# Modelos semiparamétricos com erros da classe de distribuições de mistura na escala normal: uma abordagem Bayesiana

Luz Marina Rondon Poveda

TESE APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
DOUTOR EM CIÊNCIAS

Programa: Estatística
Orientador: Prof. Dr. Heleno Bolfarine

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do Programa Estudantes-Convênio de Pós-Graduação– PEC-PG, da CAPES/CNPq - Brasil

São Paulo, maio de 2015

# Modelos semiparamétricos com erros da classe de distribuições de mistura na escala normal: uma abordagem Bayesiana

Esta é a versão original da tese elaborada pelo candidato Luz Marina Rondon Poveda, tal como submetida à Comissão Julgadora.

# Resumo

Neste trabalho estudamos modelos de regressão semiparamétricos sob a abordagem Bayesiana, em que sua componente aleatória segue distribuições de mistura normal na escala, as quais incluem distribuições bem conhecidas como a normal, *t*-Student, slash, normal contaminada, Laplace e hiperbólica simétrica. Na primeira parte do trabalho, estudamos a inferência e diagnóstico nos modelos semiparamétricos aditivos elípticos generalizados, em que o parâmetro de localização bem como o de dispersão incluem componentes não paramétricas aditivas aproximadas usando *B*-splines. Em seguida, estudamos a versão estrutural de modelos com erros nas variáveis homocedásticos e heterocedásticos. A componente sistemática destes modelos admite variáveis explicativas com e sem erro de medição bem como a presenca de efeitos não lineares aproximados usando *B*-splines. Com o objetivo de gerar amostras da distribuição a posteriori dos parâmetros dos modelos estudados, propomos algoritmos MCMC eficientes. Adicionalmente, desenvolvemos o pacote **BayesGESM** na linguagem R, o qual é uma ferramenta computacional para aplicar os métodos estatísticos estudados neste trabalho. Com o intuito de ilustrar as metodologias propostas, estudos de simulação são conduzidos e várias aplicações a conjuntos de dados reais são apresentadas.

**Palavras-chave:** Modelos semiparamétricos, modelos com erros nas variáveis, *B*-splines, algoritmo MCMC, distribuições de mistura normal na escala, heterocedasticidade.

# Abstract

We study the statistical inference under the Bayesian approach for semi-parametric models, where the random component belongs to the class of scale mixtures of normal distributions. Members of this class include some well known symmetric distributions such as the normal, Student-$t$, slash, contaminated normal, Laplace and symmetric hyperbolic. First, we study the generalized elliptical semi-parametric models, where both location and dispersion parameters of the response variable distribution include nonparametric additive components approximated by using $B$-splines. Also, we study the structural version of the flexible measurement error models under the presence of homocedastic and heterocedastic random errors. The systematic components of these models include explanatory variable vectors with and without measurement errors, as well as nonlinear effects that are approximated by using $B$-splines. To draw samples of the posterior distribution of the model parameters, efficient MCMC algorithms are proposed. The proposed methodology is illustrated by using simulations experiments and by analyzing several real data sets. Additionally, the proposed methods are implemented in the R package **BayesGESM**.

**Keywords:** Bayesian analysis, measurement error models, semi-parametric models, MCMC algorithm, $B$-splines, scale mixtures of normal distributions.

# Sumário

# Lista de Figuras

# Lista de Tabelas

Capítulo 1

# Introdução

Modelos de regressão com erros normais independentes e dispersão variável constituem uma ferramenta muito flexível para a análise estatística de dados, pois, eles admitem que os parâmetros de localização bem como os de dispersão dependam de variáveis explicativas, o que permite que este tipo de modelos possa ser aplicado numa classe bastante abrangente de situações práticas. A inferência nesta classe de modelos foi desenvolvida por Aitkin (1987) e Verbyla (1993) sob o enfoque clássico, e por Cepeda e Gamerman (2001) sob o enfoque Bayesiano. Xu e Zhang (2013) estendem o trabalho de Cepeda e Gamerman (2001) incluindo um efeito não paramétrico aditivo na componente sistemática do parâmetro de localização, ou seja, admitindo que a forma funcional da dependência entre a média ou mediana da variável resposta e uma das variáveis explicativas contínuas é desconhecida. No entanto, na prática podem existir conjuntos de dados onde o efeito de uma variável explicativa sobre o parâmetro de dispersão possui uma forma funcional que também é desconhecida. Além disso, como é bem conhecido, a inferência em modelos sob a suposição de erros normalmente distribuídos pode ser altamente influenciada pelas observações extremas na variável resposta (veja Maronna *et al.*, 2006). Com estas motivações, estudamos no capítulo 2 deste trabalho, a inferência e diagnóstico sob o enfoque Bayesiano de modelos de regressão com erros independentes seguindo distribuição normal, $t$-Student, slash, normal contaminada, Laplace e hiperbólica simétrica, em que o parâmetro de localização bem como o de dispersão incluem efeitos não paramétricos aditivos aproximados através de $B$-splines.

Por outro lado, na prática existem conjuntos de dados com a presença de variáveis sujetas a erros de medição. A presença de variáveis deste tipo pode afectar significativamente o bom desempenho dos estimadores dos parâmetros. Por exemplo, segundo Fuller (1987) e Cheng e VanNess (1999), a presença de erro de medição pode introduzir viés e produzir intervalos de confiança com taxas de cobertura baixa. Os modelos de regressão em que são levados em conta os erros de medição, são chamados de modelos com erros nas variáveis. Estes modelos têm sido estudados por muitos autores. Por exemplo, sob a aborgagem clássica, Arellano-Valle *et al.* (1996) estudaram modelos em que a componente aleatória segue uma distribuição $t$-Student; Carroll *et al.* (1999) utilizaram misturas da normal para estudar um modelo flexível paramétrico; Kulathinal *et al.* (2002) estudam o modelo com erros nas variáveis heterocedástico assumindo que a componente aleatória do modelo segue distribuição normal; Liang *et al.* (2004) estudaram um modelo em que a distribuição da variável explicativa é Poisson; Patriota *et al.* (2009) apresentam a teoria assintótica de um modelo com erros nas variáveis herecedástico com erro na equação, supondo para a componente aleatória a distribuição normal; Cao *et al.* (2012) estudaram modelos com erros nas variáveis heterocedáticos em que sua componente aleatória segue distribuições de mistura normal na escala. Por outro lado, sob a abordagem Bayesiana, Kelly (2007) descreve um modelo para dados astronômicos;

Carroll *et al.* (2006, capítulo 9) apresentam modelos com erros nas variávies lineares e não lineares; e de Castro *et al.* (2013) estudam modelos com erros nas variáveis heterocedásticos, em que as matrizes de variâncias e covariâncias dos erros aleatórios são estimadas usando réplicas. A maioria das propostas apresentadas nestes trabalhos estão baseadas na suposição de normalidade para a componente aleatória do modelo. Além disso, elas não permitem a presença simultânea de variáveis explicativas com e sem erro de medição nem a presença de efeitos não lineares cuja forma funcional é desconhecida.

Sendo assim, nos capítulos 3 e 4 deste trabalho estudamos, sob o enfoque Bayesiano, modelos com erros nas variáveis homocedásticos e heterocedásticos, respectivamente. A componente sistemática destes modelos admite variáveis explicativas com e sem erro de medição bem como a presenca de efeitos não lineares aproximados usando $B$-splines. Nos dois casos, o modelo estudado é a versão estrutural (ou seja, assume-se que as variaveis explicativas com erro de medicao são variáveis aleatórias identicamente distribuidas), e sua componente aleatória segue distribuições da classe de mistura normal na escala, o qual proporciona flexibilidade bem como robustez na presença de observacões extremas.

Com o objetivo de fornecer uma ferramenta computacional para realizar a inferência estatística baseada na abordagem Bayesiana para os modelos apresentados nesta tese, desenvolvimos o pacote **BayesGESM** (Rondon e Bolfarine, 2014) na linguagem R (R Core Team, 2014). Este pacote encontra-se disponível no "Comprehensive R Archive Network" (CRAN) em http://CRAN.R-project.org/package=BayesGESM e pode ser instalado livremente. Uma descrição completa deste pacote é apresentada no capítulo 5.

## 1.1 Família de Distribuições de Mistura Normal na Escala

A classe de distribuições de mistura normal na escala, introduzida por Andrews e Mallows (1974), fornece um amplo grupo de distribuições simétricas, algumas das quais têm caudas mais pesadas/leves do que a distribuição normal, bem como distribuições com diferentes níveis de curtose. Na verdade, as distribuições com caudas mais pesadas do que a normal podem ser usadas para obter inferência robusta em conjuntos de dados com observações extremas ou atípicas. Na família de distribuições de mistura normal na escala ($\mathcal{SMN}$) podemos encontrar distribuições bastante conhecidas tais como $t$-Student, slash (veja Rogers e Tukey, 1972), normal contaminada, Laplace (veja Box e Tiao, 1973) e hiperbólica simétrica (veja Barndoff-Nielsen , 1977).

Seguindo Andrews e Mallows (1974), um vetor aleatório $r$-dimensional $\mathbf{Y} = (Y_1, \ldots, Y_r)^T$ segue uma distribuição de mistura normal na escala multivariada, denotada por $\mathcal{SMN}_r$, se ele pode ser escrito como

$$\mathbf{Y} = \boldsymbol{\mu} + \kappa^{\frac{1}{2}}(U)\mathbf{Z},$$

em que

- $\boldsymbol{\mu} \in \boldsymbol{R}^r$ é o parâmetro de localização;
- $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, com $\boldsymbol{\Sigma}$ uma matriz definida positiva de dimensão $(r \times r)$;
- $\kappa(\cdot)$ é uma função estritamente positiva;
- $U$ é uma variável aleatória positiva independente de $\mathbf{Z}$, com $H(u; \boldsymbol{\eta})$ sua função de distribuição acumulada (fda); e
- $\boldsymbol{\eta}$ é um parâmetro ou um vetor de parâmetros que indexa a distribuição de $U$.

Então, $\mathbf{Y}$ tem uma distribuição $\mathcal{SMN}$ se sua função de densidade de probabilidade (fdp) pode ser escrita como

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}) = \int_0^\infty \phi_r(\mathbf{y}|\boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma}) dH(u; \boldsymbol{\eta}), \tag{1.1}$$

em que $\phi_r(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \exp\left\{-\frac{1}{2}\delta^2\right\} \times (2\pi)^{-r/2}|\boldsymbol{\Sigma}|^{-1/2}$ e $\delta^2 = (\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$. Neste caso denotamos $\mathbf{Y} \sim \mathcal{SMN}_r(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\eta})$. Usando uma representação hierárquica, a distribuição de $\mathbf{Y}$ pode ser especificada como $\mathbf{Y}|U = u \sim \mathcal{N}_r(\boldsymbol{\mu},\kappa(u)\boldsymbol{\Sigma})$ e $U \sim \mathcal{H}(u;\boldsymbol{\eta})$. A variável aleatória $U$ pode ser discreta ou contínua e a forma da distribuição $\mathcal{SMN}_r$ é determinada pela distribuição desta variável. A seguer apresentamos alguns exemplos de distribuições que pertencem à família $\mathcal{SMN}_r$:

- *Distribuição normal multivariada*
  Temos que $P[U = 1] = 1$, ou seja, $U$ é uma variável aleatória degenerada. Portanto, a distribuição de $\mathbf{Y}$ se reduz a
  $$f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \phi(y|\boldsymbol{\mu},\boldsymbol{\Sigma}).$$
  Assim, $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ e $\mathrm{Var}(\mathbf{Y}) = \zeta\boldsymbol{\Sigma}$, com $\zeta = 1$.

- *Distribuição t-Student multivariada.*
  Neste caso temos que $U \sim \mathcal{G}ama(\eta/2,\eta/2)$, com $\eta > 0$ e $\kappa(u) = 1/u$. Então, de acordo com (1.1) a densidade do vector aleatório $\mathbf{Y}$ é dada por
  $$f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\eta) = \frac{\Gamma\left(\frac{\eta+r}{2}\right)}{\Gamma\left(\frac{\eta}{2}\right)|\boldsymbol{\Sigma}|^{1/2}(\pi\eta)^{r/2}}\left[1+\frac{\delta^2}{\eta}\right]^{-\left(\frac{\eta+r}{2}\right)}.$$
  Daí, $\mathbf{Y} \sim t_r(\boldsymbol{\mu},\boldsymbol{\Sigma},\eta)$. Além disso, temos que $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ e $\mathrm{Var}(\mathbf{Y}) = \zeta\boldsymbol{\Sigma}$, com $\zeta = \frac{\eta}{\eta-2}$ para $\eta > 2$.

- *Distribuição slash multivariada.*
  Seja $\kappa(u) = 1/u$ e $U \sim \mathcal{B}eta(\eta,1)$, $\eta > 0$. Daí, a fpd de $\mathbf{Y}$ pode ser escrita como
  $$f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\eta) = \frac{\eta}{(2\pi)^{r/2}|\boldsymbol{\Sigma}|^{1/2}}\int_0^1 u^{\eta+\frac{r}{2}-1}\exp[-u\delta^2/2]du.$$
  Então, $\mathbf{Y} \sim \mathcal{S}l_r(\boldsymbol{\mu},\boldsymbol{\Sigma},\eta)$. Temos que $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ e $\mathrm{Var}(\mathbf{Y}) = \zeta\boldsymbol{\Sigma}$, com $\zeta = \frac{\eta}{\eta-1}$ para $\eta > 1$.

- *Distribuição normal contaminada multivariada.*
  Neste caso, $\kappa(u) = 1/u$ e $U$ é uma variável aleatória discreta que toma o valor de $\eta_2$ com probabilidade $\eta_1$ e o valor de 1 com probabilidade $(1-\eta_1)$, ou seja, a fdp de $U$ é dada por
  $$h(u|\boldsymbol{\eta} = (\eta_1,\eta_2)^T) = \eta_1\,\mathbb{I}_{(u=\eta_2)} + (1-\eta_1)\,\mathbb{I}_{(u=1)}.$$
  Segundo (1.1), a densidade do vetor aleatório $\mathbf{Y}$ tem a seguinte forma
  $$f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\eta}) = \eta_1\,\phi_r(\mathbf{y}|\boldsymbol{\mu},\eta_2^{-1}\boldsymbol{\Sigma}) + (1-\eta_1)\,\phi_r(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma}),$$
  em que $0 < \eta_1 < 1$ e $0 < \eta_2 < 1$. Então, $\mathbf{Y} \sim \mathcal{CN}_r(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\eta})$. Temos que $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ e $\mathrm{Var}(\mathbf{Y}) = \zeta\boldsymbol{\Sigma}$, com $\zeta = \frac{\eta_1}{\eta_2} + (1-\eta_1)$.

- *Distribuição Laplace multivariada.*
  Seja $U \sim \mathcal{E}xp\,(1/8)$ e $\kappa(u) = u$, então
  $$f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\mathbf{K}_a\left(\sqrt{\delta^2/4}\right)\left(\delta^2\right)^{-\frac{r+2}{4}}}{2^{r+1}\pi^{r/2}|\boldsymbol{\Sigma}|^{1/2}},$$

em que $a = -\frac{r}{2} + 1$ e $\mathbf{K}_a(b) = \frac{1}{2}\int_0^\infty x^{a-1}\exp\left(-\frac{1}{2}b(x+x^{-1})\right)\partial x$ é a função modificada de Bessel tipo três de ordem $a$ (veja Watson, 1995). Por exemplo, para $a = \frac{1}{2}$ temos que $\mathbf{K}_a(b) = \sqrt{\pi/2b}\exp(-b)$. Assim, $\mathbf{Y} \sim \mathcal{L}aplace_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Além disso, temos que $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ e $\mathrm{Var}(\mathbf{Y}) = \zeta\boldsymbol{\Sigma}$, com $\zeta = 8$.

- *Distribuição hiperbólica simétrica multivariada.*
  Neste caso, $\kappa(u) = u$ e $U$ segue uma distribuição Gaussiana inversa generalizada, isto é, $U \sim \mathcal{GIG}(1, 1, \eta^2)$ (veja Jørgensen, 1982). A fdp da variável aleatória $U \sim \mathcal{GIG}(a, b, c)$ é dada por
  $$h(u|a, b, c) = \frac{(c/b)^{\frac{a}{2}}}{2\mathbf{K}_a(\sqrt{bc})}u^{a-1}\exp\left[-\frac{1}{2}(bu^{-1} + cu)\right].$$
  Então, deacordo com (1.1) a fdp de $\mathbf{Y}$ pode ser expressa por
  $$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta) = \frac{\mathbf{K}_a\left(\eta\sqrt{\delta^2+1}\right)\eta^{r/2}\left(\delta^2+1\right)^{-\frac{r}{4}+\frac{1}{2}}}{2^{r/2}\pi^{r/2}|\boldsymbol{\Sigma}|^{1/2}\mathbf{K}_1(\eta)},$$
  em que $a = -\frac{r}{2} + 1$. Portanto, $\mathbf{Y} \sim \mathcal{SH}_r(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta)$, com $\eta > 0$. Daí, $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ e $\mathrm{Var}(\mathbf{Y}) = \zeta\boldsymbol{\Sigma}$, com $\zeta = \frac{\mathbf{K}_2(\eta)}{\eta\mathbf{K}_1(\eta)}$.

Como casos particulares das distribuições $t$-Student, slash, normal contaminada, hiperbólica simétrica e Laplace temos distribuições com caudas mais pesadas que as da normal, bem como distribuições com graus de curtose mais acentuados que os da normal, o que torna bastante flexível os modelos que consideran esta classe para descrever sua componente aleatória. Além disso, como foi mostrado anteriormente, estas distribuições podem ser escritas de forma hierárquica a partir da distribuição normal, o qual facilita o desenvolvimento dos algoritmos para amostrar da distribuição a posteriori dos parâmetros de interesse em modelos de regressão com erros seguindo este tipo de distribuições. Propriedades adicionais das distribuições da classe $\mathcal{SMN}$ podem ser encontradas em Andrews e Mallows (1974); Fang *et al.* (1990) e Lange e Sinsheimer (1993).

Na Figura 1.1 apresentamos os gráficos das densidades de distribuições da classe $\mathcal{SMN}$ univariada padrão ($\mu = 0$ e $\sigma^2 = 1$) quando é comparada com a distribuição normal padrão. Consideramos as distribuições $t$-Student, slash, normal contaminada, Laplace e hiperbólica simétrica, com vários valores do parâmetro extra $\eta$. Além disso, na Figura 1.2 podemos observar os gráficos de algumas densidades da família $\mathcal{SMN}_2$ padrão ($\boldsymbol{\mu} = \mathbf{0}$ e $\boldsymbol{\Sigma} = \mathbf{I}$) quando é comparada com a distribuição normal padrão bivariada.

## 1.2 Métodos MCMC

Sob o enfoque Bayesiano a inferência sobre os parâmetros de interesse baseia-se na distribuição a posteriori. Comummente esta distribuição possui uma forma complexa e não pode ser aproximada usando as técnicas de integração numérica. Então, uma forma eficiente de fazer inferência Bayesiana nestes casos consiste em gerar amostras desta distribuição usando métodos baseados em simulação estocástica na teoria de cadeias de Markov, chamados métodos de Monte Carlo via Cadeias de Markov (MCMC). Estes métodos procuram determinar quais são as condições sob as quais existe uma distribuição invariante e condições em que iterações do núcleo da transição da cadeia convergem para esta distribuição invariante (Gamerman e Lopes, 2006). Os métodos MCMC mais usados são o amostrador de Gibbs e o algoritmo de Metropolis-Hastings.

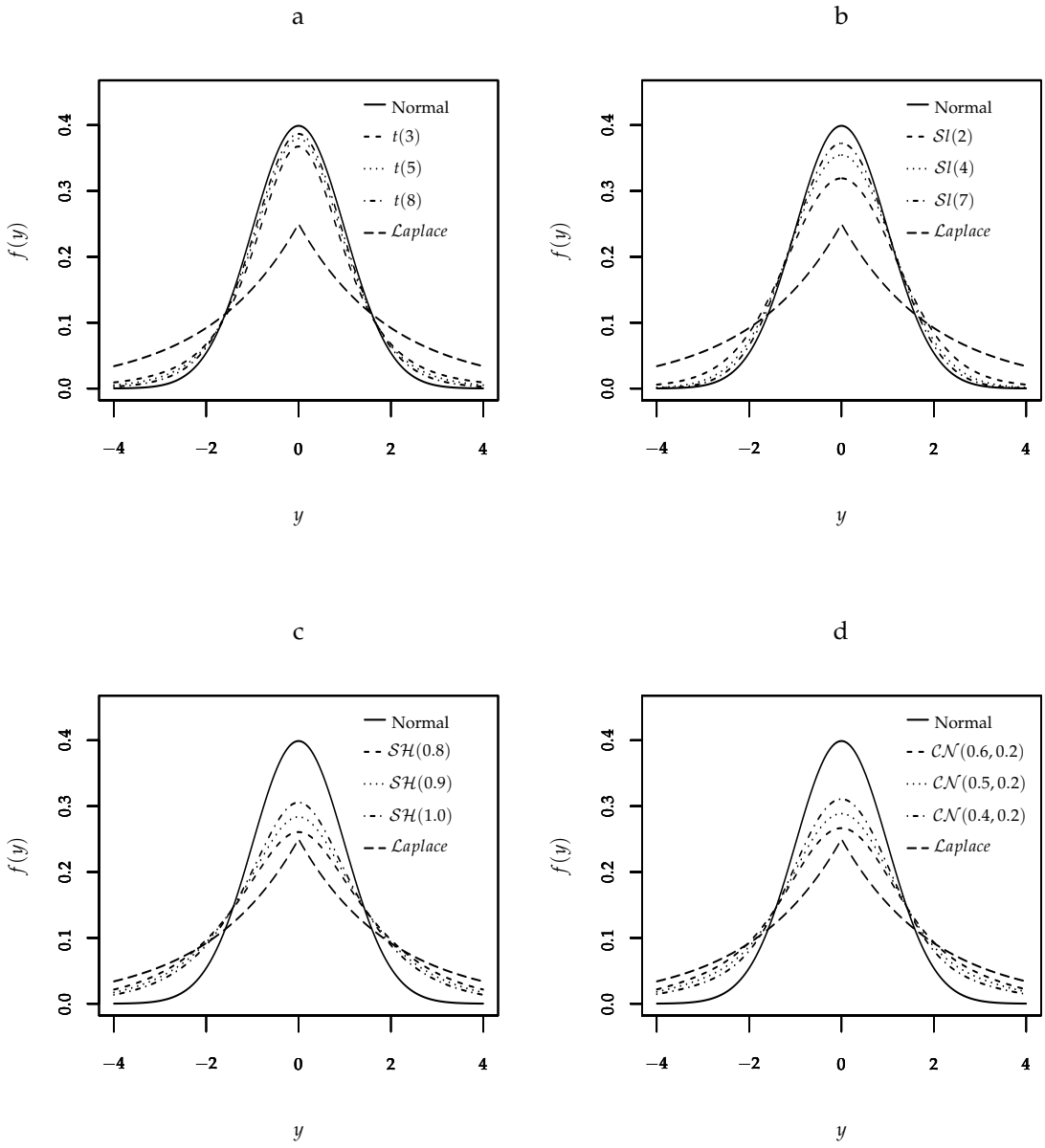


**Figura 1.1:** *Gráficos das densidades de distribuições padrão da classe $\mathcal{SMN}$ univariada: t-Student (a), Slash (b), hiperbólica simétrica (c) e normal contaminada (d).*

O algoritmo de Metropolis-Hastings proposto por Metropolis *et al.* (1953) e estendido por Hastings (1970) é um procedimento de simulação eficiente, poderoso e versátil, pois permite amostrar de qualquer densidade de probabilidade, sem que seja necessário conhecer as distribuições condicionais. Este algoritmo gera valores para o parâmetro de interesse $\theta$ usando uma distribuição proposta, em que cada valor pode ser aceito ou rejeitado para fazer parte da amostra da distribuição a posteriori. O uso deste algoritmo é conveniente quando a distribuição condicional completa não é fácil de identificar.

O amostrador de Gibbs pode ser considerado um caso particular do algoritmo Metropolis-Hastings (Gelman e Rubin, 1992), em que todos os valores gerados são aceitos. O amostrador de Gibbs é um método de simulação de distribuições multivariadas de natureza bastante complexa, baseado na sua caracterização através das distribuições condicionais completas. O algoritmo de simulação proposto define uma cadeia de Markov que, sob condições muito gerais,

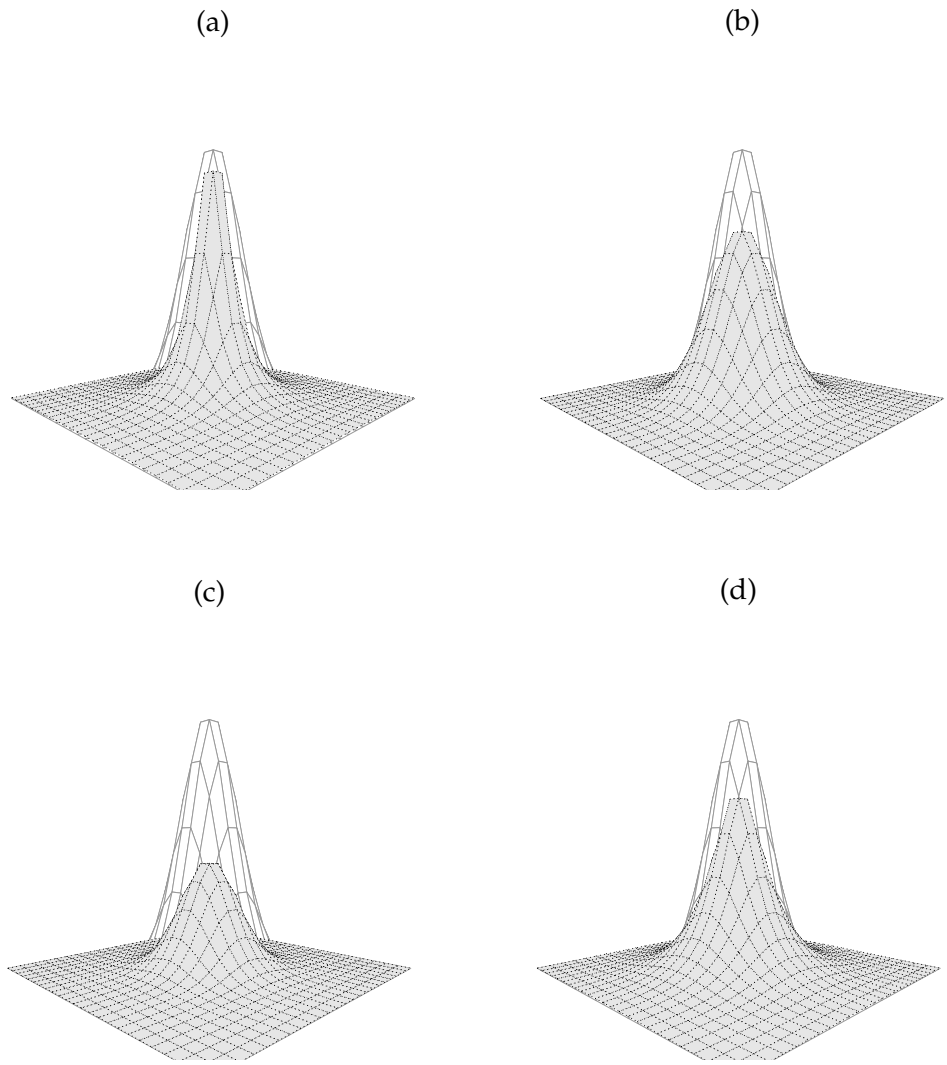


**Figura 1.2:** *Gráficos das densidades de distribuições padrão da classe $\mathcal{SMN}$ multivariada: t-Student$(\eta = 1)$ (a), Slash$(\eta = 2)$ (b), normal contaminada$(\boldsymbol{\eta} = (0.6, 0.2)^T)$ (c) e hipérbolica simétrica$(\eta = 1)$ (d).*

tem como distribuição limite a distribuição conjunta que se pretende simular.

Neste trabalho usaremos os métodos MCMC para construir uma cadeia que tenha como distribuição estacionária a distribuição a posteriori. Uma vez a convergência seja atingida, usaremos os valores da cadeia como uma amostra da distribuição a posteriori dos parâmetros dos modelos.

## 1.3 *B*-splines

Os splines são oriundos do cálculo numérico, que ganharam atenção na área da estatística pelo seu poder adaptativo na aproximação de funções usando uma série de polinômios, os quais estão definidos em sub-intervalos que cumprem determinadas condições de suavização. Os extremos destes intervalos são chamados nós. O espaço dos splines é determinado pela ordem dos polinómios e a localização dos nós. Os splines vêm em diversas variedades: splines de suavização, splines de regressão, *B*-splines, entre outros.

Neste trabalho, consideramos os *B*-splines para aproximar uma função não paramétrica, pois como aproximadores de uma função suave, os *B*-splines (veja de Boor, 1978) têm duas propriedades desejáveis. As funções base são suportadas localmente, assim esta aproximação adapta-se muito bem ao comportamento local da função a ser aproximada. Além disso, freqüentemente fornecem boas aproximações com um pequeno número de nós e apresentam eficiência computacional e estabilidade. Portanto, os *B*-splines são um bom método de alisamento. A seguer apresentamos algumas das propriedades de um *B*-spline de grau $m$ (veja Eilers e Marx, 1996):

- é composto por $m+1$ troços polinomiais, cada um de ordem $m$;
- se unen em $m$ nós internos;
- nos pontos de união as derivadas até a ordem $m-1$ são contínuas;
- é positivo num dominio definido por $m+2$ nós adjacentes, em qualquer outro lugar é zero;
- exceto nas fronteiras, intersecta até $2m$ peças polinomiais vizinhas;
- para cada valor de $v$, $m+1$ *B*-spline são não nulos.

Suponhamos que nosso objetivo é estimar uma função de $v$ no intervalo $[0,1]$, denotada $f(v)$. Então, a função $f(v)$ aproximada usando *B*-splines é dada por

$$\sum_{k=1}^{K} \alpha_k b_k(v;m),$$

em que $b_k(v;m)$ são as funções de base do *B*-spline de grau $m$ que constituem a base da aproximação de $f(v)$ e $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$ representa o vetor que contém os coeficientes do *B*-spline. Seja $0 = s_0 < s_1 < \ldots < s_k < s_{k+1} = 1$ uma partição do intervalo. Usando $s_i$ $(i = 1, \ldots, k)$ como nós internos, temos que $K = m + k$ são as funções (normalizadas) B-splines de ordem $m$ que formam a base para o espaço do spline linear. As funções de base podem ser expressas pelo vetor $\mathbf{b}(v) = (b_1(v), \ldots, b_K(v))^T$. de Boor (1978) apresenta um algoritmo para o calculo dos *B*-splines.

De certa forma, a escolha dos nós controla a suavidade da curva, e por isso a seleção do número de nós é um ponto importante no splines. Seguindo a proposta de He *et al.* (2005), neste trabalho consideramos $k = [n^{1/5}]$ como o número de nós internos, em que $n$ é o tamanho da amostra e $[x]$ é a função parte inteira de $x$. De forma similar, os nós internos são selecionados como $\{q(v, 1/(k+1)), \ldots, q(v, k/(k+1))\}$, em que $q(x,p)$ representa o quantil de ordem $0 < p < 1$ de $x$.

Capítulo 2

# Modelo semiparamétrico aditivo elíptico generalizado

Neste capítulo estudamos a inferência e diagnóstico sob o enfoque Bayesiano dos modelos semiparamétricos aditivos elípticos generalizados. Nestes modelos, os erros do modelo seguem distribuições de mistura normal na escala ($\mathcal{SMN}$). Como exemplos desta classe temos as distribuições normal, $t$-Student, slash, normal contaminada, Laplace e hiperbólica simétrica. Nos Modelos Semiparamétricos Aditivos Elípticos Generalizados (MSAEG), o parâmetro de localização bem como o de dispersão incluem componentes não paramétricas aditivas aproximadas usando $B$-splines. Vale salientar que, estes modelos generalizam as componentes sistemática (uma vez que eles consideram simultaneamente efeitos paramétricos lineares e não paramétricos) e aleatória (pois eles consideram para o erro aleatório distribuições da classe $\mathcal{SMN}$) dos modelos estudados por Aitkin (1987), Verbyla (1993), Cepeda e Gamerman (2001) e Xu e Zhang (2013).

Na primeira parte deste capítulo formulamos o MSAEG. Em seguida, descrevemos as distribuições a priori e desenvolvemos o algoritmo MCMC para obter amostras da distribuição a posteriori dos parâmetros de interesse. Na seção 2.3 apresentamos critérios que permitem a comparação de modelos e algumas ferramentas de diagnóstico no contexto Bayesiano como, por exemplo, medidas de influência baseadas em deleção de casos e resíduos. Na Seção 2.4 apresentamos um estudo de simulação em que ilustramos o desempenho do algoritmo MCMC proposto. Finalmente, na Seção 2.5 a metodologia proposta é aplicada a um conjunto de dados reais.

A implementação computacional da metodologia desenvolvida foi realizada na linguagem `R` através da função `gesm()` disponível no pacote **BayesGESM** (Rondon e Bolfarine, 2014). Uma descrição completa desta função é apresentada no capítulo 5.

## 2.1 Formulação do modelo

Suponha que temos um conjunto de $n$ observações $y_1, \ldots, y_n$ que foram geradas a través do seguinte mecanismo:

$$y_i = \mu_i + \sigma_i \epsilon_i, \qquad i = 1, \ldots, n,$$

em que $\epsilon_i \sim \mathcal{SMN}(0, 1; H, \kappa)$ são variáveis aleatórias independentes. Além disso, assumimos que os parâmetros de localização e dispersão podem ser escritos como

$$\begin{cases} \mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + f(v_i) & \text{e} \\ \sigma_i^2 = h(\mathbf{z}_i^T \boldsymbol{\gamma} + g(w_i)), \end{cases} \tag{2.1}$$

em que

- $(\mathbf{x}_i^T, v_i)^T$ e $(\mathbf{z}_i^T, w_i)^T$ são vetores de variáveis explicativas associadas ao parâmetro de localização e dispersão do indivíduo $i$, respectivamente;
- $f(\cdot)$ e $g(\cdot)$ são funções suaves arbitrárias e desconhecidas das variáveis explicativas contínuas $v$ e $w$, respectivamente;
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$ são vetores de parâmetros desconhecidos a serem estimados;
- $h^{-1}(\cdot)$ é uma função de ligação para o parâmetro de dispersão cuja inversa é monótona e estritamente positiva (ou seja, $h(x) > 0$ para $x \in \mathbb{R}$). Uma escolha conveniente é $h(\cdot) = \exp(\cdot)$, mas outras escolhas são possíveis.

Neste trabalho, as funções não paramétricas $f(v)$ e $g(w)$ no modelo (2.1) são aproximadas usando $B$-splines (para detalhes dos $B$-splines veja a seção 1.3). Portanto, as funções $f(v)$ e $g(w)$ podem ser aproximadas por $\mathbf{b}(v)^T\boldsymbol{\alpha}$ e $\mathbf{d}(w)^T\boldsymbol{\lambda}$, respectivamente, em que $\mathbf{b}(v) = (b_1(v), \ldots, b_{K_1}(v))^T$ e $\mathbf{d}(w) = (d_1(w), \ldots, d_{K_2}(w))^T$ são os vetores das funções de base, $\boldsymbol{\alpha} \in \mathbb{R}^{K_1}$ e $\boldsymbol{\lambda} \in \mathbb{R}^{K_2}$ são os vetores que contém os coeficientes dos $B$-splines, e $K_j = M_j + k_j$ $(j = 1, 2)$, onde $M_j$ e $k_j$ são o grau e o número de nós internos do $B$-spline, respectivamente.

Sob a abordagem freqüentista, termos de penalização para $\boldsymbol{\alpha}$ e $\boldsymbol{\lambda}$ são introduzidos para evitar "overfitting", os quais são dados por $\frac{1}{2\tau_\alpha^2}\boldsymbol{\alpha}^T\boldsymbol{\alpha}$ e $\frac{1}{2\tau_\lambda^2}\boldsymbol{\lambda}^T\boldsymbol{\lambda}$, em que $\tau_\alpha^2$ e $\tau_\lambda^2$ são os parâmetros de suavização. De fato, estas aproximações de $f(v)$ e $g(w)$ podem ser consideradas como sendo $P$-splines (veja por exemplo Eilers e Marx, 1996) com um termo de penalidade de ordem zero. Então, sob a abordagem bayesiana, prioris para $\boldsymbol{\alpha}$ e $\boldsymbol{\lambda}$ proporcionais a $\exp(\frac{1}{2\tau_\alpha^2}\boldsymbol{\alpha}^T\boldsymbol{\alpha})$ e $\exp(\frac{1}{2\tau_\lambda^2}\boldsymbol{\lambda}^T\boldsymbol{\lambda})$ podem ser introduzidas.

Usando a notação anterior, temos que as componentes não paramétricas de $\mu_i$ e $\sigma_i^2$ do modelo (2.1) podem ser escritas como

$$\mu_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{b}(v_i)^T\boldsymbol{\alpha} \quad \text{and} \quad \sigma_i^2 = h(\mathbf{z}_i^T\boldsymbol{\gamma} + \mathbf{d}(w_i)^T\boldsymbol{\lambda}), \qquad i = 1 \ldots, n. \tag{2.2}$$

Definindo $\mathbf{B} = (\mathbf{b}(v_1), \ldots, \mathbf{b}(v_n))^T$ e $\mathbf{D} = (\mathbf{d}(w_1), \ldots, \mathbf{d}(w_n))^T$, e usando a expressão (2.2), temos que a função de verossimilhança (completa) de $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)^T$ pode ser escrita como

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda} | \mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{w}, \mathbf{u}) \propto |\boldsymbol{\Sigma}^*|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{B}\boldsymbol{\alpha})^T(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{B}\boldsymbol{\alpha})\right] \tag{2.3}$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$, $\mathbf{v} = (v_1, \ldots, v_n)^T$, $\mathbf{w} = (w_1, \ldots, w_n)^T$, $\boldsymbol{\Sigma} = \text{diag}\{h(\mathbf{z}_1^T\boldsymbol{\gamma} + \mathbf{d}(w_1)^T\boldsymbol{\lambda}), \ldots, h(\mathbf{z}_n^T\boldsymbol{\gamma} + \mathbf{d}(w_n)^T\boldsymbol{\lambda})\}$, $\mathbf{u} = (u_1, \ldots, u_n)^T$ é o vetor de variáveis latentes (ou vetor que contém as variáveis da mistura), $\mathbf{L}_{(u)} = \text{diag}\{\kappa(u_1), \ldots, \kappa(u_n)\}$ e $\boldsymbol{\Sigma}^* = \mathbf{L}_{(u)}\boldsymbol{\Sigma}$.

## 2.2 Inferência Bayesiana

Nesta seção especificamos as distribuições a priori para os parâmetros do modelo (2.1) e descrevemos a implementação do algoritmo MCMC proposto com o objetivo de obter amostras da distribuição a posteriori. Além disso, apresentamos alguns critérios para a comparação de modelos.

### 2.2.1 Distribuições a priori

Com objetivo de obter inferências a posteriori para o modelo (2.1), precisamos especificar a distribuição a priori para $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)^T$. Assumimos que $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ e $\boldsymbol{\lambda}$ são independentes a priori. Daí, a distribuição a priori conjunta pode ser escrita como $\pi(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\beta})\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\lambda})$.

Vamos a supor distribuições a priori proprias, e por simplicidade assumimos que $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ e $\boldsymbol{\lambda}$ seguem distribuição normal assim:

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\beta}_0, \mathbf{S}_\beta), \quad \boldsymbol{\alpha} \sim \mathcal{N}_{K_1}(\boldsymbol{\alpha}_0, \tau_\alpha^2 \mathbf{I}_{K_1}), \quad \boldsymbol{\gamma} \sim \mathcal{N}_q(\boldsymbol{\gamma}_0, \mathbf{S}_\gamma), \quad \text{e} \quad \boldsymbol{\lambda} \sim \mathcal{N}_{K_2}(\boldsymbol{\lambda}_0, \tau_\lambda^2 \mathbf{I}_{K_2}),$$

com hiperparâmetros $\boldsymbol{\beta}_0$, $\boldsymbol{\alpha}_0$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\lambda}_0$, $\mathbf{S}_\beta$ e $\mathbf{S}_\gamma$ conhecidos; em que $\mathbf{I}_K$ é a matriz identidade de ordem $K$, e $\tau_\alpha^2 \sim \mathcal{IG}(a_{\tau_\alpha}, b_{\tau_\alpha})$ e $\tau_\lambda^2 \sim \mathcal{IG}(a_{\tau_\lambda}, b_{\tau_\lambda})$ seguem distribuições gama inversa com funções de densidade de probabilidades dadas por

$$p(\tau^2 | a_\tau, b_\tau) \propto (\tau^2)^{-a_\tau - 1} \exp\left(-\frac{b_\tau}{\tau^2}\right)$$

em que $a_{\tau_\alpha} > 0$, $b_{\tau_\alpha} > 0$, $a_{\tau_\lambda} > 0$ e $b_{\tau_\lambda} > 0$ são quantidades assumidas conhecidas.

### 2.2.2 Algoritmo MCMC

A seguir apresentamos a implementação do algoritmo MCMC com o objetivo de obter a inferência a posteriori para os parâmetros do modelo (2.1). Usando a função de verossimilhança definida em (2.3) e as distribuições a priori apresentadas anteriormente, é possível obter as distribuições condicionais completas para cada um dos parâmetros do modelo (2.1). Se a distribuição condicional segue uma forma simples ou reconhecível, geramos amostras diretamente; caso contrário, utilizamos o algoritmo Metropolis-Hastings. Portanto, o algoritmo proposto consiste em gerar iterativamente amostras dessas distribuições condicionais completas.

Inicialmente, usamos o esquema de dados aumentados (veja Tanner e Wong, 1987). Seja $y_i \sim \mathcal{SMN}(\mu_i, \sigma_i^2; H, \kappa)$ como no modelo (2.1), então temos que

$$y_i | U_i = u_i \sim \mathcal{N}(\mu_i, \kappa(u_i)\sigma_i^2), \qquad U_i \sim \mathcal{H}(\cdot|\boldsymbol{\eta}), \qquad i = 1, \ldots, n,$$

em que $u_i$ é uma variável latente não observável. O algoritmo MCMC segue os seguintes passos:

*Passo 1:* Inicializar os valores dos parâmetros $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\lambda}^{(0)})$.

*Passo 2:* Baseado em $\boldsymbol{\theta}^{(l)} = (\boldsymbol{\beta}^{(l)}, \boldsymbol{\alpha}^{(l)}, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\lambda}^{(l)})$ calcular $\boldsymbol{\mu}^{(l)} = (\mu_1^{(l)}, \ldots, \mu_n^{(l)})^\top$, em que $\mu_i^{(l)} = \mathbf{x}_i^\top \boldsymbol{\beta}^{(l)} + \mathbf{b}(v_i)^\top \boldsymbol{\alpha}^{(l)}$ para $i = 1, \ldots, n$.

*Passo 3:* Baseado em $\boldsymbol{\theta}^{(l)} = (\boldsymbol{\beta}^{(l)}, \boldsymbol{\alpha}^{(l)}, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\lambda}^{(l)})$ calcular $\boldsymbol{\Sigma}^{(l)} = \text{diag}\{\sigma_1^{2\,(l)}, \ldots, \sigma_n^{2\,(l)}\}$, em que $\sigma_i^{2\,(l)} = h(\mathbf{z}_i^\top \boldsymbol{\gamma}^{(l)} + \mathbf{d}(w_i)^\top \boldsymbol{\lambda}^{(l)})$ para $i = 1, \ldots, n$.

*Passo 4:* Gerar $u_i^{(l+1)}$, $i = 1, \ldots, n$ independentes de $p(u_i|\boldsymbol{\theta}^{(l)})$, a qual depende da distribuição de $U$ usada para obter a distribuição $\mathcal{SMN}$ da seguinte forma:

$(a)$ Distribuição Normal: $P[u_i = 1|\boldsymbol{\theta}^{(l)}] = 1$.

$(b)$ Distribuição $t$-Student:

$$p(u_i|\boldsymbol{\theta}^{(l)}) \propto u_i^{\frac{\eta+1}{2}-1} \exp\left\{-u_i\left[\frac{(y_i - \mu_i^{(l)})^2}{2\sigma_i^{2\,(l)}} + \frac{\eta}{2}\right]\right\}.$$

Logo, $u_i|\boldsymbol{\theta}^{(l)} \sim \mathcal{G}amma\left(\frac{\eta+1}{2}, \frac{(y_i - \mu_i^{(l)})^2}{2\sigma_i^{2\,(l)}} + \frac{\eta}{2}\right)$.

$(c)$ Distribuição Slash:

$$p(u_i|\boldsymbol{\theta}^{(l)}) \propto u_i^{\eta-\frac{1}{2}} \exp\left\{-u_i\left[\frac{(y_i-\mu_i^{(l)})^2}{2\sigma_i^{2\,(l)}}\right]\right\} I_{(0,1)}(u_i).$$

Então, $u_i|\boldsymbol{\theta} \sim TrunGamma\left(\eta+\frac{1}{2}, \frac{(y_i-\mu_r^{(l)})^2}{2\sigma_i^{2\,(l)}}; (0,1)\right)$, isto é, $u_i|\boldsymbol{\theta}$ segue uma distribuição gamma truncada (veja Nadarajah e Kotz, 2006).

$(d)$ Distribuição Normal Contaminada:

$$p(u_i|\boldsymbol{\theta}^{(l)}) = p_\eta \mathbb{I}_{(u_i=\eta_2)} + (1-p_\eta)\mathbb{I}_{(u_i=1)},$$

onde

$$p_\eta \propto \eta_1\eta_2^{1/2} \exp\left\{-\frac{\eta_2}{2}\left[\frac{(y_i-\mu_i^{(l)})^2}{\sigma_i^{2\,(l)}}\right]\right\} \quad \text{e}$$

$$(1-p_\eta) \propto (1-\eta_1)\exp\left\{-\frac{(y_i-\mu_i^{(l)})^2}{2\sigma_i^{2\,(l)}}\right\}.$$

$(e)$ Distribuição Laplace:

$$p(u_i|\boldsymbol{\theta}^{(l)}) \propto u_i^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\frac{(y_i-\mu_i^{(l)})^2}{\sigma_i^{(l)}\,u_i}+u_i\right]\right\}.$$

Portanto, $u_i|\boldsymbol{\theta}^{(l)} \sim \mathcal{GIG}\left(\frac{1}{2}, \frac{(y_i-\mu_i^{(l)})^2}{\sigma_i^{(l)}}, 1\right)$.

$(f)$ Distribuição Hiperbólica Simétrica:

$$p(u_i|\boldsymbol{\theta}^{(l)}) \propto u_i^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\frac{1}{u_i}\left(\frac{(y_i-\mu_i^{(l)})^2}{\sigma_i^{2\,(l)}}+1\right)+\eta u_i\right]\right\}.$$

Note que $u_i|\boldsymbol{\theta}^{(l)} \sim \mathcal{GIG}\left(\frac{1}{2}, \frac{(y_i-\mu_i^{(l)})^2}{\sigma_i^{2\,(l)}}+1, \eta\right)$.

*Passo 5:* Baseado em $\boldsymbol{\theta}^{(l)} = (\boldsymbol{\beta}^{(l)}, \boldsymbol{\alpha}^{(l)}, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\lambda}^{(l)})$ obter $\tau_\alpha^{2\,(l+1)}$ e $\tau_\lambda^{2\,(l+1)}$ da seguinte forma

$(a)$ Gerar $\tau_\alpha^{2\,(l+1)}$ de

$$p(\tau_\alpha^2|\boldsymbol{\alpha}^{(l)}) \propto (\tau_\alpha^2)^{-\frac{K_1}{2}-a_{\tau_\alpha}-1} \exp\left\{-\frac{1}{2\tau_\alpha^2}\left[2b_{\tau_\alpha}+(\boldsymbol{\alpha}^{(l)}-\boldsymbol{\alpha}_0)^T(\boldsymbol{\alpha}^{(l)}-\boldsymbol{\alpha}_0)\right]\right\}.$$

Logo, $\tau_\alpha^2|\boldsymbol{\alpha}^{(l)} \sim \mathcal{IG}\left(\frac{K_1}{2}+a_{\tau_\alpha}, \frac{2b_{\tau_\alpha}+(\boldsymbol{\alpha}^{(l)}-\boldsymbol{\alpha}_0)^T(\boldsymbol{\alpha}^{(l)}-\boldsymbol{\alpha}_0)}{2}\right)$.

$(b)$ Gerar $\tau_\lambda^{2\,(l+1)}$ de

$$p(\tau_\lambda^2|\boldsymbol{\lambda}^{(l)}) \propto (\tau_\lambda^2)^{-\frac{K_2}{2}-a_{\tau_\lambda}-1} \exp\left\{-\frac{1}{2\tau_\lambda^2}\left[2b_{\tau_\lambda}+(\boldsymbol{\lambda}^{(l)}-\boldsymbol{\lambda}_0)^T(\boldsymbol{\lambda}^{(l)}-\boldsymbol{\lambda}_0)\right]\right\}.$$

$$\text{Então, } \tau_\lambda^2|\boldsymbol{\lambda}^{(l)} \sim \mathcal{IG}\left(\frac{K_2}{2}+a_{\tau_\lambda}, \frac{2b_{\tau_\lambda}+(\boldsymbol{\lambda}^{(l)}-\boldsymbol{\lambda}_0)^T(\boldsymbol{\lambda}^{(l)}-\boldsymbol{\lambda}_0)}{2}\right).$$

*Passo 6:* Obter $\boldsymbol{\beta}^{(l+1)}$ e $\boldsymbol{\alpha}^{(l+1)}$ da seguinte maneira

$(a)$ Gerar $\boldsymbol{\beta}^{(l+1)}$ de

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}^{(l)},\boldsymbol{\gamma}^{(l)},\boldsymbol{\lambda}^{(l)},\mathbf{u}^{(l+1)}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0^*)^T\mathbf{S}_\beta^{*\,-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0^*)\right\},$$

em que $\mathbf{S}_\beta^* = \left[\mathbf{S}_\beta^{-1}+\mathbf{X}^T(\mathbf{L}_{(u)}^{(l+1)}\boldsymbol{\Sigma}^{(l)})^{-1}\mathbf{X}\right]^{-1}$ e $\boldsymbol{\beta}_0^* = \mathbf{S}_\beta^*\left[\mathbf{S}_\beta^{-1}\boldsymbol{\beta}_0+\mathbf{X}^T(\mathbf{L}_{(u)}^{(l+1)}\boldsymbol{\Sigma}^{(l)})^{-1}(\mathbf{y}-\mathbf{B}\boldsymbol{\alpha}^{(l)})\right]$. Logo, $\boldsymbol{\beta}|\boldsymbol{\alpha}^{(l)},\boldsymbol{\gamma}^{(l)},\boldsymbol{\lambda}^{(l)},\mathbf{u}^{(l+1)} \sim \mathcal{N}_p\left(\boldsymbol{\beta}_0^*,\mathbf{S}_\beta^*\right)$.

$(b)$ Gerar $\boldsymbol{\alpha}^{(l+1)}$ de

$$p(\boldsymbol{\alpha}|\tau_\alpha^{2\,(l+1)},\boldsymbol{\beta}^{(l+1)},\boldsymbol{\gamma}^{(l)},\boldsymbol{\lambda}^{(l)},\mathbf{u}^{(l+1)}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0^*)^T\mathbf{S}_\alpha^{*\,-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0^*)\right\},$$

onde $\mathbf{S}_\alpha^* = \left[(\tau_\alpha^{2\,(l+1)})^{-1}\mathbf{I}_{K_1}+\mathbf{B}^T(\mathbf{L}_{(u)}^{(l+1)}\boldsymbol{\Sigma}^{(l)})^{-1}\mathbf{B}\right]^{-1}$ e $\boldsymbol{\alpha}_0^* = \mathbf{S}_\alpha^*\left[(\tau_\alpha^{2\,(l+1)})^{-1}\mathbf{I}_{K_1}\boldsymbol{\alpha}_0+\mathbf{B}^T(\mathbf{L}_{(u)}^{(l+1)}\boldsymbol{\Sigma}^{(l)})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(l+1)})\right]$. Então, $\boldsymbol{\alpha}|\tau_\alpha^{2\,(l+1)},\boldsymbol{\beta}^{(l+1)},\boldsymbol{\gamma}^{(l)},\boldsymbol{\lambda}^{(l)},\mathbf{u}^{(l+1)} \sim \mathcal{N}_{K_1}(\boldsymbol{\alpha}_0^*,\mathbf{S}_\alpha^*)$.

*Passo 7:* Gerar $\boldsymbol{\gamma}^{(l+1)}$ e $\boldsymbol{\lambda}^{(l+1)}$ da seguinte forma

$(a)$ Gerar $\boldsymbol{\gamma}^{(l+1)}$ de

$$p(\boldsymbol{\gamma}|\boldsymbol{\lambda}^{(l)},\mathbf{u}^{(l+1)},\boldsymbol{\alpha}^{(l+1)},\boldsymbol{\beta}^{(l+1)}) \propto |\boldsymbol{\Sigma}_\gamma^{(l)}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(l+1)}-\mathbf{B}\boldsymbol{\alpha}^{(l+1)})^T (\mathbf{L}_{(u)}^{(l+1)}\boldsymbol{\Sigma}_\gamma^{(l)})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(l+1)}-\mathbf{B}\boldsymbol{\alpha}^{(l+1)})-\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^T\mathbf{S}_\gamma^{-1}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\right\}. \quad (2.4)$$

em que $\boldsymbol{\Sigma}_\gamma^{(l)} = \text{diag}\{h(\mathbf{z}_1^T\boldsymbol{\gamma}+\mathbf{d}(w_1)^T\boldsymbol{\lambda}^{(l)}),\ldots,h(\mathbf{z}_n^T\boldsymbol{\gamma}+\mathbf{d}(w_n)^T\boldsymbol{\lambda}^{(l)})\}$.

$(b)$ Gerar $\boldsymbol{\lambda}^{(l+1)}$ de

$$p(\boldsymbol{\lambda}|\mathbf{u}^{(l+1)},\boldsymbol{\alpha}^{(l+1)},\boldsymbol{\beta}^{(l+1)},\boldsymbol{\gamma}^{(l+1)}) \propto |\boldsymbol{\Sigma}_\lambda^{(l)}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(l+1)}-\mathbf{B}\boldsymbol{\alpha}^{(l+1)})^T (\mathbf{L}_{(u)}^{(l+1)}\boldsymbol{\Sigma}_\lambda^{(l)})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(l+1)}-\mathbf{B}\boldsymbol{\alpha}^{(l+1)})-\frac{1}{2\tau_\lambda^2}(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0)^T(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0)\right\}, \quad (2.5)$$

onde $\boldsymbol{\Sigma}_\lambda^{(l)} = \text{diag}\{h(\mathbf{z}_1^T\boldsymbol{\gamma}^{(l+1)}+\mathbf{d}(w_1)^T\boldsymbol{\lambda}),\ldots,h(\mathbf{z}_n^T\boldsymbol{\gamma}^{(l+1)}+\mathbf{d}(w_n)^T\boldsymbol{\lambda})\}$.

*Passo 8:* Repetir os pasos 2 - 7.

Pode-se observar que as distribuições condicionais apresentadas em (2.4) e (2.5) não tem uma forma reconhecida, o que torna complicado gerar observações a partir destas distribuições. Então, para obter observações geradas a partir destas distribuições usamos o algoritmo Metropolis-Hastings da seguinte forma

*Passo 1:* Para gerar $\boldsymbol{\gamma}^{(l+1)}$ consideramos a distribuição $\text{N}_q(\boldsymbol{\gamma}^{(l)},\sigma_\gamma^2[\boldsymbol{\Omega}_\gamma^{(l)}]^{-1})$ como distribuição proposta (veja Roberts, 1996), com $\sigma_\gamma^2$ considerada tal que a taxa média de aceitação esteja

entre 0.25 and 0.45 como sugerido em Gelman *et al.* (1995) e

$$\boldsymbol{\Omega}_{\gamma}^{(l)} = \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(l+1)} - \mathbf{b}(v_i)^T\boldsymbol{\alpha}^{(l+1)})^2}{h(\mathbf{z}_i^T\boldsymbol{\gamma}^{(l)} + \mathbf{d}(w_i)^T\boldsymbol{\lambda}^{(l)})}\mathbf{z}_i\mathbf{z}_i^T + \mathbf{S}_{\gamma}^{-1}.$$

O algoritmo Metropolis-Hastings é implementado para a iteração $(l+1)$ usando o valor atual $\boldsymbol{\gamma}^{(l)}$. Assim, o novo valor $\boldsymbol{\gamma}^*$ é gerado de $\mathrm{N}_q(\boldsymbol{\gamma}^{(l)}, \sigma_{\gamma}^2[\boldsymbol{\Omega}_{\gamma}^{(l)}]^{-1})$ o qual é aceito com a seguinte probabilidade

$$\min\left\{1, \frac{p(\boldsymbol{\gamma}^*|\boldsymbol{\lambda}^{(l)}, \mathbf{u}^{(l+1)}, \boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}}{p(\boldsymbol{\gamma}^{(l)}|\boldsymbol{\lambda}^{(l)}, \mathbf{u}^{(l+1)}, \boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}}\right\}.$$

*Passo 2:* De forma similar, para gerar $\boldsymbol{\lambda}^{(l+1)}$ consideramos $\mathrm{N}_{K_2}(\boldsymbol{\lambda}^{(l)}, \sigma_{\lambda}^2[\boldsymbol{\Omega}_{\lambda}^{(l)}]^{-1})$ como distribuição proposta, em que $\sigma_{\lambda}^2$ é tal que a taxa média de aceitação pertence ao intervalo (0.25, 0.45) e

$$\boldsymbol{\Omega}_{\lambda}^{(l)} = \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(l+1)} - \mathbf{b}(v_i)^T\boldsymbol{\alpha}^{(l+1)})^2}{h(\mathbf{z}_i^T\boldsymbol{\gamma}^{(l+1)} + \mathbf{d}(w_i)^T\boldsymbol{\lambda}^{(l)})}\mathbf{d}(w_i)\mathbf{d}(w_i)^T + (\tau_{\lambda}^{2\,(l+1)}\,\mathbf{I}_{K_2})^{-1}.$$

Logo, o algoritmo Metropolis-Hastings é implementado para a iteração $(l+1)$ usando o valor atual $\boldsymbol{\lambda}^{(l)}$. Então, um novo candidato $\boldsymbol{\lambda}^*$ é gerado de $\mathrm{N}_{K_2}(\boldsymbol{\lambda}^{(l)}, \sigma_{\lambda}^2[\boldsymbol{\Omega}_{\lambda}^{(l)}]^{-1})$ que é aceito com a probabilidade:

$$\min\left\{1, \frac{p(\boldsymbol{\lambda}^*|\mathbf{u}^{(l+1)}, \boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\gamma}^{(l+1)})}{p(\boldsymbol{\lambda}^{(l)}|\mathbf{u}^{(l+1)}, \boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\gamma}^{(l+1)})}\right\}.$$

Portanto, na convergência do algoritmo proposto acima é possível obter uma amostra de tamanho $J$ da distribuição a posteriori conjunta de $\boldsymbol{\theta}$. Essa amostra é usada para obter as inferências no modelo (2.1). Neste trabalho, consideramos as médias a posteriori de $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$, $\tau_{\alpha}^2$ e $\tau_{\lambda}^2$ como medidas de resumo, denotadas por $\bar{\boldsymbol{\beta}}$, $\bar{\boldsymbol{\alpha}}$, $\bar{\boldsymbol{\gamma}}$, $\bar{\boldsymbol{\lambda}}$, $\bar{\tau}_{\alpha}^2$ and $\bar{\tau}_{\lambda}^2$. Geyer (1992) mostrou que $\hat{\boldsymbol{\theta}} = (\bar{\boldsymbol{\beta}}^T, \bar{\boldsymbol{\alpha}}^T, \bar{\boldsymbol{\gamma}}^T, \bar{\boldsymbol{\lambda}}^T, \bar{\tau}_{\alpha}^2, \bar{\tau}_{\lambda}^2)^T$ é uma estimativa consistente da média a posteriori do vetor $\boldsymbol{\theta}$ quando $J$ vai para infinito. Da mesma forma, o desvio padrão para cada parâmetro pode ser calculado como uma medida de resumo da distribuição a posteriori de $\boldsymbol{\theta}$.

## 2.3 Seleção de modelos e medidas de influência

Uma análise de dados deve considerar a avaliação e escolha do modelo que melhor descreve a situação em estudo. O objetivo desta seção é discutir, como do ponto do vista Bayesiano, é possível responder as seguintes perguntas:

$(i)$ O modelo estudado é adequado?

$(ii)$ Dada uma série de modelos, qual deles é o melhor?

### 2.3.1 Critérios de comparação de modelos

Na literatura existem várias metodologias para a comparação de modelos alternativos e não existe um acordo em relação a qual delas é a melhor perspectiva para a seleção de modelos na inferência Bayesiana. Neste trabalho, consideramos o DIC (Deviance Information Criterion) apresentado por Spiegelhalter *et al.* (2002), o EAIC (Expected Aikaike Information Criterion), EBIC (Expected Bayesian Information Criterion) ambos os dois propostos por Brooks (2002) e

CPO (Conditional Predictive Ordinate) estudado por Gelfand *et al.* (1992). As medidas anteriores podem ser estimadas usando uma única amostra da distribuição a posteriori gerada pelo algoritmo MCMC proposto na seção anterior.

O DIC é uma generalização do AIC (Akaike Information Criterion) que está baseado na média a posteriori do desvío. É possível estimar o DIC da seguinte forma

$$\widehat{\text{DIC}} = 2\bar{\text{D}} - \text{D}(\bar{\theta}), \tag{2.6}$$

em que $\bar{\text{D}} = J^{-1}\sum_{j=1}^{J}\text{D}(\theta^{(j)})$ e $\bar{\theta} = J^{-1}\sum_{j=1}^{J}\theta^{(j)}$, com $\text{D}(\theta) = -2\sum_{i=1}^{n}\log f(y_i|\theta)$ e $\theta^{(j)}$ o $j$-ésimo elemento da amostra a posteriori de $\theta$, $j = 1, \ldots, J$.

Os critérios EAIC e EBIC podem ser estimados usando

$$\widehat{\text{EAIC}} = \bar{\text{D}} + 2(p + q + \text{tr}(\mathbf{H}_\alpha) + \text{tr}(\mathbf{H}_\lambda)) \quad \text{e}$$
$$\widehat{\text{EBIC}} = \bar{\text{D}} + \log(n)(p + q + \text{tr}(\mathbf{H}_\alpha) + \text{tr}(\mathbf{H}_\lambda)),$$

em que $\mathbf{H}_\alpha = \mathbf{B}[\mathbf{B}^T\mathbf{B} + (1/\bar{\tau}_\alpha^2)\mathbf{I}_{K_1}]^{-1}\mathbf{B}^T$, $\mathbf{H}_\lambda = \mathbf{D}[\mathbf{D}^T\mathbf{D} + (1/\bar{\tau}_\lambda^2)\mathbf{I}_{K_2}]^{-1}\mathbf{D}^T$ e $\text{tr}(\mathbf{H})$ é o traço de $\mathbf{H}$. Nos três critérios anteriores o melhor ajuste é considerado pelo menor valor da estatística.

O CPO é uma estatística baseada no critério de validação cruzada para a comparação de modelos. Esta medida é a densidade preditiva de uma observação condicionada ao restante dos dados, sendo assím, para $i$-ésima observação o $\text{CPO}_i$ pode ser escrito como

$$\text{CPO}_i = p(y_i|y_{(-i)}) = \int f(y_i|\theta)\pi(\theta|y_{(-i)})\partial\theta = \left[\int \frac{\pi(\theta)|\mathbf{y}}{f(y_i|\theta)}\right]^{-1}. \tag{2.7}$$

Uma estimação de Monte Carlo do $\text{CPO}_i$ para uma amostra de tamanho $J$ é dada por

$$\widehat{\text{CPO}_i} = \left[\frac{1}{J}\sum_{j=1}^{J}\frac{1}{f(y_i|\theta^j)}\right]^{-1}.$$

Note que existe um valor da medida CPO para cada observação da amostra. Então, podemos definir uma estatística que resume os valores $\text{CPO}_i$ chamada de log-pseudo verossimilhança marginal e é expressa por:

$$\text{LMPL} = \sum_{i=1}^{n}\log(\widehat{\text{CPO}_i}),$$

em que o melhor ajuste é obtido pelo maior valor desta medida. Se o objetivo é comparar dois modelos $\text{M}_1$ e $\text{M}_2$, podemos usar a LMPL para calcular o pseudo fator de Bayes (PBF) definido por

$$\text{PBF}(\text{M}_1, \text{M}_2) = \exp(\text{LMPL}_1 - \text{LMPL}_2),$$

em que $\text{LMPL}_i$ é a log-verossimilhança marginal para o modelo $\text{M}_i$, $i = 1, 2$.

### 2.3.2 Resíduos

Consideramos o resíduo quantil, proposto por Dunn e Smyth (1996) no contexto de inferência clássica para modelos de regressão com respostas independentes do tipo contínuo ou discreto, mas adaptado para o caso Bayesiano neste trabalho. O resíduo quantil pode ser escrito como

$$r_{q,i} = \Phi^{-1}[F(y_i; \mu_i(\bar{\theta}), \sigma_i^2(\bar{\theta}))],$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada (fda) da normal padrão e $F(y_i;\cdot,\cdot)$ é a fda de $y_i$. Se $\bar{\theta}$ e $\theta$ coincidem, então $r_{q,1},\ldots,r_{q,n}$ representam uma amostra aleatória de uma distribuição normal padrão. Os resíduos quantil podem ser usados de forma semelhante que na inferência clássica para avaliar a qualidade de ajuste do modelo estudado.

### 2.3.3 Influência

Neste trabalho usamos as medidas de influência global para a análise de modelos de regressão sob a abordagem Bayesiano propostas por Weiss e Cook (1992) e Peng e Dey (1995), com o objetivo de desenvolver o diagnóstico para o modelo (2.1).

Para a análise de diagnóstico considerando deleção de casos utilizaremos a função de perturbação proposta Weiss (1996) e a medida de divergência-g proposta por Csiszar (1967). Temos que $\pi(\boldsymbol{\theta}|\mathbf{y})$ e $\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})$ são as distribuições a posteriori de $\boldsymbol{\theta}$ com todos os dados e sem a $i$-ésima observação, respectivamente. Daí, a função de perturbação é definida por

$$m_i(\theta) = \frac{\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})}{\pi(\boldsymbol{\theta}|\mathbf{y})},$$

e a medida de divergência-g entre duas densidades $\pi_1$ e $\pi_2$ avaliadas em $\boldsymbol{\theta}$ é dada por

$$d_g(\pi_1,\pi_2) = \mathrm{E}_{\theta|y}\left[g\left(\frac{\pi_1(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})}\right)\right],$$

em que $g(\cdot)$ é uma função convexa tal que $g(1) = 0$. Usando a anterior medida é possível definir algumas medidas de divergência específicas, considerando casos particulares para a função $g(\cdot)$. Alguns exemplos destas medidas são apresentados a seguer.

**Divergência de Kullback-Leibler** ($\mathrm{K}(\pi_1,\pi_2)$)

Esta medida de divergência entre $\pi_1$ e $\pi_2$ tem-se quando $g(\pi_1,\pi_2) = -\log\left(\frac{\pi_1}{\pi_2}\right)$, em que $\pi_1 = \pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})$ e $\pi_2 = \pi(\boldsymbol{\theta}|\mathbf{y})$, então

$$\mathrm{K}(\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)}),\pi(\boldsymbol{\theta}|\mathbf{y})) = \mathrm{E}_{\theta|y}\left[-\log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\right)\right].$$

**Distância-J** ($\mathrm{J}(\pi_1,\pi_2)$)

Esta medida é a versão simétrica da medida de divergência Kullback-Leibler, neste caso consideramos a função

$$g(\pi_1,\pi_2) = \left(\frac{\pi_1}{\pi_2} - 1\right)\log\left(\frac{\pi_1}{\pi_2}\right),$$

então, esta medida pode ser expressa por

$$\mathrm{J}(\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)}),\pi(\boldsymbol{\theta}|\mathbf{y})) = \mathrm{E}_{\theta|y}\left[\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})}{\pi(\boldsymbol{\theta}|\mathbf{y})} - 1\right)\log\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\right)\right].$$

**Distância-$\chi^2$** ($\chi^2(\pi_1,\pi_2)$)

Podemos obter esta medida quando $g(\pi_1,\pi_2) = \left(\frac{\pi_1}{\pi_2} - 1\right)^2$, logo

$$\chi^2(\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)}),\pi(\boldsymbol{\theta}|\mathbf{y})) = \mathrm{E}_{\theta|y}\left[\left(\frac{\pi(\boldsymbol{\theta}|\mathbf{y}_{(-i)})}{\pi(\boldsymbol{\theta}|\mathbf{y})} - 1\right)^2\right].$$

As medidas de influência apresentadas anteriormente podem ser calculadas usando uma

única amostra da distribuição a posterior de $\boldsymbol{\theta}$ gerada pelo algoritmo MCMC.

## 2.4 Estudo de simulação

Nesta seção apresentamos um estudo de simulação com o objetivo de ilustrar o desempenho do algoritmo MCMC proposto em 2.2.2, em particular, queremos avaliar o comportamento das estimativas Bayesianas (médias a posteriori) dos parâmetros e das funções não paramétricas do modelo (2.1). Geramos $n = 100$ observações de acordo com o seguinte mecanismo:

$$\begin{cases} y_i \overset{ind}{\sim} \mathcal{SMN}(\mu_i, \sigma_i^2; H, \kappa), \\ \mu_i = \mathbf{x}_i^T\boldsymbol{\beta} + \frac{1}{2}\text{sin}(2\pi v_i), \\ \log(\sigma_i^2) = \mathbf{z}_i^T\boldsymbol{\gamma} + \frac{1}{2}\text{sin}(2\pi w_i), \end{cases} \tag{2.8}$$

em que $\mathbf{x}_i$ é um vetor $(3 \times 1)$ com elementos independentes gerados seguindo $\mathcal{U}(-1, 1)$, $v_i \sim \mathcal{U}(0, 1)$, $\mathbf{z}_i$ é um vetor $(3 \times 1)$ com elementos independentes gerados seguindo $\mathcal{U}(-1, 1)$, $w_i \sim \mathcal{U}(0, 1)$, $\boldsymbol{\beta} = (1, -0.5, 0.5)$ e $\boldsymbol{\gamma} = (1, -0.5, 0.5)$. Para $y_i$ foram consideradas as seguintes distribuições:

- $\mathcal{N}(\mu_i, \sigma_i^2)$;
- $t(\mu_i, \sigma_i^2, \eta)$ para $\eta = 3, 5, 8, 12$;
- $\mathcal{S}l(\mu_i, \sigma_i^2, \eta)$ para $\eta = 2, 4, 7, 11$;
- $\mathcal{S}H(\mu_i, \sigma_i^2, \eta)$ para $\eta = 0.8, 0.9, 1.0, 1.1$;
- $\mathcal{CN}(\mu_i, \sigma_i^2, \boldsymbol{\eta})$ para $\eta_1 = 0.6, 0.5, 0.55, 0.4$ and $\eta_2 = 0.2$; e
- $\mathcal{L}aplace(\mu_i, \sigma_i^2)$.

Logo, amostramos a distribuição a posteriori dos parâmetros de interesse do seguinte modelo

$$\begin{cases} y_i \overset{ind}{\sim} \mathcal{SMN}(\mu_i, \sigma_i^2; H, \kappa), \\ \mu_i = \mathbf{x}_i^T\boldsymbol{\beta} + f(v_i), \\ \log(\sigma_i^2) = \mathbf{z}_i^T\boldsymbol{\gamma} + g(w_i), \end{cases} \tag{2.9}$$

em que $f(\cdot)$ e $g(\cdot)$ são aproximadas usando os $B$-splines. As distribuições a priori para os parâmetros foram consideradas como foram especificadas em 2.2.1, usando os seguintes valores para os hiperparâmetros: $\boldsymbol{\beta}_0 = \mathbf{0}_3$, $\mathbf{S}_\beta = 10^5 \times \mathbf{I}_3$, $\boldsymbol{\alpha}_0 = \mathbf{0}_{K_1}$, $\boldsymbol{\gamma}_0 = \mathbf{0}_3$, $\mathbf{S}_\gamma = 10^5 \times \mathbf{I}_3$, $\boldsymbol{\lambda}_0 = \mathbf{0}_{K_2}$, $a_{\tau_\alpha} = b_{\tau_\alpha} = a_{\tau_\lambda} = b_{\tau_\lambda} = 10^{-5}$. No procedimento de estimação via MCMC foi considerado um aquecimento (*burn-in*) de tamanho 10000 e em seguida gerada uma amostra de tamanho 100000 com saltos de tamanho 20. O anterior com o objetivo de reduzir a autocorrelação entre as cadeias para assim obter uma amostra de Monte Carlo aproximadamente independente de tamanho $J = 5000$. Este processo foi replicado $R = 100$ vezes mantendo fixos os valores de $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{v}$ e $\mathbf{w}$. Vale salientar que, o estudo de simulação foi implementado usando a função `gesm(·)` do pacote **BayesGESM** (para mais detalhes deste pacote do R veja o capítulo 5). Como medidas de resumo usamos

$$\text{M}(\boldsymbol{\beta}) = \frac{1}{R}\sum_{r=1}^{R}\bar{\boldsymbol{\beta}}^{(r)}, \qquad \text{D}(\boldsymbol{\beta}) = \left\{\frac{1}{R-1}\sum_{r=1}^{R}[\bar{\boldsymbol{\beta}}^{(r)} - \text{M}(\boldsymbol{\beta})]^2\right\}^{1/2},$$

$$\text{M}(\boldsymbol{\gamma}) = \frac{1}{R}\sum_{r=1}^{R}\bar{\boldsymbol{\gamma}}^{(r)}, \qquad \text{D}(\boldsymbol{\gamma}) = \left\{\frac{1}{R-1}\sum_{r=1}^{R}[\bar{\boldsymbol{\gamma}}^{(r)} - \text{M}(\boldsymbol{\gamma})]^2\right\}^{1/2},$$

em que $\bar{\boldsymbol{\beta}}^{(r)}$ e $\bar{\boldsymbol{\gamma}}^{(r)}$ são as médias a posteriori de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ na réplica $r$ ($r = 1, \ldots, R$), respectivamente. Para estimar as componentes não paramétricas do modelo 2.9, utilizamos as seguintes medidas de resumo

$$\bar{\hat{f}}(v) = \frac{1}{R}\sum_{r=1}^{R} \mathbf{B}\,\bar{\boldsymbol{\alpha}}^{(r)}, \qquad \text{and} \qquad \bar{\hat{g}}(w) = \frac{1}{R}\sum_{r=1}^{R} \mathbf{D}\,\bar{\boldsymbol{\lambda}}^{(r)},$$

em que $\bar{\boldsymbol{\alpha}}^{(r)}$ e $\bar{\boldsymbol{\lambda}}^{(r)}$ são as médias a posterioro de $\boldsymbol{\alpha}$ e $\boldsymbol{\lambda}$ na réplica $r$ ($r = 1, \ldots, R$), respectivamente.

Na Tabela 2.1 apresentamos os valores das medidas de resumo $\mathrm{M}(\cdot)$ e $\mathrm{D}(\cdot)$ para todos os cenários de simulação. Podemos notar que em todos os casos os valores de $\mathrm{M}(\cdot)$ estão muito próximos dos verdadeiros valores dos parâmetros do modelo. Além disso, em todos os casos os valores de $\mathrm{D}(\cdot)$ aumentam à medida que as caudas da distribuição do erro aleatório se tornam mais pesadas. Em forma geral, observamos que os valores da medida de resumo $\mathrm{D}(\cdot)$ são maiores para os parâmetros de dispersão do que para os parâmetros de localização.

Por outro lado, com o intuito de investigar a precisão das estimativas das funções $f(v)$ e $g(w)$, apresentamos nas Figuras 2.1 e 2.2 o verdadeiro valor dessas funções e as suas estimativas, representadas pelas linhas pontilhadas, nos diferentes cenários de simulação. Podemos observar que em todos os casos, os valores das funções não paramétricas estimadas estão muito próximos dos verdadeiros valores. Entretanto, observa-se que na maioria dos casos o comportamento da função não paramétrica estimada para o parâmetro de localização está mais próximo à verdadeira função do que para o parâmetro de dispersão. Estes resultados indicam um bom desempenho do algoritmo MCMC proposto.

## 2.5 Aplicação

Nesta seção aplicamos a metodologia proposta neste capítulo a um conjunto de dados reias. Este conjunto de dados foi apresentado originalmente em Dudzinski e Mykytowycz (1961) e analisado por Ratkowsky (1983) usando um modelo de regressão normal não linear. Os dados consistem de 71 observações de coelhos europeus (*Oryctolagus Cuniculus*) na Austrália, em que a variável resposta $y$ representa o peso das lentes (em mg) dos olhos dos coelhos e a variável explicativa $x$ corresponde à idade (em dias) do animal. No gráfico de dispersão de $y$ versus $x$ (omitido aqui), podemos observar que a variância não é constante, isto é, a variação do peso das lentes dos olhos aumenta com a idade do animal. Portanto, propomos que a distribuição do peso das lentes do olhos seja descrita usando um modelo em que a distribuição do erro pertence à classe geral de distribuições $\mathcal{SMN}$, onde seus parâmetros de localização e dispersão são descritos usando funções não paramétricas da idade do coelho. Então, supomos que o peso das lentes dos olhos do $i$-ésimo coelho pode ser descrita por

$$y_i = \mu_i + \sigma_i^2\,\epsilon_i, \qquad i = 1, \ldots, 71, \tag{2.10}$$

em que $\epsilon_i \sim \mathcal{SMN}(0, 1; H, \kappa)$, $\mu_i = f(x_i)$ e $\sigma_i^2 = \exp(g(x_i))$, onde $f(\cdot)$ e $g(\cdot)$ são funções não paramétricas aproximadas usando os $B$-splines. Note que $f(x_i)$ e $g(x_i)$ podem ser expressas por $f(x_i) = \mathbf{b}(x_i)^T\boldsymbol{\alpha}$ e $g(x_i) = \mathbf{d}(x_i)^T\boldsymbol{\lambda}$, respectivamente. As distribuições a priori consideradas para os parâmetros do modelo são as seguintes:

$$\begin{gathered}\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \tau_\alpha^2\mathbf{I}_{K_1}), \quad \boldsymbol{\lambda} \sim \mathcal{N}(\mathbf{0}, \tau_\lambda^2\mathbf{I}_{K_1}),\\ \tau_\alpha^2 \sim \mathcal{GI}(10^{-5}, 10^{-5}) \quad \text{and} \quad \tau_\lambda^2 \sim \mathcal{GI}(10^{-5}, 10^{-5}),\end{gathered}$$

em que os valores dos parâmetros foram considerados para obter prioris não informativas. A seguir, usamos o algoritmo MCMC proposto para obter uma amostra da distribuição a pos-

**Tabela 2.1:** *Valores das medidas de resumo* $\mathrm{M}(\cdot)$ *e* $\mathrm{D}(\cdot)$ *para todos os cenários de simulação.*

| Distribuição | Medida | Parâmetro | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ |
| $\mathcal{N}ormal$ | M | 0.9800 | -0.5433 | 0.5056 | 1.0226 | -0.5375 | 0.4827 |
| | D | 0.1321 | 0.1647 | 0.1728 | 0.2493 | 0.2432 | 0.3270 |
| $t(3)$ | M | 0.9968 | -0.5371 | 0.5126 | 1.0217 | -0.4622 | 0.6018 |
| | D | 0.1662 | 0.1928 | 0.1816 | 0.3726 | 0.3589 | 0.3735 |
| $t(5)$ | M | 1.0205 | -0.5599 | 0.5238 | 0.9837 | -0.5077 | 0.5251 |
| | D | 0.1518 | 0.1852 | 0.2008 | 0.3461 | 0.4074 | 0.3634 |
| $t(8)$ | M | 1.0072 | -0.5464 | 0.4945 | 1.0470 | -0.4262 | 0.5410 |
| | D | 0.1469 | 0.1772 | 0.1631 | 0.2834 | 0.3059 | 0.3062 |
| $t(12)$ | M | 0.9939 | -0.6060 | 0.5128 | 1.0040 | -0.5099 | 0.5744 |
| | D | 0.1563 | 0.1823 | 0.1452 | 0.2641 | 0.3150 | 0.2910 |
| $\mathcal{S}l(2)$ | M | 0.9577 | -0.5840 | 0.4778 | 0.9345 | -0.4291 | 0.4973 |
| | D | 0.2121 | 0.2365 | 0.2324 | 0.4441 | 0.4523 | 0.4026 |
| $\mathcal{S}l(4)$ | M | 0.9447 | -0.5594 | 0.5332 | 0.9849 | -0.5086 | 0.4881 |
| | D | 0.1625 | 0.1901 | 0.1712 | 0.3712 | 0.3698 | 0.3904 |
| $\mathcal{S}l(7)$ | M | 0.9898 | -0.5716 | 0.4968 | 0.9435 | -0.3486 | 0.5797 |
| | D | 0.1639 | 0.1804 | 0.1503 | 0.3509 | 0.3112 | 0.3324 |
| $\mathcal{S}l(11)$ | M | 0.9838 | -0.5573 | 0.5048 | 1.0150 | -0.4547 | 0.5887 |
| | D | 0.1582 | 0.1798 | 0.1541 | 0.3219 | 0.2973 | 0.3141 |
| $\mathcal{S}H(0.8)$ | M | 0.9970 | -0.5075 | 0.5364 | 1.0781 | -0.4525 | 0.5428 |
| | D | 0.2414 | 0.2520 | 0.2773 | 0.2725 | 0.3259 | 0.3477 |
| $\mathcal{S}H(0.9)$ | M | 0.9632 | -0.5432 | 0.5543 | 1.0510 | -0.4935 | 0.5903 |
| | D | 0.2624 | 0.2954 | 0.2958 | 0.2735 | 0.2772 | 0.3069 |
| $\mathcal{S}H(1.0)$ | M | 0.9988 | -0.5615 | 0.5271 | 1.0090 | -0.5432 | 0.5378 |
| | D | 0.2108 | 0.2636 | 0.2192 | 0.3018 | 0.2815 | 0.3015 |
| $\mathcal{S}H(1.1)$ | M | 0.9622 | -0.5815 | 0.5086 | 1.0086 | -0.5126 | 0.6004 |
| | D | 0.2289 | 0.2553 | 0.2066 | 0.2982 | 0.2826 | 0.2999 |
| $\mathcal{CN}(0.4, 0.2)$ | M | 0.9903 | -0.5334 | 0.5268 | 1.0347 | -0.5094 | 0.5304 |
| | D | 0.2368 | 0.2496 | 0.2888 | 0.3086 | 0.3590 | 0.3422 |
| $\mathcal{CN}(0.5, 0.2)$ | M | 0.9557 | -0.4679 | 0.5217 | 1.0359 | -0.4746 | 0.5695 |
| | D | 0.2778 | 0.2761 | 0.2608 | 0.3608 | 0.3859 | 0.3528 |
| $\mathcal{CN}(0.55, 0.2)$ | M | 0.9574 | -0.5598 | 0.6007 | 1.0145 | -0.4525 | 0.5219 |
| | D | 0.2665 | 0.3104 | 0.2815 | 0.3610 | 0.3234 | 0.3142 |
| $\mathcal{CN}(0.6, 0.2)$ | M | 0.9774 | -0.5147 | 0.5328 | 1.0948 | -0.5475 | 0.4912 |
| | D | 0.3008 | 0.3065 | 0.2767 | 0.3616 | 0.3320 | 0.3289 |
| Laplace | M | 0.9836 | -0.5982 | 0.4841 | 0.9532 | -0.4496 | 0.6042 |
| | D | 0.1820 | 0.1820 | 0.1708 | 0.3871 | 0.3916 | 0.3763 |

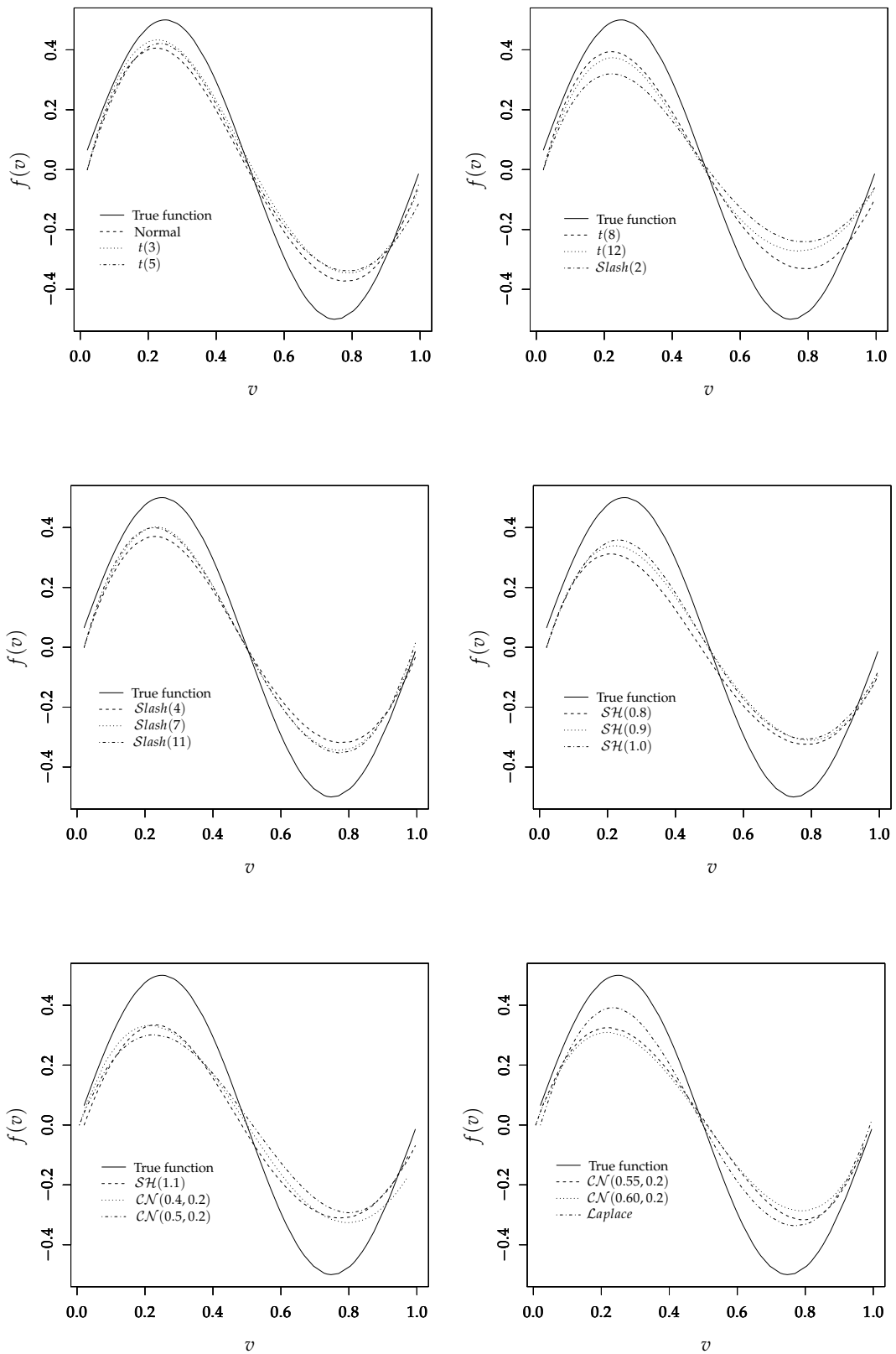


**Figura 2.1:** *Verdadeiro valor da função $f(v)$ contra suas estimativas (linhas pontilhadas) nos diferentes cenários de simulação.*

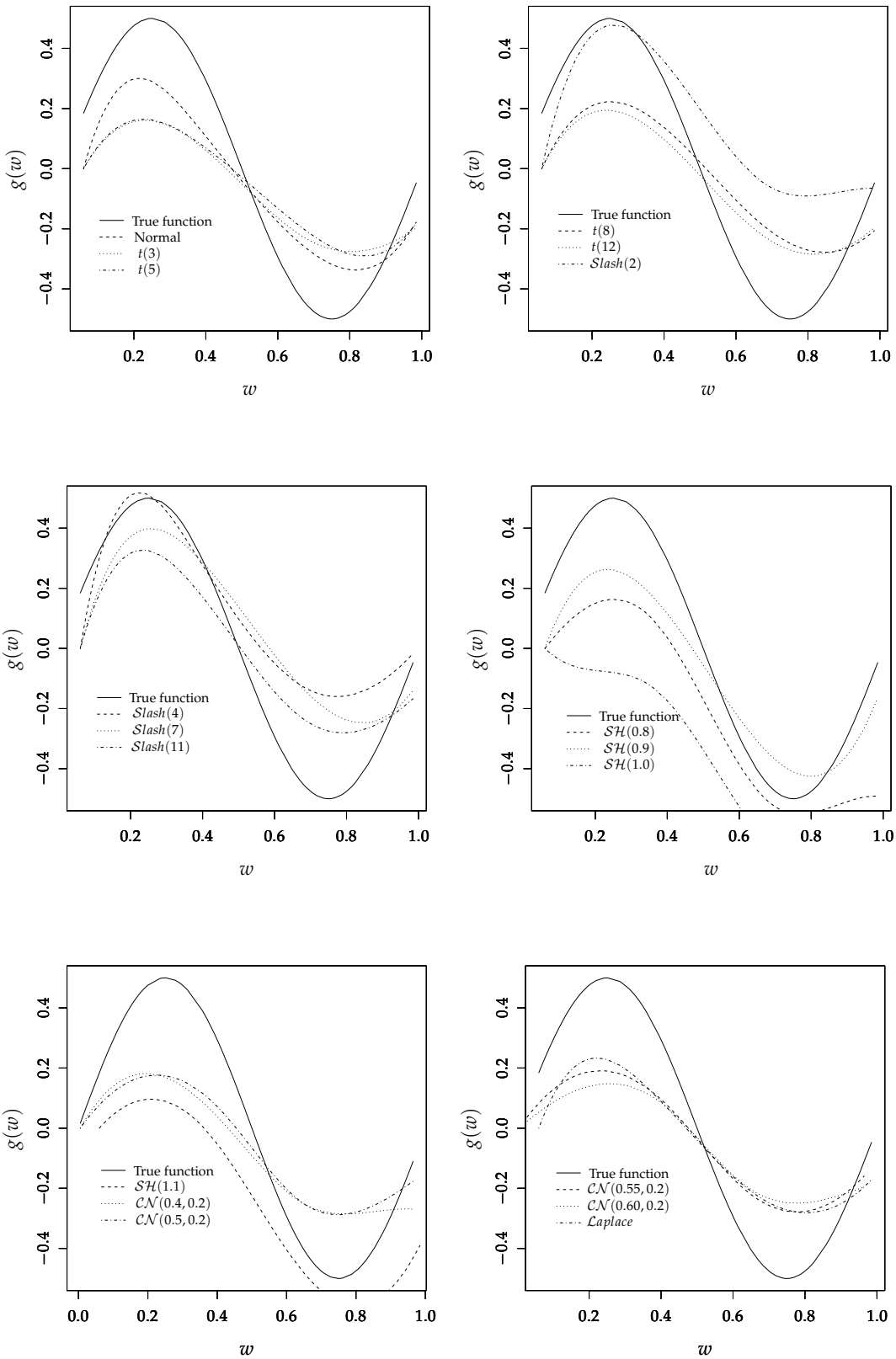


**Figura 2.2:** *Verdadeiro valor da função $g(w)$ contra suas estimativas (linhas pontilhadas) nos diferentes cenários de simulação.*

**Tabela 2.2:** *Critérios de seleção de modelos para o conjunto de dados dos Coelhos Europeus*

| Model | DIC | EAIC | EBIC | LMPL |
|---|---|---|---|---|
| $\mathcal{N}ormal$ | 495.9911 | 508.1499 | 500.5532 | -251.2792 |
| $\mathcal{S}tudent\text{-}t(14)$ | 497.1567 | 509.0596 | 501.4666 | -251.8771 |
| $\mathcal{S}lash(19)$ | 497.0911 | 508.7571 | 501.1643 | -252.4791 |
| $\mathcal{SH}(1)$ | 497.2766 | 509.7742 | 502.2335 | -252.0562 |
| $\mathcal{L}aplace$ | 499.2906 | 512.5431 | 505.1554 | -253.5180 |
| $\mathcal{CN}(0.8, 0.9)$ | **495.3543** | **507.8627** | **500.2670** | **-249.9310** |

teriori dos parâmetros de interesse. Simulamos uma cadeia de tamanho 55000, incluindo um período de aquecimento (*burn.in*) de 5000 iterações para eliminar o efeito dos valores iniciais, consideramos um espaçamento de 10 iteraçoes, obtendo uma amostra aproximadamente independente de tamanho $J = 5000$. Foram consideradas as distribuições normal, $t$-Student, slash, hiperbólica, normal contaminada e Laplace para o erro do modelo.

Na Tabela 2.2 apresentamos os valores dos critérios de comparação, apenas para os modelos que apresentaram o melhor ajuste (de acordo com as medidas apresentadas na seção 2.3.1) para cada distribuição do erro aleatório considerada. Em cada caso, vários valores do parâmetro extra foram considerados, e selecionamos aquele que apresentou os melhores valores nos critérios de comparação. Observa-se que todos os modelos selecionados apresentam caudas mais pesadas do que a distribuição normal. O melhor ajuste é obtido para o modelo *Normal Contaminado* com parâmetro extra $\boldsymbol{\eta} = c(0.8, 0.9)$ (isto é, $\mathcal{CN}(0.8, 0.9)$).

As Figuras 2.3(a) - 2.3(b) apresentam os gráficos das funções não paramétricas da média e da dispersão para o modelo ajustado $\mathcal{CN}(0.8, 0.9)$, podemos observar que o efeito da idade do animal sob a localização e dispersão da distribuição do peso das lentes do olhos é não linear. Neste gráfico também pode-se observar intervalos de credibilidade do 95% para as funções estimadas. Além disso, podemos observar que a dispersão do peso das lentes dos olhos não é monótona em relação à idade do animal. Considerando a amostra a posteriori dos parâmetros do modelo ajustado, calculamos os resíduos Bayesianos (veja a seção 2.3.2) e as medidas divergência, descritas na seção 2.3.3. Na Figura 2.3(c) apresentamos o QQ-plot dos resíduos do modelo $\mathcal{CN}$. O gráfico mostra o comportamento esperado, o que indica que este modelo descreve adequadamente o conjunto de dados. Além disso, o gráfico dos resíduos versus a variável explicativa (omitido aqui) sugere que não existem nem observações atípicas nem qualquer padrão ou tendência. Por outro lado, a Figura 2.3(d) apresenta o gráfico da medida de influência global, usando a medida de divergência Kullback-Leibler ($\mathrm{K}(\pi_1, \pi_2)$). As observações 4, 70, 71 são consideradas como influentes no modelo $\mathcal{CN}$. Os gráficos de outras medidas de divergência (omitidos aqui) permitem identificar as mesmas observações como influentes no modelo. Portanto, os métodos de diagnóstico indicam que o modelo $\mathcal{CN}$ permite obter um melhor ajuste do que o modelo normal. Note que o modelo ajustado considera o fato da variância não ser constante e caudas mais pesadas do que a normal, como sugerido por Wei (1998).

## 2.6 Conclusões

Neste capítulo estudamos a inferência estatística baseada na abordagem Bayesiana para modelos de regressão sob a suposição de erros independentes e aditivos seguindo distribuição normal, $t$-Student, slash, normal contaminada, Laplace e hipérbolica simétrica; em que os parâmetros de localização e dispersão da distribuição da variável resposta incluem componentes aditivos paramétricos e não paramétricos. Portanto, a metodologia desenvolvida neste capítulo estende as propostas de Cepeda e Gamerman (2001) e Xu e Zhang (2013).

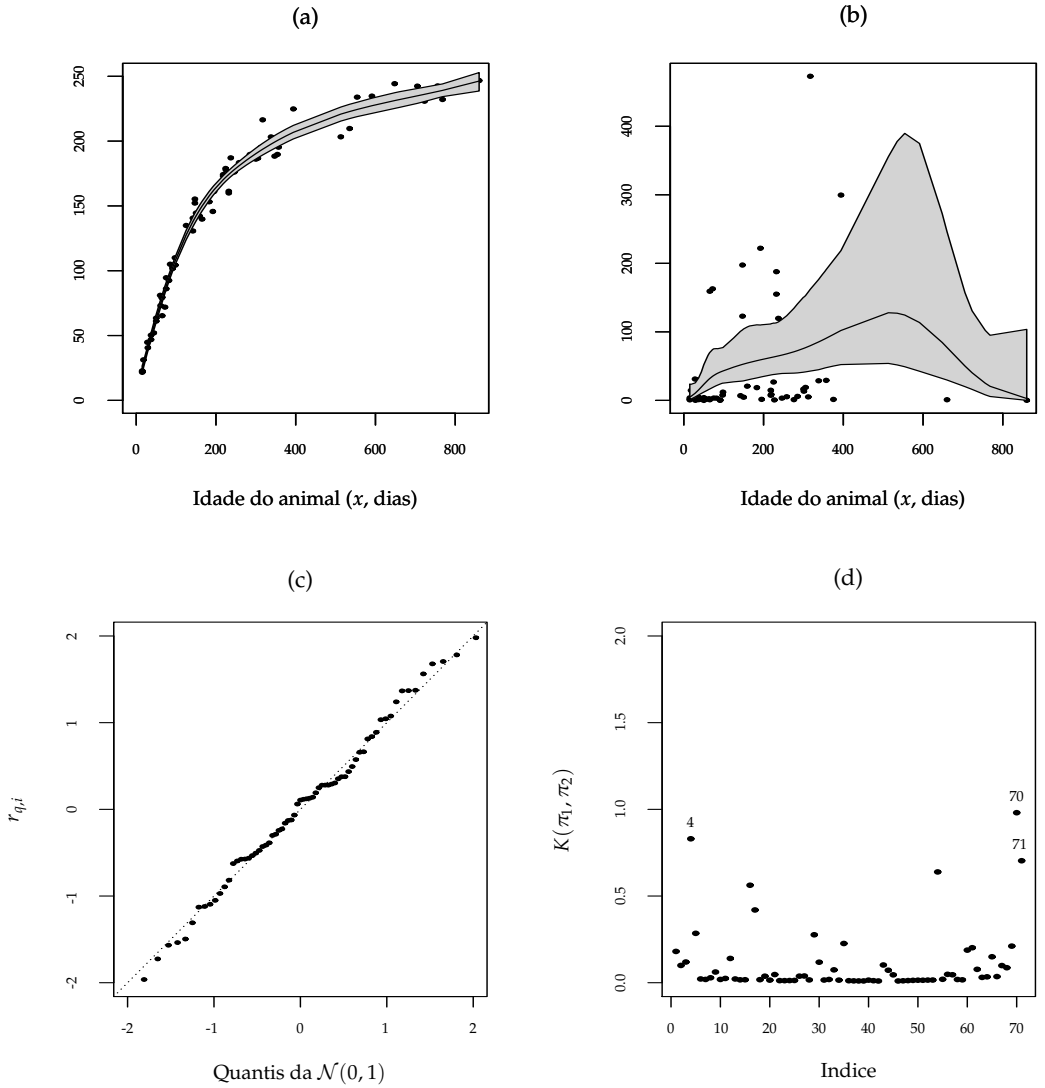


**Figura 2.3:** *Gráficos das funções não paramétricas para (a) média e (b) dispersão; (c) QQplot para os resíduos $r_{q,i}$ e (d) gráfico da medida $K(\pi_1, \pi_2)$ do modelo $\mathcal{CN}(\boldsymbol{\eta} = (0.8, 0.9))$ para os dados dos Coelhos Europeus*

Capítulo 3

# Modelo flexível com erros nas variáveis

Modelos de regressão com erros nas variáveis são uma extensão dos modelos clássicos de regressão, onde algumas das variáveis explicativas do modelo estão sujeitas a erro de medição, ou seja, no lugar de observarmos o vetor de covariáveis $\mathbf{m}_i$, observamos $\mathbf{M}_i = \mathbf{m}_i + \boldsymbol{\xi}_i$ para o indivíduo $i$, $i = 1, \ldots, n$, em que $\boldsymbol{\xi}_i$ é um vetor de erros aleatórios e $n$ é o tamanho da amostra. Fuller (1987) e Cheng e VanNess (1999) apresentam em detalhe vários modelos com erros nas variáveis sob a suposição de normalidade, tais como os modelos funcional (em que os $\mathbf{m}_i$ são fixados), estrutural (em que os $\mathbf{m}_i$ são variáveis aleatórias identicamente distribuidas), e ultra-estructural (em que a média bem como a variância dos $\mathbf{m}_i$ variam de um indivíduo para outro).

Neste capítulo apresentamos um modelo de regressão que admite variáveis explicativas com e sem erro de medição, bem como a presenca de efeitos não lineares aproximados através de $B$-splines. Este modelo recebe o nome de modelo flexível com erros nas variáveis. A componente aleatória deste modelo considera distribuições com caudas mais pesadas do que a distribuição normal multivariada, pois, esta componente é descrita usando vetores aleatórios obtidos como mistura normal na escala da distribuição normal multivariada, o qual proporciona flexibilidade bem como robustez na presença de observacões extremas. Como exemplos desta classe podemos citar as distribuições multivariadas $t$-Student, slash, Laplace, hiperbolica simétrica e normal contaminada. Nós estudamos a inferência nesta classe de modelos usando o enfoque Bayesiano, onde um algoritmo do tipo MCMC é desenvolvido para amostrar da distribuição a posteriori dos parâmetros de interesse. O algoritmo MCMC que propomos usa o esquema de dados aumentados (veja Tanner e Wong, 1987) para aproveitar o fato da distribuição do vetor aleatório ser obtida como uma mistura normal na escala da distribuição normal multivariada ($\mathcal{SMN}_r$). Também apresentamos alguns critérios para a comparação de modelos. Um estudo de simulação é desenvolvido com o intuito de ilustrar o desempenho do algoritmo MCMC proposto. A Seção 3.4 apresenta uma aplicação da metodologia proposta a dois conjuntos de dados reais. Na função `fmem()` do pacote **BayesGESM** (Rondon e Bolfarine, 2014) do R (R Core Team, 2014) implementamos computacionalmente a metodologia proposta. Uma descrição completa desta função pode ser encontrada no capítulo 5.

## 3.1 Formulação do modelo

Inicialmente, definimos o modelo semiparamétrico com erros nas variáveis para relacionar a variável resposta ($y$) com as variáveis explicativas, incluindo $s$ efeitos não paramétricos, os quais são aproximados usando $B$-splines. Mais especificamente, suponha que temos um con-

junto de $n$ observações geradas através do seguinte mecanismo

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{m}_i^T\boldsymbol{\rho} + \sum_{j=1}^{s} f_j(v_{ij}) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

em que $(\mathbf{x}_i^T, \mathbf{m}_i^T, v_{i1}, \ldots, v_{is})^T$ representa o vetor das variáveis explicativas associadas ao indivíduo $i$; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ e $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_q)^T$ são vetores de parâmetros desconhecidos a serem estimados; $f_j(\cdot)$ $(j = 1, \ldots, s)$ é uma função suave e contínua, mas desconhecida, que é aproximada usando $B$-splines e $\epsilon_i$ é o erro aleatório. Entretanto, o vetor de variáveis explicativas $\mathbf{m}_i$ não é observado diretamente para nenhum indivíduo (veja Cheng e VanNess, 1999), e no seu lugar só podemos obter uma "estimativa" desta variável, denotada por $\mathbf{M}_i$ e representada pela seguinte equação

$$\mathbf{M}_i = \mathbf{m}_i + \boldsymbol{\xi}_i, \qquad i = 1, \ldots, n,$$

em que $\boldsymbol{\xi}_i$ é um erro aleatório. Adicionalmente, assumimos que

$$\begin{pmatrix} \epsilon_i \\ \boldsymbol{\xi}_i \\ \mathbf{m}_i \end{pmatrix} \sim \mathcal{SMN}_{2q+1}\left[\begin{pmatrix} 0 \\ \mathbf{0} \\ \boldsymbol{\mu}_m \end{pmatrix}; \begin{pmatrix} \sigma_\epsilon^2 & 0 & 0 \\ 0 & \sigma_\xi^2\mathbf{I}_q & \mathbf{0} \\ 0 & \mathbf{0} & \boldsymbol{\Sigma}_m \end{pmatrix}\right], \quad i = 1, \ldots, n,$$

são vetores aleatórios independentes, onde $\boldsymbol{\mu}_m$, $\sigma_y^2$, $\sigma_\xi^2$ e $\boldsymbol{\Sigma}_m$ são parâmetros desconhecidos a serem estimados. O modelo definido acima é chamado de modelo flexível com erros nas variáveis (MFEV) e corresponde à versão estrutural do modelo com erros nas variáveis. Neste trabalho estamos supondo que $\omega = \sigma_\epsilon^2/\sigma_\xi^2$ é uma quantidade conhecida, o qual torna o modelo identificável. As componentes não paramétricas do modelo (3.1) são aproximadas usando $B$-splines (para detalhes veja de Boor, 1978). Então, os valores $f_j(v_{ij})$ podem ser aproximados por $\mathbf{b}_{ij}^T\boldsymbol{\alpha}_j$, em que $\mathbf{b}_{ij} = (b_{1j}(v_{ij}), \ldots, b_{K_jj}(v_{ij}))^T$ é o vetor das funções de base, $\boldsymbol{\alpha}_j \in \mathbb{R}^{K_j}$ é o vetor que contém os coeficientes do $B$-spline para o $j$ efeito não linear, e $K_j = M_j + k_j$, onde $M_j$ e $k_j$ são o grau e o número de nós internos que considera o $B$-spline, respectivamente.

## 3.2 Inferência Bayesiana

Sob a abordagem Bayesiana a inferência sobre os parâmetros de interesse baseia-se na distribuição a posteriori. Nesta seção, as distribuições a priori para os parâmetros do modelo são descritas, e um algoritmo MCMC é desenvolvido para gerar amostras da distribuição a posteriori dos parâmetros.

### 3.2.1 Distribuições a priori

Uma etapa importante na abordagem Bayesiana é a especificação das distribuições a priori para os parâmetros do modelo. Vamos supor a priori que os parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, $\boldsymbol{\alpha}_j$, $\boldsymbol{\mu}_m^T$, $\boldsymbol{\Sigma}_m$ e $\sigma_\epsilon^2$ são independentes e suas distribuições são:

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\beta}_0, \mathbf{S}_\beta), \quad \boldsymbol{\rho} \sim \mathcal{N}_q(\boldsymbol{\rho}_0, \mathbf{S}_\rho), \quad \boldsymbol{\alpha}_j \sim \mathcal{N}_{K_j}(\boldsymbol{\alpha}_{j0}, \tau_{\alpha_j}^2\,\mathbf{I}_{K_j}), \quad \boldsymbol{\mu}_m \sim \mathcal{N}_q(\boldsymbol{\mu}_{\mu_{m0}}, \boldsymbol{\Sigma}_{\mu_{m0}}),$$
$$\boldsymbol{\Sigma}_m^{-1} \sim \mathcal{W}ishart(q, \boldsymbol{\Omega}_m), \quad \sigma_\epsilon^2 \sim \mathcal{IG}\left(\frac{a}{2}, \frac{b}{2}\right), \quad \tau_{\alpha_j}^2 \sim \mathcal{IG}(a_{\tau_{\alpha_j}}, b_{\tau_{\alpha_j}}),$$

onde os hiperparâmetros $\boldsymbol{\beta}_0$, $\boldsymbol{\rho}_0$, $\boldsymbol{\alpha}_{j0}$, $\boldsymbol{\mu}_{\mu_{m0}}$, $\mathbf{S}_\beta > 0$, $\mathbf{S}_\rho > 0$, $\boldsymbol{\Sigma}_{\mu_{m0}} > 0$, $\boldsymbol{\Omega}_m > 0$, $a > 0$, $b > 0$, $a_{\tau_{\alpha_j}} > 0$ e $b_{\tau_{\alpha_j}} > 0$ $(j = 1, \ldots, s)$ são assumidos conhecidos. Note que $\tau_{\alpha_j}^2$ pode ser considerado como um parâmetro de suavização para a aproximação da função $f_j(\cdot)$ usando os $B$-splines. Valores "grandes" deste parâmetro indicam que a aproximação de $f_j(\cdot)$ é bastante "ondulada" ou "complexa". De forma similar, valores "pequenos" deste parâmetro indicam que a aproximação de $f_j(\cdot)$ está bastante próxima de uma linha reta.

### 3.2.2 Algoritmo MCMC

O algoritmo proposto a seguir aproveita o fato da distribuição do vetor aleatório ser obtida como uma mistura na escala da distribuição normal multivariada. De fato, o algoritmo usa o esquema de dados aumentados, supondo em cada iteração a presença das variáveis aleatórias não observáveis $u_i$ e $\mathbf{m}_i$ $(i = 1, \ldots, n)$. Portanto, de acordo com as especificações acima, a função de verossimilhança aumentada para o vetor de parâmetros $(\boldsymbol{\beta}^T, \boldsymbol{\rho}^T, \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_s^T, \boldsymbol{\mu}_m^T, \boldsymbol{\Sigma}_m, \sigma_\epsilon^2)$ pode ser escrita como

$$L(\boldsymbol{\beta}^T, \boldsymbol{\rho}^T, \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_s^T, \boldsymbol{\mu}_m^T, \boldsymbol{\Sigma}_m, \sigma_\epsilon^2 | \mathbf{y}, \mathbf{X}, \mathbf{M}, \mathbf{v}, \mathbf{u}, \mathbf{m}) \propto \prod_{i=1}^{n} (\sigma_\epsilon^2)^{-\frac{1}{2}} (\sigma_\xi^2)^{-\frac{q}{2}} \kappa(u_i)^{-q-\frac{1}{2}} |\boldsymbol{\Sigma}_m|^{-\frac{1}{2}}$$

$$\exp\left[ -\frac{(\mathbf{M}_i - \mathbf{m}_i)^T (\mathbf{M}_i - \mathbf{m}_i)}{2\sigma_\xi^2 \kappa(u_i)} - \frac{\left( y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{m}_i^T \boldsymbol{\rho} - \sum_{j=1}^{s} \mathbf{b}_{ij}^T \boldsymbol{\alpha}_j \right)^2}{2\sigma_\epsilon^2 \kappa(u_i)} \right.$$

$$\left. - \frac{(\mathbf{m}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\mathbf{m}_i - \boldsymbol{\mu}_m)}{2\kappa(u_i)} \right].$$

Combinando a função de verossimilhança com as distribuições a priori especificadas anteriormente, é possível obter as distribuições condicionais completas para cada um dos parâmetros do modelo (3.1). O algoritmo MCMC segue os seguintes passos:

*Passo 1:* Inicializar os valores dos parâmetros $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\rho}^{(0)}, \boldsymbol{\mu}_m^{(0)}, \boldsymbol{\Sigma}_m^{(0)}, \boldsymbol{\alpha}_1^{(0)}, \ldots, \boldsymbol{\alpha}_s^{(0)}, \sigma_\epsilon^{2\,(0)})$;

*Passo 2:* Calcular a quantidade $S_i^{(l)}$ $(i = 1, \ldots, n)$,

$$S_i = \frac{\left( y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{m}_i^T \boldsymbol{\rho} - \sum_{j=1}^{s} \mathbf{b}_{ij}^T \boldsymbol{\alpha}_j \right)^2}{\sigma_\epsilon^2} + \frac{(\mathbf{M}_i - \mathbf{m}_i)^T (\mathbf{M}_i - \mathbf{m}_i)}{\omega \sigma_\epsilon^2} + (\mathbf{m}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\mathbf{m}_i - \boldsymbol{\mu}_m),$$

em que $S_i^{(l)}$ representa $S_i$ avaliado em $\boldsymbol{\theta}^{(l)}$.

*Passo 3:* Gerar $u_i^{(l+1)} \sim p(u_i | S_i^{(l)})$, $i = 1, \ldots, n$, independentes, deacordo à distribuição da componente aleatória $(\epsilon, \boldsymbol{\xi}^T, \mathbf{m}^T)^T$:

$(a)$ Distribuição normal: $P(u_i = 1 | S_i^{(l)}) = 1$.

$(b)$ Distribuição $t$-Student:

$$p(u_i | S_i^{(l)}) \propto u_i^{\frac{\eta+1}{2}+q-1} \exp\left[ -\frac{u_i}{2} \left( S_i^{(l)} + \eta \right) \right],$$

Logo, $u_i | S_i^{(l)} \sim \mathcal{G}amma\left( \frac{\eta+1}{2} + q, \frac{S_i^{(l)} + \eta}{2} \right)$.

$(c)$ Distribuição slash:

$$p(u_i | S_i^{(l)}) \propto u_i^{q+\eta+\frac{1}{2}-1} \exp\left[ -\frac{u_i}{2}\ S_i^{(l)} \right] I_{(0,1)}(u_i).$$

Portanto, $u_i | S_i^{(l)} \sim Trun\mathcal{G}amma\left( \eta + q + \frac{1}{2}, \frac{S_i^{(l)}}{2}; (0,1) \right)$.

$(c)$ Distribuição normal contaminada:

$$p(u_i|S_i^{(l)}) = p_\eta \mathbb{I}_{(u_i=\eta_2)} + (1-p_\eta)\mathbb{I}_{(u_i=1)},$$

em que

$$p_\eta \propto \ \eta_1 \eta_2^{\frac{1}{2}+q} \exp\left\{-\frac{\eta_2\, S_i^{(l)}}{2}\right\} \qquad \text{e} \qquad (1-p_\eta) \propto \ (1-\eta_1)\exp\left\{-\frac{S_i^{(l)}}{2}\right\}.$$

$(d)$ Distribuição hiperbólica simétrica:

$$p(u_i|S_i^{(l)}) \propto u_i^{-\frac{1}{2}-q} \exp\left\{-\frac{1}{2}\left[\frac{S_i^{(l)}+1}{u_i} + \eta^2 u_i\right]\right\},$$

ou seja, $u_i|S_i^{(l)} \sim \mathcal{GIG}\left(-q+\frac{1}{2}, S_i^{(l)}+1, \eta^2\right)$.

$(e)$ Distribuição Laplace:

$$p(u_i|S_i^{(l)}) \propto u_i^{-\frac{1}{2}-q} \exp\left\{-\frac{1}{2}\left[\frac{S_i^{(l)}}{u_i} + \frac{u_i}{4}\right]\right\},$$

por conseguinte, $u_i|S_i^{(l)} \sim \mathcal{GIG}\left(-q+\frac{1}{2}, S_i^{(l)}, \frac{1}{4}\right)$.

*Passo 4:* Calcular a matriz $\mathbf{L}_u^{(l+1)} = \text{diag}\left\{u_1^{(l+1)}, \ldots, u_n^{(l+1)}\right\}$.

*Passo 5:* Gerar $\tilde{\boldsymbol{\beta}}^{(l+1)} \sim \mathcal{N}_{p+q}\left(\boldsymbol{\mu}_{\tilde{\beta}}, \boldsymbol{\Sigma}_{\tilde{\beta}}\right)$, em que

$$\boldsymbol{\Sigma}_{\tilde{\beta}} = \left[\begin{pmatrix} \mathbf{S}_\beta^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\rho^{-1} \end{pmatrix} + \frac{1}{\sigma_\epsilon^{2\,(l)}} \overline{\mathbf{X}}^{T(l)} \left[\mathbf{L}_u^{(l+1)}\right]^{-1} \overline{\mathbf{X}}^{(l)}\right]^{-1}, \qquad \text{e}$$

$$\boldsymbol{\mu}_{\tilde{\beta}} = \boldsymbol{\Sigma}_{\tilde{\beta}} \left[\begin{pmatrix} \mathbf{S}_\beta^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\rho^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\rho}_0 \end{pmatrix} + \frac{1}{\sigma_\epsilon^{2\,(l)}} \overline{\mathbf{X}}^{T(l)} \left[\mathbf{L}_u^{(l+1)}\right]^{-1} \left(\mathbf{y} - \sum_{j=1}^{s} \mathbf{B}_j \boldsymbol{\alpha}_j^{(l)}\right)\right],$$

com $\overline{\mathbf{X}}^{(l)} = \left[\mathbf{X}, \mathbf{m}^{(l)}\right]$, $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^T, \boldsymbol{\rho}^T)^T$ e $\mathbf{B}_j = (\mathbf{b}_{1j}, \ldots, \mathbf{b}_{nj})^T$.

*Passo 6:* Gerar a $i$-ésima linha de $\mathbf{m}^{(l+1)}$, denotada por $\mathbf{m}_i^{(l+1)}$, em que $\mathbf{m}_i^{(l+1)} \sim \mathcal{N}_q\left(\boldsymbol{\mu}_{m_i}, \boldsymbol{\Sigma}_{m_i}\right)$, com

$$\boldsymbol{\Sigma}_{m_i} = \left[\frac{\left(\boldsymbol{\Sigma}_m^{(l)}\right)^{-1}}{\kappa\left(u_i^{(l+1)}\right)} + \frac{\mathbf{I}_q}{\omega\sigma_\epsilon^{2\,(l)}\kappa\left(u_i^{(l+1)}\right)} + \frac{\boldsymbol{\rho}^{(l+1)}\boldsymbol{\rho}^{T\,(l+1)}}{\sigma_\epsilon^{2\,(l)}\kappa\left(u_i^{(l+1)}\right)}\right]^{-1} \qquad \text{e}$$

$$\boldsymbol{\mu}_{m_i} = \boldsymbol{\Sigma}_{m_i} \left[\frac{\left(\boldsymbol{\Sigma}_m^{(l)}\right)^{-1} \boldsymbol{\mu}_m^{(l)}}{\kappa\left(u_i^{(l+1)}\right)} + \frac{\mathbf{M}_i}{\omega\sigma_\epsilon^{2\,(l)}\kappa\left(u_i^{(l+1)}\right)} + \frac{\boldsymbol{\rho}^{(l+1)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(l+1)} - \sum_{j=1}^{s} \mathbf{b}_{ij}^T \boldsymbol{\alpha}_j^{(l)}\right)}{\sigma_\epsilon^{2\,(l)}\kappa\left(u_i^{(l+1)}\right)}\right].$$

*Passo 7:* Gerar $\boldsymbol{\mu}_m^{(l+1)} \sim \mathcal{N}_q\left(\boldsymbol{\mu}_{\mu_m}, \boldsymbol{\Sigma}_{\mu_m}\right)$, em que

$$\boldsymbol{\Sigma}_{\mu_m} = \left[\left(\boldsymbol{\Sigma}_m^{(l)}\right)^{-1}\left(\sum_{r=1}^{n}\frac{1}{\kappa\left(u_i^{(l+1)}\right)}\right) + \boldsymbol{\Sigma}_{\mu_{m0}}^{-1}\right]^{-1} \quad \text{e}$$

$$\boldsymbol{\mu}_{\mu_m} = \boldsymbol{\Sigma}_{\mu_m}\left[\left(\boldsymbol{\Sigma}_m^{(l)}\right)^{-1}\left(\sum_{r=1}^{n}\frac{\mathbf{m}_i^{(l+1)}}{\kappa\left(u_i^{(l+1)}\right)}\right) + \boldsymbol{\Sigma}_{\mu_{m0}}^{-1}\boldsymbol{\mu}_{\mu_{m0}}\right].$$

*Passo 8:* Gerar $\left(\boldsymbol{\Sigma}_m^{(l+1)}\right)^{-1} \sim \mathcal{W}ishart\,(q+n, \boldsymbol{\Omega}_m^*)$, em que

$$\boldsymbol{\Omega}_m^* = \left[\boldsymbol{\Omega}_m^{-1} + \sum_{r=1}^{n}\frac{\left(\mathbf{m}_i^{(l+1)} - \boldsymbol{\mu}_m^{(l+1)}\right)\left(\mathbf{m}_i^{(l+1)} - \boldsymbol{\mu}_m^{(l+1)}\right)^T}{\kappa\left(u_i^{(l+1)}\right)}\right]^{-1}.$$

*Passo 9:* Gerar $\tau_{\alpha_j}^{2\,(l+1)} \sim \mathcal{IG}\left(\frac{K_j}{2} + a_{\tau_{\alpha_j}}, \frac{2b_{\tau_{\alpha_j}} + \left(\boldsymbol{\alpha}_j^{(l)} - \boldsymbol{\alpha}_{j0}\right)^T\left(\boldsymbol{\alpha}_j^{(l)} - \boldsymbol{\alpha}_{j0}\right)}{2}\right)$, $j = 1, \ldots, s$.

*Passo 10:* Gerar $\boldsymbol{\alpha}_j^{(l+1)} \sim \mathcal{N}_{K_j}\left(\boldsymbol{\mu}_{\alpha_j}, \boldsymbol{\Sigma}_{\alpha_j}\right)$, $j = 1, \ldots, s$, em que

$$\boldsymbol{\Sigma}_{\alpha_j} = \left[\frac{1}{\tau_{\alpha_j}^{2\,(l+1)}}\mathbf{I}_{K_j} + \frac{1}{\sigma_\epsilon^{2\,(l)}}\mathbf{B}_j^T\left[\mathbf{L}_u^{(l+1)}\right]^{-1}\mathbf{B}_j\right]^{-1} \quad \text{e}$$

$$\boldsymbol{\mu}_{\alpha_j} = \boldsymbol{\Sigma}_{\alpha_j}\left[\frac{1}{\tau_{\alpha_j}^{2\,(l+1)}}\boldsymbol{\alpha}_{j0} + \frac{1}{\sigma_\epsilon^{2\,(l)}}\mathbf{B}_j^T\left[\mathbf{L}_u^{(l+1)}\right]^{-1}\left(\mathbf{y} - \overline{\mathbf{X}}^{(l+1)}\tilde{\boldsymbol{\beta}}^{(l+1)} - \sum_{0<i<j}\mathbf{B}_i\boldsymbol{\alpha}_i^{(l+1)} - \sum_{j<i\leq s}\mathbf{B}_i\boldsymbol{\alpha}_i^{(l)}\right)\right].$$

*Passo 11:* Gerar

$$\sigma_\epsilon^{2\,(l+1)} \sim \mathcal{IG}\left(\frac{n(1+q)+a}{2}, \frac{1}{2}\left[\sum_{r=1}^{n} b_i^{(l+1)} + b\right]\right),$$

em que $b_i = \frac{\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{m}_i^T\boldsymbol{\rho} - \sum_{j=1}^{s}\mathbf{b}_{ij}^T\boldsymbol{\alpha}_j\right)^2}{\kappa(u_i)} + \frac{(\mathbf{M}_i - \mathbf{m}_i)^T(\mathbf{M}_i - \mathbf{m}_i)}{\omega\kappa(u_i)}$, e $b_i^{(l+1)}$ é $b_i$ avaliado em $\boldsymbol{\beta}^{(l+1)}$, $\boldsymbol{\rho}^{(l+1)}$, $\mathbf{m}_i^{(l+1)}$, $\boldsymbol{\alpha}_1^{(l+1)}$, ..., $\boldsymbol{\alpha}_s^{(l+1)}$ e $u_i^{(l+1)}$.

*Passo 12:* Repetir os passos 2 - 11 até obter a convergência.

Portanto, seguindo o algoritmo MCMC é possível amostrar das distribuições a posteriori marginais dos parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_s$, $\boldsymbol{\mu}_m$ e $\sigma_\epsilon^2$. Sendo assim, com a amostra gerada de tamanho $R$, podemos resumir as distribuições a posteriori usando as seguintes medidas

$$\bar{\boldsymbol{\beta}} = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\beta}^{(r)}, \quad \bar{\boldsymbol{\rho}} = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\rho}^{(r)}, \quad \bar{\boldsymbol{\alpha}}_j = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\alpha}_j^{(r)}, \quad \bar{\boldsymbol{\mu}}_m = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\mu}_m^{(r)} \quad \text{e} \quad \bar{\sigma}_\epsilon^2 = \frac{1}{R}\sum_{r=1}^{R}\sigma_\epsilon^{2\,(r)}.$$

em que "(r)" representa o $r$-ésimo elemento da amostra a posteriori.

### 3.2.3 Parâmetro extra $\eta$ desconhecido

No algoritmo MCMC descrito na seção anterior assumimos que o parâmetro $\eta$ das distribuições $\mathcal{SMN}_r$ é conhecido. No entanto, quando ele não for conhecido é possível introduzir

um novo passo no algoritmo MCMC, denotado por *Passo 11.b*, para amostrar da distribuição a posteriori marginal deste parâmetro. A seguir descrevemos este passo para algumas das distribuições $\mathcal{SMN}_r$.

*Passo 11.b* Gerar $\boldsymbol{\eta}^{(l+1)} \sim p(\eta|\boldsymbol{\theta}^{(l)})$ segundo a distribuição da componente aleatória $(\epsilon, \boldsymbol{\xi}^T, \mathbf{m}^T)^T$:

$(a)$ Distribuição slash:

$$p(\eta|\boldsymbol{\theta}) \propto \eta^{n+a_\eta-1} \exp\left[-\eta\left(b_\eta - \sum_{i=1}^{n} \log u_i\right)\right].$$

Então, $\eta|\boldsymbol{\theta} \sim \mathcal{G}amma\left(n + a_\eta, b_\eta + \sum_{i=1}^{n} \log u_i\right)$. Neste caso, a distribuição a priori para $\eta$ é $\eta \sim \mathcal{G}amma(a_\eta, b_\eta)$, em que os hiperparâmetros $a_\eta > 0$ e $b_\eta > 0$ são assumidos conhecidos.

$(b)$ Distribuição normal contaminada: Temos que $\pi(\eta_1, \eta_2) = \pi(\eta_1)\pi(\eta_2)$, com $\eta_1 \sim \mathcal{B}eta(a_{\eta_1}, b_{\eta_1})$ e $\eta_2 \sim Trun\mathcal{G}amma(\frac{a_{\eta_2}}{2}, \frac{b_{\eta_2}}{2}; (0,1))$ são as distribuições a priori para $\boldsymbol{\eta}$, e

$$p(\eta_1|\boldsymbol{\theta}) \propto \eta_1^{a_{\eta_1}+g_\eta-1}(1-\eta_1)^{b_{\eta_1}+n-g_\eta-1}, \quad \text{e}$$
$$p(\eta_2|\boldsymbol{\theta}) \propto \eta_2^{\frac{g_\eta+a_{\eta_2}}{2}-1} \exp\left[-\frac{\eta_2}{2}\left(\sum_{i:u_i\in g(\eta)} S_i + b_{\eta_2}\right)\right] I_{(0,1)}(\eta_2),$$

em que $g_\eta = \sum_{i=1}^{n} U_i$ e $U_i = \begin{cases} 1 & \text{if} \quad u_i = \eta_2 \\ 0 & \text{if} \quad u_i = 1 \end{cases}$. Portanto, $\eta_1|\boldsymbol{\theta} \sim \mathcal{B}eta(a_{\eta_1} + g_\eta, b_{\eta_1} + n - g_\eta)$ e $\eta_2|\boldsymbol{\theta} \sim Trun\mathcal{G}amma\left(\frac{g_\eta+a_{\eta_2}}{2}, \frac{\sum_{i:u_i\in g(\eta)} S_i + b_{\eta_2}}{2}; (0,1)\right)$, em que os hiperparâmetros $a_{\eta_1} > 0$, $b_{\eta_1} > 0$, $a_{\eta_2} > 0$ e $b_{\eta_2} > 0$ são assumidos conhecidos.

$(c)$ Para as distribuições $t$-Student e hiperbólica simétrica é necessário introduzir um passo Metropolis-Hastings.

### 3.2.4 Critérios de comparação de modelos

Como medidas da qualidade do ajuste consideramos o critério de informação do desvio (DIC) e o CPO (Conditional Predictive Ordinate). Estes critérios ja foram definidos na seção 2.3.1, mas precisam ser adaptados para o modelo definido em (3.1). Os critérios DIC e CPO podem ser calculados usando as equações (2.6) e (2.7), respectivamente. Note que $\mathbf{y}_i^* = (y_i, \mathbf{M}_i^T)^T \sim \mathcal{SMN}_{q+1}(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*)$, em que

$$\boldsymbol{\mu}_i^* = \begin{pmatrix} \mathbf{x}_i^T\boldsymbol{\beta} + \sum_{j=1}^{s} \mathbf{b}_{ij}^T\boldsymbol{\alpha}_j + \boldsymbol{\mu}_m^T\boldsymbol{\rho} \\ \boldsymbol{\mu}_m \end{pmatrix} \quad \text{e} \quad \boldsymbol{\Sigma}_i^* = \begin{pmatrix} \sigma_\epsilon^2 + \boldsymbol{\rho}^T\boldsymbol{\Sigma}_m\boldsymbol{\rho} & \boldsymbol{\rho}^T\boldsymbol{\Sigma}_m \\ \boldsymbol{\rho}^T\boldsymbol{\Sigma}_m & \omega\sigma_\epsilon^2\mathbf{I}_q + \boldsymbol{\Sigma}_m \end{pmatrix}.$$

## 3.3 Estudo de simulação

Nesta seção apresentamos um estudo de simulação com o objetivo principal de ilustrar o desempenho do algoritmo proposto para o modelo (3.1). Uma amostra de tamanho $n = 500$ é gerada através do seguinte mecanismo:

$$\begin{cases} y_i = \beta_1 + \beta_2 \mathrm{x}_i + \rho \mathrm{m}_i + \frac{1}{2}\sin(2\pi v_i) + \epsilon_i, \\ \mathrm{M}_i = \mathrm{m}_i + \xi_i, \qquad i = 1, \ldots, n, \end{cases} \tag{3.2}$$

em que $(\epsilon_1, \xi_1, m_1)^T, \ldots, (\epsilon_n, \xi_n, m_n)^T$ são vetores aleatórios independentes e identicamente distribuidos cuja distribuição é $\mathcal{SMN}_3$; $x_i$ é gerada seguindo $\mathcal{U}(-1,1)$; $v_i \sim \mathcal{U}(0,1)$; $\beta_1 = \beta_2 = 1$, $\rho = 0.5$, $\sigma_\epsilon^2 = 1$, $\mu_m = 1$, $\sigma_m^2 = 0.5$ e $\omega = 1$. Para o termo aleatório $(\epsilon_i, \xi_i, m_i)^T \sim \mathcal{SMN}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta})$ foram consideradas as seguintes distribuições:

- $\mathcal{N}_3(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$;
- $t_3(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \eta)$ para $\eta = 3, 5, 8$, e 12;
- $\mathcal{S}l_3(\mu^*, \boldsymbol{\Sigma}^*, \eta)$ para $\eta = 2, 4, 7$, e 11;
- $SH_3(\mu^*, \boldsymbol{\Sigma}^*, \eta)$ para $\eta = 0.8, 1.0, 1.2$ e 1.4;
- $\mathcal{L}aplace_3(\mu^*, \boldsymbol{\Sigma}^*)$; e
- $\mathcal{C}N_3(\mu^*, \boldsymbol{\Sigma}^*, \boldsymbol{\eta})$ para $\boldsymbol{\eta} = (0.4, 0.2)^T$, $(0.5, 0.2)^T$, $(0.55, 0.2)^T$ e $(0.6, 0.2)^T$,

onde $\boldsymbol{\mu}^* = (0, 0, \mu_m)^T$ e $\boldsymbol{\Sigma}^* = \text{diag}(\sigma_\epsilon^2, \sigma_\xi^2, \sigma_m^2)$.

Então, amostramos a distribuição a posteriori dos parâmetros de interesse do modelo (3.2), em que a forma funcional de $f(v)$ é assumida como desconhecida mas aproximada usando $B$-splines cúbicos, com $[n^{1/5}]$ como o número de nós internos. Além disso, assumimos que os valores de $m_i$ são desconhecidos e as distribuições a priori para os parâmetros são consideradas como especificadas na seção 3.2.1, usando os seguintes hiperparâmetros: $\beta_0 = \rho_0 = \mu_{\mu_m} = 0$, $S_\beta = S_\rho = \Sigma_{\mu_m} = \Omega_m = 10^3$, $\boldsymbol{\alpha}_0 = \mathbf{0}_K$, e $a_{\tau_\alpha} = b_{\tau_\alpha} = a = b = 0.001$. Consideramos dois cenários, o primeiro em que $\boldsymbol{\eta}$ é conhecido. O segundo, em que assumimos que $\boldsymbol{\eta}$ é desconhecido. O procedimento MCMC com 55000 iterações foi implementado, o qual inclui um aquecimento de tamanho 5000 e saltos de tamanho 10, de modo que uma amostra de tamanho $R = 5000$ foi obtida. O procedimento foi réplicado $R = 100$ vezes, onde os valores de $\mathbf{x}$ e $\mathbf{v}$ são mantidos fixos. Como medidas de resumo consideramos as seguintes

$$\text{M}(\theta_j) = \frac{1}{100}\sum_{r=1}^{100}\bar{\theta}_j^{(r)}, \qquad \text{D}(\theta_j) = \left\{\frac{1}{99}\sum_{r=1}^{100}[\bar{\theta}_j^{(r)} - \text{M}(\theta_j)]^2\right\}^{1/2}, \tag{3.3}$$

em que $\theta_1 = \beta_1$, $\theta_2 = \beta_2$, $\theta_3 = \rho$, $\theta_4 = \mu_m$, $\theta_5 = \sigma_m^2$, $\theta_6 = \sigma_y^2$ e $\bar{\theta}_j^{(r)}$ são as médias a posteriori de $\theta_j$ na réplica $j$, $j = 1, \ldots, 100$. Para a componente não paramétrica consideramos a seguinte estatística de resumo

$$\hat{f}(v) = \frac{1}{100}\sum_{j=1}^{100}\mathbf{B}\,\bar{\boldsymbol{\alpha}}^{(j)}, \tag{3.4}$$

onde $\bar{\boldsymbol{\alpha}}^{(j)}$ é a média a posteriori de $\boldsymbol{\alpha}$ na réplica $j$.

As Tabelas 3.1 e 3.2 apresentam os valores das medidas resumo $\text{M}(\cdot)$ e $\text{D}(\cdot)$ para cada um dos parâmetros do modelo (3.2) nos cenários de simulação considerados. Note que os valores das estimativas dos parâmetros estão próximos dos valores verdadeiros. Uma exceção parece ser o parâmetro $\sigma_m^2$, em que para obter uma estimativa mais próxima do verdadeiro valor amostras maiores parecem ser necessárias. Além disso, em todos os casos os valores de $\text{D}(\cdot)$ aumentam à medida que as caudas da distribuição do termo do erro são mais pesadas. Em geral, observa-se que os valores de $\text{D}(\cdot)$ são maiores quando o parâmetro extra $\boldsymbol{\eta}$ é considerado desconhecido.

Nosso objetivo agora é estudar o desempenho das estimativas da função $f(v)$. Com essa perspectiva, as Figuras 3.1 e 3.2 apresentam a verdadeira função (linha cheia) e as suas estimativas (linhas pontilhadas), sob os diferentes cenários de simulação. Podemos concluir que as estimativas da função não paramétrica apresentam comportamento semelhante ao verdadeiro,

**Tabela 3.1:** *Valores das medidas de resumo* $\mathrm{M}(\cdot)$ *e* $\mathrm{D}(\cdot)$ *em que o parâmetro* $\boldsymbol{\eta}$ *é considerado conhecido.*

| Distribuição | Medida | Parâmetro | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\beta_1$ | $\beta_2$ | $\rho$ | $\mu_m$ | $\sigma_m^2$ | $\sigma_\epsilon^2$ |
| $\mathcal{N}ormal$ | M | 0.9760 | 1.0300 | 0.5287 | 0.9873 | 0.5088 | 1.0174 |
| | D | 0.1179 | 0.0822 | 0.1081 | 0.0504 | 0.1015 | 0.0733 |
| $t(3)$ | M | 0.9343 | 0.9618 | 0.5384 | 1.0089 | 0.6515 | 1.0752 |
| | D | 0.1476 | 0.1071 | 0.1319 | 0.0584 | 0.1793 | 0.1181 |
| $t(5)$ | M | 0.9391 | 1.0010 | 0.5632 | 0.9902 | 0.4731 | 1.0657 |
| | D | 0.2024 | 0.0888 | 0.2064 | 0.0561 | 0.1764 | 0.1225 |
| $t(8)$ | M | 0.8818 | 1.0111 | 0.6204 | 1.0159 | 0.4128 | 1.0630 |
| | D | 0.2713 | 0.0880 | 0.2603 | 0.0494 | 0.1794 | 0.1115 |
| $t(12)$ | M | 0.9337 | 0.9656 | 0.5668 | 0.9942 | 0.4699 | 1.0242 |
| | D | 0.1947 | 0.0965 | 0.1788 | 0.0492 | 0.1644 | 0.1079 |
| $\mathcal{S}l(2)$ | M | 0.8962 | 0.9762 | 0.5926 | 1.0242 | 0.3854 | 1.0589 |
| | D | 0.2051 | 0.0978 | 0.2009 | 0.0613 | 0.1559 | 0.1078 |
| $\mathcal{S}l(4)$ | M | 0.9191 | 1.0393 | 0.6019 | 0.9998 | 0.4361 | 1.0190 |
| | D | 0.2883 | 0.0885 | 0.2884 | 0.0537 | 0.1481 | 0.0803 |
| $\mathcal{S}l(7)$ | M | 0.9042 | 1.0012 | 0.5968 | 0.9968 | 0.3828 | 1.0321 |
| | D | 0.2705 | 0.0822 | 0.2709 | 0.0572 | 0.1508 | 0.0973 |
| $\mathcal{S}l(11)$ | M | 0.9074 | 0.9951 | 0.6047 | 0.9880 | 0.3859 | 1.0340 |
| | D | 0.2429 | 0.0808 | 0.2459 | 0.0477 | 0.1553 | 0.0970 |
| $\mathcal{S}H(0.8)$ | M | 0.8743 | 0.9508 | 0.6065 | 0.9817 | 0.3981 | 1.0589 |
| | D | 0.3390 | 0.1228 | 0.2948 | 0.0783 | 0.2037 | 0.1238 |
| $\mathcal{S}H(1.0)$ | M | 0.8890 | 0.9777 | 0.5996 | 1.0291 | 0.4293 | 1.0522 |
| | D | 0.2627 | 0.1116 | 0.2433 | 0.0811 | 0.1822 | 0.1215 |
| $\mathcal{S}H(1.2)$ | M | 0.9173 | 1.0390 | 0.5697 | 0.9940 | 0.4716 | 1.0432 |
| | D | 0.1995 | 0.0956 | 0.1906 | 0.0572 | 0.1487 | 0.0989 |
| $\mathcal{S}H(1.4)$ | M | 0.9082 | 0.9888 | 0.5954 | 1.0011 | 0.4375 | 1.0512 |
| | D | 0.2597 | 0.0864 | 0.2429 | 0.0576 | 0.1826 | 0.1123 |
| $\mathcal{CN}(0.4, 0.2)$ | M | 0.9504 | 1.0311 | 0.5472 | 0.9399 | 0.6946 | 1.1564 |
| | D | 0.1754 | 0.1121 | 0.1513 | 0.0732 | 0.1591 | 0.1083 |
| $\mathcal{CN}(0.5, 0.2)$ | M | 0.9427 | 0.9530 | 0.5483 | 1.0501 | 0.5042 | 1.0400 |
| | D | 0.1765 | 0.1216 | 0.1464 | 0.0836 | 0.1657 | 0.1069 |
| $\mathcal{CN}(0.55, 0.2)$ | M | 0.9504 | 0.9570 | 0.5418 | 1.0595 | 0.5315 | 1.0334 |
| | D | 0.1791 | 0.1291 | 0.1407 | 0.0859 | 0.1584 | 0.1009 |
| $\mathcal{CN}(0.6, 0.2)$ | M | 0.9525 | 0.9510 | 0.5377 | 1.0603 | 0.5523 | 1.0294 |
| | D | 0.1793 | 0.1335 | 0.1362 | 0.0888 | 0.1486 | 0.0903 |
| Laplace | M | 0.9620 | 0.9797 | 0.5435 | 1.0397 | 0.5746 | 1.0466 |
| | D | 0.1835 | 0.1597 | 0.1498 | 0.1333 | 0.1959 | 0.1184 |

**Tabela 3.2:** *Valores das medidas de resumo* $\mathrm{M}(\cdot)$ *e* $\mathrm{D}(\cdot)$ *em que o parâmetro* $\boldsymbol{\eta}$ *é considerado desconhecido.*

| Distribuição | Medida | Parâmetro | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $\beta_1$ | $\beta_2$ | $\rho$ | $\mu_m$ | $\sigma_m^2$ | $\sigma_\epsilon^2$ | $\boldsymbol{\eta}$ |
| $\mathcal{N}ormal$ | M | 0.9760 | 1.0300 | 0.5287 | 0.9873 | 0.5088 | 1.0174 | |
| | D | 0.1179 | 0.0822 | 0.1081 | 0.0504 | 0.1015 | 0.0733 | |
| $t(3)$ | M | 0.9245 | 0.9823 | 0.5593 | 1.0153 | 0.6662 | 1.1172 | 3.3528 |
| | D | 0.1888 | 0.0853 | 0.1735 | 0.0566 | 0.1924 | 0.1403 | 0.4223 |
| $t(5)$ | M | 0.8841 | 0.9894 | 0.5987 | 1.0036 | 0.4656 | 1.1312 | 6.6527 |
| | D | 0.2755 | 0.1057 | 0.2527 | 0.0546 | 0.1890 | 0.1341 | 1.5486 |
| $t(8)$ | M | 0.8733 | 1.0093 | 0.6288 | 1.0101 | 0.4032 | 1.1121 | 11.8453 |
| | D | 0.2839 | 0.0814 | 0.2782 | 0.0484 | 0.1944 | 0.1344 | 4.7581 |
| $t(12)$ | M | 0.9261 | 0.9814 | 0.5688 | 1.0013 | 0.4914 | 1.0615 | 18.9004 |
| | D | 0.2227 | 0.0818 | 0.2085 | 0.0554 | 0.1567 | 0.1075 | 9.3567 |
| $\mathcal{S}l(2)$ | M | 0.8949 | 0.9906 | 0.5950 | 1.0307 | 0.4034 | 1.0782 | 2.1688 |
| | D | 0.2547 | 0.0961 | 0.2353 | 0.0690 | 0.1610 | 0.1241 | 0.3137 |
| $\mathcal{S}l(4)$ | M | 0.8518 | 1.0528 | 0.6454 | 1.0152 | 0.3784 | 1.1131 | 6.6906 |
| | D | 0.2757 | 0.0989 | 0.2549 | 0.0596 | 0.1823 | 0.1382 | 2.3979 |
| $\mathcal{S}l(7)$ | M | 0.9175 | 1.0052 | 0.5904 | 1.0032 | 0.3670 | 1.0229 | 9.1147 |
| | D | 0.2138 | 0.0777 | 0.1943 | 0.0535 | 0.1395 | 0.1036 | 2.5640 |
| $\mathcal{S}l(11)$ | M | 0.8787 | 1.0261 | 0.6222 | 1.0002 | 0.3475 | 0.9897 | 10.0281 |
| | D | 0.2695 | 0.0901 | 0.2583 | 0.0542 | 0.1455 | 0.1039 | 2.1096 |
| $\mathcal{S}H(0.8)$ | M | 0.9075 | 0.9602 | 0.6000 | 0.9856 | 1.3221 | 3.8715 | 1.7738 |
| | D | 0.2858 | 0.1509 | 0.3129 | 0.0867 | 0.9182 | 2.3020 | 0.6884 |
| $\mathcal{S}H(1.0)$ | M | 0.8642 | 0.9850 | 0.6238 | 1.0287 | 1.0874 | 2.9591 | 1.8222 |
| | D | 0.3687 | 0.1232 | 0.3530 | 0.0757 | 0.9026 | 1.8072 | 0.8076 |
| $\mathcal{S}H(1.2)$ | M | 0.8868 | 1.0149 | 0.6154 | 0.9799 | 1.1252 | 2.6549 | 2.2248 |
| | D | 0.2742 | 0.0977 | 0.2733 | 0.0625 | 1.0149 | 1.6078 | 0.9998 |
| $\mathcal{S}H(1.4)$ | M | 0.9041 | 0.9915 | 0.6040 | 1.0056 | 0.9530 | 2.4467 | 2.4691 |
| | D | 0.2868 | 0.0842 | 0.2681 | 0.0499 | 0.6733 | 1.3370 | 0.9488 |
| $\mathcal{CN}(0.4, 0.2)$ | M | 0.9122 | 0.9330 | 0.5814 | 1.0375 | 0.5308 | 1.1591 | 0.4496; 0.0343 |
| | D | 0.2318 | 0.1081 | 0.2090 | 0.0692 | 0.1972 | 0.2115 | 0.0878; 0.2548 |
| $\mathcal{CN}(0.5, 0.2)$ | M | 0.8987 | 0.9501 | 0.6070 | 1.0484 | 0.5349 | 1.2796 | 0.5236; 0.0439 |
| | D | 0.3284 | 0.1279 | 0.3129 | 0.0755 | 0.2360 | 0.3183 | 0.0982; 0.2623 |
| $\mathcal{CN}(0.55, 0.2)$ | M | 0.9462 | 0.9542 | 0.5723 | 1.0586 | 0.6319 | 1.3270 | 0.5575; 0.0599 |
| | D | 0.1968 | 0.1178 | 0.1677 | 0.0750 | 0.2669 | 0.3700 | 0.1087; 0.2711 |
| $\mathcal{CN}(0.6, 0.2)$ | M | 0.9307 | 0.9465 | 0.5829 | 1.0387 | 0.6894 | 1.3375 | 0.6019; 0.0558 |
| | D | 0.2666 | 0.1244 | 0.2350 | 0.0727 | 0.2886 | 0.3882 | 0.1113; 0.2697 |
| Laplace | M | 0.9620 | 0.9797 | 0.5435 | 1.0397 | 0.5746 | 1.0466 | |
| | D | 0.1835 | 0.1597 | 0.1498 | 0.1333 | 0.1959 | 0.1184 | |

independentemente da distribuição assumida para o termo de erro $(\epsilon_i, \xi_i, \mathrm{m}_i)^T$. As estimativas dos parâmetros e a estimativa da função não paramétrica estão muito próximos dos valores reais, mesmo quando o parâmetro extra é considerado desconhecido.

## 3.4 Aplicações

Nesta seção, apresentamos duas aplicações a dados reais dos modelos propostos neste capítulo. A primeira aplicação corresponde a um conjunto de dados do nível de pólen. Na segunda, usamos um conjunto de dados disponível no pacote **MASS** do R.

### 3.4.1 Nível de pólen

A metodologia proposta é usada para analizar os dados RAGWEED POLLEN (veja Stark *et al.* , 1997). A variável de interesse é o nível de pólen na erva da ambrosia. Os dados foram coletados em 1993 na temporada da ambrósia em Kalamazoo, Michigan. Uma vez que a prevenção desempenha um papel importante no tratamento de alergias relacionadas com o pólen, um dos principais objectivos em aerobiologia é o desenvolvimento de modelos de previsão para níveis diários de pólen. Os dados são 87 observações diarias das seguintes variáveis

- `ragweed`: nível de ambrósia para esse dia (grãos/$\mathrm{m}^3$);
- `temperature`: temperatura do dia seguinte ($^oF$);
- `rain`: indicador de chuva significativa para o dia seguinte (1 = pelo menos 3 horas de chuva constante, ou breve, mas intensa; 0 = não teve chuva);
- `Wind`: velocidade do vento para o dia seguinte (nós);
- `day.in.seas`: número do dia na atual temporada de pólen.

Para a variável resposta `ragweed` foi considerada a transformação $y = \sqrt{\texttt{ragweed}}$ como foi sugerido por Ruppert *et al.* (2003). O gráfico de dispersão (omitido aqui) da resposta $y$ contra `day.in.seas` sugere uma relação não linear, portanto, o efeito desta variável será aproximado usando os $B$-splines. Então, o modelo considerado para a análise dos dados é o seguinte:

$$\begin{cases} y_i = \beta_1 + \beta_2 \texttt{rain}_i + \beta_3 \texttt{temperature}_i + \rho_1 \texttt{wind}_i + f(\texttt{day.in.seas}_i) + \epsilon_i \\ \texttt{Wind}_i = \texttt{wind}_i + \xi_i \qquad i = 1, \ldots, 87, \end{cases}$$

onde supomos que a variável `wind` tem erro de medição, pois este tipo de variáveis são propensas a erros introduzidos pelos instrumentos de medição e $(\epsilon_1, \xi_1, \texttt{wind}_1)^T, \ldots, (\epsilon_n, \xi_n, \texttt{wind}_n)^T$ são vetores aleatórios independentes e identicamente distribuidos, cuja distribuição é $\mathcal{SMN}_3$. No modelo consideramos que $\omega = 1$. Usamos as seguintes prioris independentes para os parâmetros do modelo

$$\beta_l \sim \mathcal{N}(0, 10^3), \quad l = 1, 2, 3, \qquad \rho_1 \sim \mathcal{N}(0, 10^3), \qquad \boldsymbol{\alpha} \sim \mathcal{N}_K(\mathbf{0}, \tau_\alpha^2 \mathbf{I}_K), \quad \mu_{\texttt{wind}} \sim \mathcal{N}(0, 10^3)$$
$$\tau_\alpha^2 \sim \mathcal{GI}(0.001, 0.001), \quad \sigma_y^2 \sim \mathcal{GI}(0.0005, 0.0005).$$

Para o termo aleatório $(\epsilon_i, \xi_i, \texttt{wind}_i)^T$ foram consideradas as distribuições normal, $t$-Student, slash, hiperbólica simétrica, Laplace e normal contaminada. Na aproximação do efeito não linear é considerado um $B$-spline cúbico com 3 nós internos. Então, usamos o algoritmo MCMC proposto para obter as estimativas Bayesianas dos parâmetros. No algoritmo MCMC foi considerado um *burn-in* de tamanho 10000 e, em seguida, foi gerada uma amostra de tamanho 200000 com saltos de tamanho 10. O anterior com o objetivo de reduzir a autocorrelação entre as cadeias para assim obter uma amostra de Monte Carlo aproximadamente independente de tamanho $J = 20000$. Assumimos que o parâmetro extra é conhecido uma vez que o tamanho

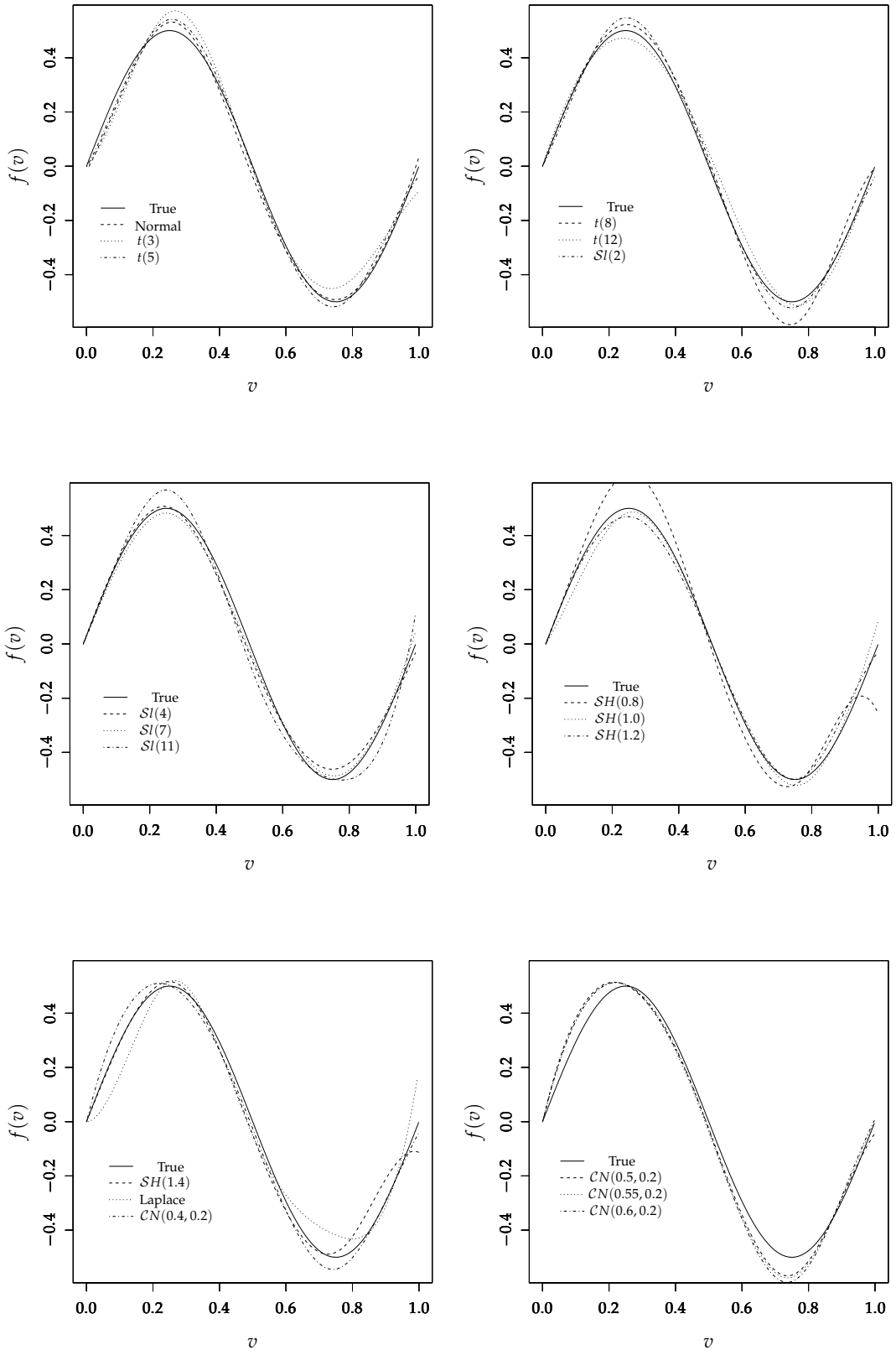


**Figura 3.1:** *Verdadeiro valor da função $f(v)$ contra suas estimativas (linhas pontilhadas), em que o parâmetro $\boldsymbol{\eta}$ é considerado sendo conhecido.*

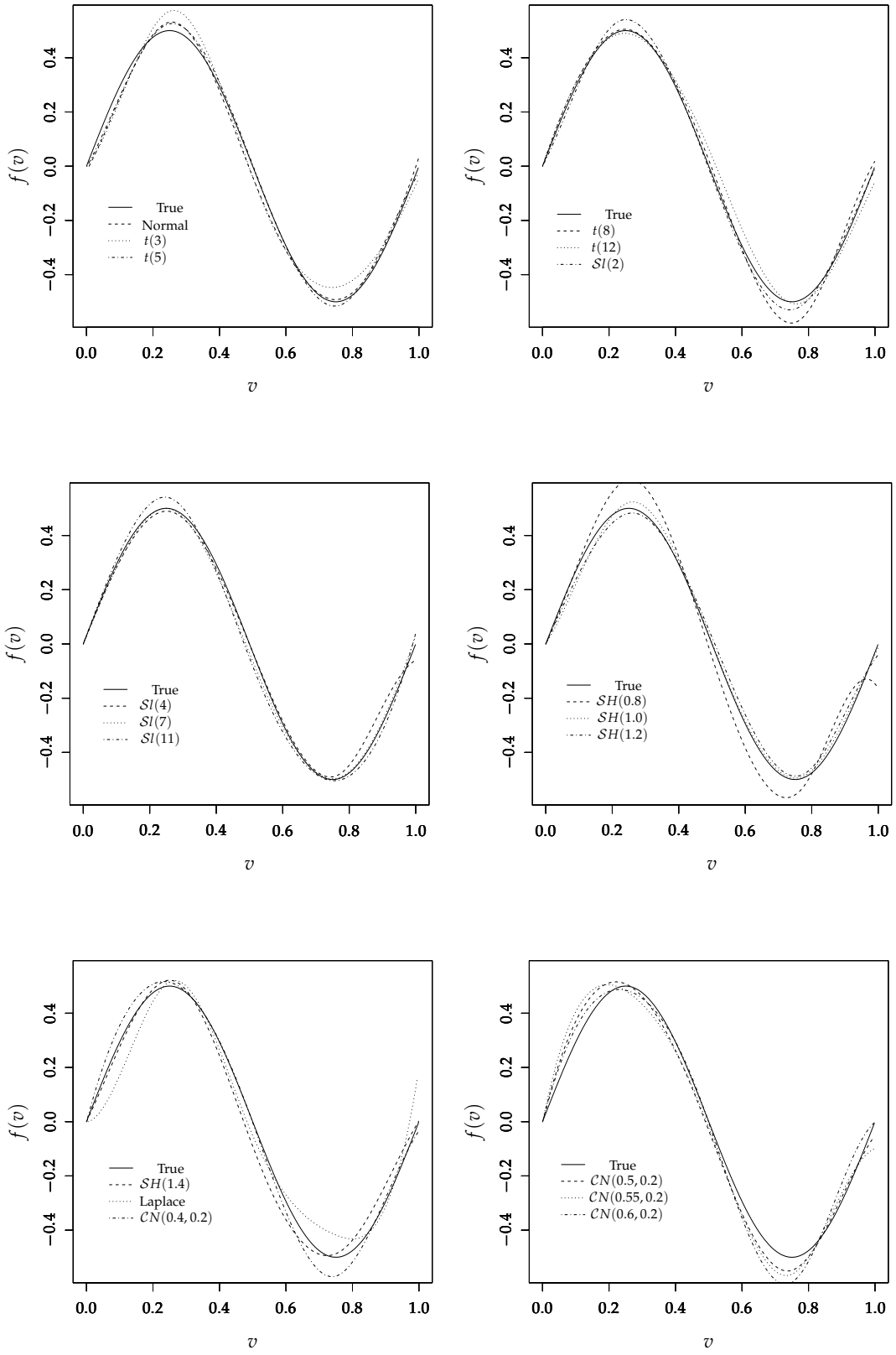


**Figura 3.2:** *Verdadeiro valor da função $f(v)$ contra suas estimativas (linhas pontilhadas), em que o parâmetro $\boldsymbol{\eta}$ é considerado sendo desconhecido.*

**Tabela 3.3:** *Critérios de seleção de modelos para os dados do nível de pólen.*

| Distribuição | DIC | LMPL |
|---|---|---|
| Normal | 826.314 | -412.377 |
| $t$-Student($\eta = 5$) | 813.832 | -407.478 |
| Slash($\eta = 5$) | 814.958 | -407.906 |
| Hiperbólica simétrica ($\eta = 1$) | **812.451** | **-406.636** |
| N. Contaminada($\boldsymbol{\eta} = (0.9, 0.5)$) | 814.968 | -408.141 |
| Laplace | 813.822 | -407.174 |

**Tabela 3.4:** *Média a posteriori, desvio padrão e intervalo de credibilidade para os parâmetros do modelo hiperbólico simétrico ($\mathcal{SH}(1)$), nos dados do nível de pólen.*

| Parâmetro | Média | Desvio Padrão | Intervalo de credibilidade 95% |
|---|---|---|---|
| $\beta_1$ | -5.6392 | 2.2424 | (-10.1635, -1.3199) |
| $\beta_2$ | 1.8056 | 0.6451 | (0.6074, 3.1545) |
| $\beta_3$ | 0.0793 | 0.0315 | (0.0177, 0.1407) |
| $\rho_1$ | 0.2870 | 0.1109 | (0.0817, 0.5200) |
| $\mu_{\texttt{wind}}$ | 8.8444 | 0.3282 | (8.2047, 9.4821) |
| $\sigma^2_{\texttt{wind}}$ | 2.9353 | 0.8424 | (1.5127, 4.8287) |
| $\sigma^2_\epsilon$ | 1.2540 | 0.2703 | (0.8190, 1.8732) |

da amostra é pequeno. Portanto, usamos diferentes valores do parâmetro extra $\eta$ e realizamos a seleção de um modelo para cada uma das distribuições consideradas usando os critérios de comparação DIC e LMPL. Na Tabela 3.3 são apresentados os valores dos critérios de comparação para cada distribuição considerada para descrever o vetor $(\epsilon_i, \xi_i, \texttt{wind}_i)^T$. Desta forma, o modelo selecionado, ou seja, aquele com o menor valor do DIC, foi o que segue distribuição hiperbólica simétrica ($\mathcal{HS}(1)$). Para ajustar este modelo usamos a função `fmem()` na seguinte forma

```
model <- fmem(sqrt(ragweed) ~ wind.speed |  rain + temperature +
         bsp(day.in.seas,3), data=ragweedn, family="Hyperbolic",
         eta=1, burn.in=10000, post.sam.s=20000, thin=10)
summary(model)
```

As Figuras 3.3 e 3.4 apresentam o comportamento das cadeias e as densidades marginais a posteriori aproximadas para os parâmetros $\beta_1$, $\beta_2$, $\beta_3$, $\rho_1$, $\mu_{\texttt{wind}}$ e $\sigma^2_\epsilon$ sob o modelo $\mathcal{SH}(1)$. Estas Figuras sugerem que a convergência das cadeias foi atingida e que as densidades marginais posteriores são aproximadamente simétricas para os parâmetros $\beta_1$, $\beta_2$, $\beta_3$, $\rho_1$ e $\mu_{\texttt{wind}}$. A Figura 3.5 apresenta o comportamento da função não paramétrica $f(\texttt{day.in.seas})$ estimada usando os $B$-splines para o modelo $\mathcal{SH}(1)$. Pode-se observar um comportamento não linear nas observações e na função estimada. A Tabela 3.4 apresenta a média a posteriori, o desvio padrão e o intervalo de credibilidade para os parâmetros do modelo $\mathcal{SH}(1)$. Na Figura 3.6 é apresentado o gráfico dos residuos $r_i$ para o modelo $\mathcal{SH}(1)$, o qual indica que este modelo descreve adequadamente o conjunto de dados, em que $r_i = \Phi^{-1}\{F(y_i; \bar{\boldsymbol{\theta}})\}$, $i = 1, \ldots, n$, com $\Phi(\cdot)$ como a distribuição acumulada da normal padrão e $F(\cdot; \bar{\boldsymbol{\theta}})$ como a função de distribuição acumulada da distribuição marginal de $y_i$ avaliada na média a posteriori de $\boldsymbol{\theta}$.

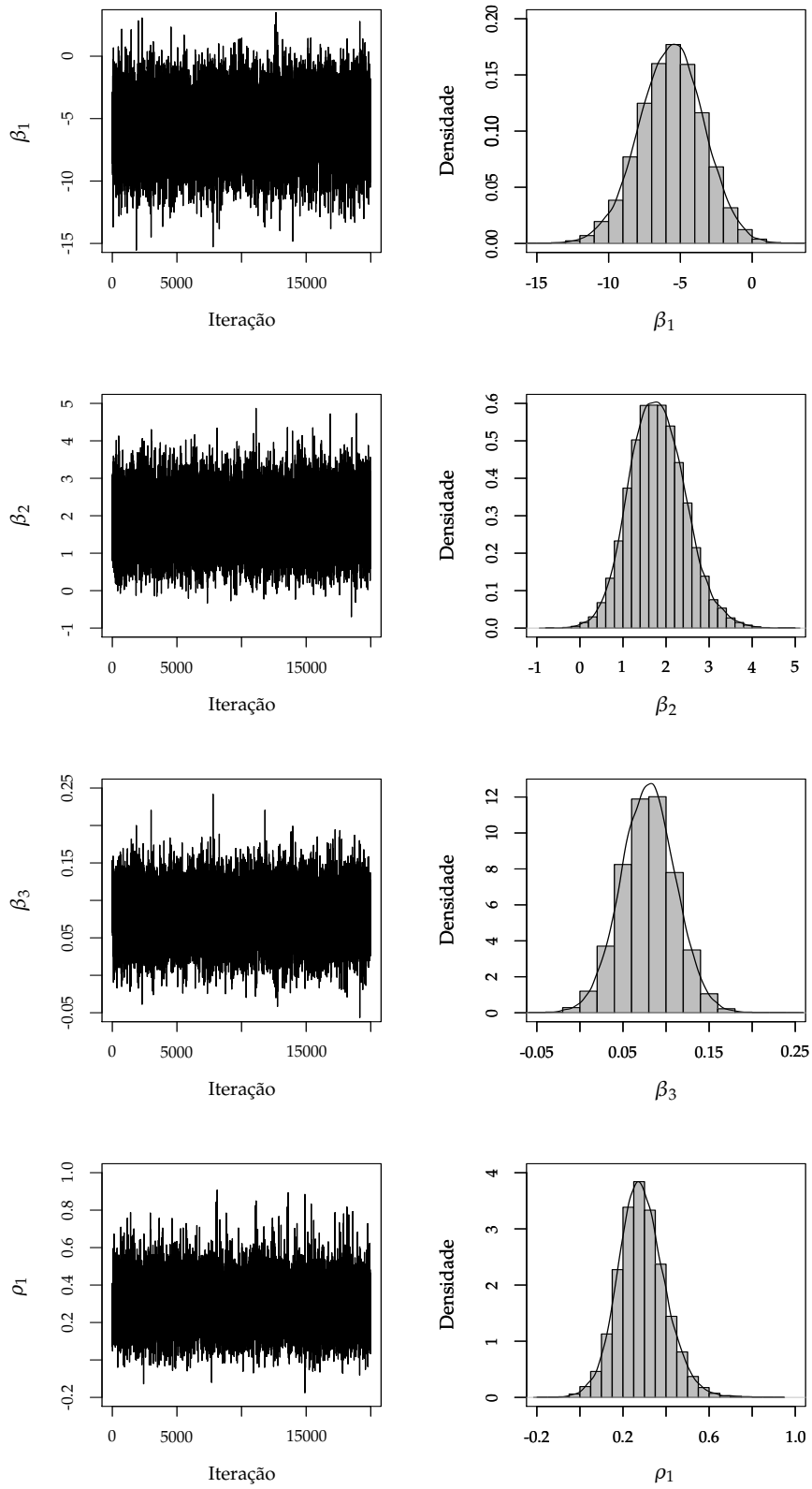


**Figura 3.3:** *Comportamento das cadeias e densidades marginais a posteriori dos parâmetros $\beta_1$, $\beta_2$, $\beta_3$ e $\rho_1$ no modelo $\mathcal{SH}(1)$, nos dados do nível de pólen.*

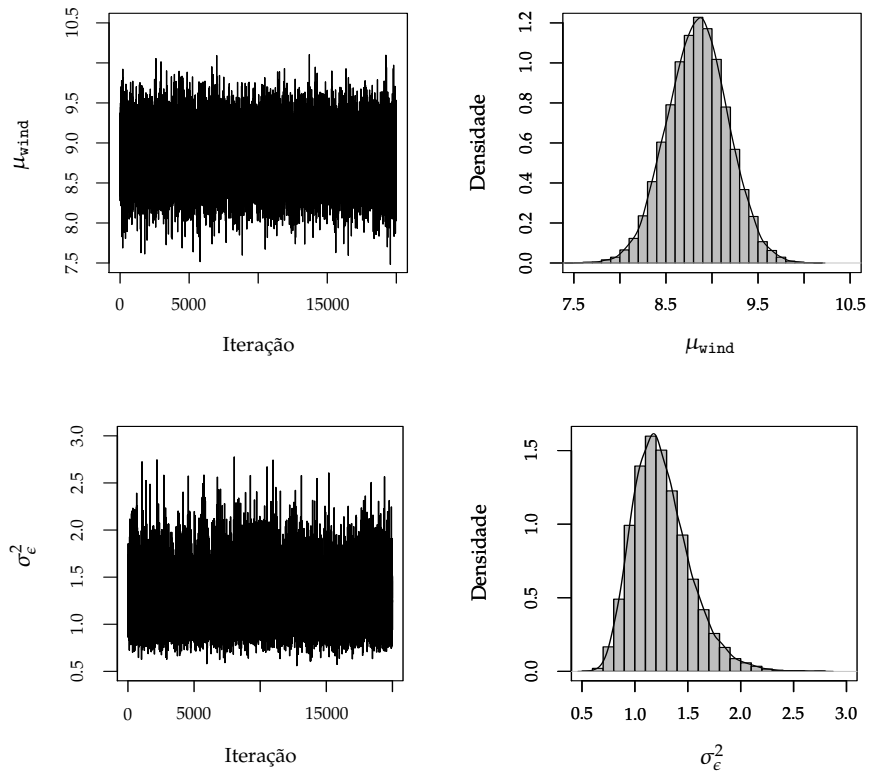


**Figura 3.4:** *Comportamento das cadeias e densidades marginais a posteriori dos parâmetros* $\mu_{\texttt{wind}}$ *e* $\sigma^2_\epsilon$ *do modelo* $\mathcal{SH}(1)$*, nos dados do nível de pólen.*

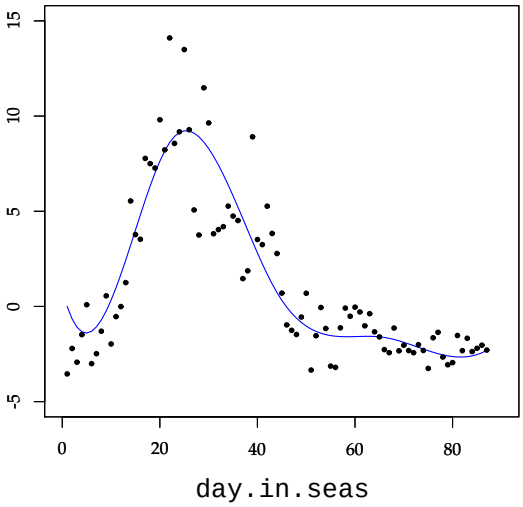


**Figura 3.5:** *Gráfico da função* $f(\texttt{day.in.seas})$ *ajustada usando o modelo* $\mathcal{SH}(1)$*, nos dados do nível de pólen.*

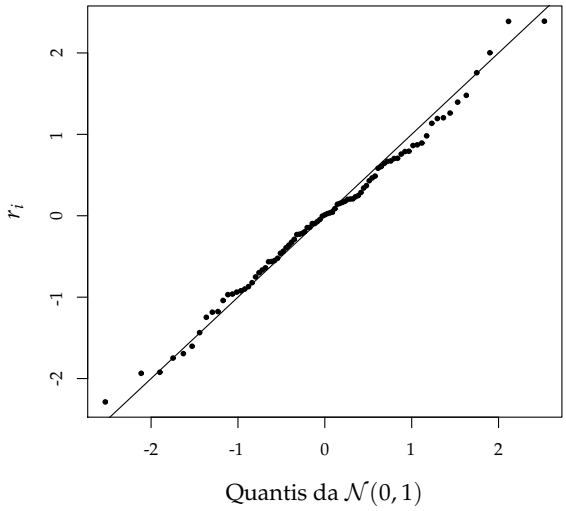


**Figura 3.6:** *Grafico dos resíduos no modelo $\mathcal{SH}(1)$ ajustado nos dados do nível de pólen.*

### 3.4.2 Boston

Este conjunto de dados consiste de 506 indivíduos e relaciona o impacto da poluição do ar e outras variáveis explicativas sobre o preço das casas ocupadas pelos proprietários em Boston. Este conjunto de dados foi apresentado por Harrison (1978) e analisado por Belsley *et al.* (2004). A variável resposta é o logaritmo do valor mediano das casas ocupadas pelos proprietários (`log(medv)`), e está relacionada com 14 variáveis explicativas, seis delas são definidas a partir do censo e as restantes são definidas para clusters. Para esta ilustração são consideradas as seguintes variáveis:

- `Nox`: concentração de óxido de nitrogênio (10 partes por milhão);
- `crim`: taxa de criminalidade per capita por cidade;
- `rm`: número médio de quartos;
- `lstat`: população sem educação média (%);
- `dis`: média ponderada das distâncias de cinco centros de emprego em Boston.

Os gráficos de dispersão da variável `log(medv)` contra as variáveis explicativas `lstat` e `dis` sugerem uma relação não-linear entre elas, de modo que, essas relações são descritas usando funções não paramétricas. Assumimos que a variável `nox` tem erro de medição, pois este tipo de variáveis são propensas a erros introduzidos pelos instrumentos de medição. Portanto, o seguinte modelo é proposto para analisar o conjunto de dados:

$$\begin{cases} \texttt{log(medv)}_i = \beta_0 + \beta_1 \texttt{crim}_i + \beta_2 \texttt{rm}_i + \rho_1 \texttt{nox}_i + f_1(\texttt{lstat}_i) + f_2(\texttt{dis}_i) + \epsilon_i \\ \texttt{Nox}_i = \texttt{nox}_i + \xi_i, \qquad i = 1, \ldots, 506, \end{cases} \tag{3.5}$$

em que $(\epsilon_1, \xi_1, \texttt{nox}_1)^T, \ldots, (\epsilon_n, \xi_n, \texttt{nox}_n)^T$ são vetores aleatórios independentes e identicamente distribuídos com $\mathcal{SMN}_3$. Para evitar problemas de identificabilidade o valor de $\omega$ é fixado em

**Tabela 3.5:** *Critérios de seleção de modelos para os dados* BOSTON.

| Distribution | DIC | LMPL |
|---|---|---|
| Normal | -987.34 | 489.03 |
| $t$-Student | -1011.78 | 501.83 |
| $\mathcal{S}l$ | -1006.99 | 498.27 |
| $\mathcal{SH}$ | -1010.62 | 501.39 |
| $\mathcal{CN}$ | **-1017.56** | **504.63** |
| Laplace | -972.60 | 480.89 |

**Tabela 3.6:** *Comportamento das estimativas dos parâmetros para os diferentes valores de $\omega$ sob o modelo $\mathcal{CN}$.*

| Model | Medida | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\rho_1$ | $\sigma_y^2$ | DIC | LMPL |
|---|---|---|---|---|---|---|---|---|
| com erro de | Média | 3.6988 | -0.0127 | 0.1773 | -1.1192 | 0.0156 | -1017.565 | 504.631 |
| medição* | DP | 0.2213 | 0.0016 | 0.0185 | 0.2441 | 0.0015 | | |
| sem erro de | Média | 3.4241 | -0.0127 | 0.1774 | -0.6132 | 0.0180 | -1017.301 | 504.528 |
| medição** | DP | 0.2046 | 0.0016 | 0.0184 | 0.1492 | 0.0014 | | |

$^*\omega = 4$, $^{**}\omega = 10000$.

4. Além disso, as seguintes distribuições a priori são especificadas

$$\beta_0 \sim \mathcal{N}(0, 10^3), \quad \beta_j \sim \mathcal{N}(0, 10^3), \quad \rho_1 \sim \mathcal{N}(0, 10^3), \quad \boldsymbol{\alpha}_j \sim \mathcal{N}_{K_j}(\mathbf{0}, \tau_{\alpha_j}^2 \mathbf{I}_{K_j}),$$
$$\mu_{\texttt{nox}} \sim \mathcal{N}(0, 10^3), \quad \tau_{\alpha_j}^2 \sim \mathcal{GI}(0.001, 0.001) \quad \text{e} \quad \sigma_\epsilon^2 \sim \mathcal{GI}(0.0005, 0.0005).$$

60000 iterações do algoritmo MCMC (descrito na seção 3.2.2) são simuladas, que incluem um período de aquecimento de 10000 e saltos de 10 iterações, de modo que uma amostra independente de tamanho $R = 5000$ é obtida. Para o termo $(\epsilon_i, \xi_i, \texttt{nox}_i)^T$, as distribuições normal, $t$-Student, slash, hiperbólica simétrica, Laplace e normal contaminada foram consideradas. Em todos os casos, o parâmetro extra é assumido desconhecido. A comparação dos modelos ajustados foi realizada usando os critérios DIC e LMPL.

A Tabela 3.5 apresenta os valores desses critérios para cada modelo considerado. O modelo em que o termo do erro segue distribuição normal contaminada apresenta o menor DIC e o maior valor de LMPL. Assim, podemos considerá-lo como o melhor modelo para descrever o conjunto de dados.

As Figuras 3.7 e 3.8 revelam o comportamento das cadeias, bem como as densidades marginais a posteriori aproximadas para os parâmetros $\beta_1$, $\beta_2$, $\rho_1$, $\mu_{\texttt{nox}}$, $\sigma_{\texttt{nox}}^2$ e $\sigma_\epsilon^2$ no modelo normal contaminado. Estes gráficos sugerem que a convergência foi atingida e que as densidades marginais a posteriori são aproximadamente simétricas para os parâmetros $\beta_1$, $\beta_2$, $\rho_1$ e $\mu_{\texttt{nox}}$. A Figura 3.9 mostra o comportamento das funções não paramétricas $f_1(\texttt{lstat})$ e $f_2(\texttt{dis})$ estimadas usando os $B$-splines sob o modelo normal contaminado. Note que, um comportamento não-linear é revelado não só pelos dados, mas também pelas funções estimadas.

Para investigar como o erro de medição na covariável nox influencia os resultados na modelagem, comparamos dois métodos de estimação. O primeiro, baseia-se no modelo flexível com erros nas variáveis apresentado neste capítulo, isto é, levando em conta o erro de medição usando a seguinte equação $\texttt{Nox}_i = \texttt{nox}_i + \xi_i$. O outro é o método ingênuo (*naive*), isto é, usando a variável $\texttt{Nox}_i$ (observada) diretamente para substituir o verdadeiro valor (não observado) $\texttt{nox}_i$. A comparação dos dois métodos é apresentada na Tabela 3.6. Podemos observar que existem diferenças importantes nas estimativas do parâmetro $\rho_1$. O método ingênuo pode subestimar o efeito das variáveis medidas com erro. Além disso, o modelo que melhor se ajusta aos dados é o modelo MFEV, pois este apresenta o menor valor do DIC.

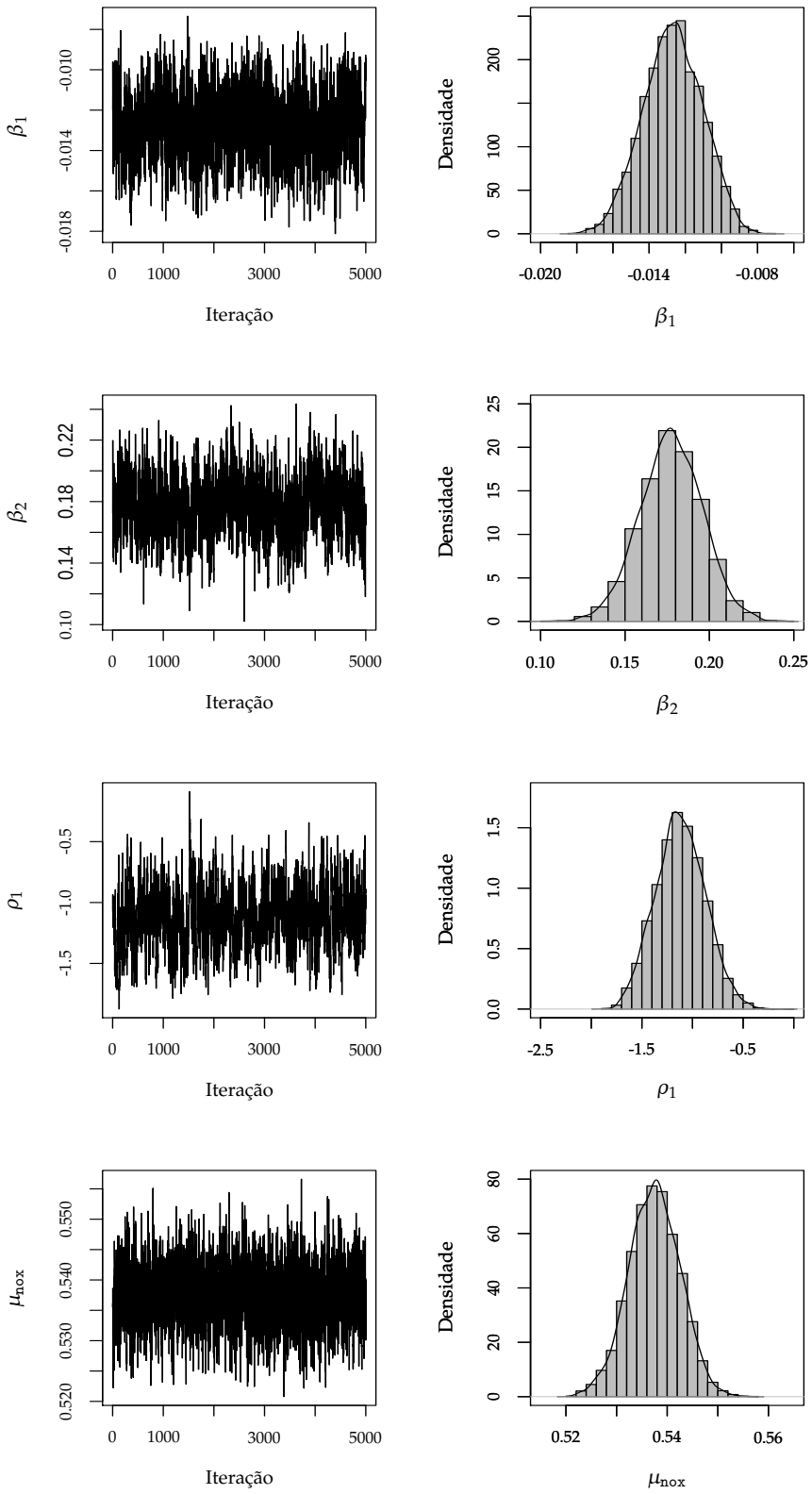


**Figura 3.7:** *Comportamento das cadeias e densidades marginais a posteriori dos parâmetros $\beta_1$, $\beta_2$, $\rho_1$ e $\mu_{\mathrm{nox}}$ considerando o modelo normal contaminado nos dados* BOSTON.

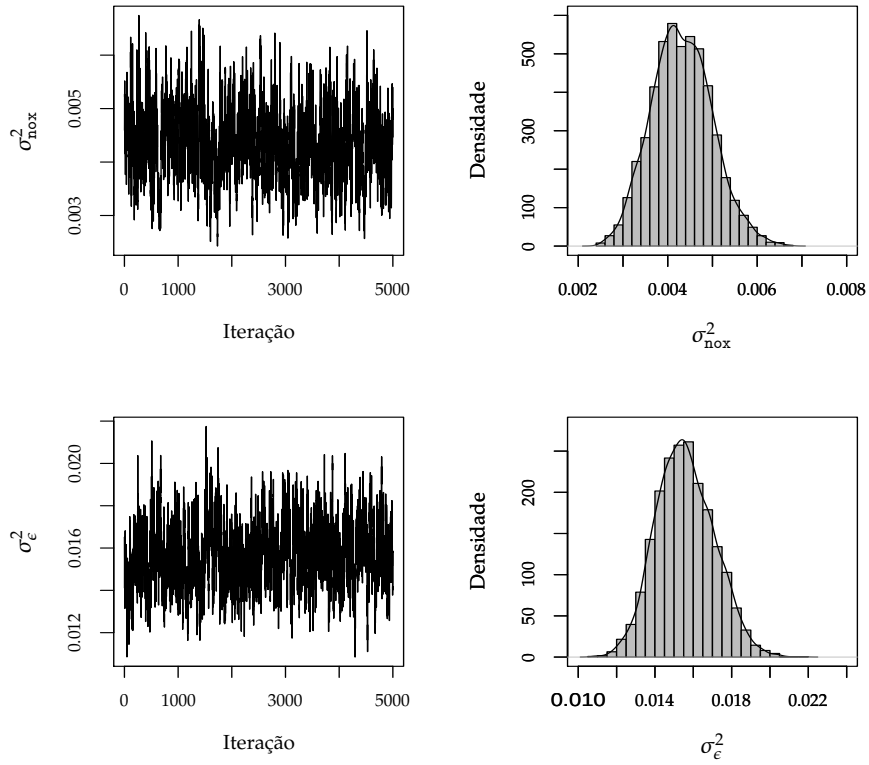


**Figura 3.8:** *Comportamento das cadeias e densidades marginais a posteriori dos parâmetros* $\sigma^2_{\texttt{nox}}$ *e* $\sigma^2_\epsilon$*, considerando o modelo normal contaminado nos dados* BOSTON.

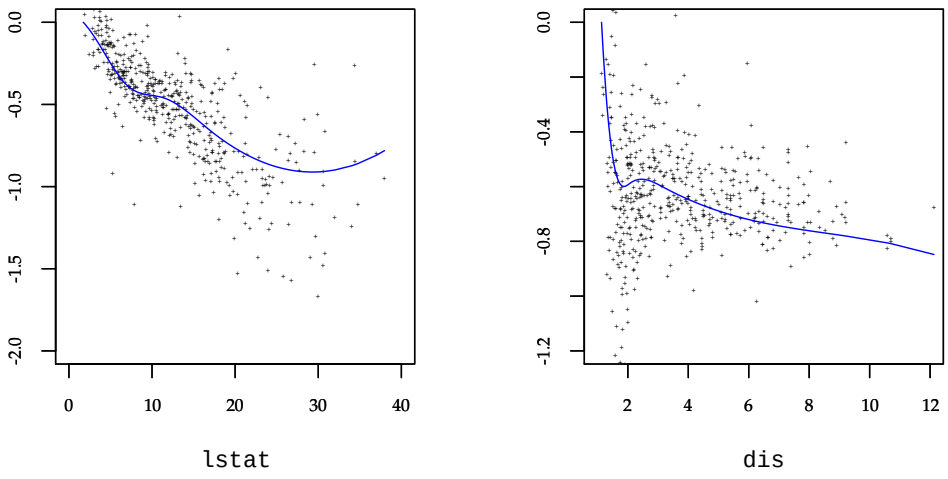


**Figura 3.9:** *Gráficos de* $\hat{f}_1(\texttt{lstat})$ $(a)$ *e* $\hat{f}_2(\texttt{dis})$ $(b)$ *sob o modelo normal contaminado para os dados* BOSTON.

## 3.5 Conclusões

Neste capítulo estudamos, sob a abordagem Bayesiana, um modelo de regressão que admite variáveis explicativas com e sem erro de medição, bem como a presenca de efeitos não lineares aproximados usando $B$-splines. Neste modelo, a componente aleatória considera distribuições da família $\mathcal{SMN}$. Vale salientar que, o modelo apresentado neste capítulo estende as propostas encontradas na literatura estatística para modelos com erro de medição.

Capítulo 4

# Modelo flexível com erros nas variáveis heterocedástico

No modelo com erros nas variáveis descrito em (3.1), assume-se que as variâncias dos erros aleatórios $\epsilon_i$ e $\boldsymbol{\xi}_i$ são as mesmas para todos os indivíduos. Porém, em muitas situações, as variâncias dos erros podem mudar de uma observação para outra, devido às diferenças entre indivíduos ou por outros motivos. Como resposta a este problema, os modelos com erros nas variáveis heteroscedásticos têm sido desenvolvidos nos últimos anos para lidar com este tipo de dados, os quais estão presentes em diversas áreas. Estes modelos têm sido estudados por vários autores como, por exemplo, de Castro *et al.* (2008); Kelly (2007); Kulathinal *et al.* (2002) e Patriota *et al.* (2009).

Neste capítulo estudamos um modelo com erros nas variáveis que assume erros aleatórios heteroscedásticos, onde, da mesma forma que no modelo estudado no capítulo anterior, a componente sistemática admite variáveis explicativas com e sem erro de medição, bem como a presenca de efeitos não lineares aproximados através de $B$-splines. Então, assume-se que as variâncias $\sigma^2_{\epsilon_i} > 0$ e $\boldsymbol{\Sigma}_{\xi_i} > 0$ são conhecidas para todo $i$ $(i = 1, \ldots, n)$. Portanto, o modelo estudado neste capítulo recebe o nome de modelo flexível com erros nas variáveis heterocedástico. A componente aleatória deste modelo considera distribuições que pertencem a la classe de mistura na escala da normal multivariada. Desenvolvemos a inferência Bayesiana nesta classe de modelos, propondo um algoritmo MCMC que permite obter amostras da distribuição a posteriori dos parâmetros do modelo. Na seção 4.3 um estudo de simulação é apresentado com o objetivo de avaliar o desempenho do algoritmo MCMC. Além disso, na seção 4.4, aplicamos a metodologia estudada em dois conjuntos de dados reais.

## 4.1 Formulação do modelo

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes que representam uma medição de interesse realizada em $n$ objetos ou indivíduos. Assume-se que $y_i$ pode ser escrita como

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{m}_i^T \boldsymbol{\rho} + \sum_{j=1}^{s} f_j(v_{ij}) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{4.1}$$

em que $\epsilon_1, \ldots, \epsilon_n$ são erros aleatórios independentes seguindo distribuição simétrica de média zero e variância $\sigma^2_{\epsilon_i} > 0$ (conhecida); $(\mathbf{x}_i^T, \mathbf{m}_i^T, v_{i1}, \ldots, v_{is})^T$ é um vetor de variáveis explicativas associadas ao indivíduo $i$; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ e $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_q)^T$ são vetores de parâmetros desconhecidos a serem estimados; $f_j(\cdot)$ $(j = 1, \ldots, s)$ é uma função suave e contínua, mas desconhecida, que é aproximada usando $B$-splines. Porém, o vetor de variáveis explicativas $\mathbf{m}_i$

não é observável diretamente para nenhum indivíduo, e no seu lugar é observada uma aproximação dele que está contaminada por um erro aditivo seguindo uma distribuição que pertence à classe de distribuições de mistura normal na escala normal multivariada ($\mathcal{SMN}_q$), com média zero e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}_{\xi_i} > 0$ (conhecida). Então, assume-se que para cada indivíduo observa-se a resposta $y_i$ e o vetor de variáveis explicativas $(\mathbf{x}_i^T, \mathbf{M}_i^T, v_{i1}, \ldots, v_{is})^T$, em que

$$\mathbf{M}_i = \mathbf{m}_i + \boldsymbol{\xi}_i, \tag{4.2}$$

e $(\epsilon_i, \boldsymbol{\xi}_i^T, \mathbf{m}_i^T)$ são vetores aleatórios independentes tais que

$$\begin{pmatrix} \epsilon_i \\ \boldsymbol{\xi}_i \\ \mathbf{m}_i \end{pmatrix} \sim \mathcal{SMN}_{2q+1} \left[ \begin{pmatrix} 0 \\ \mathbf{0} \\ \boldsymbol{\mu}_m \end{pmatrix} ; \begin{pmatrix} \sigma^2_{\epsilon_i}/\zeta & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{\xi_i}/\zeta & \mathbf{0} \\ 0 & \mathbf{0} & \boldsymbol{\Sigma}_m \end{pmatrix} \right], \tag{4.3}$$

em que $\boldsymbol{\mu}_m$ e $\boldsymbol{\Sigma}_m$ são parâmetros desconhecidos a serem estimados; $\zeta = \zeta(\eta)$ é um valor de padronização selecionado para satisfazer que $Var(\epsilon_i) = \sigma^2_{\epsilon_i}$ e $Var(\boldsymbol{\xi}_i) = \boldsymbol{\Sigma}_{\xi_i}$ (para detalhes veja a seção 1.1). Então, o modelo definido por (4.1), (4.2) e (4.3) recebe neste trabalho o nome de modelo flexível com erros nas variáveis heterocedástico (MFEVH).

## 4.2 Inferência bayesiana

Nesta seção especificamos as distribuições a priori para os parâmetros do modelo e descrevemos a implementação do algoritmo MCMC para gerar amostras da distribuição a posteriori.

### 4.2.1 Distribuições a priori

Supomos que os parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, $\boldsymbol{\alpha}_j$, $\boldsymbol{\mu}_m$ e $\boldsymbol{\Sigma}_m^{-1}$ são independentes a priori, com as seguintes distribuições

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\beta}_0, \mathbf{S}_\beta), \quad \boldsymbol{\rho} \sim \mathcal{N}_q(\boldsymbol{\rho}_0, \mathbf{S}_\rho), \quad \boldsymbol{\alpha}_j \sim \mathcal{N}_{K_j}(\boldsymbol{\alpha}_0, \tau^2_{\alpha_j}\mathbf{I}_{K_j}), \quad \boldsymbol{\mu}_m \sim \mathcal{N}_q(\boldsymbol{\mu}_{\mu_{m0}}, \boldsymbol{\Sigma}_{\mu_{m0}}),$$
$$\boldsymbol{\Sigma}_m^{-1} \sim \mathcal{W}ishart(q, \boldsymbol{\Omega}_m) \quad \text{e} \quad \tau^2_{\alpha_j} \sim \mathcal{G}I(a_{\tau_{\alpha_j}}, b_{\tau_{\alpha_j}}),$$

em que os hiperparâmetros $\boldsymbol{\beta}_0$, $\boldsymbol{\rho}_0$, $\boldsymbol{\alpha}_0$, $\boldsymbol{\mu}_{\mu_{m0}}$, $\mathbf{S}_\beta$, $\mathbf{S}_\rho$, $\boldsymbol{\Sigma}_{\mu_{m0}}$, $\boldsymbol{\Omega}_m$, $a_{\tau_{\alpha_j}} > 0$ e $b_{\tau_{\alpha_j}} > 0$ são assumidos conhecidos.

### 4.2.2 Algoritmo MCMC

O algoritmo MCMC proposto para obter as distribuições a posteriori dos parâmetros do modelo é construído de forma similar ao algoritmo MCMC apresentado na seção 3.2.2. Então, neste algoritmo consideramos as variáveis $u_i$ e $\mathbf{m}_i$ $(i = 1, \ldots, n)$ como variáveis latentes e usamos o algoritmo de dados aumentados. Assim, a função de verossimilhança aumentada para o vetor de parâmetros $(\boldsymbol{\beta}^T, \boldsymbol{\rho}^T, \boldsymbol{\alpha}^T, \ldots, \boldsymbol{\alpha}_s^T, \boldsymbol{\mu}_m^T, \boldsymbol{\Sigma}_m)$ é dada por

$$L(\boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_s^T, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \sigma^2_\epsilon | \mathbf{y}, \mathbf{X}, \mathbf{M}, \mathbf{v}, \mathbf{u}, \mathbf{m}) \propto \prod_{i=1}^n (\sigma^2_{\epsilon_i})^{-\frac{1}{2}} |\boldsymbol{\Sigma}_{\xi_i}|^{-\frac{1}{2}} \kappa(u_i)^{-q-\frac{1}{2}} |\boldsymbol{\Sigma}_m|^{-\frac{1}{2}}$$
$$\exp\left[ -\frac{(\mathbf{M}_i - \mathbf{m}_i)^T \boldsymbol{\Sigma}_{\xi_i}^{-1} (\mathbf{M}_i - \mathbf{m}_i)}{2\kappa(u_i)} - \frac{\left( y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{m}_i^T\boldsymbol{\rho} - \sum_{j=1}^s \mathbf{b}_{ij}^T \boldsymbol{\alpha}_j \right)^2}{2\sigma^2_{\epsilon_i}\kappa(u_i)} \right.$$
$$\left. - \frac{(\mathbf{m}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\mathbf{m}_i - \boldsymbol{\mu}_m)}{2\kappa(u_i)} \right],$$

em que $\boldsymbol{\Sigma}_{\xi_i}$ e $\sigma^2_{\epsilon_i}$ são quantidades conhecidas para todo $i$ ($i = 1, \ldots, n$). Usando a verossimilhança e as distribuições a priori apresentadas acima podemos obter as distribuições condicionais completas para cada um dos parâmetros do modelo MFEVH. O algoritmo MCMC segue os seguintes passos:

*Passo 1:* Inicializar os valores dos parâmetros $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\rho}^{(0)}, \boldsymbol{\mu}_m^{(0)}, \boldsymbol{\Sigma}_m^{(0)}, \boldsymbol{\alpha}_1^{(0)}, \ldots, \boldsymbol{\alpha}_s^{(0)})$;

*Passo 2:* Calcular a quantidade $S_i^{(l)}$ ($i = 1, \ldots, n$),

$$S_i = \frac{\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{m}_i^T\boldsymbol{\rho} - \sum_{j=1}^{s} \mathbf{b}_{ij}^T\boldsymbol{\alpha}_j\right)^2}{\sigma^2_{\epsilon_i}} + (\mathbf{M}_i - \mathbf{m}_i)^T \boldsymbol{\Sigma}_{\xi_i}^{-1} (\mathbf{M}_i - \mathbf{m}_i) + (\mathbf{m}_i - \boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1}(\mathbf{m}_i - \boldsymbol{\mu}_m),$$

em que $S_i^{(l)}$ representa $S_i$ avaliado em $\boldsymbol{\theta}^{(l)}$.

*Passo 3:* Gerar $u_i^{(l+1)} \sim p(u_i|S_i^{(l)})$, $i = 1, \ldots, n$, independentes, deacordo à distribuição da componente aleatória $(\epsilon, \boldsymbol{\xi}^T, \mathbf{m}^T)^T$:

($a$) Distribuição normal: $P(u_i = 1|S_i^{(l)}) = 1$.

($b$) Distribuição $t$-Student: $u_i|S_i^{(l)} \sim \mathcal{G}amma\left(\frac{\eta+1}{2} + q, \frac{S_i^{(l)} + \eta}{2}\right)$.

($c$) Distribuição slash: $u_i|S_i^{(l)} \sim Trun\mathcal{G}amma\left(\eta + q + \frac{1}{2}, \frac{S_i^{(l)}}{2}; (0,1)\right)$.

($c$) Distribuição normal contaminada:

$$p(u_i|S_i^{(l)}) = p_\eta \mathbb{I}_{(u_i=\eta_2)} + (1 - p_\eta)\mathbb{I}_{(u_i=1)},$$

em que

$$p_\eta \propto \eta_1\eta_2^{\frac{1}{2}+q} \exp\left\{-\frac{\eta_2\, S_i^{(l)}}{2}\right\} \quad \text{e} \quad (1 - p_\eta) \propto (1 - \eta_1) \exp\left\{-\frac{S_i^{(l)}}{2}\right\}.$$

($d$) Distribuição hiperbólica simétrica: $u_i|S_i^{(l)} \sim \mathcal{GIG}\left(-q + \frac{1}{2}, S_i^{(l)} + 1, \eta^2\right)$.

($e$) Distribuição Laplace: $u_i|S_i^{(l)} \sim \mathcal{GIG}\left(-q + \frac{1}{2}, S_i^{(l)}, \frac{1}{4}\right)$.

*Passo 4:* Calcular a matriz $\mathbf{L}_u^{(l+1)} = \text{diag}\left\{u_1^{(l+1)}, \ldots, u_n^{(l+1)}\right\}$.

*Passo 5:* Gerar $\tilde{\boldsymbol{\beta}}^{(l+1)} \sim \mathcal{N}_{p+q}\left(\boldsymbol{\mu}_{\tilde{\beta}}, \boldsymbol{\Sigma}_{\tilde{\beta}}\right)$, em que

$$\boldsymbol{\Sigma}_{\tilde{\beta}} = \left[\begin{pmatrix} \mathbf{S}_\beta^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\rho^{-1} \end{pmatrix} + \overline{\mathbf{X}}^{T(l)} \left[\boldsymbol{\Sigma}_\epsilon \mathbf{L}_u^{(l+1)}\right]^{-1} \overline{\mathbf{X}}^{(l)}\right]^{-1}, \quad \text{e}$$

$$\boldsymbol{\mu}_{\tilde{\beta}} = \boldsymbol{\Sigma}_{\tilde{\beta}} \left[\begin{pmatrix} \mathbf{S}_\beta^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\rho^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\rho}_0 \end{pmatrix} + \overline{\mathbf{X}}^{T(l)} \left[\boldsymbol{\Sigma}_\epsilon \mathbf{L}_u^{(l+1)}\right]^{-1} \left(\mathbf{y} - \sum_{j=1}^{s} \mathbf{B}_j \boldsymbol{\alpha}_j^{(l)}\right)\right],$$

com $\overline{\mathbf{X}}^{(l)} = \left[\mathbf{X}, \mathbf{m}^{(l)}\right]$, $\boldsymbol{\Sigma}_\epsilon = \text{diag}\{\sigma^2_{\epsilon_1}, \ldots, \sigma^2_{\epsilon_n}\}$, $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^T, \boldsymbol{\rho}^T)^T$ e $\mathbf{B}_j = (\mathbf{b}_{1j}, \ldots, \mathbf{b}_{nj})^T$.

*Passo 6:* Gerar a $i$-ésima linha de $\mathbf{m}^{(l+1)}$, denotada por $\mathbf{m}_i^{(l+1)}$, em que $\mathbf{m}_i^{(l+1)} \sim \mathcal{N}_q\left(\boldsymbol{\mu}_{m_i}, \boldsymbol{\Sigma}_{m_i}\right)$, com

$$\boldsymbol{\Sigma}_{m_i} = \left[\frac{\left(\boldsymbol{\Sigma}_m^{(l)}\right)^{-1}}{\kappa\left(u_i^{(l+1)}\right)} + \frac{\boldsymbol{\Sigma}_{\xi_i}^{-1}}{\kappa\left(u_i^{(l+1)}\right)} + \frac{\boldsymbol{\rho}^{(l+1)}\boldsymbol{\rho}^{T\,(l+1)}}{\sigma_{\epsilon_i}^2\kappa\left(u_i^{(l+1)}\right)}\right]^{-1} \quad \text{e}$$

$$\boldsymbol{\mu}_{m_i} = \boldsymbol{\Sigma}_{m_i}\left[\frac{\left(\boldsymbol{\Sigma}_m^{(l)}\right)^{-1}\boldsymbol{\mu}_m^{(l)}}{\kappa\left(u_i^{(l+1)}\right)} + \frac{\mathbf{M}_i\boldsymbol{\Sigma}_{\xi_i}^{-1}}{\kappa\left(u_i^{(l+1)}\right)} + \frac{\boldsymbol{\rho}^{(l+1)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(l+1)} - \sum_{j=1}^{s}\mathbf{b}_{ij}^T\boldsymbol{\alpha}_j^{(l)}\right)}{\sigma_{\epsilon_i}^2\kappa\left(u_i^{(l+1)}\right)}\right].$$

*Passo 7:* Gerar $\boldsymbol{\mu}_m^{(l+1)} \sim \mathcal{N}_q\left(\boldsymbol{\mu}_{\mu_m}, \boldsymbol{\Sigma}_{\mu_m}\right)$, em que

$$\boldsymbol{\Sigma}_{\mu_m} = \left[\left(\boldsymbol{\Sigma}_m^{(l)}\right)^{-1}\left(\sum_{r=1}^{n}\frac{1}{\kappa\left(u_i^{(l+1)}\right)}\right) + \boldsymbol{\Sigma}_{\mu_{m0}}^{-1}\right]^{-1} \quad \text{e}$$

$$\boldsymbol{\mu}_{\mu_m} = \boldsymbol{\Sigma}_{\mu_m}\left[\left(\boldsymbol{\Sigma}_m^{(l)}\right)^{-1}\left(\sum_{r=1}^{n}\frac{\mathbf{m}_i^{(l+1)}}{\kappa\left(u_i^{(l+1)}\right)}\right) + \boldsymbol{\Sigma}_{\mu_{m0}}^{-1}\boldsymbol{\mu}_{\mu_{m0}}\right].$$

*Passo 8:* Gerar $\left(\boldsymbol{\Sigma}_m^{(l+1)}\right)^{-1} \sim \mathcal{W}ishart\left(q+n, \boldsymbol{\Omega}_m^*\right)$, em que

$$\boldsymbol{\Omega}_m^* = \left[\boldsymbol{\Omega}_m^{-1} + \sum_{r=1}^{n}\frac{\left(\mathbf{m}_i^{(l+1)} - \boldsymbol{\mu}_m^{(l+1)}\right)\left(\mathbf{m}_i^{(l+1)} - \boldsymbol{\mu}_m^{(l+1)}\right)^T}{\kappa\left(u_i^{(l+1)}\right)}\right]^{-1}.$$

*Passo 9:* Gerar $\tau_{\alpha_j}^{2\,(l+1)} \sim \mathcal{IG}\left(\frac{K_j}{2} + a_{\tau_{\alpha_j}}, \frac{2b_{\tau_{\alpha_j}} + \left(\boldsymbol{\alpha}_j^{(l)} - \boldsymbol{\alpha}_{j0}\right)^T\left(\boldsymbol{\alpha}_j^{(l)} - \boldsymbol{\alpha}_{j0}\right)}{2}\right)$, $j = 1, \ldots, s$.

*Passo 10:* Gerar $\boldsymbol{\alpha}_j^{(l+1)} \sim \mathcal{N}_{K_j}\left(\boldsymbol{\mu}_{\alpha_j}, \boldsymbol{\Sigma}_{\alpha_j}\right)$, $j = 1, \ldots, s$, em que

$$\boldsymbol{\Sigma}_{\alpha_j} = \left[\frac{1}{\tau_{\alpha_j}^{2\,(l+1)}}\mathbf{I}_{K_j} + \mathbf{B}_j^T\left[\boldsymbol{\Sigma}_\epsilon\mathbf{L}_u^{(l+1)}\right]^{-1}\mathbf{B}_j\right]^{-1} \quad \text{e}$$

$$\boldsymbol{\mu}_{\alpha_j} = \boldsymbol{\Sigma}_{\alpha_j}\left[\frac{1}{\tau_{\alpha_j}^{2\,(l+1)}}\boldsymbol{\alpha}_{j0} + \mathbf{B}_j^T\left[\boldsymbol{\Sigma}_\epsilon\mathbf{L}_u^{(l+1)}\right]^{-1}\left(\mathbf{y} - \overline{\mathbf{X}}^{(l+1)}\tilde{\boldsymbol{\beta}}^{(l+1)} - \sum_{0<i<j}\mathbf{B}_i\boldsymbol{\alpha}_i^{(l+1)} - \sum_{j<i\leq s}\mathbf{B}_i\boldsymbol{\alpha}_i^{(l)}\right)\right].$$

*Passo 11:* Repetir os passos 2 - 10 até obter a convergência.

Com a amostra gerada de tamanho $R$ usando o algoritmo MCMC, podemos resumir as distribuições a posteriori usando

$$\bar{\boldsymbol{\beta}} = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\beta}^{(r)}, \quad \bar{\boldsymbol{\rho}} = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\rho}^{(r)}, \quad \bar{\boldsymbol{\alpha}}_j = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\alpha}_j^{(r)}, \quad \bar{\boldsymbol{\mu}}_m = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\mu}_m^{(r)}.$$

em que "(r)" representa o $r$-ésimo elemento da amostra a posteriori.

## 4.3 Estudo de simulação

Nesta seção conduzimos um estudo de simulação com objetivo de avaliar o desempenho do algoritmo MCMC descrito na seção anterior. Geramos uma amostra de tamanho $n = 200$ da seguinte forma:

$$\begin{cases} y_i = \beta_1 + \beta_2 \mathrm{x}_i + \rho \mathrm{m}_i + \frac{1}{2}\sin(2\pi v_i) + \epsilon_i, \\ \mathrm{M}_i = \mathrm{m}_i + \xi_i, \qquad i = 1, \ldots, n, \end{cases} \tag{4.4}$$

onde $(\epsilon_1, \xi_1, \mathrm{m}_1)^T, \ldots, (\epsilon_n, \xi_n, \mathrm{m}_n)^T$ são vetores aleatórios independentes seguindo distribuição $\mathcal{SMN}_3$; $\mathrm{x}_i \sim \mathcal{U}(0,1)$; $v_i \sim \mathcal{U}(0,1)$; $\sigma^2_{\epsilon_i} \overset{iid}{\sim} \mathcal{U}(0.2, 0.5)$; $\sigma^2_{\xi_i} \overset{iid}{\sim} \mathcal{U}(0.2, 0.5)$; $\beta_1 = \beta_2 = 1$, $\rho = 0.5$, $\mu_m = 2$ e $\sigma^2_m = 1$. Para a componente aleatória do modelo $(\epsilon_i, \xi_i, \mathrm{m}_i)^T \sim \mathcal{SMN}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta)$ foram consideradas as seguintes distribuições:

- normal;
- $t$-Student com $\eta = 3, 5, 8, 12$;
- slash com $\eta = 2, 4, 7, 11$;
- hiperbólica simétrica com $\eta = 0.8, 0.9, 1.0, 1.1$;
- normal contaminada com $\boldsymbol{\eta} = (0.4, 0.2)^T$,$(0.5, 0.2)^T$, $(0.55, 0.2)^T$, $(0.6, 0.2)^T$;
- Laplace,

em que $\boldsymbol{\mu} = (0, 0, \mu_m)$ e $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma^2_{\epsilon_i}, \sigma^2_{\xi_i}, \sigma^2_m)$. Geramos uma amostra da distribuição a posteriori dos parâmetros do modelo (4.4), em que a forma funcional de $f(v)$ é assumida como desconhecida mas aproximada usando $B$-splines cúbicos, com $[n^{1/5}]$ como o número de nós internos. Usamos as distribuições a priori descritas na seção 4.2 para os parâmetros do modelo, em que os valores dos hiperparâmetros são $\beta_0 = \rho_0 = \mu_{\mu_m} = 0$, $S_\beta = S_\rho = \Sigma_{\mu_m} = \Omega_m = 10^3$, $\boldsymbol{\alpha}_0 = \mathbf{0}_K$, e $a_{\tau_\alpha} = b_{\tau_\alpha} = 0.001$.

Em todos os casos foram geradas 110000 iterações do algoritmo MCMC, que incluem um período de aquecimento (*burn.in*) de 10000 e, com o objetivo de minimizar a autocorrelação das cadeias geradas, consideramos um espaçamento de 10 iteraçoes, de modo que obtemos uma amostra aproximadamente independente de tamanho $R = 10000$. O processo anterior foi replicado 100 vezes, mantendo os valores de $\mathbf{x}$ e $\mathbf{v}$ fixos. Vale salientar que, o estudo de simulação foi desenvolvido usando a função `fmem()` do pacote **BayesGESM**. Como medidas de resumo consideramos as definidas em (3.3) e (3.4), em que $\theta_1 = \beta_1, \theta_2 = \beta_2, \theta_3 = \rho, \theta_4 = \mu_m$ e $\theta_5 = \sigma^2_m$.

Na Tabela 4.1 apresentamos os valores das medidas $\mathrm{M}(\cdot)$ e $\mathrm{D}(\cdot)$ para os parâmetros do modelo (4.4). Note que os valores da medida $\mathrm{M}(\cdot)$ estão próximos dos valores dos parâmetros do modelo. Embora, em alguns cénarios, uma exceção parece ser o parâmetro $\sigma^2_m$, em que para obter uma estimativa mais próxima do verdadeiro valor precisamos uma amostra de tamanho maior. Além disso, em todos os casos os valores da medida $\mathrm{D}(\cdot)$ são maiores para os parâmetros associados às variáveis sem erro do modelo. O comportamento desta medida parece não depender das caudas da distribuição do termo de erro.

Por outro lado, estudamos a precisão das estimativas da função $f(v)$ apresentando na Figura 4.1 o verdadeiro valor da função e suas estimativas (linhas pontilhadas) nos diferentes cénarios de simulação. Podemos concluir que, as estimativas da função $f(v)$ apresentam um comportamento semelhante ao verdadeiro independentemente da distribuição considerada para a componente aleatória do modelo.

**Tabela 4.1:** *Valores das medidas de resumo* $\mathrm{M}(\cdot)$ *e* $\mathrm{D}(\cdot)$ *nos diferentes cenários de simulação.*

| Distribuição | Medida | Parâmetro | | | | |
|---|---|---|---|---|---|---|
| | | $\beta_1$ | $\beta_2$ | $\rho$ | $\mu_m$ | $\sigma_m^2$ |
| Normal | M | 0.9898 | 0.9699 | 0.5150 | 1.9793 | 0.9752 |
| | D | 0.2186 | 0.1160 | 0.1032 | 0.0560 | 0.1604 |
| $t(\eta = 3)$ | M | 1.0358 | 1.0130 | 0.4928 | 1.9472 | 1.0934 |
| | D | 0.1531 | 0.1901 | 0.0448 | 0.0626 | 0.1316 |
| $t(\eta = 5)$ | M | 0.9902 | 1.0049 | 0.5079 | 1.9888 | 1.0260 |
| | D | 0.1839 | 0.1849 | 0.0604 | 0.0469 | 0.1175 |
| $t(\eta = 8)$ | M | 1.0120 | 0.9618 | 0.5075 | 2.0006 | 1.2081 |
| | D | 0.1613 | 0.1831 | 0.0449 | 0.0547 | 0.0897 |
| $t(\eta = 12)$ | M | 1.0054 | 0.9313 | 0.5116 | 2.0082 | 1.1733 |
| | D | 0.1810 | 0.1832 | 0.0563 | 0.0379 | 0.1090 |
| $\mathcal{S}l(\eta = 2)$ | M | 0.9776 | 1.0559 | 0.5062 | 1.8251 | 1.1892 |
| | D | 0.1519 | 0.2033 | 0.0467 | 0.0636 | 0.1031 |
| $\mathcal{S}l(\eta = 4)$ | M | 1.0089 | 0.9612 | 0.5096 | 1.9478 | 1.1505 |
| | D | 0.1506 | 0.1616 | 0.0512 | 0.0451 | 0.0849 |
| $\mathcal{S}l(\eta = 7)$ | M | 1.0365 | 0.9436 | 0.5101 | 1.9978 | 1.0429 |
| | D | 0.1304 | 0.1561 | 0.0473 | 0.0376 | 0.0945 |
| $\mathcal{S}l(\eta = 11)$ | M | 1.0018 | 0.9545 | 0.5118 | 2.0104 | 1.0330 |
| | D | 0.1514 | 0.1425 | 0.0513 | 0.0367 | 0.0815 |
| $\mathcal{HS}(\eta = 0.8)$ | M | 1.0196 | 0.9701 | 0.5114 | 1.9500 | 0.9964 |
| | D | 0.1990 | 0.2908 | 0.0438 | 0.0979 | 0.0858 |
| $\mathcal{HS}\eta = 0.9)$ | M | 1.0050 | 0.9541 | 0.5103 | 1.9738 | 1.0878 |
| | D | 0.2020 | 0.2549 | 0.0403 | 0.1025 | 0.0991 |
| $\mathcal{HS}(\eta = 1.0)$ | M | 0.9804 | 1.0847 | 0.4971 | 2.0053 | 1.1320 |
| | D | 0.1864 | 0.2402 | 0.0495 | 0.0815 | 0.1206 |
| $\mathcal{HS}(\eta = 1.1)$ | M | 1.0335 | 0.9982 | 0.5027 | 1.9886 | 1.1183 |
| | D | 0.1702 | 0.2204 | 0.0520 | 0.0954 | 0.1116 |
| $\mathcal{NC}(0.4, 0.2)$ | M | 1.0199 | 0.9433 | 0.5040 | 1.9171 | 1.2732 |
| | D | 0.1870 | 0.2086 | 0.0493 | 0.0868 | 0.1253 |
| $\mathcal{NC}(0.5, 0.2)$ | M | 1.0119 | 0.9528 | 0.5057 | 1.9133 | 1.2654 |
| | D | 0.1980 | 0.2330 | 0.0492 | 0.0885 | 0.1131 |
| $\mathcal{NC}(0.55, 0.2)$ | M | 1.0114 | 0.9546 | 0.5056 | 1.9093 | 1.3283 |
| | D | 0.2034 | 0.2467 | 0.0458 | 0.0924 | 0.1107 |
| $\mathcal{NC}(0.6, 0.2)$ | M | 1.0167 | 0.9458 | 0.5052 | 1.8973 | 1.3294 |
| | D | 0.2116 | 0.2591 | 0.0437 | 0.0973 | 0.1049 |
| Laplace | M | 0.9293 | 1.1132 | 0.4998 | 1.8841 | 0.8699 |
| | D | 0.2337 | 0.3347 | 0.0367 | 0.1709 | 0.0763 |

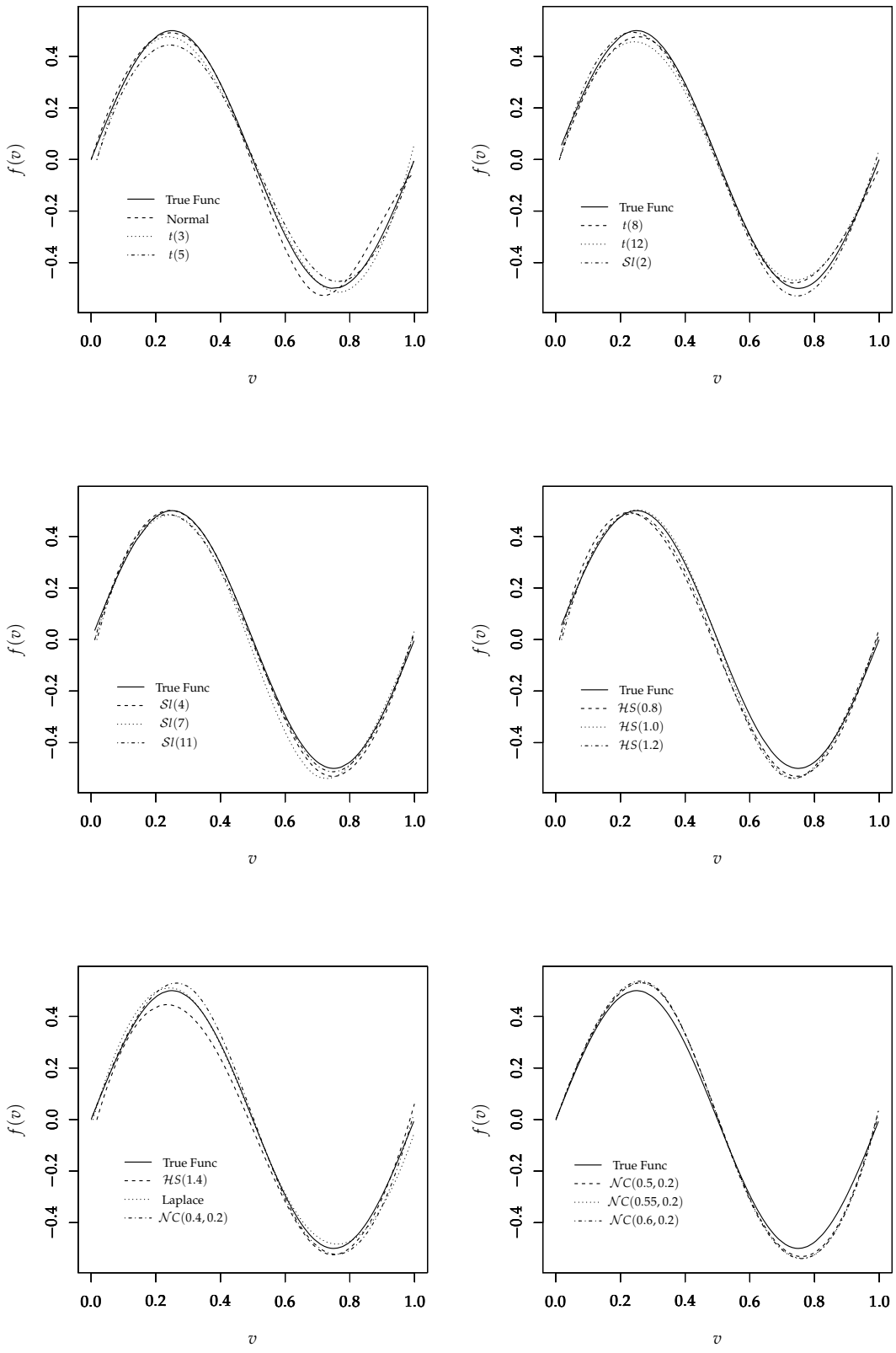


**Figura 4.1:** *Verdadeiro valor da função $f(v)$ contra suas estimativas (linhas pontilhadas) nos diferentes cenários de simulação.*

## 4.4 Aplicações

Com o objetivo de ilustrar o modelo proposto, consideramos um banco de dados construido apartir da pesquisa *American Community Survey* (http://www.census.gov/acs/www/) e do censo dos Estados Unidos e para o estado do Texas. Tambem, analisamos o conjunto de dados do projeto WHO MONICA (http://www.thl.fi/publications/monica/), o qual motivou os primeiros trabalhos desenvolvidos para a abordagem dos modelos heteroscedásticos com erros nas variáveis.

### 4.4.1 Renda média das famílias no Texas

Este conjunto de dados relaciona a renda média das famílias com a composição demográfica e algumas variáveis da força e do mercado de trabalho nos municipios do estado de Texas, nos Estados Unidos. A composição demográfica foi obtida do Censo dos Estados Unidos de 2010, enquanto que a renda média das famílias e as variáveis da força e do mercado de trabalho foram obtidas da pesquisa *American Community Survey* (ACS) realizada no período de 2009 a 2013. Portanto, os indivíduos correspondem aos 254 municipios do estado de Texas. Este banco de dados contém as seguintes variáveis:

*Fonte* : Pesquisa *American Community Survey*.

- `MeanIng`: Renda média da família no ultimo ano em centenas de dólares.
- `Plabor16`: Porcentagem de pessoas na força de trabalho (maiores de 16 anos).
- `PDesemp`: Taxa de desemprego.

*Fonte* : Censo do ano 2010.

- `DenPop`: Densidade populacional.
- `PHisp`: Porcentagem da população de origem hispânica.
- `PNegra`: Porcentagem da população de origem negra.
- `PFem`: Porcentagem de mulheres no total da população.

O erros amostrais das estimativas reportadas pela *ACS* são considerados aqui como erros de medição. Sendo assim, usamos o modelo MFEVH para analisar este conjunto de dados em que as variâncias amostrais reportadas pela *ACS* são tratadas como as variâncias dos erros de medição. Então, as variáveis com erro de medição foram obtidas da *ACS*, e as variáveis sem erro correspondem às obtidas do censo. O gráfico de dispersão (omitido aqui) da variável resposta (`meanIng`) contra `PHisp` sugere uma relação não-linear. Essa relação é descrita usando uma função não paramétrica. Então, usamos o seguinte modelo para analisar o conjunto de dados:

$$\begin{cases} \texttt{MeanIng}_i &= \beta_1 + \beta_2 \texttt{log}(\texttt{DenPop}_i) + \beta_3 \texttt{PNegra}_i + \beta_4 \texttt{PFem}_i + \mathbf{m}_i^T \boldsymbol{\rho} + f(\texttt{PHisp}_i) + \epsilon_i \\ \mathbf{M}_i &= \mathbf{m}_i + \boldsymbol{\xi}_i, \qquad i = 1, \ldots, 254, \end{cases} \tag{4.5}$$

em que $\mathbf{M}_i = (\texttt{Plabor16}_i, \texttt{PDesemp}_i)^T$ representa o vetor das variáveis explicativas observadas com erro para o indivíduo $i$, $\boldsymbol{\rho} = (\rho_1, \rho_2)^T$ e $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ são os parâmetros a serem estimados. Além disso, $(\epsilon_1, \boldsymbol{\xi}_1^T, \mathbf{m}_1^T), \ldots, (\epsilon_n, \boldsymbol{\xi}_n^T, \mathbf{m}_n^T)$ são vetores aleatórios independentes com distribuição $\mathcal{SMN}_5$.

As distribuições a priori consideradas para os parâmetros do modelo são as seguintes:

$$\boldsymbol{\beta} \sim \mathcal{N}_4(\mathbf{0}, 10^4 \times \mathbf{I}), \qquad \boldsymbol{\rho} \sim \mathcal{N}_2(\mathbf{0}, 10^4 \times \mathbf{I}), \qquad \boldsymbol{\alpha} \sim \mathcal{N}_K(\mathbf{0}, \tau_\alpha^2 \times \mathbf{I}_K),$$
$$\tau_\alpha^2 \sim \mathcal{GI}(0.001, 0.001), \quad \boldsymbol{\mu}_m \sim \mathcal{N}_2(\mathbf{0}, 10^4 \times \mathbf{I}), \quad \text{e} \quad \boldsymbol{\Sigma}_m^{-1} \sim \mathcal{W}ishart(2, 10^4 \times \mathbf{I}).$$

Foram geradas 110000 iterações do algoritmo MCMC, que incluem um período de aquecimento (*burn.in*) de 10000 e, com o objetivo de minimizar a autocorrelação das cadeias geradas, consideramos saltos de 10 iteraçoes, de modo que obtemos uma amostra aproximadamente independente de tamanho $R = 10000$. Para a componente aleátoria $(\epsilon_i, \boldsymbol{\xi}_i^T, \mathbf{m}_i^T)^T$ consideramos as distribuições normal, $t$-Student, slash, hiperbólica simétrica, normal contaminada e Laplace. Em todos os casos, o parâmetro extra é assumido conhecido. O ajuste sob o modelo normal sugere que para este conjunto de dados pode ser mais adequado um modelo em que a distribuição do erro aleatório apresenta caudas pesadas. Portanto, para os modelos $t$-Student, slash, hiperbólico simétrico e normal contaminado o ajuste fornecido usando vários valores do parâmetro extra foi comparado usando as medidas DIC e LMPL, de modo que o modelo com o "melhor desempenho" foi escolhido em cada caso.

Na Tabela 4.2 apresentamos a média a posteriori, o desvio padrão e o intervalo de credibilidade de 95% para os parâmetros dos diferentes modelos obtidos. Podemos observar que, o modelo em que o termo de erro segue distribuição slash com parâmetro extra $\eta = 3$ (i.e., $\mathcal{S}l(3)$) apresenta o menor DIC. Portanto, ele pode ser considerado como o melhor modelo para descrever o conjunto de dados. As Figuras 4.2 e 4.3 apresentam o comportamento das cadeias, bem como a aproximação das densidades a posteriori marginais para os parâmetros $\beta_1$, $\beta_2$, $\rho_1$, $\rho_2$, $\mu_{m_1}$, $\mu_{m_2}$, $\sigma^2_{m_1}$ e $\sigma^2_{m_2}$ sob o modelo $\mathcal{S}l(3)$. Nestes gráficos é possível observar que a convergência foi atingida e que a densidade a posteriori marginal é aproximadamente simétrica para os parâmetros. O ajuste da função estimada não paramétrica para a variável porcentagem da população com origem hispânica ($\hat{f}(\texttt{PHisp})$) é apresentado na Figura 4.4. Neste gráfico também pode-se observar intervalos de credibilidade do 95% para cada valor de $\hat{f}(\texttt{PHisp})$.

Os resultados indicam que o ingresso médio por família no estado de Texas não está muito influenciado pela composição demográfica e sim pelas caracteristicas do mercado e da força de trabalho. Por exemplo, os intervalos de credibillidade para $\beta_3$ e $\beta_4$ contem o zero e estes parâmetros estão associados às porcentagens de mulheres e de pessoas de raça negra que compõem a população do municipio, respectivamente. Contudo, os resultados sugerem que o ingresso médio por família é maior em municipios mais urbanos, ou seja, em municipios com uma densidade populacional alta.

### 4.4.2 Projeto WHO MONICA

O conjunto de dados do projeto WHO MONICA sobre doenças cardiovasculares é considerado. Este conjunto de dados tem sido analisado por varios autores usando modelos com erros nas variáveis (veja, por exemplo, de Castro *et al.*, 2008; Kulathinal *et al.*, 2002; Patriota *et al.* , 2009). Usamos o modelo MFEVH para relacionar as tendências do escore de risco médio anual ($x$) e as tendências em variação anual nas taxas de eventos coronários ($y$) para homens e mulheres. O escore de risco médio foi definido como uma combinação linear de tabagismo, pressão arterial sistólica, índice de massa corporal e o nível de colesterol total. Os coeficientes desta combinação foram obtidos a partir de um estudo de acompanhamento, utilizando o modelo de riscos proporcionais de Cox. Informações detalhadas sobre os dados podem ser encontrados em http://www.thl.fi/publications/monica/. A tendência nas taxas de eventos coronários é o coeficiente de uma tendência linear das taxas anuais ao longo de um período de cinco anos. Os erros amostrais das estimativas de tendências variam de população para população e são considerados como erros de medição. Portanto, os desvios de amostragem determinados a partir das estimativas das tendências podem ser consideradas como os erros de medição conhecidos (Kulathinal *et al.*, 2002). O seguinte modelo é usado para analisar o conjunto de dados para cada uma das populações:

$$\begin{cases} Y_i = \beta + \rho x_i + \epsilon_i \\ X_i = x_i + \xi_i, \qquad i = 1, \ldots, n. \end{cases} \tag{4.6}$$

**Tabela 4.2:** *Média a posteriori, desvio padrão, intervalo de credibilidade de 95% para os parâmetros dos modelos $\mathcal{SMN}_5$ para conjunto de dados* TEXAS.

| Parâmetro | Normal | | | *t*-Student(12) | | | Slash(3) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Média | DP | IC (95%) | Média | DP | IC (95%) | Média | DP | IC (95%) |
| $\beta_1$ | 101.14 | 30.33 | (42.27, 160.58) | 82.12 | 30.63 | (22.05, 142.09) | 83.04 | 31.24 | (21.48, 144.37) |
| $\beta_2$ | 36.91 | 3.36 | (30.32, 43.52) | 30.55 | 3.52 | (23.76, 37.63) | 29.67 | 3.54 | (22.70, 36.92) |
| $\beta_3$ | 2.60 | 1.33 | (0.06, 5.27) | 1.64 | 1.66 | (-1.60, 4.91) | 1.50 | 1.65 | (-1.78, 4.79) |
| $\beta_4$ | 2.27 | 0.64 | (1.04, 3.55) | 0.30 | 0.77 | (-1.21, 1.85) | -0.17 | 0.77 | (-1.68, 1.35) |
| $\rho_1$ | 14.48 | 0.98 | (12.53, 16.40) | 15.45 | 1.14 | (13.20, 17.68) | 15.54 | 1.12 | (13.25, 17.70) |
| $\rho_2$ | -77.26 | 3.26 | (-83.89, -71.10) | -69.75 | 4.85 | (-79.48, -60.40) | -68.90 | 5.31 | (-80.09, -59.04) |
| $\mu_{m_1}$ | 58.21 | 0.44 | (57.34, 59.08) | 58.36 | 0.42 | (57.52, 59.19) | 58.34 | 0.42 | (57.49, 59.18) |
| $\mu_{m_2}$ | 6.64 | 0.12 | (6.41, 6.88) | 6.94 | 0.13 | (6.69, 7.20) | 6.97 | 0.12 | (6.72, 7.21) |
| $\sigma^2_{m_1}$ | 46.83 | 4.65 | (38.55, 56.71) | 29.63 | 3.31 | (23.69, 36.63) | 23.45 | 2.70 | (18.58, 29.26) |
| $\sigma^2_{m_2}$ | 2.46 | 0.27 | (1.98, 3.06) | 1.59 | 0.22 | (1.19, 2.08) | 1.23 | 0.19 | (0.90, 1.65) |
| DIC | 5484.838 | | | 5214.271 | | | 5163.332 | | |

| Parâmetro | Hiperbolica(2.4) | | | N.Contaminada(0.1,0.4) | | | Laplace | | |
|---|---|---|---|---|---|---|---|---|---|
| | Média | DP | IC (95%) | Média | DP | IC (95%) | Média | DP | IC (95%) |
| $\beta_1$ | 76.74 | 30.95 | (16.53, 138.30) | 92.46 | 33.23 | (24.95, 156.33) | 62.66 | 31.378 | (1.12, 123.69) |
| $\beta_2$ | 29.54 | 3.61 | (22.63, 36.72) | 35.46 | 3.66 | (28.12, 42.57) | 25.82 | 3.57 | (19.07, 33.16) |
| $\beta_3$ | 1.19 | 1.51 | (-1.74, 4.21) | 2.13 | 1.44 | (-0.75, 4.95) | 0.31 | 1.65 | (-2.82, 3.57) |
| $\beta_4$ | 0.41 | 0.77 | (-1.11, 1.95) | 1.95 | 0.68 | (0.56, 3.26) | -0.15 | 0.79 | (-1.65, 1.44) |
| $\rho_1$ | 15.84 | 1.13 | (13.63, 18.15) | 14.86 | 1.08 | (12.75, 17.03) | 16.56 | 1.17 | (14.25, 18.81) |
| $\rho_2$ | -68.86 | 4.36 | (-77.67, -60.70) | -75.20 | 3.92 | (-82.80, -67.31) | -64.50 | 4.90 | (-74.44, -55.03) |
| $\mu_{m_1}$ | 58.39 | 0.41 | (57.57, 59.23) | 58.22 | 0.47 | (57.28, 59.17) | 58.39 | 0.40 | (57.60, 59.20) |
| $\mu_{m_2}$ | 6.95 | 0.12 | (6.70, 7.21) | 6.67 | 0.13 | (6.41, 6.92) | 7.04 | 0.14 | (6.75, 7.33) |
| $\sigma^2_{m_1}$ | 49.02 | 5.71 | (38.75, 61.22) | 35.38 | 8.02 | (18.57, 49.09) | 3.51 | 0.44 | (2.73, 4.45) |
| $\sigma^2_{m_2}$ | 2.68 | 0.38 | (2.00, 3.52) | 1.86 | 0.42 | (1.00, 2.62) | 0.20 | 0.03 | (0.15, 0.27) |
| DIC | 5304.171 | | | 5209.740 | | | 5484.944 | | |

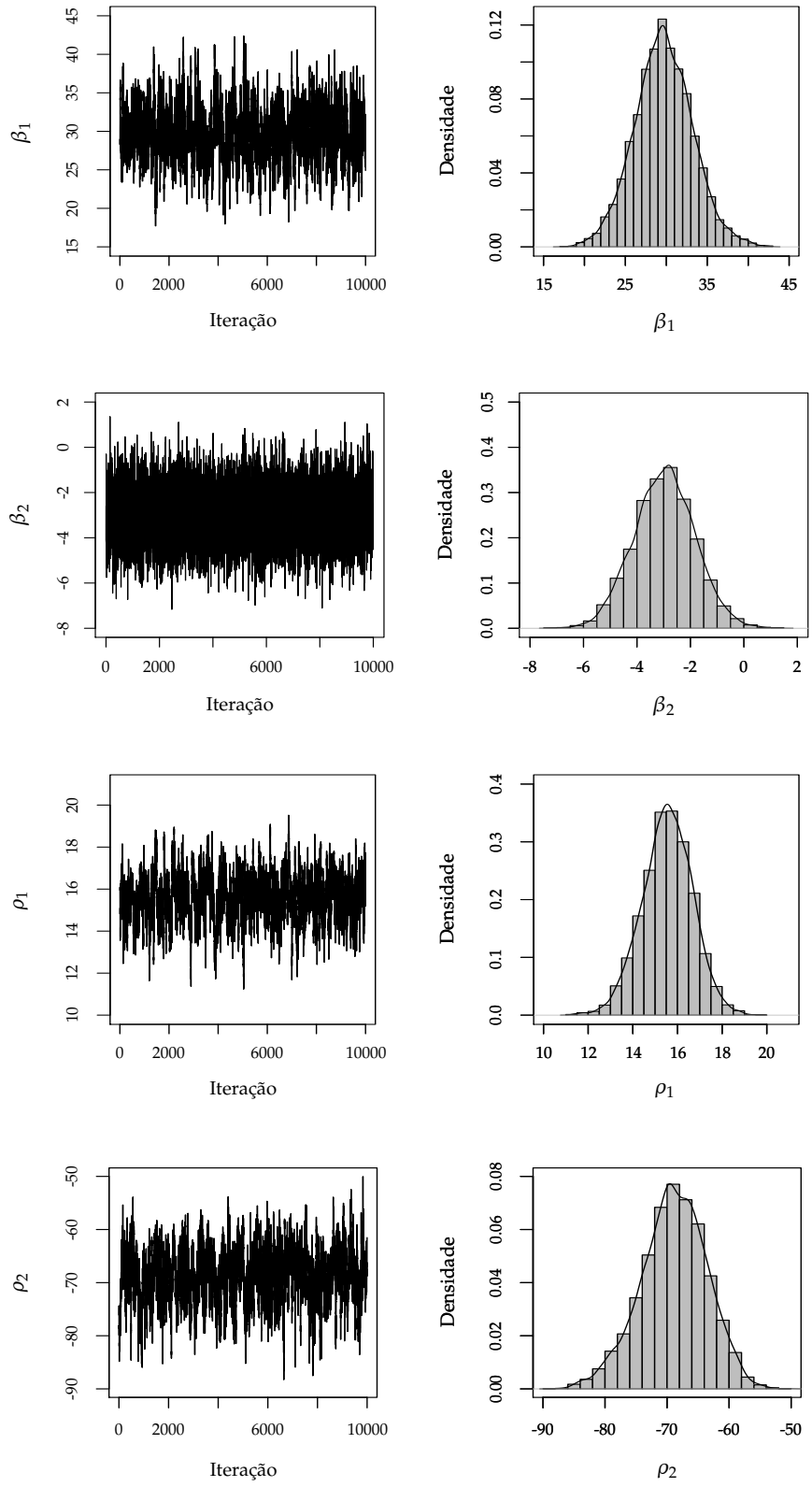


**Figura 4.2:** *Comportamento das cadeias e densidades marginais a posteriori dos parâmetros $\beta_1$, $\beta_2$, $\rho_1$ e $\rho_2$ sob o modelo $\mathcal{S}l(3)$ para o conjunto de dados* TEXAS.

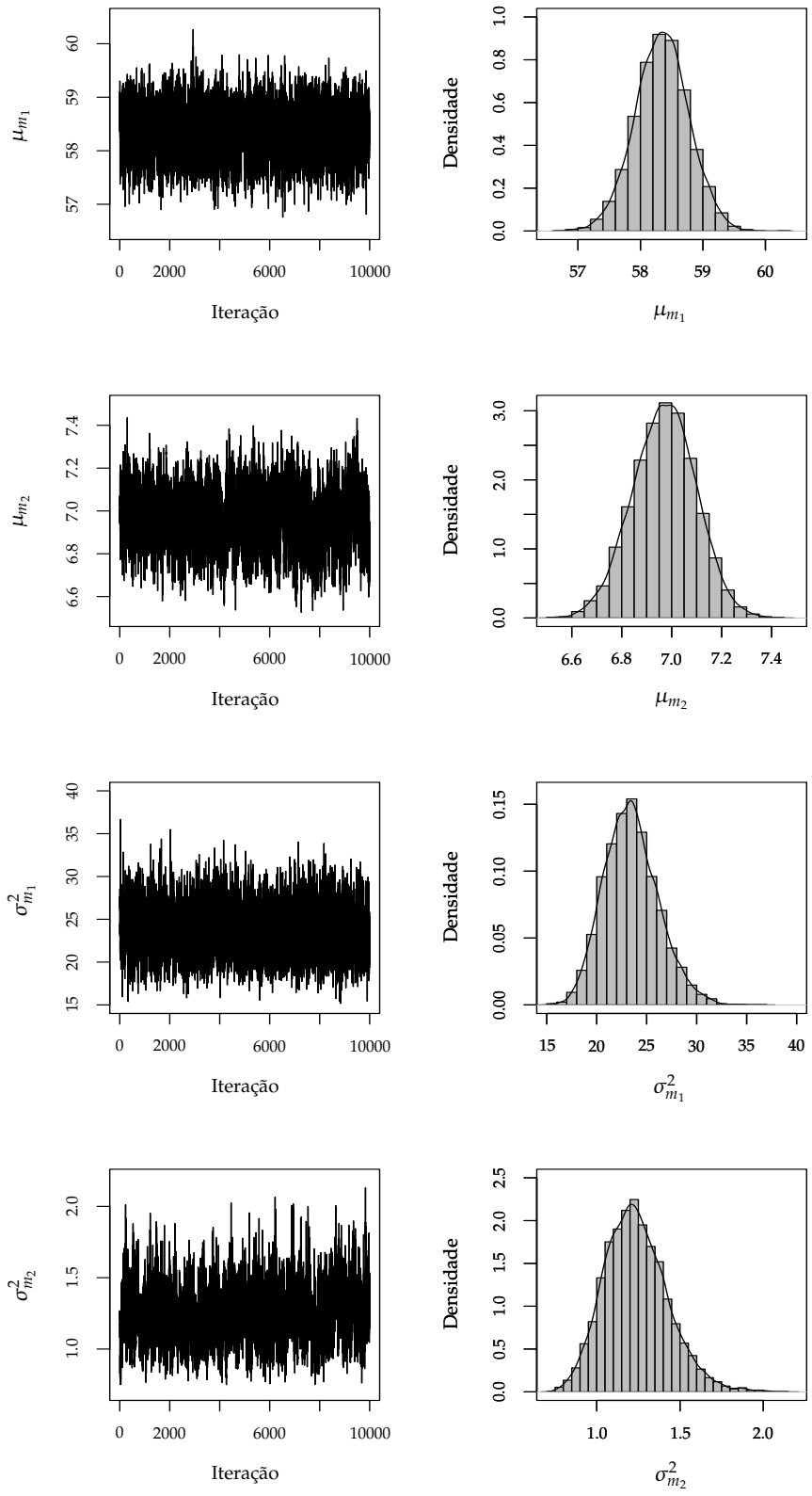


**Figura 4.3:** *Comportamento das cadeias e densidades marginais a posteriori dos parâmetros $\mu_{m_1}$, $\mu_{m_2}$, $\sigma^2_{m_1}$ e $\sigma^2_{m_2}$ sob o modelo $\mathcal{S}l(3)$ para o conjunto de dados* TEXAS.

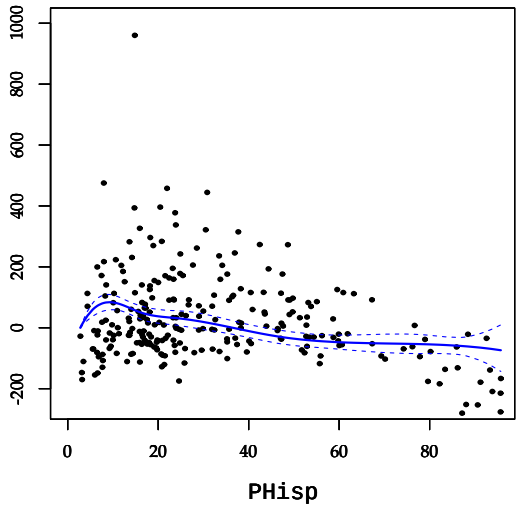


**Figura 4.4:** *Gráfico do ajuste da função não paramétrica* $\hat{f}(\texttt{PHisp})$*(linha contínua) e intervalos de credibilidade do 95% (linhas pontilhadas) para o conjunto de dados* TEXAS *sob o modelo* $\mathcal{S}l(3)$.

As distribuições a priori consideradas para os parâmetros do modelo são as seguintes:

$$\beta \sim \mathcal{N}(0, 10^4), \qquad \rho \sim \mathcal{N}(0, 10^4), \quad \mu_x \sim \mathcal{N}(0, 10^4).$$

Usando o algoritmo MCMC proposto obtemos uma amostra a posteriori dos parâmetros do modelo, geramos 210000 iterações, incluindo um período de aquecimento (*burn.in*) de 10000 e considerando um espaçamento de 10 iteraçoes, obtendo uma amostra aproximadamente independente de tamanho $R = 20000$. Foram consideradas as distribuições normal, $t$-Student, slash, hiperbólica, normal contaminada e Laplace para a componente aleatória do modelo, isto é, $(\epsilon_i, \xi_i, x_i)^T \sim \mathcal{SMN}_3$. Nas Tabelas 4.3 e 4.4 apresentamos a média a posteriori, o desvio padrão e o intervalo de credibilidade de 95% para os parâmetros dos diferentes modelos. Em cada caso, vários valores do parâmetro extra foram considerados, e selecionamos aquele que apresentou os melhores valores nos critérios de comparação DIC e LMPL. Como sugerido por Cao *et al.* (2012), foram considerados valores do parâmetro extra que induzem distribuições com caudas mais pesadas do que a distribuição normal. Dentro destes modelos, o modelo em que a componente aleatória segue distribuição normal contaminada apresenta o menor DIC e o maior valor do LMPL. Portanto, ele pode ser considerado como o melhor modelo para descrever o conjunto de dados dos homens quanto das mulheres. O modelo selecionado para o conjunto de dados dos homens pode ser ajustado com o uso da função `fmem()`, usando

```
model <- fmem(y ~ x, data=data_h, family="ContNormal", eta=c(0.15,0.3),
        burn.in=10000, post.sam.s=20000, thin=10, heter=heter)
summary(model)
```

As Figuras 4.5 e 4.6 apresentam o comportamento das cadeias, bem como as densidades marginais a posteriori aproximadas para os parâmetros $\beta$, $\rho$, $\mu_x$ e $\sigma_y^2$ sob o modelo normal contaminado. Estes gráficos sugerem que a convergência foi atingida e que a densidade a posteriori marginal é aproximadamente simétrica para os parâmetros $\beta$, $\rho$ e $\mu_x$, e assimetrica para $\sigma_x^2$.

**Tabela 4.3:** *Média a posteriori, desvio padrão, intervalo de credibilidade de 95% para os parâmetros dos modelos $\mathcal{SMN}_3$ para os homens do projeto WHO MONICA*

| Parâmetro | Normal | | | $t$-Student(12) | | | Slash(3) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Média | DP | IC (95%) | Média | DP | IC (95%) | Média | DP | IC (95%) |
| $\beta$ | -1.931 | 0.333 | (-2.57, -1.27) | -1.911 | 0.385 | (-2.63, -1.14) | -1.910 | 0.404 | (-2.65, -1.05) |
| $\rho$ | 1.207 | 0.141 | (0.94, 1.50) | 1.236 | 0.194 | (0.88, 1.65) | 1.252 | 0.205 | (0.89, 1.69) |
| $\mu_x$ | -1.052 | 0.341 | (-1.73, -0.39) | -1.002 | 0.325 | (-1.63, -0.35) | -1.016 | 0.328 | (-1.65, -0.36) |
| $\sigma_x^2$ | 4.046 | 1.101 | (2.38, 6.65) | 2.684 | 0.888 | (1.36, 4.85) | 2.138 | 0.728 | (1.04, 3.84) |
| DIC | 395.569 | | | 380.410 | | | 382.329 | | |
| LMPL | -208.729 | | | -201.911 | | | -203.591 | | |

| Parâmetro | Hiperbolica(2.3) | | | NormalCont.(0.15,0.3) | | | Laplace | | |
|---|---|---|---|---|---|---|---|---|---|
| | Média | DP | IC (95%) | Média | DP | IC (95%) | Média | DP | IC (95%) |
| $\beta$ | -1.250 | 0.665 | (-2.28, 0.25) | -1.449 | 0.411 | (-2.20, -0.58) | -1.878 | 0.313 | (-2.47, -1.23) |
| $\rho$ | 1.690 | 0.451 | (1.08, 2.71) | 1.455 | 0.203 | (1.10, 1.89) | 1.241 | 0.216 | (0.85, 1.69) |
| $\mu_x$ | -1.023 | 0.289 | (-1.58, -0.44) | -1.048 | 0.320 | (-1.68, -0.42) | -0.751 | 0.297 | (-1.33, -0.18) |
| $\sigma_x^2$ | 2.752 | 1.263 | (0.88, 5.81) | 2.897 | 0.933 | (1.48, 5.09) | 0.375 | 0.142 | (0.17, 0.72) |
| DIC | 345.572 | | | 339.191 | | | 394.744 | | |
| LMPL | -193.746 | | | -192.911 | | | -204.300 | | |

**Tabela 4.4:** *Média a posteriori, desvio padrão, intervalo de credibilidade de 95% para os parâmetros dos modelos $\mathcal{SMN}_3$ para as mulheres do projeto WHO MONICA*

| Parâmetro | Normal | | | $t$-Student(10) | | | Slash(3) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Média | DP | IC (95%) | Média | DP | IC (95%) | Média | DP | IC (95%) |
| $\beta$ | 1.800 | 0.528 | (0.83, 2.90) | 1.479 | 0.739 | (0.16, 3.07) | 1.421 | 0.713 | (0.12, 2.93) |
| $\rho$ | 1.120 | 0.214 | (0.74, 1.58) | 1.395 | 0.291 | (0.90, 2.04) | 1.390 | 0.280 | (0.91, 2.01) |
| $\mu_x$ | -2.059 | 0.337 | (-2.73, -1.40) | -2.033 | 0.332 | (-2.68, -1.38) | -2.014 | 0.335 | (-2.67, -1.34) |
| $\sigma_x^2$ | 3.607 | 1.042 | (2.05, 6.09) | 2.563 | 0.868 | (1.27, 4.66) | 2.182 | 0.721 | (1.10, 3.90) |
| DIC | 393.319 | | | 372.985 | | | 369.887 | | |
| LMPL | -211.779 | | | -201.797 | | | -201.653 | | |

| Parâmetro | Hiperbolica(2.4) | | | NormalCont.(0.1,0.3) | | | Laplace | | |
|---|---|---|---|---|---|---|---|---|---|
| | Média | DP | IC (95%) | Média | DP | IC (95%) | Média | DP | IC (95%) |
| $\beta$ | 3.664 | 2.128 | (1.06, 8.43) | 3.316 | 1.070 | (1.57, 5.75) | 3.441 | 1.863 | (1.05, 8.21) |
| $\rho$ | 2.470 | 0.963 | (1.35, 4.56) | 2.081 | 0.508 | (1.27, 3.24) | 2.329 | 0.824 | (1.25, 4.42) |
| $\mu_x$ | -2.100 | 0.288 | (-2.65, -1.52) | -2.078 | 0.297 | (-2.66, -1.50) | -2.034 | 0.315 | (-2.65, -1.47) |
| $\sigma_x^2$ | 2.656 | 1.314 | (0.74, 5.77) | 2.149 | 0.808 | (0.96, 4.07) | 0.258 | 0.138 | (0.06, 0.59) |
| DIC | 344.806 | | | 334.779 | | | 369.777 | | |
| LMPL | -196.217 | | | -194.800 | | | -197.028 | | |

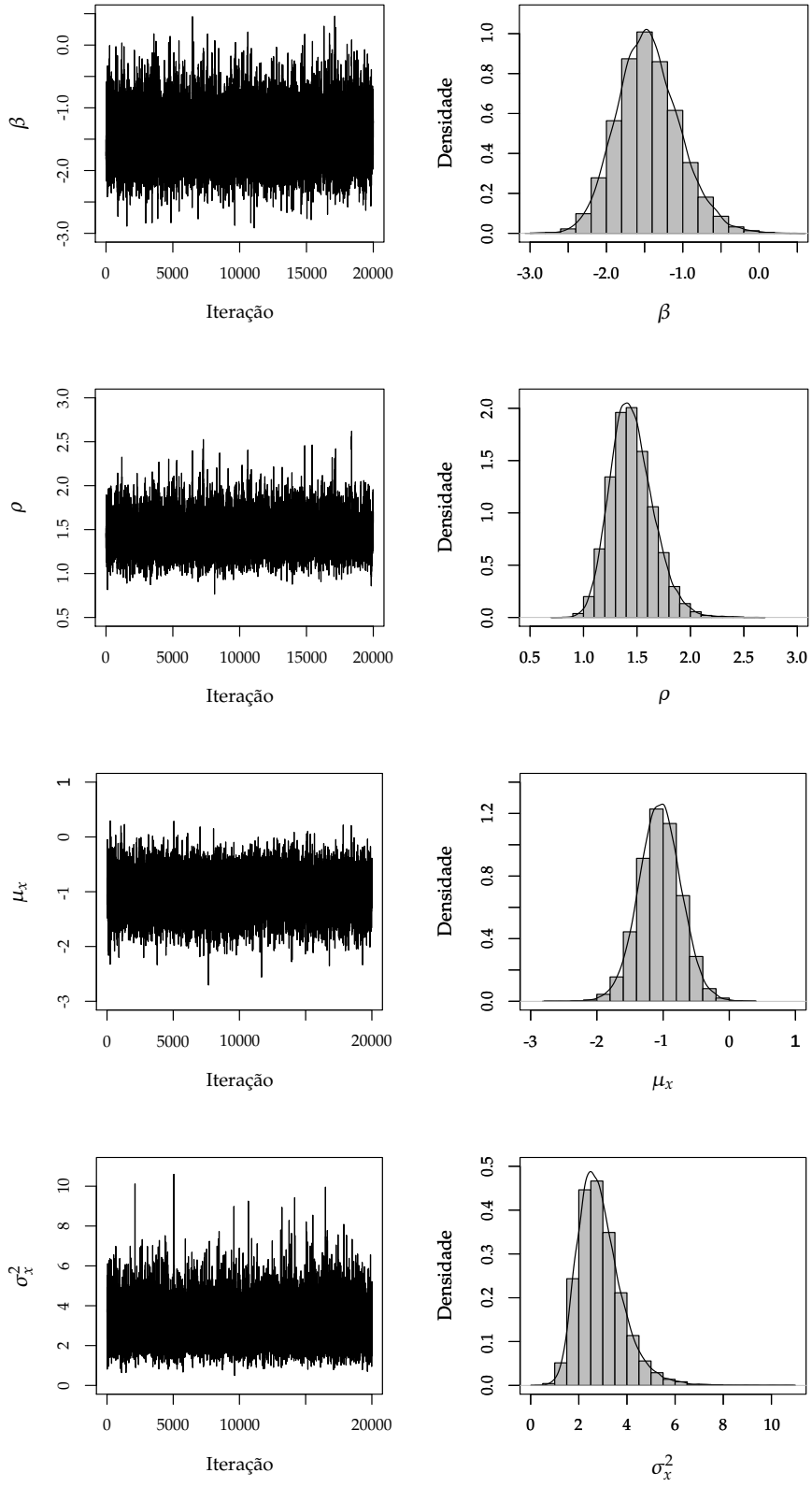


**Figura 4.5:** *Comportamento das cadeias e densidades marginais a posteriori dos parâmetros $\beta$, $\rho$, $\mu_x$ e $\sigma_x^2$ sob o modelo normal contaminado ($\mathcal{NC}(0.15, 0.3)$), para o conjunto de dados dos homens.*

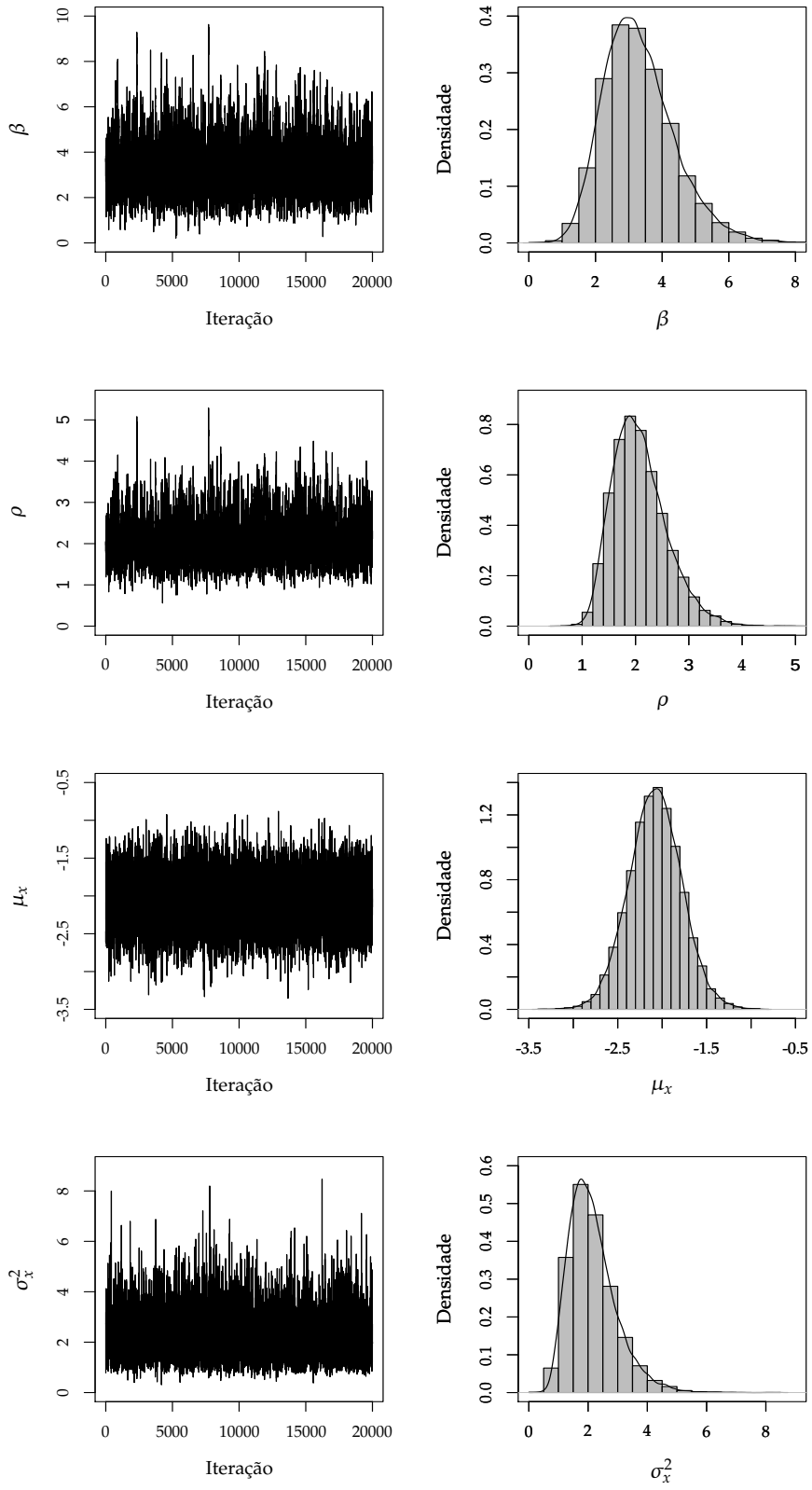


**Figura 4.6:** *Comportamento das cadeias e densidades marginais a posteriori dos parâmetros $\beta$, $\rho$, $\mu_x$ e $\sigma_x^2$ sob o modelo normal contaminado ($\mathcal{NC}(0.1, 0.3)$) para o conjunto de dados das mulheres.*

Capítulo 5

# Pacote BayesGESM

Neste capítulo apresentamos as principais funções do pacote **BayesGESM** (Rondon e Bolfarine , 2014). Este pacote foi desenvolvido na linguagem R (R Core Team, 2014), com o objetivo de fornecer uma ferramenta para realizar a inferência estatística baseada na abordagem Bayesiana para os modelos apresentados nesta tese. O pacote **BayesGESM** encontra-se disponível no "Comprehensive R Archive Network" (CRAN) em http://CRAN.R-project.org/package=BayesGESM e pode ser instalado livremente.

As funções principais do pacote, `gesm()` e `fmem()`, permiten obter infêrencias a posteriori dos parâmetros dos modelos estudados nesta tese. Além disso, estas funções fornecem medidas da qualidade do ajuste do modelo, como o DIC e a LMPL, bem como algumas medidas de diagnóstico, como resíduos e medidas de influência global baseadas na divergência de Kullback-Leibler e na distância $\chi^2$ (veja Peng e Dey, 1995; Weiss e Cook, 1992).

## 5.1 Função `gesm()`

Esta função permite ajustar o modelo semiparamétrico aditivo elíptico generalizado apresentado no capítulo 2. Os argumentos da função `gesm()` são os seguintes

```
gesm(formula, data, family, eta, burn.in, post.sam.s, thin),
```

em que

`formula`: permite descrever a componente sistemática do modelo de interesse, usando funções paramétricas e não paramétricas para os parâmetros de localização e dispersão do modelo. Este argumento é composto por três partes, a saber: $(i)$ variável resposta; $(ii)$ variáveis para descrever o parâmetro de localização; $(iii)$ variáveis para descrever o parâmetro de dispersão. Em $(ii)$ e $(iii)$ é possível incluir uma componente não paramétrica aproximada usando $B$-splines. As duas primeiras partes são separadas pelo símbolo "~" e a segunda e terceira parte são separadas pelo símbolo "`|`". As componentes não paramétricas podem ser especificadas usando a função `bsp()`.

`data`: objeto do tipo *data.frame*, onde estão armazenadas as variáveis resposta e explicativas.

`family`: distribuição considerada para o erro do modelo apresentado em 2.1. As distribuições que podem ser consideradas são normal, $t$-Student, slash, hiperbólica simétrica, Laplace e normal contaminada. Na função `gesm()` estas distribuições são especificadas por *"Normal"*, *"Student-t"*, *"Slash"*, *"Hyperbolic"*, *"Laplace"* e *"ContNormal"*, respectivamente.

`eta`: um valor ou vetor numérico que representa o parâmetro extra da distribuição considerada para o erro do modelo.

`burn.in`: o número de iterações consideradas para o período de aquecimento do algoritmo MCMC 2.2.2.

`post.sam.s`: o tamanho da amostra a posteriori de $\theta$.

`thin`: (opcional) tamanho dos saltos usados no algoritmo MCMC para obter o tamanho da amostra a posteriori requerido.

A função `gesm()` retorna os seguintes valores

`chains`: uma matriz que contém a amostra a posteriori dos parâmetros do modelo. Cada coluna representa a amostra da distribuição marginal a posteriori de um parâmetro.

`res`: um vetor que contém os resíduos descritos em 2.3.2 para o modelo ajustado.

`K-L`: um vetor que contém a medida de divergência Kullback-Leibler para o modelo ajustado.

`X_2`: um vetor que contém a distância $\chi^2$.

`DIC`: o valor do critério de seleção de modelos DIC.

`EBIC`: o valor do critério de seleção de modelos EBIC.

`EAIC`: o valor do critério de seleção de modelos EAIC.

`LMPL`: o valor da log-verossimilhança marginal para o modelo ajustado.

## 5.2 Função `fmem()`

A função `fmem()` permite ajustar o modelo flexível com erros nas variáveis apresentado no capítulo 3, e o modelo flexível com erros nas variáveis heterocedástico estudado no capítulo 4. Os argumentos da função `fmem()` são

```
fmem(formula, data, family, eta, burn.in, post.sam.s,
     thin, omeg, heter),
```

em que

`formula`: permite descrever a componente sistemática do modelo de interesse. Este argumento é composto por três partes, a saber: $(i)$ variável resposta; $(ii)$ variáveis com erro de medição $(iii)$ variáveis sem erro de medição e efeitos não lineares que podem ser especificados usando a função `bsp()`. As duas primeiras partes são separadas pelo símbolo "~" e a segunda e terceira parte são separadas pelo símbolo "`|`".

`data`: objeto do tipo *data.frame*, onde estão armazenadas as variáveis usadas no modelo.

`family`: distribuição considerada para a componente aleatória dos modelos. As distribuições que podem ser consideradas são normal, *t*-Student, slash, hiperbólica simétrica, Laplace e normal contaminada. Na função `fmem` estas distribuições são especificadas por *"Normal"*, *"Student-t"*, *"Slash"*, *"Hyperbolic"*, *"Laplace"* e *"ContNormal"*, respectivamente.

`eta`: (opcional) um valor ou vetor numérico que representa o parâmetro extra da distribuição considerada para a componente aleatória do modelo. Se este parâmetro não é fornecido pelo usuário, ele será estimado dos dados introduzindo um passo adicional no algoritmo MCMC.

`burn.in`: o número de iterações consideradas para o período de aquecimento do algoritmo MCMC 3.2.2 ou 4.2.2.

`post.sam.s`: o tamanho da amostra a posteriori de $\theta$.

`thin`: (opcional) tamanho dos saltos usados no algoritmo MCMC para obter o tamanho da amostra a posteriori requerido.

`omeg`: (opcional) o valor considerado para $\omega = \sigma_\epsilon^2/\sigma_\xi^2$, este valor deve ser especificado quando o modelo de interesse é o apresentado no capítulo 3, ou seja, o modelo flexível com erros nas variáveis homocedástico. Se este valor não é especificado é assumido como 1, isto é, $\sigma_\epsilon^2 = \sigma_\xi^2$.

`heter`: (opcional) objeto do tipo *list* que contém os valores de $\sigma_{\epsilon_i}^2$ e $\mathbf{\Sigma}_{\xi_i}$ para todo $i$ ($i = 1, \ldots, n$). Os objetos deven ser especificados como `sigma2y` e `sigma2xi`, isto é, `heter <- list(sigma2y, sigma2xi)`. Se o argumento `heter` não é especificado o modelo ajustado pela funcão `fmem()` será o modelo homocedástico.

Aplicando a função `fmem()` no R, obtemos

`chains`: uma matriz que contém a amostra a posteriori dos parâmetros do modelo. Cada coluna representa a amostra da distribuição marginal a posteriori de um parâmetro, que pode ser usada para avaliar se a convergência foi atingida.

`res`: um vetor que contém os resíduos para o modelo ajustado.

`K-L`: um vetor que contém a medida de divergência Kullback-Leibler para o modelo ajustado.

`X_2`: um vetor que contém a distância $\chi^2$.

`DIC`: o valor do critério de seleção de modelos DIC.

`LMPL`: o valor da log-verossimilhança marginal para o modelo ajustado.

## 5.3 Outras funções

No pacote **BayesGESM** implementamos adicionalmente as seguintes funções:

`summary()`: Esta função apresenta o resumo da análise Bayesiana do modelo ajustado com as funções `gesm()` e `fmem()`. Este resumo contém uma tabela com as estatísticas de resumo: média, mediana, desvio padrão e intervalo de credibilidade de 95% da distribuição a posteriori dos parâmetros do modelo. Além disso, esta função exibe medidas da qualidade de ajuste do modelo tais como DIC e LMPL.

`bsp()`: Esta função é usada para calcular a matriz de base do spline para aproximar as componentes não paramétricas do modelo. Os argumentos desta função são: `x`, que corresponde aos valores da variável explicativa, e `kn` (opcional) que permite especificar o número de nós internos do *B*-spline. O valor por defecto para os nós internos é $[n^{1/5}]$. Por exemplo, suponha que queremos que a variável $v$ seja aproximada usando *B*-splines, com 4 nós internos. Então, no argumento `formula` temos `bsp(v,4)` . Os *B*-splines implementados nesta função são cúbicos.

`bsp.graph.gesm()`: Esta função mostra os gráficos dos efeitos não paramétricos de um objecto da classe `gesm()`. Os argumentos desta função são `object` e `which`, em que `object` é um objeto da classe `gesm()` e `which` é um valor inteiro, onde 1 indica que a componente não paramétrica foi especificada no parâmetro de localização e 2 indica que a componente não paramétrica foi especificada no parâmetro de dispersão.

`bsp.graph.fmem()`: Esta função mostra os gráficos dos efeitos não paramétricos de um objecto da classe `fmem()`. Os argumentos desta função são `object` e `which`, em que `object` é um objeto da classe `fmem()` e `which` é um valor inteiro que indica qual é a componente não paramétrica requerida.

## 5.4 Uso das funções

Nesta seção o objetivo é apresentar os códigos no R para o ajuste dos modelos usando as funções descritas anteriormente.

### 5.4.1 Coelhos europeus

Este conjunto de dados foi apresentado na seção 2.5. O modelo proposto para descrever o peso das lentes dos olhos do $i$-ésimo coelho foi dado em (2.10). Usando a função `gesm()` podemos ajustar o modelo através de

```
### Leitura dos dados
data("Erabbits", package="ssym")
### Para ajustar o modelo
fit <- gesm(wlens ~ bsp(age) | bsp(age), data=Erabbits,
       family="ContNormal", eta=c(0.8,0.9), burn.in=5000,
       post.sam.s=5000, thin=10)
### Resultado obtidos
summary(fit)

       Error distribution: ContNormal ( 0.8 0.9 )
              Sample size: 71
 Size of posterior sample: 5000

 =================   Location Submodel   =================
 =========   Nonparametric part
           Mean    Median     SD         C.I. 95%
alpha 1  23.6786  23.6167   1.3431    21.1088  26.607
alpha 2  56.6431  56.6413   2.6990    51.4161  62.086
alpha 3 145.5083 145.5955   3.9813   137.5404 153.252
alpha 4 228.7898 228.6964   8.0047   212.8565 244.720
alpha 5 228.6626 228.4163   7.9634   213.4230 245.521
alpha 6 246.3082 246.5567   3.4333   238.6283 252.867

   Graphs of the nonparametric effects are provided by
   using the function 'bsp.graph.gesm(fit,1)'

 =================  Dispersion Submodel  =================
 =========   Nonparametric part
           Mean   Median    SD         C.I. 95%
lambda 1  1.5794  1.5202  0.7360     0.3037  3.1734
lambda 2  3.4082  3.3714  0.6651     2.2154  4.8126
lambda 3  4.0359  4.0206  0.6835     2.7288  5.4225
lambda 4  4.2628  4.2666  1.3769     1.6466  6.9849
lambda 5  6.3306  6.2305  2.0477     2.6377 10.6762
lambda 6  1.0075  1.0137  1.8529    -2.7591  4.6409

   Graphs of the nonparametric effects are provided by
   using the function 'bsp.graph.gesm(fit,2)'

 ==================     Model Selection Criteria     ==================
 DIC= 495.3543    EAIC= 507.8627    EBIC= 500.267    LMPL= -249.931

### Gráfico da função não paramétrica da média
bsp.graph.gesm(fit,1)
### Gráfico da função não paramétrica da dispersão
```

```
bsp.graph.gesm(fit,2)
```

Note que o modelo ajustado para este conjunto de dados não considera variáveis explicativas paramétricas. Entretanto, relações lineares de variáveis explicativas podem ser consideradas para o parâmetro de localização, bem como para o parâmetro de dispersão. Estas relações podem ser especificadas no argumento `formula` da função `gesm()`.

### 5.4.2 Boston

O conjunto de dados BOSTON foi analizado na seção 3.4.2. Lembremos que, o modelo considerado para a análise foi o modelo flexível com erros nas variáveis (homocedástico). Para este conjunto consideramos o modelo (3.5), que pode ser ajustado usando a função `fmem()` da seguinte forma

```
### Leitura dos dados
library(MASS)
data(Boston)
attach(Boston)
### Ajuste do modelo normal contaminado
fit <- fmem(log(medv) ~ nox | crim + rm + bsp(lstat) + bsp(dis),
       data=Boston, family="ContNormal", burn.in=10000,
       post.sam.s=5000, omeg=4, thin=10)
### Resultados
summary(fit)
       Error distribution: ContNormal
              Sample size: 506
 Size of posterior sample: 5000

 ===============    Parametric part   ===============
 =======   Covariates measured without error
              Mean   Median    SD         C.I. 95%
(Intercept)  3.6988  4.0943  0.2213     3.2830  4.1451
crim        -0.0127 -0.0128  0.0016    -0.0156 -0.0101
rm           0.1773  0.1669  0.0185     0.1396  0.2028

 =======   Covariates measured with error
      Mean   Median    SD         C.I. 95%
nox -1.1192 -1.0952  0.2441    -1.5739 -0.6642

 ===============   Nonparametric part ===============
 Effects    internal knots
lstat             3
dis               3

 Graphs of the nonparametric effects are provided by
 using the function 'bsp.graph.fmem'

 =======   Dispersion parameter
            Mean  Median   SD         C.I. 95%
Sigma2_y   0.0156 0.0160 0.0015     0.0126  0.0193

  Ratio of the error variances:  4

 ==========   Model Selection Criteria   ==========
 DIC= -1017.565    LMPL= 504.631

##### Gráficos das componentes não paramétricas do modelo
### Gráfico para a variável lstat
bsp.graph.fmem(fit,1)
### Gráfico para a variável dis
bsp.graph.fmem(fit,2)
```

### 5.4.3 Texas

Na seção 4.4.1 analisamos o conjunto de dados referente ao estado do Texas, usando o modelo flexível com erros nas variáveis heterocedástico. No R o modelo (4.5) pode ser ajustado usando a função `fmem()` usando

```
###Leitura dos dados
TexasData <- read.table("TexasData.txt",header=TRUE)
attach(TexasData)
### Leitura das variâncias
nu <- 3
zeta <- nu/(nu-1)
hetert <- list(sigma2y=VarMeanIng/zeta, sigma2xi=cbind(VarPopLabor16,
          VarP_Desemp)/zeta)
### Ajuste do modelo
model <- fmem(MeanIng/100 ~  PopLabor16 + P_Desemp | log(DensPobla) +
         P_PbFem +  P_PNegra + bsp(P_PbHisp), data=TexasData, family="Slash",
         eta=nu, burn.in=10000, post.sam.s=10000, heter=hetert,thin=10)
### Resultados
summary(model)
       Error distribution: Slash ( 3 )
              Sample size: 254
 Size of posterior sample: 10000

 ===============    Parametric part   ===============
 =======   Covariates measured without error
                 Mean   Median    SD         C.I.  95%
(Intercept)    83.0461 83.1827 31.2474    21.4751 144.3678
log(DensPobla) 29.6750 29.7493  3.5446    22.6984  36.9191
P_PbFem         1.5033  1.5874  1.6588    -1.7762   4.7906
P_PNegra       -0.1710 -0.1795  0.7762    -1.6849   1.3479

 =======   Covariates measured with error
              Mean    Median     SD         C.I. 95%
PopLabor16  15.5434  15.4795   1.1290    13.2530  17.6969
P_Desemp   -68.9054 -69.1270   5.3162   -80.0907 -59.0365

 ===============   Nonparametric part ===============
 Effects    internal knots
P_PbHisp             3

 Graphs of the nonparametric effects are provided by
 using the function 'bsp.graph.fmem'


 ==========   Model Selection Criteria   ==========
 DIC= 5163.3320    LMPL= -3243.7610
```

Capítulo 6

# Considerações finais

Nesta trabalho foram apresentados modelos de regressão semiparamétricos sob a abordagem Bayesiana, onde sua componente aleatória segue distribuições de mistura normal na escala. No capítulo 2 propusemos os modelos semiparamétricos aditivos elípticos generalizados, em que o parâmetro de localização bem como o de dispersão incluem componentes não paramétricas aditivas aproximadas usando *B*-splines. Por otro lado, nos capítulos 3 e 4 apresentamos os modelos com erros nas variáveis homocedásticos e heterocedásticos, respectivamente. Nestes modelos a componente sistemática admite variáveis explicativas com e sem erro de medição bem como a presenca de efeitos não lineares aproximados usando *B*-splines. Desenvolvimos algoritmos MCMC, com o objetivo de gerar amostras das distribuições a posteriori dos parâmetros dos modelos estudados. Além disso, apresentamos critérios que permitem a comparação dos modelos e algumas ferramentas de diagnóstico no contexto Bayesiano. Para cada um dos algoritmos MCMC propostos, foi implementado um estudo de simulação para ilustrar o seu desempenho. As metodologias propostas são aplicadas em conjuntos de dados reias. Também, com o objetivo de fornecer uma ferramenta computacional para realizar a inferência estatística baseada na abordagem Bayesiana para os modelos apresentados neste trabalho, foi desenvolvido o pacote **BayesGESM** na linguagem R.

Como possíveis pesquisas futuras que podem ser derivadas a partir dos resultados obtidos neste trabalho, tem-se as seguintes:

- Estender os modelos estudados neste trabalho à classe de distribuições de mistura normal na escala assimétrica.
- Estender o modelo apresentado no capítulo 4, considerando modelos com erro na equação.
- Estender os modelos apresentados no capítulo 2, incluindo efeitos aleatórios aditivos na componente sistemática.
- Estender os modelos apresentados nos capítulos 3 e 4, considerando efeitos aditivos mistos.

# Referências Bibliográficas


Aitkin M (1987). Modelling variance heterogeneity in normal regression using glim. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **36**, 332–339. Citado na pág 1, 9

Andrews DF & Mallows CL (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society Series B (Methodological)*, **36**, 99–102. Citado na pág 2, 4

Arellano-Valle RB, Bolfarine H & Labra V (1996). Ultrastructural elliptical models. *Canadian Journal of Statistics*, **24**, 207–216. Citado na pág 1

Barndoff-Nielsen O (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London Series A, Mathematical & Physical Sciences*, **353**, 401–419. Citado na pág 2

Belsley DA, Kuh E & Welsch RE (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons. Citado na pág 40

Box GEP & Tiao GC (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley. Citado na pág 2

Brooks SP (2002). Discussion on the paper by spiegelhalter, best, carlin e van der linde. *Journal of the Royal Statistical Society: Series B (Methodological)*, **64**, 616–618. Citado na pág 14

Cao CZ, Lin JG & Zhu XX (2012). On estimation of a heteroscedastic measurement error model under heavy-tailed distributions. *Computational Statistics and Data Analysis*, **56**, 438–448. Citado na pág 1, 57

Carroll RJ, Roeder K & Wasserman L (1999). Flexible parametric measurement error models. *Biometrics*, **55**, 44–54. Citado na pág 1

Carroll RJ, Ruppert D, Stefanski LA & Crainiceanu CM (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. 2º edition. Chapman and Hall: Boca Raton. Citado na pág 2

Cepeda E & Gamerman D (2001). Bayesian modelling of variance heterogeneity in normal regression models. *Brazilian Journal of Probability and Statistics*, **14**, 207–221. Citado na pág 1, 9, 22

Cheng CL & VanNess JM (1999). *Statistical Regression with Measurement Error*. Arnold: London. Citado na pág 1, 25, 26

Csiszar I (1967). Information-type measures of difference of probabilily distributions and indirect observations. *Studia Scientiarum Mathematicarum Humgarica*, **2**, 299–318. Citado na pág 16

de Boor C (1978). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer-Verlag, New York. Citado na pág 7, 26

de Castro M, Bolfarine H & Galea M (2013). Bayesian inference in measurement error models for replicated data. *Environmetrics*, **24**, 22–30. Citado na pág 2

de Castro M, Galea M & Bolfarine H (2008). Hypothesis testing in an errors-in-variables model with heteroscedastic measurement errors. *Statistics in Medicine*, **27**, 5217–5234. Citado na pág 45, 53

Dudzinski ML & Mykytowycz R (1961). The eye lens as an indicator of age in the wild rabbit in australia. *CSIRO Wildlife Research*, **6**, 156–159. Citado na pág 18

Dunn PK & Smyth GK (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244. Citado na pág 15

Eilers PHC & Marx BD (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**, 89–121. Citado na pág 7, 10

Fang KT, Kotz S & Ng KW (1990). *Symmetrical Multivariate and Related Distributions*. Chapman and Hall, London. Citado na pág 4

Fuller MA (1987). *Measurement Error Models*. Wiley: New York. Citado na pág 1, 25

Gamerman D & Lopes HF (2006). *Markov Chain Monte Carlo*. 2º edition. Chapman and Hall. Citado na pág 4

Gelfand A, Dey D & Chang H (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics*, **4**, 147–167. Citado na pág 15

Gelman A, Roberts GO & Gilks WR (1995). Efficient metropolis jumping rules. *Bayesian Statistics*, **5**, 599–608. Citado na pág 14

Gelman A & Rubin DB (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, **7**, 457–472. Citado na pág 5

Geyer CJ (1992). Practical markov chain monte carlo. *Statistical Science*, **7**, 473–511. Citado na pág 14

Harrison DJ (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, **5**, 81–102. Citado na pág 40

Hastings WK (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109. Citado na pág 5

He X, Fung W & Zhu Z (2005). Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association*, **100**, 1176–1184. Citado na pág 7

Jørgensen B (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*. Springer, New York. Citado na pág 4

Kelly BC (2007). Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal*, **665**, 1489–1506. Citado na pág 1, 45

Kulathinal SB, Kuulasmaa K & Gasbarra D (2002). Estimation of an errors-in-variables regression model when the variances of the measurement errors vary betwwen the observations. *Statistics in Medicine*, **21**, 1089–1101. Citado na pág 1, 45, 53

Lange KL & Sinsheimer JS (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, **2**, 175–198. Citado na pág 4

Liang L, Palta M & Shao J (2004). A measurement error model with a poisson distributed surrogate. *Statistics in Medicine*, **23**, 2527–2536. Citado na pág 1

Maronna RA, Martin DR & Yohai VJ (2006). *Robust Statistics: Theory and Methods*. Wiley: New York. Citado na pág 1

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH & Teller E (1953). Equation of state calculations by fast computing machines. *The Jorunal of Chemical Physics*, **21**, 1087–1092. Citado na pág 5

Nadarajah S & Kotz S (2006). The exponentiated type distributions. *Acta Applicandae Mathematicae*, **92**, 97–111. Citado na pág 12

Patriota AG, Bolfarine H & de Castro M (2009). A heteroscedastic structural errors-in.variables model with equation error. *Statistical Methodology*, **6**, 408–423. Citado na pág 1, 45, 53

Peng F & Dey DK (1995). Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics*, **23**, 199–213. Citado na pág 16, 61

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. Citado na pág 2, 25, 61

Ratkowsky DA (1983). *Nonlinear Regression Modelling*. Marcel Dekker, New York. Citado na pág 18

Roberts G (1996). *Markov chain concepts related to sampling algorithms*. Chapman and Hall, London. In *Markov Chain Monte Carlo in Practice* (Gilks, W.R., Richardson, S. e Spiegelhalter, D.J., Eds), pp. 45-57. Citado na pág 13

Rogers WH & Tukey JW (1972). Understanding some long-tailed symmetrical distributions. *Statistica Neerlandica*, **26**, 211–226. Citado na pág 2

Rondon LM & Bolfarine H (2014). ***BayesGESM**: Bayesian Analysis of Generalized Elliptical Semi-Parametric Models and Flexible Measurement Error Models*. R package version 1.3. URL http://CRAN.R-project.org/package=BayesGESM. Citado na pág 2, 9, 25, 61

Ruppert D, Wand MP & Carroll RJ (2003). *Semiparametric Regression*. Cambridge University Press. Citado na pág 34

Spiegelhalter DJ, Best NG, Carlin BP & der Linde AV (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B (Methodological)*, **64**, 583–640. Citado na pág 14

Stark PC, Ryan LM, McDonald JL & Burge HA (1997). Using meteorologic data to model and predict daily ragweed pollen levels. *Aerobiologia*, **13**, 177–184. Citado na pág 34

Tanner MA & Wong WH (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–550. Citado na pág 11, 25

Verbyla AP (1993). Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society: Series B (Methodological)*, **55**, 493–508. Citado na pág 1, 9

Watson GN (1995). *A Treatise on the Theory of Bessel Functions*. Cambridge Mathematical Library. Citado na pág 4

Wei BC (1998). *Exponential Family Nonlinear Models*. Springer, New York. Lectures Notes in Statistics. Citado na pág 22

Weiss R (1996). An approach to bayesian sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 739–750. Citado na pág 16

Weiss R & Cook R (1992). A grafical case statistic for assessing posterior influence. *Biometrika*, **72**, 51–55. Citado na pág 16, 61

Xu D & Zhang Z (2013). A semiparametric bayesian approach to joint mean and variance models. *Statistics and Probability Letters*, **83**, 1624–1631. Citado na pág 1, 9, 22