UFAM

ESTIMAÇÃO BAYESIANA EM MODELOS DE REGRESSÃO T DE STUDENT COM ERROS NAS VARIÁVEIS, RESPOSTAS MULTIVARIADAS E CENSURAS

Márcia Brandão de Oliveira Martins

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática

Orientador: Celso Rômulo Barbosa Cabral

Manaus

Novembro de 2016

ESTIMAÇÃO BAYESIANA EM MODELOS DE REGRESSÃO T DE STUDENT COM ERROS NAS VARIÁVEIS, RESPOSTAS MULTIVARIADAS E CENSURAS

Márcia Brandão de Oliveira Martins

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA, DA UNIVERSIDADE FEDERAL DO AMAZONAS, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MATEMÁTICA.

Examinada por:

Prof^{a}. Larissa Avila Matos, D.Sc.

Prof. James Dean Oliveira dos Santos Júnior, D.Sc.

Prof. Celso Rômulo Barbosa Cabral, D.Sc.

MANAUS, AM – BRASIL

NOVEMBRO DE 2016

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Martins, Marcia Brandão de Oliveira
M386e Estimação Bayesiana em modelos de regressão T de student com erros nas variáveis, respostas multivariadas e censuras / Marcia Brandão de Oliveira Martins. 2016
43 f.: il.; 31 cm.

Orientador: Celso Romulo Barbosa Cabral
Dissertação (Mestrado em Matemática - Estatística) - Universidade Federal do Amazonas.

1. Modelos com erros nas variáveis. 2. Algoritmo de Gibbs. 3. Truncamento. 4. Estatística. I. Cabral, Celso Romulo Barbosa II. Universidade Federal do Amazonas III. Título

Este trabalho de dissertação é
dedicado a Deus que é GRANDE e
à Nossa Senhora das Graças.

Agradecimentos

A Deus por todas as graças concedidas ao longo dessa vida tão cheia de dificuldades e a Nossa Senhora das Graças pela intercessão.

Ao Professor Celso Rômulo por ter acreditado no meu potencial mesmo quando não me conhecia, tendo me orientado pacientemente desde os tempos de graduação, não me deixando desistir nas diversas vezes em que tropecei durante o caminho. Quando as dificuldades apareceram, me amparou, auxiliou e contribuiu diretamente para o meu crescimento profissional. Agradeço muito pela amizade e pelo grande exemplo de vida.

À minha mãe e ao meu irmão por me ajudarem a trilhar este caminho, em especial à minha mãe, que mesmo tendo muito pouco, nunca deixou de me incentivar a estudar e mudar meu destino.

Ao Leonardo, meu amado, pela paciência, dedicação, respeito e companheirismo, por me amparar principalmente nos momentos mais estressantes e difíceis deste curso.

Aos meus amigos Camila, Carina, Carla, Diego, Nelson, Regina e Vanessa por estarem sempre presentes, me ajudando com seus conhecimentos, me dando forças, conselhos, orações e por sempre acreditarem no meu melhor.

Aos professores do Departamento de Estatística da UFAM, por todos os ensinamentos transmitidos ao longo desses anos. Em especial ao professor José Raimundo, por ter me incentivado a buscar sempre um pouco mais nos meus primeiros anos nesta universidade.

À CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior) pela assistência financeira ao PPGMAT.

“O melhor ainda está por vir.”

(Autor desconhecido)

Resumo da Dissertação apresentada ao Programa de Pós-Graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do grau de Mestre em Matemática. (M.Sc.)

ESTIMAÇÃO BAYESIANA EM MODELOS DE REGRESSÃO T DE STUDENT COM ERROS NAS VARIÁVEIS, RESPOSTAS MULTIVARIADAS E CENSURAS

Márcia Brandão de Oliveira Martins

Novembro/2016

Orientador: Celso Rômulo Barbosa Cabral

Área de Concentração : Estatística

Apresentamos uma proposta de extensão para o modelo de regressão com erro nas variáveis usual em que tanto o vetor de respostas quanto a covariável estão sujeitos à censura. Assumimos que a distribuição conjunta da covariável e dos erros de observação é t de Student, que é uma alternativa ao modelo normal, porém com caudas pesadas. Um algoritmo do tipo Gibbs sampler é proposto para proceder a estimação Bayesiana dos parâmetros no modelo. Três estudos de simulação são realizados, mostrando a maior flexibilidade do modelo, em relação ao modelo sob normalidade, em ajustar dados com padrão de censura e caudas pesadas, além de uma aplicação em dados reais.

Abstract of Dissertation presented to Postgraduate in Mathematics, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

BAYESIAN INFERENCE FOR MULTIVARIATE MEASUREMENT ERRORS MODELS BASED ON T DISTRIBUTION WITH CENSORING IN BOTH VARIABLES

Márcia Brandão de Oliveira Martins

November/2016

Advisor: Celso Rômulo Barbosa Cabral

Research area: Statistics

We propose an extension of the usual normal regression model where both the vector of responses and the covariate are possibly censored. We assume that the jointly distribution of covariate and errors is Student-t, which is an alternative to the normal distribution, but with heavy tails. A Gibbs-type algorithm is proposed to carry out Bayesian estimation of the parameters in the model. Three simulation studies are conducted, showing that the proposed model is more flexible than the normal one when fitting data with censoring pattern and heavy tails, in addition to an application with real data.

Sumário

Lista de Figuras **xi**

Lista de Tabelas **xii**

1 Introdução **1**

1.1 Preliminares 1

1.1.1 Truncamento 2

1.1.2 Censura 2

1.1.3 A Distribuição t de Student Multivariada 3

1.1.4 A Distribuição t de Student Multivariada Truncada 5

1.2 O Modelo de Regressão Normal Multivariado com Erro de Medida 5

1.3 O Modelo de Regressão com Erro nas Variáveis Normal Multivariado com Ambas Variáveis Censuradas 8

1.4 Organização do trabalho 8

2 Modelos de Regressão t de Student com Erros nas Variáveis, Respostas Multivariadas e Censuras. **10**

2.1 Modelos de Regressão t de Student com Erros nas Variáveis, Respostas Multivariadas e Censuras (t-MEMC) 11

2.1.1 A Função de Verossimilhança 12

3 Estimação via MCMC **14**

3.1 Distribuições a Priori 14

3.2 Um Algoritmo do Tipo Gibbs 15

3.2.1 Detalhes do Algoritmo 16

3.3 Critérios de Seleção de Modelos 24

3.3.1 O DIC observado 24
3.3.2 WAIC 26

4 Aplicação com Dados Simulados e Reais **28**
4.1 Dados Simulados 28
4.1.1 Estudo de Simulação 1 28
4.1.2 Estudo de Simulação 2 30
4.1.3 Estudo de Simulação 3 32
4.2 Aplicação em Dados Reais 33

5 Conclusão **39**

Referências Bibliográficas **40**

Lista de Figuras

4.1 Estimativas de μ_x e σ_x^2 para os casos propostos no estudo de simulação 3 . . 32

4.2 Estimativas de ω_1^2, ω_2^2 e ω_3^2 para os casos propostos no estudo de simulação 3 32

4.3 Estimativas de α_1, α_2, β_1 e β_2 para os casos propostos no estudo de simulação 3 33

4.4 Estimativas de ν para os casos propostos no estudo de simulação 3 34

4.5 *Traceplots* e histogramas das estimativas MCMC dos parâmetros μ_x, σ_x^2, ω_1^2, α_1, β_1 e ν no ajuste do t-MEMC para o conjunto de dados Chipkevitch. . . 38

Lista de Tabelas

4.1 Estimativas dos vícios e dos erros quadráticos médios para os diferentes níveis de censura do t-MEMC no estudo de simulação 1. 30

4.2 Cobertura dos intervalos de credibilidade construídos ao nível de 95% de credibilidade para as estimativas MCMC do t-MEMC no estudo de simulação 1. 30

4.3 Média e Percentual de vezes em que cada modelo foi melhor segundo os métodos de comparação DIC_{obs} e WAIC, para diferentes níveis de censura do estudo de simulação 2. 31

4.4 Dados reais extraídos de Chipkevitch *et al.* (1996) sobre medições do volume testicular de 42 adolescentes sob 5 diferentes instrumentos. 35

4.5 Estimativas MCMC para os parâmetros nos ajustes dos modelos t-MEMC e N-MEMC para o conjunto de dados Chipkevitch. 36

4.6 Intervalos de credibilidade obtidos nos ajustes dos modelos t-MEMC e N-MEMC para o conjunto de dados Chipkevitch. 37

4.7 Critérios de seleção obtidos nos ajustes dos modelos t-MEMC e N-MEMC para o conjunto de dados Chipkevitch. 37

Capítulo 1

Introdução

1.1 Preliminares

Inicialmente, vamos definir algumas notações que serão utilizadas ao longo deste trabalho. Seja $X \sim \mathrm{N}(\mu, \sigma^2)$ uma variável aleatória com distribuição normal com média μ e variância σ^2. Então, $\phi(\cdot|\mu, \sigma^2)$ denota a sua função densidade de probabilidade e $\Phi(\cdot|\mu, \sigma^2)$ a sua respectiva função de distribuição. Quando fazemos $\mu = 0$ e $\sigma^2 = 1$, dizemos que a variável aleatória X tem distribuição normal padrão e sua função de distribuição pode ser denotada por $\Phi(\cdot|0,1)$. Seguindo a convenção tradicional denotamos uma variável aleatória por letra maiúscula (X) e sua realização por letra minúscula (x). Vetores aleatórios e matrizes são indicados por letras maiúsculas em negrito $\mathbf{X}$. A operação transposta de um vetor ou matriz será denotada por $\mathbf{X}^\top$ e $\mathbf{X} \perp Y$ indica que os vetores ou matrizes $\mathbf{X}$ e $\mathbf{Y}$ são independentes. A notação $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indica um vetor aleatório com distribuição normal p-variada com vetor de médias $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^\top$ e matriz de covariâncias de dimensão $(p \times p)$ dada por

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}^2 \end{bmatrix}.$$

Regressão é uma das ferramentas mais amplamente utilizadas em estatística para analisar a influência de algumas variáveis sobre alguns outros fatores, com o objetivo de

descobrir padrões explicativos e preditivos.

Assim, para um vetor de parâmetros $\boldsymbol{\theta}$ e um vetor de variáveis regressoras ou explanatórias $\mathbf{x}$ desejamos encontrar uma representação adequada da distribuição condicional, $f(\mathbf{y}|\boldsymbol{\theta},\mathbf{x})$, de uma variável resposta $\mathbf{Y}$ observável, com base em uma amostra de $\mathbf{x}$ e $\mathbf{y}$ (Marin & Robert, 2014). No contexto Bayesiano, isto é feito através da análise da distribuição a posteriori $f(\boldsymbol{\theta}|\mathbf{y},\mathbf{x})$.

Há situações em que a medição de $\mathbf{y}$ e $\mathbf{x}$ podem estar sujeitas a um limite de quantificação, isto é, um certo limite abaixo ou acima em que a medição não é avaliada. Essas medidas podem ser submetidas a um limite de detecção superior ou inferior (por isso, censurados à direita ou à esquerda). Contudo, sob este aspecto a modelagem pode levar a vícios. Há uma diferença muito tênue nos conceitos sobre as principais causas de ocorrência de observações incompletas em conjunto de dados, são conceitos muito parecidos, porém não coincidentes: *truncamento* e *censura*.

1.1.1 Truncamento

O truncamento ocorre quando algumas observações ou indivíduos envolvidos no estudo são completamente excluídos por algum motivo relacionado à natureza da investigação, ou seja, a ocorrência do truncamento se caracteriza especificamente por uma condição imposta ao conjunto de dados que inviabiliza a participação de certos indivíduos no estudo.

Um exemplo onde há a ocorrência de truncamento é apresentado em Colosimo & Giolo (2006), onde é analisada a distribuição do tempo de vida dos moradores de uma certa região, observando-se uma amostra extraída do banco de dados da previdência local. Neste caso, somente aqueles moradores que chegaram a atingir a aposentadoria fazem parte da amostra. Assim, as observações são truncadas à esquerda. Outros exemplos sobre truncamento em um conjunto de dados podem ser encontrados em Nelson (1990), Kalbfleisch & Lawless (1992) e Massuia *et al.* (2015).

1.1.2 Censura

Uma observação é dita ser censurada quando informações sobre a variável de interesse não estão completamente disponíveis para algumas unidades no conjunto de dados. A ocorrência de censura pode ser devida a limitações dos equipamentos de medição, plano

amostral, etc. Assim, quando a censura ocorre, temos disponível na base de dados apenas parte da informação sobre um evento de interesse.

Um exemplo bastante esclarecedor é o de uma agulha de um instrumento utilizado para aferir a massa corpórea que não fornece uma leitura acima de 200 kg, esta mostrará 200 kg para todos os indivíduos que ultrapassem este limite. Breen (1996) apresenta outro exemplo interessante: em um exame escolar o percentual mínimo de acertos para obter aprovação é de 40%. O certificado contendo a situação do aluno (aprovado ou reprovado) é dado a todos, porém só aqueles alunos que atingiram aprovação tem especificada sua pontuação exata; no caso em que se deseja estudar a pontuação do aluno como função de outras variáveis explicativas, saberemos somente que este valor é menor que 40% para os reprovados.

1.1.3 A Distribuição t de Student Multivariada

A distribuição t de Student multivariada é uma extensão da tradicional distribuição t de Student univariada que, como é de amplo conhecimento, tem um papel central na teoria estatística. Ela tem similaridades com a distribuição normal, como a simetria, porém por pertencer a uma familia de distribuições de caudas mais pesadas, pode gerar valores mais extremos. Como uma alternativa ao uso da distribuição normal, em situações onde a distribuição dos dados apresenta caudas pesadas, é aplicada na solução de problemas em áreas que incluem, por exemplo, análise discriminante, classificação, regressão multivariada e análise de dados com observações faltantes. Alguns excelentes textos que tratam da distribuição t de Student são os de Kotz & Nadarajah (2004) e Ahsanullah & Kibria (2014).

Dizemos que um vetor aleatório p-dimensional $\mathbf{Y}$ tem distribuição t de Student p-variada com vetor de médias $\boldsymbol{\mu}$, matriz de escala $\boldsymbol{\Sigma}$ e ν graus de liberdade, quando a sua densidade é dada por

$$\mathrm{t}_p(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma(\frac{p+\nu}{2})}{\Gamma(\frac{\nu}{2})\pi^{p/2}}\nu^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\left(1+\frac{\mathrm{d}_{\boldsymbol{\Sigma}}(\mathbf{y},\boldsymbol{\mu})}{\nu}\right)^{-(p+\nu)/2},$$

onde $\Gamma(\cdot)$ é a função gama e $\mathrm{d}_{\boldsymbol{\Sigma}}(\mathbf{y},\boldsymbol{\mu}) = (\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ é a distância de Mahalanobis. Neste caso, usaremos a notação $\mathbf{Y} \sim \mathrm{t}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$. Podemos ainda escrever a variável $\mathbf{Y}$ da

seguinte forma

$$\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}, \quad \mathbf{Z} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma}), \quad U \sim \mathrm{Gamma}(\nu/2, \nu/2), \tag{1.1}$$

onde $\mathrm{Gamma}(a,b)$ denota a distribuição Gama com média a/b e variância a/b^2 e U e $\mathbf{Z}$ são independentes.

Se em (1.1) fizermos $U \sim \mathrm{Beta}(\nu, 1)$ teremos a distribuição Slash, com densidade dada por:

$$\mathrm{SL}_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = 2\nu \int_0^1 \phi(\mathbf{y}|\boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma})\Phi(u^{1/2}\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu}))du, \tag{1.2}$$

que pode ser avaliada usando utilizando métodos tradicionais de integração numérica. Esta distribuição será utilizada nas aplicações com dados simulados que faremos no Capítulo 4.

Denotamos a função de distribuição de $\mathbf{Y} \sim \mathrm{t}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ por $\mathrm{T}_p(.|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Se $p = 1$, então não utilizamos o índice p nas notações t_p e T_p. Se $\nu > 1$, então $\mathrm{E}[\mathbf{y}] = \boldsymbol{\mu}$. Se $\nu > 2$, então $\mathrm{Cov}[\mathbf{y}] = \nu(\nu-2)^{-1}\boldsymbol{\Sigma}$. Quando $\nu \to \infty$, a variável latente U tende para 1 com probabilidade 1, e então, temos que a variável aleatória em (1.1) com distribuição t de Student tende para uma variável aleatória com distribuição $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

O resultado a seguir mostra que a classe das distribuições t de Student é fechada para marginalizações e condicionamentos. A demonstração do item (i) é imediata, e a demonstração do item (ii) segue da Proposição 4 apresentada em Arellano-Valle & Genton (2010) e da Proposição 1 de Matos *et al.* (2013)

Proposição 1. *Seja* $\mathbf{Y} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. *Considere a partição* $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top)^\top$, *com* $\mathbf{Y}_1 : p_1 \times 1$ *e* $\mathbf{Y}_2 : p_2 \times 1$. *De maneira conforme, considere as partições* $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$ *e* $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{ij})$, $i, j = 1, 2$. *Então*

(i) $\mathbf{Y}_1 \sim \mathrm{t}_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \nu)$;

(ii) $\mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1 \sim \mathrm{t}_{p_2}(\boldsymbol{\mu}_{2.1}, \widetilde{\boldsymbol{\Sigma}}_{22.1}, \nu + p_1)$,

onde $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$, $\widetilde{\boldsymbol{\Sigma}}_{22.1} = \frac{\nu + \mathrm{d}_{\boldsymbol{\Sigma}_{11}}(\mathbf{y}_1, \boldsymbol{\mu}_1)}{\nu + p_1}\boldsymbol{\Sigma}_{22.1}$ *e* $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

1.1.4 A Distribuição t de Student Multivariada Truncada

Seja $\mathbf{Y} \sim \mathrm{t}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Seja $\mathbb{D}$ um Boreliano de $\mathbb{R}^p$. Dizemos que o vetor aleatório $\mathbf{Z}$ tem *distribuição t de Student truncada em* $\mathbb{D}$ quando a distribuição de $\mathbf{Z}$ for a mesma de $\mathbf{Y}|(\mathbf{Y} \in \mathbb{D})$. Neste caso, a densidade de $\mathbf{Z}$ é dada por

$$\mathrm{Tt}_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{D}) = \frac{\mathrm{t}_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}{P(\mathbf{Y} \in \mathbb{D})} \mathbb{I}_{\mathbb{D}}(\mathbf{z}),$$

onde $\mathbb{I}_{\mathbb{D}}(\cdot)$ é a função indicadora do conjunto $\mathbb{D}$, ou seja, $\mathbb{I}_{\mathbb{D}}(\mathbf{z}) = 1$ se $\mathbf{z} \in \mathbb{D}$ e $\mathbb{I}_{\mathbb{D}}(\mathbf{z}) = 0$ caso contrário. Usamos a notação $\mathbf{Z} \sim \mathrm{Tt}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{D})$. Se $\mathbb{D}$ é da forma

$$\mathbb{D} = \{(x_1, \ldots, x_p) \in \mathbb{R}^p;\ x_1 \leq d_1, \ldots, x_p \leq d_p\},$$

usamos a notação $(\mathbf{Y} \in \mathbb{D}) = (\mathbf{Y} \leq \mathbf{d})$, onde $\mathbf{d} = (d_1, \ldots, d_p)^\top$. Neste caso, temos que $P(\mathbf{Y} \leq \mathbf{d}) = \mathrm{T}_p(\mathbf{d}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Observe que podemos ter $d_i = +\infty$, $i = 1, \ldots, p$.

1.2 O Modelo de Regressão Normal Multivariado com Erro de Medida

Em modelos de regressão usuais os valores dos regressores são fixados. Quando os regressores são considerados como variáveis aleatórias observadas com erro, utilizar os procedimentos de estimação tradicionais pode levar a vícios (Fuller, 1987, Cap. 1).

Existem muitas situações na prática em que a covariável não pode ser observada diretamente, mas com erros. Dentre muitas aplicações neste contexto, temos a medição de massa com uma balança. Nesse caso, é sabido que fatores aleatórios, como correntes de ar, desregulagem do equipamento ou até mesmo vibrações, introduziriam erros nas mensurações. Posto isto, o verdadeiro valor da variável explanatória é considerado como um valor não observado, pois assume-se que a ele deve ser adicionada alguma informação inerente ao erro de medida.

O modelo de regressão linear multivariado simples é definido por

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta} x_i + \mathbf{e}_i, \tag{1.3}$$

onde $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ir})^\top$ é um vetor aleatório de medições feitas no indivíduo i, os vetores $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_r)^\top$ e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_r)^\top$ são vetores de parâmetros de regressão desconhecidos, $x_i, i = 1, ..., n$, é o valor da variável regressora para o indivíduo i, e $\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{ir})^\top$ é um vetor de erros aleatórios.

Suponha que x_i não pode ser observada diretamente, mas em vez disto observamos

$$X_i = x_i + \xi_i. \tag{1.4}$$

Então a variável observada x_i é mensurada com um erro aleatório ξ_i associado à observação i. Este modelo é conhecido na literatura como o modelo de regressão com erros nas variáveis. Uma outra denominação é modelo de regressão com erro de medida.

Um estudo pioneiro neste tipo de modelagem foi introduzido no século XIX. A descrição de seu desenvolvimento histórico pode ser encontrado em Sprent (1990), e um estudo detalhado é descrito por Kendall & Stuart (1961), Fuller (1987) e Cheng & Van Ness (1999). Um estudo mais recente pode ser encontrado em Buonaccorsi (2010).

O modelo definido em (1.3) e (1.4) é também conhecido como modelo de calibração comparativa, onde X_i corresponde a mensuração realizada por um instrumento padrão. Outros r instrumentos são utilizados para gerar o vetor resposta $\mathbf{Y}_i = (Y_{i1}, ..., Y_{ir})^\top$, ou seja, Y_{ij} é a i-ésima observação feita pelo j-ésimo instrumento.

Popularmente X_i é definida como variável substituta, enquanto que a variável x_i, que não é observada diretamente, é chamada de variável latente.

Dependendo da natureza de x_i, podemos considerar diferentes formas de modelos com erros nas variáveis. O primeiro caso é quando x_i são constantes desconhecidas, o segundo, quando os x_i são variáveis aleatórias independentes e identicamente distribuídas e o terceiro, quando os x_i são variáveis aleatórias com diferentes médias, porém com variâncias iguais. No primeiro caso, o modelo descrito pela equação (1.3) é dito *modelo funcional*. No segundo caso, o chamamos de *modelo estrutural*. O terceiro caso é denominado *modelo ultraestrutural* e foi proposto por Dolby (1976), que é uma generalização do modelo funcional e estrutural: quando consideramos que as médias da covariável x_i são iguais, o modelo se reduz ao modelo estrutural; quando assumimos que a variância da covariável x_i é nula, o modelo fica reduzido ao modelo funcional.

As diferenças entre os modelos funcional e estrutural estão bem esclarecidas em Ken-

dall (1951, 1952), no Capítulo 1 de Cheng & Van Ness (1999) e mais recentemente em Buonaccorsi (2010). Exemplos sobre modelos funcionais e estruturais podem ser encontrados em Fuller (1987). Neste trabalho daremos atenção ao modelo estrutural.

Seja $\boldsymbol{\varepsilon}_i = (\xi_i, \mathbf{e}_i^\top)^\top$ o vetor de erros de mensuração para o indivíduo i. Definindo $\mathbf{r}_i = (x_i, \boldsymbol{\varepsilon}_i^\top)^\top$, temos que a suposição a seguir é feita usualmente:

$$\mathbf{r}_i \overset{\text{iid}}{\sim} \mathrm{N}_{1+p}\left(\begin{pmatrix}\mu_x \\ \mathbf{0}_p\end{pmatrix}, \begin{pmatrix}\sigma_x^2 & \mathbf{0}_p^\top \\ \mathbf{0}_p & \boldsymbol{\Omega}\end{pmatrix}\right), \quad i = 1, \ldots, n, \tag{1.5}$$

onde $\mathbf{0}_p = (0, \ldots, 0)^\top : p \times 1$, $p = 1 + r$ é um vetor p-dimesional de zeros, $\boldsymbol{\Omega} = \text{diag}\{\omega_1^2, \ldots, \omega_p^2\}$ é uma matriz diagnonal de ordem $p \times p$ e a notação $\overset{\text{iid}}{\sim}$ denota vetores aleatórios independentes e identicamente distribuídos. Marginalmente, temos que $x_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu_x, \sigma_x^2)$ e $\boldsymbol{\varepsilon}_i \overset{\text{iid}}{\sim} \mathrm{N}_r(\mathbf{0}, \boldsymbol{\Omega})$ são independentes para todo $i = 1, \ldots, n$. O modelo definido pelas Equações (1.3), (1.4) e (1.5) será denominado *modelo de regressão normal multivariado com erros nas variáveis*, ou simplesmente N-MEM (normal measurement error model). Para mais detalhes ver, por exemplo, Fuller (1987, Seção 4.1).

Seja $\mathbf{Z}_i = (X_i, \mathbf{Y}_i^\top)^\top \overset{\text{def}}{=} (Z_{i1}, \ldots, Z_{ip})^\top$ o vetor de variáveis observáveis para o indivíduo i. As Equações (1.3) e (1.4) podem ser escritas de maneira unificada como

$$\begin{aligned}\mathbf{Z}_i &= \mathbf{a} + \mathbf{b}x_i + \boldsymbol{\varepsilon}_i \\ &= \mathbf{a} + \mathbf{B}\mathbf{r}_i,\ i = 1, \ldots, n,\end{aligned} \tag{1.6}$$

onde $\mathbf{a} = (0, \boldsymbol{\alpha}^\top)^\top$ e $\mathbf{b} = (1, \boldsymbol{\beta}^\top)^\top$ são vetores p- dimensionais e $\mathbf{B} = [\mathbf{b}; \mathbf{I}_p]$ é uma matriz de ordem $p \times (p+1)$ com $\mathbf{I}_p$ como sendo uma matriz identidade de ordem $p \times p$.. A partir de (1.6) em conjunto com (1.5), temos que

$$\mathbf{Z}_i \overset{\text{iid}}{\sim} \mathrm{N}_p(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \quad i = 1, \ldots, n,$$

onde

$$\boldsymbol{\mu}_z = \mathbf{a} + \mathbf{b}\mu_x \ \text{ e } \ \boldsymbol{\Sigma}_z = \sigma_x^2 \mathbf{b}\mathbf{b}^\top + \boldsymbol{\Omega}. \tag{1.7}$$

1.3 O Modelo de Regressão com Erro nas Variáveis Normal Multivariado com Ambas Variáveis Censuradas

Há situações em que as informações sobre as variáveis em estudo não estão completamente disponíveis para todas as unidades do conjunto de dados, tanto da variável resposta como da variável regressora, ou seja, apresentam padrão de censura. Isto pode fazer com que os resultados possam levar à inferências errôneas acerca do comportamento das variáveis sob investigação.

Seja um vetor $Z_i = (X_i, \mathbf{Y}_i)^\top$ definido como em (1.6). Para incorporar a possibilidade de observações censuradas, consideramos que Z_{ij} não é diretamente observável para todo $i = 1,2,\ldots,n$ e $j = 1,2,\ldots,p$. O que observamos efetivamente, para cada $i = 1,\ldots,n$, é o vetor aleatório $\mathbf{V}_i = (V_{i1},\ldots,V_{ip})^\top$, de tal forma que $V_{ij} = \text{máx}\{Z_{ij}, \kappa_{ij}\}$, onde κ_{ij} é um nível de censura, ou seja,

$$V_{ij} = \begin{cases} Z_{ij} & \text{se} \quad Z_{ij} > \kappa_{ij} \\ \kappa_{ij} & \text{se} \quad Z_{ij} \leq \kappa_{ij}. \end{cases} \tag{1.8}$$

Observe que por conveniência definição acima trata da ocorrência de censura à esquerda, vale ressaltar que estes resultados são facilmente estendíveis para outros tipos de censura. No contexto de regressão com erros nas variáveis, o modelo definido pelas Equações (1.3), (1.5) e (1.8) é denominado *modelo de regressão normal com erros nas variáveis, respostas multivariadas e censuras*, ou simplesmente N-MEMC.

1.4 Organização do trabalho

As etapas de desenvolvimento deste trabalho são resumidas a seguir:

1. Apresentação do modelo de regressão t de Student multivariado com erro nas variáveis e censuras;

2. Desenvolvimento de um algoritmo amostrador de Gibbs para estimação dos parâmetros do modelo de regressão com erro nas variáveis com ambas variáveis censuradas via inferência Bayesiana;

3. Estudos de simulação conduzidos com a finalidade de avaliar a consistência das estimativas dos parâmetros do modelo, analisar a performance do modelo t de Student

no ajuste para um conjunto de dados censurado e verificar o ganho em se trabalhar com uma distribuição de caudas pesadas, como a t de Student, quando um conjunto de dados com caudas mais pesadas é considerado.

4. Aplicação em dados reais.

Capítulo 2

Modelos de Regressão t de Student com Erros nas Variáveis, Respostas Multivariadas e Censuras.

No contexto de censuras, os modelos de regressão com erros de medida vem sendo objeto de estudo de muitos pesquisadores na literatura nas últimas décadas. Temos, por exemplo, o trabalho de Stapleton & Young (1984), que provaram a inconsistência dos estimadores de máxima verossimilhança para a classe dos modelos com erro de medição com censuras e esperança zero, estudando ainda a distribuição assintótica de estimadores baseados na função de expectativa (EF) ou a função de expectativa condicional (CEF) para os valores da variável dependente. Weiss (1993), que utilizando métodos numéricos de estimação, comparou o tradicional estimador de máxima verossimilhança dos parâmetros do modelo a um estimador construído baseado somente na mediana da distribuição das variáveis dependentes. Wang (1998), que propôs um modelo linear com erros nas variáveis onde a variável resposta é censurada, sugerindo um procedimento de duas etapas para estimar o modelo e a correspondente matriz de covariância assintótica. A estrutura abrange o modelo Tobit usual como um caso especial. Neste mesmo texto é mostrado que, sob normalidade e uma determinada condição de identificabilidade, é possível reduzir este modelo para um modelo de regressão censurada sem erros e, portanto, os estimadores existentes para o modelo Tobit podem ser usados para obter estimativas para este modelo, como o estimador de máxima verossimilhança, em particular. Existem alguns trabalhos na literatura que estendem

o N-MEM, com o objetivo de obter modelagem robusta, substituindo a hipótese de normalidade em (1.5) pela hipótese de que o vetor aleatório $\mathbf{r}_i$ tem distribuição t de Student como é o caso, por exemplo, de Bolfarine & Arellano-Valle (1994), Galea *et al.* (2002) e de Castro & Galea (2010), mas nenhum destes lida com respostas censuradas, além de considerarem somente o caso de respostas univariadas. O trabalho de Rocha *et al.* (2016) considera uma extensão Bayesiana usando a distribuição t de Student e respostas sujeitas à censura, mas não considera covariáveis censuradas, além de tratar somente do caso em que as respostas são univariadas. Recentemente, Matos *et al.* (2016) propuseram modelagem para Modelos de Regressão t de Student com Erros nas Variáveis, Respostas Multivariadas e Censuras, porém sob o enfoque frequentista.

2.1 Modelos de Regressão t de Student com Erros nas Variáveis, Respostas Multivariadas e Censuras (t-MEMC)

Neste capítulo propomos uma extensão robusta do N-MEMC. Especificamente, a suposição de normalidade em (1.5) é substituída por

$$\mathbf{r}_i \overset{\text{iid}}{\sim} \mathrm{t}_{1+p}\left(\begin{pmatrix}\mu_x\\ \mathbf{0}_p\end{pmatrix}, \begin{pmatrix}\sigma_x^2 & \mathbf{0}_p^\top\\ \mathbf{0}_p & \boldsymbol{\Omega}\end{pmatrix}, \nu\right),\ i=1,\ldots,n. \tag{2.1}$$

Pela equação (1.1), é possível perceber que $\mathbf{r}_i$ possui a seguinte representação hierárquica

$$\begin{pmatrix} x_i\\ \varepsilon_i\end{pmatrix}\Big| U_i=u_i \sim \mathrm{N}_{1+p}\left(\begin{pmatrix}\mu_x\\ \mathbf{0}_p\end{pmatrix}, u_i^{-1}\begin{pmatrix}\sigma_x^2 & \mathbf{0}_p^\top\\ \mathbf{0}_p & \boldsymbol{\Omega}\end{pmatrix}\right),$$
$$U_i \sim \text{Gamma}\left(\nu/2, \nu/2\right),\ i=1,\ldots,n.$$

Dado que $U_i = u_i$, temos que ε_i e x_i são independentes, então os seguintes condicio-

namentos podem ser facilmente verificados

$$x_i|U_i = u_i \overset{\text{ind}}{\sim} \mathrm{N}(\mu_x, u_i^{-1}\sigma_x^2) \ \ \mathbf{e} \tag{2.2}$$

$$\varepsilon_i|U_i = u_i \overset{\text{ind}}{\sim} \mathrm{N}_p(\mathbf{0}_p, u_i^{-1}\boldsymbol{\Omega}). \tag{2.3}$$

Marginalmente, temos que $\varepsilon_i \sim \mathrm{t}_p(\mathbf{0}, \boldsymbol{\Omega}, \nu)$ e $x_i \sim \mathrm{t}_1(\mu_x, \sigma_x^2, \nu)$. Ao contrário do que ocorre no modelo normal, ε_i e x_i, por serem indexadas pelo mesmo fator de escala U_i, não são independentes em geral (a menos que $U_i = 1$, ou seja, o caso normal). No entanto, ε_i e x_i são não correlacionadas, pois

$$\begin{aligned}\mathrm{Cov}(\varepsilon_i, x_i) &= \mathrm{E}\{\mathrm{E}[\varepsilon_i x_i|U_i]\} \\ &= \mathrm{E}\{\mathrm{E}[\varepsilon_i|U_i]\mathrm{E}[x_i|U_i]\} = \mathbf{0}.\end{aligned}$$

Usando a relação (1.3) e propriedades usuais da distribuição t de Student multivariada, temos que a distribuição marginal de $\mathbf{Z}_i$ é dada por

$$\mathbf{Z}_i \sim \mathrm{t}_p(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}, \nu), \quad i = 1, \ldots, n, \tag{2.4}$$

onde $\boldsymbol{\mu}_{z_i}$ e $\boldsymbol{\Sigma}_{z_i}$ são dados em (1.7).

Assim como no caso normal, para incorporarmos a possibilidade de observações censuradas ao modelo, assumimos que as variáveis Z_{ij}, para algumas unidades do conjunto de dados, não são observadas completamente. O que observamos efetivamente, para cada $i = 1, \ldots, n$, é o vetor aleatório $\mathbf{V}_i = (V_{i1}, \ldots, V_{ip})^\top$, tal que $V_{ij} = \text{máx}\{Z_{ij}, \kappa_{ij}\}$, onde κ_{ij} é um nível de censura, como apresentado na equação (1.8). O modelo definido pelas Equações (1.3), (1.4),(2.1) e (1.8) é denominado *modelo de regressão t de Student com erros nas variáveis, respostas multivariadas e censuras*, ou simplesmente t-MEMC.

2.1.1 A Função de Verossimilhança

Nesta seção obteremos a função de verossimilhança associada à uma amostra proveniente do t-MEMC. A expressão desta função será importante para gerarmos amostras da distribuição a posteriori do parâmetro de graus de liberdade e também para calcularmos os critérios de seleção de modelos, a fim de podermos comparar os diferentes modelos ajusta-

dos.

Para cada unidade experimental i, sabemos exatamente quais são os índices $j \in \{1,\ldots,p\}$ correspondentes aos elementos não censurados ou censurados do vetor $\mathbf{Z}_i = (Z_{i1},\ldots,Z_{ip})$. Assim, vamos particionar o vetor $\mathbf{Z}_i$ em dois subvetores: $\mathbf{Z}_i^o : p_i^o \times 1$, que é o vetor correspondente ao primeiro caso e $\mathbf{Z}_i^c : p_i^c \times 1$, que é o vetor correspondente ao segundo caso. Vamos escrever, sem perda de generalidade, $\mathbf{Z}_i = (\mathbf{Z}_i^{o\top}, \mathbf{Z}_i^{c\top})^\top$ e, de maneira conforme, vamos considerar $\mathbf{V}_i = (\mathbf{V}_i^{o\top}, \mathbf{V}_i^{c\top})^\top$ e, lembrando que $\mathbf{Z}_i \sim t_p(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}, \nu)$, $\boldsymbol{\mu}_{z_i} = (\boldsymbol{\mu}_{z_i}^{o\top}, \boldsymbol{\mu}_{z_i}^{c\top})^\top$ e a partição em blocos $\boldsymbol{\Sigma}_{z_i} = \left(\begin{smallmatrix} \boldsymbol{\Sigma}_{z_i}^{oo} & \boldsymbol{\Sigma}_{z_i}^{oc} \\ \boldsymbol{\Sigma}_{z_i}^{co} & \boldsymbol{\Sigma}_{z_i}^{cc} \end{smallmatrix}\right)$. $\boldsymbol{\kappa}_i^c$ denotará o vetor com os níveis de censura correspondentes a $\mathbf{Z}_i^c$. Pela Proposição 1, temos que

$$\mathbf{Z}_i^o \sim \mathrm{t}_{p_i^o}(\boldsymbol{\mu}_{z_i}^o, \boldsymbol{\Sigma}_{z_i}^{oo}, \nu) \;\text{ e }\; \mathbf{Z}_i^c|\mathbf{Z}_i^o \sim \mathrm{t}_{p_i^c}(\boldsymbol{\mu}_{z_i}^{co}, \mathbf{S}_{z_i}^{co}, \nu + p_i^o), \tag{2.5}$$

onde

$$\boldsymbol{\mu}_{z_i}^{co} = \boldsymbol{\mu}_{z_i}^c + \boldsymbol{\Sigma}_{z_i}^{co}(\boldsymbol{\Sigma}_{z_i}^{oo})^{-1}(\mathbf{Z}_i^o - \boldsymbol{\mu}_{z_i}^o), \tag{2.6}$$

$$\mathbf{S}_{z_i}^{co} = \left(\frac{\nu + \mathrm{d}_{\boldsymbol{\Sigma}_{z_i}^{oo}}(\mathbf{Z}_i^o, \boldsymbol{\mu}_{z_i}^o)}{\nu + p^o}\right)\boldsymbol{\Sigma}_{z_i}^{cc.o}, \tag{2.7}$$

$$\boldsymbol{\Sigma}_{z_i}^{cc.o} = \boldsymbol{\Sigma}_{z_i}^{cc} - \boldsymbol{\Sigma}_{z_i}^{co}\boldsymbol{\Sigma}_{z_i}^{oo-1}\boldsymbol{\Sigma}_{z_i}^{oc}. \tag{2.8}$$

Temos que $\{\mathbf{z}_i^o, \boldsymbol{\kappa}_i^c\}$ é a amostra observada para a unidade experimental i. A verossimilhança associada é dada por

$$L_i(\boldsymbol{\theta}) = P(\mathbf{V}_i^c = \boldsymbol{\kappa}_i^c|\mathbf{Z}_i^o = \mathbf{z}_i^o)f(\mathbf{z}_i^o),$$

onde $f(\cdot)$ é a densidade marginal de $\mathbf{Z}_i^o$. Observe que o lado direito da igualdade acima é a densidade conjunta dos vetores $\mathbf{V}_i^c$ e $\mathbf{Z}_i^o$. Ocorre que $\mathbf{V}_i^c = \boldsymbol{\kappa}_i^c$ se, e somente se, $\mathbf{Z}_i^c \leq \boldsymbol{\kappa}_i^c$. Por (2.5), vem que

$$L_i(\boldsymbol{\theta}) = \mathrm{T}_{p_i^c}(\boldsymbol{\kappa}_i^c|\boldsymbol{\mu}_{z_i}^{co}, \mathbf{S}_{z_i}^{co}, \nu + p_i^o)\mathrm{t}_{p_i^o}(\mathbf{z}_i^o|\boldsymbol{\mu}_{z_i}^o, \boldsymbol{\Sigma}_{z_i}^{oo}, \nu). \tag{2.9}$$

Calculando a verossimilhança $L_i(\boldsymbol{\theta})$ para cada unidade experimental i, $i = 1,\ldots,n$, obtemos a log-verossimilhança associada aos dados observados, dada por

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta}).$$

Capítulo 3

Estimação via MCMC

Nesta seção apresentaremos um algoritmo do tipo MCMC. A base para a construção de nosso algoritmo é a seguinte representação em dados aumentados do modelo, obtida a partir da equação (1.6) e da representação estocástica dada em (2.2) e (2.3):

$$\mathbf{Z}_i \mid x_i, U_i = u_i \overset{\text{ind}}{\sim} \mathrm{N}_p(\mathbf{a} + \mathbf{b}x_i, u_i^{-1}\boldsymbol{\Omega}); \tag{3.1}$$

$$x_i \mid U_i = u_i \overset{\text{ind}}{\sim} \mathrm{N}(\mu_x, u_i^{-1}\sigma_x^2); \tag{3.2}$$

$$U_i \overset{\text{iid}}{\sim} \text{Gamma}\left(\nu/2, \nu/2\right),\ i = 1, \ldots, n. \tag{3.3}$$

3.1 Distribuições a Priori

No contexto Bayesiano, é necessário especificarmos as distribuições a priori a fim de realizarmos inferência para o vetor com todos os parâmetros a serem estimados $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\omega}^\top, \mu_x, \sigma_x^2, \nu)^\top$, onde $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_r)^\top$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_r)^\top$ e $\boldsymbol{\omega} = (\omega_1^2, \omega_1^2, \ldots, \omega_p^2)^\top$.

Uma distribuição a priori para os parâmetros contidos em $\boldsymbol{\theta}$ deverá quantificar em termos de probabilidade a incerteza a respeito vetor de parâmetros sob investigação, baseado em um conhecimento prévio a respeito do problema. Neste trabalho, usaremos distribuições a priori não informativas. Esta escolha está diretamente ligada à classe que cada parâmetro pertence, ou seja, se o parâmetro é de locação, escala ou forma. Definimos então a seguinte

especificação a priori:

$$\boldsymbol{\alpha} \sim \mathrm{N}_r(\mathbf{c}_\alpha, \mathbf{D}_\alpha),\ \boldsymbol{\beta} \sim \mathrm{N}_r(\mathbf{c}_\beta, \mathbf{D}_\beta),\ \tau_j^2 = \omega_j^{-2} \sim \mathrm{Gamma}(l,m),\ \mu_x \sim \mathrm{N}(c_x, d_x^2),$$
$$\gamma_x^2 = \sigma_x^{-2} \sim \mathrm{Gamma}(e,f),\ \nu \sim \exp(1/\lambda)\ \text{ e } \lambda \sim \mathrm{Unif}(\lambda_0, \lambda_1),$$

onde os hiperparâmetros c_x, d_x^2, l, m, e, f, λ_0, λ_1, $\mathbf{c}_\alpha$, $\mathbf{c}_\beta$, $\mathbf{D}_\alpha$ e $\mathbf{D}_\beta$, são conhecidos – assumindo que as matrizes de covariância são diagonais e positivas definidas. Esta especificação é usual em estudos envolvendo a distribuição t de Student, ver por exemplo Rocha *et al.* (2016). A especificação a priori para ν foi adotada com sucesso em outros trabalhos como, por exemplo, Cabral *et al.* (2012). Para um estudo de sensibilidade comparando diferentes distribuições a priori para ν em um modelo de regressão com erros de observação t de Student, veja Garay *et al.* (2015).

Assim, supondo que os parâmetros de $\boldsymbol{\theta}$ são independentes entre si, a especificação a priori completa é dada por

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\mu_x)\pi(\nu|\lambda)\pi(\lambda)\left(\prod_{j=1}^{p}\omega_j^2\right)\pi(\sigma_j^2).$$

3.2 Um Algoritmo do Tipo Gibbs

Nesta seção vamos propor um algoritmo do tipo Gibbs para estimação Bayesiana que consiste de um amostrador marginal dos parâmetros a partir das distribuições condicionais completas calculadas usando apenas a representação estocástica do modelo definidas nas Equações (3.1), (3.2) e (3.3); isto significa que os parâmetros são atualizados um de cada vez. Mais detalhes sobre o amostrador de Gibbs e referências envolvendo resultados teóricos mais gerais sobre algoritmos do tipo MCMC podem ser encontrados em Gelfand *et al.* (1992) e Gamerman & Lopes (2006)

Antes de apresentarmos o algoritmo, vamos apresentar alguns resultados que serão necessários para a sua implementação. No que segue, dados dois vetores aleatórios $\mathbf{X}_1$ e $\mathbf{X}_2$, denotaremos a densidade condicional de $\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2$ por $\pi(\mathbf{x}_1|\mathbf{x}_2)$ o que, embora sendo um abuso de notação, simplifica os cálculos. Também vamos definir $\mathbf{x} = \{x_1, \ldots, x_n\}$, $\mathbf{u} = \{u_1, \ldots, u_n\}$ e $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$.

3.2.1 Detalhes do Algoritmo

Para obtermos uma amostra da distribuição a posteriori do modelo t-MEMC basta seguirmos os passos do seguinte algoritmo:

Passo 1. Como definido na Seção 2.1.1, para cada $i = 1, \ldots, n$, vamos considerar a partição de $\mathbf{Z}_i$ em dois subvetores: $\mathbf{Z}_i^o : p_i^o \times 1$, como sendo o vetor correspondente aos valores efetivamente observados e $\mathbf{Z}_i^c : p_i^c \times 1$, o vetor correspondente aos valores censurados. Neste caso, $\mathbf{Z}_i^c$ é uma variável latente, e atribuiremos a este vetor uma distribuição condicional completa. Pela equação (2.4), temos que

$$\begin{aligned} \mathbf{Z}_i | U_i = u_i &\sim \mathrm{N}_p(\boldsymbol{\mu}_z, u_i^{-1}\boldsymbol{\Sigma}_z); \\ U_i &\sim \mathrm{Gamma}(\nu/2, \nu/2). \end{aligned}$$

Usando a partição $\mathbf{Z}_i = (\mathbf{Z}_i^{o\top}, \mathbf{Z}_i^{c\top})^\top$ e a conhecida propriedade de que a distribuição normal é fechada para condicionamentos, temos que

$$\mathbf{Z}_i^c | \mathbf{Z}_i^o = \mathbf{z}_i^o, U_i = u_i \sim \mathrm{N}_{p_i^c}(\boldsymbol{\mu}_{z_i}^{co}, u_i^{-1}\boldsymbol{\Sigma}_{z_i}^{cc.o}),$$

com $\boldsymbol{\mu}_{z_i}^{co}$ e $\boldsymbol{\Sigma}_{z_i}^{cc.o}$ definidos pelas Equações (2.6) e (2.8). Para geramos valores a partir da distribuição condicional completa de $\mathbf{Z}_i^c$, devemos observar que se $\mathbf{V}_i = \kappa_i$, então $\mathbf{Z}_i^c \leq \kappa_i$, deste modo

$$(\mathbf{Z}_i^c | \mathbf{V}_i = (\kappa_i, \mathbf{z}_i^o), U_i = u_i) \overset{d}{=} (\mathbf{Z}_i^c | \mathbf{Z}_i^c \leq \kappa_i, \mathbf{Z}_i^o = \mathbf{z}_i^o, U_i = u_i) \sim \mathrm{TN}_{p_i^c}(\boldsymbol{\mu}_{z_i}^{co}, u_i^{-1}\boldsymbol{\Sigma}_{z_i}^{cc.o}; \mathbb{D}_i^c) \tag{3.4}$$

onde $\overset{d}{=}$ significa "tem a mesma distribuição que" e

$$\mathbb{D}_i^c = \prod_{i \in \mathscr{C}} (-\infty, \kappa_{ij}],$$

onde $\mathscr{C}$ é o conjunto de índices das componentes censuradas. Note que para cada $i = 1, 2, \ldots, n$ a dimensão da distribuição em (3.4) varia de acordo com o número p_i^c de componentes censuradas em $\mathbf{Z}_i$, ou seja, de acordo com a dimensão de $\mathbf{Z}_i^c$.

Seja $\mathbf{Z}_i^{c*}$ a amostra gerada neste passo do algoritmo, com $i = 1,\ldots,n$. Para dar continuidade ao processo, completamos o vetor de amostras para o indivíduo i, imputando $\mathbf{Z}_i^{c*}$ no lugar de $\mathbf{Z}_i^c$, obtendo o vetor completo $\mathbf{Z}_i = (\mathbf{Z}_i^{o\top}, \mathbf{Z}_i^{c*\top})^\top$, de modo que a partir deste passo, utilizaremos o novo vetor $\mathbf{Z}_i$ para encontrarmos as demais distribuições condicionais completas.

Passo 2. Para $i = 1,2,\ldots,n$, geramos u_i independentemente a partir da distribuição de $\pi(u_i|\mathbf{Z}_i, x_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \nu)$. Através da representação estocástica do t-MEMC pelas Equações (3.1) e (3.3), é possível observar que

$$\begin{aligned}
\pi(u_i|\mathbf{Z}_i, x_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \mu_x, \sigma_x^2, \nu) &\propto \pi(\mathbf{Z}_i|u_i, x_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \nu)\pi(x_i|u_i, \mu_x, \sigma_x^2)\pi(u_i|\nu) \\
&\propto |u_i^{-1}\boldsymbol{\Omega}|^{-1/2} \exp\left\{-\frac{u_i}{2}(\mathbf{z}_i - (\mathbf{a}+\mathbf{b}x_i))^\top \boldsymbol{\Omega}^{-1}(\mathbf{z}_i - (\mathbf{a}+\mathbf{b}x_i))\right\} \\
&\quad \times u_i^{\frac{\nu+1}{2}-1} \exp\left\{-\frac{u_i}{2\sigma_x^2}(x_i-\mu_x)^2\right\} \exp\left\{-\frac{\nu}{2}u_i\right\} \\
&\propto u_i^{\frac{p+\nu+1}{2}-1} \exp\left\{-\frac{u_i}{2}(\mathbf{z}_i - (\mathbf{a}+\mathbf{b}x_i))^\top \boldsymbol{\Omega}^{-1}(\mathbf{z}_i - (\mathbf{a}+\mathbf{b}x_i)) - \frac{u_i}{2}\nu \right. \\
&\quad \left. -\frac{u_i}{2\sigma_x^2}(x_i-\mu_x)^2\right\} \\
&\propto u_i^{\frac{p+\nu+1}{2}-1} \exp\left\{-\frac{u_i}{2}\left((\mathbf{z}_i - (\mathbf{a}+\mathbf{b}x_i))^\top \boldsymbol{\Omega}^{-1}(\mathbf{z}_i - (\mathbf{a}+\mathbf{b}x_i)) \right.\right. \\
&\quad \left.\left. + \frac{(x_i-\mu_x)^2}{\sigma_x^2} + \nu\right)\right\}.
\end{aligned}$$

Então temos que $u_i|\mathbf{Z}_i, x_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \nu \sim \text{Gamma}\left(\frac{p+\nu+1}{2}, \frac{1}{2}\left[\text{d}_{\boldsymbol{\Omega}}(\mathbf{z}_i, \mathbf{a}+\mathbf{b}x_i) + \frac{(x_i-\mu_x)^2}{\sigma_x^2} + \nu\right]\right)$.

Passo 3. Para $i = 1,2,\ldots,n$ geramos x_i independentemente de $\pi(x_i|\mathbf{Z}_i, u_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \mu_x, \sigma_x^2, \nu)$. Atráves das Equações (3.1) e (3.2) do t-MEMC, temos que esta distribuição condicional completa de variável latente x_i é encontrada da seguinte forma

$$\begin{aligned}
\pi(x_i|\mathbf{Z}_i, u_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \mu_x, \sigma_x^2, \nu) &\propto \pi(\mathbf{Z}_i|u_i, x_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \mu_x, \sigma_x^2, \nu)\pi(x_i|u_i, \mu_x, \sigma_x^2) \\
&\propto \exp\left\{-\frac{u_i}{2}(\mathbf{z}_i - (\mathbf{a}+\mathbf{b}x_i))^\top \boldsymbol{\Omega}^{-1}(\mathbf{z}_i - (\mathbf{a}+\mathbf{b}x_i))\right\} \\
&\quad \times \exp\left\{-\frac{u_i}{2\sigma_x^2}(x_i-\mu_x)^2\right\} \\
&\propto \exp\left\{-\frac{u_i}{2}\left[-2\mathbf{z}_i^\top \boldsymbol{\Omega}^{-1}(\mathbf{a}+\mathbf{b}x_i) + (\mathbf{a}+\mathbf{b}x_i)\boldsymbol{\Omega}^{-1}(\mathbf{a}+\mathbf{b}x_i)\right]\right\} \\
&\quad \times \exp\left\{-\frac{u_i}{2\sigma_x^2}(x_i^2 - 2x_i\mu_x)\right\},
\end{aligned}$$

implicando em

$$\begin{aligned}\pi(x_i|\mathbf{Z}_i,u_i,\boldsymbol{\alpha},\boldsymbol{\beta},\omega,\mu_x,\sigma_x^2,\nu) &\propto \exp\left\{-\frac{u_i}{2}\left[-2x_i\mathbf{b}^\top\boldsymbol{\Omega}^{-1}\mathbf{z}_i+2x_i\mathbf{b}^\top\boldsymbol{\Omega}^{-1}\mathbf{a}+x_i^2\mathbf{b}^\top\boldsymbol{\Omega}^{-1}\mathbf{b}\right]\right\}\\ &\quad\times\exp\left\{-\frac{u_i}{2\sigma_x^2}(x_i^2-2x_i\mu_x)\right\}\\ &\propto \exp\left\{-\frac{u_i}{2}\left[-2x_i\left(\mathbf{b}^\top\boldsymbol{\Omega}^{-1}\mathbf{z}_i-\mathbf{b}^\top\boldsymbol{\Omega}^{-1}\mathbf{a}\right)+x_i^2\mathbf{b}^\top\boldsymbol{\Omega}^{-1}\mathbf{b}\right]\right.\\ &\quad\left.-\frac{1}{\sigma_x^2}(x_i^2-2x_i\mu_x)\right\}\\ &\propto \exp\left\{-\frac{u_i}{2}\left[x_i^2\left(\frac{1+\sigma_x^2\mathbf{b}^\top\boldsymbol{\Omega}^{-1}\mathbf{b}}{\sigma_x^2}\right)\right.\right.\\ &\quad\left.\left.-2x_i\left(\mathbf{b}^\top\boldsymbol{\Omega}^{-1}(\mathbf{z}_i-\mathbf{a})+\frac{\mu_x}{\sigma_x^2}\right)\right]\right\}.\end{aligned}$$

Portanto, $x_i|\mathbf{Z}_i,u_i,\boldsymbol{\alpha},\boldsymbol{\beta},\omega,\mu_x,\sigma_x^2,\nu \sim \mathrm{N}\left(\xi_{x_i},u_i^{-1}\delta_{x_i}^2\right)$, onde

$$\delta_{x_i}^2=\left(\frac{1+\sigma_x^2\mathbf{b}^\top\boldsymbol{\Omega}^{-1}\mathbf{b}}{\sigma_x^2}\right)^{-1} \quad \text{e} \quad \xi_{x_i}=\delta_{x_i}^2\left(\mathbf{b}^\top\boldsymbol{\Omega}^{-1}(\mathbf{z}_i-\mathbf{a})+\frac{\mu_x}{\sigma_x^2}\right).$$

Passo 4. Gere μ_x a partir da distribuição de $\pi(\mu_x|\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\omega,\sigma_x^2,\nu)$. Para encontrarmos essa distribuição, devemos utilizar o segundo nível da representação estocástica do modelo t-MEMC dado pela equação (3.2), de modo que

$$\begin{aligned}\pi(\mu_x|\mathbf{u},\mathbf{x},\sigma_x^2,\nu) &\propto \pi(\mathbf{x}|\mathbf{u},\mu_x,\sigma_x^2,\nu)\pi(\mu_x)\\ &\propto \exp\left\{-\sum_{i=1}^n\frac{u_i}{2\sigma_x^2}(x_i-\mu_x)^2\right\}\exp\left\{-\frac{1}{2d_x^2}(\mu_x-c_x)^2\right\}\\ &= \exp\left\{-\frac{1}{2}\sum_{i=1}^n\left[\frac{u_i}{\sigma_x^2}(x_i-2\mu_x x_i+\mu_x^2)\right]-\frac{1}{2d_x^2}(\mu_x^2-2\mu_x c_x+c_x^2)\right\}\\ &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n\frac{u_i}{\sigma_x^2}(-2\mu_x x_i+\mu_x^2)+\frac{1}{d_x^2}(\mu_x^2-2\mu_x c_x)\right]\right\}\\ &= \exp\left\{-\frac{1}{2}\left[\mu_x^2\left(\sigma_x^{-2}\sum_{i=1}^n u_i+d_x^{-2}\right)-2\mu_x\left(\sigma_x^{-2}\sum_{i=1}^n u_i x_i+d_x^{-2}c_x\right)\right]\right\}.\end{aligned}$$

Assim, temos que $\mu_x|\mathbf{u},\mathbf{x},\sigma_x^2,\nu \sim \mathrm{N}(\xi_{\mu_x},\delta_{\mu_x}^2)$, onde

$$\delta_{\mu_x}^2=\left(\sigma_x^{-2}\sum_{i=1}^n u_i+d_x^{-2}\right) \quad \text{e} \quad \xi_{\mu_x}=\delta_{\mu_x}^2\left(\sigma_x^{-2}\sum_{i=1}^n u_i x_i+d_x^{-2}c_x\right).$$

Passo 5. Gere $\gamma_x^2 = \sigma_x^{-2}$ a partir da distribuição de $\pi(\gamma_x^2|\mathbf{u},\mathbf{x},\mu_x)$. Utilizando a distribuição apresentada na equação (3.2) da representação estocástica do modelo t-MEMC , bem como a distribuição a priori para γ_x^2, temos que

$$\begin{aligned}
\pi(\gamma_x^2|\mathbf{u},\mathbf{x},\mu_x) &\propto \pi(\mathbf{x}|u_i,\mu_x,\sigma_x^2,\nu)\pi(\gamma_x^2) \\
&\propto (\gamma_x^{-2})^{-n/2}\exp\left\{-\sum_{i=1}^{n}\frac{u_i\gamma_x^2}{2}(x_i-\mu_x)^2\right\}\frac{f^e}{\Gamma(e)}\left(\gamma_x^2\right)^{e-1}\exp\left\{-f\left(\gamma_x^2\right)\right\} \\
&\propto (\gamma_x^2)^{n/2}\exp\left\{-\gamma_x^2\sum_{i=1}^{n}\frac{u_i}{2}(x_i-\mu_x)^2\right\}(\gamma_x^2)^{e-1}\exp\left\{-f\gamma_x^2\right\} \\
&\propto (\gamma_x^2)^{n/2+e-1}\exp\left\{-\gamma_x^2\left[\sum_{i=1}^{n}\frac{u_i}{2}(x_i-\mu_x)^2+f\right]\right\}.
\end{aligned}$$

Logo, temos que

$$\gamma_x^2|\mathbf{u},\mathbf{x},\mu_x \sim \text{IG}\left(e+\frac{n}{2}, f+\frac{1}{2}\sum_{i=1}^{n}u_i(x_i-\mu_x)^2\right).$$

Passo 6. Para $j = 1,2,\ldots,p$ amostraremos $\tau_j^2 = \omega_j^{-2}$ a partir da distribuição condicional completa de $\tau_j^2|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\mu_x,\sigma_x^2,\nu$. Com a distribuição apresentada na equação (3.1) e a distribuição priori atribuída para τ_j^2, temos

$$\begin{aligned}
\pi(\tau_j^2|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}) &\propto \pi(\mathbf{Z}|\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta})\pi(\tau_j^2) \\
&\propto |\Omega|^{-1/2}\prod_{i=1}^{n}\left[\exp\left\{-\frac{u_i}{2}(\mathbf{z}_i-\mathbf{a}-\mathbf{b}x_i)^\top\Omega^{-1}(\mathbf{z}_i-\mathbf{a}-\mathbf{b}x_i)\right\}\right] \\
&\quad\times(\tau_j^2)^{l-1}\exp\{-m\tau_j^2\} \\
&\propto (\tau_j^2)^{n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[u_i(\mathbf{z}_i-\mathbf{a}-\mathbf{b}x_i)^\top\Omega^{-1}(\mathbf{z}_i-\mathbf{a}-\mathbf{b}x_i)\right]\right\} \\
&\quad\times(\tau_j^2)^{l-1}\exp\{-m\tau_j^2\}.
\end{aligned}$$

Para o i-ésimo indivíduo e $j \in \{1,2,\ldots,p\}$, vamos denotar a j-ésima coordenada dos vetores

$\mathbf{z}_i$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, por z_{ij}, α_j e β_j, respectivamente, obtendo a seguinte igualdade

$$\begin{aligned}(\mathbf{z}_i-\mathbf{a}-\mathbf{b}x_i)^\top\Omega^{-1}(\mathbf{z}_i-\mathbf{a}-\mathbf{b}x_i) &= (X_i-x_i, Y_{i1}-\alpha_1-\beta_1x_i,\ldots,Y_{ip}-\alpha_p-\beta_px_i)^\top\\ &\quad\times \mathrm{diag}\{\tau_1^2,\ldots,\tau_p^2\}\times(X_i-x_i,Y_{i1}-\alpha_1-\beta_1x_i,\\ &\quad Y_{i2}-\alpha_2-\beta_2x_i,\ldots,Y_{ip}-\alpha_p-\beta_px_i)\\ &= \tau_1^2(X_i-x_i)^2+\tau_2^2(Y_{i1}-\alpha_1-\beta_1x_i)^2\\ &\quad+\tau_3^2(Y_{i2}-\alpha_2-\beta_2x_i)^2+\ldots+\tau_p^2(Y_{ir}-\alpha_r-\beta_rx_i)^2,\end{aligned}$$

de tal modo que

$$\begin{aligned}\pi(\tau_j^2|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}) &\propto (\tau_j^2)^{n/2}\exp\left\{\tau_1^2\sum_{i=1}^n u_i(X_i-x_i)^2+\sum_{j=2}^p\tau_j^2\sum_{i=1}^n u_i(Y_{i(j-1)}-\alpha_{j-1}+\beta_{j-1}x_i)^2\right\}\\ &\quad\times(\tau_j^2)^{l-1}\exp\{-m\tau_j^2\}.\end{aligned}$$

Então, para $j=1$, temos que

$$\begin{aligned}\pi(\tau_1^2|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}) &\propto (\tau_1^2)^{n/2}\exp\left\{-\tau_1^2\left[\frac{1}{2}\sum_{i=1}^n u_i(X_i-x_i)^2\right]\right\}\times(\tau_1^2)^{l-1}\exp\{-m\tau_1^2\}\\ &\propto (\tau_1^2)^{\frac{n}{2}+l-1}\exp\left\{-\tau_1^2\left[\frac{1}{2}\sum_{i=1}^n u_i(X_i-x_i)^2+m\right]\right\},\end{aligned}$$

portanto,

$$\tau_1^2|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}\sim\mathrm{IG}\left(\frac{n}{2}+l, m+\frac{1}{2}\sum_{i=1}^n u_i(X_i-x_i)^2\right).$$

Para $j=2,3,\ldots,p$, a distribuição de τ_j^2 é encontrada de modo análogo:

$$\begin{aligned}\pi(\tau_j^2|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}) &\propto (\tau_j^2)^{n/2}\exp\left\{-\tau_j^2\left[\frac{1}{2}\sum_{i=1}^n u_i(Y_{i(j-1)}-\alpha_j-\beta_jx_i)^2\right]\right\}\times(\tau_j^2)^{l-1}\exp\{-m\tau_j^2\}\\ &\propto (\tau_j^2)^{\frac{n}{2}+l-1}\exp\left\{-\tau_j^2\left[\frac{1}{2}\sum_{i=1}^n u_i(Y_{i(j-1)}-\alpha_j-\beta_jx_i)^2+m\right]\right\}.\end{aligned}$$

Assim, temos que

$$\tau_j^2|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}\sim\mathrm{IG}\left(\frac{n}{2}+l, m+\frac{1}{2}\sum_{i=1}^n u_i(Y_{i(j-1)}-\alpha_j-\beta_jx_i)^2\right).$$

Passo 7. Gere $\boldsymbol{\alpha}$ a partir da distribuição de $\boldsymbol{\alpha}|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\beta},\omega$. Com a distribuição condicional definida na equação (3.1) e a priori atribuída para $\boldsymbol{\alpha}$, temos que

$$\begin{aligned}
\pi(\boldsymbol{\alpha}|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\beta},\omega) &\propto \pi(\mathbf{Z}|\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\omega)\pi(\boldsymbol{\alpha}) \\
&\propto \prod_{i=1}^{n}\left[\exp\left\{-\frac{u_i}{2}(\mathbf{z}_i-(\mathbf{a}-\mathbf{b}x_i))^\top\Omega^{-1}(\mathbf{z}_i-(\mathbf{a}-\mathbf{b}x_i))\right\}\right] \\
&\quad\times\exp\left\{-\frac{1}{2}(\boldsymbol{\alpha}-\mathbf{c}_\alpha)^\top\mathbf{D}_\alpha^{-1}(\boldsymbol{\alpha}-\mathbf{c}_\alpha)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(-2u_i\mathbf{a}^\top\Omega^{-1}(\mathbf{z}_i-\mathbf{b}x_i)+u_i\mathbf{a}^\top\Omega^{-1}\mathbf{a}\right)\right\} \qquad (3.5)\\
&\quad\times\exp\left\{-\frac{1}{2}(\boldsymbol{\alpha}^\top\mathbf{D}_\alpha^{-1}\boldsymbol{\alpha}-2\boldsymbol{\alpha}^\top\mathbf{D}_\alpha^{-1}\mathbf{c}_\alpha)\right\}.
\end{aligned}$$

Sabendo que $\mathbf{Z}_i=(X_i,\mathbf{Y}_i)$, $\mathbf{a}=(0,\boldsymbol{\alpha}^\top)^\top$ e $\mathbf{b}=(1,\boldsymbol{\beta}^\top)^\top$, então temos que a expressão no somatório na equação (3.5) pode ser escrita da seguinte forma:

$$\begin{aligned}
\left(-2u_i\mathbf{a}^\top\Omega^{-1}(\mathbf{z}_i-\mathbf{b}x_i)+u_i\mathbf{a}^\top\Omega^{-1}\mathbf{a}\right) &= \big(-2u_i\left[0\times\omega_1^{-2}(X_i-x_i)+\alpha_1\omega_2^{-2}(Y_{i1}-\beta_1x_i)\right. \\
&\quad\left.+\alpha_2\omega_3^{-2}(Y_{i2}-\beta_2x_i)+\ldots+\alpha_r\omega_p^{-2}(Y_{ir}-\beta_rx_i)\right] \\
&\quad+u_i\boldsymbol{\alpha}^\top\Omega_\star^{-1}\boldsymbol{\alpha}\big) \\
&= \left(-2u_i\boldsymbol{\alpha}^\top\Omega_\star^{-1}(\mathbf{Y}_i-\boldsymbol{\beta}x_i)+u_i\boldsymbol{\alpha}^\top\Omega_\star^{-1}\boldsymbol{\alpha}\right).
\end{aligned}$$

Assim, temos que

$$\begin{aligned}
\pi(\boldsymbol{\alpha}|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\beta},\omega) &\propto \exp\left\{-\frac{1}{2}\left[-2\left(\sum_{i=1}^{n}u_i\boldsymbol{\alpha}^\top\Omega_\star^{-1}(\mathbf{Y}_i-\boldsymbol{\beta}x_i)+\boldsymbol{\alpha}^\top\mathbf{D}_\alpha^{-1}\mathbf{c}_\alpha\right)\right.\right. \\
&\quad\left.\left.+\left(\sum_{i=1}^{n}u_i\boldsymbol{\alpha}^\top\Omega_\star^{-1}\boldsymbol{\alpha}\right)+\boldsymbol{\alpha}^\top\mathbf{D}_\alpha^{-1}\boldsymbol{\alpha}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\alpha}^\top\left(\sum_{i=1}^{n}u_i\Omega_\star^{-1}+\mathbf{D}_\alpha^{-1}\right)\boldsymbol{\alpha}\right.\right. \\
&\quad\left.\left.-2\boldsymbol{\alpha}^\top\left(\sum_{i=1}^{n}u_i\Omega_\star^{-1}(\mathbf{Y}_i-\boldsymbol{\beta}x_i)+\mathbf{D}_\alpha^{-1}\mathbf{c}_\alpha\right)\right]\right\}.
\end{aligned}$$

Logo, $\boldsymbol{\alpha}|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\beta},\boldsymbol{\omega} \sim \mathrm{N}_r(\zeta_\alpha,\Delta_\alpha)$, onde

$$\Delta_\alpha = \left(\sum_{i=1}^{n} u_i \Omega_\star^{-1} + \mathbf{D}_\alpha^{-1}\right)^{-1} \qquad \text{e}$$
$$\zeta_\alpha = \Delta_\alpha \left(\sum_{i=1}^{n} u_i \Omega_\star^{-1}(\mathbf{Y}_i - \beta x_i) + \mathbf{D}_\alpha^{-1}\mathbf{c}_\alpha\right),$$

em que $\Omega_\star = \mathrm{diag}\{\omega_2^2, \omega_3^2, \ldots, \omega_p^2\}$ é uma matriz diagonal de dimensão $r \times r$.

Passo 8. Gere $\boldsymbol{\beta}$ a partir da distribuição de $\boldsymbol{\beta}|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\omega}$. Com a distribuição condicional definida na equação (3.1) e a priori atribuída a $\boldsymbol{\beta}$, temos

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\omega}) &\propto \pi(\mathbf{Z}|\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\omega})\pi(\boldsymbol{\beta}) \\
&\propto \prod_{i=1}^{n}\left[\exp\left\{-\frac{u_i}{2}((\mathbf{z}_i-\mathbf{a})-\mathbf{b}x_i)^\top \Omega^{-1}((\mathbf{z}_i-\mathbf{a})-\mathbf{b}x_i)\right\}\right] \\
&\quad \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{c}_\beta)^\top \mathbf{D}_\beta^{-1}(\boldsymbol{\beta}-\mathbf{c}_\beta)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} u_i\left(x_i^2\mathbf{b}^\top\Omega^{-1}\mathbf{b} - 2x_i\mathbf{b}^\top\Omega^{-1}(\mathbf{z}_i-\mathbf{a})\right)\right\} \qquad (3.6) \\
&\quad \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^\top\mathbf{D}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbf{D}_\beta^{-1}\mathbf{c}_\beta)\right\}.,
\end{aligned}$$

Observe que em termos de $\boldsymbol{\beta}$ o segundo termo no somatório da equação (3.6) pode ser escrito da seguinte forma:

$$\begin{aligned}
-2x_i\mathbf{b}^\top\Omega^{-1}(\mathbf{z}_i-\mathbf{a}) &= -2x_i(1,\beta_1,\beta_2,\ldots,\beta_r)\mathrm{diag}\left\{\omega_1^{-2},\omega_2^{-2},\ldots,\omega_p^{-2}\right\}(X_i-0, Y_{i1}-\alpha_1, \\
&\quad Y_{i2}-\alpha_2,\ldots,Y_{ir}-\alpha_r)^\top \\
&\propto -2x_i[\beta_1\omega_2^{-2}(Y_{i1}-\alpha_1)+\beta_2\omega_3^{-2}(Y_{i2}-\alpha_2)+\ldots+\beta_r\omega_p^{-2}(Y_{ir}-\alpha_r)] \\
&= -2x_i\boldsymbol{\beta}^\top\Omega_\star^{-1}(\mathbf{Y}_i-\boldsymbol{\alpha}).
\end{aligned}$$

Analogamente, o primeiro termo no somatório pode ser escrito como

$$x_i^2\mathbf{b}^\top\Omega^{-1}\mathbf{b} \propto x_i^2\boldsymbol{\beta}^\top\Omega_\star^{-1}\boldsymbol{\beta}.$$

Assim, a distribuição condicional completa para $\boldsymbol{\beta}$ tem a seguinte expressão

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\omega}) &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n} u_i\left(x_i^2\boldsymbol{\beta}^\top\Omega_\star^{-1}\boldsymbol{\beta} - 2x_i\boldsymbol{\beta}^\top\Omega_\star^{-1}(\mathbf{Y}_i-\boldsymbol{\alpha})\right)\right.\right.\\ &\qquad \left.\left.+\boldsymbol{\beta}^\top\mathbf{D}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbf{D}_\beta^{-1}\mathbf{c}_\beta\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^\top\left(\sum_{i=1}^{n} u_i x_i^2\Omega_\star^{-1} + \mathbf{D}_\beta^{-1}\right)\boldsymbol{\beta}\right.\right.\\ &\qquad \left.\left.-2\boldsymbol{\beta}\left(\sum_{i=1}^{n} u_i x_i\Omega_\star^{-1}(\mathbf{Y}_i-\boldsymbol{\alpha}) + \mathbf{D}_\beta^{-1}\mathbf{c}_\beta\right)\right]\right\}.\end{aligned}$$

Logo, temos que $\boldsymbol{\beta}|\mathbf{Z},\mathbf{u},\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\omega} \sim \mathrm{N}_r\left(\zeta_\beta,\Delta_\beta\right)$, onde

$$\begin{aligned}\Delta_\beta &= \left(\sum_{i=1}^{n} u_i x_i^2\Omega_\star^{-1} + \mathbf{D}_\beta^{-1}\right)^{-1} \qquad \text{e}\\ \zeta_\beta &= \Delta_\beta\left(\sum_{i=1}^{n} u_i x_i\Omega_\star^{-1}(\mathbf{Y}_i-\boldsymbol{\alpha}) + \mathbf{D}_\beta^{-1}\mathbf{c}_\beta\right).\end{aligned}$$

Passo 9. A amostragem dos graus de liberdade ν será realizada em dois passos:

1) Gerar λ a partir da densidade $\pi(\lambda|\nu)$, que é a distribuição $\mathrm{TG}(2;\nu;(\lambda_0,\lambda_i))$;

2) Usar um passo de Metropolis-Hastings para gerar ν a partir da seguinte distribuição marginal condicional completa

$$\pi(\nu|\mathbf{Z},\mu_x,\sigma_x^2,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\omega}) \propto \exp(-\lambda\nu)\prod_{i=1}^{n}\left[\mathrm{T}_{p_i^c}(\boldsymbol{\kappa}_i^c|\boldsymbol{\mu}_{z_i}^{co},\mathbf{S}_{z_i}^{co},\nu+p_i^o)\mathrm{t}_{p_i^o}(\mathbf{z}_i^o|\boldsymbol{\mu}_{z_i}^{o},\boldsymbol{\Sigma}_{z_i}^{oo},\nu)\right]. \tag{3.7}$$

Dada a observação $\nu^{(j-1)}$ obtida na iteração $(j-1)$ do algoritmo, é gerado um novo candidato para $\nu^\star$ através da distribuição lognormal

$$\mathrm{LN}(\log\nu^{(j-1)};\phi_\nu^2).$$

A nova observação é aceita com probabilidade:

$$\text{mín}\left\{\frac{\pi(\nu^\star)\nu^\star}{\pi(\nu^{(j-1)})\nu^{(j-1)}},1\right\}$$

onde $\pi(\nu^{(j-1)})$ é a densidade 3.7 avaliada nos valores atualizados de $\mu_x,\ \sigma_x^2,\ \boldsymbol{\alpha},\ \boldsymbol{\beta}$ e ω.

3.3 Critérios de Seleção de Modelos

Neste trabalho apresentamos modelagem de dados com censura na resposta e na covariável via modelo de regressão com erros de medida utilizando as distribuições Normal e t de Student. Assim, em uma análise com dados reais ou simulados, a questão da escolha de modelos é de grande importância. Na literatura existem vários critérios para comparação de modelos. No contexto frequentista, critérios como o AIC (Akaike, 1974) e o BIC (Schwarz, 1978) são geralmente empregados para este fim. Para a abordagem é Bayesiana, talvez o critério mais popular seja o chamado DIC (deviance information criterion) (Spiegelhalter *et al.*, 2002) (ver também Celeux *et al.* (2006) e Spiegelhalter *et al.* (2014)), que usaremos neste trabalho em conjunto com o WAIC (Watanabe-Akaike information criterion) proposto por Watanabe (2010).

3.3.1 O DIC observado

Como primeira proposta para seleção de modelos, utilizaremos o chamado *DIC observado*, denotado por DIC_{obs}. Ele é uma reformulação do DIC para o contexto de dados aumentados, uma vez que o DIC tradicional que foi proposto por Spiegelhalter *et al.* (2002) não é adequado para esta situação, pois a verossimilhança dos dados completos não é regular, como consequência, os argumentos assintóticos que validam o DIC usual não podem ser verificados (Li *et al.*, 2012). Esta versão do DIC foi proposta por Celeux *et al.* (2006, Seção 3.1). O DIC_{obs} usa em sua formulação a verossimilhança integrada, ou seja, a verossimilhança obtida integrando as variáveis latentes. O seu cálculo é baseado nas amostras a posteriori MCMC.

Seja $D(\Theta) = -2\log L(\Theta)$ o *desvio*, onde $L(\Theta) = \prod_{i=1}^{n} L_i(\Theta)$ e $L_i(\Theta)$ é a verossimilhança definida em (2.9). Seja

$$\overline{D(\Theta)} = \text{E}[D(\Theta)|\mathbf{z}], \tag{3.8}$$

o desvio médio a posteriori. A medida

$$\tau_D = \overline{D(\boldsymbol{\theta})} - D(\tilde{\Theta}), \tag{3.9}$$

onde $\tilde{\Theta}$ é qualquer estimador de Θ, é denominada a *dimensão efetiva de* Θ (também chamada de número efetivo de parâmetros, ver Spiegelhalter *et al.* (2002)).

O DIC é definido analogamente ao critério de seleção clássico AIC proposto por Akaike (1974). Ele é dado por

$$\text{DIC} = D(\overline{\Theta}) + 2\tau_D,$$

onde $\overline{\Theta} = \text{E}[\Theta|\mathbf{z}]$ é a esperança a posteriori de Θ. Isto é, o primeiro termo no DIC é uma medida da qualidade do ajuste e a segunda é uma penalidade pela complexidade do modelo. Se usarmos $\tilde{\Theta} = \overline{\Theta}$ na equação (3.9), então

$$\begin{aligned} \text{DIC} &= -2\log L(\overline{\Theta}) + 2\text{E}[D(\Theta)|\mathbf{z}] + 4\log L(\overline{\Theta}) \\ &= 2\text{E}[D(\Theta)|\mathbf{z}] + 2\log L(\overline{\Theta}) \\ &= \overline{D(\Theta)} + \tau_D. \end{aligned} \tag{3.10}$$

Como comentamos anteriormente, existem problemas com a definição usual do DIC. Primeiramente, como abordado por Spiegelhalter *et al.* (2014), τ_D não é invariante à reparametrizações. Ou seja, diferentes parametrizações podem levar a diferentes valores de τ_D e, portanto, a diferentes valores do DIC. Uma outra questão é que, na prática, frequentemente usamos $\tilde{\Theta} = \overline{\Theta} = \text{E}[\Theta|\mathbf{z}]$. Se a distribuição de Θ é acentuadamente não normal, temos que $\overline{\Theta}$ não é um bom estimador para Θ. Com isso, podemos ter um valor negativo para τ_D.

Para solucionar este problema, observe que a verossimilhança associada à i-ésima observação $L_i(\Theta)$ é invariante à reparametrizações. Uma estimativa de $L_i(\Theta)$ é a *densidade preditiva a posteriori*, dada por $\text{E}[L_i(\Theta)|\mathbf{z}]$. Esta integral pode ser aproximada usando amostras a posteriori MCMC. Seja $\Theta^{(l)}$ a amostra MCMC gerada no l-ésima iteração do algoritmo. Para $l = 1, \ldots, m$, vamos aproximar a densidade preditiva a posteriori por

$$\hat{p}(\mathbf{z}_i) = \frac{1}{m}\sum_{l=1}^{m} L_i(\Theta^{(l)}).$$

Assim, um estimador de $L(\Theta)$ é

$$\hat{p}(\mathbf{z}) = \prod_{i=1}^{n} \hat{p}(\mathbf{z}_i), \tag{3.11}$$

e um estimador para $D(\Theta)$ é $-2\log\hat{p}(\mathbf{z})$. Observe que procedendo desta maneira temos uma expressão para τ_D que é invariante à reparametrizações, ao mesmo tempo que evitamos a utilização da média a posteriori como estimador para Θ. Substituindo em (3.9), temos

$$\tau_D = \overline{D(\boldsymbol{\theta})} + 2\log\hat{p}(\mathbf{z}).$$

A esperança a posteriori $\overline{D(\boldsymbol{\theta})} = \mathrm{E}[D(\Theta)|\mathbf{z}]$ pode aproximada por

$$\overline{D} = -\frac{2}{m}\sum_{l=1}^{m}\sum_{i=1}^{n}\log L_i(\Theta^{(l)}).$$

Finalmente, a aproximação do DIC, que será utilizada neste trabalho é

$$\mathrm{DIC}_{\mathrm{obs}} = \overline{D} + \tau_D, \text{ onde } \tau_D = \overline{D} + 2\sum_{i=1}^{n}\log\hat{p}(\mathbf{z}_i).$$

Para comparação de diferentes modelos, aquele com menor valor do $\mathrm{DIC}_{\mathrm{obs}}$ será o preferido. Para mais detalhes, veja a Seção 3.1 de Celeux *et al.* (2006).

3.3.2 WAIC

O *critério de informação Watanabe-Akaike* (WAIC) foi introduzido por Watanabe (2010). Sua definição é similar a definição dos critérios AIC e DIC, isto é,

$$\text{WAIC = medida de ajuste} + 2 \times \text{penalidade}.$$

Neste caso, utilizaremos

$$\begin{aligned}\text{medida de ajuste} &= -2\sum_{i=1}^{n}\log\pi(\mathbf{z}_i|\mathbf{z}) \\ &= -2\sum_{i=1}^{n}\log\int\pi(\mathbf{z}_i|\Theta)\pi(\Theta|\mathbf{z})d\Theta \\ &= -2\sum_{i=1}^{n}\log\mathrm{E}[L_i(\Theta)|\mathbf{z}],\end{aligned} \tag{3.12}$$

onde $\pi(\mathbf{z}_i|\mathbf{z})$ é a densidade preditiva a posteriori de $\mathbf{Z}_i$. As penalidades são definidas de duas formas:

$$\rho_{\text{WAIC}_1} = \overline{D(\boldsymbol{\theta})} + 2\sum_{i=1}^{n} \log \pi(\mathbf{z}_i|\mathbf{z}) \text{ e } \rho_{\text{WAIC}_2} = \sum_{i=1}^{n} \text{Var}\left[\log L_i(\boldsymbol{\theta})|\mathbf{z}\right],$$

onde $\overline{D(\boldsymbol{\theta})}$ é dado em (3.8). Um expressão para (3.12) não pode ser encontrada de forma fechada. Uma aproximação usando amostras MCMC é dada por

$$-2\sum_{i=1}^{n} \log\left(\frac{1}{m}\sum_{l=1}^{m} L_i(\Theta^{(l)})\right) = -2\log \hat{p}(\mathbf{z}),$$

onde $\hat{p}(\mathbf{z})$ é dado em (3.11). Assim, temos que

$$\begin{aligned}
\text{WAIC}_1 &= -2\log \hat{p}(\mathbf{z}) + 2\rho_{\text{WAIC}_1} \\
&= -2\log \hat{p}(\mathbf{z}) + 2\overline{D(\boldsymbol{\theta})} + 4\log \hat{p}(\mathbf{z}) \\
&= 2\log \hat{p}(\mathbf{z}) + 2\overline{D(\boldsymbol{\theta})} \\
&= \text{DIC}_{\text{obs}}.
\end{aligned}$$

Assim, neste trabalho, usaremos somente o WAIC_2, definido como

$$\text{WAIC}_2 = -2\log \hat{p}(\mathbf{z}) + 2\rho_{\text{WAIC}_2}.$$

Assim como no primeiro critério apresentado, para comparação de modelos, aquele que apresentar o menor valor do WAIC é escolhido como o melhor modelo.

Capítulo 4

Aplicação com Dados Simulados e Reais

Neste capítulo apresentamos três diferentes estudos de simulação com o propósito de avaliar a performance da metodologia proposta neste trabalho. O primeiro estudo analisa o comportamento do vício, do erro quadrático médio (EQM) e da cobertura dos intervalos de credibilidade das estimativas MCMC em 100 (cem) réplicas, ajustando o t-MEMC sob quatro diferentes níveis de censura: sem censura, 3% (leve), 10% (moderado) e 30% (severo). O segundo estudo compara os dois modelos propostos via critérios de seleção, também sob os quatro diferentes níveis de censura, analisando seus comportamentos para 100 (cem) réplicas. No terceiro estudo, 100 conjuntos de dados gerados a partir do t-MEMC com 20% de censura são estudados sob dois pontos de vista: Caso 1: considerando que não há censura no conjunto de dados; Caso 2: ajustando o t-MEMC. Também analisamos um conjunto de dados reais. Os procedimentos computacionais foram implementados utilizando o software R (R Core Team, 2016).

4.1 Dados Simulados

4.1.1 Estudo de Simulação 1

A proposta deste estudo de simulação é verificar o comportamento do vício e do erro quadrático médio (EQM) das estimativas MCMC obtidas via o algoritmo proposto no Capítulo 3. Além disso, avaliamos a taxa de cobertura dos intervalos de credibilidade como função dos níveis de censura. Para isso, geramos amostras de tamanho $n = 200$ do t-MEMC. Com os parâmetros fixados em $\mu_x = 4$, $\sigma_x^2 = 2$, $\Omega = \text{diag}(0.1, 0.1, 0.1)$, $\boldsymbol{\alpha} = (0.1, 0.4)^\top$,

$\boldsymbol{\beta} = (0.8, 1.2)^\top$ e $\nu = 5$. Os níveis de censura fixados foram: sem censura (0%), 3% (leve), 10% (moderado) e 30% (severo). Em relação ao passo de Metropolis-Hastings utilizado para obter amostras de ν, o parâmetro de escala para as propostas foi ajustado com o propósito de obtermos uma taxa de aceitação no intervalo (0.15;0.4).

Para cada nível de censura, foram simulados 100 conjuntos de dados. Para cada um deles, ajustamos o modelo t-MEMC e armazenamos as estimativas MCMC dos parâmetros. Então, calculamos o vício e o EQM para as estimativas. Para obter as estimativas partir de um determinado conjunto de dados fixado, foram consideradas 10000 amostras MCMC, descontando-se as primeiras 1000 como um período de burn-in. Neste caso para o vetor de parâmetros $\boldsymbol{\theta} = (\mu_x, \sigma_x^2, \omega, \boldsymbol{\alpha}, \boldsymbol{\beta}, \nu)^\top$, temos que as estimativas do vício e do EQM das estimativas de $\boldsymbol{\theta}$ para a i-ésima coordenada do vetor de parâmetros, com $i = 1, 2, \ldots, 3p$ e $p = r + 1$, são dadas pelas Equações a seguir

$$\widehat{\text{Vício}}(\hat{\theta}_i) = \frac{1}{100}\sum_{j=1}^{100}(\widehat{\theta}_i^{(j)} - \theta_i),$$

$$\widehat{\text{EQM}}(\hat{\theta}_i) = \frac{1}{100}\sum_{j=1}^{100}(\widehat{\theta}_i^{(j)} - \theta_i)^2,$$

onde $\hat{\theta}_i^{(j)}$ é a estimativa de θ_i para a j-ésima amostra gerada. Além disto, foram calculados intervalos de credibilidade utilizando as estimativas MCMC, ao final calculamos o percentual de vezes que o intervalo calculado conteve o valor real do parâmetro sob estimação.

Os valores do vício, do EQM e dos percentuais para os intervalos de credibilidade para diferentes níveis de censura são apresentados nas Tabelas 4.1 e 4.2, e representados nas Figuras **??** e **??**.

A Tabela 4.1 apresentam os resultados obtidos para o do vício e EQM para os dados simulados. É possível perceber que estes apresentam valores muito pequenos para as estimativas do vício e do EQM, os gráficos não mostram um padrão de variação bem definido tanto para o vício, quanto para o EQM, quando observados como função dos níveis de censura pré-fixados. Na Tabela 4.2 são apresentados os percentuais obtidos para os intervalos de credibilidade para cada nível de censura proposto. Observe que, em todas as combinações de nível de censura, os percentuais de vezes em que os intervalos contiveram os valores reais dos parâmetros permaneceram dentro do esperado na literatura.

Tabela 4.1: Estimativas dos vícios e dos erros quadráticos médios para os diferentes níveis de censura do t-MEMC no estudo de simulação 1.

	Sem censura		3%		10%		30%	
parâmetros	Vício	EQM	Vício	EQM	Vício	EQM	Vício	EQM
α_1	-0,00256	0,01645	-0,00405	0,01558	-0,02782	3,00e-05	-0,01782	0,02063
α_2	-0,00457	0,15336	-0,00302	0,11903	-0,02171	0,15321	0,01109	0,09632
β_1	0,00210	0,00208	0,00125	0,00100	0,00606	$< 10^{-6}$	0,00281	0,00131
β_2	0,00096	0,14406	0,00194	0,16065	0,00475	0,14408	-0,00200	0,15353
μ_x	0,00010	0,06567	-0,01823	0,00283	0,00869	0,04426	-0,01525	0,00641
σ_x^2	0,06567	0,00374	0,02633	0,03288	0,01075	0,06304	0,04641	0,25984
ω_1^2	0,00120	0,00003	0,00441	5,00e-05	0,00496	$< 10^{-6}$	0,00973	1,00e-05
ω_2^2	0,00340	0,00035	0,00654	0,00053	0,00490	$< 10^{-6}$	0,00839	0,00234
ω_3^2	0,00181	0,00080	-0,00021	0,00184	0,00084	0,00064	-0,00207	2,00e-04
ν	0,40958	1,55352	0,80305	3,1728 0	0,84219	4,11463	0,77479	3,61579

Tabela 4.2: Cobertura dos intervalos de credibilidade construídos ao nível de 95% de credibilidade para as estimativas MCMC do t-MEMC no estudo de simulação 1.

	Sem			
Parâmetros	Sem censura	3%	10%	30%
α_1	0 ,95	0 ,93	0 ,99	0 ,97
α_2	0 ,94	0 ,97	0 ,94	0 ,93
β_1	0 ,94	0 ,94	0 ,96	0 ,97
β_2	0 ,95	0 ,93	0 ,95	0 ,92
μ_x	0 ,93	0 ,90	0 ,95	0 ,98
σ_x^2	0 ,96	0 ,98	0 ,97	0 ,95
ω_1^2	0 ,96	0 ,93	0 ,96	0 ,95
ω_2^2	0 ,95	0 ,96	0 ,93	0 ,96
ω_3^2	0 ,97	0 ,95	0 ,93	0 ,94
ν	0 ,98	0 ,95	0 ,98	0 ,98

4.1.2 Estudo de Simulação 2

No segundo estudo de simulação, verificamos a performance dos dois modelos propostos utilizando os critérios de seleção de modelos DIC_{obs} e WAIC, apresentados no Capítulo 3.3. Para isto, geramos 100 amostras de tamanho $n = 200$ provenientes da distribuição Slash, definida como em (1.2). Os parâmetros foram fixados em $\mu_x = 4$, $\sigma_x^2 = 2$, $\Omega = \text{diag}(0.1, 0.1, 0.1)$, $\boldsymbol{\alpha} = (0.1, 0.4)^\top$, $\boldsymbol{\beta} = (0.8, 1.2)^\top$ e $\nu = 2$. Os níveis de censura considerados foram: sem censura, 3% (leve), 10% (moderado) e 30% (severo). Em relação ao passo de Metropolis-Hastings utilizado para obter amostras de ν, o parâmetro de escala para as propostas foi ajustado com o propósito de obtermos uma taxa de aceitação no intervalo (0.15;0.4). Comparamos para cada conjunto de dados gerado o ajuste dos modelos N-MEMC e t-MEMC segundo os critérios de seleção de modelos. Os percentuais vezes em

Tabela 4.3: Média e Percentual de vezes em que cada modelo foi melhor segundo os métodos de comparação DIC_{obs} e WAIC, para diferentes níveis de censura do estudo de simulação 2.

		DIC_{obs}		WAIC	
% de censura	Estatísticas	Normal	T	Normal	T
Sem censura	Percentual	0%	100%	0%	100%
	Média	1564,40	1521,24	1566,96	1521,70
3%	Percentual	2%	98%	1%	99%
	Média	1528,91	1497,83	1530,88	1498,28
10%	Percentual	3%	97%	2%	98%
	Média	1481,91	1450,25	1484,05	1450,75
30%	Percentual	4%	96%	4%	96%
	Média	1311,96	1287,82	1314,21	1288,41

que cada modelo foi escolhido como o melhor para em um dado nível de censura fixado foram calculados e são apresentados na Tabela 4.3, além disso foram calculados os valores médios dos critérios de seleção.

Através da Tabela 4.3 é possível perceber que a modelagem segundo a distribuição t de Student apresenta melhor desempenho, o que ocorre para todas as combinações de nível de censura.

4.1.3 Estudo de Simulação 3

No terceiro estudo, geramos 100 (cem) conjuntos de dados com 20 (vinte)% de observações censuradas provenientes do modelo t-MEMC e realizamos dois ajustes. Caso1: ajustando o modelo de regressão t de Student com erros nas váriáveis (t-MEM). Caso 2: ajustando o t-MEMC. Neste caso, geramos 100 amostras de tamanho $n = 200$ provenientes do t-MEMC. A configuração utilizada para o valores dos parâmetros foi a mesma adotada no Estudo 1. A comparação das estimativas encontradas para os parâmetros segundo esses dois casos pode ser observada nas Figuras 4.1, 4.2, 4.3 e 4.4, em que as linhas tracejadas indicam o verdadeiro valor do parâmetro sob estimação.

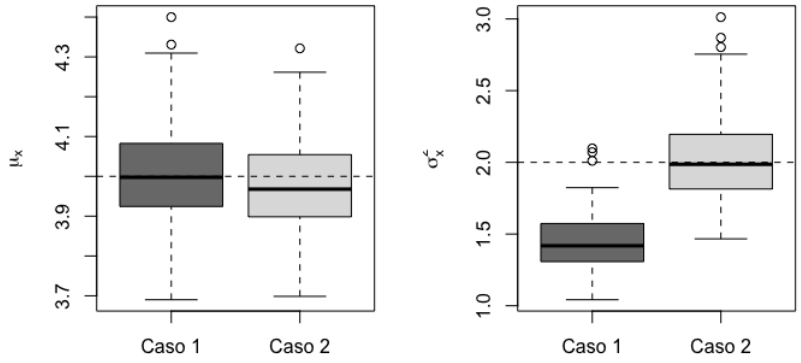

Figura 4.1: Estimativas de μ_x e σ_x^2 para os casos propostos no estudo de simulação 3

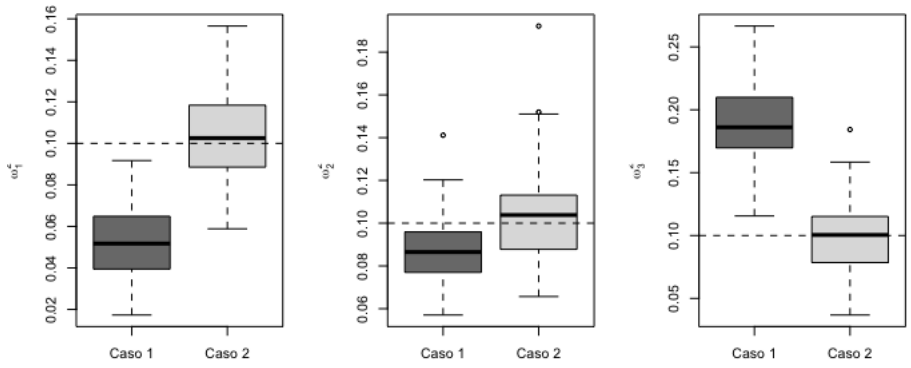

Figura 4.2: Estimativas de ω_1^2, ω_2^2 e ω_3^2 para os casos propostos no estudo de simulação 3

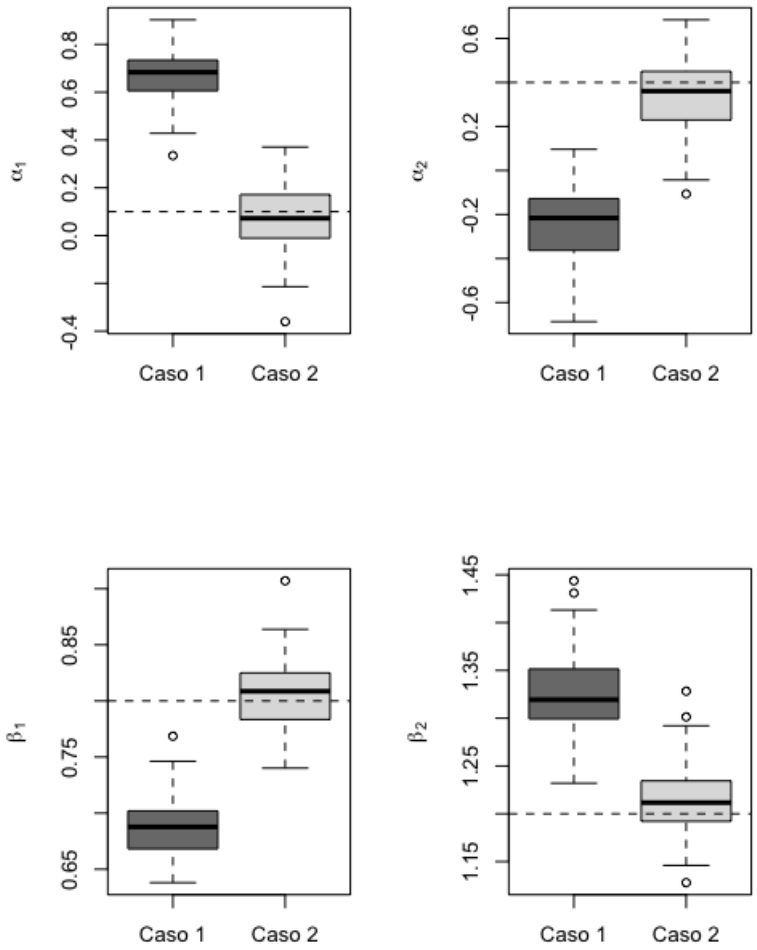

Figura 4.3: Estimativas de α_1, α_2, β_1 e β_2 para os casos propostos no estudo de simulação 3

Observando as estimativas dos parâmetros apresentadas nos boxplots das Figuras 4.1, 4.3, 4.2 e 4.4 é possível perceber que as estimativas obtidas no caso 2 são em geral mais precisas que as do caso 1. Também é possível ver que, no caso 2, a variabilidade das estimativas são menores. Essas características apresentadas no Caso 2 evidenciam a importância de considerar e trabalhar adequadamente as censuras em um conjunto de dados.

4.2 Aplicação em Dados Reais

A fim de ilustrar o método apresentado neste trabalho, utilizaremos um conjunto de dados do trabalho de Chipkevitch *et al.* (1996) apresentados na Tabela 4.4. Este conjunto

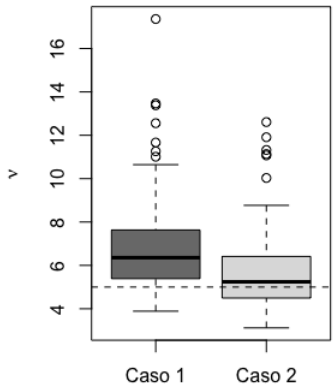

Figura 4.4: Estimativas de ν para os casos propostos no estudo de simulação 3

de dados é constituído de medições do volume testicular de 42 adolescentes, usando cinco diferentes técnicas: ultra-som(US), método gráfico proposto pelos autores (I), a medição dimensional (II), orquidômetro de Prader (III) e orquidômetro anel (IV). A abordagem do ultra-som é assumida como instrumento de medição de referência (X).

Galea *et al.* (2002) analisaram os mesmos dados ajustando um modelo com erros nas variáveis assumindo normalidade e além disso, recomendaram que uma transformação de raiz cúbica fosse feita para uma melhor aproximação ao modelo normal. Lachos *et al.* (2010) também analisaram este conjunto de dados, detectando fortes evidências de que a distribuição da variável latente apresenta um comportamento assimétrico e de cauda pesada, realizando um ajuste tentando evitar possíveis desnecessárias transformações no dados, modelando conjuntamente a variável latente e os erros observacionais pelas distribuições da família SNI (Normal Assimétrica Independente). Uma análise posterior destes dados, com conclusões semelhantes, também pode ser encontrada em Cabral *et al.* (2014).

Para ilustrar nossa metodologia usando este conjunto de dados, censuramos aleatoriamente 10% (21 observações). Ao censurá-los, o limite inferior ou ponto de corte encontrado foi de 4.4. Na Tabela 4.4 apresentamos os dados do volume testicular com o verdadeiro valor entre parênteses para as observações censuradas. Realizamos dois ajustes, um considerando o modelo t-MEMC e outro considerando o N-MEMC, para cada modelo geramos 25000 amostras Gibbs, descontando as 2000 primeiras como período de *burn-in*, selecionando os valores estimados utilizando um *lag* de 5 em 5 para eliminar o efeito da autocorrelação da

cadeia de Markov.

Tabela 4.4: Dados reais extraídos de Chipkevitch *et al.* (1996) sobre medições do volume testicular de 42 adolescentes sob 5 diferentes instrumentos.

	Métodos utilizados				
Adolescente	US	(I)	(II)	(III)	(IV)
1	5,0	7,5	5,9	8,0	9,0
2	5,7	5,0	4,8	6,0	10,0
3	7,4	5,0	6,8	9,0	12,0
4	4,4(2,6)	4,4(3,5)	4,4(3,1)	4,4(4,0)	4,4(4,0)
5	5,7	5,0	5,0	6,0	7,0
6	6,1	5,0	4,4(4,4)	7,0	8,0
7	6,2	5,0	6,0	8,0	9,0
8	10,4	10,0	8,8	10,0	10,0
9	9,1	7,5	7,9	10,0	11,0
10	14,8	10,0	13,0	12,0	15,0
11	16,4	12,5	10,3	17,5	17,5
12	9,6	7,5	8,2	10,0	11,0
13	15,7	15,0	19,8	20,0	20,0
14	4,4(3,0)	4,4(2,0)	4,4(2,0)	4,4(3,0)	4,4(4,0)
15	16,4	15,0	17,3	20,0	20,0
16	17,6	15,0	17,3	20,0	22,5
17	10,0	7,5	7,9	12,0	12,0
18	4,4(4,1)	4,4(3,5)	4,4(4,4)	4,4(4,0)	6,0
19	12,7	10,0	11,4	12,0	12,0
20	4,4(2,7)	4,4(3,5)	4,4(4,1)	4,4(2,5)	6,0
21	10,2	10,0	11,1	12,0	13,5
22	16,5	10,0	15,3	15,0	15,0
23	4,5	4,4(3,5)	4,4(3,9)	6,0	7,0
24	5,6	5,0	4,5	4,5	6,0
25	11,0	7,5	9,7	9,0	11,0
26	9,2	10,0	11,3	12,0	13,5
27	8,5	7,5	8,8	12,0	12,0
28	5,4	5,0	6,1	8,0	8,0
29	6,7	7,5	7,2	10,0	8,0
30	5,3	5,0	5,9	8,0	10,0
31	20,0	20,0	16,3	25,0	22,5
32	18,8	15,0	16,3	20,0	25,0
33	13,9	12,5	12,2	15,0	17,5
34	9,4	10,0	10,3	12,0	13,5
35	9,1	7,5	10,8	12,0	12,0
36	14,1	15,0	13,0	13,5	15,0
37	9,3	10,0	8,4	10,0	10,0
38	20,9	20,0	22,1	25,0	25,0
39	11,5	10,0	10,6	15,0	13,5
40	9,7	10,0	9,7	11,0	12,0
41	13,7	12,5	11,6	17,5	15,0
42	8,9	10,0	8,1	12,0	12,0

As Tabelas 4.5 e 4.6 apresentam, com exceção dos parâmetros de dispersão ω, σ_x^2 e dos graus de liberdade ν, os valores para a média e desvio padrão das estimativas MCMC encontradas, além dos intervalos de credibilidade construídos para cada parâmetro segundo

Tabela 4.5: Estimativas MCMC para os parâmetros nos ajustes dos modelos t-MEMC e N-MEMC para o conjunto de dados Chipkevitch.

	t-MEMC		N-MEMC	
Parâmetros	Estimativa	Desvio Padrão	Estimativa	Desvio Padrão
α_1	-0,0597	0,7655	-0,1501	0,7997
α_2	-0,6591	0,7560	-0,4591	0,7491
α_3	0,1866	0,7551	0,0095	0,7995
α_4	1,7044	0,6769	1,5811	0,7040
β_1	0,9086	0,0751	0,9060	0,0713
β_2	1,0221	0,0763	0,9849	0,0659
β_3	1,1487	0,0738	1,1482	0,0724
β_4	1,0820	0,0646	1,0821	0,0621
μ_x	9,1210	0,7652	9,8910	0,8250
σ_x^2	18,0776	5,9071	24,6471	6,6910
ω_1^2	1,1990	0,4431	1,5571	0,5143
ω_2^2	1,2294	0,4262	1,5869	0,4717
ω_3^2	1,2403	0,5039	2,0187	0,6008
ω_4^2	0,9872	0,4028	1,2311	0,4732
ω_5^2	1,2029	0,4595	1,6114	0,5117
ν	6,403296	5,6330	-	-

os modelos t-MEMC e N-MEMC. Para ω, σ_x^2 e ν consideramos na Tabela 4.5 a estimativa da mediana, pelo fato das distribuições de suas estimativas apresentarem padrão de assimetria, como pode ser visto na Figura 4.5. Na Tabela 4.7 são apresentados os valores dos critérios de seleção de modelos DIC_{obs} e WAIC calculados, observe que tanto o DIC_{obs} quanto o WAIC selecionaram o t-MEMC como o melhor modelo.

A Figura 4.5 apresenta os *traceplots* e o histogramas construídos com as estimativas MCMC obtidas no ajuste do modelo t-MEMC para os parâmetros μ_x, σ_x^2, ω_1^2, α_1, β_1 e ν. Observa-se que a distribuição a posteriori dos parâmetros σ_x^2, ω_1^2 e ν apresentam padrão assimétrico, o que indica que a mediana para esses parâmetros é um melhor estimador de seus valores.

Tabela 4.6: Intervalos de credibilidade obtidos nos ajustes dos modelos t-MEMC e N-MEMC para o conjunto de dados Chipkevitch.

	t-MEMC		N-MEMC	
Parâmetros	LIM INF	LIM SUP	LIM INF	LIM SUP
α_1	-1,4735	1,4678	-1,6786	1,4027
α_2	-2,0669	0,7646	-1,9452	0,9322
α_3	-1,3176	1,5758	-1,6220	1,4895
α_4	0,4332	3,0501	0,2457	2,9934
β_1	0,7610	1,0597	0,7801	1,0554
β_2	0,8811	1,1717	0,8594	1,1137
β_3	1,0098	1,2954	1,0043	1,2877
β_4	0,9536	1,2041	0,9664	1,2065
μ_x	7,5960	10,5680	8,2134	11,4356
σ_x^2	9,0902	30,9030	13,5360	38,3422
ω_1^2	0,5314	2,1645	0,7955	2,6953
ω_2^2	0,6066	2,1483	0,8409	2,5989
ω_3^2	0,4562	2,3251	1,1384	3,3568
ω_4^2	0,3314	1,8275	0,4599	2,2129
ω_5^2	0,4773	2,1518	0,8354	2,7260
ν	2,4372	16,7117	-	-

Tabela 4.7: Critérios de seleção obtidos nos ajustes dos modelos t-MEMC e N-MEMC para o conjunto de dados Chipkevitch.

	DIC	WAIC
t-MEMC	**829,0587**	**830,5854**
N-MEMC	833,4913	838,0598

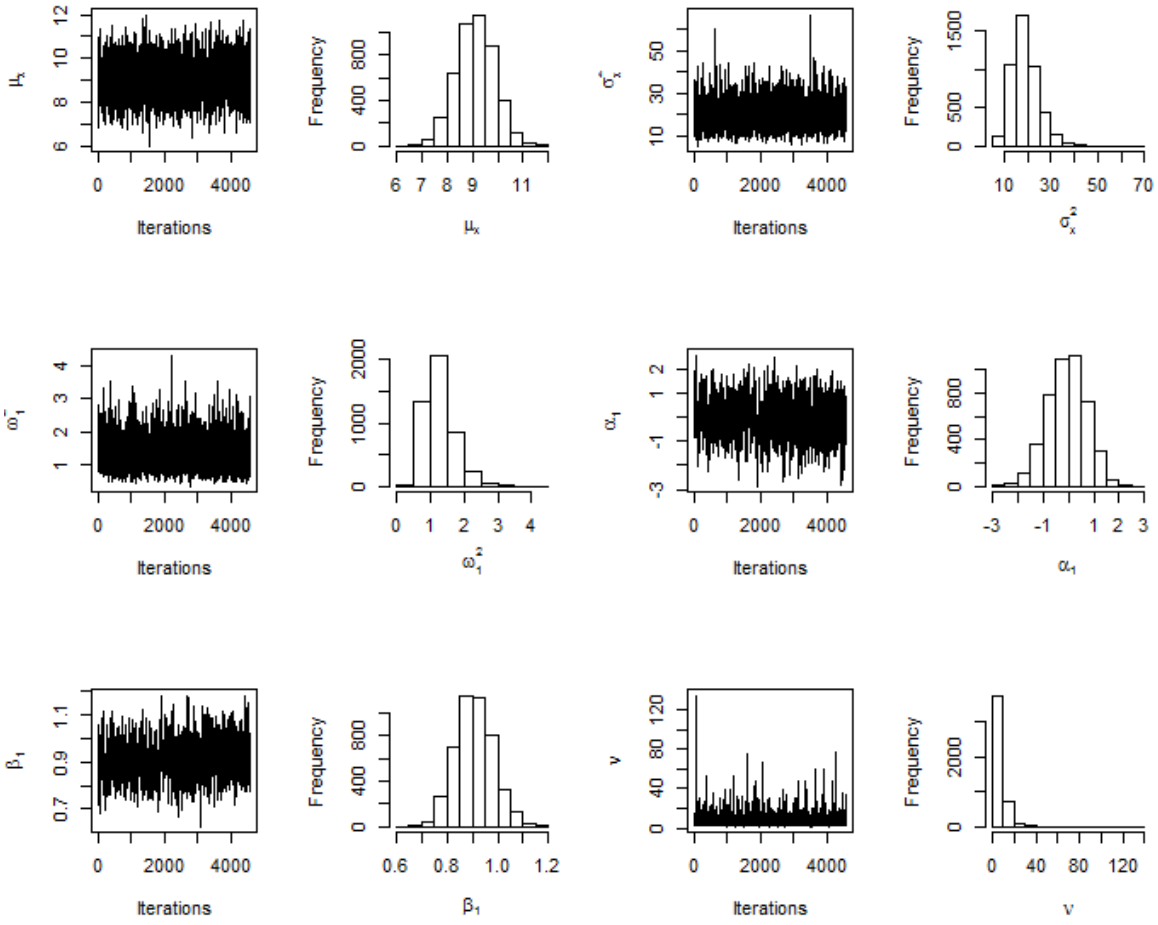

Figura 4.5: *Traceplots* e histogramas das estimativas MCMC dos parâmetros μ_x, σ_x^2, ω_1^2, α_1, β_1 e ν no ajuste do t-MEMC para o conjunto de dados Chipkevitch.

Capítulo 5

Conclusão

Neste trabalho apresentamos uma proposta de estimação Bayesiana para tratar de situações onde se tem um conjunto de observações que apresenta erros nas variáveis, censuras para algumas unidades, além de uma distribuição dos erros com caudas pesadas, como é o caso da distribuição t de Student.

Neste sentido, desenvolvemos um algoritmo do tipo Gibbs para estimação no modelo de regressão t de Student com erros nas variáveis, respostas multivariadas e censuras. Estudos de simulação foram conduzidos com a finalidade de avaliar a eficácia da nossa metodologia. Alguns cenários foram delineados com o propósito de verificar a consistência das estimativas do t-MEMC.

Esses estudos avaliaram o desempenho do modelo t-MEMC em situações onde há a necessidade de modelar um conjunto de dados cuja distribuição dos erros de observação sejam provenientes de uma distribuição com caudas mais pesadas que as da distribuição Normal e quando há nesta base de dados, diferentes percentuais de observações censuradas. Também foi observada a importância do tratamento adequado de observações censuradas. Além de uma aplicação em um conjunto de dados reais, conduzido com o intuito de ilustrar a metodologia proposta.

Para pesquisas futuras, pretendemos desenvolver métodos de análise de diagnóstico para os parâmetros do t-MEMC. Outra propostaa é a extensão da suposição de que os erros deste modelo são provenientes da distribuição t de Student para a suposição de que são provenientes de uma família de distribuições como é o caso da família SMN (misturas de escala normal) ou da SMSN(misturas de escala normal assimétrica).

Referências Bibliográficas

Ahsanullah, M. & Kibria, B. M. G. (2014). *Normal and Student's t Distributions and Their Applications*, volume 4. Atlantis Studies in Probability and Statistics.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Cont.*, **19**, 716–723.

Arellano-Valle, R. B. & Genton, M. G. (2010). Multivariate extended skew-t distributions and related families. *Metron LXVIII*, pages 201–234.

Bolfarine, H. & Arellano-Valle, R. B. (1994). Robust modelling in measurement error models using the t-distribution. *Brazilian Journal of Probability and Statistics*, **8**, 67–84.

Breen, R. (1996). *Regression Models: Censored, Sample Selected, or Truncated Data*. Sage publications.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, And Applications*. Chapman & Hall/CRC.

Cabral, C. R. B., Lachos, V. H. & Madruga, M. R. (2012). Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population. *Journal of Statistical Planning and Inference*, **142**, 181–200.

Cabral, C. R. B., Zeller, C. B. & Lachos, V. H. (2014). Multivariate measurement error models using finite mixtures of skew-Student t distributions. *Journal of Multivariate Analysis*, **124**, 179–198.

Celeux, G., Forbes, F., Robert, C. P. & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651–674.

Cheng, C. L. & Van Ness, J. W. (1999). *Statistical regression with measurement error*. Kendall's Library of Statistics. John Wiley & Sons, New York, NY.

Chipkevitch, N., Tu, D. G. S. & Galea-Rojas, M. (1996). Clinical measurement of Comparison of the reliability of 5 methods. *Journal of Urology*, **156**, 2050–2053.

Colosimo, E. A. & Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. ABE-Projeto Fisher.

de Castro, M. & Galea, M. (2010). Robust inference in an heteroscedastic measurement error model. *Journal of the Korean Statistical Society*, **39**, 439–447.

Dolby, G. R. (1976). The ultrastructural relation: A synthesis of the functional and structural relations. *Biometrika*, **63**, 39–50.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley and Sons, New York.

Galea, M., , Bolfarine, H. & Vilcalabra, F. (2002). Influence diagnostics for the structural errors-in-variables model under the Student-t distribution. *Journal of Applied Statistics*, **29**, 1191–1204.

Gamerman, D. & Lopes, F. H. (2006). *Markov Chain Monte Carlo*. Chapman & Hall, second edition.

Garay, A. M., Bolfarine, H., Lachos, V. H. & Cabral, C. R. (2015). Bayesian analysis of censored linear regression models with scale mixtures of normal distributions. *Journal of Applied Statistics*. http://dx.doi.org/10.1080/02664763.2015.1048671.

Gelfand, A. E., Dey, D. K. & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4 (Peñíscola, 1991)*, pages 147–167. Oxford Univ. Press, New York.

Kalbfleisch, J. & Lawless, J. (1992). Some useful statistical methods for truncated data. *Journal of Quality and Technology*, **24**, 145–152.

Kendall, M. G. (1951). Regression, structure and functional relationship. part i. *Biometrika*, **38**, 11–25.

Kendall, M. G. (1952). Regression, structure and functional relationship. part ii. *Biometrika*, **39**, 96–108.

Kendall, M. G. & Stuart, A. (1961). *The Advanced Theory of Statistics*, volume 2. Griffin, London.

Kotz, S. & Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge.

Lachos, V. H., Labra, F. V., Bolfarine, H. & Ghosh, P. (2010). Multivariate measurement error models based on scale mixtures of the skew-normal distribution. *Statistics*, **44**, 541–556.

Li, Y., Zeng, T. & Yu, J. (2012). Robust deviance information criterion for latent variable-models. *Economics and statistics working paper series, SMU*.

Marin, J. & Robert, C. P. (2014). *Bayesian Essentials with R*. Springer New York, second edition. ISBN 978-1-4614-8687-9.

Massuia, M. B., Garay, A. M., Lachos, V. H. & Cabral, C. R. B. (2015). *Bayesian Analysis of Censored Linear Regression Models with Scale Mixtures of Skew-Normal Distributions*. Ph.D. thesis, Universidade Estadual de Campinas, Oxford.

Matos, L. A., Prates, M. O., Chen, M. H. & Lachos, V. H. (2013). Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica*, **23**, 1323–1342.

Matos, L. A., Castro, L. M., Cabral, C. R. B. & Lachos, V. H. (2016). *Multivariate Measurement Error Models Based on Student-t Distribution under Censored Responses*. Ph.D. thesis, Universidade Estadual de Campinas, Oxford.

Nelson, W. (1990). Hazard plotting of left truncated life data. *Journal of Quality and Technology*, **22**, 230–238.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rocha, G. H., Loschi, R. H. & Arellano-Valle, R. B. (2016). Bayesian mismeasurement t-models for censored responses. *Statistics*, **50**, 841–869.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, **64**, 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 485–493.

Sprent, P. (1990). Some history of functional and structural relashionships. *Contemporary Mathematics, American Society*, **112**, 3–15.

Stapleton, D. C. & Young, D. J. (1984). Censored normal regression with measurement error on the dependent variable. *Econometrica*, **52**, 737–760.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. Ph.D. thesis, Magdalen College, Oxford.

Wang, L. (1998). Estimation of censored linear errors-in-variables models. *Journal of Econometrics*, **84**, 383–400.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, **11**, 3571–3594.

Weiss, A. (1993). Some aspects of measurement error in a censored regression model. *Journal of Econometrics*, pages 169–188.