UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelo de mistura de regressão: uma abordagem Bayesiana

Luiz Gabriel Fernandes Cotrim

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

ICMC USP SÃO CARLOS

Luiz Gabriel Fernandes Cotrim

Modelo de mistura de regressão: uma abordagem Bayesiana

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 Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
Março de 2020

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

F363m

Fernandes Cotrim, Luiz Gabriel
Modelo de mistura de regressão: uma abordagem Bayesiana / Luiz Gabriel Fernandes Cotrim; orientadora Daiane Aparecida Zuanetti. -- São Carlos, 2020.
109 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. Modelos de Mistura. 2. Modelos de Mistura de Regressão. 3. Inferência Bayesiana. 4. MCMC. 5. Reversible Jump. 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

Luiz Gabriel Fernandes Cotrim

Regression mixture model: a Bayesian approach

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 Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
March 2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Luiz Gabriel Fernandes Cotrim, realizada em 14/04/2020:

Daiane Aparecida Zuanetti
Profa. Dra. Daiane Aparecida Zuanetti
UFSCar

Daiane Aparecida Zuanetti
Profa. Dra. Rosineide Fernando da Paz
UFC

Daiane Aparecida Zuanetti
Prof. Dr. Michel Helcias Montoril
UFSCar

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Daiane Aparecida Zuanetti Rosineide Fernando da Paz, Michel Helcias Montoril e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Daiane Aparecida Zuanetti
Profa. Dra. Daiane Aparecida Zuanetti

Este trabalho é dedicado a minha vó, Martinha Rodrigues Fernandes.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Miriam Fernandes e João Roberto Cotrim, que, além de apoio, carinho e amor, sempre me proporcionaram um ambiente fértil e plural de oportunidades.

Agradeço aos meus irmãos, Luiza, Karolina e Paulo, pelo amor e companheirismo.

À toda minha família por sempre estarem juntos e me mostrarem a verdadeira magia da união.

Aos meus amigos pelo suporte emocional e pelos momentos de alegria, especialmente Lucas Sales Fidelis, Vitor de Araujo, Franco Simões e Raquel Gruppi.

À professora e orientadora Daiane Aparecida Zuanetti por ter acreditado em mim, pelo conhecimento compartilhado e pela dedicação empenhada neste trabalho.

Aos professores, Mario de Castro, Mariana Curi, Vicente G. Cancho, Rafael Izbicki, Luis Aparecido Milan, Gustavo Henrique de Araujo Pereiro e, especialmente, Vera Lucia Damasceno Tomazella pelas disciplinas ministradas e por todo conhecimento compartilhado.

Por fim, agradeço à CAPES, pois o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

COTRIM, L. G. F. **Modelo de mistura de regressão: uma abordagem Bayesiana**. 2020. 109 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, 2020.

Nesse trabalho, estudamos os modelos de mistura de regressão e apresentamos duas metodologias Bayesianas para a estimação deles. A primeira considerando que o número de componentes é conhecido e propomos a utilização de critérios de seleção de modelos com enfoque Bayesiano, *DIC* e *EBIC*, para estimar o número de componentes da mistura. Na segunda, propomos um algoritmo *reversible jump* com passos de *split-merge* que estima conjuntamente os parâmetros do modelo e o número de componentes da mistura. Aplicamos as metodologias propostas e também o algoritmo EM, já disponível em pacote R, em dados simulados e em dados educacionais brasileiros, estudando a relação entre o Índice de Desenvolvimento da Educação Básica e alguns dados socioeconômicos e demográficos.

Palavras-chave: Modelo de mistura, modelo de mistura de regressão, inferência Bayesiana, MCMC, *DIC*, *EBIC*, IDEB, *Data-driven reversible jump*.

ABSTRACT

COTRIM, L. G. F. **Regression mixture model: a Bayesian approach**. 2020. 109 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, 2020.

In the current dissertation, we study the mixture regression models and present two Bayesian methodologies for their estimation. The first one considers the number of components is known and we propose the use of two Bayesian model selection *criteria*, *DIC* and *EBIC*, to identify the number of components. In the other one, we propose a reversible jump algorithm with split-merge steps that estimates parameters and the number of components. We apply the proposed methodologies and also the EM algorithm, already available in R package, for simulated dataset and for Brazilian educational data, studying the relationship among the Basic Education Development Index and some socioeconomic and demographic data.

Keywords: Mixture model, mixture regression model, Bayesian approach, MCMC, *DIC*, *EBIC*, IDEB, Data-driven reversible jump.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

É comum nos depararmos com dados que não podem ser adequadamente modelados por famílias paramétricas de distribuições padrões e os modelos com misturas de distribuições, denominados modelos de mistura, fornecem-nos uma conveniente forma de modelar estes dados.

Além de fornecer uma estrutura para modelar distribuições mais complexas, os modelos de mistura são utilizados para modelar dados cujas observações são provenientes de uma população composta por K subpopulações. Cada subpopulação recebe o nome de componente da mistura e é ponderada por sua frequência relativa na população, que denominamos de pesos da mistura. Desta maneira, identificando de qual das K subpopulações cada observação é proveniente, o modelo de mistura também se torna uma importante ferramenta para agrupamento de dados, além de permitir outras inferências sobre a população. Para uma ampla revisão sobre modelos de mistura, ver McLachlan e Peel (2000) e Frühwirth-Schnatter (2006).

Os modelos de mistura de regressão, propostos por Goldfeld e Quandt (1973), e conhecidos em econometria como regressões de comutação, são modelos de regressão nos quais a distribuição da variável resposta condicionada as covariáveis é um modelo de mistura. Em outras palavras, utilizamos um modelo de mistura de regressão quando a variável resposta se relaciona diferentemente com as covariáveis de acordo com qual componente não-observável e desconhecida o elemento amostral é proveniente.

Do ponto de vista da inferência frequentista, a estimação de um modelo de mistura de regressão é usualmente realizada através do algoritmo *Expectation-Maximization* (EM), que é abordado por McLachlan e Krishnan (2007). Bai, Yao e Boyer (2012) demonstram que as estimativas de máxima verossimilhança são sensíveis a *outliers* nos modelos de mistura de regressões normais e propõem um algoritmo robusto de estimação dos parâmetros, alterando o algoritmo EM existente. Huang e Yao (2012) estudam a classe de modelos de mistura de regressão normal semi-paramétrico, permitindo que as proporções da mistura dependam de covariáveis. Mostra, ainda, que tais modelos, sob certas condições, são identificáveis e, portanto,

não possuem o problema de *label switching*, que será discutido no decorrer deste trabalho. Por fim, propõem um passo adicional no método de estimação *backfitting* para o modelo proposto e investiga modificações no algoritmo EM. Wang *et al.* (1996) estudam modelos de mistura de regressão Poisson permitindo que as proporções da mistura também dependam de covariáveis e propõe um processo de estimação dos parâmetros baseado no algoritmo EM e no algoritmo *quasi-Newton*. Quandt e Ramsey (1978) estima os parâmetros dos modelos de mistura e dos modelos de mistura de regressões normais através do *moment generating function estimator* definido como o estimador que minimiza a soma dos quadrados da diferença entre a função geradora de momentos teórica e amostral.

Como é sabido na literatura, o algoritmo EM é geralmente sensível ao ponto inicial, pode não convergir ao ponto de máximo global e, algumas vezes, apresenta lenta convergência. Devido a estes problemas, este trabalho adota a abordagem Bayesiana para a estimação de modelos de mistura, em especial os modelos de mistura de regressão que, em modelos multiparamétricos, geralmente utiliza os algoritmos *Gibbs Sampling* (CASELLA; GEORGE, 1992), *Metropolis-Hastings* (CHIB; GREENBERG, 1995) ou *Reversible Jump* (GREEN, 1995) para simular amostras da distribuição *a posteriori* conjunta dos parâmetros e estimar o modelo em questão.

A escolha do algoritmo depende do caso a ser estudado. Frühwirth-Schnatter (2001) propõe um algoritmo *Markov Chain Monte Carlo* (MCMC) denominado *Permutation sampling* que é baseado em sucessivas simulações dos parâmetros através da distribuição *a posteriori* e permuta aleatoriamente, a cada passo, a marcação atual dos estados, tratando o problema de *label switching* com identificabilidade artificial. Hurn, Justel e Robert (2003) revisa brevemente os métodos *Gibbs Sampling* e *Metropolis–Hastings* para a estimação dos parâmetros de modelos de mistura de regressão normal, logístico e Poisson. Também propõe um método de estimação MCMC *Reversible Jump* baseado na técnica de nascimento e morte para estimação dos parâmetros quando o número de componentes da mistura é desconhecido, mostrando como a escolha adequada da função de perda pode resolver o problema de *label switching*.

Aplicações são encontradas em diversas áreas, como Economia (WEDEL; DESARBO, 1993), Biologia (WANG *et al.*, 1996) e Genética (QIN; SELF, 2006). Com o objetivo de melhor entender a realidade do nível educacional brasileiro, este trabalho aplica a metodologia proposta em dados educacionais do Brasil, considerando o Índice de Desenvolvimento da Educação Básica (IDEB) como nossa variável de interesse.

Neste trabalho, apresentamos e discutimos algumas metodologias Bayesianas de estimação e seleção de modelos para o caso em que temos uma mistura de regressões normais. Analisamos a performance das metodologias propostas e do algoritmo EM, já disponível em pacote no R, em dados simulados e, por fim, aplicamos em dados reais sobre a qualidade da educação brasileira.

Este trabalho está organizado como segue: No Capítulo 2, introduzimos os modelos de mistura na forma original e hierárquica, abordando sua ampla aplicabilidade e flexibilidade

para tratar dados que não são facilmente modelados por famílias de distribuições paramétricas conhecidas. Em seguida, apresentamos os modelos de mistura Gaussiana e um algoritmo MCMC de estimação dos parâmetros considerando o número de componentes da mistura conhecido. No Capítulo 3, abordamos o tema principal dessa dissertação de mestrado, que são os modelos de mistura de regressão. Estudamos com detalhes os modelos de mistura de regressões normais, apresentando duas metodologias Bayesianas de estimação dos parâmetros. No Capítulo 4, aplicamos as metodologias apresentadas no Capítulo 3 a dados simulados e reais, comparando as estimativas obtidas com as obtidas pelo algoritmo EM. Por fim, no Capítulo 5 apresentamos nossas conclusões e propostas futuras.

CAPÍTULO

2

MODELO DE MISTURA DE DISTRIBUIÇÕES

Os modelos de mistura de distribuições são utilizados para modelar dados provenientes de uma população composta por K subpopulações, onde K pode ser conhecido ou desconhecido. Formalmente, dizemos que a densidade $f(\cdot|\boldsymbol{w},\boldsymbol{\theta})$ é uma mistura de distribuições se ela for uma combinação linear convexa de K densidades, sendo cada uma pertencente a alguma família de distribuições indexadas pelo parâmetro $\boldsymbol{\theta}_k$, isto é,

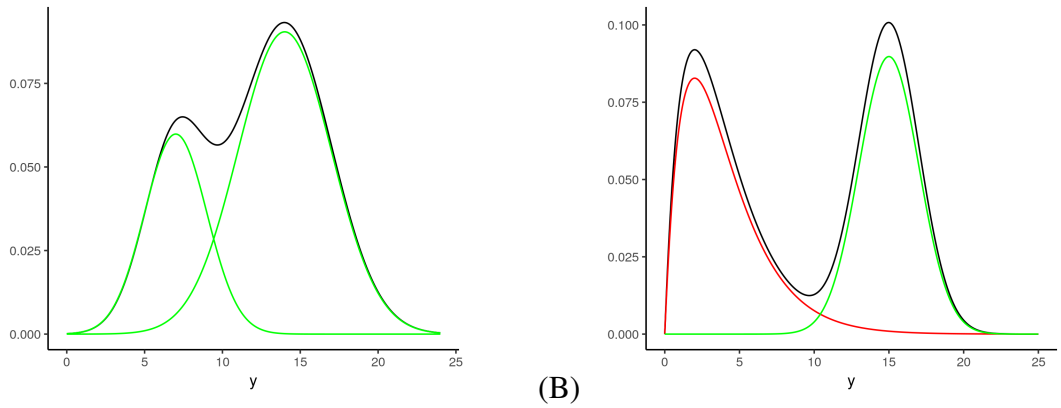
$$f(\cdot|\boldsymbol{w},\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k f(.|\boldsymbol{\theta}_k), \tag{2.1}$$

onde $w_k \geq 0$, $\sum_{k=1}^{K} w_k = 1$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$ e $\boldsymbol{w} = (w_1, ..., w_K)$. Atribuímos aos w_k's o nome de pesos ou ponderações e denominamos de componentes ou subpopulações da mistura as densidades $f(\cdot|\boldsymbol{\theta}_k)$.

Como, no geral, não conhecemos a função densidade de probabilidade (ou função de probabilidade) da população, cada subpopulação é adequadamente modelada por uma densidade pertencente à alguma família de distribuições paramétricas conhecida de forma que facilite o tratamento dos dados. Note que as subpopulações de um mesmo modelo de mistura podem pertencer a famílias de distribuições distintas.

A Figura 1 ilustra dois exemplos de modelos de mistura. Na Figura 1 (A), a densidade em preto representa a distribuição dada por $f(y|\boldsymbol{w},\boldsymbol{\theta}) = 0.3N(7,4) + 0.7N(14,9)$, onde $N(\mu,\sigma^2)$ simboliza a distribuição Normal com média μ e variância σ^2. Neste caso, $\boldsymbol{\theta} = (\boldsymbol{\mu},\boldsymbol{\sigma}^2)$ onde $\boldsymbol{\mu} = (7,14)$ e $\boldsymbol{\sigma}^2 = (4,9)$. As densidades em verde representam, da esquerda para a direita, a distribuição $N(7,4)$ e $N(14,9)$. Quando todas as componentes da mistura são distribuições Normais, chamamos o modelo de Mistura Gaussiana ou Mistura de Normais e estes modelos desempenham um papel muito importante na prática. Já a Figura 1 (B) representa a distribuição dada por $f(y|\boldsymbol{w},\boldsymbol{\theta}) = 0.5Gama(2,0.5) + 0.5N(15,4)$, ou seja, é uma mistura entre uma distribuição Gama e uma distribuição Normal. Em vermelho temos a densidade da distribuição Gama

e em verde a densidade da Normal. A parametrização utilizada da distribuição Gama foi tal que se $Z \sim Gama(a,b)$, então $E(Z) = a/b$.



(A) (B)

Figura 1 – Modelos de Mistura: (A) em preto temos a representação do modelo de mistura dado pela função densidade de probabilidade $f(y|\boldsymbol{w},\boldsymbol{\theta}) = 0.3N(7,4) + 0.7N(14,9)$ e em verde temos a representação da densidade das componentes da mistura; (B) em preto temos a representação do modelo de mistura dado pela função densidade de probabilidade $f(y|\boldsymbol{w},\boldsymbol{\theta}) = 0.5Gama(2,0.5) + 0.5N(15,4)$, em vermelho temos a representação da densidade de sua componente Gama e em verde a representação de sua componente Normal.

Pelas ilustrações acima, fica evidente que os modelos de mistura são uma forma conveniente de modelar dados que não são provenientes de uma mesma distribuição. Além disso, se para cada uma das observações, conseguimos identificar de qual das K subpopulações ela é proveniente, os modelos de mistura de distribuições se tornam uma importante ferramenta de agrupamento.

2.1 Modelo de Mistura de Distribuições

Considere $\boldsymbol{Y} = (Y_1,...,Y_n)$ uma amostra aleatória proveniente de uma população formada por K diferentes subpopulações e com função densidade de probabilidade (ou função de probabilidade, no caso discreto) dada por

$$f(y_i|\boldsymbol{w},\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k f(y_i|\boldsymbol{\theta}_k), \tag{2.2}$$

onde $f(y_i|\boldsymbol{\theta}_k)$ é a distribuição da subpopulação k indexada pelo vetor de parâmetros $\boldsymbol{\theta}_k$, w_k é a frequência relativa da subpopulação k na população, $\boldsymbol{w} = (w_1,...,w_K)$ é o vetor de pesos, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_K)$ é o vetor de parâmetros contendo os vetores paramétricos de cada componente, $k = 1,...,K$ e $i = 1,...,n$ e $\boldsymbol{y} = (y_1,...,y_n)$ o vetor de observações de $\boldsymbol{Y}$.

A função de verossimilhança de $\boldsymbol{w}$ e $\boldsymbol{\theta}$ é dada por

$$L(\boldsymbol{\theta},\boldsymbol{w}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta},\boldsymbol{w}) = \prod_{i=1}^{n}\sum_{k=1}^{K} w_k f(y_i|\boldsymbol{\theta}_k). \tag{2.3}$$

Considerando $\pi(\boldsymbol{\theta}, \boldsymbol{w})$ a distribuição *a priori* conjunta de $\boldsymbol{\theta}$ e $\boldsymbol{w}$, a distribuição *a posteriori* é dada por

$$\pi(\boldsymbol{\theta}, \boldsymbol{w}|\boldsymbol{y}) \propto L(\boldsymbol{\theta}, \boldsymbol{w}|\boldsymbol{y})\,\pi(\boldsymbol{\theta}, \boldsymbol{w}) = \left[\prod_{i=1}^{n}\sum_{k=1}^{K} w_k f(y_i|\boldsymbol{\theta}_k)\right]\pi(\boldsymbol{\theta}, \boldsymbol{w}). \tag{2.4}$$

Note que obter o estimador de máxima verossimilhança de (2.3) e o estimador de Bayes de (2.4) não são tarefas computacionalmente triviais, pois, em ambas, temos a soma de K^n termos.

2.2 Forma Hierárquica do Modelo de Mistura

Com o objetivo de simplificar a estimação dos parâmetros do modelo (2.2), podemos reescrevê-lo na forma hierárquica adicionando uma variável aleatória não-observável com K possíveis valores que indica, para cada uma das observações, de qual componente da mistura ela é proveniente.

Seja $\boldsymbol{S} = (S_1, ..., S_n)$ um conjunto de variáveis aleatórias discretas independentes e não observáveis, tal que $S_i \sim Discreta(\boldsymbol{w})$ para $i = 1, ..., n$ e $\boldsymbol{w} = (w_1, ..., w_K)$, em que $Discreta(\boldsymbol{w})$ representa uma distribuição Discreta com K possíveis valores e uma única observação. Dessa maneira, $P(S_i = k|\boldsymbol{w}) = w_k$ e $\sum_{k=1}^{K} w_k = 1$.

Consideramos que o valor assumido pela variável aleatória S_i identifica de qual das subpopulações a i-ésima observação é proveniente, isto é, $Y_i|S_i = k \sim f(y_i|\boldsymbol{\theta}_k)$. A Figura 2 ilustra essa relação. Note que, da mesma forma que os y's são independentes entre si, os S's também são independentes entre si, no entanto, as setas representam a relação de dependência entre as variáveis aleatórias Y_i e S_i para $i = 1, ..., n$.

Figura 2 – Relação entre a variável não-observável e a variável a ser, realmente, observada na amostra.

Desta forma, podemos reescrever o modelo (2.2) da seguinte maneira

$$\begin{cases} S_i \sim Discreta(\boldsymbol{w} = (w_1, ..., w_K)) \\ \quad Y_i | S_i = k \sim f(y_i|\boldsymbol{\theta}_k), \end{cases} \tag{2.5}$$

que chamamos de forma hierárquica do modelo de mistura.

Por essa estrutura hierárquica, a distribuição marginal de Y_i é dada por

$$\begin{aligned} f(y_i|\boldsymbol{\theta}, \boldsymbol{w}) &= \sum_{k=1}^{K} f(y_i, S_i = k|\boldsymbol{\theta}, \boldsymbol{w}) \\ &= \sum_{k=1}^{K} f(y_i|S_i = k, \boldsymbol{\theta}) P(S_i = k|\boldsymbol{w}) \\ &= \sum_{k=1}^{K} w_k f(y_i|\boldsymbol{\theta}_k) \end{aligned} \tag{2.6}$$

que é exatamente igual a Equação (2.2).

Observando que $P(S_i = k|\boldsymbol{w}) = \prod_{k=1}^{K} {w_k}^{I_{S_i}(k)}$ e $f(y_i|S_i = k, \boldsymbol{\theta}) = \prod_{k=1}^{K} f(y_i|\boldsymbol{\theta}_k)^{I_{S_i}(k)}$, onde $I_{S_i}(k)$ é uma função indicadora que assume valor um se $S_i = k$ e zero caso contrário, temos que a distribuição conjunta de Y_i e S_i é dada por

$$\begin{aligned} f(y_i, S_i = k|\boldsymbol{\theta}, \boldsymbol{w}) &= P(S_i = k|\boldsymbol{w}) f(y_i|S_i = k, \boldsymbol{\theta}) \\ &= \prod_{k=1}^{K} {w_k}^{I_{S_i}(k)} \prod_{k=1}^{K} f(y_i|\boldsymbol{\theta}_k)^{I_{S_i}(k)} \\ &= \prod_{k=1}^{K} {w_k}^{I_{S_i}(k)} f(y_i|\boldsymbol{\theta}_k)^{I_{S_i}(k)}. \end{aligned} \tag{2.7}$$

E a função de verossimilhança aumentada de $\boldsymbol{w}$ e $\boldsymbol{\theta}$ é

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{s}) &= \prod_{i=1}^{n} \prod_{k=1}^{K} {w_k}^{I_{s_i}(k)} f(y_i|\boldsymbol{\theta}_k)^{I_{s_i}(k)} \\ &= \prod_{k=1}^{K} \prod_{i=1}^{n} {w_k}^{I_{s_i}(k)} f(y_i|\boldsymbol{\theta}_k)^{I_{s_i}(k)} \\ &= \prod_{k=1}^{K} \left[{w_k}^{\sum_{i=1}^{n} I_{s_i}(k)} \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}_k)^{I_{s_i}(k)} \right] \\ &= \prod_{k=1}^{K} \left[w_k^{n_k} \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}_k)^{I_{s_i}(k)} \right] \\ &= w_1^{n_1} \dots w_K^{n_K} \prod_{i:s_i=1} f(y_i|\boldsymbol{\theta}_1) \cdots \prod_{i:s_i=K} f(y_i|\boldsymbol{\theta}_K), \end{aligned} \tag{2.8}$$

onde $n_k = \sum_{i=1}^{n} I_{s_i}(k)$ é o número de observações na k-ésima componente da mistura.

Na abordagem frequentista, a estimação de um modelo de mistura é usualmente realizada através do algoritmo EM a partir da equação (2.8). Na abordagem Bayesiana, para estimar os parâmetros geralmente utilizamos os algoritmos *Gibbs Sampling, Metropolis-Hastings* ou *Reversible Jump* para simular amostras da distribuição *a posteriori* conjunta dos parâmetros. A escolha do algoritmo depende do caso a ser estudado.

2.3 Mistura Gaussiana

Quando todas as componentes da mistura pertencem a família de distribuições Normal, nos referimos ao modelo de mistura como mistura Gaussiana ou mistura de normais. Segundo Fraley e Raftery (2002) esse caso especial dos modelos de mistura possui bom desempenho em muitas aplicações, inclusive em algumas situações onde os dados são provenientes de um modelo de mistura de distribuições contínuas não-normais. De acordo com Bishop (2006), eles são comumente usados em mineração de dados, reconhecimento de padrões, aprendizado de máquina e outras análises estatísticas.

Seja $\boldsymbol{y} = (y_1, ..., y_n)$ um vetor de observações de uma amostra aleatória proveniente de uma mistura de distribuições normais com K componentes, isto é, $Y_i|S_i = k \sim N(\mu_k, \sigma_k^2)$ para $k = 1, ..., K$. Da Equação (2.8) temos que a função de verossimilhança aumentada é dada por

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{s}) &= w_1^{n_1} \dots w_K^{n_K} \prod_{i:s_i=1} N(y_i; \mu_1, \sigma_1^2) \cdots \prod_{i:s_i=K} N(y_i; \mu_K, \sigma_K^2) \\ &= w_1^{n_1} \dots w_K^{n_K} \\ &\times \prod_{i:s_i=1} \frac{1}{\sqrt{2\pi}\sigma_1} exp\left[-\frac{1}{2\sigma_1^2}(y_i - \mu_1)^2\right] \\ &\times \prod_{i:s_i=2} \frac{1}{\sqrt{2\pi}\sigma_2} exp\left[-\frac{1}{2\sigma_2^2}(y_i - \mu_2)^2\right] \\ &\times \dots \\ &\times \prod_{i:s_i=K} \frac{1}{\sqrt{2\pi}\sigma_K} exp\left[-\frac{1}{2\sigma_K^2}(y_i - \mu_K)^2\right], \end{aligned} \tag{2.9}$$

onde $\boldsymbol{\theta} = (\mu_1 ... \mu_K, \sigma_1^2, ..., \sigma_K^2)$, $\boldsymbol{w} = (w_1, ..., w_K)$ e $n_k = \sum_{i=1}^n I_{s_i}(k)$ é o número de observações na k-ésima componente da mistura.

2.3.1 Estimação Bayesiana

A princípio vamos supor que o número de componentes da mistura, K, seja conhecido. Também assumiremos que, *a priori*, os parâmetros são independentes, isto é $\pi(\boldsymbol{\theta}, \boldsymbol{w}) = \pi(\mu_1) \dots \pi(\mu_K)\pi(\sigma_1^2) \dots \pi(\sigma_K^2)\pi(\boldsymbol{w})$.

Neste trabalho consideramos as seguintes distribuições *a priori* marginais: $\mu_k \sim N(\mu_{\mu_k}, \sigma_{\mu_k}^2)$, $1/\sigma_k^2 \sim Gama(a_k, b_k)$ para $k = 1, ..., K$ e $\boldsymbol{w} \sim Dirichlet(\gamma_1, ..., \gamma_K)$, onde $\gamma_1, ..., \gamma_K, \mu_{\mu_k}, \sigma_{\mu_k}^2, a_k$ e

b_k são hiperparâmetros conhecidos. A parametrização utilizada da distribuição Gama foi tal que se $Z \sim Gama(a,b)$, então $E(Z) = a/b$.

Com essa configuração, temos que as distribuições *a posteriori* condicionais completas dos parâmetros são conhecidas e dadas pelas mesmas famílias das distribuições *a priori*, como segue:

$$\boldsymbol{w}|\cdots \sim Dirichlet(\gamma_1 + n_1, ..., \gamma_K + n_K), \tag{2.10}$$

$$\mu_k|\cdots \sim N\left(\left(\frac{\mu_{\mu_k}}{\sigma^2_{\mu_k}} + \frac{\sum_{i:s_i=k} y_i}{\sigma^2_k}\right)\left(\frac{1}{\sigma^2_{\mu_k}} + \frac{n_k}{\sigma^2_k}\right)^{-1}, \left(\frac{1}{\sigma^2_{\mu_k}} + \frac{n_k}{\sigma^2_k}\right)^{-1}\right) \tag{2.11}$$

e

$$\frac{1}{\sigma^2_k}|\cdots \sim Gama\left(a_k + \frac{n_k}{2}, b_k + \frac{\sum_{i:S_i=k}(y_i - \mu_k)^2}{2}\right), \tag{2.12}$$

para $k = 1, ..., K$, onde $\cdots$ contempla os dados e todos os parâmetros, exceto o parâmetro em questão. Os cálculos detalhados são apresentados no Apêndice A.

A probabilidade *a posteriori* de $S_i = k$ é dada por

$$\begin{aligned} P(S_i = k|y_i, \boldsymbol{\theta}, \boldsymbol{w}) &= \frac{f(S_i = k, y_i|\boldsymbol{\theta}, \boldsymbol{w})}{f(y_i|\boldsymbol{\theta}, \boldsymbol{w})} \\ &= \frac{f(S_i = k, y_i|\boldsymbol{\theta}, \boldsymbol{w})}{\sum_{t=1}^{K} f(S_i = t, y_i|\boldsymbol{\theta}, \boldsymbol{w})} \\ &= \frac{P(S_i = k|\boldsymbol{w}) f(y_i|S_i = k, \boldsymbol{\theta})}{\sum_{t=1}^{K} P(S_i = t|\boldsymbol{w}) f(y_i|S_i = t, \boldsymbol{\theta})} \\ &= \frac{P(S_i = k|\boldsymbol{w}) f(y_i|\boldsymbol{\theta}_k)}{\sum_{t=1}^{K} P(S_i = t|\boldsymbol{w}) f(y_i|\boldsymbol{\theta}_t)} \\ &= \frac{w_k f(y_i|\boldsymbol{\theta}_k)}{\sum_{t=1}^{K} w_t f(y_i|\boldsymbol{\theta}_t)} \end{aligned} \tag{2.13}$$

para $k = 1, ..., K$, onde $f(y_i|\boldsymbol{\theta}_k)$ é a função densidade da $N(\mu_k, \sigma^2_k)$ calculada no ponto y_i.

Como as distribuições *a posteriori* condicionais completas dos parâmetros são distribuições padrões conhecidas, implementamos o algoritmo *Gibbs Sampling* para simular valores dessas distribuições e gerar amostras da distribuição conjunta. Assumimos como estimativas pontuais dos parâmetros a média *a posteriori* da amostra MCMC obtida para cada parâmetro já considerando o salto e o *burn-in* da cadeia gerada.

Descartamos as primeiras iterações da cadeia gerada considerando um *burn-in* para permitir que a cadeia MCMC atinja convergência e guardamos uma iteração a cada determinado

valor, convenientemente escolhido, de iterações para garantir amostras de valores independentes ou com correlação muito baixa.

2.3.1.1 Algoritmo de Estimação Gibbs Sampling

Considere nosso vetor de parâmetros $\boldsymbol{\theta} = (\mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2)$ e $\boldsymbol{w}$ com distribuição conjunta $\pi(\boldsymbol{\theta}, \boldsymbol{w})$ e distribuições condicionais dadas por (2.10), (2.11) e (2.12). O algoritmo *Gibbs Sampling* é baseado em sucessivas simulações dos parâmetros contidos em $\boldsymbol{\theta}$ e em $\boldsymbol{w}$ e da variável não-observável $\boldsymbol{S}$ através das distribuições *a posteriori* condicionais utilizando as obervações de $\boldsymbol{y}$. Podemos resumir o algoritmo pelos seguintes passos:

Passo 1: para $i = 1, ..., n$, inicialize aleatoriamente S_i com valores de 1 até K, considerando inicialmente que $P(S_i = k) = 1/K$ para $k = 1, ..., K$;

Passo 2: Inicialize arbitrariamente $\mu_k^{(0)}$ para $k = 1, ..., K$;

Passo 3: para $k = 1, ..., K$, gere $1/(\sigma_k^2)^{(0)}$ de

$$Gama\left(a_k + \frac{n_k}{2}, b_k + \frac{\sum_{i:s_i=k}(y_i - \mu_k^{(0)})^2}{2}\right)$$

onde $n_k = \sum_{i=1}^{n} I_{s_i}(k)$ e a_k e b_k são hiperparâmetros escolhidos anteriormente;

Passo 4: para ℓ-ésima iteração, $\ell = 1, ..., L$, onde L é o número total de iterações, faça:

- gere $\mu_k^{(\ell)}$ de

$$N\left(\left(\frac{\mu_{\mu_k}}{\sigma_{\mu_k}^2} + \frac{\sum_{i:s_i=k} y_i}{(\sigma_k^2)^{(\ell-1)}}\right)\left(\frac{1}{\sigma_{\mu_k}^2} + \frac{n_k}{(\sigma_k^2)^{(\ell-1)}}\right)^{-1}, \left(\frac{1}{\sigma_{\mu_k}^2} + \frac{n_k}{(\sigma_k^2)^{(\ell-1)}}\right)^{-1}\right)$$

  para $k = 1, ..., K$;

- gere $1/(\sigma_k^2)^{(\ell)}$ de

$$Gama\left(a_k + \frac{n_k}{2}, b_k + \frac{\sum_{i:s_i=k}(y_i - \mu_k^{(\ell)})^2}{2}\right)$$

  para $k = 1, ..., K$;

- gere $\boldsymbol{w}^{(\ell)} \sim Dirichlet(\gamma_1 + n_1, ..., \gamma_K + n_K)$;

- para $i = 1, ..., n$ atualize o valor de S_i considerando a probabilidade de ocorrência (2.13), isto é,

$$P(S_i = k|\boldsymbol{y}, \boldsymbol{\theta}^{(\ell)}, \boldsymbol{w}^{(\ell)}) = \frac{P(S_i = k|\boldsymbol{w}^{(\ell)}) f(y_i|S_i = k, \boldsymbol{\theta}^{(\ell)})}{\sum_{t=1}^{K} P(S_i = t|\boldsymbol{w}^{(\ell)}) f(y_i|S_i = t, \boldsymbol{\theta}^{(\ell)})}$$

  para $k = 1, ..., K$, onde $f(y_i|\boldsymbol{\theta}_k^{(\ell)})$ é a função densidade da $N(\mu_k^{(\ell)}, (\sigma_k^2)^{(\ell)})$ avaliada no ponto y_i e $P(S_i = k|\boldsymbol{w}^{(\ell)}) = w_k^{(\ell)}$.

Ao final das L iterações, descartamos as B primeiras iterações como *burn-in* e guardamos uma iteração a cada "*saltos*" iterações, obtendo uma cadeia final de tamanho $L_{final} = [(L-B)/saltos]$, onde [] representa a parte inteira da divisão. Consideramos como estimativas pontuais dos parâmetros as médias dos valores gerados, isto é,

$$\hat{\mu}_k = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} \mu_k^{(\ell)}, \tag{2.14}$$

$$\hat{\sigma}_k^2 = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} (\sigma_k^2)^{(\ell)} \tag{2.15}$$

e

$$\hat{w}_k = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} w_k^{(\ell)} \tag{2.16}$$

para $k = 1, ..., K$.

Considerando N_{ik} o número de vezes em que a observação y_i foi associada a componente k nas L_{final} iterações, então, assumimos $P(S_i = k|\cdots) = \frac{N_{ik}}{L_{final}}$ como a probabilidade *a posteriori* de y_i ser proveniente da componente k. Fazemos uso desta probabilidade *a posteriori* para determinar o valor predito de S_i de forma aleatória, para $i = 1, ..., n$. Uma alternativa a esta abordagem, seria considerar como estimativa para S_i a moda *a posteriori* dos valores gerados pela cadeia MCMC. No entanto, ocorrem casos em que a frequencia de dois ou mais valores para S_i são muito próximas e, por isso, não optamos por essa segunda abordagem.

A convergência da cadeia MCMC gerada após *burn-in* e "*saltos*" pode ser verificada por gráficos de traço ou algumas estatísticas de convergência como: estatística de Gelman e Rubin, estatística de Geweke, entre outras (COWLES; CARLIN, 1996). Neste trabalho não detalharemos estes métodos, mas os utilizamos para verificar a convergência das cadeias MCMC nos dados simulados e reais.

Um problema muito frequente e discutido na literatura na estimação de modelos de mistura é o problema conhecido como *label switching*, identificado por alguns autores como problema de não-identificabilidade do modelo. O *label switching* se caracteriza pela permutação dos rótulos das componentes no decorrer do processo de estimação e isso ocorre porque a função de verossimilhança é invariante sob diferentes rotulações das componentes (STEPHENS, 2000). A Figura 3 exemplifica como as cadeias MCMC se comportam na presença de *label switching*. Note que próximo a iteração 8000 ocorre a permutação dos rótulos entre as componentes 1 e 2.

No caso de ocorrência de *label switching* é muito difícil sumarizar a cadeia MCMC para determinar as estimativas dos parâmetros. Alguns métodos tem sido propostos na literatura para corrigir esse problema em modelos de mistura de distribuições, mas a simples inclusão da restrição $\mu_1 < \mu_2 < ... < \mu_K$ tem se mostrado eficiente para modelos de mistura Normal. Para implementar essa restrição na amostra MCMC simulada, basta rearranjar o rótulo das

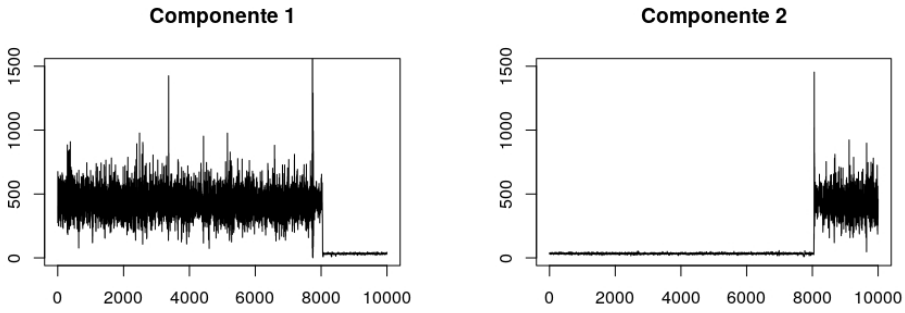


Figura 3 – *Label switching*: troca de rótulos entre a componente 1 e a componente 2 próximo a iteração 8000.

componentes em cada iteração MCMC de modo que $\mu_1^{(\ell)} < \mu_2^{(\ell)} < ... < \mu_K^{(\ell)}$. No próximo capítulo, discutiremos com mais detalhes este assunto.

CAPÍTULO

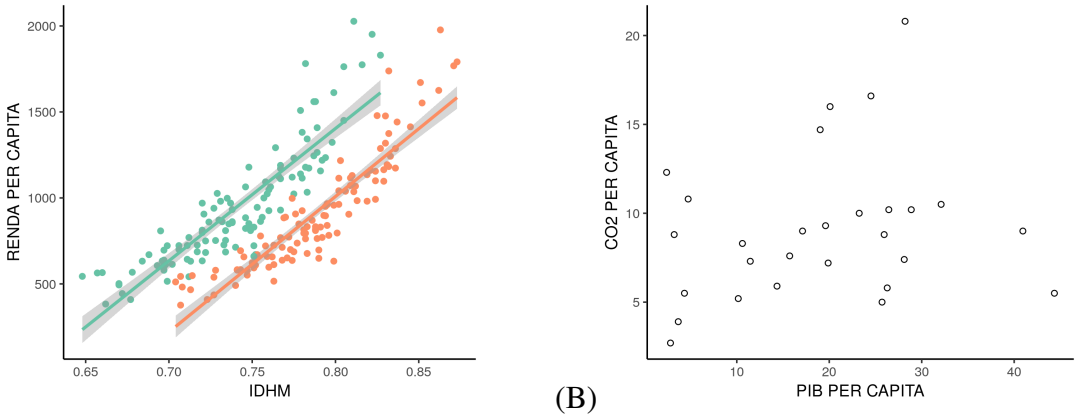
3

MODELO DE MISTURA DE REGRESSÕES

É extremamente relevante entender a relação entre duas ou mais variáveis de interesse. Por exemplo, entender como a temperatura de ebulição da água se altera com a altitude, como a poluição do ar é afetada pela emissão de CO_2, ou como a renda per capita de determinada região é influenciada pelo índice de desenvolvimento humano (IDH). No entanto, é comum que os dados disponíveis para estudo apresentem heterogeneidade, ou seja, que a variável resposta não se relacione com as covariáveis de forma homogênea para toda a população.

A Figura 4 (A) representa a relação entre a renda per capita e o índice de desenvolvimento humano municipal (IDHM) de 111 municípios brasileiros, disponível no censo 2010 realizado pelo IBGE. Note que a relação entre essas variáveis é diferente para homens e para mulheres e, se não tivéssemos a informação sobre o sexo, provavelmente um único modelo de regressão entre renda per capita e IDHM não seria adequado para essa amostra e, consequentemente, para a população. A Figura 4 (B) representa a relação entre o nível de emissão per capita de CO_2 e o *PIB* para 28 países, conjunto de dados *CO2data* publicado pelo *World Resource Institute* e disponível no pacote *mixtools* do software R (YOUNG *et al.*, 2009). Note que a relação entre essas variáveis não é homogênea entre todos os elementos da amostra, havendo dois ou mais grupos que se relacionam de maneira distinta em relação às variáveis consideradas.

Uma maneira interessante de se modelar esses dados é considerar que a heterogeneidade observada surgiu devido ao fato de que a população é composta por K subpopulações e que as variáveis a serem analisadas se relacionam de maneira diferente em cada subpopulação. Se as subpopulações são desconhecidas, podemos assumir que a distribuição da variável resposta condicionada às covariáveis é um modelo de mistura. Então, pelo que foi visto anteriormente, esse modelo é capaz de identificar que a variável resposta se relaciona diferentemente com as covariáveis de acordo com qual componente não-observável e desconhecida o elemento amostral é proveniente. Denominamos esta classe de modelos de modelos de mistura de regressão. Nesse trabalho, a princípio, o foco é o modelo de mistura de regressões normais.



(A) (B)

Figura 4 – (A) representação da renda per capita e do IDHM pra 111 munícipios brasileiros disponibilizados pelo Censo Demográfico de 2010 realizado pelo IBGE, onde os pontos em vermelho representam o sexo feminino e os pontos em verde representam o sexo masculino; (B) representação entre a emissão per capita de CO_2 e o PIB per capita de 28 países do conjunto de dados *CO2data* do pacote *mixtools* do software R.

3.1 Modelo de Mistura de Regressões Normais

Considere $(Y_1, \boldsymbol{X}_1), ..., (Y_n, \boldsymbol{X}_n)$ uma amostra de tamanho n, onde, para $i = 1, ..., n$, Y_i é a variável resposta ou de interesse do i-ésimo indivíduo da amostra e $\boldsymbol{X}_i$ é um vetor $(p+1)$-dimensional que contempla as p variáveis explicativas, ou covariáveis, da i-ésima observação e o valor 1 que acompanha o intercepto da regressão linear. Seja S_i uma variável aleatória não-observável que identifica de qual das K componentes ou subpopulações a observação y_i é proveniente, ou seja, S_i assume o valor k se a observação y_i é proveniente da k-ésima componente. Suponha que $P(S_i = k) = w_k$ para $k = 1, ..., K$, onde K é o número de componentes na mistura, $0 < w_k < 1$ e $\sum_{k=1}^{K} w_k = 1$.

Tomando $S_i = k$, suponha que Y_i dependa linearmente de $\boldsymbol{X}_i$, ou seja,

$$Y_i = \boldsymbol{X}_i \boldsymbol{\beta}_k + \varepsilon_i, \tag{3.1}$$

onde $\boldsymbol{\beta}_k = (\beta_{0k}, ..., \beta_{pk})^T$ e $\varepsilon_i \sim N(0, \sigma_k^2)$, sendo $N(\mu, \sigma^2)$ a função densidade da distribuição normal com média μ e variância σ^2. Desta maneira, temos que a distribuição conjunta de Y_i e S_i condicionada à $\boldsymbol{X}_i = \boldsymbol{x}_i$ pode ser escrita como

$$(Y_i, S_i = k) | \boldsymbol{X}_i = \boldsymbol{x}_i \sim w_k N(\boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2) \tag{3.2}$$

e, consequentemente, a distribuição marginal de Y_i condicionada à $\boldsymbol{X}_i = \boldsymbol{x}_i$ é dada por

$$Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i \sim \sum_{k=1}^{K} w_k N(\boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2). \tag{3.3}$$

Sendo $\boldsymbol{w} = (w_1, ..., w_k)$, $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_K^T, \sigma_1^2, ..., \sigma_K^2)$ e $I_{S_i(k)}$ uma função indicadora que assume valor 1 se $S_i = k$ e 0 caso contrário, podemos escrever a função de verossimilhança aumentada do modelo da seguinte maneira:

$$
\begin{aligned}
L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x}) &= \prod_{i=1}^{n}\prod_{k=1}^{K}(w_k N(\boldsymbol{x}_i\boldsymbol{\beta}_k,\sigma_k^2))^{I_{s_i(k)}} \\
&= \prod_{i=1}^{n}\prod_{k=1}^{K}\left[w_k^{I_{s_i(k)}}\left(\frac{1}{\sqrt{2\pi\sigma_k^2}}exp\left\{-\frac{1}{2\sigma_k^2}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right)^{I_{s_i(k)}}\right] \\
&= \prod_{k=1}^{K}\prod_{i=1}^{n}\left[w_k^{I_{s_i(k)}}\left(\frac{1}{\sqrt{2\pi\sigma_k^2}}exp\left\{-\frac{1}{2\sigma_k^2}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right)^{I_{s_i(k)}}\right] \\
&= \prod_{k=1}^{K}\left[w_k^{n_k}\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma_k^2}}exp\left\{-\frac{1}{2\sigma_k^2}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right)^{I_{s_i(k)}}\right] \\
&= \prod_{k=1}^{K}\left[w_k^{n_k}\prod_{i:s_i=k}\left(\frac{1}{(2\pi\sigma_k^2)^{\frac{1}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right)\right] \\
&= \prod_{k=1}^{K}\left[w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right],
\end{aligned}
$$

onde n_k é a quantidade de elementos da amostra que são provenientes da componente k. Assim,

$$
L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x}) = \prod_{k=1}^{K}\left[w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right]. \tag{3.4}
$$

3.2 Estimação Bayesiana com K Conhecido

A estimação do modelo via abordagem Bayesiana requer que especifiquemos a distribuição *a priori* dos parâmetros do modelo de mistura, $\pi(\boldsymbol{\theta})$. Neste trabalho, consideramos que os parâmetros são independentes *a priori* e assumimos as seguintes distribuições conjugadas

$$
\begin{aligned}
\boldsymbol{w} &\sim Dirichlet(\gamma_1,...,\gamma_K), &(3.5)\\
\beta_{jk} &\sim N(\mu_{\beta_{jk}},\sigma^2_{\beta_{jk}}) &(3.6)
\end{aligned}
$$

e

$$
1/\sigma_k^2 \sim Gama(a_k,b_k), \tag{3.7}
$$

para $j = 0,1,...,p$ e $k = 1,...,K$ e onde $\gamma_1,...,\gamma_K,\mu_{\beta_{jk}},\sigma^2_{\beta_{jk}},a_k$ e b_k são hiperparâmetros conhecidos. A parametrização utilizada da distribuição Gama foi tal que, para $Z \sim Gama(a,b)$, $E(Z) = a/b$. As distribuições *a priori* especificadas dessa maneira, como será visto posteriormente na Seção 3.2.1, nos permitem utilizar o algoritmo *Gibbs Sampling* para a estimação do modelo. Porém, segundo Hurn, Justel e Robert (2003), outras escolhas de distribuições *a priori* são igualmente válidas, uma vez que podemos utilizar o algoritmo *Metropolis-Hastings*

para simular de distribuições não conhecidas. Hurn, Justel e Robert (2003) também reforça que distribuições *a priori* impróprias, como a de Jeffreys, nao são aceitáveis nesse modelo porque elas conduzem a distribuições *a posteriori* impróprias, independentemente do tamanho da amostra.

Considerando que o número de componentes, K, é conhecido, as distribuições *a posteriori* condicionais completas dos parâmetros são conhecidas e dadas por

$$\boldsymbol{w}|\cdots \sim Dirichlet(\gamma_1 + n_1, ..., \gamma_K + n_K), \tag{3.8}$$

$$\beta_{jk}|\cdots \sim N\left(\left(\frac{m_{ij}}{\sigma_k^2} + \frac{\mu_{\beta_{jk}}}{\sigma_{\beta_{jk}}^2}\right)\left(\frac{1}{\sigma_{\beta_{jk}}^2} + \frac{v_{ij}}{\sigma_k^2}\right)^{-1}, \left(\frac{1}{\sigma_{\beta_{jk}}^2} + \frac{v_{ij}}{\sigma_k^2}\right)^{-1}\right), \tag{3.9}$$

para $j = 0, ..., p$ e $k = 1, ..., K$ e

$$\frac{1}{\sigma_k^2}|\cdots \sim Gama\left(a_k + \frac{n_k}{2}, b_k + \frac{\sum_{i:s_i=k}(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_k)^2}{2}\right), \tag{3.10}$$

para $k = 1, ..., K$, onde $\cdots$ contempla os dados e todos os parâmetros, exceto o parâmetro em questão, $m_{ij} = \sum_{i:s_i=k} x_{ij}(y_i - \sum_{l\neq j} x_{il}\beta_{lk})$ e $v_{ij} = \sum_{i:s_i=k} x_{ij}^2$. Os cálculos detalhados são apresentados no Apêndice B.

A probabilidade *a posteriori* de $S_i = k$ é dada por

$$\begin{aligned} P(S_i = k|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) &= \frac{f(S_i = k, y_i|\boldsymbol{x}_i, \boldsymbol{\theta})}{f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})} \\ &= \frac{f(S_i = k, y_i|\boldsymbol{x}_i, \boldsymbol{\theta})}{\sum_{t=1}^K f(S_i = t, y_i|\boldsymbol{x}_i, \boldsymbol{\theta})} \\ &= \frac{P(S_i = k|\boldsymbol{w}) f(y_i|S_i = k, \boldsymbol{x}_i, \boldsymbol{\theta})}{\sum_{t=1}^K P(S_i = t|\boldsymbol{w}) f(y_i|S_i = t, \boldsymbol{x}_i, \boldsymbol{\theta})} \\ &= \frac{P(S_i = k|\boldsymbol{w}) f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k)}{\sum_{t=1}^K P(S_i = t|\boldsymbol{w}) f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_t)} \\ &= \frac{w_k f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k)}{\sum_{t=1}^K w_t f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_t)} \end{aligned} \tag{3.11}$$

para $k = 1, ..., K$, onde $f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k)$ é a função densidade da $N(\boldsymbol{x}_i\boldsymbol{\beta}_k, \sigma_k^2)$ calculada no ponto y_i.

Como as distribuições *a posteriori* condicionais completas são distribuições padrões conhecidas, implementamos o algoritmo *Gibbs Sampling* para simular valores dessas distribuições e gerar uma amostra da distribuição *a posteriori* conjunta. Assumimos como estimativas pontuais dos parâmetros a média *a posteriori* da amostra MCMC obtida para cada parâmetro já considerando o salto e o *burn-in* da amostra MCMC gerada.

3.2.1 Algoritmo de Estimação Gibbs Sampling

Considere o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_K^T, \sigma_1^2, ..., \sigma_K^2)$ com distribuições condicionais dadas por (3.8), (3.9) e (3.10). O algoritmo *Gibbs Sampling* é baseado em sucessivas simulações dos parâmetros contidos em $\boldsymbol{\theta}$ e da variável não-observável $\boldsymbol{S}$ através de suas distribuições *a posteriori* condicionais. Podemos resumir o algoritmo pelos seguintes passos:

Passo 1: para $i = 1, ..., n$, inicialize aleatoriamente S_i com valores de 1 até K, considerando inicialmente que $P(S_i = k) = 1/K$ para $k = 1, ..., K$;

Passo 2: Inicialize arbitrariamente $\beta_{jk}^{(0)}$ para $j = 0, ..., p$ e $k = 1, ..., K$;

Passo 3: para $k = 1, ..., K$, gere $1/(\sigma_k^2)^{(0)}$ de

$$Gama\left(a_k + \frac{n_k}{2}, b_k + \frac{\sum_{i:s_i=k}(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_k^{(0)})^2}{2}\right)$$

onde $n_k = \sum_{i=1}^n I_{s_i}(k)$ e a_k e b_k são hiperparâmetros escolhidos anteriormente;

Passo 4: para ℓ-ésima iteração, $\ell = 1, ..., L$, onde L é o número total de iterações, faça:

- gere $\beta_{jk}^{(\ell)}$ de

$$N\left(\left(\frac{m_{ij}^{(\ell-1)}}{(\sigma_k^2)^{(\ell-1)}} + \frac{\mu_{\beta_{jk}}}{\sigma_{\beta_{jk}}^2}\right)\left(\frac{1}{\sigma_{\beta_{jk}}^2} + \frac{v_{ij}^{(\ell-1)}}{(\sigma_k^2)^{(\ell-1)}}\right)^{-1}, \left(\frac{1}{\sigma_{\beta_{jk}}^2} + \frac{v_{ij}^{(\ell-1)}}{(\sigma_k^2)^{(\ell-1)}}\right)^{-1}\right)$$

  para $j = 0, ..., p$ e $k = 1, ..., K$;

- gere $1/(\sigma_k^2)^{(\ell)}$ de

$$Gama\left(a_k + \frac{n_k}{2}, b_k + \frac{\sum_{i:s_i=k}(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_k^{(\ell)})^2}{2}\right)$$

  para $k = 1, ..., K$;

- gere $\boldsymbol{w}^{(\ell)} \sim Dirichlet(\gamma_1 + n_1, ..., \gamma_K + n_K)$;

- para $i = 1, ..., n$ atualizar o valor de S_i considerando a probabilidade de ocorrência (3.11), isto é,

$$P(S_i = k|\boldsymbol{\theta}^{(\ell)}, \boldsymbol{y}, \boldsymbol{x}_i) = \frac{P(S_i = k|\boldsymbol{w}^{(\ell)})f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(\ell)})}{\sum_{t=1}^K P(S_i = t|\boldsymbol{w}^{(\ell)})f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_t^{(\ell)})}$$

  para $k = 1, ..., K$, onde $f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(\ell)})$ é a função densidade de probabilidade da $N(\boldsymbol{x}_i\boldsymbol{\beta}_k^{(\ell)}, (\sigma_k^2)^{(\ell)})$ calculada no ponto y_i e $P(S_i = k|\boldsymbol{w}^{(\ell)}) = w_k^{(\ell)}$.

Ao final das L iterações, descartamos as B primeiras iterações como *burn-in* e guardamos uma iteração a cada "*saltos*" iterações, obtendo uma cadeia final de tamanho $L_{final} = [(L-B)/saltos]$. Como habitualmente é feito na análise Bayesiana, consideramos como estimativas pontuais dos parâmetros a média *a posteriori* da cadeia MCMC resultante, isto é,

$$\hat{\beta}_{jk} = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} \beta_{jk}^{(\ell)} \tag{3.12}$$

para $j = 0, ..., p$ e $k = 1, ..., K$,

$$\hat{\sigma}_k^2 = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} (\sigma_k^2)^{(\ell)} \tag{3.13}$$

e

$$\hat{w}_k = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} w_k^{(\ell)} \tag{3.14}$$

para $k = 1, ..., K$.

Considerando N_{ik} o número de vezes em que a observação $(y_i, \boldsymbol{x}_i)$ foi associada a componente k nas L_{final} iterações, então, assumimos $P(S_i = k|\cdots) = \frac{N_{ik}}{L_{final}}$ como a probabilidade *a posteriori* de $(y_i, \boldsymbol{x}_i)$ ser proveniente da componente k. Fazemos uso desta probabilidade *a posteriori* para determinar o valor predito de S_i de forma aleatória, para $i = 1, ..., n$.

3.2.2 Seleção do Número de Componentes

Para o ajuste do modelo utilizando o algoritmo *Gibbs Sampling* assumimos que K é conhecido, isto é, consideramos que sabemos o número de componentes da mistura.

Para dados experimentais, o número verdadeiro de grupos pode ser desconhecido, neste caso buscamos uma boa escolha do valor de K que resulte em resultados estáveis e nos forneça um bom ajuste aos dados. Assim, podemos ajustar o modelo para diferentes valores de K e utilizar técnicas de seleção de modelos para definir o K que apresentou o melhor ajuste.

Existem diversos critérios de informação para seleção de modelos na metodologia Bayesiana que são utilizados quando as amostras da distribuição *a posteriori* para os parâmetros do modelo são obtidas por meio de métodos MCMC. Este trabalho propõe a utilização do *Deviance Information Criterion (DIC)* e do *Expected Bayesian Information Criterion (EBIC)* para a estimação do número de componentes na mistura, K.

O *DIC*, proposto por Spiegelhalter *et al.* (2014), é definido como

$$DIC = 2\bar{D}(\boldsymbol{\theta}^{\ell}) - D(\bar{\boldsymbol{\theta}}), \tag{3.15}$$

onde $\bar{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X})$ é a média *a posteriori* dos parâmetros, $\boldsymbol{\theta}^{\ell}$ é o valor de $\boldsymbol{\theta}$ na ℓ-ésima interação da cadeia MCMC desconsiderando *burn-in* e os *saltos*,

$$D(\boldsymbol{\theta}) = -2\log L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x}) - 2\log(h(\boldsymbol{y})), \tag{3.16}$$

calculado para cada interação MCMC, é denominado *deviance* e $\bar{D}(\boldsymbol{\theta}^{\ell})$ é a média de $D(\boldsymbol{\theta}^{\ell})$, $\ell = 1,...,L_{final}$.

O *EBIC*, proposto por Chen e Chen (2008), é definido como

$$EBIC = \bar{D}(\boldsymbol{\theta}^{\ell}) + |\boldsymbol{\theta}|\log(n), \tag{3.17}$$

onde n é o número de observações e $|\boldsymbol{\theta}|$ é o número de parâmetros em $\boldsymbol{\theta}$ a serem estimados.

Podemos simplificar o cálculo da *deviance*, dada por (3.16), assumindo que $h(\boldsymbol{y}) = 1$. Dessa maneira, a *deviance* pode ser reescrita como $D(\boldsymbol{\theta}) = -2\log L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x})$, mas

$$\begin{aligned}
\log L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x}) &= \log\prod_{k=1}^{K}\left[w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right] \\
&= \sum_{k=1}^{K}\log\left[w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right] \\
&= \sum_{k=1}^{K}\left\{n_k\log w_k + \frac{n_k}{2}\log\frac{1}{2\pi\sigma_k^2} + \log\left[\exp\left[-\frac{1}{2\sigma_k^2}\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right]\right]\right\} \\
&= \sum_{k=1}^{K}\left\{n_k\log w_k - \frac{n_k}{2}\log 2\pi\sigma_k^2 - \frac{1}{2\sigma_k^2}\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\},
\end{aligned}$$

então,

$$\begin{aligned}
D(\boldsymbol{\theta}) &= -2\left(\sum_{k=1}^{K}\left\{n_k\log w_k - \frac{n_k}{2}\log 2\pi\sigma_k^2 - \frac{1}{2\sigma_k^2}\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\}\right) \\
&= \sum_{k=1}^{K}\left\{n_k\log 2\pi\sigma_k^2 - 2n_k\log w_k + \frac{1}{\sigma_k^2}\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k)^2\right\} \\
&= \sum_{k=1}^{K}\left\{n_k\log 2\pi\sigma_k^2 - 2n_k\log w_k + \frac{1}{\sigma_k^2}\sum_{i:s_i=k}\varepsilon_i^2\right\},
\end{aligned}$$

onde ε_i é o erro da observação i. Consideramos como preditor de ε_i o resíduo da observação i, $\hat{\varepsilon}_i = e_i = (y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_k)$.

O K estimado será o valor de K associado ao modelo que melhor explica os dados observados, isto é, associado ao modelo que apresentar o menor valor *DIC* e *EBIC*.

3.3 Estimação Bayesiana com K desconhecido

Como mencionado na seção anterior, supor que o número de componentes da mistura é conhecido muitas vezes não é plausível na prática. Dessa forma, como contribuição à literatura, propomos um algoritmo MCMC guiado pelos dados denominado *Data Driven Reversible Jump* - DDRJ que, conjuntamente com a estimação do modelo, seleciona o número de componentes da mistura. O DDRJ é composto por três etapas: no primeiro passo, os valores atuais dos parâmetros são atualizados via algoritmo *Gibbs Sampling* desenvolvido na seção anterior; na segunda etapa do algoritmo, o número de componentes é virtualmente atualizado através de movimentos de *split-merge*; e, por fim, na terceira etapa avaliamos a transição de um modelo para o outro.

Considere que o número de componentes, K, é desconhecido e possui distribuição Discreta com K_{max} possíveis valores. Vamos supor que os parâmetros são independentes *a priori* e assumiremos as distribuições conjugadas dadas por (3.5), (3.6) e (3.7) para $k = 1, ..., K_{max}$, onde K_{max} é o valor máximo para K. Da mesma forma, para cada K neste intervalo, temos que as distribuições *a posteriori* condicionais completas dos parâmetros são conhecidas e dadas por (3.8), (3.9) e (3.10). A probabilidade *a posteriori* de $S_i = k$ é dada por (3.11).

Para atualizar o valor de K realizamos movimentos entre modelos através de procedimentos de *split* e *merge*. O *split* divide uma componente arbitrária, s_{t^*}, em duas componentes, s_{t_1} e s_{t_2}, incrementando o valor de K em uma unidade. Já o *merge* junta duas componentes arbitrárias, s_{t_1} e s_{t_2}, em uma única componente, s_{t^*}, diminuindo o valor de K em uma unidade.

Considere $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_K^T, \sigma_1^2, ..., \sigma_K^2)$ o estado atual dos parâmetros de um modelo com K componentes e $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{\beta}_1^{*T}, ..., \boldsymbol{\beta}_{K^*}^{*T}, \sigma_1^{*2}, ..., \sigma_{K^*}^{*2})$ o estado do movimento proposto com K^* componentes, onde "*" representa tanto um *split* como um *merge*. O movimento proposto é aceito de acordo com a probabilidade de aceitação $\psi(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = min(1, A)$, onde

$$A = \frac{L(\boldsymbol{\theta}^*|\boldsymbol{y}, \boldsymbol{s}^*, \boldsymbol{x})}{L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{x})} \frac{\pi(\boldsymbol{w}^*, \boldsymbol{\beta}_1^{*T}, ..., \boldsymbol{\beta}_{K^*}^{*T}, \sigma_1^{*2}, ..., \sigma_{K^*}^{*2})}{\pi(\boldsymbol{w}, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_K^T, \sigma_1^2, ..., \sigma_K^2)} \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta})} \tag{3.18}$$

e $q(\cdot|\cdot)$ é a distribuição da qual construímos o modelo candidato com K^* componentes.

A escolha entre os movimentos de *split* e *merge* é dada de acordo com as seguintes probabilidades:

$$\begin{cases} \text{Se } K = 1, \text{ então } P(split|K) = 1 \text{ e } P(merge|K) = 0 \\ \text{Se } K = K_{max}, \text{ então } P(split|K) = 0 \text{ e } P(merge|K) = 1 \\ \text{Caso contrário, } P(split|K) = P(merge|K) = 0.5 \end{cases}$$

Dado que um movimento de *merge* foi escolhido, assumimos que todas as componentes possuem a mesma probabilidade de se juntarem, isto é, as componentes s_{t_1} e s_{t_2} serão unidas

com probabilidade $p_{s_{t_1}s_{t_2}} = \frac{1}{\binom{K}{2}} = \frac{2}{K(K-1)}$. Por outro lado, dado que um movimento de *split* foi definido, a componente s_{t^*} é escolhida com probabilidade $p_{s_{t^*}} = \frac{1}{K}$.

3.3.1 Split

No *split* da componente s_{t^*}, distribuímos suas observações em duas novas componentes, $s_{t_1} = s_{t^*}$ e $s_{t_2} = K+1$, de forma arbitrária e, consecutivamente, estimamos os parâmetros para estas novas duas componentes. Considere que $\boldsymbol{s} = (s_1, ..., s_n)$ seja a configuração de $\boldsymbol{S}$ antes do *split* e $\boldsymbol{s}^{sp} = (s_1^{sp}, ..., s_n^{sp})$ seja sua configuração após o *split*. O passo de *split* pode ser descrito como segue:

1. **Realocação das observações:** Para $i = 1, ..., n$,

   - se $s_i \neq s_{t^*}$, então mantenha $(Y_i, \boldsymbol{X}_i)$ na mesma componente, $s_i^{sp} = s_i$; e
   - se $s_i = s_{t^*}$, então aloque $(Y_i, \boldsymbol{X}_i)$ na componente $K+1$, $s_i^{sp} = K+1 = s_{t_2}$, se $e_i \leq e^*$, ou mantenha na componente s_{t^*}, $s_i^{sp} = s_{t^*} = s_{t_1}$, se $e_i > e^*$, onde e^* é o valor do resíduo que separa as observações em dois grupos, um contendo 15% das observações cujos resíduos são os menores dentro da componente a ser quebrada e outro com as 85% restantes.

2. **Simulação dos parâmetros das novas componentes:** Condicionado a $\boldsymbol{S} = \boldsymbol{s}^{sp}$, simulamos valores candidatos para $\boldsymbol{\theta}^{sp} = (\boldsymbol{w}^{sp}, \boldsymbol{\beta}_1^{sp^T}, ..., \boldsymbol{\beta}_{K+1}^{sp^T}, \sigma_1^{sp2}, ..., \sigma_{K+1}^{sp2})$ a partir de suas distribuições *a posteriori* dadas por (3.8), (3.9) e (3.10).

A distribuição proposta do *split* é dada por

$$q(\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}) = P(split|K)\, p_{s_{t^*}}\, \pi(\boldsymbol{w}^{sp}|\cdots)\pi(\boldsymbol{\beta}_{s_{t_1}}^{sp}|\cdots)\pi(\boldsymbol{\beta}_{s_{t_2}}^{sp}|\cdots)\pi(\sigma_{s_{t_1}}^{sp^2}|\cdots)\pi(\sigma_{s_{t_2}}^{sp^2}|\cdots), \quad (3.19)$$

onde $\pi(\cdot|\cdot)$ é a distribuição condicional *a posteriori* usada para simular os valores candidatos.

A probabilidade de aceitação do movimento de *split* é dada por $\psi(\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}) = min(1, A^{sp})$, onde A^{sp} é dado pela equação (3.18) como o produto da razão das funções de verossimilhança

$$\begin{aligned}\frac{L(\boldsymbol{\theta}^{sp}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{s}^{sp})}{L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{s})} &= \frac{(2\pi\sigma^2_{s_{t^*}})^{\frac{n_{s_{t^*}}}{2}}}{(2\pi\sigma^{sp^2}_{s_{t_1}})^{\frac{n_{s_{t_1}}}{2}}(2\pi\sigma^{sp^2}_{s_{t_2}})^{\frac{n_{s_{t_2}}}{2}}} \\ &\times \frac{exp\left\{-\frac{1}{2\sigma^{sp^2}_{s_{t_1}}}\sum_{i:s_i^{sp}=s_{t_1}}\left(y_i-\boldsymbol{x}_i\boldsymbol{\beta}^{sp}_{s_{t_1}}\right)^2\right\} exp\left\{-\frac{1}{2\sigma^{sp^2}_{s_{t_2}}}\sum_{i:s_i^{sp}=s_{t_2}}\left(y_i-\boldsymbol{x}_i\boldsymbol{\beta}^{sp}_{s_{t_2}}\right)^2\right\}}{exp\left\{-\frac{1}{2\sigma^2_{s_{t^*}}}\sum_{i:s_i=s_{t^*}}\left(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_{s_{t^*}}\right)^2\right\}} \\ &\times \frac{\prod_{k=1}^{K+1} w_k^{sp^{n_k^{sp}}}}{\prod_{k=1}^{K} w_k^{n_k}},\end{aligned}$$

pela razão das densidades *a priori*

$$\begin{aligned}\frac{\pi(\boldsymbol{\theta}^{sp})}{\pi(\boldsymbol{\theta})} &= \frac{\pi(\boldsymbol{w}^{sp})\pi(\boldsymbol{\beta}^{sp}_1)...\pi(\boldsymbol{\beta}^{sp}_{K+1})\pi(\sigma_1^{sp^2})...\pi(\sigma^{sp^2}_{K+1})}{\pi(\boldsymbol{w})\pi(\boldsymbol{\beta}_1)...\pi(\boldsymbol{\beta}_K)\pi(\sigma_1^2)...\pi(\sigma^2_K)} \\ &= \frac{\pi(\boldsymbol{w}^{sp})\pi(\boldsymbol{\beta}^{sp}_{s_{t_1}})\pi(\boldsymbol{\beta}^{sp}_{s_{t_2}})\pi(\sigma^{sp^2}_{s_{t_1}})\pi(\sigma^{sp^2}_{s_{t_2}})}{\pi(\boldsymbol{w})\pi(\boldsymbol{\beta}_{s_{t^*}})\pi(\sigma^2_{s_{t^*}})}\end{aligned}$$

e pela razão das distribuições propostas

$$\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}^{sp})}{q(\boldsymbol{\theta}^{sp}|\boldsymbol{\theta})} = \frac{P(merge|K+1)\frac{1}{\binom{K+1}{2}}\pi(\boldsymbol{w}|\cdots)\pi(\boldsymbol{\beta}_{s_{t^*}}|\cdots)\pi(\sigma^2_{s_{t^*}}|\cdots)}{P(split|K)p_{s_{t^*}}\pi(\boldsymbol{w}^{sp}|\cdots)\pi(\boldsymbol{\beta}^{sp}_{s_{t_1}}|\cdots)\pi(\boldsymbol{\beta}^{sp}_{s_{t_2}}|\cdots)\pi(\sigma^{sp^2}_{s_{t_1}}|\cdots)\pi(\sigma^{sp^2}_{s_{t_2}}|\cdots)}.$$

3.3.2 Merge

No movimento de *merge*, juntamos as observações de duas componentes selecionadas, s_{t_1} e s_{t_2}, em uma única componente, s_{t^*}, e estimamos os parâmetros para esta nova componente. Considere que $\boldsymbol{s} = (s_1,...,s_n)$ seja a configuração de $\boldsymbol{S}$ antes do *merge* e que $\boldsymbol{s}^{mg} = (s_1^{mg},...,s_n^{mg})$ seja sua configuração após o *merge*. O *merge* das componentes s_{t_1} e s_{t_2} é implementado de acordo com o seguinte procedimento:

1. **Realocação das observações:** para $i = 1,...,n$

   - se $s_i < max(s_{t_1}, s_{t_2})$, faça $s_i^{mg} = s_i$, mantendo $(Y_i, \boldsymbol{X}_i)$ na mesma componente;
   - se $s_i = max(s_{t_1}, s_{t_2})$, faça $s_i^{mg} = min(s_{t_1}, s_{t_2})$, alocando os elementos $(Y_i, \boldsymbol{X}_i)$ de s_{t_1} e s_{t_2} em uma única componente; e
   - se $s_i > max(s_{t_1}, s_{t_2})$, faça $s_i^{mg} = s_i - 1$.

2. **Simulando os parâmetros da nova componente:** Dado $\boldsymbol{S} = \boldsymbol{s}^{mg}$, simulamos os valores candidatos para o novo conjunto de parâmetros $\boldsymbol{\theta}^{mg} = (\boldsymbol{w}^{mg}, \boldsymbol{\beta}^{mg}, \boldsymbol{\sigma}^{mg^2})$ a partir de suas distribuições *a posteriori* dadas por (3.8), (3.9) e (3.10). Movendo-se de $\boldsymbol{\theta}$ para $\boldsymbol{\theta}^{mg}$ só precisamos atualizar $\boldsymbol{\theta}^{mg}_{s_{t^*}}$, os parâmetros da componente cuja configuração foi alterada.

A distribuição proposta do *merge* é dada por

$$q(\boldsymbol{\theta}^{mg}|\boldsymbol{\theta}) = P(merge|K)\frac{1}{\binom{K}{2}}\pi(\boldsymbol{w}^{mg}|\cdots)\pi(\boldsymbol{\beta}^{mg}_{s_{t^*}}|\cdots)\pi(\sigma^{mg}_{s_{t^*}}). \tag{3.20}$$

A probabilidade de aceitação do movimento de *merge* é dada por $\psi(\boldsymbol{\theta}^{sp}|\boldsymbol{\theta}) = min(1, A^{mg})$, onde $A^{mg} = 1/A^{sp}$.

O algoritmo proposto é um caso especial do *reversible jump* quando os parâmetros do modelo proposto são extraídos da distribuição proposta e o Jacobiano é igual a 1.

3.3.3 Algoritmo de Estimação DDRJ

Considere nosso vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_K^T, \sigma_1^2, ..., \sigma_K^2)$ com distribuições condicionais dadas por (3.8), (3.9) e (3.10). O algoritmo DDRJ é especificado pelos seguintes passos:

Passo 1: inicialize K arbitrariamente com $K^{(0)}$;

Passo 2: para $i = 1, ..., n$, inicialize aleatoriamente S_i com valores de 1 até $K^{(0)}$, considerando inicialmente que $P(S_i = k) = 1/K^{(0)}$ para $k = 1, ..., K^{(0)}$;

Passo 3: para $k = 1, ..., K^{(0)}$, gere $1/(\sigma_k^2)^{(0)}$ de

$$Gama\left(a_k + \frac{n_k}{2}, b_k + \frac{\sum_{i:s_i=k}(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_k^{(0)})^2}{2}\right)$$

onde $n_k = \sum_{i=1}^n I_{s_i}(k)$ e a_k e b_k são hiperparâmetros escolhidos anteriormente;

Para ℓ-ésima iteração, $\ell = 1, ..., L$, onde L é o número total de iterações, faça:

Passo 4: atualize os parâmetros via *Gibbs Sampling*:

- gere $\beta_{jk}^{(\ell)}$ de

$$N\left(\left(\frac{m_{ij}^{(\ell-1)}}{(\sigma_k^2)^{(\ell-1)}} + \frac{\mu_{\beta_{jk}}}{\sigma^2_{\beta_{jk}}}\right)\left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{v_{ij}^{(\ell-1)}}{(\sigma_k^2)^{(\ell-1)}}\right)^{-1}, \left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{v_{ij}^{(\ell-1)}}{(\sigma_k^2)^{(\ell-1)}}\right)^{-1}\right)$$

para $j = 0, ..., p$ e $k = 1, ..., K^{(\ell-1)}$;

- gere $1/(\sigma_k^2)^{(\ell)}$ de

$$Gama\left(a_k+\frac{n_k}{2}, b_k+\frac{\sum_{i:s_i=k}(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_k^{(\ell)})^2}{2}\right)$$

para $k=1,...,K^{(\ell-1)}$;

- gere $\boldsymbol{w}^{(\ell)} \sim Dirichlet(\gamma_1+n_1,...,\gamma_{K^{(\ell-1)}}+n_{K^{(\ell-1)}})$;
- para $i=1,...,n$ atualize o valor de S_i considerando a probabilidade de ocorrência (3.11), isto é,

$$P(S_i=k|\boldsymbol{\theta}^{(\ell)},\boldsymbol{y},\boldsymbol{x}_i)=\frac{P(S_i=k|\boldsymbol{w}^{(\ell)})f(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(\ell)})}{\sum_{t=1}^{K}P(S_i=t|\boldsymbol{w}^{(\ell)})f(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}_t^{(\ell)})},$$

para $k=1,...,K^{(\ell-1)}$, onde $f(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}_k^{(\ell)})$ é a função densidade da $N(\boldsymbol{x}_i\boldsymbol{\beta}_k^{(\ell)},(\sigma_k^2)^{(\ell)})$ calculada no ponto y_i e $P(S_i=k|\boldsymbol{w}^{(\ell)})=w_k^{(\ell)}$.

Passo 5: atualize o valor de K através do processo *split-merge* guiado pelos dados:

(A) Decida entre *split* e *merge* de acordo com as probabilidades definidas anteriormente:

$$\begin{cases} \text{Se } K^{(\ell-1)}=1, \text{ então } P(split|K^{(\ell-1)})=1 \text{ e } P(merge|K^{(\ell-1)})=0 \\ \text{Se } K^{(\ell-1)}=K_{max}, \text{ então } P(split|K^{(\ell-1)})=0 \text{ e } P(merge|K^{(\ell-1)})=1 \\ \text{Caso contrário, } P(split|K^{(\ell-1)})=P(merge|K^{(\ell-1)})=0.5 \end{cases}$$

onde K_{max} é o número máximo para K definido de forma arbitrária;

(B) Construa o vetor $\boldsymbol{S}$ candidato, $\boldsymbol{s}^*$, e amostre valores candidatos para $\boldsymbol{\theta}^*$ através de suas distribuições *a posteriori*;

(C) Aceite o movimento proposto com probabilidade $\psi(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, onde $*$ representa tanto *split* quanto *merge*:

- se o *split* é aceito, faça $K^{(\ell)}=K^{(\ell-1)}+1$ e considere $\boldsymbol{\theta}^{sp}$;
- se o *merge* é aceito, faça $K^{(\ell)}=K^{(\ell-1)}-1$ e considere $\boldsymbol{\theta}^{mg}$;
- se nenhum movimento é aceito, faça $K^{(\ell)}=K^{(\ell-1)}$ e considere $\boldsymbol{\theta}$.

Ao final das L iterações, descartamos as B primeiras iterações como *burn-in* e guardamos uma iteração a cada "*saltos*" iterações obtendo uma cadeia intermediária de tamanho $L_1=[(L-B)/saltos]$. Consideramos como probabilidade *a posteriori* de K os valores $P(K=j)=\frac{N_{K=j}}{L_1}$, onde $N_{K=j}$ é o número de vezes que K assumiu o valor j, para $j=1,...,K_{max}$, nas L_1 iterações e a estimativa para K é dada pelo valor com maior probabilidade *a posteriori*. Após isso, descartamos as T iterações onde este valor estimado para K não foi obtido, obtendo uma cadeia final de tamanho $L_{final}=L_1-T$. Como habitualmente é feito na análise Bayesiana,

consideramos como estimativas pontuais dos parâmetros a média *a posteriori* da cadeia MCMC, isto é,

$$\hat{\beta}_{jk} = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} \beta_{jk}^{(\ell)} \tag{3.21}$$

para $j = 0, ..., p$ e $k = 1, ..., K$,

$$\hat{\sigma}_k^2 = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} (\sigma_k^2)^{(\ell)} \tag{3.22}$$

e

$$\hat{w}_k = \frac{1}{L_{final}} \sum_{\ell=1}^{L_{final}} w_k^{(\ell)} \tag{3.23}$$

para $k = 1, ..., K$.

Considerando N_{ik} o número de vezes em que a observação $(y_i, \boldsymbol{x}_i)$ foi associada a componente k nas L_{final} iterações, então, assumimos $P(S_i = k|\cdots) = \frac{N_{ik}}{L_{final}}$ como a probabilidade *a posteriori* de $(y_i, \boldsymbol{x}_i)$ ser proveniente da componente k. Fazemos uso desta probabilidade *a posteriori* para determinar o valor predito de S_i de forma aleatória, para $i = 1, ..., n$.

3.4 O Problema do *Label Switching*

Como já discutido brevemente no capítulo anterior, quando se adota a abordagem Bayesiana para a estimação dos parâmetros de um modelo de mistura, um problema comum é a troca de rótulos dos componentes durante o processo de estimação, conhecido pelo termo *label switching*.

A raiz do problema do *label switching* está no fato da função de verossimilhança ser invariante a qualquer permutação de $\boldsymbol{\theta} = (w_1, ..., w_k, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_K^T, \sigma_1^2, ..., \sigma_K^2)$. De fato, considere qualquer permutação ρ de $1, ..., K$ e defina a correspondente permutação do vetor de parâmetros $\boldsymbol{\theta}$ por

$$\begin{aligned} \rho(\boldsymbol{\theta}) &= \rho(w_1, ..., w_K, \boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_K^T, \sigma_1^2, ..., \sigma_K^2) \\ &= (w_{\rho(1)}, ..., w_{\rho(K)}, \boldsymbol{\beta}_{\rho(1)}^T, ..., \boldsymbol{\beta}_{\rho(K)}^T, \sigma_{\rho(1)}^2, ..., \sigma_{\rho(K)}^2). \end{aligned} \tag{3.24}$$

Note que a função de verossimilhança (3.4) é a mesma para qualquer que seja (3.24). Essa propriedade faz com que a distribuição *a posteriori* dos parâmetros seja, no geral, similarmente simétrica entre as componentes.

Algumas abordagens tentam solucionar este problema, sendo a mais comum delas a imposição de uma restrição artificial de identificabilidade no espaço paramétrico $\boldsymbol{\theta}$. No caso de modelo de mistura de regressões poderíamos escolher o parâmetro que parece mais se diferenciar

entre as componentes, β_{1k} por exemplo, impondo a restrição $\beta_{11} < \beta_{12} < \cdots < \beta_{1K}$ e rearranjar as componentes de cada iteração MCMC tal que essa restrição seja respeitada. No entanto, Stephens (2000) demonstra que, apesar de muito popular e simples, este método falha em geral, especialmente para modelos de mistura de regressões, e descreve uma série de abordagens alternativas.

Para solucionar o problema de *label switching*, este trabalho faz uso do algoritmo *Equivalence Classes Representatives* (ECR) proposto por Papastamoulis e Iliopoulos (2010), no qual sua superioridade frente aos outros métodos fica aparente em Papastamoulis (2014). A ideia básica do ECR é renomear as componentes de cada iteração MCMC de modo que os valores preditos de $\boldsymbol{S}$ de cada iteração sejam os mais próximos possíveis de uma sequência $\boldsymbol{s}$ da primeira iteração MCMC após $burn-in$ e "$saltos$" e com convergência garantida. O algoritmo ECR está implementado no pacote *label.switching* do software estatístico R (PAPASTAMOULIS, 2015) e que não é exclusivo para o modelo de mistura de regressões normais.

CAPÍTULO

4

APLICAÇÕES EM DADOS SIMULADOS E REAIS

Neste capítulo, aplicamos as metodologias de estimação propostas e o algoritmo EM para modelos de mistura de regressão normal em nove conjuntos de dados: oito conjuntos de dados artificiais gerados via simulação estocástica e um conjunto de dados reais referente ao IDEB de 2011.

4.1 Dados Simulados

Com o objetivo de testar as metodologias propostas em diferentes cenários, propomos a seguir oito conjuntos de dados simulados, assumimos em todos eles $K = 3$ componentes.

Nos seis primeiros, consideramos uma única variável explicativa e a cada simulação vamos dificultando o processo inferencial diminuindo a distância entre as componentes, isto é, deixando as observações simuladas cada vez mais próximas, independentemente de qual componente ela pertence. Na Seção 4.1.1, apresentamos um conjunto de dados mais simples de ser estimado, onde as componentes são totalmente esparsas. Nas Seções 4.1.2, 4.1.3, 4.1.4 e 4.1.5 consideramos os mesmos valores para os coeficientes de regressão, porém vamos aumentando as variâncias das componentes gradativamente. Na Seção 4.1.6, consideramos o cenário onde a única diferença entre as componentes da mistura está na variância, exemplificando um cenário de modelo de regressão heterocedástico.

Nos dois últimos cenários simulados, consideramos duas covariáveis. Na Seção 4.1.7, assumimos componentes esparsas. Já na Seção 4.1.8, consideramos o cenário onde a única diferença entre as componentes da mistura está na variância, também exemplificando um cenário de modelo de regressão heterocedástico.

4.1.1 Conjunto de Dados Simulado 1

Na geração do primeiro conjunto de dados artificiais, consideramos um modelo de mistura de regressão normal com $K = 3$ componentes e $p = 1$ covariável,

$$Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i \sim \sum_{k=1}^{3} w_k N(\boldsymbol{x}_i\boldsymbol{\beta}_k, \sigma_k^2),$$

onde $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k})^T$, para $k = 1, ..., 3$.

Consideramos uma amostra de tamanho $n = 2000$ e fixamos $w_1 = 0.3$, $w_2 = 0.4$, $w_3 = 0.3$, $\sigma_1^2 = 2$, $\sigma_2^2 = 16$, $\sigma_3^2 = 4$, $\boldsymbol{\beta}_1 = (5,3)^T$, $\boldsymbol{\beta}_2 = (2,-5)^T$ e $\boldsymbol{\beta}_3 = (-1,5)^T$. Seja $\mathbf{a} = (a_1, a_2, a_3)$, onde $a_j = \sum_{k=1}^{j} w_k$ para $j = 1, ..., 3$, o vetor acumulado de $\boldsymbol{w}$, o procedimento de simulação utilizado é dado pelos seguintes passos:

Passo 1: para $i = 1, ..., n$, gere $x_{i1} \sim N(100, 1)$ e $u_i \sim U(0, 1)$, onde $U(0, 1)$ é a função densidade da distribuição Uniforme com parâmetros 0 e 1; e

Passo 2: para $i = 1, ..., n$, se $u_i \leq a_1$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_1, \sigma_1^2)$ e faça $S_i = 1$; se $a_1 < u_i \leq a_2$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_2, \sigma_2^2)$ e faça $S_i = 2$; e se $a_2 < u_i \leq a_3$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_3, \sigma_3^2)$ e faça $S_i = 3$, onde $\boldsymbol{x}_i = (1, x_{i1})$ e S_i é nossa variável de alocação, que identifica de qual componente a i-ésima observação é proveniente.

A Figura 5 representa o gráfico de dispersão entre a variável resposta e a variável explicativa: os pontos em rosa representam as observações simuladas oriundas da componente 1, os pontos em verde as observações simuladas oriundas da componente 2 e os pontos em azul as observações simuladas oriundas da componente 3. Como podemos observar na Figura 5, as 3 componentes desse conjunto de dados simulado são bem evidentes e distintas. Na Figura 6, temos o histograma da variável resposta simulada, y, e sua densidade estimada pelo método Kernel.

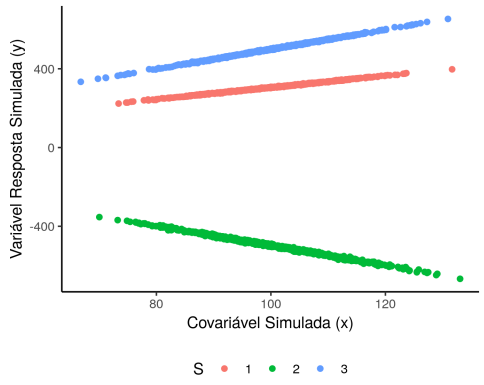


Figura 5 – Simulação 1: gráfico de dispersão.

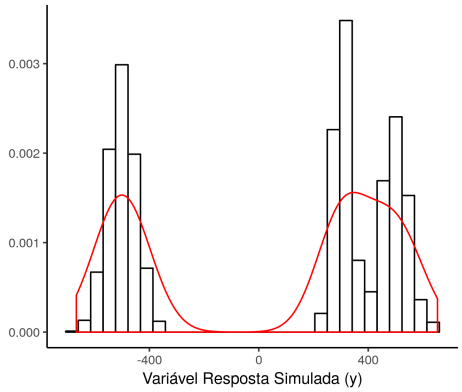


Figura 6 – Simulação 1: histograma da variável resposta simulada.

Aqui vale uma observação que não foi feita anteriormente no texto: é possível que a distribuição da variável resposta seja um modelo de mistura de distribuições, mas um único modelo de regressão seja suficiente para explicar a relação entre a variável resposta e as covariáveis.

4.1.1.1 Estimação via Gibbs Sampling, DIC e EBIC

Cinco modelos com diferentes números de componentes, $K = 1, \ldots, 5$, foram estimados pelo algoritmo *Gibbs Sampling* com passos discutidos na Seção 3.2.1. Os valores dos hiperparâmetros das distribuições *a priori* escolhidos foram: $\gamma_1 = ... = \gamma_K = 1$, $\mu_{\beta_{jk}} = 0$, $\sigma^2_{\beta_{jk}} = 100$ e $a_k = b_k = 0.1$, para $j = 0, 1, ..., p$ e $k = 1, ..., K$. Como a variância das distribuições *a priori* especificadas são altas, essas distribuições são vagas e trazem pouca informação sobre os valores *a priori* dos parâmetros.

Para cada modelo rodamos 80000 iterações, descartamos as 30000 primeiras como *burn-in* e registramos uma a cada 5 iterações. As amostras MCMC que apresentaram problema de *label switching* foram renomeadas usando o algoritmo ECR. Os resultados apresentados a seguir são então oriundos de uma amostra MCMC final de tamanho 10000.

Para cada modelo estimado calculamos o *EBIC* e o *DIC* que servirão para a escolha do valor de K mais adequado para os dados analisados.

A convergência das cadeias MCMC foi verificada via gráficos de traço e estatística de diagnóstico de Geweke (COWLES; CARLIN, 1996), cujos valores e gráficos para o modelo escolhido serão apresentação a seguir.

Os valores do *DIC* e do *EBIC* são encontrados na Tabela 1. Note que o modelo com $K = 3$ componentes obteve o menor valor para o *DIC* e para o *EBIC* e, portanto, nesta simulação, obtivemos uma boa predição para o número de componentes da mistura.

Assim, adotando o modelo de mistura de regressão com $K = 3$ componentes, temos na Tabela 2 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais, que são a média *a posteriori* de cada parâmetro na amostra. Também apresentamos nela o intervalo de

Tabela 1 – Simulação 1: critérios de informação - *DIC* e *EBIC*.

K	1	2	3	4	5
DIC	30113.90	21672.34	13509.97	13750.58	13677.71
EBIC	30142.28	21727.09	13590.35	13884.32	13818.51

credibilidade de máxima densidade *a posteriori* (HPD) de 95% e o diagnóstico de Geweke para convergência.

Como podemos observar na Tabela 2, as estimativas pontuais dos parâmetros estão bem próximas dos seus valores reais e, como esperado, os intervalos de 95% contém os valores verdadeiros dos parâmetros. O valor da estatística Geweke pertence a $(-2,2)$ para todos os parâmetros, indicando convergência. Na Figura 7, apresentamos os gráficos ergódicos ou gráficos de traço dos valores gerados, que também evidenciam convergência do algoritmo. Além disso, o algoritmo classificou corretamente nas componentes 100% das observações.

Tabela 2 – Simulação 1: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke	Estimativa EM
β_{01}	5	5.14	$(3.94, 6.26)$	1.04	5.14
β_{11}	3	3.00	$(2.99, 3.01)$	-1.04	2.99
β_{02}	2	3.07	$(0.36, 5.73)$	-0.39	3.12
β_{12}	-5	-5.01	$(-5.04, -4.99)$	0.39	-5.01
β_{03}	-1	-1.25	$(-2.71, 0.31)$	-0.01	-1.27
β_{13}	5	5.00	$(4.99, 5.02)$	-0.03	5.00
w_1	0.3	0.30	$(0.28, 0.32)$	0.75	0.30
w_2	0.4	0.40	$(0.37, 0.42)$	-0.44	0.40
w_3	0.3	0.30	$(0.28, 0.32)$	-0.38	0.30
σ_1^2	2	2.03	$(1.80, 2.26)$	0.10	2.01
σ_2^2	16	16.56	$(14.94, 18.22)$	-0.47	16.49
σ_3^2	4	3.87	$(3.45, 4.31)$	0.46	3.84

4.1.1.2 Estimação via EM

Da mesma forma que na estimação via *Gibbs Sampling*, cinco modelos com diferentes números de componentes, $K = 1, ..., 5$, foram ajustados, mas, agora, via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools*, (BENAGLIA *et al.*, 2009), implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos pelo pacote.

Semelhante ao processo de estimação do K via *EBIC* e *DIC*, o pacote fornece, para cada modelo estimado, os valores do *Akaike Information Criterion (AIC)* e do *Bayesian Information Criterion (BIC)* para a escolha do valor de K mais adequado. Detalhes sobre o *AIC* e o *BIC* podem ser encontrado em Akaike (1974) e Akaike (1998), respectivamente. Os valores do *AIC* e do *BIC* são encontrados na Tabela 3. Note que o modelo com $K = 5$ obteve o menor valor para o

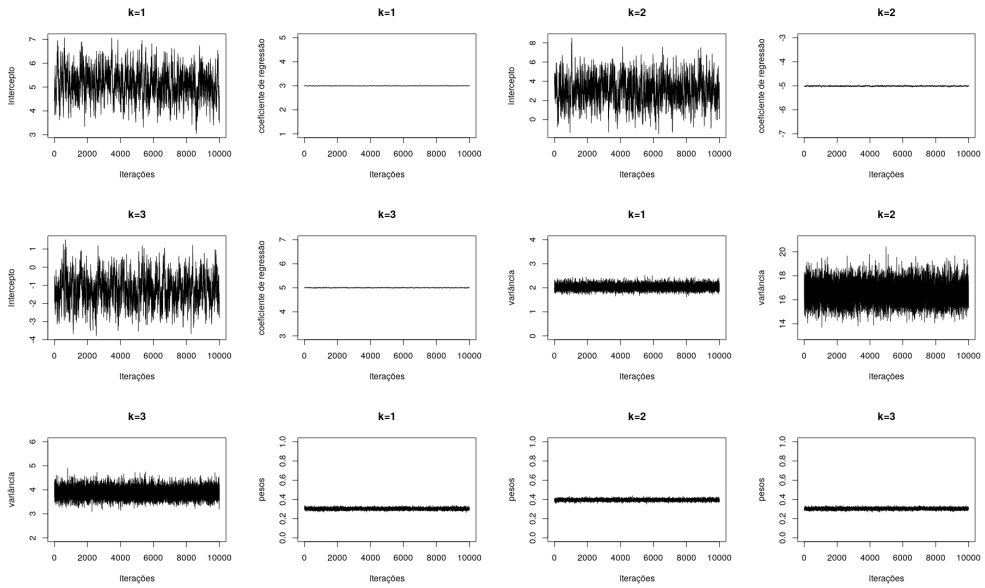


Figura 7 – Simulação 1: gráficos de traço - *Gibbs Sampling*.

AIC e o modelo com $K = 3$ obteve o menor valor para o *BIC*. Apesar dos critérios de informação divergirem, o modelo selecionado pelo *AIC* apresentou duas componentes com pesos estimados zero, o que torna o modelo com o mesmo número de componentes que o escolhido pelo *BIC*.

Tabela 3 – Simulação 1: critérios de informação - *AIC* e *BIC*.

K	1	2	3	4	5
AIC	inf	12826.15	6754.13	6756.90	6751.72
BIC	inf	12842.95	6782.13	6796.11	6802.12

Assim, adotando o modelo de mistura de regressão com $K = 3$ componentes, as estimativas pontuais obtidas são apresentadas na última coluna da Tabela 2 e podemos notar que os valores obtidos são semelhantes aos obtidos pelo algoritmo *Gibbs Sampling* desenvolvido neste trabalho. O algoritmo EM também classificou corretamente 100% das observações.

Ainda considerando $K = 3$, também realizamos a estimação para diferentes valores iniciais e para valores iniciais muito distantes dos valores reais dos parâmetros e o algoritmo EM não apresentou boas estimativas, constatando a sensibilidade do algoritmo EM aos valores iniciais.

4.1.1.3 Estimação via DDRJ

Considerando que K é desconhecido, estimamos o modelo via algoritmo DDRJ com passos discutidos na Seção 3.3.3. Desta forma, tentamos estimar o valor de K conjuntamente com os parâmetros do modelo. Os valores dos hiperparâmetros das distribuições *a priori* escolhidos foram os mesmos adotados na estimação via *Gibbs Sampling*, isto é: $\gamma_1 = ... = \gamma_K = 1$, $\mu_{\beta_{jk}} = 0$, $\sigma^2_{\beta_{jk}} = 100$ e $a_k = b_k = 0.1$, para $j = 0, 1, ..., p$ e $k = 1, ..., K_{max}$. Como as variâncias das

distribuições *a priori* especificadas são altas, essas distribuições são vagas e trazem pouca informação sobre os valores *a priori* dos parâmetros.

Rodamos 80000 iterações, descartamos as 30000 primeiras como *burn-in* e registramos uma a cada 5 iterações. As amostras MCMC que apresentaram problema de *label switching* foram renomeadas usando o algoritmo ECR. Portanto, os resultados apresentados a seguir são oriundos de uma amostra MCMC final de tamanho 10000.

Consideramos como probabilidade *a posteriori* de K os valores $P(K = j) = \frac{N_{K=j}}{L_1}$, onde $N_{K=j}$ é o número de vezes que K assumiu o valor j, para $j = 1, ..., K_{max}$, nas $L_1 = 10000$ iterações e a estimativa para o número de componentes, K, é dada pelo valor com maior probabilidade *a posteriori*. Consideramos para esta simulação $K_{max} = 5$.

A convergência das cadeias MCMC foi verificada via gráficos de traço e estatística de diagnóstico de Geweke, cujos valores e gráficos serão apresentação a seguir.

A Tabela 4 apresenta os valores obtidos para a probabilidade *a posteriori* de K. O máximo valor *a posteriori* é obtido em $K = 5$ e esta deveria ser nossa estimativa para o valor de K. No entanto, duas componentes apresentam pesos estimados zero, tornando o modelo com $K = 3$ componentes.

Tabela 4 – Simulação 1: probabilidade *a posteriori* para K.

K	1	2	3	4	5
$P(K\|\cdots)$	0.0	0.0	0.0	0.0	1.0

Temos na Tabela 5 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais, que são a média *a posteriori* de cada parâmetro da amostra. Também apresentamos nela o intervalo de credibilidade de máxima densidade *a posteriori* (HPD) de 95% e o diagnóstico de Geweke para convergência.

Tabela 5 – Simulação 1: estimativas para os parâmetros - DDRJ.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	5	11.87	$(1.81, 22.55)$	0.42
β_{11}	3	2.93	$(2.83, 3.03)$	-0.42
β_{02}	2	2.88	$(-11.98, 19.45)$	-2.11
β_{12}	-5	-5.01	$(-5.18, -4.87)$	2.12
β_{03}	-1	-0.79	$(-12.70, 9.37)$	0.34
β_{13}	5	5.00	$(4.90, 5.12)$	-0.34
w_1	0.3	0.30	$(0.28, 0.32)$	0.45
w_2	0.4	0.40	$(0.37, 0.42)$	-0.29
w_3	0.3	0.30	$(0.28, 0.32)$	-1.44
σ_1^2	2	2.03	$(1.81, 2.26)$	-0.72
σ_2^2	16	16.57	$(14.92, 18.28)$	-0.75
σ_3^2	4	3.86	$(3.43, 4.31)$	0.50

Como podemos observar na Tabela 5, as estimativas pontuais dos parâmetros estão próximas dos seus valores reais e, como esperado, os intervalos de 95% contém os valores verdadeiros dos parâmetros. Por sua vez, podemos notar que as estimativas obtidas pelo *Gibbs Sampling* são levemente melhores. Para a grande maioria dos parâmetros o valor da estatística Geweke pertence a $(-2,2)$ indicando convergência. Na Figura 8, apresentamos os gráficos de traço gerados e podemos notar dificuldade na convergência de alguns poucos parâmetros, também concluindo superioridade na convergência do *Gibbs Sampling*.

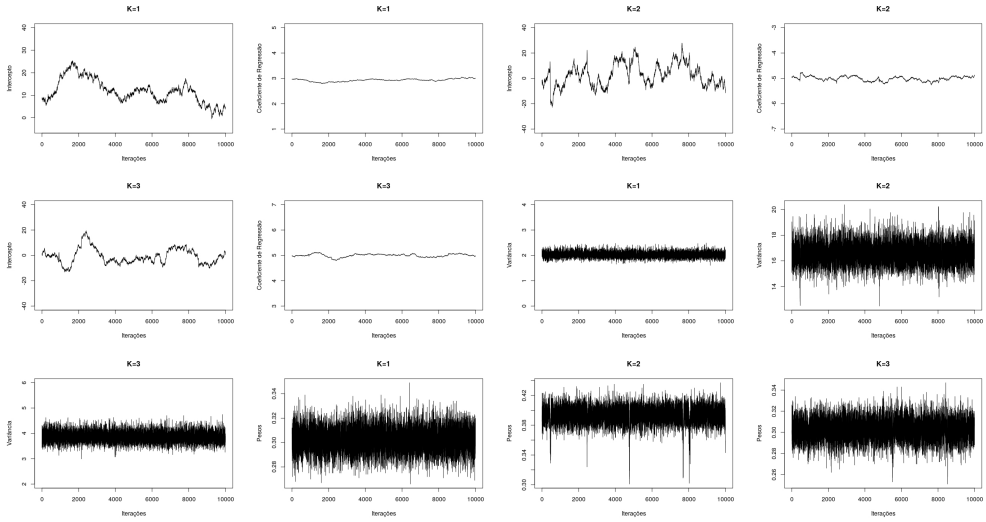


Figura 8 – Simulação 1: gráficos de traço - DDRJ.

4.1.2 Conjunto de Dados Simulado 2

O conjunto de dados artificial 2 foi simulado de maneira semelhante ao conjunto de dados simulado 1, com exceção de que os valores das variâncias e dos betas foram alterados para $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 2$, $\boldsymbol{\beta}_1 = (50,0)^T$, $\boldsymbol{\beta}_2 = (100,-5)^T$ e $\boldsymbol{\beta}_3 = (-1,5)^T$ e o procedimento de simulação utilizado foi dado pelos seguintes passos:

Passo 1: para $i = 1,...,n$, gere $x_{i1} \sim N(10,10)$ e $u_i \sim U(0,1)$, onde $U(0,1)$ é a função densidade da distribuição Uniforme com parâmetros 0 e 1; e

Passo 2: para $i = 1,...,n$, se $u_i \leq a_1$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_1, \sigma_1^2)$ e faça $S_i = 1$; se $a_1 < u_i \leq a_2$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_2, \sigma_2^2)$ e faça $S_i = 2$; e se $a_2 < u_i \leq a_3$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_3, \sigma_3^2)$ e faça $S_i = 3$, onde $\boldsymbol{x}_i = (1, x_{i1})$ e S_i é nossa variável de alocação, que identifica de qual componente a i-ésima observação é proveniente.

A Figura 9 representa o gráfico de dispersão entre a variável resposta e a variável explicativa: os pontos em rosa representam as observações simuladas oriundas da componente 1, os pontos em verde as observações simuladas oriundas da componente 2 e os pontos em azul as

observações simuladas oriundas da componente 3. Como podemos observar, as componentes são muito diferentes e distintas exceto quando o valor da covariável é próximo de 10. Na Figura 10, temos o histograma da variável resposta simulada, y, e sua densidade estimada pelo método Kernel.

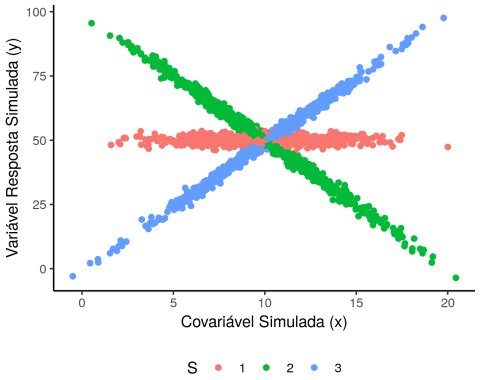


Figura 9 – Simulação 2: gráfico de dispersão.

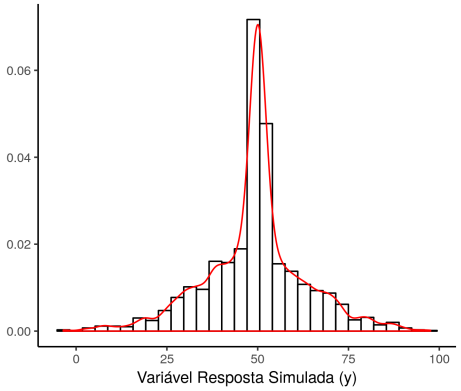


Figura 10 – Simulação 2: histograma da variável resposta simulada.

4.1.2.1 *Estimação via Gibbs Sampling, DIC e EBIC*

O processo de estimação e as amostras MCMC foram geradas respeitando as mesmas características da simulação 1, bem como a análise de convergência.

Encontramos os valores do *DIC* e do *EBIC* na Tabela 6. Note que o modelo com $K = 3$ componentes obteve o menor valor para o *DIC* e para o *EBIC* e, portanto, nesta simulação, também obtivemos uma boa predição para o número de componentes da mistura.

Tabela 6 – Simulação 2: critérios de informação - *DIC* e *EBIC*.

K	1	2	3	4	5
DIC	16080.11	19087.88	11484.26	11559.53	11615.99
EBIC	16107.50	16606.78	11548.15	11622.99	11673.70

Assim, adotando o modelo de mistura de regressão com $K = 3$ componentes, temos na Tabela 7 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais. Também apresentamos nela os intervalos de HPD de 95% e os valores das estatísticas do diagnóstico de Geweke. Como podemos observar, as estimativas pontuais dos parâmetros estão bem próximas dos seus valores reais e, como esperado, os intervalos de 95% contém os valores reais dos parâmetros. O critério de Geweke pertence a $(-2,2)$ para praticamente todos os parâmetros, indicando convergência. Na Figura 11, apresentamos os gráficos de traço dos valores gerados. Além disso, o algoritmo classificou corretamente nas componentes 90% das observações.

Tabela 7 – Simulação 2: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke	Estimativa EM
β_{01}	50	49.84	$(49.42, 50.25)$	-2.01	50.04
β_{11}	0	0.02	$(-0.03, 0.05)$	1.59	0.00
β_{02}	100	100.11	$(99.76, 100.46)$	-0.27	100.08
β_{12}	-5	-5.02	$(-5.05, -4.98)$	0.21	-5.01
β_{03}	-1	-0.98	$(-1.37, -0.61)$	0.30	-1.06
β_{13}	5	5.00	$(4.96, 5.04)$	-0.09	5.00
w_1	0.3	0.30	$(0.28, 0.32)$	0.34	0.30
w_2	0.4	0.40	$(0.38, 0.42)$	0.52	0.40
w_3	0.3	0.30	$(0.28, 0.32)$	-0.89	0.30
σ_1^2	2	2.08	$(1.81, 2.37)$	1.75	2.06
σ_2^2	2	2.10	$(1.87, 2.31)$	-1.49	2.08
σ_3^2	2	1.92	$(1.70, 2.16)$	1.60	1.91

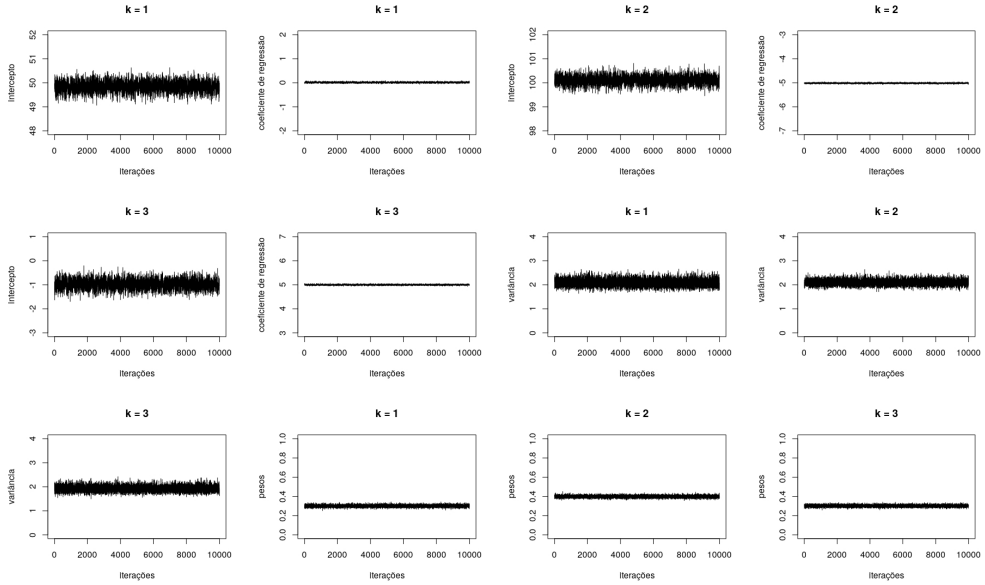


Figura 11 – Simulação 2: gráficos de traço - *Gibbs Sampling*.

4.1.2.2 Estimação via EM

Da mesma forma que na seção anterior, cinco modelos com diferentes números de componentes, $K = 1, ..., 5$, foram ajustados, mas, agora, via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools* implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos pelo pacote.

Os valores do *AIC* e do *BIC* são encontrados na Tabela 8. Note que o modelo com $K = 3$ componentes obteve o menor valor para o *AIC* e para *BIC* e, portanto, nesta simulação obtivemos uma boa predição para o número de componentes da mistura através do algoritmo EM combinado com métodos de seleção de modelos.

Tabela 8 – Simulação 2: critérios de informação - *AIC* e *BIC*.

K	1	2	3	4	5
AIC	inf	8023.42	5423.78	5426.90	5423.85
BIC	inf	8040.23	5451.78	5466.11	5474.26

Adotando o modelo de mistura de regressão com $K = 3$ componentes, as estimativas pontuais obtidas são apresentadas na última coluna da Tabela 7 e podemos notar que os valores obtidos são semelhantes aos obtidos pelo algoritmo *Gibbs Sampling* desenvolvido neste trabalho. O algoritmo EM classificou corretamente 91% das observações.

Da mesma forma que na simulação anterior, também realizamos a estimação para diferentes valores iniciais e para valores iniciais muito distantes dos valores reais dos parâmetros e o algoritmo EM não apresentou boas estimativas, constatando, novamente, a sensibilidade do algoritmo EM aos valores iniciais.

4.1.2.3 Estimação via DDRJ

O processo de estimação via DDRJ e as amostras MCMC foram geradas respeitando as mesmas características da simulação 1, bem como a análise de convergência.

A Tabela 9 apresenta os valores obtidos para a probabilidade *a posteriori* de K. O máximo valor *a posteriori* é obtido em $K = 1$ e, portanto, o algoritmo não foi capaz de estimar corretamente o valor de K para este cenário simulado.

Tabela 9 – Simulação 2: probabilidade *a posteriori* para K.

K	1	2	3	4	5
$P(K\|\cdots)$	1.0	0.0	0.0	0.0	0.0

Temos na Tabela 10 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais, que são a média *a posteriori* de cada parâmetro da amostra. Também apresentamos nela o intervalo de credibilidade de máxima densidade *a posteriori* (HPD) de 95% e o diagnóstico de Geweke para convergência.

Tabela 10 – Simulação 2: estimativas para os parâmetros - DDRJ.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	50	54.88	$(54.08, 56.90)$	0.42
β_{11}	0	-0.56	$(-0.74, -0.38)$	-0.30
β_{02}	100	-	-	-
β_{12}	-5	-	-	-
β_{03}	-1	-	-	-
β_{13}	5	-	-	-
w_1	0.3	1.00	-	-
w_2	0.4	-	-	-
w_3	0.3	-	-	-
σ_1^2	2	181.47	$(170.33, 193.00)$	-0.80
σ_2^2	2	-	-	-
σ_3^2	2	-	-	-

Desta forma, pelos resultados obtidos, concluímos que o algoritmo DDRJ proposto não foi adequado para estimação deste modelo e os resultados obtidos pelo *Gibbs Sampling* e pelo algoritmo EM foram superiores. O algoritmo só foi capaz de identificar a primeira componente.

Na Figura 12, apresentamos os gráficos de traço gerados que, juntamente com o diagnóstico de Geweke apresentado na Tabela 10, atestam convergência.

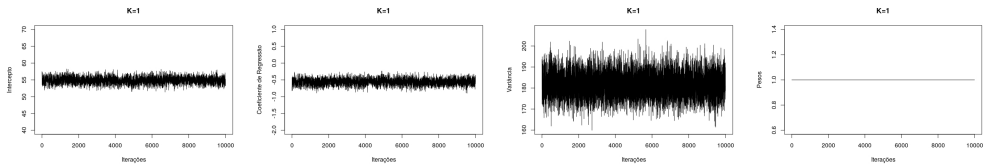


Figura 12 – Simulação 2: gráficos de traço - DDRJ.

4.1.3 Conjunto de Dados Simulado 3

O conjunto de dados artificial 3 foi simulado respeitando as mesmas características do conjunto de dados simulado 2, porém, alteramos os valores das variâncias para $\sigma_1^2 = 2$, $\sigma_2^2 = 16$, $\sigma_3^2 = 4$. Na Figura 13 temos a representação do gráfico de dispersão entre a variável resposta e a variável explicativa: os pontos em rosa representam as observações simuladas oriundas da componente 1, os pontos em verde as observações simuladas oriundas da componente 2 e os pontos em azul as observações simuladas oriundas da componente 3. As componentes continuam bem diferentes e distintas, no entanto, diminuímos a distância entre elas quando o valor da covariável está próximo de 10. Na Figura 14, temos o histograma da variável resposta simulada, y, e sua densidade estimada pelo método Kernel.

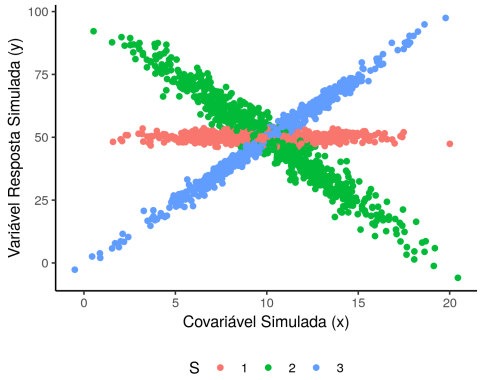


Figura 13 – Simulação 3: gráfico de dispersão.

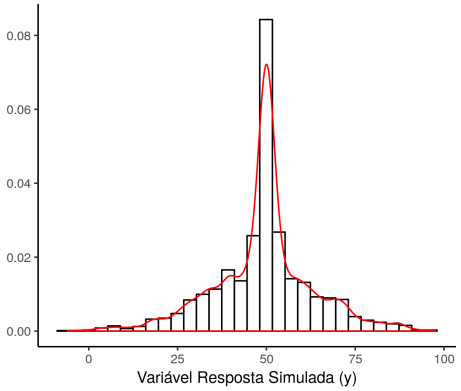


Figura 14 – Simulação 3: histograma da variável resposta simulada.

4.1.3.1 *Estimação via Gibbs Sampling, DIC e EBIC*

O processo de estimação e as amostras MCMC foram geradas respeitando as mesmas características da simulação 1, bem como a análise de convergência.

Encontramos os valores do *DIC* e do *EBIC* na Tabela 11. O modelo com $K = 3$ componentes obteve o menor valor para o *EBIC*, enquanto que o modelo com $K = 5$ componentes obteve o menor valor para o *DIC*. Apesar dos critérios de informação divergirem, o modelo selecionado pelo *DIC* apresentou duas componentes com pesos estimados zero, o que torna o modelo com o mesmo número de componentes que o escolhido pelo *EBIC*. Desta maneira, optamos pela escolha do modelo com $K = 3$ componentes e apontamos pela instabilidade na utilização do critério de informação *DIC*.

Tabela 11 – Simulação 3: critérios de informação - *DIC* e *EBIC*.

K	1	2	3	4	5
DIC	16159.24	16569.24	14909.66	14723.55	14400.46
EBIC	16186.62	16245.79	14263.53	14330.39	14404.92

Assim, adotando o modelo de mistura de regressão com $K = 3$ componentes, temos na Tabela 12 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais.

Também apresentamos nela os intervalos de HPD de 95% e os valores das estatísticas do diagnóstico de Geweke. Como podemos observar, as estimativas pontuais dos parâmetros estão bem próximas dos seus valores reais e, como esperado, os intervalos de HPD de 95% contêm os valores verdadeiros dos parâmetros. O valor da estatística do critério de Geweke pertence a $(-2,2)$ para todos os parâmetros, indicando convergência. Na Figura 15, apresentamos os gráficos de traço dos valores gerados que também evidenciam a convergência das cadeias. Além disso, o algoritmo classificou corretamente nas componentes 86.7% das observações.

Tabela 12 – Simulação 3: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke	Estimativa EM
β_{01}	50	50.02	$(49.52, 50.53)$	1.19	50.03
β_{11}	0	0.00	$(-0.05, 0.05)$	-1.34	0.00
β_{02}	100	100.14	$(99.15, 101.10)$	-1.13	100.28
β_{12}	-5	-5.03	$(-5.12, -4.94)$	1.07	-5.04
β_{03}	-1	-1.26	$(-1.84, -0.71)$	0.15	-1.11
β_{13}	5	5.02	$(4.97, 5.08)$	-0.43	5.01
w_1	0.3	0.30	$(0.28, 0.33)$	0.40	0.30
w_2	0.4	0.40	$(0.37, 0.42)$	-1.55	0.40
w_3	0.3	0.30	$(0.28, 0.32)$	1.25	0.30
σ_1^2	2	2.12	$(1.82, 2.45)$	-0.21	2.10
σ_2^2	16	16.47	$(14.64, 18.34)$	0.98	16.35
σ_3^2	4	3.80	$(3.30, 4.27)$	0.84	3.77

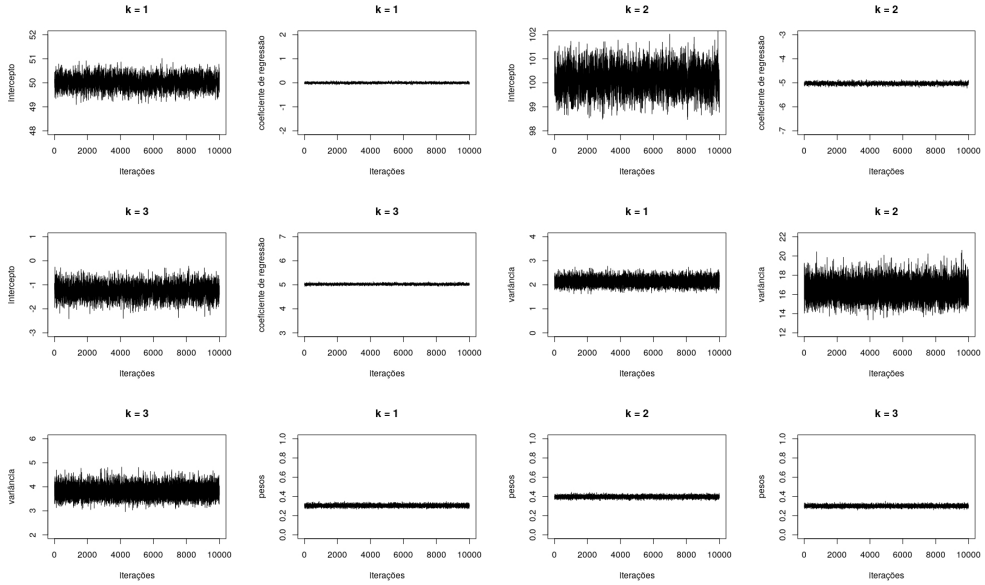


Figura 15 – Simulação 3: gráficos de traço - *Gibbs Sampling*.

4.1.3.2 Estimação via EM

Como anteriormente, cinco modelos com diferentes números de componentes, $K = 1, \ldots, 5$, foram ajustados via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools*

implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos pelo pacote.

Os valores do *AIC* e do *BIC* são encontrados na Tabela 13. Note que o modelo com $K = 5$ componentes obteve o menor valor para o *AIC* e o modelo com $K = 3$ componentes obteve o menor valor para o *BIC*. Diferentemente da estimação via *Gibbs Sampling*, o modelo com $K = 5$ componentes estimado via algoritmo EM apresentou apenas uma componente com peso estimado zero. Unindo este resultado aos apresentados na Simulação 1, podemos notar dificuldades do *AIC* em estimar o valor de K quando o modelo é ajustado via algoritmo EM mesmo em casos cujas componentes são bem evidentes.

Tabela 13 – Simulação 3: critérios de informação - *AIC* e *BIC*.

K	1	2	3	4	5
AIC	inf	8063.16	6294.35	6298.08	6293.65
BIC	inf	8079.96	6322.35	6337.29	6344.06

Adotando o modelo de mistura de regressão com $K = 3$ componentes, as estimativas pontuais obtidas são apresentadas na última coluna da Tabela 12 e podemos notar que os valores obtidos são semelhantes aos obtidos pelo algoritmo *Gibbs Sampling* desenvolvido neste trabalho. O algoritmo EM classificou corretamente 85.8% das observações.

Da mesma forma que nas simulações anteriores, considerando um modelo de mistura de regressão com $K = 3$ componentes, também realizamos a estimação para diferentes valores iniciais e para valores iniciais muito distantes dos valores reais dos parâmetros e o algoritmo EM não apresentou boas estimativas, constatando, novamente, a sensibilidade do algoritmo EM aos valores iniciais.

4.1.3.3 Estimação via DDRJ

O processo de estimação via DDRJ e as amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência.

A Tabela 14 apresenta os valores obtidos para a probabilidade *a posteriori* de K. O máximo valor *a posteriori* é obtido em $K = 1$ e, portanto, novamente o algoritmo não foi capaz de estimar corretamente o valor de K.

Tabela 14 – Simulação 3: probabilidade *a posteriori* para K.

K	1	2	3	4	5
$P(K\|\cdots)$	1.0	0.0	0.0	0.0	0.0

Temos na Tabela 15 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais, que são a média *a posteriori* de cada parâmetro da amostra. Também apresentamos nela o intervalo de credibilidade de máxima densidade *a posteriori* (HPD) de 95% e o diagnóstico de Geweke para convergência.

Tabela 15 – Simulação 3: estimativas para os parâmetros - DDRJ.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	50	54.93	$(52.98, 56.91)$	0.42
β_{11}	0	-0.57	$(-0.76, -0.39)$	-0.30
β_{02}	100	-	-	-
β_{12}	-5	-	-	-
β_{03}	-1	-	-	-
β_{13}	5	-	-	-
w_1	0.3	1.00	-	-
w_2	0.4	-	-	-
w_3	0.3	-	-	-
σ_1^2	2	188.79	$(177.26, 200.61)$	-0.80
σ_2^2	16	-	-	-
σ_3^2	4	-	-	-

O algoritmo só foi capaz de identificar a primeira componente e novamente observamos a superioridade dos resultados obtidos pelo *Gibbs Sampling* e pelo algoritmo EM.

Na Figura 16, apresentamos os gráficos de traço gerados que, juntamente com o diagnóstico de Geweke apresentado na Tabela 15, atestam convergência.

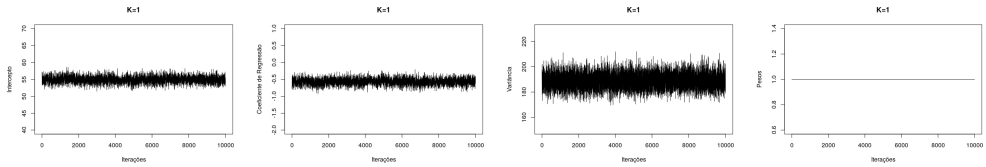


Figura 16 – Simulação 3: gráficos de traço - DDRJ.

4.1.4 Dados Simulados 4

Na geração do quarto conjunto de dados artificiais, consideramos um cenário semelhante aos conjuntos de dados simulados 2 e 3, porém, alteramos novamente os valores das variâncias para $\sigma_1^2 = 30$, $\sigma_2^2 = 50$ e $\sigma_3^2 = 40$. A Figura 17 representa o gráfico de dispersão entre a variável resposta e a variável explicativa: os pontos em rosa representam as observações simuladas oriundas da componente 1, os pontos em verde as observações simuladas oriundas da componente 2 e os pontos em azul as observações simuladas oriundas da componente 3. Aqui percebemos que apesar de nas extremidades dos valores da covariável as diferenças entre as componentes serem mais evidentes, no intervalo $(5, 15)$ para o valor da covariável, o valor da variável resposta para diferentes componente já está muito parecido e as componentes pouco evidentes. Na Figura 18, temos o histograma da variável resposta simulada, y, e sua densidade estimada.

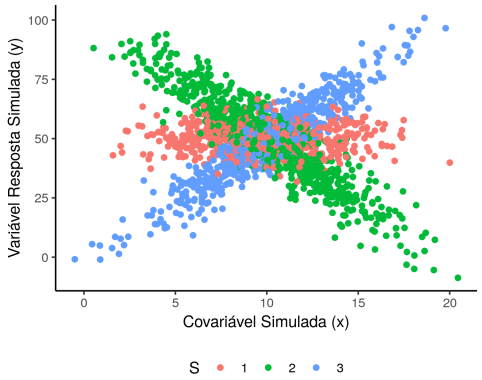


Figura 17 – Simulação 4: gráfico de dispersão.

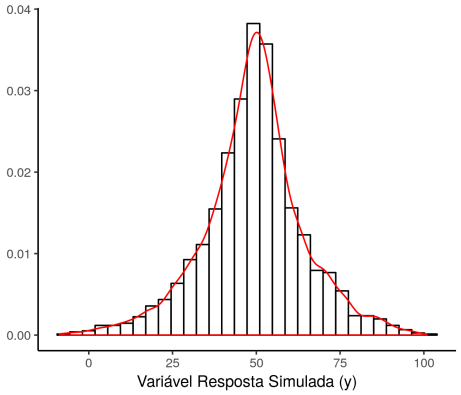


Figura 18 – Simulação 4: histograma da variável resposta simulada.

4.1.4.1 *Estimação via Gibbs Sampling, DIC e EBIC*

As amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência.

Encontramos os valores do *DIC* e do *EBIC* na Tabela 16. Podemos observar que o modelo com $K = 1$ componente é escolhido levando em consideração os dois critérios. Além disso, o valor comparativamente alto para o *DIC* obtido pelo modelo com $K = 5$ deve-se ao fato de que não observamos convergência da cadeia MCMC, fato que já pode nos indicar que esse modelo não é adequado aos dados.

Tabela 16 – Simulação 4: critérios de informação - *DIC* e *EBIC*.

K	1	2	3	4	5
DIC	16495.67	17406.93	17513.86	57910.77	439052.9
EBIC	16523.04	17310.26	18329.84	18505.94	18774.14

Adotando o modelo de mistura de regressão com $K = 1$ componentes, temos na Tabela 17 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais. Também apresentamos nela os intervalos de credibilidade de HPD de 95% e o diagnóstico de Geweke. A reta de regressão simples obtida pelo modelo está representada na Figura 19.

Os valores das estatísticas do critério de Geweke estão próximos ou contido no intervalo $(-2,2)$ para todos os parâmetros, indicando convergência. Na Figura 20, apresentamos os gráficos de traço dos valores gerados, também indicando convergência.

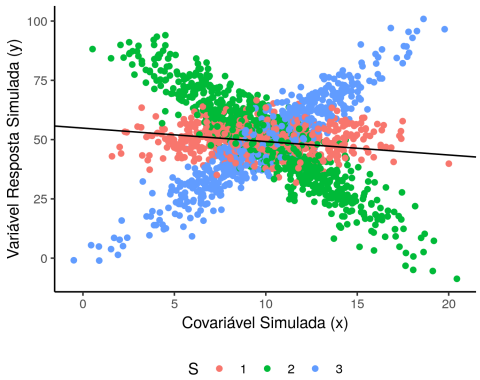


Figura 19 – Regressão simples obtida para o conjunto de dados simulado 4.

Tabela 17 – Simulação 4: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke	Estimativa EM
β_{01}	50	54.84	$(52.70, 57.08)$	-2.10	49.30
β_{11}	0	-0.57	$(-0.78, -0.36)$	1.96	0.08
β_{02}	100	-	-	-	100.74
β_{12}	-5	-	-	-	-5.10
β_{03}	-1	-	-	-	-2.06
β_{13}	5	-	-	-	5.11
w_1	0.3	1.00	-	-	0.31
w_2	0.4	-	-	-	0.41
w_3	0.3	-	-	-	0.28
σ_1^2	30	223.36	$(209.70, 237.68)$	0.2058	31.66
σ_2^2	50	-	-	-	48.60
σ_3^2	40	-	-	-	35.24

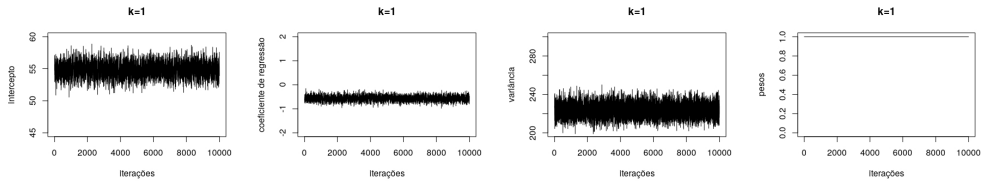


Figura 20 – Simulação 4: gráficos de traço - *Gibbs Sampling*.

Vale observar que, assumindo o número de componentes do modelo de mistura de regressão conhecido, $K = 3$, o algoritmo *Gibbs Sampling* desenvolvido neste trabalho apresentou boas estimativa para os parâmetros do modelo. Assim, concluímos com esta simulação que a proposta de seleção de modelos já não funciona para o caso em que as componentes são pouco evidentes, porém o algoritmo *Gibbs Sampling* ainda é capaz de fornecer boas estimativas neste cenário quando o número de componentes é conhecido.

4.1.4.2 Estimação via EM

Cinco modelos com diferentes números de componentes, $K = 1, ..., 5$, foram ajustados via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools* implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos pelo pacote.

Os valores do *AIC* e do *BIC* são encontrados na Tabela 18. Note que o modelo com $K = 5$ componentes obteve o menor valor para o *AIC* e o modelo com $K = 3$ componentes obteve o menor valor para o *BIC*. Novamente constatamos a superioridade do critério de informação *BIC* frente ao *AIC* quando a estimação dos modelos é feita via algoritmo EM.

Tabela 18 – Simulação 4: critérios de informação - *AIC* e *BIC*.

K	1	2	3	4	5
AIC	inf	8236.18	7591.56	7592.99	7589.06
BIC	inf	8252.98	7619.57	7632.20	7639.46

Adotando o modelo de mistura de regressão com $K = 3$ componentes, as estimativas pontuais obtidas são apresentadas na última coluna da Tabela 17 e podemos notar que são obtidas boas estimativas pontuais para os parâmetros. O algoritmo EM classificou corretamente 86.6% das observações.

Da mesma forma que nas simulações anteriores, considerando um modelo de mistura de regressão com $K = 3$ componentes, também realizamos a estimação para diferentes valores iniciais e para valores iniciais muito distantes dos valores reais dos parâmetros e o algoritmo EM não apresentou boas estimativas.

4.1.4.3 Estimação via DDRJ

O processo de estimação via DDRJ e as amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência.

A Tabela 19 apresenta os valores obtidos para a probabilidade *a posteriori* de K. O máximo valor *a posteriori* é obtido em $K = 1$ obtendo o mesmo resultado que o algoritmo *Gibbs Sampling*. Como esperado devido aos resultados apresentados nas simulações anteriores, o algoritmo não foi capaz de estimar corretamente o valor de K.

Tabela 19 – Simulação 4: probabilidade *a posteriori* para K.

K	1	2	3	4	5
$P(K\|\cdots)$	1.0	0.0	0.0	0.0	0.0

Temos na Tabela 20 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais, que são a média *a posteriori* de cada parâmetro da amostra. Também apresentamos nela

o intervalo de credibilidade de máxima densidade *a posteriori* (HPD) de 95% e o diagnóstico de Geweke para convergência.

Tabela 20 – Simulação 4: estimativas para os parâmetros - DDRJ.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	50	54.88	$(52.77, 57.04)$	0.13
β_{11}	0	-0.57	$(-0.77, -0.37)$	-0.14
β_{02}	100	-	-	-
β_{12}	-5	-	-	-
β_{03}	-1	-	-	-
β_{13}	5	-	-	-
w_1	0.3	1.00	-	-
w_2	0.4	-	-	-
w_3	0.3	-	-	-
σ_1^2	30	223.37	$(209.73, 237.36)$	-0.15
σ_2^2	50	-	-	-
σ_3^2	40	-	-	-

O algoritmo só foi capaz de identificar a primeira componente. Para este cenário, as estimativas obtidas via algoritmo *Gibbs Sampling* foram muito próximas das obtidas via algoritmo DDRJ. Neste cenário, o algoritmo EM foi capaz de estimar melhor tanto o número de componentes como os parâmetros do modelo.

Na Figura 21, apresentamos os gráficos de traço gerados que, juntamente com o diagnóstico de Geweke apresentado na Tabela 20, atestam convergência.

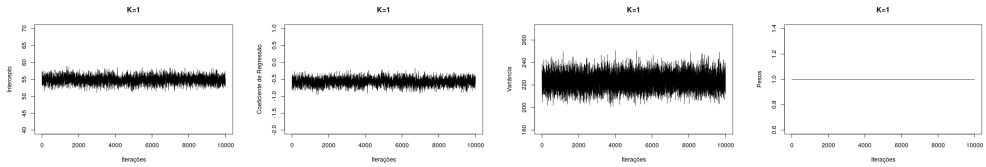


Figura 21 – Simulação 4: gráficos de traço - DDRJ.

4.1.5 Conjunto de Dados Simulado 5

Na geração do quinto conjunto de dados simulado seguimos a mesma estrutura anterior. Os valores permanecem inalterados, exceto para as variâncias que passam a ser $\sigma_1^2 = 300$, $\sigma_2^2 = 450$ e $\sigma_3^2 = 340$. A Figura 22 representa o gráfico de dispersão entre a variável resposta e a variável explicativa: os pontos em rosa representam as observações simuladas oriundas da componente 1, os pontos em verde as observações simuladas oriundas da componente 2 e os pontos em azul as observações simuladas oriundas da componente 3. Para esse conjunto de dados simulados observamos que as observações geradas por diferentes componentes estão todas misturadas e as componentes são bem pouco evidentes. Na Figura 23, temos o histograma da variável resposta simulada, y, e sua densidade estimada pelo método Kernel.

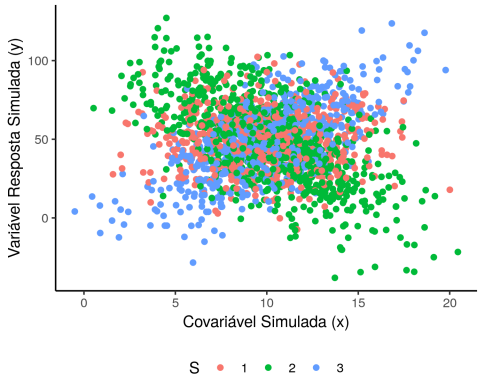


Figura 22 – Simulação 5: gráfico de dispersão.

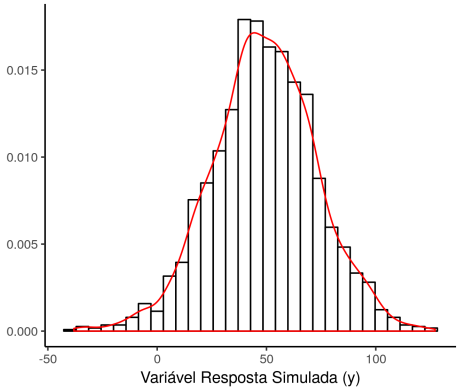


Figura 23 – Simulação 5: histograma da variável resposta simulada.

4.1.5.1 *Estimação via Gibbs Sampling, DIC e EBIC*

Para este conjunto de dados, adotamos o mesmo processo de estimação dos conjuntos de dados simulados anteriores. Observamos que o método proposto não atingiu a convergência para $K \geq 2$, evidenciando a não adequação do modelo para modelos de mistura de regressão com componentes pouco esparsas.

Assim, adotamos o modelo de mistura de regressão com $K = 1$ componentes. Temos na Tabela 21 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais. Também apresentamos nela o intervalo de credibilidade de HPD de 95% e o diagnóstico de Geweke. A reta de regressão simples obtida pelo modelo está representada na Figura 24. Os valores das estatísticas do critério de Geweke estão contidos no intervalo $(-2, 2)$ para todos os parâmetros, indicando convergência. Na Figura 25, apresentamos os gráficos de traço dos valores gerados, também indicando convergência.

4.1.5.2 *Estimação via EM*

Cinco modelos com diferentes números de componentes, $K = 1, \ldots, 5$, foram ajustados via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools* implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos

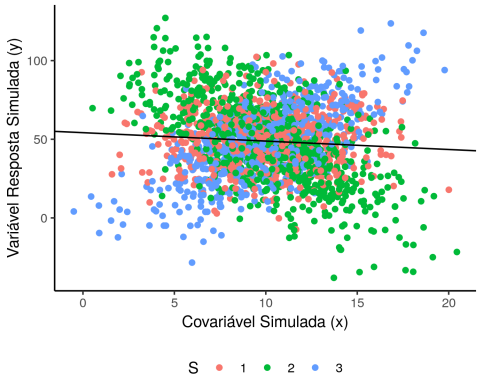


Figura 24 – Regressão simples obtida para o conjunto de dados simulado 5.

Tabela 21 – Simulação 5: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke	Estimativa EM
β_{01}	50	54.18	$(50.76, 57.63)$	-0.56	52.65
β_{11}	0	-0.54	$(-0.88, -0.22)$	0.54	0.14
β_{02}	100	-	-	-	94.01
β_{12}	-5	-	-	-	-4.49
β_{03}	-1	-	-	-	6.11
β_{13}	5	-	-	-	4.30
w_1	0.3	1.00	-	-	0.01
w_2	0.4	-	-	-	0.56
w_3	0.3	-	-	-	0.43
σ_1^2	300	563.18	$(529.42, 600.22)$	-0.46	0.00
σ_2^2	450	-	-	-	415
σ_3^2	340	-	-	-	318

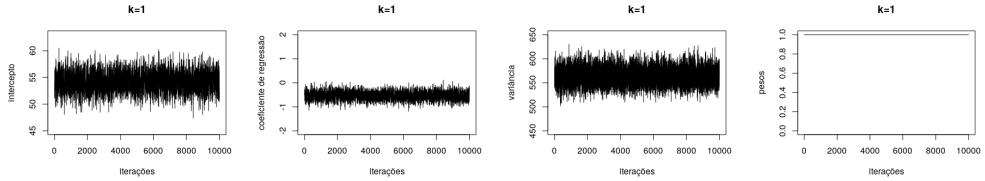


Figura 25 – Simulação 5: gráficos de traço - *Gibbs Sampling*.

pelo pacote.

Os valores do *AIC* e do *BIC* são encontrados na Tabela 22. Note que o modelo com $K = 5$ componentes obteve o menor valor para o *AIC* e o modelo com $K = 3$ componentes obteve o menor valor para o *BIC*. Mais uma vez notamos a superioridade do critério de informação *BIC* frente ao *AIC* quando a estimação dos modelos é feita via algoritmo EM.

Adotando o modelo de mistura de regressão com $K = 3$ componentes, as estimativas pontuais obtidas são apresentadas na última coluna da Tabela 21. O algoritmo EM apresenta boas estimativas para a componente 1 que é a componente intermediária. Também apresenta boas estimativas para a componente 2, porém apenas 1% das observações são alocadas nela e

Tabela 22 – Simulação 5: critérios de informação - *AIC* e *BIC*.

K	1	2	3	4	5
AIC	inf	9170.20	9078.42	9070.85	9066.71
BIC	inf	9187.00	9106.43	9110.06	9117.12

isso não é verdade nos dados simulados. Já os coeficientes de regressão da componente 3 não são bem estimados. O algoritmo EM já não foi capaz de classificar bem as observações em suas reais componentes, classificando corretamente apenas 45% das observações.

Considerando um modelo de mistura de regressão com $K = 3$ componentes, também realizamos a estimação para diferentes valores iniciais e para valores iniciais muito distantes dos valores reais dos parâmetros e o algoritmo EM não apresentou boas estimativas.

4.1.5.3 Estimação via DDRJ

O processo de estimação via DDRJ e as amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência.

A Tabela 23 apresenta os valores obtidos para a probabilidade *a posteriori* de K. O máximo valor *a posteriori* é obtido em $K = 1$ obtendo o mesmo resultado que o algoritmo *Gibbs Sampling*.

Tabela 23 – Simulação 5: probabilidade *a posteriori* para K.

K	1	2	3	4	5
$P(K\|\cdots)$	1.0	0.0	0.0	0.0	0.0

Temos na Tabela 24 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais, que são a média *a posteriori* de cada parâmetro da amostra. Também apresentamos nela o intervalo de credibilidade de máxima densidade *a posteriori* (HPD) de 95% e o diagnóstico de Geweke para convergência.

O algoritmo só foi capaz de identificar a primeira componente. Novamente, as estimativas obtidas via algoritmo *Gibbs Sampling* foram muito próximas das obtidas via algoritmo DDRJ. A metodologia apresentada na estimação via algoritmo EM foi capaz de estimar melhor tanto o número de componentes como os parâmetros do modelo.

Na Figura 26, apresentamos os gráficos de traço gerados que, juntamente com o diagnóstico de Geweke apresentado na Tabela 24, atestam convergência.

4.1.6 Dados Simulados 6

No sexto conjunto de dados simulados, utilizamos o mesmo processo de geração dos dados utilizado até o momento, porém alteramos os valores dos betas e das variâncias para $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = (2,5)^T$, $\sigma_1^2 = 5$, $\sigma_2^2 = 50$ e $\sigma_3^2 = 500$.

Tabela 24 – Simulação 5: estimativas para os parâmetros - DDRJ.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	50	54.23	$(50.70, 57.42)$	0.13
β_{11}	0	-0.54	$(-0.87, -0.23)$	-0.14
β_{02}	100	-	-	-
β_{12}	-5	-	-	-
β_{03}	-1	-	-	-
β_{13}	5	-	-	-
w_1	0.3	1.00	-	-
w_2	0.4	-	-	-
w_3	0.3	-	-	-
σ_1^2	300	563.29	$(528.70, 598.38)$	-0.16
σ_2^2	450	-	-	-
σ_3^2	340	-	-	-

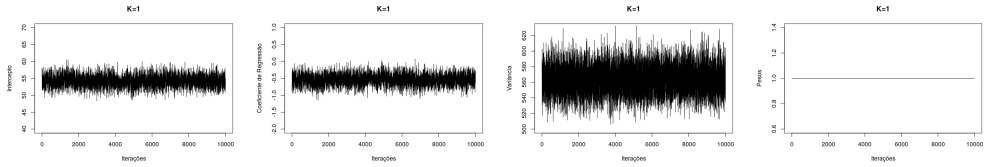


Figura 26 – Simulação 5: gráficos de traço - DDRJ.

A Figura 27 representa o gráfico de dispersão entre a variável resposta e a variável explicativa: os pontos em rosa representam as observações simuladas oriundas da componente 1, os pontos em verde as observações simuladas oriundas da componente 2 e os pontos em azul as observações simuladas oriundas da componente 3. Esta situação simulada se trata de um modelo de regressão linear simples apenas com variâncias diferentes entre as componentes. Na Figura 28, temos o histograma da variável resposta simulada, y, e sua densidade estimada pelo método Kernel.

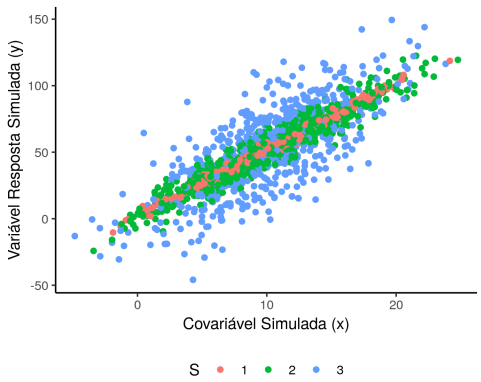


Figura 27 – Simulação 6: gráfico de dispersão.

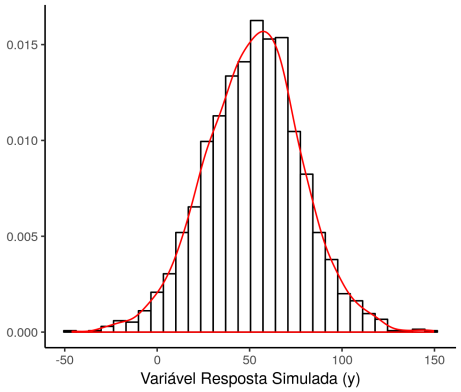


Figura 28 – Simulação 6: histograma da variável resposta simulada.

4.1.6.1 *Estimação via Gibbs Sampling, DIC e EBIC*

As amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência. No entanto, o método apresentou dificuldades de convergência para valores de K maiores ou iguais a dois e, assim, adotamos o modelo de mistura de regressão com $K = 1$ componentes.

Temos na Tabela 25 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais. Também apresentamos nela o intervalor de credibilidade de HPD de 95% e o diagnóstico de Geweke. A reta de regressão simples obtida pelo modelo está representada na Figura 29.

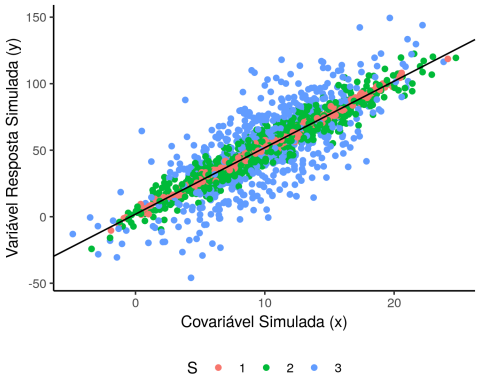


Figura 29 – Regressão simples obtida pela simulação 6.

As estimativas obtidas para os coeficientes de regressão são muito boas, proporcionando um bom ajuste aos dados. Os valores das estatísticas do critério de Geweke estão contidos no intervalo $(-2, 2)$ para todos os parâmetros, indicando convergência. Na Figura 30, apresentamos os gráficos de traço dos valores gerados, também indicando convergência.

Tabela 25 – Simulação 6: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke	Estimativa EM
β_{01}	2	1.94	$(0.56, 3.32)$	-1.80	2.17
β_{11}	5	5.00	$(4.87, 5.12)$	1.76	4.96
β_{02}	2	-	-	-	-
β_{12}	5	-	-	-	-
β_{03}	2	-	-	-	1.44
β_{13}	5	-	-	-	5.06
w_1	0.3	1.00	-	-	0.52
w_2	0.4	-	-	-	-
w_3	0.3	-	-	-	0.48
σ_1^2	5	167.79	$(157.74, 178.78)$	0.59	10
σ_2^2	50	-	-	-	-
σ_3^2	500	-	-	-	338

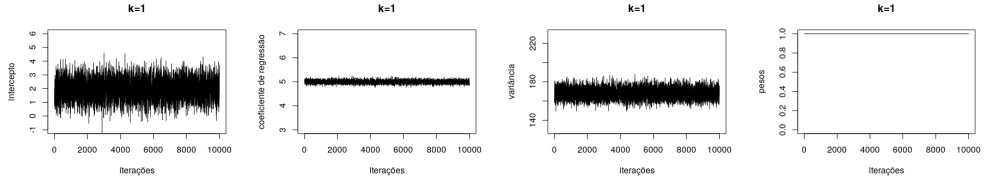


Figura 30 – Simulação 6: gráficos de traço - *Gibbs Sampling*.

4.1.6.2 Estimação via EM

Cinco modelos com diferentes números de componentes, $K = 1, ..., 5$, foram ajustados via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools* implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos pelo pacote.

Os valores do *AIC* e do *BIC* são encontrados na Tabela 26. Note que o modelo com $K = 5$ componentes obteve o menor valor para o *AIC* e o modelo com $K = 3$ componentes obteve o menor valor para o *BIC*.

Tabela 26 – Simulação 6: critérios de informação - *AIC* e *BIC*.

K	1	2	3	4	5
AIC	inf	7901.89	7483.55	7484.58	7478.18
BIC	inf	7918.69	7511.56	7523.78	7528.58

Adotando o modelo de mistura de regressão com $K = 3$ componentes, as estimativas pontuais obtidas são apresentadas na última coluna da Tabela 25. Note que apesar de o critério de informação *BIC* ter selecionado o modelo com $K = 3$, o algoritmo EM só conseguiu identificar duas componentes e estimar de forma razoável os coeficientes de regressão. Os demais parâmetros não foram bem estimados. Além disso, o algoritmo EM já não foi capaz de classificar bem as observações em suas devidas componentes: as observações oriundas da componente dois foram

distribuídas entre as componentes 1 e 3; 18% das observações oriundas da componente 1 foram alocadas na componente 3;e 20% das observações oriundas da componente 3 foram alocadas na componente 1.

4.1.6.3 Estimação via DDRJ

O processo de estimação via DDRJ e as amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência.

A Tabela 27 apresenta os valores obtidos para a probabilidade *a posteriori* de K. O máximo valor *a posteriori* é obtido em $K = 1$ obtendo o mesmo resultado que o algoritmo *Gibbs Sampling*.

Tabela 27 – Simulação 6: probabilidade *a posteriori* para K.

K	1	2	3	4	5
$P(K\|\cdots)$	1.0	0.0	0.0	0.0	0.0

Temos na Tabela 28 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais, que são a média *a posteriori* de cada parâmetro da amostra. Também apresentamos nela o intervalo de credibilidade de máxima densidade *a posteriori* (HPD) de 95% e o diagnóstico de Geweke para convergência.

Tabela 28 – Simulação 6: estimativas para os parâmetros - DDRJ.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	2	1.96	$(0.60, 3.34)$	-0.02
β_{11}	5	5.00	$(4.86, 5.11)$	-0.03
β_{02}	2	-	-	-
β_{12}	5	-	-	-
β_{03}	2	-	-	-
β_{13}	5	-	-	-
w_1	0.3	1.00	-	-
w_2	0.4	-	-	-
w_3	0.3	-	-	-
σ_1^2	5	167.78	$(157.72, 178.45)$	-0.15
σ_2^2	50	-	-	-
σ_3^2	500	-	-	-

As estimativas obtidas via algoritmo *Gibbs Sampling* foram muito próximas das obtidas via algoritmo DDRJ.

Na Figura 31, apresentamos os gráficos de traço gerados que, juntamente com o diagnóstico de Geweke apresentado na Tabela 28, atestam convergência.

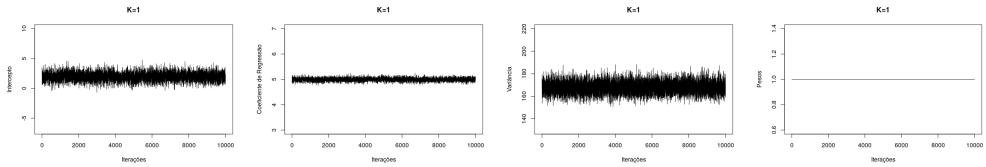


Figura 31 – Simulação 6: gráficos de traço - DDRJ.

4.1.7 Conjunto de Dados Simulado 7

Na geração do sétimo conjunto de dados artificiais, consideramos um modelo de mistura de regressão normal com $K = 3$ componentes e $p = 2$ covariáveis,

$$Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i \sim \sum_{k=1}^{3} w_k N(\boldsymbol{x}_i\boldsymbol{\beta}_k, \sigma_k^2),$$

onde $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \beta_{2k})^T$, para $k = 1, ..., 3$.

Consideramos uma amostra de tamanho $n = 2000$ e fixamos $w_1 = 0.2$, $w_2 = 0.3$, $w_3 = 0.5$, $\sigma_1^2 = 5$, $\sigma_2^2 = 50$, $\sigma_3^2 = 500$ e $\boldsymbol{\beta}_1 = (5, 3, -4)^T$, $\boldsymbol{\beta}_2 = (0, -8, 0.7)^T$ e $\boldsymbol{\beta}_3 = (-16, 8, 3)^T$. Seja $\mathbf{a} = (a_1, a_2, a_3)$, onde $a_j = \sum_{k=1}^{j} w_k$ para $j = 1, ..., 3$, o vetor acumulado de $\boldsymbol{w}$, o procedimento de simulação utilizado é dado pelos seguintes passos:

Passo 1: para $i = 1, ..., n$, gere $x_{i1} \sim N(10, 20)$, $x_{i2} \sim N(20, 4)$ e $u_i \sim U(0, 1)$; e

Passo 2: para $i = 1, ..., n$, se $u_i \leq a_1$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_1, \sigma_1^2)$ e faça $S_i = 1$; se $a_1 < u_i \leq a_2$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_2, \sigma_2^2)$ e faça $S_i = 2$; e se $a_2 < u_i \leq a_3$, gere $Y_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_3, \sigma_3^2)$ e faça $S_i = 3$, onde $\boldsymbol{x}_i = (1, x_{i1}, x_{i2})$ e S_i é nossa variável de alocação.

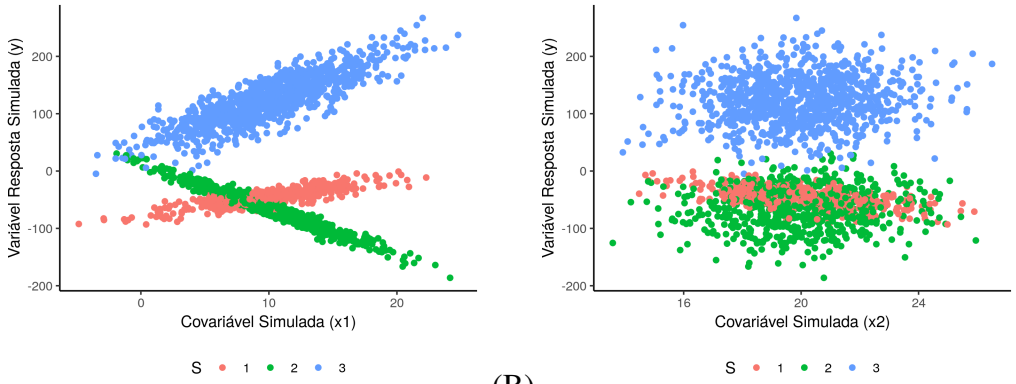
A Figura 32 representa os gráficos de dispersão entre a variável resposta e as covariáveis: os pontos em rosa representam as observações simuladas oriundas da componente 1, os pontos em verde as observações simuladas oriundas da componente 2 e os pontos em azul as observações simuladas oriundas da componente 3. A primeira covariável parece separar bem as 3 componentes, ao passo que a segunda covariável não separa as observações da componente 1 e 2. Na Figura 33, temos o histograma da variável resposta simulada, y, e sua densidade estimada pelo método Kernel.

4.1.7.1 Estimação via Gibbs Sampling, DIC e EBIC

O processo de estimação e as amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência.

Encontramos os valores do *DIC* e do *EBIC* na Tabela 29. Temos que os dois critérios selecionam o modelo com $K = 3$ componentes como o modelo com melhor ajuste.

Adotamos o modelo de mistura de regressão com $K = 3$ componentes. Temos na Tabela 30 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais. Também



(A) (B)

Figura 32 – Simulação 7: gráfico de dispersão. (A) gráfico de dispersão da variável resposta com a primeira variável explicativa. (B) gráfico de dispersão da variável resposta com a segunda variável explicativa.

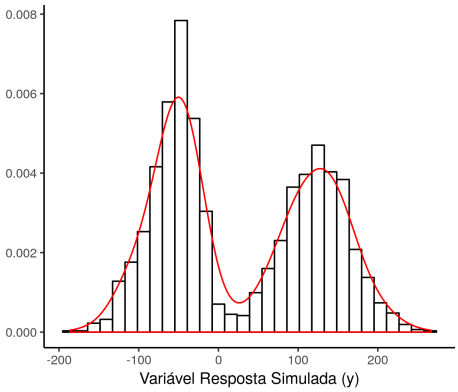


Figura 33 – Simulação 7: histograma da variável resposta simulada.

Tabela 29 – Simulação 7: critérios de informação - *DIC* e *EBIC*.

K	1	2	3	4	5
DIC	24030.19	21327.00	19700.34	2589100.00	3725428.00
EBIC	24065.06	21376.03	19433.97	1926637.00	2426899.00

apresentamos nela o intervalo de credibilidade de máxima densidade *a posteriori* de 95% e o diagnóstico de Geweke.

Com a excessão de alguns poucos parâmetros, alcançamos boas estimativas pontuais para os parâmetros. Além disso, os intervalos HPD de 95% contêm os valores verdadeiros para a maioria dos parâmetros. O algoritmo classificou corretamente nas componentes 96.85% das observações, o que consideramos como um ótimo ajuste aos dados. Os valores das estatísticas de Geweke estão contidos no intervalo $(-2,2)$ para a maioria dos parâmetros, indicando convergência. Na Figura 34, apresentamos os gráficos de traço dos valores gerados, que também indicam convergência.

Tabela 30 – Simulação 7: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke	Estimativa EM
β_{01}	5	3.85	$(1.24, 6.58)$	1.40	4.85
β_{11}	3	2.99	$(2.92, 3.05)$	-1.37	2.98
β_{11}	-4	-3.94	$(-4.06, -3.82)$	-1.37	-3.98
β_{02}	0	-18.47	$(-24.24, -12.48)$	-0.29	-0.26
β_{12}	-8	-7.85	$(-7.98, -7.71)$	0.73	-7.94
β_{22}	0.7	1.54	$(1.27, 1.84)$	0.26	0.68
β_{03}	-16	-5.46	$(-17.55, 6.70)$	2.51	-20.97
β_{13}	8	7.83	$(7.52, 8.13)$	-0.69	7.93
β_{23}	3	2.57	$(1.96, 3.14)$	-2.63	3.29
w_1	0.2	0.19	$(0.17, 0.21)$	0.29	0.19
w_2	0.3	0.33	$(0.31, 0.35)$	0.85	0.33
w_3	0.5	0.48	$(0.46, 0.50)$	-1.08	0.48
σ_1^2	5	4.93	$(4.18, 5.72)$	-0.48	4.85
σ_2^2	50	51.47	$(45.61, 57.98)$	0.55	47.75
σ_3^2	500	483.14	$(439.63, 527.28)$	0.72	479.34

4.1.7.2 Estimação via EM

Cinco modelos com diferentes números de componentes, $K = 1, ..., 5$, foram ajustados via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools* implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos pelo pacote.

Os valores do *AIC* e do *BIC* são encontrados na Tabela 31. Note que o modelo com $K = 5$ componentes obteve o menor valor para o *AIC* e o modelo com $K = 3$ componentes obteve o menor valor para o *BIC*. Mais uma vez verificando a superioridade do critério *BIC* para a seleção do número de variáveis quando os modelos são ajustados via algoritmo EM.

Tabela 31 – Simulação 7: critérios de informação - *AIC* e *BIC*.

K	1	2	3	4	5
AIC	inf	10990.20	9360.07	10592.41	9357.60
BIC	inf	11012.60	9396.48	10642.81	9422.01

Adotando o modelo de mistura de regressão com $K = 3$ componentes, as estimativas pontuais obtidas são apresentadas na última coluna da Tabela 30 e podemos observar boa estimação dos parâmetros. O algoritmo EM classificou corretamente 97% das observações.

Considerando um modelo de mistura de regressão com $K = 3$ componentes, também realizamos a estimação via algoritmo EM para diferentes valores iniciais e para valores iniciais muito distantes dos valores reais dos parâmetros e o algoritmo EM não apresentou boas estimativas, constatando novamente a sensibilidade do algoritmo aos pontos iniciais.

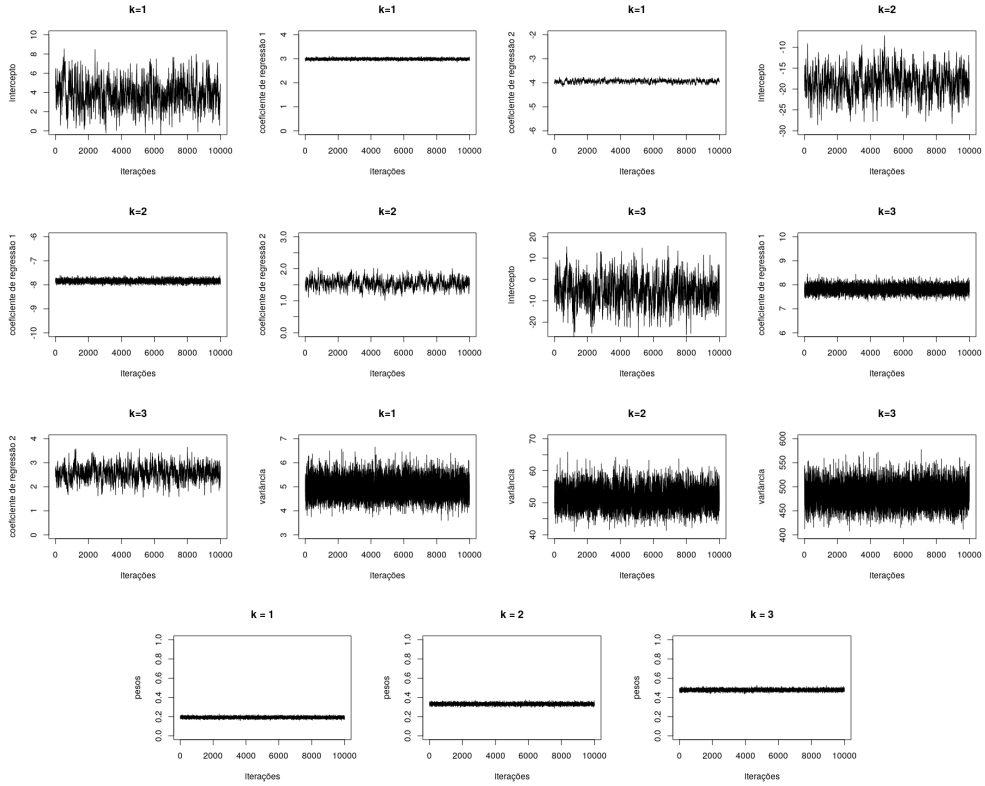


Figura 34 – Simulação 7: gráficos de traço - *Gibbs Sampling*.

4.1.7.3 Estimação via DDRJ

O processo de estimação via DDRJ e as amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência.

A Tabela 32 apresenta os valores obtidos para a probabilidade *a posteriori* de K. O máximo valor *a posteriori* é obtido em $K = 5$, porém duas delas apresentam estimativas para os pesos iguais a zero, tornando o modelo de mistura estimado com $K = 3$ componentes. Assumindo

Tabela 32 – Simulação 7: probabilidade *a posteriori* para K.

K	1	2	3	4	5
$P(K\|\cdots)$	0.0	0.0	0.0	0.0	1.0

o modelo de mistura de regressão com $K = 3$ componentes, temos na Tabela 33 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais. Também apresentamos nela o intervalo de credibilidade de máxima densidade *a posteriori* de 95% e o diagnóstico de Geweke. Podemos notar na Tabela 33 que, com exceção do intercepto da segunda componente, o algoritmo DDRJ encontrou boas estimativas para os parâmetros do modelo. Além disso, os intervalos HPD de 95% contêm os valores verdadeiros para a maioria dos parâmetros. Os valores das estatísticas de Geweke estão contidos no intervalo $(-2,2)$ para a maioria dos parâmetros, indicando convergência. Na Figura 35, apresentamos os gráficos de traço dos valores gerados, que também indicam convergência.

Tabela 33 – Simulação 7: estimativas para os parâmetros - DDRJ.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	5	3.83	$(1.16, 6.35)$	-1.42
β_{11}	3	2.99	$(2.92, 3.05)$	1.35
β_{21}	-4	-3.94	$(-4.06, -3.82)$	2.37
β_{02}	0	-18.32	$(-24.96, -11.85)$	2.16
β_{12}	-8	-7.84	$(-7.98, -7.71)$	-0.97
β_{22}	0.7	1.53	$(1.22, 1.85)$	-2.19
β_{03}	-16	-15.50	$(-137.78, 89.80)$	0.70
β_{13}	8	7.70	$(6.30, 9.07)$	0.60
β_{23}	3	3.14	$(-1.58, 8.62)$	-0.78
w_1	0.2	0.21	$(0.17, 0.23)$	1.63
w_2	0.3	0.34	$(0.31, 0.36)$	-2.00
w_3	0.5	0.45	$(0.24, 0.49)$	0.69
σ_1^2	5	4.98	$(4.16, 5.71)$	1.02
σ_2^2	50	51.29	$(44.75, 57.63)$	-1.83
σ_3^2	500	454.15	$(354.73, 532.66)$	0.69

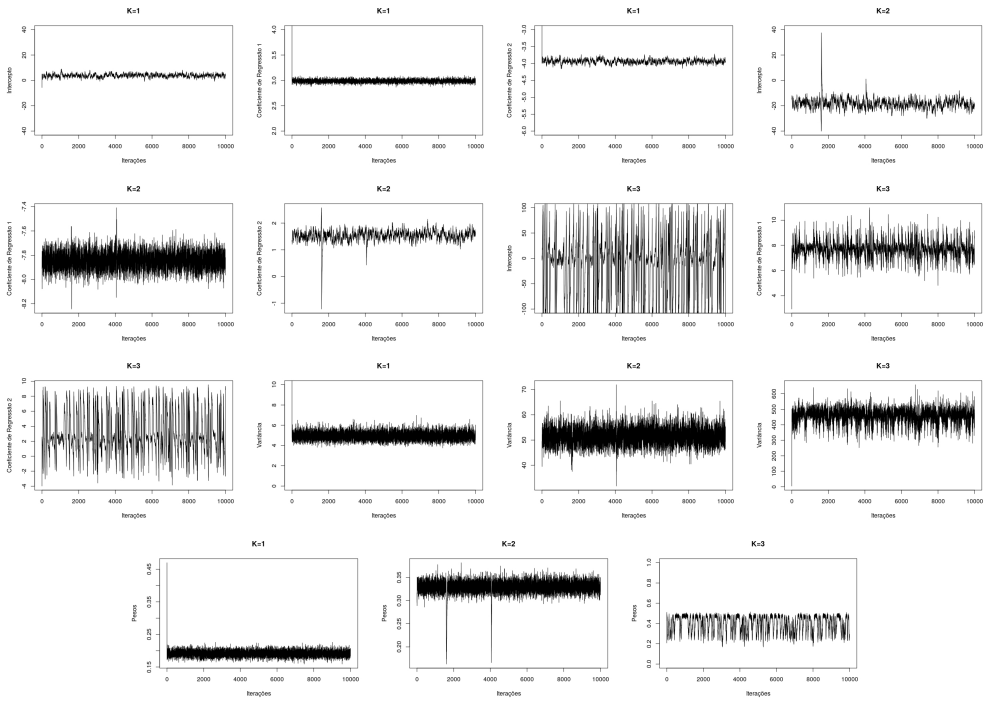


Figura 35 – Simulação 7: gráficos de traço - DDRJ.

4.1.8 Conjunto de Dados Simulado 8

Na geração do oitavo conjunto de dados simulado, utilizamos o mesmo processo da seção anterior, no entanto, alteramos os valores dos betas para $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = (5, 3, -4)^T$.

A Figura 36 representa os gráficos de dispersão entre a variável resposta e as covariáveis: os pontos em rosa representam as observações simuladas oriundas da componente 1, os pontos em verde as observações simuladas oriundas da componente 2 e os pontos em azul as observações

simuladas oriundas da componente 3. As componentes do conjunto de dados simulados 8 se diferenciam apenas pelo valor das suas variâncias. Na Figura 37, temos o histograma da variável resposta simulada, y, e sua densidade estimada.

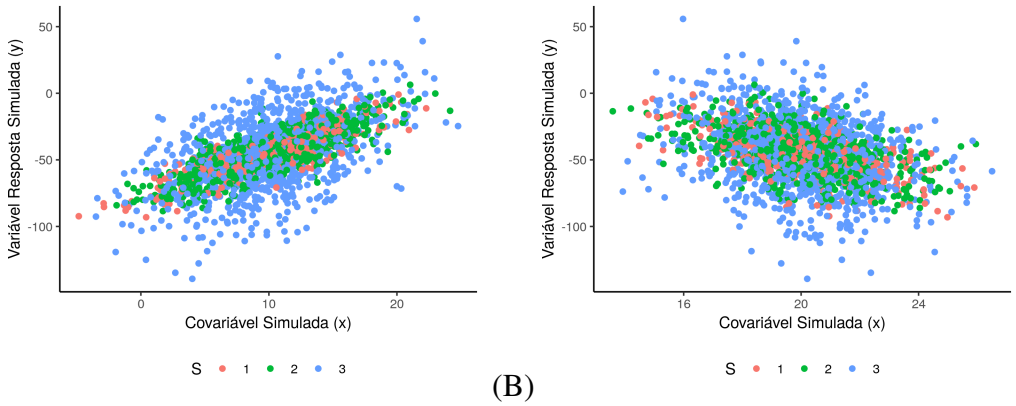


Figura 36 – Simulação 8: gráfico de dispersão. (A) gráfico de dispersão da variável resposta com a primeira variável explicativa. (B) gráfico de dispersão da variável resposta com a segunda variável explicativa.

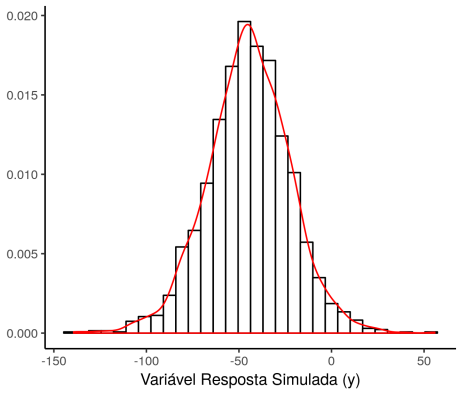


Figura 37 – Simulação 8: histograma da variável resposta simulada.

4.1.8.1 *Estimação via Gibbs Sampling, DIC e EBIC*

Seguimos o mesmo processo de estimação descrito nas seções anteriores. Encontramos os valores do *DIC* e do *EBIC* na Tabela 34. Temos que os dois critérios selecionam o modelo com $K = 1$ componente como o modelo com melhor ajuste. Apesar dos dados simulados terem sido gerados de uma mistura de regressão com $K = 3$, de fato, o modelo de regressão linear geral parece ser a melhor escolha para prever a resposta média.

Tabela 34 – Simulação 8: critérios de informação - *DIC* e *EBIC*.

K	1	2	3	4	5
DIC	16720.27	17811.18	18264.07	324570.6	1457287
EBIC	16754.39	18073.49	28446.25	186300.9	767061

Adotamos, então, o modelo de mistura de regressão com $K = 1$ componente. Temos na Tabela 35 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais. Também apresentamos nela o intervalo de credibilidade de HPD de 95% e o diagnóstico de Geweke.

Tabela 35 – Simulação 8: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke	Estimativa EM
β_{01}	5	2.55	$(-4.27, 9.27)$	-0.05	4.68
β_{11}	3	3.00	$(2.85, 3.16)$	0.26	2.98
β_{11}	-4	-3.88	$(-4.20, -3.56)$	0.04	-3.97
β_{02}	5	-	-	-	-
β_{12}	3	-	-	-	-
β_{22}	-4	-	-	-	-
β_{03}	5	-	-	-	1.12
β_{13}	3	-	-	-	3.01
β_{23}	-4	-	-	-	-3.81
w_1	0.2	1.00	-	-	0.47
w_2	0.3	-	-	-	-
w_3	0.5	-	-	-	0.53
σ_1^2	5	249.77	$(234.74, 265.68)$	0.2815	18.94
σ_2^2	50	-	-	-	-
σ_3^2	500	-	-	-	451.13

Assim como na simulação 6 apresentada na Seção 4.1.6, o modelo estima bem os coeficientes de regressão.

Os valores das estatísticas do diagnóstico de Geweke estão contidos no intervalo $(-2, 2)$ para todos os parâmetros, indicando convergência. Na Figura 38, apresentamos os gráficos de traço dos valores gerados, que também indicam convergência.

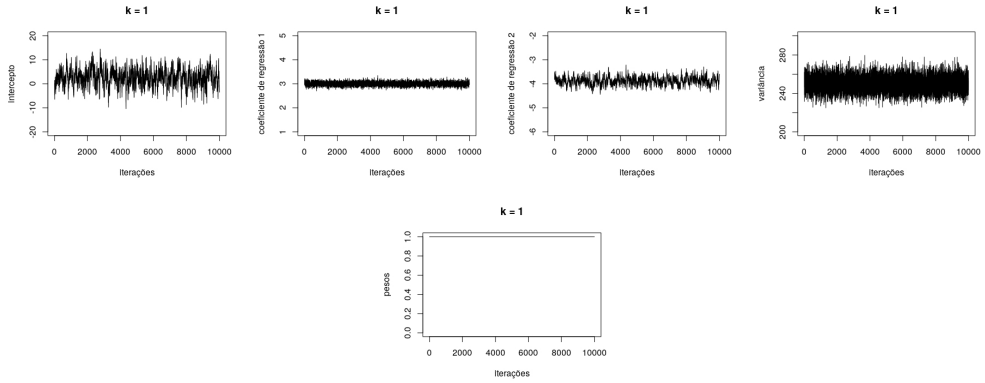


Figura 38 – Simulação 8: gráficos de traço - *Gibbs Sampling*.

4.1.8.2 Estimação via EM

Assim como nas seções anteriores, cinco modelos com diferentes números de componentes, $K = 1, ..., 5$, foram ajustados via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools* implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos pelo pacote.

Os valores do *AIC* e do *BIC* são encontrados na Tabela 36. Note que o modelo com $K = 5$ componentes obteve o menor valor para o *AIC* e o modelo com $K = 2$ componentes obteve o menor valor para o *BIC*. Dada a superioridade do *BIC* nas demais simulação, adotamos o modelo de mistura de regressão com $K = 2$ componentes e as estimativas pontuais obtidas são apresentados na última coluna da Tabela 35. Note que, assim como o algoritmo *Gibbs Sampling*, o algoritmo EM estima bem os coeficientes de regressão.

Tabela 36 – Simulação 8: critérios de informação - *AIC* e *BIC*.

K	1	2	3	4	5
AIC	inf	8057.30	8047.75	8049.84	8041.82
BIC	inf	8079.70	8084.16	8100.25	8106.23

4.1.8.3 Estimação via DDRJ

O processo de estimação via DDRJ e as amostras MCMC foram geradas respeitando as mesmas características das simulações anteriores, bem como a análise de convergência.

A Tabela 37 apresenta os valores obtidos para a probabilidade *a posteriori* de K. O máximo valor *a posteriori* é obtido em $K = 1$.

Tabela 37 – Simulação 8: probabilidade *a posteriori* para K.

K	1	2	3	4	5
$P(K\|\cdots)$	1.0	0.0	0.0	0.0	0.0

Temos na Tabela 38 a comparação entre os valores reais dos parâmetros e suas estimativas pontuais. Também apresentamos nela o intervalo de credibilidade de HPD de 95% e o diagnóstico de Geweke.

Podemos notar na Tabela 38 que os coeficientes de regressão são bem estimados. Além disso, as estimativas são bem próximas das estimativas obtidas via algoritmo *Gibbs Sampling*.

Os valores das estatísticas do diagnóstico de Geweke estão contidos no intervalo $(-2, 2)$ para todos os parâmetros, indicando convergência. Na Figura 39, apresentamos os gráficos de traço dos valores gerados, que também indicam convergência.

Tabela 38 – Simulação 8: estimativas para os parâmetros - DDRJ.

Parâmetro	Valor Real	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	5	2.54	$(-4.27, 8.99)$	0.41
β_{11}	3	3.00	$(2.84, 3.14)$	-0.51
β_{11}	-4	-3.88	$(-4.20, -3.56)$	-0.43
β_{02}	5	-	-	-
β_{12}	3	-	-	-
β_{22}	-4	-	-	-
β_{03}	5	-	-	-
β_{13}	3	-	-	-
β_{23}	-4	-	-	-
w_1	0.2	1.00	-	-
w_2	0.3	-	-	-
w_3	0.5	-	-	-
σ_1^2	5	249.92	$(234.36, 265.52)$	1.43
σ_2^2	50	-	-	-
σ_3^2	500	-	-	-

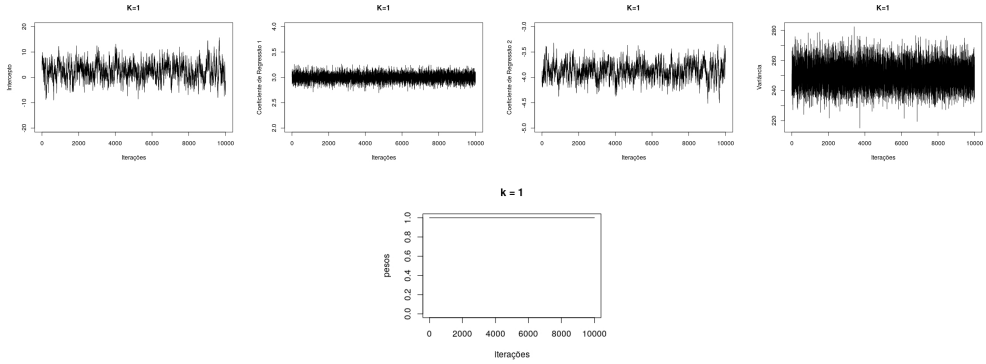


Figura 39 – Simulação 8: gráficos de traço - DDRJ.

Os resultados obtidos com o DDRJ foram, em geral, ruins principalmente na escolha e seleção do número de componentes. Os possíveis motivos e soluções para os problemas encontrados serão melhor discutidos na conclusão desse trabalho.

4.2 Conjunto de Dados Reais

Com o objetivo de contribuir com a discussão de políticas públicas em prol da educação básica brasileira, aplicaremos as metodologias descrita neste capítulo à dados reais referentes ao nível de qualidade da educação brasileira.

4.2.1 Descrição do Dados

4.2.1.1 Índice de Desenvolvimento da Educação Básica

Criado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) em 2007, o Índice de Desenvolvimento da Educação Básica (IDEB) é o principal indicador de qualidade da educação básica brasileira.

O IDEB leva em consideração duas variáveis importantes para a qualidade do ensino: fluxo e aprendizado. O fluxo (F) é representado pela taxa de aprovação dos alunos, variando de 0 a 100%. Este indicador é obtido no Censo Escolar realizado todos os anos pelo Inep. O aprendizado (A) é constituído pela média obtida pelos alunos nas avaliações de desempenho do Inep, a Prova Brasil e o Sistema de Avaliação da Educação Básica (SAEB), variando de 0 a 10. A Prova Brasil é utilizada para o cálculo dos IDEBs de escolas e municípios e o SAEB é utilizado para o cálculo dos IDEBs dos estados e nacional.

Este índice está numa escala que vai de zero a dez e quanto mais próximo de dez, melhor é a qualidade da educação. Obtemos o IDEB multiplicando a taxa de aprovação pela média obtida nas avaliações de desempenho, isto é,

$$\text{IDEB}_{ji} = A_{ji}F_{ji}, \tag{4.1}$$

sendo j o índice que identifica a unidade de interesse, i o ano de aplicação das provas e do censo escolar, A_{ji} a nota média da unidade j no ano i e F_{ji} a taxa de aprovação da unidade j no ano i. Devido a esta formulação, para melhorar o valor do IDEB é necessário que o sistema de ensino em questão melhore simultaneamente as duas dimensões do indicador. Caso o sistema de ensino retenha mais seus alunos para melhorar as notas obtidas nas avaliações, o fator fluxo terá valores mais baixos impossibilitando uma melhora no indicador do IDEB. Por outro lado, se o sistema de ensino apressar a aprovação de seus alunos para melhorar a taxa de aprovação, provavelmente as notas obtidas pelos alunos vão diminuir, novamente impossibilitando uma melhora no indicador do IDEB.

As avaliações de desempenho do Inep são aplicadas a cada dois anos aos alunos do 5º e 9º ano do ensino fundamental e do 3º ano do ensino médio. Portanto, o IDEB é calculado a cada dois anos e separado em três níveis: anos iniciais do ensino fundamental, anos finais do ensino fundamental e ensino médio.

O Brasil possui 5.570 municípios mais o Distrito Federal. Destes, 4.690 possuem IDEBs iniciais e finais para todos os anos apurados e divulgados até hoje (2007, 2009, 2011, 2013 e 2014). A Figura 40 representa a distribuição estimada pelo método de Kernel dos IDEBs iniciais dos 4.690 municípios brasileiros em todos os anos e a Figura 41 representa a distribuição estimada pelo método de Kernel dos IDEBs finais.

Por ser um índice comparável nacionalmente, o IDEB é um condutor de políticas públicas

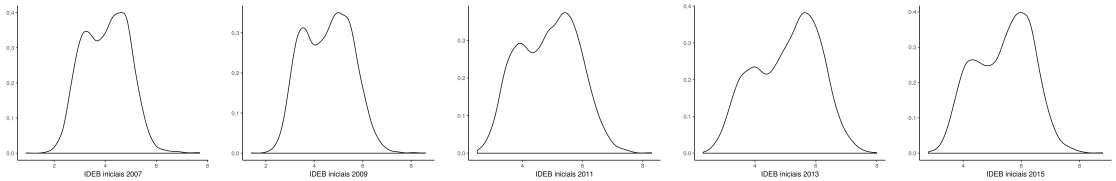


Figura 40 – Distribuição estimada pelo método de Kernel do IDEB inicial, ou seja, IDEB de provas de desempenho aplicadas aos alunos de 5º ano do ensino fundamental, para os anos de 2007, 2009, 2011, 2013 e 2015.

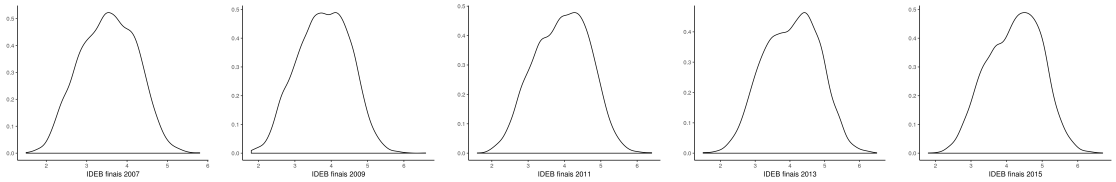


Figura 41 – Distribuição estimada pelo método de Kernel do IDEB final, ou seja, IDEB de provas de desempenho aplicadas aos alunos de 9º ano do ensino fundamental, para os anos de 2007, 2009, 2011, 2013 e 2015.

educacionais no Brasil. Devido à tamanha importância desse indicador e por sua distribuição ser, aparentemente, um modelo de mistura, este trabalho busca aplicar a metodologia proposta para estudar os fatores socioeconômicos e demográficos que o influenciam, bem como eles se relacionam, contribuindo assim com a discussão de políticas públicas em prol da educação brasileira.

4.2.1.2 Atlas do Desenvolvimento Humano no Brasil

O Atlas do Desenvolvimento Humano no Brasil é uma plataforma digital de consulta do IDHM e de mais de 200 indicadores de demografia, educação, renda, trabalho, habitação e vulnerabilidade, com dados extraídos dos Censos Demográficos de 1991, 2000 e 2010 de 5.565 municípios brasileiros. O Atlas é realizado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD), pelo Instituto de Pesquisa Econômica Aplicada (Ipea) e pela Fundação João Pinheiro (FJP).

É com base neste portal que extraímos nossas covariáveis de interesse. As covariáveis utilizadas neste estudo foram:

1) IDHM: O Índice de Desenvolvimento Humano Municipal (IDHM) é um indicador de qualidade de vida composto por três dimensões: longevidade, educação e renda. Segundo o PNUD, o IDHM adequa a metodologia do Índice de Desenvolvimento Humano (IDH) global à realidade brasileira e aos dados disponíveis. O IDHM varia entre 0 e 1: quanto mais próximo de 1, maior é o desenvolvimento humano da região. A Figura 42 mostra como é distribuído o IDHM de 2010 para os 5.109 municípios brasileiros que possuem IDEB de 2011 iniciais e finais. Por essa figura observamos que o IDHM também parece ser proveniente de um modelo de mistura;

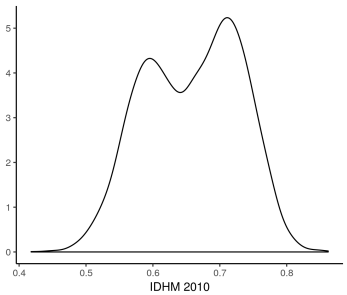


Figura 42 – Distribuição estimada pelo método de Kernel do IDHM de 2010 para os 5.109 municípios brasileiros que possuem IDEB de 2011 iniciais e finais.

2) MORT1: é um indicador de mortalidade infantil, representando o número de crianças que não deverão sobreviver ao primeiro ano de vida em cada 1000 crianças nascidas vivas. Podemos ver na Figura 43 como esse indicador de mortalidade infantil de 2010 é distribuído para os 5.109 municípios brasileiros que possuem IDEB de 2011 iniciais e finais;

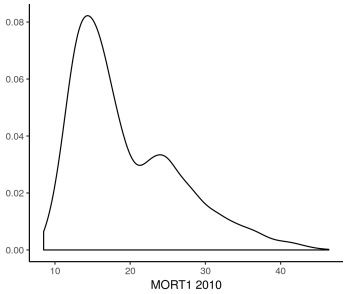


Figura 43 – Distribuição estimada pelo método de Kernel do índice MORT1 de 2010 para os 5.109 municípios brasileiros que possuem IDEB 2011 iniciais e finais.

3) RDPC: representa a renda per capita média da região. É a razão entre o somatório da renda de todos os indivíduos residentes em domicílios particulares permanentes e o número total desses indivíduos de uma específica região. Os valores, em Reais, são de 01/agosto de 2010. A Figura 44 mostra como é distribuída a renda per capita média de 2010 para os 5.109 municípios brasileiros que possuem IDEB de 2011 iniciais e finais.

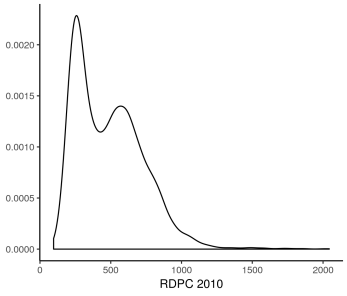


Figura 44 – Distribuição estimada pelo método de Kernel da renda per capita média de 2010 para os 5.109 municípios brasileiros que possuem IDEB 2011 iniciais e finais.

4) GINI: desenvolvido pelo estatístico e demógrafo italiano Corrado Gini, o índice de GINI mede o grau de desigualdade, isto é, o grau de concentração de renda de determinada região. O índice varia entre 0 e 1, onde 0 representa a completa igualdade e 1 a completa desigualdade. A Figura 45 mostra como é distribuído o índice de GINI de 2010 para os 5.109 municípios brasileiros que possuem IDEB de 2011 iniciais e finais. Ao contrário das covariáveis analisadas anteriormente, o GINI parece ter uma distribuição simétrica em torno da sua média;

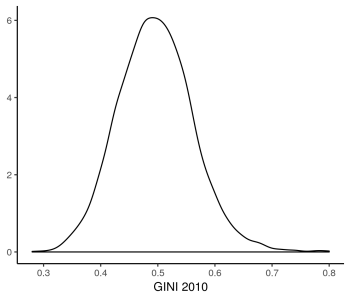


Figura 45 – Distribuição estimada pelo método de Kernel do índice Gini de 2010 para os 5.109 municípios brasileiros que possuem IDEB 2011 iniciais e finais.

É importante ter clareza de que todos estes índices são calculados de forma indireta por meio do Censo Demográfico realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

4.2.2 Aplicação

Seja $n = 5109$ o número de municípios que possuem IDEBs iniciais e finais do ano de 2011. Considere $(y_1, z_1, \boldsymbol{x}_1), \ldots, (y_n, z_n, \boldsymbol{x}_n)$ de tal forma que y_i é o valor do IDEB de 2011 referente aos anos iniciais do ensino fundamental para o município i, z_i é o valor do IDEB de 2011 referente aos anos finais do ensino fundamental para o município i e $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})$ é o vetor de covariáveis do município i, onde x_{i1} é o valor do IDHM, x_{i2} o valor do GINI, x_{i3} o valor do indicador de mortalidade infantil, MORT1, e x_{i4} é o valor da renda per capita média, RDPC, todos referentes ao ano de 2010 e ao município i.

Queremos, não somente saber como os IDEBs finais e iniciais se relacionam com as covariáveis descritas acima, mas também avaliar se essa relação é única para todo o Brasil. Dessa maneira, vamos assumir que tanto o IDEB para os anos iniciais como o IDEB para os anos finais condicionados às covariáveis de interesse são misturas de regressões normais com K_1 e K_2 componentes, respectivamente, isto é,

$$Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i \sim \sum_{k=1}^{K_1} w_k N(\boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2) \tag{4.2}$$

e

$$Z_i|\boldsymbol{X}_i = \boldsymbol{x}_i \sim \sum_{k=1}^{K_2} w'_k N(\boldsymbol{x}_i\boldsymbol{\delta}_k, \alpha_k^2). \tag{4.3}$$

No modelo (4.2), temos que $\boldsymbol{\beta}_k = (\beta_{0k}, ..., \beta_{4k})$, σ_k^2 e w_k são, respectivamente, os coeficientes de regressão, a variância e o peso da componente k, para $k = 1, ..., K_1$, e no modelo (4.3), temos que $\boldsymbol{\delta}_k = (\delta_{0k}, ..., \delta_{4k})$, α_k^2 e w'_k são, respectivamente, os coeficientes de regressão, a variância e o peso da componente k, para $k = 1, ..., K_2$.

4.2.2.1 Estimação via Gibbs Sampling, DIC e EBIC

Para cada uma das relações, cinco modelos com diferentes números de componentes, $K_1 = 1, \ldots, 5$ e $K_2 = 1, \ldots, 5$, foram estimados pelo algoritmo *Gibbs Sampling* com passos discutidos na Seção 3.2.1. Os valores dos hiperparâmetros das distribuições *a priori* escolhidos foram: $\gamma_1 = ... = \gamma_K = 1$, $\mu_{\beta_{jk}} = 0$, $\sigma^2_{\beta_{jk}} = 100$ e $a_k = b_k = 0.1$, para $j = 0, 1, ..., p$ e $k = 1, ..., K$. Como discutido anteriormente, como a variância das distribuições *a priori* especificadas são altas, essas distribuições são vagas e trazem pouca informação sobre os valores *a priori* dos parâmetros.

Para cada modelo rodamos 170000 iterações, descartamos as 100000 primeiras como *burn-in* e registramos uma a cada 7 iterações. As amostras MCMC que apresentaram problema de *label switching* foram renomeadas usando o algoritmo ECR. Os resultados apresentados a seguir são então oriundos de uma amostra MCMC final de tamanho 10000.

A Tabela 39 se refere a estimação do número de componentes, K_1, do Modelo (4.2). Note que, o modelo de mistura com $K_1 = 1$ componente apresentou o menor valor para o *DIC* e para o *EBIC* e, portanto, refutamos a suposição de que os municípios brasileiros apresentam heterogeneidade na relação entre o IDEB referente aos anos iniciais e as covariáveis apresentadas. As estimativas para os parâmetros são apresentadas na Tabela 40, juntamente com os intervalos de credibilidade de HPD de 95% e o diagnóstico de Geweke para verificarmos a convergência. Com estes resultados, o modelo ajustado é dado por:

$$\hat{y}_i = 4.07 + 4.37x_{i1} - 2.89x_{i2} - 0.04x_{i3} + 0.0002x_{i4}. \tag{4.4}$$

Tabela 39 – IDEBs iniciais: critérios de informação - *DIC* e *EBIC*.

K_1	1	2	3	4	5
DIC	10339.87	53258.55	149750.90	254012.80	221810.80
EBIC	10393.57	31963.40	80273.36	132498.9	116410.1

De forma análoga, observamos na Tabela 41 que o modelo com $K_2 = 1$ também é o que melhor se ajusta aos dados, descartando a possibilidade de heterogeneidade na relação entre o

Tabela 40 – IDEBs iniciais: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	4.07	$(3.56, 4.71)$	0.30
β_{11}	4.37	$(3.51, 5.14)$	-0.32
β_{21}	-2.89	$(-3.21, -2.58)$	-0.16
β_{31}	-0.04	$(-0.05, -0.04)$	-0.57
β_{41}	0.0002	$(0.0000, 0.0004)$	0.37
w_1	1.00	-	-
σ_1^2	0.44	$(0.43, 0.46)$	-1.76

IDEB para os anos finais do ensino fundamental e as covariáveis. Os parâmetros estimados são apresentados na Tabela 42 juntamente com os intervalos de HPD de 95% e o diagnóstico de Geweke. O modelo ajustado para os anos finais é dado por:

$$\hat{z}_i = 2.74 + 3.68x_{i1} - 1.53x_{i2} - 0.03x_{i3} + 0.0001x_{i4}. \tag{4.5}$$

Tabela 41 – IDEBs finais: critérios de informação - *DIC* e *EBIC*.

K_2	1	2	3	4	5
DIC	8545.42	21223.68	44355.88	25870.96	37667.09
EBIC	8599.16	14999.10	26626.00	17444.26	23403.13

Tabela 42 – IDEBs finais: estimativas para os parâmetros - *Gibbs Sampling*.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. Geweke
δ_{01}	2.74	$(2.33, 3.24)$	0.59
δ_{11}	3.68	$(2.99, 4.26)$	-0.52
δ_{21}	-1.53	$(-1.79, -1.26)$	-0.73
δ_{31}	-0.03	$(-0.03, -0.02)$	-2.34
δ_{41}	0.0001	$(-0.0001, 0.0002)$	0.50
w_1	1.00	-	-
α_1^2	0.31	$(0.30, 0.32)$	0.95

O sinal das estimativas dos parâmetros se deu como esperado: β_{11}, β_{41}, δ_{11} e δ_{41} são positivos, indicando que uma melhora na qualidade de vida da população e/ou no nível de renda da população refletem numa melhora do IDEB; e β_{21}, β_{31}, δ_{21} e δ_{31} são negativos indicando que um aumento na desigualdade social ou a piora nos níveis de saneamento (aumento da taxa de mortalidade) induzem a uma piora do IDEB.

Das Equações (4.4) e (4.5), também podemos inferir que:

1) mantendo todas as demais variáveis constantes, um aumento de 0.1 no IDHM gera um acréscimo de 0.437 na nota média do IDEB dos anos iniciais e de 0.368 na nota média do IDEB dos anos finais;

2) a *ceteris paribus*, a redução de 0.1 no índice de Gini gera um acréscimo de 0.289 na nota média do IDEB dos anos iniciais e de 0.153 na nota média do IDEB dos anos finais;

3) a redução de 10 pontos percentuais na taxa de mortalidade infantil, implica num aumento de 0.004 na nota média do IDEB dos anos iniciais e de 0.003 na nota média do IDEB dos anos finais, quando todas as demais variáveis são mantidas constantes;

4) o aumento de *R*$100.00 na renda per capita média, implica no aumento de 0.02 na nota média do IDEB dos anos iniciais e de 0.01 na nota média do IDEB dos anos finais.

Podemos observar na Tabela 40 e na Tabela 42, que os valores das estatísticas do diagnóstico de Geweke estão entre $(-2,2)$, indicando convergência. Na Figura 46 apresentamos os gráficos de traço para o Modelo (4.2) e na Figura 46 para o Modelo (4.3), que também indicam convergência.

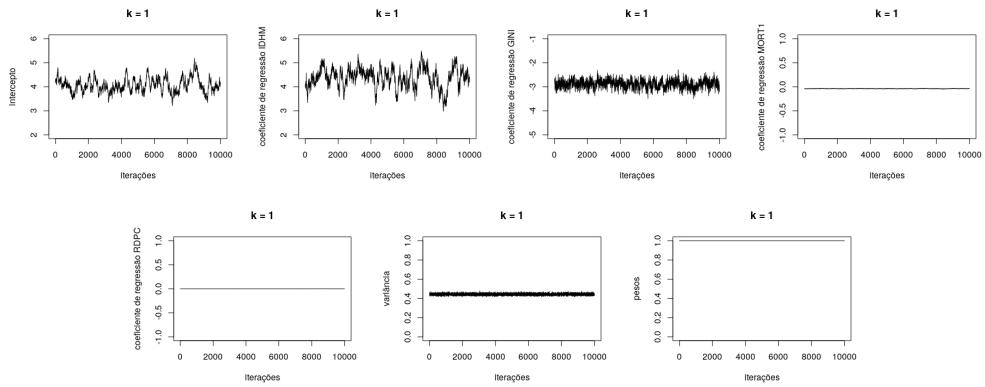


Figura 46 – IDEB anos iniciais: gráficos de traço - *Gibbs Sampling*.

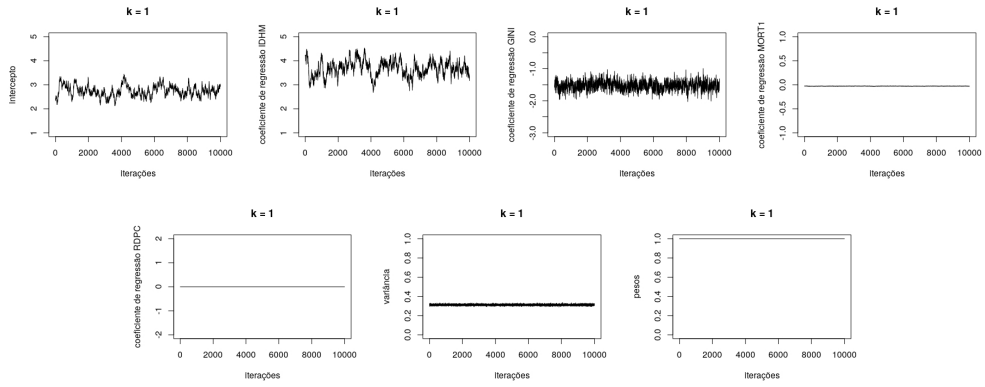


Figura 47 – IDEB anos finais: gráficos de traço - *Gibbs Sampling*.

4.2.2.2 Estimação via EM

Da mesma forma que na estimação via *Gibbs Sampling*, para cada uma das relações cinco modelos com diferentes números de componentes, $K = 1,...,5$, foram ajustados, mas,

agora, via algoritmo EM. Para este procedimento utilizamos o pacote *mixtools* implementado no software estatístico R. Consideramos como valores iniciais dos parâmetros os valores de *default* sugeridos pelo pacote. O algoritmo EM não conseguiu convergência para $K \geq 2$.

4.2.2.3 Estimação via DDRJ

Para cada uma das relações estimamos o modelo via algoritmo DDRJ com passos discutidos na Seção 3.3.3. Desta forma, tentamos estimar o valor de K conjuntamente com os parâmetros do modelo. Os valores dos hiperparâmetros das distribuições *a priori* escolhidos foram os mesmos adotados na estimação via *Gibbs Sampling*, isto é: $\gamma_1 = ... = \gamma_K = 1$, $\mu_{\beta_{jk}} = 0$, $\sigma^2_{\beta_{jk}} = 100$ e $a_k = b_k = 0.1$, para $j = 0, 1, ..., p$ e $k = 1, ..., K_{max}$. Como as variâncias das distribuições *a priori* especificadas são altas, essas distribuições são vagas e trazem pouca informação sobre os valores *a priori* dos parâmetros.

Rodamos 170000 iterações para cada relação, descartamos as 100000 primeiras como *burn-in* e registramos uma a cada 7 iterações. As amostras MCMC que apresentaram problema de *label switching* foram renomeadas usando o algoritmo ECR. Os resultados apresentados a seguir são então oriundos de uma amostra MCMC final de tamanho 10000.

A Tabela 43 apresenta os valores obtidos para a probabilidade *a posteriori* do número de componentes, K_1, do Modelo (4.2). O máximo valor *a posteriori* é obtido em $K_1 = 1$.

Tabela 43 – IDEB anos iniciais: probabilidade *a posteriori* para K_1.

K_1	1	2	3	4	5
$P(K_1 \mid \cdots)$	1.0	0.0	0.0	0.0	0.0

As estimativas para os parâmetros são apresentadas na Tabela 44, juntamente com os intervalos de credibilidade de HPD de 95% e o diagnóstico de Geweke para verificarmos a convergência. Podemos notar bastante similaridade com os resultados obtidos via *Gibbs Sampling*, no entanto, descartando a relação do IDEB com a renda per capita. Com estes resultados, o modelo estimado via DDRJ é dado por:

$$\hat{y}_i = 4.11 + 4.32x_{i1} - 2.90x_{i2} - 0.04x_{i3}. \tag{4.6}$$

De forma análoga, a Tabela 45 apresenta os valores obtidos para a probabilidade *a posteriori* do número de componentes, K_2, do Modelo (4.3). O máximo valor *a posteriori* é obtido em $K_2 = 1$.

Os parâmetros estimados são apresentados na Tabela 46 juntamente com os intervalos de HPD de 95% e o diagnóstico de Geweke e também podemos notar similaridade com a estimação realizada via *Gibbs Sampling* e novamente excluindo a relação entre o IDEB e a renda per capita

Tabela 44 – IDEB anos iniciais: estimativas para os parâmetros - DDRJ.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. Geweke
β_{01}	4.11	$(3.59, 4.62)$	-0.80
β_{11}	4.32	$(3.61, 5.02)$	0.80
β_{21}	-2.90	$(-3.22, -2.58)$	1.42
β_{31}	-0.04	$(-0.05, -0.04)$	1.43
β_{41}	0.00	-	-
w_1	1.00	-	-
σ_1^2	0.44	$(0.43, 0.46)$	-0.34

Tabela 45 – IDEB anos finais: probabilidade *a posteriori* para K_2.

K_2	1	2	3	4	5
$P(K_2 \mid \cdots)$	1.0	0.0	0.0	0.0	0.0

uma vez que estimativa pontual e intervalar foram basicamente zero. O modelo estimado para os anos finais via DDRJ é dado por:

$$\hat{z}_i = 2.70 + 3.75x_{i1} - 1.52x_{i2} - 0.03x_{i3}. \tag{4.7}$$

Tabela 46 – IDEBs finais: estimativas para os parâmetros - DDRJ.

Parâmetro	Estimativa	Int. Cred. 95%	Diag. Geweke
δ_{01}	2.70	$(2.21, 3.16)$	1.35
δ_{11}	3.75	$(3.15, 4.49)$	-1.52
δ_{21}	-1.52	$(-1.79, -1.26)$	-1.82
δ_{31}	-0.03	$(-0.03, -0.02)$	-1.77
δ_{41}	0.00	-	-
w_1	1.00	-	-
α_1^2	0.31	$(0.30, 0.32)$	-0.30

Podemos observar na Tabela 44 e na Tabela 46, que os valores das estatísticas do diagnóstico de Geweke estão entre $(-2, 2)$, indicando convergência. Na Figura 48 apresentamos os gráficos de traço para o Modelo (4.2) e na Figura 48 para o Modelo (4.3), que também indicam convergência.

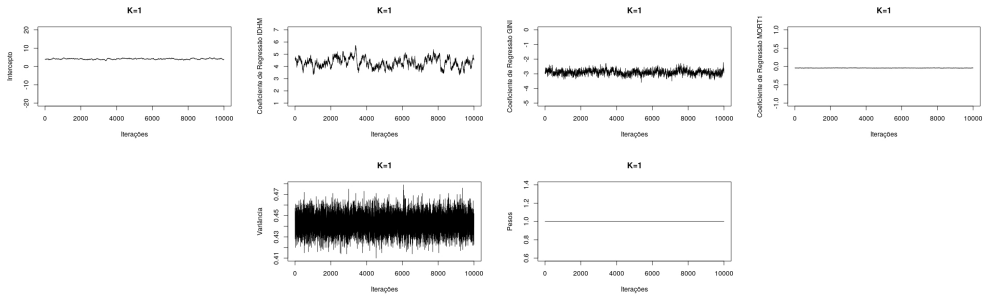


Figura 48 – IDEB anos iniciais: gráficos de traço - DDRJ.

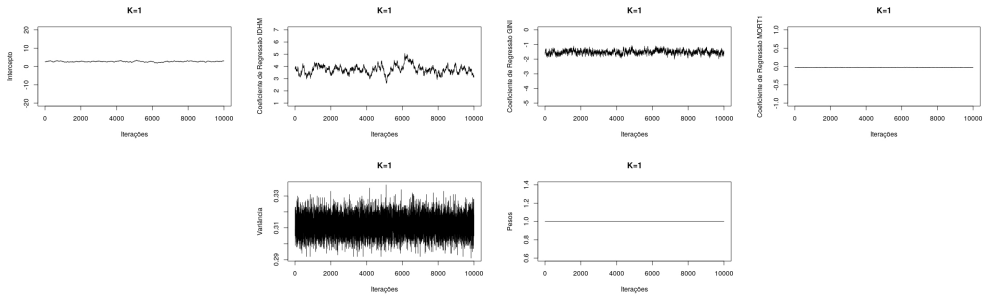


Figura 49 – IDEB anos finais: gráficos de traço - DDRJ.

CAPÍTULO

5

CONSIDERAÇÕES FINAIS E PERSPECTIVAS FUTURAS

Apresentamos neste trabalho duas ferramentas inferenciais, através da ótica Bayesiana, para modelos de mistura de regressões normais considerando o número de componentes da mistura desconhecido, e comparamos com o algoritmo EM desenvolvido sob a ótica frequentista.

O algoritmo *Gibbs Sampling* se mostrou eficaz quando as componentes da mistura são esparsas, porém, para dados poucos esparsos entre as componentes, a metodologia apresentou dificuldades de estimação e a utilização do critério *DIC* se mostrou menos estável e precisa do que a utlização do critério *EBIC*. Outro problema enfrentado é que, para inferir o valor do número de componentes, é necessário o ajuste de um número grande de modelos com diferentes valores de K e isso pode se tornar computacionalmente custoso dependendo do tamanho da amostra e número de covariáveis.

O algoritmo EM se mostrou eficaz e apresentou bons resultados mesmo para dados poucos esparsos entre as componentes quando a escolha dos valores iniciais eram otimizadas. No entanto, o algoritmo se mostrou muito sensível a estes valores iniciais, apresentando resultados ruins para valores iniciais diferentes dos valores otimizados. Além disso, para o cenário com mais covariáveis, como a aplicação em dados reais apresentada neste trabalho, o algoritmo EM não convergiu quando o número de componentes era superior a 2. O critério *AIC* tende a selecionar modelos que são super parametrizados e, portanto, na maioria dos casos selecionou os modelos com o número máximo de componentes, mostrando-se não adequado para a estimação de K em modelos de mistura de regressões normais. Já o critério *BIC* mostrou-se eficiente mesmo em cenários mais complicados de se identificar as componentes. Assim como o *Gibbs Sampling*, para inferir o valor do número de componentes é necessário o ajuste de um número grande de modelos com diferentes valores de K e isso pode se tornar computacionalmente custoso dependendo do tamanho da amostra e do número de covariáveis. Outra desvantagem do algoritmo EM é não fornecer estimativas intervalares para os parâmetros.

O algoritmo DDRJ não apresentou bons resultados mesmo em situações nas quais as diferentes componentes eram muito evidentes. No caso mais simples, onde as componentes não se cruzam, foi capaz de estimar bem os parâmetros do modelo, porém superestimou o número de componentes, mesmo atribuindo estimativas nulas para os pesos. Para o cenário onde as componentes se cruzam, o algoritmo não foi capaz de identificá-las, estimando sempre modelos de regressão linear. Como era esperado e muitas vezes já relatado na literatura, o algoritmo mostrou-se sensível à maneira como construímos os candidatos de *split* e *merge*. Vale ressaltar que testamos outras funções de transição para construir o modelo candidato e elas também não funcionaram bem.

Dados os resultados obtidos, este trabalho indica a utilização do método *Gibbs Sampling* combinado ao *DIC* e *EBIC*.

Para os dados do IDEB, as metodologias estudadas não identificaram componentes distintas no modelo de regressão, sendo o modelo de regressão geral o modelo escolhido para ajustar os dados. Conforme esperávamos, percebemos uma relação positiva entre o IDEB, IDHM e RDPC e uma relação negativa entre o IDEB, MORT1 e GINI. Os fatores com maiores influências sobre a qualidade de ensino da educação básica são o IDHM e o índice de Gini.

Como propostas futuras pretendemos aperfeiçoar o algoritmo DDRJ, buscando encontrar melhores propostas de *split* e de *merge* e acrescentar passos de nascimento e morte das componentes vazias. Vários autores tem trabalhado no aperfeiçoamento do DDRJ, principalmente na escolha da função de transição entre os modelos, para selecionar e estimar modelos de mistura (SARAIVA, 2009), (ZUANETTI; MILAN, 2017), (JAIN; NEAL, 2004) e obtido bons resultados. O objetivo desses autores é propor novos modelos de uma função de transição que torne a cadeia MCMC mais dinâmica e que mais rapidamente convirja. Além disso, podemos atualizar o modelo proposto algumas vezes, via passos de *Gibbs sampling*, antes de calcularmos a probabilidade de aceitação. Green e Mira (2001) propõem um algoritmo que, na rejeição, uma segunda tentativa de movimento é feita. Zuanetti e Milan (2016) aprimoraram o modelo proposto por 10 iterações antes de avaliarem a sua aceitação.

Os algoritmos de estimação desenvolvidos neste trabalho e os dados utilizados estão disponíveis em <https://github.com/lgfcotrim>.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 54.

____. A Bayesian analysis of the minimum AIC procedure. In: **Selected Papers of Hirotugu Akaike**. [S.l.]: Springer, 1998. p. 275–280. Citado na página 54.

BAI, X.; YAO, W.; BOYER, J. E. Robust fitting of mixture regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 7, p. 2347–2359, 2012. Citado na página 23.

BENAGLIA, T.; CHAUVEAU, D.; HUNTER, D.; YOUNG, D. mixtools: An R package for analyzing finite mixture models. 2009. Citado na página 54.

BISHOP, C. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006. ISBN 9780387310732. Citado na página 31.

CASELLA, G.; GEORGE, E. I. Explaining the Gibbs sampler. **The American Statistician**, Taylor & Francis, v. 46, n. 3, p. 167–174, 1992. Citado na página 24.

CHEN, J.; CHEN, Z. Extended Bayesian information criteria for model selection with large model spaces. **Biometrika**, Oxford University Press, v. 95, n. 3, p. 759–771, 2008. Citado na página 43.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings algorithm. **The American Statistician**, Taylor & Francis Group, v. 49, n. 4, p. 327–335, 1995. Citado na página 24.

COWLES, M. K.; CARLIN, B. P. Markov chain Monte Carlo convergence diagnostics: a comparative review. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 91, n. 434, p. 883–904, 1996. Citado nas páginas 34 e 53.

FRALEY, C.; RAFTERY, A. E. Model-based clustering, discriminant analysis, and density estimation. **Journal of the American Statistical Association**, Taylor & Francis, v. 97, n. 458, p. 611–631, 2002. Citado na página 31.

FRÜHWIRTH-SCHNATTER, S. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. **Journal of the American Statistical Association**, Taylor & Francis, v. 96, n. 453, p. 194–209, 2001. Citado na página 24.

____. **Finite mixture and Markov switching models**. [S.l.]: Springer Science & Business Media, 2006. Citado na página 23.

GOLDFELD, S. M.; QUANDT, R. E. A Markov model for switching regressions. **Journal of Econometrics**, Elsevier, v. 1, p. 3–15, 1973. Citado na página 23.

GREEN, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, Oxford University Press, v. 82, n. 4, p. 711–732, 1995. Citado na página 24.

GREEN, P. J.; MIRA, A. Delayed rejection in reversible jump Metropolis–Hastings. **Biometrika**, Biometrika Trust, v. 88, n. 4, p. 1035–1053, 2001. Citado na página 98.

HUANG, M.; YAO, W. Mixture of regression models with varying mixing proportions: a semiparametric approach. **Journal of the American Statistical Association**, Taylor & Francis, v. 107, n. 498, p. 711–724, 2012. Citado na página 23.

HURN, M.; JUSTEL, A.; ROBERT, C. P. Estimating mixtures of regressions. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 12, n. 1, p. 55–79, 2003. Citado nas páginas 24, 39 e 40.

JAIN, S.; NEAL, R. M. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. **Journal of Computational and Graphical Statistics**, v. 13, n. 1, 2004. Citado na página 98.

MCLACHLAN, G.; KRISHNAN, T. **The EM algorithm and extensions**. [S.l.]: John Wiley & Sons, 2007. v. 382. Citado na página 23.

MCLACHLAN, G.; PEEL, D. **Finite mixture models**. [S.l.]: New York : Wiley, c2000, 2000. Citado na página 23.

PAPASTAMOULIS, P. Handling the label switching problem in latent class models via the ECR algorithm. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 43, n. 4, p. 913–927, 2014. Citado na página 50.

____. label. switching: An r package for dealing with the label switching problem in mcmc outputs. **arXiv preprint arXiv:1503.02271**, 2015. Citado na página 50.

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**, Taylor & Francis, v. 19, n. 2, p. 313–331, 2010. Citado na página 50.

QIN, L.-X.; SELF, S. G. The clustering of regression models method with applications in gene expression data. **Biometrics**, Wiley Online Library, v. 62, n. 2, p. 526–533, 2006. Citado na página 24.

QUANDT, R. E.; RAMSEY, J. B. Estimating mixtures of normal distributions and switching regressions. **Journal of the American Statistical Association**, Taylor & Francis, v. 73, n. 364, p. 730–738, 1978. Citado na página 24.

SARAIVA, E. F. Modelo de mistura com número de componentes desconhecido: estimação via método split-merge. Universidade Federal de São Carlos, 2009. Citado na página 98.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. The deviance information criterion: 12 years on. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 76, n. 3, p. 485–493, 2014. Citado na página 42.

STEPHENS, M. Dealing with label switching in mixture models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 62, n. 4, p. 795–809, 2000. Citado nas páginas 34 e 50.

WANG, P.; PUTERMAN, M. L.; COCKBURN, I.; LE, N. Mixed Poisson regression models with covariate dependent rates. **Biometrics**, JSTOR, p. 381–400, 1996. Citado na página 24.

WEDEL, M.; DESARBO, W. S. A latent class binomial logit methodology for the analysis of paired comparison choice data. **Decision Sciences**, Wiley Online Library, v. 24, n. 6, p. 1157–1170, 1993. Citado na página 24.

YOUNG, D. S.; HUNTER, D. R.; CHAUVEAU, D.; BENAGLIA, T. mixtools: an r package for analyzing mixture models. **Journal of Statistical Software**, American Statistical Association, v. 32, n. 06, 2009. Citado na página 37.

ZUANETTI, D. A.; MILAN, L. A. Data-driven reversible jump for QTL mapping. **Genetics**, Genetics Soc America, v. 202, n. 1, p. 25–36, 2016. Citado na página 98.

____. A generalized mixture model applied to diabetes incidence data. **Biometrical Journal**, Wiley Online Library, v. 59, n. 4, p. 826–842, 2017. Citado na página 98.

APÊNDICE

A

TÓPICOS ADICIONAIS: MODELOS DE MISTURA DE DISTRIBUIÇÕES

A.1 Distribuição *a posteriori* condicional completa para as médias

Nesta seção, apresentamos o cálculo da distribuição *a posteriori* condicional completa para as médias de um modelo de mistura Gaussiana.

$$
\begin{aligned}
\pi(\mu_k|\cdots) &\propto L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x})\pi(\mu_k) \\
&= w_1^{n_1}\ldots w_K^{n_K} \\
&\times \prod_{i:S_i=1}\frac{1}{\sqrt{2\pi}\sigma_1}exp\left[-\frac{1}{2\sigma_1^2}(y_i-\mu_1)^2\right] \\
&\times \ldots \\
&\times \prod_{i:S_i=k}\frac{1}{\sqrt{2\pi}\sigma_k}exp\left[-\frac{1}{2\sigma_k^2}(y_i-\mu_k)^2\right] \\
&\times \ldots \\
&\times \prod_{i:S_i=K}\frac{1}{\sqrt{2\pi}\sigma_K}exp\left[-\frac{1}{2\sigma_K^2}(y_i-\mu_K)^2\right] \\
&\times \frac{1}{\sqrt{2\pi}\sigma_{\mu_k}}exp\left[-\frac{1}{2\sigma_{\mu_k}^2}\left(\mu_k-\mu_{\mu_k}\right)^2\right] \\
&\propto \prod_{i:S_i=k}\frac{1}{\sqrt{2\pi}\sigma_k}exp\left[-\frac{1}{2\sigma_k^2}(y_i-\mu_k)^2\right]\frac{1}{\sqrt{2\pi}\sigma_{\mu_k}}exp\left[-\frac{1}{2\sigma_{\mu_k}^2}\left(\mu_k-\mu_{\mu_k}\right)^2\right] \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma_k}\right)^{n_k}\left(\frac{1}{\sqrt{2\pi}\sigma_{\mu_k}}\right)exp\left[-\frac{1}{2\sigma_{\mu_k}^2}\left(\mu_k-\mu_{\mu_k}\right)^2\right]\prod_{i:S_i=k}exp\left[-\frac{1}{2\sigma_k^2}(y_i-\mu_k)^2\right] \\
&\propto exp\left[-\frac{1}{2\sigma_{\mu_k}^2}\left(\mu_k-\mu_{\mu_k}\right)^2\right]exp\left[-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-\mu_k)^2\right] \\
&= exp\left[-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-\mu_k)^2-\frac{1}{2\sigma_{\mu_k}^2}\left(\mu_k-\mu_{\mu_k}\right)^2\right] \\
&= exp\left[-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}\left(y_i^2-2y_i\mu_k+\mu_k^2\right)-\frac{1}{2\sigma_{\mu_k}^2}\left(\mu_k^2-2\mu_k\mu_{\mu_k}+\mu_{\mu_k}^2\right)\right] \\
&\propto exp\left[-\frac{1}{2\sigma_k^2}\left(-\sum_{i:S_i=k}2y_i\mu_k+\sum_{i:S_i=k}\mu_k^2\right)-\frac{1}{2\sigma_{\mu_k}^2}\left(\mu_k^2-2\mu_k\mu_{\mu_k}\right)\right] \\
&= exp\left[-\frac{1}{2\sigma_k^2}\left(-2\mu_k\sum_{i:S_i=k}y_i+n_k\mu_k^2\right)-\frac{1}{2\sigma_{\mu_k}^2}\left(\mu_k^2-2\mu_k\mu_{\mu_k}\right)\right] \\
&= exp\left[\left(\frac{\mu_k}{\sigma_k^2}\sum_{i:S_i=k}y_i\right)-\left(\frac{n_k\mu_k^2}{2\sigma_k^2}\right)-\left(\frac{\mu_k^2}{2\sigma_{\mu_k}^2}\right)+\left(\frac{\mu_k\mu_{\mu_k}}{\sigma_{\mu_k}^2}\right)\right] \\
&= \exp\left[-\frac{\mu_k^2}{2}\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)+\mu_k\left(\frac{\sum_{i:S_i=k}y_i}{\sigma_k^2}+\frac{\mu_{\mu_k}}{\sigma_{\mu_k}^2}\right)\right] \\
&= \exp\left[-\frac{1}{2}\left\{\mu_k^2\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)-2\mu_k\left(\frac{\sum_{i:S_i=k}y_i}{\sigma_k^2}+\frac{\mu_{\mu_k}}{\sigma_{\mu_k}^2}\right)\right\}\right] \\
&= \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)\left\{\mu_k^2-2\mu_k\left(\frac{\sum_{i:S_i=k}y_i}{\sigma_k^2}+\frac{\mu_{\mu_k}}{\sigma_{\mu_k}^2}\right)\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)^{-1}\right\}\right] \\
&\propto \exp\left[-\frac{1}{2\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)^{-1}}\left\{\mu_k-\left(\frac{\mu_{\mu_k}}{\sigma_{\mu_k}^2}+\frac{\sum_{i:S_i=k}y_i}{\sigma_k^2}\right)\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)^{-1}\right\}^2\right] \\
&\propto \frac{1}{\sqrt{2\pi\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)^{-1}}}\exp\left[-\frac{1}{2\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)^{-1}}\left\{\mu_k-\left(\frac{\mu_{\mu_k}}{\sigma_{\mu_k}^2}+\frac{\sum_{i:S_i=k}y_i}{\sigma_k^2}\right)\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)^{-1}\right\}^2\right] \\
&\propto N\left(\left(\frac{\mu_{\mu_k}}{\sigma_{\mu_k}^2}+\frac{\sum_{i:S_i=k}y_i}{\sigma_k^2}\right)\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)^{-1},\left(\frac{1}{\sigma_{\mu_k}^2}+\frac{n_k}{\sigma_k^2}\right)^{-1}\right)
\end{aligned}
$$

A.2 Distribuição *a posteriori* condicional completa para as variâncias

Nesta seção, apresentamos o cálculo da distribuição *a posteriori* condicional completa para as variâncias de um modelo de mistura Gaussiana.

$$
\begin{aligned}
\pi(\sigma_k^2|\cdots) &\propto L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x})\pi(\sigma_k^2) \\
&= w_1^{n_1}\ldots w_K^{n_K} \\
&\times \prod_{i:S_i=1}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{1}{2\sigma_1^2}(y_i-\mu_1)^2\right] \\
&\times \ldots \\
&\times \prod_{i:S_i=k}\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left[-\frac{1}{2\sigma_k^2}(y_i-\mu_k)^2\right] \\
&\times \ldots \\
&\times \prod_{i:S_i=K}\frac{1}{\sqrt{2\pi}\sigma_K}\exp\left[-\frac{1}{2\sigma_K^2}(y_i-\mu_K)^2\right] \\
&\times \frac{b_k^{a_k}}{\Gamma(a_k)}(\sigma_k^2)^{-a_k-1}\exp\left\{-\frac{b_k}{\sigma_k^2}\right\} \\
&\propto \left(\frac{1}{\sqrt{2\pi}\sigma_k}\right)^{n_k}\exp\left[-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-\mu_k)^2\right]\frac{b_k^{a_k}}{\Gamma(a_k)}(\sigma_k^2)^{-a_k-1}\exp\left\{-\frac{b_k}{\sigma_k^2}\right\} \\
&= \left(\frac{1}{2\pi\sigma_k^2}\right)^{n_k/2}\exp\left[-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-\mu_k)^2\right]\frac{b_k^{a_k}}{\Gamma(a_k)}(\sigma_k^2)^{-a_k-1}\exp\left\{-\frac{b_k}{\sigma_k^2}\right\} \\
&\propto \left(\sigma_k^2\right)^{-\frac{n_k}{2}-a_k-1}\exp\left[-\frac{b_k+\frac{1}{2}\sum_{i:S_i=k}(y_i-\mu_k)^2}{\sigma_k^2}\right] \\
&\propto IG\left(a_k+\frac{n_k}{2}, b_k+\frac{\sum_{i:S_i=k}(y_i-\mu_k)^2}{2}\right)
\end{aligned}
$$

A.3 Distribuição *a posteriori* condicional completa para os pesos

Nesta seção, apresentamos o cálculo da distribuição *a posteriori* condicional completa para os pesos de um modelo de mistura Gaussiana.

$$
\begin{aligned}
\pi(\boldsymbol{w}|...) &\propto L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{w})\pi(\boldsymbol{w}) \\
&= w_1^{n_1}\dots w_K^{n_K}\prod_{i:S_i=1}N(y_i;\mu_1,\sigma_1^2)\cdots\prod_{i:S_i=K}N(y_i;\mu_K,\sigma_K^2)\frac{1}{B(\boldsymbol{\gamma})}w_1^{\gamma_1-1}...w_k^{\gamma_k-1} \\
&\propto w_1^{n_1}...w_k^{n_k}w_1^{\gamma_1-1}...w_k^{\gamma_k-1} \\
&= w_1^{\gamma_1+n_1-1}...w_k^{\gamma_k+n_k-1} \\
&\propto Dirichlet(\gamma_1+n_1,...,\gamma_k+n_k)
\end{aligned}
$$

APÊNDICE

B

TÓPICOS ADICIONAIS: MODELOS DE MISTURA DE REGRESSÕES

B.1 Distribuição *a posteriori* condicional completa para os coeficientes de regressão

Nesta seção, apresentamos o cálculo da distribuição *a posteriori* condicional completa para os coeficientes de regressão de um modelo de mistura de regressões Normais considerando que o número de componentes, K, é conhecido.

$$
\begin{aligned}
\pi(\beta_{jk}|...) &\propto L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x})\pi(\beta_{jk}) \\
&= \prod_{k=1}^{K}\left[w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x_i}\boldsymbol{\beta}_k))^2\right\}\right]\frac{1}{(2\pi\sigma_{\beta_{jk}}^2)^{\frac{1}{2}}}exp\left\{-\frac{1}{2\sigma_{\beta_{jk}}^2}(\beta_{jk}-\mu_{\beta_{jk}})^2\right\} \\
&\propto w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x_i}\boldsymbol{\beta}_k))^2\right\}\frac{1}{(2\pi\sigma_{\beta_{jk}}^2)^{\frac{1}{2}}}exp\left\{-\frac{1}{2\sigma_{\beta_{jk}}^2}(\beta_{jk}-\mu_{\beta_{jk}})^2\right\} \\
&\propto exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x_i}\boldsymbol{\beta}_k))^2\right\}exp\left\{-\frac{1}{2\sigma_{\beta_{jk}}^2}(\beta_{jk}-\mu_{\beta_{jk}})^2\right\} \\
&= exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}[y_i-(x_{i0}\beta_{0k}+x_{i1}\beta_{1k}+...+x_{ip}\beta_{pk})]^2\right\}exp\left\{-\frac{1}{2\sigma_{\beta_{jk}}^2}(\beta_{jk}-\mu_{\beta_{jk}})^2\right\} \\
&= exp\left\{-\frac{1}{2\sigma_{\beta_{jk}}^2}(\beta_{jk}-\mu_{\beta_{jk}})^2-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-x_{i0}\beta_{0k}-x_{i1}\beta_{1k}-...-x_{ip}\beta_{pk})^2\right\}
\end{aligned}
$$

Mas,

$$\begin{aligned}
(y_i - x_{i0}\beta_{0k} - ... - x_{ip}\beta_{pk})^2 &= (y_i - x_{i0}\beta_{0k} - ... - x_{ip}\beta_{pk})(y_i - x_{i0}\beta_{0k} - ... - x_{ip}\beta_{pk}) \\
&= (y_i^2 - 2y_i x_{i0}\beta_{0k} - ... - 2y_i x_{ij}\beta_{jk} - ... - 2y_i x_{ip}\beta_{pk} \\
&+ 2x_{i0}\beta_{0k}x_{i1}\beta_{1k} + ... + 2x_{i0}\beta_{0k}x_{ij}\beta_{jk} + ... + 2x_{i0}\beta_{0k}x_{ip}\beta_{pk} \\
&+ ... + 2x_{ip}\beta_{pk}x_{i0}\beta_{0k} + ... + 2x_{ip}\beta_{pk}x_{ij}\beta_{jk} + ... + 2x_{ip}\beta_{pk}x_{ip-1}\beta_{p-1k} \\
&+ x_{i0}^2\beta_{0k}^2 + ... + x_{ij}^2\beta_{jk}^2 + ... + x_{ip}^2\beta_{pk}^2) \\
&\propto -2y_i x_{ij}\beta_{jk} + 2x_{i0}\beta_{0k}x_{ij}\beta_{jk} + ... + 2x_{ip}\beta_{pk}x_{ij}\beta_{jk} + x_{ij}^2\beta_{jk}^2 \\
&= x_{ij}^2\beta_{jk}^2 - 2x_{ij}\beta_{jk}(y_i - x_{i0}\beta_{0k} - ... - x_{ij-1}\beta_{j-1k} - x_{ij+1}\beta_{j+1k} - ... \\
&- x_{ip}\beta_{pk}) \\
&= x_{ij}^2\beta_{jk}^2 - 2x_{ij}\beta_{jk}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)
\end{aligned}$$

Então,

$$\begin{aligned}
\pi(\beta_{jk}|...) &\propto exp\left\{-\frac{1}{2\sigma^2_{\beta_{jk}}}(\beta_{jk} - \mu_{\beta_{jk}})^2 - \frac{1}{2\sigma_k^2}\sum_{i:S_i=k}\left[x_{ij}^2\beta_{jk}^2 - 2x_{ij}\beta_{jk}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)\right]\right\} \\
&= exp\left\{-\frac{1}{2\sigma^2_{\beta_{jk}}}(\beta_{jk}^2 - 2\beta_{jk}\mu_{\beta_{jk}} + \mu^2_{\beta_{jk}}) - \frac{1}{2\sigma_k^2}\sum_{i:S_i=k}\left[x_{ij}^2\beta_{jk}^2 - 2x_{ij}\beta_{jk}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)\right]\right\} \\
&\propto exp\left\{-\frac{1}{2\sigma^2_{\beta_{jk}}}(\beta_{jk}^2 - 2\beta_{jk}\mu_{\beta_{jk}}) - \frac{1}{2\sigma_k^2}\sum_{i:S_i=k}\left[x_{ij}^2\beta_{jk}^2 - 2x_{ij}\beta_{jk}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)\right]\right\} \\
&= exp\left\{-\frac{1}{2\sigma^2_{\beta_{jk}}}(\beta_{jk}^2 - 2\beta_{jk}\mu_{\beta_{jk}}) - \frac{1}{2\sigma_k^2}\left[\beta_{jk}^2\sum_{i:S_i=k}x_{ij}^2 - 2\beta_{jk}\sum_{i:S_i=k}x_{ij}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)\right]\right\} \\
&= exp\left\{-\frac{1}{2\sigma_k^2}\left[\frac{\sigma_k^2\beta_{jk}^2}{\sigma^2_{\beta_{jk}}} - \frac{2\sigma_k^2\beta_{jk}\mu_{\beta_{jk}}}{\sigma^2_{\beta_{jk}}}\right]\right\} \\
&\times exp\left\{-\frac{1}{2\sigma_k^2}\left[\beta_{jk}^2\sum_{i:S_i=k}x_{ij}^2 - 2\beta_{jk}\sum_{i:S_i=k}x_{ij}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)\right]\right\} \\
&= exp\left\{-\frac{1}{2\sigma_k^2}\left[\beta_{jk}^2\left(\frac{\sigma_k^2}{\sigma^2_{\beta_{jk}}} + \sum_{i:S_i=k}x_{ij}^2\right)\right]\right\} \\
&\times exp\left\{\frac{1}{2\sigma_k^2}\left[-2\beta_{jk}\left(\sum_{i:S_i=k}x_{ij}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right) + \frac{\sigma_k^2\mu_{\beta_{jk}}}{\sigma^2_{\beta_{jk}}}\right)\right]\right\} \\
&= exp\left\{-\frac{1}{2}\left[\beta_{jk}^2\left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{\sum_{i:S_i=k}x_{ij}^2}{\sigma_k^2}\right) - 2\beta_{jk}\left(\frac{\sum_{i:S_i=k}x_{ij}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)}{\sigma_k^2} + \frac{\mu_{\beta_{jk}}}{\sigma^2_{\beta_{jk}}}\right)\right]\right\} \\
&= exp\left\{-\frac{1}{2\left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{\sum_{i:S_i=k}x_{ij}^2}{\sigma_k^2}\right)^{-1}}\left[\beta_{jk}^2 - 2\beta_{jk}\left(\frac{\sum_{i:S_i=k}x_{ij}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)}{\sigma_k^2} + \frac{\mu_{\beta_{jk}}}{\sigma^2_{\beta_{jk}}}\right)\left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{\sum_{i:S_i=k}x_{ij}^2}{\sigma_k^2}\right)^{-1}\right]\right\} \\
&\propto exp\left\{-\frac{1}{2\left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{\sum_{i:S_i=k}x_{ij}^2}{\sigma_k^2}\right)^{-1}}\left[\beta_{jk} - \left(\frac{\sum_{i:S_i=k}x_{ij}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)}{\sigma_k^2} + \frac{\mu_{\beta_{jk}}}{\sigma^2_{\beta_{jk}}}\right)\left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{\sum_{i:S_i=k}x_{ij}^2}{\sigma_k^2}\right)^{-1}\right]^2\right\} \\
&\propto N\left(\left(\frac{\sum_{i:S_i=k}x_{ij}\left(y_i - \sum_{\substack{l=0,\\ l\neq j}}^{p} x_{il}\beta_{lk}\right)}{\sigma_k^2} + \frac{\mu_{\beta_{jk}}}{\sigma^2_{\beta_{jk}}}\right)\left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{\sum_{i:S_i=k}x_{ij}^2}{\sigma_k^2}\right)^{-1}, \left(\frac{1}{\sigma^2_{\beta_{jk}}} + \frac{\sum_{i:S_i=k}x_{ij}^2}{\sigma_k^2}\right)^{-1}\right)
\end{aligned}$$

B.2 Distribuição *a posteriori* condicional completa para as variâncias

Nesta seção, apresentamos o cálculo da distribuição *a posteriori* condicional completa para as variâncias de um modelo de mistura de regressões Normais considerando que o número de componentes, K, é conhecido.

$$\begin{aligned}
\pi(\sigma_k^2|...) &\propto L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x})\pi(\sigma_k^2) \\
&= \prod_{k=1}^{K}\left[w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x}_i\boldsymbol{\beta}_k))^2\right\}\right]\frac{b_k^{a_k}}{\Gamma(a_k)}(\sigma_k^2)^{-a_k-1}exp\left\{-\frac{b_k}{\sigma_k^2}\right\} \\
&\propto w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x}_i\boldsymbol{\beta}_k))^2\right\}\frac{b_k^{a_k}}{\Gamma(a_k)}(\sigma_k^2)^{-a_k-1}exp\left\{-\frac{b_k}{\sigma_k^2}\right\} \\
&\propto (\sigma_k^2)^{-\frac{n_k}{2}-a_k-1}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x}_i\boldsymbol{\beta}_k))^2-\frac{b_k}{\sigma_k^2}\right\} \\
&= (\sigma_k^2)^{-(\frac{n_k}{2}+a_k+1)}exp\left\{-\frac{\frac{1}{2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x}_i\boldsymbol{\beta}_k))^2+b_k}{\sigma_k^2}\right\} \\
&\propto IG\left(\frac{n_k}{2}+a_k,\frac{1}{2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x}_i\boldsymbol{\beta}_k))^2+b_k\right)
\end{aligned}$$

para todo $k=1,...K$.

B.3 Distribuição *a posteriori* condicional completa para os pesos das componentes

Nesta seção, apresentamos o cálculo da distribuição *a posteriori* condicional completa para os pesos das componentes de um modelo de mistura de regressões Normais considerando que o número de componentes, K, é conhecido.

$$\begin{aligned}
\pi(\boldsymbol{w}|...) &\propto L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{x})\pi(\boldsymbol{w}) \\
&= \prod_{k=1}^{K}\left[w_k^{n_k}\frac{1}{(2\pi\sigma_k^2)^{\frac{n_k}{2}}}exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i-(\boldsymbol{x}_i\boldsymbol{\beta}_k))^2\right\}\right]\frac{1}{B(\boldsymbol{\gamma})}w_1^{\gamma_1-1}...w_k^{\gamma_k-1} \\
&\propto w_1^{n_1}...w_k^{n_k}w_1^{\gamma_1-1}...w_k^{\gamma_k-1} \\
&= w_1^{\gamma_1+n_1-1}...w_k^{\gamma_k+n_k-1} \\
&\propto Dirichlet(\gamma_1+n_1,...,\gamma_k+n_k)
\end{aligned}$$