal &= (1,-1,0,1,0), \\ \boldsymbol{\beta}_2^\intercal &= (-1,0,1,0,1) \quad \text{and} \\ \boldsymbol{\beta}_3^\intercal &= (-0.5,0,-0.5,0,-0.5). \end{aligned} \tag{5.7}$$

The sample size was fixed as $n = 200$ and the weights as $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$. In the variable selection, we consider only the spike and slab prior distribution to the regression coefficients,

fixing $\sigma_i^2 = 10$ for $i = 1, \cdots, K$, which showed better performance when the number of covariates is small.

According to the Geweke's convergence diagnostic, most replications showed convergence. Regarding the estimation of the number of components, the EBIC criterion selected the model with $K = 1$ component in all replications. The DIC criterion selected the correct model in 71% of the replications, selecting the model with $K = 4$ in 14% of the replications and the models with $K = 1, 2$ in 7% of the replications. The mean of both criteria are shown in Table 25.

Table 25 – Binary model: Average of criteria to estimate K and their correct estimation percentage.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	274.83	271.48	233.26	254.38	71
EBIC	306.02	461.56	602.79	728.48	0

When analysing the results of the model with $K = 1$ component we observed that no covariate was selected in the model. Consequently, the estimates of all regression coefficients are close to zero. In the model with $K = 3$ components, the results shows that only in the component 1 the covariates were correctly selected. In the components 2 and 3, all covariates had an average presence of 45% and thus no covariate was selected. Table 26 shows the estimates for the parameters of the selected covariates of the model with $K = 3$ components.

Table 26 – Binary model: Estimates, true value and credibility interval for the parameters of the selected variables.

	Component 1	Component 2	Component 3
β_0	2.39 (1) (1.11, 3.41)	-1.8 (-1) (-3.09, -0.36)	-1.07 (-0.5) (-3.01, 0.29)
β_1	-0.82 (-1) (-3.07, 0)	-	-
β_2	-	-	-
β_3	0.91 (1) (-0.22, 1.80)	-	-
β_4	-	-	-
π	0.38 (0.33) (0.09, 0.66)	0.33 (0.33) (0.04, 0.61)	0.29 (0.33) (0.03, 0.60)

Figure 31 shows the boxplots of TPR and FPR obtained over the replications.

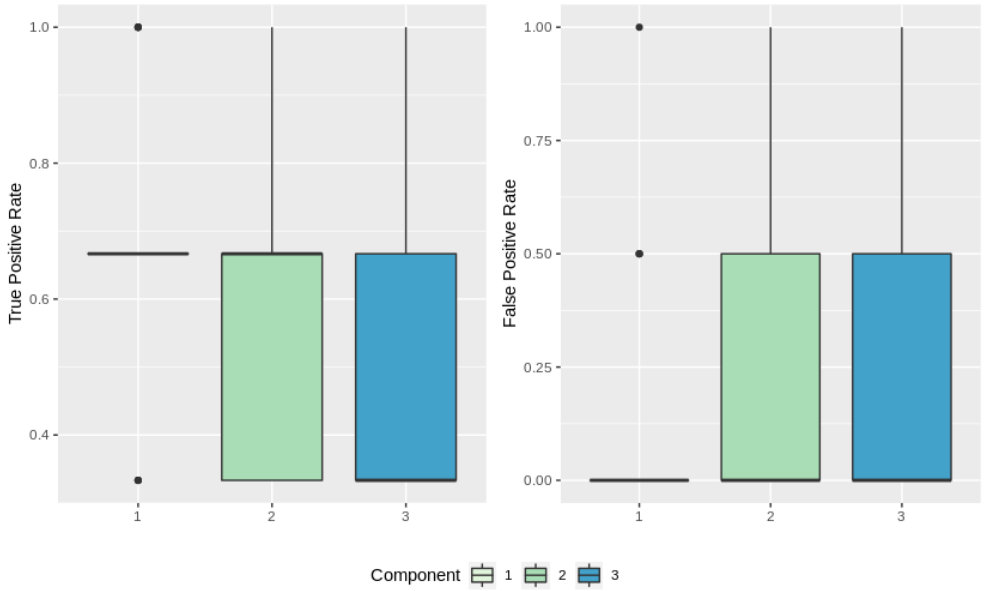

Figure 31 – Binary model: False and True Positive Rate obtained from the selection with the spike and slab prior.

The median of the TCO in model with $K = 3$ component was 36%, with credibility interval of $(32, 46)\%$.

According to these results we conclude that the binary model presents the identifiability problem. The EBIC criterion tends to select the model with only one component and does not select any covariate. Moreover, even when the DIC selected the correct model, only the model of component 1 was meaningful. However, the variables selected for the largest component are in agreement with the relevant variables in the first component. Here again, we observe the DIC tendency to choose models with a greater number of components and probably allocate atypical observations to smaller components.

In order of assessing the selection model performance under identifiability condition for mixture of Bernoulli models, we run another simulation considering that only the intercept varies across the components. We generate data from a mixture of $K = 3$ logistic regressions with Bernoulli response and $p = 5$ covariates generated from a standard normal distribution, two of them being active. The regression coefficients are given by,

$$\begin{aligned} \boldsymbol{\beta}_1^\intercal &= (2, -1, 0, 1, 0), \\ \boldsymbol{\beta}_2^\intercal &= (-3, -1, 0, 1, 0) \quad \text{and} \\ \boldsymbol{\beta}_3^\intercal &= (-0.5, -1, 0, 1, 0). \end{aligned} \tag{5.8}$$

Since we set $K_{max} = 4$, we repeated an observation 25 times in the data set, allowing only the value of one covariate to differ among them, to have $N_{11} = 25$.

Figure 32 shows the boxplot of the Geweke's convergence diagnostic, in which we see that, the most of replications showed convergence. To perform the variable selection, we apply the method with only the spike and slab prior, which showed better results when p is small.

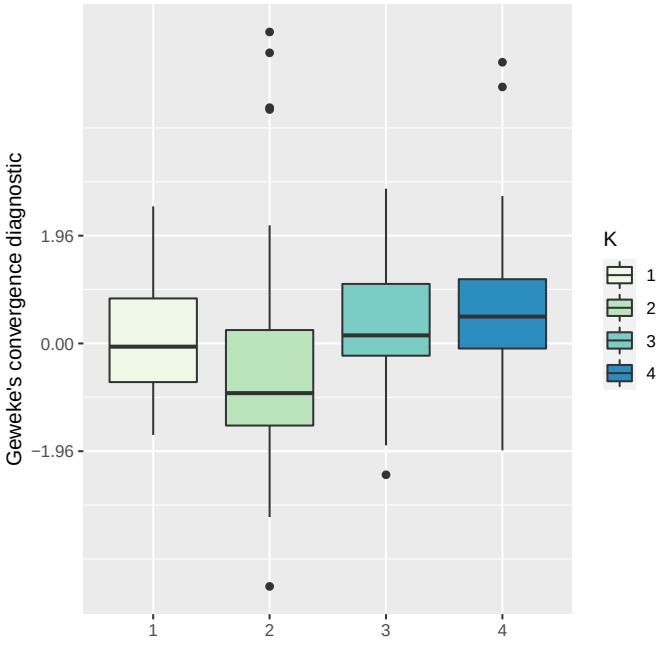

Figure 32 – Binary model: Geweke convergence diagnostic of the model under identifiability condition.

In the estimation of the number of components of the mixture, the DIC criterion selected the model with $K = 3$ components mostly. Nevertheless, the EBIC criterion selected the model with $K = 1$ component in all replications, the same behavior of the model under non-identifiability. When checking the model with $K = 1$ component, we observed that the covariates were correctly selected, since the true subset of active variables and their regression coefficients are the same among the components. However, the estimates of the regression coefficients were close to zero. The mean and percentage of correct selection of criteria are shown in Table 27.

Table 27 – Binary model: Average of criteria to estimate K and their correct estimation percentage in the model under the identifiability condition.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$	%
DIC	249.32	267.29	209.57	220.73	50
EBIC	277.67	455.78	587.84	698.31	0

Below we present the results of the model with $K = 3$ components. Table 28 shows the estimates of the regression coefficients of the selected covariates, considering their posterior inclusion probability in the replications. Through these results we see that the covariates were correctly selected and the difference among the intercept's values is captured.

Table 28 – Binary model: Estimates, true value and credibility interval for the parameters of the model under the identifiability condition.

	Component 1	Component 2	Component 3
β_0	1.87 (2) (-1.36, 5.94)	-2.40 (-3) (-6.76, 1.48)	-1.07 (-0.5) (-5.51, 4.13)
β_1	-1.09 (-1) (-5.76, 2.14)	-0.63 (-1) (-4.80, 2.44)	-1.21 (-1) (-6.14, 2.68)
β_2	-	-	-
β_3	1.11 (1) (-2.39, 5.35)	0.67 (1) (-2.85, 4.62)	1.16 (1) (-2.79, 6.08)
β_4	-	-	-
π	0.35 (0.33) (0.05, 0.63)	0.34 (0.33) (0.05, 0.62)	0.30 (0.33) (0.02, 0.64)

Regarding the variable selection performance, the mean of the TPR in each component were 0.78, 0.81 and 0.99, respectively. The mean of the FPR obtained in the model of each component were 0.13, 0.15 and 0.15, respectively. Figure 33 shows the boxplots of the TPR and FPR obtained in the replications.

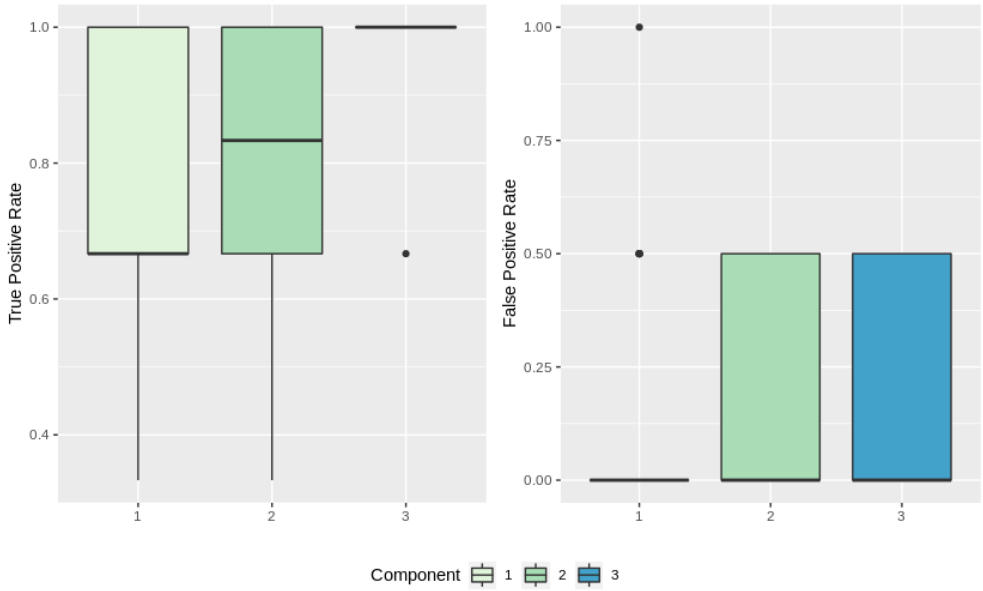

Figure 33 – Binary model: False and True Positive Rate obtained from the selection with the spike and slab prior under the identifiability condition.

Despite the good results presented above, the classification of the observations had not a good performance in this scenario. The median of the TCO was 38.2% with credibility interval of $(33, 46)\%$.

Under the identifiability condition proposed by Follmann and Lambert (1991), the methodology could select the correct covariates and provided good estimates to the regression coefficients and weights. However, the EBIC still shows the tendency of selecting the model with the smallest number of components and, consequently, does not identify the difference among the intercept values. The DIC criterion, in the other hand, were able to identify the correct number of components and estimate different intercepts. Although these results, the percentage of correct classification of the observations was low, probably because, except for the intercept's values, the components and their memberships are overlapped under the identifiability condition.

CHAPTER

6

ANALYSING AN EDUCATION DATA SET

In this chapter, we apply the methodologies discussed in this work to select variables in a real data set. This data was first analysed by Cortez and Silva (2008), where a classification tree was applied to classify student's grades and selected relevant covariates. The data set contains the final grades of $n = 395$ students of secondary school with ages between 15 and 22 years from public schools in the Alentejo region of Portugal during the period of 2005-2006. These grades are provided with respect to the Math and Portuguese exams. In our application we consider only the grades of the Math exam. The data attributes include age, gender, mother's and father's job, mother's and father's education, weekly study time and some other demographic, social and school related features. Table 29 presents a description of each covariate considered.

The response variable Y is the student's final grade, which takes integer values from 0 to 20. A student is approved if the final grade is greater or equal 10. The aim is to select the covariates that affect the final grade. The categorical covariates were dummy-coded so a total of $p = 68$ covariates were considered in the selection. Figure 34 shows the bar plot of the final grades.We observe a inflated number of zeros and also a concentration of grades at 18.

Table 29 – Description of the covariates.

Covariates	Description
sex	student's sex (binary: 1 - female or 0 - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 1 - urban or 0 - rural)
famsize	family size (binary: 1 - less or equal to 3 or 0 - greater than 3)
Pstatus	parent's cohabitation status (binary: 1 - living together or 0 - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education, 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother's job (nominal: "teacher", "health" care related, civil "services", "at_home" or "other")
Fjob	father's job (nominal: "teacher", "health" care related, civil "services", "at_home" or "other")
reason	reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
guardian	student's guardian (nominal: "mother", "father" or "other")
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: 1 - yes or 0 - no)
famsup	family educational support (binary: 1 - yes or 0 - no)
paid	extra paid classes of Math (binary: 1 - yes or 0 - no)
activities	extra-curricular activities (binary: 1-yes or 0-no)
nursery	attended nursery school (binary: 1-yes or 0 - no)
higher	wants to take higher education (binary: 1 - yes or 0 - no)
internet	Internet access at home (binary: 1 - yes or 0 - no)
relationship	with a romantic relationship (binary: 1 - yes or 0 - no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)

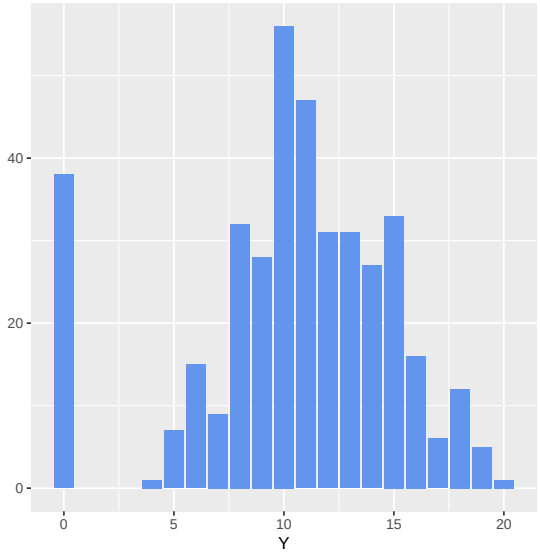

Figure 34 – Bar plot of the student's final grades (response variable).

Considering that each Y_j assumes values in $\Omega = \{0,1,2,\ldots,20\}$ for $j = 1,\ldots,n$, a mixture of Binomial distributions with $N = 20$ Bernoulli trials is considered to model each Y_j. As in the simulations with synthetic data, we run the model for $K = \{1,2,3,4\}$ components and apply the criteria DIC and EBIC to select the model that better fit the data. To perform variable selection we run the model with both spike and slab prior and g-prior. For the spike and slab prior, we set $\sigma_i^2 = 10$ for $i = 1,\ldots,K$. For the g-prior, we set $\sigma_i^2 = 1$ and $g_i = n_i$ for $i = 1,\ldots,K$. The indicator variables associated with the regression coefficients were initialized as zero in both

cases, which means no covariate in the model initially. To check the convergence, we used the Geweke's convergence diagnostic for the log-likelihood.

In Table 30 we see the values of the Geweke's diagnostic of each model with the spike and slab prior, which indicates convergence. In the estimation of the number of components, the EBIC criterion was minimized in the model with $K = 2$ components. The DIC criterion, in the other hand, was minimized in the model with $K = 4$ components. However, when checking the model with $K = 4$ components, it was observed that the estimates for π_3 and π_4 were very close to zero with a few observations allocated to them. Moreover, the estimates of the regression coefficients associated with the selected covariates in the components 3 and 4 were close to zero, and their credibility interval were large with the zero value in. Considering these results and also taking into account the performance of the DIC criterion in the scenarios with simulated data, we assumed that the final grade of the students are better modelled by a mixture of two logistic distributions. Table 31 presents a summary of each criteria to select the number of components.

Table 30 – Geweke's diagnostic of the models.

Prior	$K = 1$	$K = 2$	$K = 3$	$K = 4$
spike and slab	1.75	-0.68	-0.92	1.49
g-prior	-0.32	1.60	1.85	0.93

Table 31 – Values of the criteria to estimate K with spike and slab prior.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$
DIC	2648.17	-364.66	69.50	-1504.85
EBIC	3156.34	2865.27	3344.75	3783.75

The results of the variable selection with spike and slab prior are summarized in Table 32, where only the selected covariates in each component are shown with their credibility interval. The estimates of the weights are presented at the bottom of the table.

Table 32 – Estimates and credibility interval for the parameters of the model with spike and slab prior.

	Component 1	Component 2
Intercept	-0.20 (-1.22, 0.82)	-0.68 (-6.76, 5.40)
Gender (F)	-	0.80 (-2.62, 4.23)
Famsize	-	-1.12 (-4.63, 2.40)
Schoolsup	0.43 (0.26, 0.60)	-
Paid	-	-1.05 (-4.53, 2.44)
Higher	-	-1.01 (-5.53, 3.51)
Relationship	-	3.63 (-1.32, 8.59)
Absences	-0.01 (-0.02, 0.00)	0.87 (-0.01, 1.75)
Medu_1	-	-0.93 (-5.05, 3.19)
Mjob_health	0.20 (-0.14, 0.53)	1.01 (-3.13, 5.16)
Mjob_services	0.18 (-0.06, 0.42)	-
Fjob_health	-	0.22 (-4.03, 4.47)
Fjob_teacher	0.31 (-0.09, 0.72)	-
Reason_other	-	1.69 (-2.99, 6.38)
Reason_reputation	-	1.49 (-2.59, 5.56)
Traveltime_2	-	-1.07 (-4.87, 2.72)
Traveltime_4	-	-0.25 (-4.39, 3.89)
Studytime_3	-	1.21 (-2.85, 5.27)
Famrel_2	-	-1.24 (-5.59, 3.10)
Famrel_4	-	1.35 (-2.48, 5.17)
Goout_2	0.10 (-0.11, 0.31)	-
Goout_5	-	-1.32 (-5.54, 2.89)
Dalc_2	-	-1.02 (-4.98, 2.95)
Dalc_5	-	-0.03 (-4.50, 4.44)
Walc_4	-0.14 (-0.40, 0.12)	-
Health_3	-	-0.93 (-4.45, 2.59)
Failures_1	-0.14 (-0.43, 0.15)	-2.23 (-7.31, 2.85)
Failures_2	-0.52 (-0.90, -0.13)	-0.90 (-5.13, 3.34)
Failures_3	-0.57 (-0.96, -0.18)	-0.88 (-5.02, 3.25)
π	0.85 (0.81, 0.89)	0.15 (0.11, 0.19)

In the classification of the observations, 338 observations were allocated to the component 1 and the remaining 57 were allocated to the component 2. Figure 35 shows the bar plot of response variable allocated in the component 1 and 2. In this figure is possible to see that the component 2 represents the students with grade equal or close to zero and the few students with the greatest grade. The component 1, on the other hand, correspond to the students that had a grade varying from 4 to 18.

According to the selected variables in the component 1, we can conclude that having extra education support have a positive impact in the grade of the student. More specifically, there is an increase of 49.18% in the odds of having greater grade in the exam when the student have extra educational support. The results also shows that when the student's father works as a teacher, there is an increase of 35% in the odds of having greater grade.

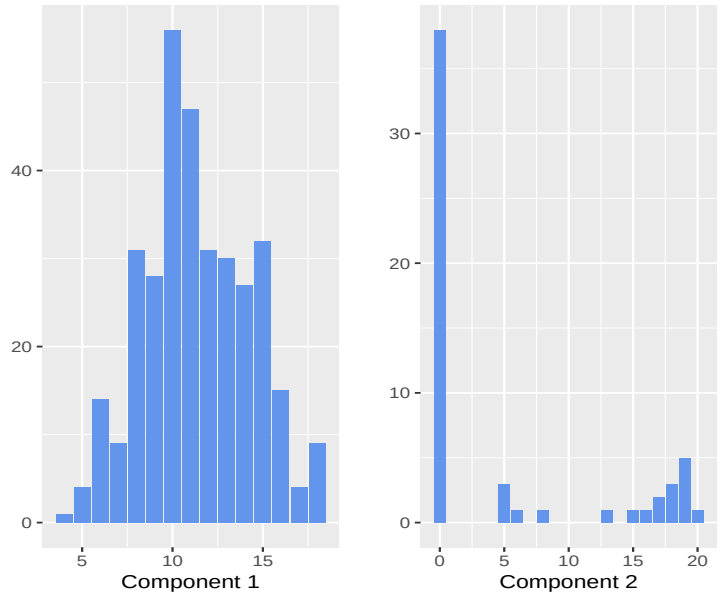

Figure 35 – Bar plot of the student's final grades classified into the components 1 and 2.

Regarding the social life of the student, results show that there is a small increase in the odds of having greater grade when the student do not go out very often. Moreover, when the student's consumption of alcohol in the weekend is high, there is a small decrease in the odds of having greater grade in the final exam. Another important finding is with respect to the past performance of the student. According to the selected covariates, past class failures have a negative impact in the final grade of the students. When the student have 3 past class failures, for example, there is a decrease of 45.11% in the odds of having greater grade in the final exam.

In the second component, many other covariates were selected, although zero belongs to the credibility interval for their regression coefficients. However, through the results we found that the second component was created to distinguish the extreme students, who present the lowest or greatest grades. The use of a mixture of logistic regression models allowed to distinguish extreme students and select relevant variables for them in one component and to model and identify relevant variables for average students in the other component. A single logistic regression model usually separates extreme observations very well, but is not good in distinguishing and selecting relevant variables for average observations.

The convergence of the method using g-prior was also observed according to Table 30 Both criteria, DIC and EBIC, selected the model with $K = 2$ components. The results of the criteria are shown in Table 33.

Table 33 – Values of the criteria to estimate K with g-prior.

Criterion	$K = 1$	$K = 2$	$K = 3$	$K = 4$
DIC	5445.85	4097.80	4583.03	4868.45
EBIC	3144.75	2893.16	3551.91	4115.08

The estimates of the regression coefficients associated with the selected covariates are shown in Table 34, where it is easy to see that the method with g-prior selected a smaller set of covariates compared to the selection in the model with the spike and slab prior. In this model, the component 1 correspond to those students that had the smallest grades and also the few greater grades, summing up to 52 observations allocated to it. Meanwhile the component 2, represents those students with grade between 4 and 19, summing up to 343 observations allocated to it.

Table 34 – Estimates and credibility interval for the parameters of the model with g-prior.

	Component 1	Component 2
Intercept	-0.45 (-1.03 0.14)	-0.07 (-0.45, 0.32)
Schoolsup	-	0.43 (0.26, 0.60)
Paid	-0.72 (-1.70, 0.27)	-
Absences	0.92 (0.05, 1.79)	-0.01 (-0.02, -0.01)
Mjob_health	-	0.29 (0.07, 0.51)
Mjob_services	-	0.22 (0.08, 0.36)
Fjob_teacher	-	0.37 (0.14, 0.60)
Studytime_3	-	0.19 (0.02, 0.37)
Goout_2	-	0.18 (0.04, 0.31)
Walc_4	-	-0.23 (-0.42, -0.05)
Failures_1	-	-0.22 (-0.41, -0.03)
Failures_2	-	-0.44 (-0.71, -0.17)
Failures_3	-	-0.45 (-0.74, -0.17)
π	0.12 (0.09, 0.16)	0.88 (0.84, 0.91)

When comparing the variable selection of both models we observe that they differ from each other only in the selection done in the component that represents the extreme students. For average students, the considered prior distributions selected, in general, the same set of relevant covariates.

For this data set, which shows a large number of covariates, the g-prior seems to be more efficient in selecting variables.

CHAPTER

7

CONCLUDING REMARKS

In this work, we developed a Bayesian method for estimating and selecting variables in a mixture of logistic regressions model. Through the data augmentation technique, using Pólya-Gamma random variables, it was possible to obtain conjugation for the distribution of the regression coefficients, simplifying the Bayesian estimation and variable selection of the model. Only a Gibbs sampling algorithm was necessary instead of other more complex approaches. To perform variable selection in this model, we investigated the performance of two prior distributions for the regression coefficients, adding a second set of latent variables to indicate the presence and absence of the predictor variables at each component of the mixture. Another benefit of the data augmentation is being able to analytically calculate the marginal likelihood and gain computational efficiency in the variable selection process.

In the estimation of the full model, without variable selection, the methodology presented a good performance in the estimation of the parameters. In the model with variable selection, both method could correctly select the variables, even in a high dimension scenario. When comparing the two variable selection methods, we see that the spike and slab prior showed a better performance under scenarios with a small number of covariates, meanwhile, the g-prior, although using the data in the prior and for estimation, showed a better performance when the number of covariates is large.

Considering the estimation of the number of components, different selection criteria select different models. The DIC, in relation to the EBIC, favors models with a greater number of components and separates atypical observations into small groups. If the number of components is not too large, the DIC behavior is good in the sense that the fitted model (component) in most observations is robust to outliers and more precise. The EBIC seems to be good for selecting the correct number of components when the components are more separated and evident.

Regarding the binary model, the variable selection with spike and slab prior could select the correct covariates in the model under the identifiability condition presented by Follmann and

Lambert (1991). Although it depends on the number of repeated observations, this identifiability condition may be a good option for models where only the intercept varies over the components and to verify and obtain first impressions of heterogeneity in data set.

For the Student's data set, the methodology identified two subgroups in the data, one to represent the average students and another to represent extreme students. For average students group, both variable selection methods selected the same relevant covariates and the g-prior seemed to be more efficient in comparison with the spike and slab prior in the group of extreme students, probably due to the large number of covariates present in the data. This application also illustrates one of the main advantages of the mixture of regression models, that is fitting good models for all observations, since it separates observations with different behaviors and selects specific predictors for each group. While, in this situation, a single logistic model would probably select only variables for separating the extreme students. The mixture of regression models allows to identify good relevant covariates also for distinguish average students.

As a future work, one interesting point would be to add a prior distribution to the prior inclusion probability p_{it}, for $i = 1, \ldots, K$ and $t = 1, \ldots, p$, and estimate it for each covariate in each component.

BIBLIOGRAPHY

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: **Selected papers of hirotugu akaike**. [S.l.]: Springer, 1998. p. 199–213. Citation on page 19.

BARAGATTI, M.; POMMERET, D. A study of variable selection using g -prior distribution with ridge parameter. **Computational Statistics & Data Analysis**, Elsevier B.V., v. 56, n. 6, p. 1920–1934, 2012. Citation on page 45.

BARBIERI, M. M.; BERGER, J. O. Optimal predictive model selection. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 32, n. 3, p. 870 – 897, 2004. Citation on page 46.

BARNDORFF-NIELSEN, O.; KENT, J.; SøRENSEN, M. Normal variance-mean mixtures and z distributions. **International Statistical Review / Revue Internationale de Statistique**, v. 50, n. 2, p. 145–159, 1982. Citation on page 31.

BELL, L.; ZHANG, J.; NIU, X. Mixture of logistic models and an ensemble approach for protein-protein interaction extraction. **2011 ACM Conference on Bioinformatics, Computational Biology and Biomedicine, BCB 2011**, p. 371–375, 2011. Citation on page 26.

BENAGLIA, T.; CHAUVEAU, D.; HUNTER, D. R.; YOUNG, D. S. mixtools: An r package for analyzing mixture models. **Journal of Statistical Software**, v. 32, n. 6, p. 1–29, 2009. Citation on page 25.

CAO, X.; LEE, K.; HUANG, Q. Bayesian variable selection in logistic regression with application to whole-brain functional connectivity analysis for parkinson's disease. **Statistical Methods in Medical Research**, v. 30, p. 826 – 842, 2020. Citation on page 42.

CASELLA, G.; GEORGE, E. Explaning the gibbs sampler. **The American Statistician**, v. 46, p. 167–174, 1992. Citation on page 20.

CHEN, B.; YE, K. Componentwise variable selection in finite mixture regression. **Statistics and Its Interface**, v. 8, p. 239–254, 2015. Citation on page 21.

CHEN, J.; CHEN, Z. Extended bayesian information critera for model selection with large model spaces. **Biometrika**, v. 95, p. 759–771, 2008. Citations on pages 20 and 40.

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **American Statistician**, v. 49, p. 327–335, 1995. Citation on page 20.

CORTEZ, P.; SILVA, A. M. G. Using data mining to predict secondary school student performance. **Proceedings of 5th Annual Future Business Technology Conference, Porto**, EUROSIS-ETI, p. 5–12, 2008. Citation on page 85.

COWLES, M. K.; CARLIN, B. P. Markov chain monte carlo convergence diagnostics: A comparative review. **Journal of the American Statistical Association**, v. 91, n. 434, p. 883–904, 1996. Citation on page 48.

DAUVIER, B.; CHEVALIER, N.; BLAYE, A. Using finite mixture of GLMs to explore variability in children's flexibility in a task-switching paradigm. **Cognitive Development**, Elsevier Inc., v. 27, n. 4, p. 440–454, 2012. Citation on page 26.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 39, n. 1, p. 1–22, 1977. Citations on pages 19 and 28.

DENG, W.; CHEN, H.; LI, Z. A logistic regression mixture model for interval mapping of genetic trait loci affecting binary phenotypes. **Genetics**, v. 172, n. 2, p. 1349–58, 2006. Citation on page 19.

DEVIJVER, E. An l1 oracle inequality for the lasso in finite mixture of multivariate gaussian regression models. **ESAIM: Probability and Statistics**, v. 19, p. 649–670, 2015. Citation on page 20.

DOBSON, A. J.; BARNETT, A. G. **An introduction to generalized linear models**. 3. ed. London: Chapman & Hall, 2008. Citation on page 25.

FOLLMANN, D.; LAMBERT, D. Identifiability of finite mixture of logistic regression models. **Journal of Statistical Planning and Inference**, v. 27, n. 3, p. 375–381, 1991. Citations on pages 78, 83, and 92.

FRüHWIRTH-SCHNATTER, S. **Finite Mixture and Markov Switching Models**. 1. ed. New York: Springer, Series in Statistics, 2006. Citations on pages 23 and 27.

FRüHWIRTH-SCHNATTER, S.; CELEUX, G. **Handbook of Mixture Analysis**. 1. ed. London: Chapman & Hall, Handbooks of Modern Statistical Methods, 2018. Citations on pages 23 and 27.

GEORGE, E.; MCCULLOCH, R. Approaches for bayesian variable selection. **Statistica Sinica**, v. 7, p. 339–373, 1997. Citation on page 42.

GEORGE, E. I.; MCCULLOCH, R. E. Variable selection via gibbs sampling. **Journal of The American Statistical Association**, v. 88, p. 881–889, 1993. Citations on pages 21, 43, and 44.

GEORGE_, E. I.; MCCULLOCH, R. E. Stochastic search variable selection. **Markov chain Monte Carlo in practice**, Chapman and Hall, v. 68, p. 203–214, 1996. Citation on page 43.

GREEN, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, v. 82, n. 4, p. 711–732, 1995. Citation on page 20.

GRÜN, B.; LEISCH, F. Finite mixtures of generalized linear regression models. In: **Recent advances in linear models and related areas**. New York: Springer, 2008. p. 205–230. Citation on page 26.

GUPTA, M.; IBRAHIM, J. G. Variable selection in regression mixture modeling for the discovery of gene regulatory networks. **Journal of the American Statistical Association**, v. 102, n. 479, p. 867–880, 2007. Citation on page 45.

ISHWARAN, H.; RAO, J. S. Spike and slab variable selection: Frequentist and Bayesian strategies. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 33, n. 2, p. 730 – 773, 2005. Citation on page 43.

KHALILI, A.; CHEN, J. Variable selection in finite mixture of regression models. **Journal of the American Statistical Association**, v. 102, p. 1025–1038, 2007. Citation on page 20.

KHALILI, A.; LIN, S. Regularization in finite mixture of regression models with diverging number of parameters. **Biometrics**, v. 69, n. 2, p. 436–446, 2013. Citation on page 20.

KONISHI, R.; NAKAMURA, F.; KIYOKI, Y. Estimating adaptive individual interests and needs based on online local variational inference for a logistic regression mixture model. **International Electronics Symposium on Knowledge Creation and Intelligent Computing, IES-KCIC 2018 - Proceedings**, IEEE, p. 164–169, 2019. Citation on page 26.

KUO, L.; MALLICK, B. Variable selection for regression models. **Sankhyā: The Indian Journal of Statistics**, v. 60, n. 1, p. 65–81, 1998. Citation on page 44.

LEE, K.-J.; CHEN, R.-B.; WU, Y. N. Bayesian variable selection for finite mixture model of linear regressions. **Computational Statistics & Data Analysis**, v. 95, p. 1–16, 2016. Citations on pages 21, 42, 45, 47, and 57.

LEE, K.-J.; FELDKIRCHER, M.; CHEN, Y.-C. Variable selection in finite mixture of regression models with an unknown number of components. **Computational Statistics & Data Analysis**, v. 158, p. 107–180, 2021. Citation on page 21.

LI, G. Application of finite mixture of logistic regression for heterogeneous merging behavior analysis. **Journal of Advanced Transportation**, v. 2018, p. 1–9, 2018. Citation on page 19.

LIANG, F.; PAULO, R.; MOLINA, G.; CLYDE, M.; BERGER, J. Mixtures of g priors for Bayesian variable selection. **Journal of the American Statistical Association**, v. 103, p. 410–423, 2008. Citation on page 45.

LLOYD-JONES, L. R.; NGUYEN, H. D.; MCLACHLAN, G. J. A globally convergent algorithm for lasso-penalized mixture of linear regression models. **Computational Statistics & Data Analysis**, v. 119, p. 19–38, 2018. Citation on page 20.

MCLACHLAN, P.; NELDER, J. **Generalized Linear Models**. 2. ed. London: Chapman & Hall, 1989. Citation on page 25.

MCLACHLAN, P.; PEEL, D. **Finite Mixture Models**. 1. ed. New York: John Wiley & Sons, 2000. Citations on pages 23 and 27.

MELNYKOV, V.; MAITRA, R. Finite mixture models and model-based clustering. **Statistics Surveys**, v. 4, p. 80–116, 2010. Citation on page 19.

MITCHELL, T. J.; BEAUCHAMP, J. J. Bayesian variable selection in linear regression. **Journal of the American Statistical Association**, Taylor & Francis, v. 83, n. 404, p. 1023–1032, 1988. Citation on page 43.

NAIK, P.; SHI, P.; TSAI, C.-L. Extending the Akaike information criterion to mixture regression models. **Journal of the American Statistical Association**, v. 102, p. 244–254, 2007. Citation on page 20.

PAPASTAMOULIS, P. label.switching: An R package for dealing with the label switching problem in MCMC outputs. **Journal of Statistical Software, Code Snippets**, v. 69, n. 1, p. 1–24, 2016. Citation on page 41.

PAPASTAMOULIS, P.; ILIOPOULOS, G. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. **Journal of Computational and Graphical Statistics**, v. 19, n. 2, p. 313–331, 2010. Citation on page 41.

POLSON, N. G.; SCOTT, J. G.; WINDLE, J. Bayesian inference for logistic models using Pólya-Gamma latent variables. **Journal of the American Statistical Association**, v. 108, n. 504, p. 1339–1349, 2013. Citations on pages 20, 21, 31, 32, 33, and 39.

RIBEIRO, L.; SARAIVA, E.; SUZUKI, A. Um tutorial sobre estimação de modelos de mistura. 2019. Citation on page 28.

SCHWARZ, G. Estimating the Dimension of a Model. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461 – 464, 1978. Citation on page 19.

SPIEGELHALTER, D.; BEST, N.; CARLIN, B.; LINDE, A. The deviance information criterion: 12 years on. **Journal of the Royal Statistical Society**, v. 76, p. 485–493, 2014. Citations on pages 19 and 40.

STÄDLER, N.; BÜHLMANN, P.; GEER, S. van de. L1-penalization for mixture regression models. **Test**, v. 19, p. 209–256, 2010. Citation on page 20.

TEICHER, H. Identifiability of finite mixtures. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 34, n. 4, p. 1265–1269, 1963. Citations on pages 27 and 78.

TüCHLER, R. Bayesian variable selection for logistic models using auxiliary mixture sampling. **Journal of Computational and Graphical Statistics**, v. 17, p. 76–94, 2012. Citation on page 20.

WAN, K.; GRIFFIN, J. An adaptive MCMC method for Bayesian variable selection in logistic and accelerated failure time regression models. **Statistics and Computing**, v. 31, n. 1, p. 1–11, 2021. Citation on page 20.

WANG, P.; PUTERMAN, M. L. Mixed logistic regression models. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 3, n. 2, p. 175–200, 1998. Citations on pages 23 and 26.

WINDLE, J. B. **Forecasting high-dimensional, time-varying variance-covariance matrices with high-frequency data and sampling Pólya-Gamma random variates for posterior distributions derived from logistic likelihoods**. Phd Thesis (PhD Thesis) — Departament of Computational Science, Engineering, and Mathematics University of Texas at Austin, 2013. Citation on page 32.

YANG, C.-C.; MUTHÉN, B. O.; YANG, C.-C. Finite mixture multivariate generalized linear models using gibbs sampling and e-m algorithms. **Proceedings-National Science Council Republic of China Part a Physical Science and Engineering**, Citeseer, v. 23, n. 6, p. 695–702, 1999. Citation on page 26.

ZELLNER, A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. **Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti**, v. 6, p. 332–365, 1986. Citations on pages 21 and 44.

APPENDIX

A

CONDITIONAL POSTERIOR DISTRIBUTIONS

A.1 Conditional Posterior Distribution of Weights

Assuming a Dirichlet prior for $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_k)$ we have that

$$\begin{aligned} p(\boldsymbol{\pi}|\cdot) &\propto f(\mathbf{y}, \mathbf{S}, \mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\pi}) p(\boldsymbol{\pi}) \\ &\propto \prod_{i=1}^{K} \pi_i^{n_i} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1} \\ &= \prod_{i=1}^{K} \pi_i^{n_i + \alpha_i - 1} \\ &\propto Dirichlet(n_1 + \alpha_1, \cdots, n_k + \alpha_k). \end{aligned}$$

A.2 Conditional Posterior Distribution of Regression Coefficients

Assuming a Normal multivariate prior with mean $\boldsymbol{\mu}_h$ and variance-covariance matrix $\boldsymbol{\Sigma}_h$ for each $\boldsymbol{\beta}_h$ we have that,

$$
\begin{aligned}
p(\boldsymbol{\beta}_h|\cdot) &\propto f(\mathbf{y},\mathbf{S},\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\pi})p(\boldsymbol{\beta}_h) \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{z}_h-\mathbf{X}_h\boldsymbol{\beta}_h)^\intercal\mathbf{W}_h(\mathbf{z}_h-\mathbf{X}_h\boldsymbol{\beta}_h)\right\}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_h-\boldsymbol{\mu}_h)^\intercal\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\beta}_h-\boldsymbol{\mu}_h)\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}_h-\boldsymbol{\mu}_h)^\intercal\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\beta}_h-\boldsymbol{\mu}_h)+(\mathbf{z}_h-\mathbf{X}_h\boldsymbol{\beta}_h)^\intercal\mathbf{W}_h(\mathbf{z}_h-\mathbf{X}_h\boldsymbol{\beta}_h)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_h^\intercal\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\beta}_h-\boldsymbol{\beta}_h^\intercal\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h-\boldsymbol{\mu}_h^\intercal\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\beta}_h+\boldsymbol{\mu}_h^\intercal\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h+(\mathbf{X}_h\boldsymbol{\beta}_h)^\intercal\mathbf{W}_h(\mathbf{X}_h\boldsymbol{\beta}_h)\right.\right. \\
&\quad \left.\left.-\ (\mathbf{X}_h\boldsymbol{\beta}_h)^\intercal\mathbf{W}_h\mathbf{z}_h-\mathbf{z}_h^\intercal\mathbf{W}_h(\mathbf{X}_h\boldsymbol{\beta}_h)+\mathbf{z}_h^\intercal\mathbf{W}_h\mathbf{z}_h\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_h^\intercal\left(\boldsymbol{\Sigma}_h^{-1}+\mathbf{X}_h^\intercal\mathbf{W}_h\mathbf{X}_h\right)\boldsymbol{\beta}_h-\boldsymbol{\beta}_h^\intercal\left(\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h+\mathbf{X}_h^\intercal\mathbf{W}_h\mathbf{z}_h\right)-\left(\boldsymbol{\mu}_h^\intercal\boldsymbol{\Sigma}_h^{-1}+\mathbf{z}_h^\intercal\mathbf{W}_h\mathbf{X}_h\right)\boldsymbol{\beta}_h\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_h^\intercal\mathbf{V}^{-1}\boldsymbol{\beta}_h-\boldsymbol{\beta}_h^\intercal\mathbf{A}-\mathbf{A}^\intercal\boldsymbol{\beta}_h\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_h^\intercal\mathbf{V}^{-1}\boldsymbol{\beta}_h-\boldsymbol{\beta}_h^\intercal\mathbf{V}^{-1}\mathbf{V}\mathbf{A}-\mathbf{A}^\intercal\mathbf{V}^{-1}\mathbf{V}\boldsymbol{\beta}_h\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_h^\intercal\mathbf{V}^{-1}\boldsymbol{\beta}_h-\boldsymbol{\beta}_h^\intercal\mathbf{V}^{-1}\mathbf{V}\mathbf{A}-\mathbf{A}^\intercal\mathbf{V}^{-1}\mathbf{V}\boldsymbol{\beta}_h+\mathbf{A}^\intercal\mathbf{V}^\intercal\mathbf{V}^{-1}\mathbf{V}\mathbf{A}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\beta}_h^\intercal-\mathbf{A}^\intercal\mathbf{V}\right)\mathbf{V}^{-1}\left(\boldsymbol{\beta}_h-\mathbf{V}\mathbf{A}\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}_h-\mathbf{m})^\intercal\mathbf{V}^{-1}\left(\boldsymbol{\beta}_h-\mathbf{m}\right)\right]\right\} \\
&\propto \mathit{Normal}\left(\mathbf{m},\mathbf{V}\right),
\end{aligned}
$$

where $\mathbf{V}=(\boldsymbol{\Sigma}_h^{-1}+\mathbf{X}_h^\intercal\mathbf{W}_h\mathbf{X}_h)^{-1}$ and $\mathbf{m}=\mathbf{V}(\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h+\mathbf{X}_h^\intercal\mathbf{W}_h\mathbf{z}_h)$.

A.3 Marginalized Likelihood Function

Consider the likelihood function given in (3.10). For a fixed component i, the likelihood function integrating out $\boldsymbol{\beta}_{\gamma_i}$ is calculated following the same steps as in Appendix A.2, obtaining

$$
\begin{aligned}
f(\mathbf{y}_i|\mathbf{S}_i,\mathbf{w}_i,\boldsymbol{\pi},\boldsymbol{\gamma}_i) &= \int f(\mathbf{y}_i|\mathbf{S}_i,\mathbf{w}_i,\boldsymbol{\pi},\boldsymbol{\beta}_{\gamma_i})p(\boldsymbol{\beta}_{\gamma_i})d\boldsymbol{\beta}_{\gamma_i} \\
&\propto \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_{\gamma_i}|}}\int \exp\left\{-\frac{1}{2}\left[(\mathbf{z}_i-\mathbf{X}_{\gamma_i}\boldsymbol{\beta}_{\gamma_i})^\intercal\mathbf{W}_i(\mathbf{z}_i-\mathbf{X}_{\gamma_i}\boldsymbol{\beta}_{\gamma_i})+(\boldsymbol{\beta}_{\gamma_i}-\boldsymbol{\mu}_{\gamma_i})^\intercal\boldsymbol{\Sigma}_{\gamma_i}^{-1}(\boldsymbol{\beta}_{\gamma_i}-\boldsymbol{\mu}_{\gamma_i})\right]\right\}d\boldsymbol{\beta}_{\gamma_i} \\
&\propto \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_{\gamma_i}|}}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}_{\gamma_i}^\intercal\Sigma_{\gamma_i}^{-1}\boldsymbol{\mu}_{\gamma_i}\right\}\int \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_{\gamma_i}^\intercal\left(\boldsymbol{\Sigma}_{\gamma_i}^{-1}+\mathbf{X}_{\gamma_i}^\intercal\mathbf{W}_i\mathbf{X}_{\gamma_i}\right)\boldsymbol{\beta}_{\gamma_i}-\boldsymbol{\beta}_{\gamma_i}^\intercal\left(\boldsymbol{\Sigma}_{\gamma_i}^{-1}\boldsymbol{\mu}_{\gamma_i}+\mathbf{X}_{\gamma_i}^\intercal\mathbf{W}_i\mathbf{z}_i\right)-\right.\right. \\
&\quad \left.\left.\left(\boldsymbol{\mu}_{\gamma_i}^\intercal\boldsymbol{\Sigma}_{\gamma_i}^{-1}+\mathbf{z}_i^\intercal\mathbf{W}i\mathbf{X}_{\gamma_i}\right)\boldsymbol{\beta}_{\gamma_i}\right]\right\}d\boldsymbol{\beta}_{\gamma_i} \\
&= \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_{\gamma_i}|}}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}_{\gamma_i}^\intercal\Sigma_{\gamma_i}^{-1}\boldsymbol{\mu}_{\gamma_i}\right\}\int \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_{\gamma_i}^\intercal\mathbf{V}^{-1}\boldsymbol{\beta}_{\gamma_i}-\boldsymbol{\beta}_{\gamma_i}^\intercal\mathbf{A}-\mathbf{A}^\intercal\boldsymbol{\beta}_{\gamma_i}\right]\right\}d\boldsymbol{\beta}_{\gamma_i} \\
&= \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_{\gamma_i}|}}\exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_{\gamma_i}^\intercal\Sigma_{\gamma_i}^{-1}\boldsymbol{\mu}_{\gamma_i}-\mathbf{A}^\intercal\mathbf{V}^\intercal\mathbf{A}\right]\right\}\int \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_{\gamma_i}^\intercal\mathbf{V}^{-1}\boldsymbol{\beta}_{\gamma_i}-\boldsymbol{\beta}_{\gamma_i}^\intercal\mathbf{V}^{-1}\mathbf{V}\mathbf{A}-\mathbf{A}^\intercal\mathbf{V}^{-1}\mathbf{V}\boldsymbol{\beta}_{\gamma_i}+\right.\right. \\
&\quad \left.\left.\mathbf{A}^\intercal\mathbf{V}^\intercal\mathbf{V}^{-1}\mathbf{V}\mathbf{A}\right]\right\}d\boldsymbol{\beta}_{\gamma_i} \\
&= \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_{\gamma_i}|}}\exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_{\gamma_i}^\intercal\Sigma_{\gamma_i}^{-1}\boldsymbol{\mu}_{\gamma_i}-\mathbf{A}^\intercal\mathbf{V}^\intercal\mathbf{A}\right]\right\}\int \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\beta}_{\gamma_i}^\intercal-\mathbf{A}^\intercal\mathbf{V}\right)\mathbf{V}^{-1}\left(\boldsymbol{\beta}_{\gamma_i}-\mathbf{V}\mathbf{A}\right)\right]\right\}d\boldsymbol{\beta}_{\gamma_i} \\
&= \left(\frac{|\mathbf{V}|}{|\boldsymbol{\Sigma}_{\gamma_i}|}\right)^{1/2}\exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_{\gamma_i}^\intercal\Sigma_{\gamma_i}^{-1}\boldsymbol{\mu}_{\gamma_i}-\mathbf{m}^\intercal\mathbf{A}\right]\right\}\int \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\beta}_{\gamma_i}^\intercal-\mathbf{m}\right)^\intercal\mathbf{V}^{-1}\left(\boldsymbol{\beta}_{\gamma_i}-\mathbf{m}\right)\right]\right\}\frac{1}{\sqrt{(2\pi)^d|\mathbf{V}|}}d\boldsymbol{\beta}_{\gamma_i} \\
&= \left(\frac{|\mathbf{V}|}{|\boldsymbol{\Sigma}_{\gamma_i}|}\right)^{1/2}\exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_{\gamma_i}^\intercal\Sigma_{\gamma_i}^{-1}\boldsymbol{\mu}_{\gamma_i}-\mathbf{m}^\intercal\mathbf{A}\right]\right\},
\end{aligned}
$$

where d is the number of relevant covariates, with $\gamma_{it}=1$, $\mathbf{V}=(\boldsymbol{\Sigma}_{\gamma_i}^{-1}+\mathbf{X}_{\gamma_i}^\intercal\mathbf{W}_i\mathbf{X}_{\gamma_i})^{-1}$, $\mathbf{A}=(\boldsymbol{\Sigma}_{\gamma_i}^{-1}\boldsymbol{\mu}_{\gamma_i}+\mathbf{X}_{\gamma_i}^\intercal\mathbf{W}_i\mathbf{z}_i)$ and $\mathbf{m}=\mathbf{V}\mathbf{A}$.


ICMC
USP
SÃO CARLOS