



UFAM

MISTURAS DE ESCALA DA DISTRIBUIÇÃO NORMAL ASSIMÉTRICA COM
DADOS FALTANTES

Camila Xavier Sá Peixoto Pinheiro

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

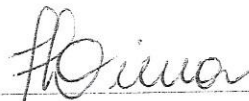
Maio de 2016

MISTURAS DE ESCALA DA DISTRIBUIÇÃO NORMAL ASSIMÉTRICA COM
DADOS FALTANTES

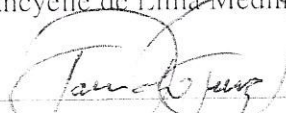
Camila Xavier Sá Peixoto Pinheiro

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. Francielle de Lima Medma, D.Sc.



Prof. José Raimundo Gomes Pereira, D.Sc.



Prof. Celso Rômulo Barbosa Cabral, D.Sc.

MANAUS, AM – BRASIL

MAIO DE 2016

Xavier Sá Peixoto Pinheiro, Camila

Misturas de Escala da Distribuição Normal Assimétrica com Dados Faltantes/Camila Xavier Sá Peixoto Pinheiro. – Manaus: UFAM/ICE, 2016.

X, 41 p.: il.; 29, 7cm.

Orientador: Celso Rômulo Barbosa Cabral

Dissertação (mestrado) – UFAM/ICE, Linha de Pesquisa: Estatística, 2016.

Referências Bibliográficas: p. 37 – 41.

1. Algoritmos MCMC. 2. Normal assimétrica. 3. Misturas de escala da normal assimétrica. 4. Dados faltantes. I. Rômulo Barbosa Cabral, Celso. II. Universidade Federal do Amazonas, UFAM, Linha de Pesquisa: Estatística. III. Título.

*Dedicado ao Professor
Celso Rômulo B. Cabral.*

Agradecimentos

Agradeço,

Às minhas tias, Diana e Amazoneida, pelo apoio, educação e exemplo de mulheres fortes e determinadas.

Ao meu pai pela educação e amor.

À minha mãe e à minha irmã pelas orações e pela torcida.

Aos amigos da UFAM Carina, Carla, Márcia e Nelson pelo eterno apoio e pela ajuda direta no trabalho.

Ao Diego pelo suporte computacional crucial para conclusão do trabalho.

À Márcia Brandão pelo incentivo diário na reta final do trabalho, tendo participação importante para que o mesmo fosse finalizado.

Aos amigos que me deram palavras de incentivo em momentos difíceis.

Ao Professor Celso pela dedicação, orientação, por acreditar em mim, mais do que eu mesma e por sempre me incentivar a seguir em frente mesmo diante das adversidades.

Ao Professor Aldo Medina por suas sugestões e auxílio na execução da nossa proposta.

Aos Professores do Departamento de Estatística da UFAM pelos ensinamentos.

Ao Professor James Dean por ter doado o computador que me acompanhou nessa trajetória e ao Professor Max por conceder alguns dias de seu computador pessoal para simulações.

Ao Guilherme Hasimoto por ceder seu computador para me ajudar nas simulações.

À CAPES pelo apoio financeiro.

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.)

MISTURAS DE ESCALA DA DISTRIBUIÇÃO NORMAL ASSIMÉTRICA COM DADOS FALTANTES

Camila Xavier Sá Peixoto Pinheiro

Maio/2016

Orientador: Celso Rômulo Barbosa Cabral

Linha de Pesquisa: Estatística

Neste trabalho estudamos uma ferramenta de estimação para modelos sob a classe de misturas de escala da distribuição normal assimétrica multivariada onde valores faltantes ocorrem nos dados. Desta forma, apresentamos uma proposta utilizando tais modelos flexíveis e algoritmos computacionais para a análise de dados multivariados com comportamento que foge do padrão usual da distribuição normal e outras distribuições simétricas usuais, apresentando forte assimetria e caudas pesadas. Além disso, mostramos a eficiência da aplicação da modelagem sugerida e do método de estimação proposto, por meio de estudos de simulação computacional, analisando a qualidade dos estimadores via estudos de vício e erro quadrático médio e comparando diferentes modelos via critérios de seleção. A abordagem inferencial utilizada foi a Bayesiana, utilizando os métodos MCMC tradicionais para obter gerações de amostras da distribuição a posterior.

Sumário

Lista de Figuras	ix
Lista de Tabelas	x
1 Introdução	1
1.1 Modelagem para Dados Multivariados com Observações Faltantes	1
1.2 Estrutura do Trabalho	3
2 Distribuição Normal Assimétrica Multivariada e Família SMSN	4
2.1 Notação	4
2.2 A Distribuição Normal Assimétrica e a Família SMSN	5
2.2.1 Distribuição t de Student Assimétrica (St)	7
2.2.2 Distribuição Slash Assimétrica (SSL)	8
2.2.3 Distribuição Normal Contaminada Assimétrica (SCN)	9
2.2.4 A Família SMN	9
2.2.5 Representação Estocástica da Família SMSN	10
3 Modelo SMSN com dados faltantes	12
3.1 A Função de Verossimilhança e Principais Resultados	12
3.2 Distribuições a Priori	16
4 O Algoritmo MCMC	18
4.1 O Amostrador de Gibbs	18
5 Comparação de modelos	23
5.1 DIC	23
5.2 WAIC	25

6	Aplicação com Dados Simulados e Reais	26
6.1	Estudo de Simulação I	26
6.2	Estudo de Simulação II	30
6.3	Análise de Dados Reais – Australian Institute of Sport Data	31
7	Conclusões e Pesquisas Futuras	36
	Referências Bibliográficas	37

Lista de Figuras

6.1	Viés e EQM das estimativas de μ_1 , μ_2 e μ_3 para os dados simulados. Primeira coluna, viés . Segunda coluna, EQM.	27
6.2	Viés e EQM das estimativas de Δ_1 , Δ_2 e Δ_3 para os dados simulados. Primeira linha, viés . Segunda linha, EQM.	29
6.3	Viés para os dados simulados. Primeira linha, Γ_{11} , Γ_{12} e Γ_{13} . Segunda linha, Γ_{22} , Γ_{23} e Γ_{33}	29
6.4	EQM para os dados simulados. Primeira linha, Γ_{11} , Γ_{12} e Γ_{13} . Segunda linha, Γ_{22} , Γ_{23} e Γ_{33}	30
6.5	Diagrama de dispersão e histogramas das variáveis BMI e LBM.	32
6.6	Primeira e segunda colunas: traceplot e histograma de μ_1 e μ_2 . Terceira e quarta colunas traceplot e histograma de Δ_1 e Δ_2	34
6.7	Traceplot e histograma de Γ_{11} , Γ_{12} e Γ_{22}	34
6.8	Traceplot e histograma de v	35

Lista de Tabelas

6.1	Viés e EQM para as estimativas para diferentes percentuais de dados faltantes e diferentes modelos.	28
6.2	Proporção de vezes em que um dos modelos foi melhor do que os demais através dos métodos de comparação DIC e WAIC, para diferentes percentuais de faltantes.	31
6.3	Critérios de seleção para os modelos da família SMNSN aplicados aos dados AIS.	33
6.4	Estimativas dos parâmetros para o conjunto de dados AIS. A sigla dp denota o desvio padrão, LI e LS representam os limites inferior e superior do intervalo de credibilidade a posteriori.	33

Capítulo 1

Introdução

1.1 Modelagem para Dados Multivariados com Observações Faltantes

O objetivo geral deste trabalho é construir uma ferramenta de estimação para modelos de misturas de escala da distribuição normal assimétrica (SMSN, *scale mixture of skew-normal*, em inglês) multivariada onde valores faltantes ocorrem nos dados, (*missing data*, em inglês).

Além disso, pretendemos mostrar a eficiência da aplicação da modelagem sugerida e do método de estimação proposto, por meio de estudos de simulação computacional, analisando a qualidade dos estimadores via estudos de vício e erro quadrático médio e comparando diferentes modelos via critérios de seleção.

Dados faltantes ocorrem com frequência na prática, nas mais diversas áreas. A presença de dados faltantes pode acarretar inúmeros problemas na análise inferencial, tais como viés nas estimativas dos parâmetros. Existem diversas técnicas para lidar com o problema, algumas não recomendáveis, como a exclusão completa de casos, e outras que se mostraram bastante eficientes em diversas aplicações, como a imputação múltipla. Para uma revisão detalhada sobre o tema, ver van Buuren (2012).

A literatura para dados multivariados faltantes com distribuição normal está bem consolidada em trabalhos como os de Anderson (1957), Hocking & Smith (1968), Rubin (1987), Liu (1999) e Sun & Sun (2006). Para a análise de dados com observações incompletas, métodos baseados em algoritmos para imputação tem sido bastante utilizados

nos últimos anos, como o EM (Dempster *et al.*, 1977, Expectation-Maximization) e o DA (Tanner & Wong, 1987, Data Augmentation).

A suposição de normalidade é uma das hipóteses utilizadas com mais frequência na modelagem estatística. Embora dados reais raramente comportem-se de acordo com esta suposição, muitas vezes é conveniente, principalmente por razões de tratabilidade matemática, assumir que os dados são provenientes de uma distribuição normal. Porém, sabemos que na prática os dados podem ser assimétricos, não-negativos, possuir múltiplas modas ou ter caudas pesadas, isso para citar alguns desvios da normalidade. No caso de observações faltantes, isso cria uma evidente incompatibilidade entre dados observados e imputados, o que pode afetar de forma negativa as estimativas de interesse.

Considerando a possibilidade de modelar dados provenientes de distribuições com caudas mais pesadas que a da normal, Liu (1995) considera o método de imputação múltipla para a distribuição t multivariada. Para avaliar a eficácia da imputação múltipla no contexto de observações multivariadas, Demirtas *et al.* (2008) analisaram o desempenho do método sob o modelo normal com covariâncias não estruturadas em uma ampla gama de conjuntos de dados incompletos simulados que apresentam diferentes características distributivas como distorção e multimodalidade, concluindo que, para amostras grandes ($n > 400$) a imputação múltipla sob a suposição de normalidade é uma ferramenta bastante razoável, mesmo quando o pressuposto de normalidade é claramente violado. No entanto, ainda permanece a questão sobre amostras de tamanho pequeno ou moderado provenientes de populações com desvios da normalidade.

Lin *et al.* (2009) propuseram algoritmos do tipo EM e MCMC para estimação dos parâmetros e imputação dos dados faltantes no modelo normal assimétrico multivariado de Azzalini & Dalla Valle (1996), que é uma distribuição que estende a normal através da incorporação de um vetor de parâmetros que regula a assimetria. Este trabalho foi estendido em Lin & Chy (2011), onde é considerada a distribuição t de Student assimétrica de Sahu *et al.* (2003), que é uma extensão da distribuição t de Student usual incorporando assimetria.

Neste trabalho propomos uma extensão das propostas de Lin *et al.* (2009) e Lin & Chy (2011), por meio da modelagem de dados com observações faltantes utilizado a classe das misturas de escala das normais assimétricas. Esta classe de distribuições contém diversos modelos que estendem o modelo normal, através da incorporação de

parâmetros que regulam ao mesmo tempo assimetria e caudas pesadas. Estão nesta classe distribuições conhecidas, como a própria normal, a slash, a normal contaminada, a t de Student, e versões assimetrizadas destas distribuições.

Neste trabalho adotaremos a abordagem inferencial Bayesiana. Uma justificativa para isto são os problemas associados ao procedimento por máxima verossimilhança (ML) no caso da distribuição normal assimétrica. Como descrito em Bayes & Branco (2007), no método ML: (i) podemos obter uma estimativa infinita para o parâmetro de forma, (ii) a matriz de informação é singular quando o parâmetro de forma é zero, e (iii) podem ocorrer máximos locais.

1.2 Estrutura do Trabalho

Esta dissertação de mestrado está dividida em sete capítulos. No Capítulo 2, é apresentada a família de distribuições multivariadas de misturas de escala da normal assimétrica, apresentando algumas propriedades úteis no contexto de análise para dados faltantes, como a representação estocástica da família. No Capítulo 3 é apresentado o modelo SMSN para análise de dados com observações faltantes. No Capítulo 4 apresentamos um algoritmo tipo MCMC para estimação a posteriori dos parâmetros do modelo proposto. No Capítulo 5 apresentamos os critérios que serão utilizados para comparar os diferentes modelos da família SMSN nas aplicações com dados reais e simulados. No Capítulo 6 mostramos dois estudos de simulação, com o objetivo de avaliar a performance da metodologia proposta, e também uma aplicação com dados reais. No Capítulo 7 são apresentadas as conclusões do trabalho e as pesquisas futuras.

Capítulo 2

Distribuição Normal Assimétrica Multivariada e Família SMSN

2.1 Notação

Primeiramente, vamos introduzir notações que serão utilizadas ao longo do trabalho. No que segue, em geral, utilizaremos a notação tradicional denotando uma variável aleatória ou vetor aleatório por letra maiúscula e a correspondente realização por letra minúscula (utilizaremos letras em negrito no caso de vetores e matrizes, aleatórios ou não). A_{ij} é a entrada (i, j) da matriz \mathbf{A} . \mathbf{A}' é a transposta de \mathbf{A} . $\mathbf{0}_{p \times q}$ é a matriz nula de ordem $p \times q$, \mathbf{I}_q é a matriz identidade de ordem q (os índices irão desaparecer quando não houver possibilidade de confusão). $\text{diag}\{a_1, \dots, a_p\}$ denota uma matriz diagonal $p \times p$ com os elementos a_1, \dots, a_p na diagonal principal. $\mathbf{D} = \text{block diag}\{\mathbf{D}_{11}, \mathbf{D}_{22}\}$ representa uma matriz particionada, denominada *block diagonal*, ou seja,

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{22} \end{pmatrix}.$$

Utilizamos os símbolos $\overset{\text{iid}}{\sim}$ e $\overset{\text{ind}}{\sim}$ para denotar variáveis aleatórias *independentes e identicamente distribuídas* and *independentes*, respectivamente.

$\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Omega})$ denota um vetor aleatório com distribuição normal p -variada com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Omega}$. A densidade correspondente é denotada por $N_p(\cdot | \boldsymbol{\mu}, \boldsymbol{\Omega})$. $\Phi(\cdot)$ é a função de distribuição da normal padrão. $\boldsymbol{\Omega}^{1/2}$ é a raiz qua-

drada de uma matriz positiva definida $\mathbf{\Omega}$, isto é, uma matriz positiva definida que satisfaz $\mathbf{\Omega}^{1/2}\mathbf{\Omega}^{1/2} = \mathbf{\Omega}$. A raiz quadrada de $\mathbf{\Omega}^{-1}$ é denotada por $\mathbf{\Omega}^{-1/2} = (\mathbf{\Omega}^{-1})^{1/2}$. A raiz quadrada é única veja, por exemplo, Gentle (2007, Sec. 5.9.1). $\det(\mathbf{\Omega})$ e $\text{tr}(\mathbf{\Omega})$ são o determinante e o traço de $\mathbf{\Omega}$, respectivamente.

2.2 A Distribuição Normal Assimétrica e a Família SMSN

A suposição de normalidade é rotineira na análise de dados, mas tal condição frequentemente é violada em situações reais, especialmente quando há a presença de forte assimetria e/ou caudas pesadas. Na prática, é comum encontrar dados com as características citadas, por exemplo, dados de renda familiar. A classe das distribuições SMSN possibilita a acomodação simultânea de assimetria e caudas pesadas, sendo assim uma alternativa viável à distribuição normal. Antes de definirmos a família SMSN, precisamos estabelecer o conceito de distribuição normal assimétrica.

O estudo de extensões da distribuição normal incorporando assimetria, ou distribuição normal assimétrica (*skew-normal*, SN), tem uma longa história na literatura de probabilidade e estatística. Contudo, não há dúvidas de que a abordagem mais famosa foi a dada por Adelchi Azzalini e colegas, quando eles apresentaram extensões para a distribuição normal univariada e multivariada em artigos como Azzalini (1985) e Azzalini & Dalla Valle (1996), respectivamente, seguidos por Arellano-Valle & Azzalini (2006).

Aqui será utilizada uma definição de distribuição normal assimétrica como membro da família de distribuição normal assimétrica fundamental, apresentada por Arellano-Valle & Genton (2005).

Definição 1. Dizemos que um vetor aleatório \mathbf{Z} p -dimensional tem distribuição normal assimétrica padrão, com notação $\mathbf{Z} \sim SN(\boldsymbol{\lambda})$, se sua densidade é dada por

$$SN_p(\mathbf{z}|\boldsymbol{\lambda}) = 2N_p(\mathbf{z}|\mathbf{0}, \mathbf{I})\Phi(\boldsymbol{\lambda}'\mathbf{z}),$$

onde, $\boldsymbol{\lambda} : p \times 1$ é um vetor de parâmetros que regula a forma da distribuição.

Parâmetros de locação e dispersão podem ser introduzidos considerando a transformação $\mathbf{Y} = \boldsymbol{\xi} + \mathbf{\Omega}^{1/2}\mathbf{Z}$, onde $\boldsymbol{\xi} : p \times 1$ é um vetor de constantes e $\mathbf{\Omega} : p \times p$ é uma

matriz positiva definida. Dessa forma, a distribuição normal assimétrica pode ser descrita de acordo com a definição que segue.

Definição 2. Um vetor aleatório p -dimensional \mathbf{X} tem distribuição normal assimétrica com vetor de locação $\boldsymbol{\mu} : p \times 1$, matriz de dispersão (positiva definida) $\boldsymbol{\Omega} : p \times p$ e vetor de forma $\boldsymbol{\lambda} : p \times 1$, com notação $\mathbf{X} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$, se sua densidade é dada por

$$SN_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}) = 2N_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Omega})\Phi(\boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})). \quad (2.1)$$

Note que o caso em que $\boldsymbol{\lambda} = \mathbf{0}$ corresponde à distribuição normal usual. A opção de utilizar essa versão é apenas uma questão de escolha e conveniência. De fato, a prática mostra que, a fim de estender os modelos baseados na normal clássica, como modelos de regressão de posto completo, modelos de efeitos mistos e modelos com erros nas variáveis a versão (2.1) da normal assimétrica apresenta algumas vantagens sobre os competidores, notavelmente quando se comparam os aspectos computacionais. Agora introduzimos o conceito mais importante desse trabalho.

Definição 3. A distribuição de um vetor p -dimensional \mathbf{Y} pertence a família de misturas de escala da normal assimétrica (SMSN), quando:

$$\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{X},$$

onde, $\boldsymbol{\mu} : p \times 1$ é um vetor de constantes, $\mathbf{X} \sim SN_p(\mathbf{0}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ e U é uma variável aleatória positiva (com probabilidade 1), independente de \mathbf{X} , com função de distribuição $H(\cdot|\mathbf{v})$.

Aqui \mathbf{v} é um parâmetro de indexação da distribuição de U , que pode ser um escalar ou um vetor. U é uma variável aleatória conhecida como *fator de escala*. $H(\cdot|\mathbf{v})$ é chamada *função de distribuição de mistura*. A notação que será utilizada é $\mathbf{Y} \sim SMSN_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \mathbf{v})$. Observe que aqui há um certo abuso de notação, por omissão de H . A família SMSN foi definida inicialmente por Branco & Dey (2001). Temos que

$$\mathbf{Y}|U = u \sim SN_p(\boldsymbol{\mu}, u^{-1}\boldsymbol{\Omega}, \boldsymbol{\lambda}), \quad U \sim H(\cdot|\mathbf{v}),$$

implicando que a densidade marginal de \mathbf{Y} é

$$SMSN_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \mathbf{v}) = 2 \int_0^\infty N_p(\mathbf{y}|\boldsymbol{\mu}, u^{-1}\boldsymbol{\Omega})\Phi(u^{1/2}\boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}))dH(u|\mathbf{v}). \quad (2.2)$$

Nas próximas sessões serão apresentados os diferentes membros da família SMSN que serão usados na nossa proposta levando em conta algumas possibilidades da distribuição do fator de escala U . Observe que a distribuição SN corresponde ao caso em que U é degenerada em 1, ou seja, $P[U = 1] = 1$. Para mais detalhes sobre a família de distribuições SMSN veja Lachos & Labra (2014).

2.2.1 Distribuição t de Student Assimétrica (St)

A distribuição t de Student multivariada é uma extensão da tradicional distribuição t de Student que, como é de amplo conhecimento, tem um papel central na teoria estatística. Possui aplicações nas mais diversas áreas, como uma alternativa ao uso da distribuição normal, em situações onde a distribuição dos dados apresenta caudas pesadas. Estas áreas incluem, por exemplo, econometria (Yang & Yuan, 2016), engenharia (Vidal & Castro, 2010), reconhecimento de padrões (Wei, 2012) e medicina (Nguyen & Wu, 2012). Para um texto com uma teoria abrangente envolvendo a distribuição t de Student multivariada, veja Kotz & Nadarajah (2004).

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

$$t_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{p+\nu}{2})}{\Gamma(\frac{\nu}{2})\pi^{p/2}} \nu^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \left(1 + \frac{d_{\boldsymbol{\Sigma}}(\mathbf{y}, \boldsymbol{\mu})}{\nu}\right)^{-(p+\nu)/2}, \quad (2.3)$$

onde $\Gamma(\cdot)$ é a função gamma e $d_{\boldsymbol{\Sigma}}(\mathbf{y}, \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ é a distância de Mahalanobis ao quadrado. Neste caso, usamos a notação $\mathbf{Y} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$.

Uma extensão da distribuição t de Student incorporando assimetria é a distribuição t de Student assimétrica (*skew-t*, St) com ν graus de liberdade, com notação $St_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu)$, que é obtida fazendo $U \sim \text{Gamma}(\nu/2, \nu/2)$, $\nu > 0$ na Definição 3. Utilizando argumentos similares aos utilizados em Gupta (2003), é possível provar que a densidade de \mathbf{Y} é dada por

$$St_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu) = 2t_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu) T \left[\left(\frac{\nu + p}{\nu + d_{\boldsymbol{\Omega}}(\mathbf{y}, \boldsymbol{\mu})} \right)^{1/2} \boldsymbol{\lambda}' \boldsymbol{\Omega}^{-1/2} (\mathbf{y} - \boldsymbol{\mu}) | \nu + p \right],$$

onde $T(\cdot | \nu + p)$ é a função de distribuição acumulada da t de Student padrão univariada

(ou seja, com média zero e escala 1) e $v + p$ são seus graus de liberdade. O vetor de médias e a matriz de covariâncias de \mathbf{Y} são dados respectivamente por

$$E[\mathbf{Y}] = \boldsymbol{\mu} + \sqrt{\frac{v}{\pi}} \frac{\Gamma((v-1)/2)}{\Gamma(v/2)} \boldsymbol{\Omega}^{1/2} \boldsymbol{\delta};$$

$$Var[\mathbf{Y}] = \frac{v}{v-2} \boldsymbol{\Omega} - (v/\pi) \left(\frac{\Gamma((v-1)/2)}{\Gamma(v/2)} \right)^2 \boldsymbol{\Omega} \boldsymbol{\delta} \boldsymbol{\delta}'.$$

onde $\boldsymbol{\delta} = \boldsymbol{\lambda} / (1 + \boldsymbol{\lambda}' \boldsymbol{\lambda})^{1/2}$.

Note que se $\boldsymbol{\lambda}$ é o vetor nulo, então recuperamos a distribuição t de Student simétrica. No contexto de modelos simétricos, a distribuição t de Student torna-se uma alternativa à distribuição normal para lidar com valores extremos, já que apresenta caudas mais pesadas que a normal, podendo proporcionar ajustes mais robustos. Isso por que a t de Student apresenta um atrativo adicional: o parâmetro extra v pode ser entendido como o responsável pela acomodação de valores extremos. Evidentemente essa interpretação estende-se ao contexto da distribuição t de Student assimétrica.

2.2.2 Distribuição Slash Assimétrica (SSL)

A distribuição Slash assimétrica (*skew slash*, SSL), denotada por $SSL_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, v)$, é mais um modelo que possui caudas pesadas e assimetria. Essa distribuição pode ser obtida utilizando a Definição 3, fazendo $U \sim Beta(v, 1)$ com $v > 0$, onde $Beta(a, b)$ denota a distribuição Beta com parâmetros $a > 0$ e $b > 0$, cuja densidade é dada por

$$f(u) = u^{a-1} (1-u)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a) + \Gamma(b)}, \quad 0 < u < 1.$$

Assim, a densidade de $\mathbf{Y} \sim SSL_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, v)$ é dada por

$$SSL_p(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, v) = 2v \int_0^1 u^{v-1} N_p(\mathbf{y} | \boldsymbol{\mu}, u^{-1} \boldsymbol{\Omega}) \Phi(u^{1/2} \boldsymbol{\lambda}' \boldsymbol{\Omega}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})) du,$$

que pode ser avaliada usando a função do R `integrate` (R Core Team, 2016). O vetor de médias e a matriz de covariâncias são respectivamente dados por

$$E[\mathbf{Y}] = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}} \frac{2\nu}{2\nu-1} \boldsymbol{\Omega}^{1/2} \boldsymbol{\delta}, \quad \nu > 1/2;$$

$$Var[\mathbf{Y}] = \frac{\nu}{\nu-1} \boldsymbol{\Omega} - (2/\pi) \left(\frac{2\nu}{2\nu-1} \right)^2 \boldsymbol{\Omega}^{1/2} \boldsymbol{\delta} \boldsymbol{\delta}' \boldsymbol{\Omega}^{1/2}, \quad \nu > 1.$$

2.2.3 Distribuição Normal Contaminada Assimétrica (SCN)

A distribuição normal contaminada assimétrica (*skew contaminated normal*, SCN), é obtida na Definição 3 quando o fator de escala U é uma variável aleatória binária com $P(U = \tau) = \rho$ e $P(U = 1) = 1 - \rho$, onde $0 < \tau, \rho < 1$ (e portanto $\mathbf{v} = (\tau, \rho)'$). Neste caso, usamos a notação $\mathbf{Y} \sim \text{SCN}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \mathbf{v})$. A densidade de \mathbf{Y} é dada por

$$\text{SCN}_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \mathbf{v}) =$$

$$2 \left\{ \rho N_p(\mathbf{y}|\boldsymbol{\mu}, \tau^{-1} \boldsymbol{\Omega}) \Phi(\tau^{1/2} \boldsymbol{\lambda}' \boldsymbol{\Omega}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})) + (1 - \rho) N_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}) \Phi(\boldsymbol{\lambda}' \boldsymbol{\Omega}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})) \right\}.$$

Os parâmetros ρ e τ indicam a proporção de outliers e o grau de contaminação, respectivamente. O vetor de médias e matriz de covariâncias são dados respectivamente por

$$E[\mathbf{Y}] = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}} \left(\frac{\rho}{\tau^{1/2}} + 1 - \rho \right) \boldsymbol{\Omega}^{1/2} \boldsymbol{\delta}$$

$$Var[\mathbf{Y}] = \left(\frac{\rho}{\tau} + 1 - \nu_1 \right) \boldsymbol{\Omega} - \frac{2}{\pi} \left(\frac{\rho}{\tau^{1/2}} + 1 - \nu_1 \right)^2 \boldsymbol{\Omega} \boldsymbol{\delta} \boldsymbol{\delta}'.$$

2.2.4 A Família SMN

Se na Definição 3 fizermos $\boldsymbol{\lambda} = \mathbf{0}$, teremos a *família de misturas de escala da distribuição normal* (SMN), como definida em Andrews & Mallows (1974) (ver também Lange & Sinsheim (1993)). Neste caso teremos, como casos particulares, as já citadas distribuições t de Student e normal, e as distribuições slash e normal contaminada.

2.2.5 Representação Estocástica da Família SMSN

Por razões computacionais utilizamos uma parametrização sugerida por Lachos & Labra (2014) que é dada da seguinte forma:

$$\mathbf{\Delta} = \mathbf{\Omega}^{1/2} \boldsymbol{\delta}, \quad \mathbf{\Gamma} = \mathbf{\Omega}^{1/2} (\mathbf{I} - \boldsymbol{\delta} \boldsymbol{\delta}') \mathbf{\Omega}^{1/2} = \mathbf{\Omega} - \mathbf{\Delta} \mathbf{\Delta}', \quad (2.4)$$

É possível recuperar os parâmetros originais $\boldsymbol{\lambda}$ e $\mathbf{\Omega}$ utilizando as relações

$$\boldsymbol{\lambda} = \frac{(\mathbf{\Gamma} + \mathbf{\Delta} \mathbf{\Delta}')^{-1/2} \mathbf{\Delta}}{[1 - \mathbf{\Delta}' (\mathbf{\Gamma} + \mathbf{\Delta} \mathbf{\Delta}')^{-1} \mathbf{\Delta}]^{1/2}}, \quad \mathbf{\Omega} = \mathbf{\Gamma} + \mathbf{\Delta} \mathbf{\Delta}'. \quad (2.5)$$

No que segue, utilizamos esta parametrização ou a parametrização original conforme a conveniência.

A proposição a seguir – uma representação estocástica para a família SMSN – é fundamental para a nossa teoria. A demonstração é consequência da Proposição 2.2 de Arellano-Valle & Genton (2005). A partir deste resultado, é possível obter um algoritmo do tipo Gibbs relativamente simples, com o objetivo de proceder estimação Bayesiana para os parâmetros dos modelos que iremos propor. Metodologia similar a que usaremos aqui foi utilizada com sucesso em outros trabalhos que estendem modelos baseados na suposição de normalidade como, por exemplo, Lachos *et al.* (2009a), Lachos *et al.* (2009b), Cancho *et al.* (2010), Cabral *et al.* (2012) e Garay *et al.* (2015).

Proposição 1. $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \mathbf{\Omega}, \boldsymbol{\lambda}, \mathbf{v})$ admite a seguinte representação estocástica

$$\begin{aligned} \mathbf{Y} | U = u, T = t &\sim \text{N}_p(\boldsymbol{\mu} + \mathbf{\Delta} t, u^{-1} \mathbf{\Gamma}); \\ T | U = u &\sim \text{TN}(0, u^{-1}, (0, \infty)); \\ U &\sim H(\cdot | \mathbf{v}), \end{aligned}$$

onde $\mathbf{\Delta}$ e $\mathbf{\Gamma}$ são como em (2.4), e $\text{TN}(\xi, \sigma^2, (a, b))$ denota a distribuição normal truncada no intervalo (a, b) , onde ξ e σ^2 são a média e a variância, respectivamente, antes do truncamento.

Note que essa proposição pode ser usada para gerar amostras artificiais de modelos SMSN, além de possibilitar obter a distribuição de uma transformação linear $A\mathbf{Y}$, onde A é uma matriz $q \times p$, segundo o próximo corolário.

Corolário 1. *Considerando a parametrização (2.4), seja $A : q \times p$ e $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \mathbf{v})$. Então $A\mathbf{Y} \sim \text{SMSN}_q(A\boldsymbol{\mu}, A\boldsymbol{\Gamma}A', A\boldsymbol{\Delta}, \mathbf{v})$.*

Para recuperar os parâmetros originais basta usar os resultados em (2.5).

Capítulo 3

Modelo SMSN com dados faltantes

3.1 A Função de Verossimilhança e Principais Resultados

Suponha que temos observações independentes em n indivíduos, $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$, onde cada $\mathbf{Y}_i \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \mathbf{v})$, $i = 1, 2, \dots, n$. Para cada i , vamos particionar o vetor \mathbf{Y}_i na forma $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$, onde $\mathbf{Y}_i^o : p_i^o \times 1$ é um vetor com as coordenadas observadas e $\mathbf{Y}_i^m : p_i^m \times 1$ é o vetor com as coordenadas que contém as observações faltantes. É importante observar que o conjunto de índices da partição depende da observação i . Ao longo da nossa exposição, vamos supor que o mecanismo gerador de observações faltantes é do tipo MAR (*missing at random*), como proposto por Rubin (1976), o que significa que o mecanismo de geração depende somente dos valores observados. Esta hipótese também foi utilizada em Lin *et al.* (2009) e Lin & Chy (2011).

Considere que $\boldsymbol{\mu} = ((\boldsymbol{\mu}_i^o)', (\boldsymbol{\mu}_i^m)')'$, $\boldsymbol{\Delta} = ((\boldsymbol{\Delta}_i^o)', (\boldsymbol{\Delta}_i^m)')'$ e $\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_i^{oo} & \boldsymbol{\Gamma}_i^{om} \\ \boldsymbol{\Gamma}_i^{mo} & \boldsymbol{\Gamma}_i^{mm} \end{pmatrix}$ são as partições correspondentes de $\boldsymbol{\mu}$, $\boldsymbol{\Delta}$ e $\boldsymbol{\Gamma}$. Além disso, vamos considerar uma representação hierárquica para \mathbf{Y}_i baseada na Proposição 1, ou seja,

$$\mathbf{Y}_i | U_i = u_i, T_i = t_i \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Delta}t_i, u_i^{-1}\boldsymbol{\Gamma}); \quad (3.1)$$

$$T_i | U_i = u_i \sim \text{TN}(0, u_i^{-1}, (0, \infty)); \quad (3.2)$$

$$U_i \sim H(\cdot | \mathbf{v}), \quad i = 1, \dots, n. \quad (3.3)$$

Temos então os seguintes resultados, que serão utilizados na elaboração do algoritmo para

estimação no modelo que iremos propor:

Proposição 2. (i) A distribuição condicional de \mathbf{Y}_i^m dado $\mathbf{Y}_i^o = \mathbf{y}_i^o$, $T_i = t_i$ e $U_i = u_i$ é

$$\mathbf{Y}_i^m | \mathbf{Y}_i^o = \mathbf{y}_i^o, T_i = t_i, U_i = u_i \sim N_{p_i^m}(\boldsymbol{\mu}_i^m + \boldsymbol{\Delta}_i^m t + \boldsymbol{\Gamma}_i^{mo} (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o - \boldsymbol{\Delta}_i^o t_i), u^{-1} \boldsymbol{\Gamma}_i^{mm.o}),$$

$$\text{onde } \boldsymbol{\Gamma}_i^{mm.o} = \boldsymbol{\Gamma}_i^{mm} - \boldsymbol{\Gamma}_i^{mo} (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Gamma}_i^{om};$$

(ii) A distribuição condicional de \mathbf{Y}_i^m dado $T_i = t_i$ e $U_i = u_i$ é

$$\mathbf{Y}_i^m | T_i = t_i, U_i = u_i \sim N_{p_i^m}(\boldsymbol{\mu}_i^m + \boldsymbol{\Delta}_i^m t_i, u_i^{-1} \boldsymbol{\Gamma}_i^{mm});$$

(iii) A distribuição condicional de T_i dado $\mathbf{Y}_i^o = \mathbf{y}_i^o$ e $U_i = u_i$ é

$$T_i | \mathbf{Y}_i^o = \mathbf{y}_i^o, U_i = u_i \sim \text{TN}(\mu_{T_i}, u_i^{-1} M_{T_i}^2, (0, \infty)),$$

$$\text{onde } \mu_{T_i} = M_{T_i}^2 (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) \text{ e } M_{T_i}^2 = 1 / (1 + (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o).$$

(iv) Se $U_i \sim \text{Gamma}(\mathbf{v}/2, \mathbf{v}/2)$, $\mathbf{v} > 0$, a distribuição condicional de T_i dado $\mathbf{Y}_i^o = \mathbf{y}_i^o$ é

$$\text{Tt} \left((\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)' \boldsymbol{\Omega}_{oo}^{-1} \boldsymbol{\Delta}_i^o, \frac{(d_{\boldsymbol{\Omega}_{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + \mathbf{v})(1 - (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Omega}_{oo})^{-1} \boldsymbol{\Delta}_i^o)}{\mathbf{v} + p_i^o}, \mathbf{v} + p_i^o, (0, \infty) \right),$$

que é uma distribuição *t* de Student truncada em $(0, \infty)$.

Demonstração. O resultado (i) é consequência de (3.1) e do clássico resultado sobre distribuição condicional de subvetores de um vetor aleatório com distribuição normal multivariada, ver Mardia *et al.* (1979, Teorema 3.2.4). O resultado (ii) é consequência de (3.1) e do fato de que subvetores de um vetor com distribuição normal multivariada também tem distribuição normal multivariada (Mardia *et al.*, 1979, Teorema 3.2.3). Os resultados (iii) e (iv) serão demonstrados em um único bloco. Novamente utilizando (3.1) e o Teorema 3.2.3 de Mardia *et al.* (1979), vem que

$$\mathbf{Y}_i^o | U_i = u_i, T_i = t_i \sim N_{p_i^o}(\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i, u_i^{-1} \boldsymbol{\Gamma}_i^{oo}). \quad (3.4)$$

No que segue, para vetores aleatórios arbitrários \mathbf{X} e \mathbf{Y} , $\pi(\mathbf{x}|\mathbf{y})$ denota a densidade condi-

cional de $\mathbf{X}|\mathbf{Y} = \mathbf{y}$. Então, a distribuição conjunta de \mathbf{Y}_i^o , T_i e U_i é dada por

$$\begin{aligned} \pi(\mathbf{y}_i^o, t_i, u_i) &= \pi(\mathbf{y}_i^o | t_i, u_i) \pi(t_i | u_i) \pi(u_i) \\ &\propto u_i^{p_i^o/2} \exp \left\{ -\frac{u_i}{2} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o - \boldsymbol{\Delta}_i^o t_i)' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o - \boldsymbol{\Delta}_i^o t_i) \right\} \end{aligned} \quad (3.5)$$

$$\times u_i^{1/2} \exp \left\{ -\frac{u_i}{2} t_i^2 \right\} \mathbb{I}_{(0,\infty)}(t_i) \quad (3.6)$$

$$\begin{aligned} &\times u_i^{v/2-1} \exp \left\{ -\frac{v}{2} u_i \right\} \\ &= u_i^{(p_i^o+1+v)/2-1} \mathbb{I}_{(0,\infty)}(t_i) \\ &\quad \times \exp \left\{ -\frac{u_i}{2} [(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o - \boldsymbol{\Delta}_i^o t_i)' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o - \boldsymbol{\Delta}_i^o t_i) + t_i^2 + v] \right\} \\ &= u_i^{(p_i^o+1+v)/2-1} \mathbb{I}_{(0,\infty)}(t_i) \\ &\quad \times \exp \left\{ -\frac{u_i}{2} [(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) \right. \\ &\quad \left. - 2(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o t_i + (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o t_i^2 + t_i^2 + v] \right\} \\ &= u_i^{(p_i^o+1+v)/2-1} \mathbb{I}_{(0,\infty)}(t_i) \\ &\quad \times \exp \left\{ -\frac{u_i}{2} [(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) \right. \\ &\quad \left. - 2(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o t_i + ((\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o + 1) t_i^2 + v] \right\}, \end{aligned} \quad (3.7)$$

onde $\mathbb{I}_B(\cdot)$ denota a função indicadora, isto é, $\mathbb{I}_B(t_i) = 1$ se $t_i \in B$ e $\mathbb{I}_B(t_i) = 0$ caso contrário. Observe que, utilizando as equações (3.5) e (3.6), podemos obter a distribuição condicional de $T_i | \mathbf{Y}_i^o = \mathbf{y}_i^o, U_i = u_i$, que é dada por

$$\begin{aligned} \pi(t_i | \mathbf{y}_i^o, u_i) &\propto \pi(\mathbf{y}_i^o | t_i, u_i) \pi(t_i | u_i) \\ &\propto \exp \left\{ -\frac{u_i}{2} [-2(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o t_i + (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o t_i^2] \right\} \\ &\quad \times \exp \left\{ -\frac{u_i}{2} t_i^2 \right\} \mathbb{I}_{(0,\infty)}(t_i) \\ &= \exp \left\{ -\frac{u_i}{2} [-2(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o t_i + (1 + (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o) t_i^2] \right\} \mathbb{I}_{(0,\infty)}(t_i), \end{aligned}$$

que é exatamente o núcleo da distribuição normal truncada dada no item (iii). Vamos agora obter o resultado (iv). Utilizando (2.5) e o exercício 2.9 de Rao (1973), obtemos

$$\begin{aligned} (\boldsymbol{\Gamma}_i^{oo})^{-1} &= (\boldsymbol{\Omega}_i^{oo} - \boldsymbol{\Delta}_i^o (\boldsymbol{\Delta}_i^o)')^{-1} = (\boldsymbol{\Omega}_i^{oo})^{-1} + (\boldsymbol{\Omega}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o (1 - (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Omega}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o)^{-1} (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Omega}_i^{oo})^{-1} \\ &= (\boldsymbol{\Omega}_i^{oo})^{-1} + (\boldsymbol{\Omega}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o (1 + (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o) (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Omega}_i^{oo})^{-1} \\ &= (\boldsymbol{\Omega}_i^{oo})^{-1} + (\boldsymbol{\Omega}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o \psi_i^{-1} (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Omega}_i^{oo})^{-1}, \end{aligned}$$

onde $\psi_i^{-1} = 1 + (\Delta_i^o)'(\Gamma_i^{oo})^{-1}\Delta_i^o$ (note que o resultado de Rao (1973) foi utilizado para encontrar as inversas de Γ_i^{oo} e $1 - (\Delta_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o$). Então, usamos esta expressão de $(\Gamma_i^{oo})^{-1}$ na equação (3.7), obtendo

$$\begin{aligned}
\pi(\mathbf{y}_i^o, t_i, u_i) &\propto u_i^{(p_i^o+1+\nu)/2-1} \mathbb{I}_{(0,\infty)}(t_i) \\
&\times \exp \left\{ -\frac{u_i}{2} \left[(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)'((\Omega_i^{oo})^{-1} + (\Omega_i^{oo})^{-1}\Delta_i^o\psi_i^{-1}(\Delta_i^o)'(\Omega_i^{oo})^{-1})(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) \right. \right. \\
&\quad \left. \left. - 2(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)'((\Omega_i^{oo})^{-1} + (\Omega_i^{oo})^{-1}\Delta_i^o\psi_i^{-1}(\Delta_i^o)'(\Omega_i^{oo})^{-1})\Delta_i^o t_i + \psi_i^{-1}t_i^2 + \nu \right] \right\} \\
&\propto u_i^{(p_i^o+1+\nu)/2-1} \mathbb{I}_{(0,\infty)}(t_i) \\
&\times \exp \left\{ -\frac{u_i}{2} \left[(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)'(\Omega_i^{oo})^{-1}(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) - 2(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o t_i + \psi_i^{-1}t_i^2 + \nu \right. \right. \\
&\quad \left. \left. + (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o\psi_i^{-1}(\Delta_i^o)'(\Omega_i^{oo})^{-1}(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) \right. \right. \\
&\quad \left. \left. - 2(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o\psi_i^{-1}(\Delta_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o t_i \right] \right\} \\
&\propto u_i^{(p_i^o+1+\nu)/2-1} \mathbb{I}_{(0,\infty)}(t_i) \\
&\times \exp \left\{ -\frac{u}{2} \left[d_{\Omega_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) - 2a_i^o t_i + \psi_i^{-1}t_i^2 + (a_i^o)^2\psi_i^{-1} - 2a_i^o\psi_i^{-1}(\Delta_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o t_i + \nu \right] \right\} \\
&\propto u_i^{(p_i^o+1+\nu)/2-1} \mathbb{I}_{(0,\infty)}(t_i) \\
&\times \exp \left\{ -\frac{u_i}{2} \left[d_{\Omega_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + \psi_i^{-1}t_i^2 - 2a_i^o\psi_i^{-1}t_i + (a_i^o)^2\psi_i^{-1} \right. \right. \\
&\quad \left. \left. - 2a_i^o t_i + 2a_i^o\psi_i^{-1}t_i - 2a_i^o\psi_i^{-1}(\Delta_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o t_i + \nu \right] \right\} \\
&\propto u_i^{(p_i^o+1+\nu)/2-1} \mathbb{I}_{(0,\infty)}(t_i) \\
&\times \exp \left\{ -\frac{u_i}{2} \left[d_{\Omega_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + (t_i - a_i^o)^2\psi_i^{-1} + 2a_i^o t_i(-1 + \psi_i^{-1} - \psi_i^{-1}(\Delta_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o) + \nu \right] \right\},
\end{aligned}$$

onde $a_i^o = (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o$. Mas $(\Delta_i^o)'(\Omega_i^{oo})^{-1}\Delta_i^o = 1 - \psi_i$, e assim o terceiro termo da última expressão é zero. Então,

$$\pi(\mathbf{y}_i^o, t_i, u_i) \propto u_i^{(p_i^o+1+\nu)/2-1} \exp \left\{ -\frac{u_i}{2} \left[d_{\Omega_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + (t_i - a_i^o)^2\psi_i^{-1} + \nu \right] \right\} \mathbb{I}_{(0,\infty)}(t_i). \quad (3.8)$$

Integrando a expressão 3.8 com relação a u_i , obtemos

$$\pi(\mathbf{y}_i^o, t_i) \propto \left(\frac{d_{\Omega_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + (t_i - a_i^o)^2\psi_i^{-1} + \nu}{2} \right)^{-\frac{p_i^o+1+\nu}{2}} \mathbb{I}_{(0,\infty)}(t_i). \quad (3.9)$$

Como temos $\pi(t_i|\mathbf{y}_i^o) = \pi(\mathbf{y}_i^o, t_i)/\pi(\mathbf{y}_i^o)$, o mesmo resultado é válido para a distribuição condicional de $T_i|\mathbf{Y}_i^o = \mathbf{y}_i^o$. Multiplicando (3.9) por $(2/(d_{\Omega_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + \nu))^{-(p_i^o+1+\nu)/2}$,

obtemos

$$\pi(t_i | \mathbf{y}_i^o) \propto \left(1 + \frac{(t_i - a_i^o)^2 \psi_i^{-1}(\nu + p_i^o)}{(\nu + p_i^o)(d_{\mathbf{a}_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + \nu)} \right)^{-\frac{p_i^o + 1 + \nu}{2}} \mathbb{I}_{(0, \infty)}(t_i), \quad (3.10)$$

que é uma distribuição t de Student-t truncada em $(0, \infty)$, com parâmetros antes do truncamento a_i^o (média), $\psi_i(d_{\mathbf{a}_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + \nu)/(\nu + p_i^o)$ (escala) e $\nu + p_i^o$ (graus de liberdade). \square

Como consequência da Proposição de 3.4 e da representação estocástica (3.1)-(3.3), temos o próximo resultado.

Proposição 3. *A distribuição de \mathbf{Y}_i^o é $\text{SMSN}_{p_i^o}(\boldsymbol{\mu}_i^o, \boldsymbol{\Gamma}_i^{oo}, \boldsymbol{\Delta}_i^o, \boldsymbol{\nu})$.*

3.2 Distribuições a Priori

Em um contexto Bayesiano, especificações das distribuições a priori são necessárias para complementar a construção do modelo e proceder inferência. Primeiramente, vamos concatenar os vetores $\boldsymbol{\mu}$ e $\boldsymbol{\Delta}$ em um único vetor $\boldsymbol{\phi}$, atribuindo-lhe uma distribuição a priori normal. Especificamente, vamos supor que

$$\boldsymbol{\phi} = (\boldsymbol{\mu}', \boldsymbol{\Delta}')' \sim N_{2p}(\mathbf{a}, \mathbf{B}), \quad (3.11)$$

onde \mathbf{a} é um vetor de médias conhecido e \mathbf{B} é uma matriz de covariâncias positiva definida conhecida. Essa concatenação dos parâmetros em um único vetor, conhecida como *blocking*, permite a construção de um algoritmo MCMC mais eficiente para gerar amostras da distribuição a posteriori, já que a atualização para os dois parâmetros é feita em um só passo do algoritmo, veja Liu (1994).

No que diz respeito à matriz de dispersão $\boldsymbol{\Gamma}$, adotamos uma especificação a priori hierárquica, definida por

$$\boldsymbol{\Gamma}^{-1} | \mathbf{D} \sim \text{Wish}_p(c, \mathbf{D}), \quad \mathbf{D} \sim \text{Wish}_p(e, \mathbf{F}),$$

onde $\text{Wish}_p(c, \mathbf{D})$ denota a distribuição Wishart, \mathbf{F} é uma matriz assimétrica conhecida, com $\det(\mathbf{F}) > 0$, c e e são hiperparâmetros conhecidas, com $c, e > (p - 1)/2$. Esta especi-

ficação foi sugerida por Richardson & Green (1997) no contexto de misturas de distribuições normais univariadas e foi estendida por Stephens (1997) para o caso multivariado. Existem muitas parametrizações para a distribuição Wishart. Aqui escolhemos a seguinte parametrização: se $\mathbf{X} \sim \text{Wish}_p(r, \mathbf{S})$, então a densidade de \mathbf{X} é dada por

$$f(\mathbf{X}|r, \mathbf{S}) = \frac{\det(\mathbf{S})^r}{\Gamma_p(r)} \det(\mathbf{X})^{r-(p+1)/2} \exp\{-\text{tr}(\mathbf{S}\mathbf{X})\},$$

onde \mathbf{X} é positiva definida e

$$\Gamma_p(r) = \pi^{p(p-1)/4} \prod_{i=1}^p \Gamma\left(\frac{2r+1-i}{2}\right)$$

é a função Gama generalizada.

Quanto ao parâmetro \mathbf{v} , cada modelo específico da família SMSN pode ser descrito de uma maneira diferente. Por exemplo, existem várias sugestões para estimar os graus de liberdade desconhecidos do modelo t de Student. Para uma discussão, veja Fonseca *et al.* (2008) e Garay *et al.* (2015). Para os casos das distribuições t assimétrica e slash assimétrica, consideramos a seguinte estrutura,

$$\mathbf{v} \sim \exp(\lambda), \quad \lambda \sim \text{Uniform}(\lambda_0, \lambda_1),$$

onde $0 < \lambda_0 < \lambda_1$ são hiperparâmetros. Para a normal contaminada assimétrica, consideramos $\rho \sim \text{Beta}(\rho_0, \rho_1)$ e $\tau \sim \text{Beta}(\tau_0, \tau_1)$, onde ρ_0, ρ_1, τ_0 e τ_1 são constantes conhecidas e positivas. Essa escolha de prioris tem sido usada com sucesso em outros trabalhos, como em Cabral *et al.* (2012).

Assumindo independência a priori entre os parâmetros, a especificação da priori completa é dada por

$$\pi(\boldsymbol{\phi}, \boldsymbol{\Psi}, \mathbf{D}, \mathbf{v}, \lambda) = \pi(\boldsymbol{\phi})\pi(\mathbf{v}|\lambda)\pi(\lambda)\pi(\boldsymbol{\Psi}|\mathbf{D})\pi(\mathbf{D}),$$

onde $\boldsymbol{\Psi} = \boldsymbol{\Gamma}^{-1}$ (retirando λ nos casos simétricos).

Capítulo 4

O Algoritmo MCMC

O objetivo deste capítulo é propor um algoritmo para estimação Bayesiana dos parâmetros nos modelos da família SMSN com dados faltantes. Para isso, a representação em dados aumentados dada em (3.1)–(3.3) tem papel fundamental.

4.1 O Amostrador de Gibbs

Primeiramente vamos considerar o caso t de Student assimétrico. A distribuição condicional completa de $T_i | \mathbf{y}_i^o, u_i, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}$ é dada pela Proposição 2, item (iii), ou seja

$$T_i | \mathbf{y}_i^o, u_i, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma} \sim \text{TN} \left(\frac{(\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)}{1 + (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o}, \frac{1}{u_i (1 + (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Gamma}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o)}, (0, \infty) \right).$$

Apesar desta estratégia ser correta do ponto de vista teórico, um procedimento mais eficiente pode ser obtido se considerarmos amostras geradas a partir de distribuições marginais parciais, integrando algumas variáveis latentes presentes no esquema de dados aumentados. Este procedimento é denominado *collapsing* principle, veja Liu (1994) para uma discussão detalhada. No caso da nossa proposta, isso é possível utilizando a Proposição 2, item (iv). Então, obtemos amostras de T_i a partir da distribuição

$$\text{Tt} \left((\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)' (\boldsymbol{\Omega}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o, \frac{(d_{\boldsymbol{\Omega}_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + \nu)(1 - (\boldsymbol{\Delta}_i^o)' (\boldsymbol{\Omega}_i^{oo})^{-1} \boldsymbol{\Delta}_i^o)}{\nu + p_i^o}, \nu + p_i^o, (0, \infty) \right).$$

Observe que, por (2.5), $\boldsymbol{\Omega}_i^{oo} = \boldsymbol{\Gamma}_i^{oo} + \boldsymbol{\Delta}_i^o (\boldsymbol{\Delta}_i^o)'$.

Para gerar amostras de U_i , consideramos a distribuição condicional completa de

$U_i | \mathbf{y}_i^o, t_i, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}, \nu$. Esta distribuição pode ser obtida usando a demonstração da Proposição 2, item (iv). Especificamente, observe que, pela equação (3.8), esta distribuição é

$$\text{Gamma} \left(\frac{p_i^o + \nu + 1}{2}, \frac{d_{\boldsymbol{\Omega}_i^{oo}}(\mathbf{y}_i^o, \boldsymbol{\mu}_i^o) + (t_i - (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)'(\boldsymbol{\Omega}_i^{oo})^{-1}\boldsymbol{\Delta}_i^o)^2(1 - (\boldsymbol{\Delta}_i^o)'(\boldsymbol{\Omega}_i^{oo})^{-1}\boldsymbol{\Delta}_i^o)^{-1} + \nu}{2} \right).$$

Para cada $i = 1, \dots, n$, o vetor com dados faltantes \mathbf{Y}_i^m é visto como um vetor latente em nossa estrutura de dados aumentados. Assim, a distribuição condicional completa deste vetor pode ser obtida com o objetivo de geramos amostras que serão imputadas no lugar dos valores faltantes. Esta distribuição pode ser obtida facilmente considerando da Proposição 2, item (i), e é dada por

$$\mathbf{Y}_i^m | \mathbf{y}_i^o, t_i, u_i, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma} \sim N_{p-p_i^o}(\boldsymbol{\mu}_i^m + \boldsymbol{\Delta}_i^m t_i + \boldsymbol{\Gamma}_i^{mo}(\boldsymbol{\Gamma}_i^{oo})^{-1}(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o - \boldsymbol{\Delta}_i^o t_i), u_i^{-1} \boldsymbol{\Gamma}_i^{mm.o}),$$

onde $\boldsymbol{\Gamma}_i^{mm.o} = \boldsymbol{\Gamma}_i^{mm} - \boldsymbol{\Gamma}_i^{mo}(\boldsymbol{\Gamma}_i^{oo})^{-1}\boldsymbol{\Gamma}_i^{om}$. Seja $\mathbf{y}_i^{m\bullet}$ a amostra gerada para o indivíduo i neste passo do algoritmo, $i = 1, \dots, n$. Para dar continuidade ao processo completamos o vetor de amostras para o indivíduo i , imputando $\mathbf{y}_i^{m\bullet}$ no lugar de \mathbf{y}_i^m , obtendo o vetor $\mathbf{y}_i = (\mathbf{y}_i^o, \mathbf{y}_i^{m\bullet})$.

Para obter a condicional completa do vetor $\boldsymbol{\phi} = (\boldsymbol{\mu}', \boldsymbol{\Delta}')'$, relembremos que, por (3.1), dados $u_i, t_i, \boldsymbol{\mu}, \boldsymbol{\Delta}$ e $\boldsymbol{\Gamma}$, temos

$$\mathbf{Y}_i^* \sim N_p(\mathbf{W}_i \boldsymbol{\phi}, \boldsymbol{\Gamma}), \quad (4.1)$$

onde $\mathbf{Y}_i^* = u_i^{1/2} \mathbf{Y}_i$ e $\mathbf{W}_i = \begin{bmatrix} u_i^{1/2} \mathbf{I}_p & u_i^{1/2} t_i \mathbf{I}_p \end{bmatrix}$, $i = 1, \dots, n$. Então, é simples obter a distribuição condicional completa de $\boldsymbol{\phi}$, desde que (4.1) é um modelo de regressão normal com vetor de parâmetros $\boldsymbol{\phi}$, com uma distribuição a priori dada por (3.11). Assim,

$$\boldsymbol{\phi} | \mathbf{y}, \mathbf{u}, \mathbf{t}, \boldsymbol{\Gamma} \sim N_{2p} \left(\boldsymbol{\Sigma}_\phi \left(\sum_{i=1}^n \mathbf{W}_i' \boldsymbol{\Gamma}^{-1} \mathbf{y}_i^* + \mathbf{B}^{-1} \mathbf{a} \right), \boldsymbol{\Sigma}_\phi \right),$$

onde $\boldsymbol{\Sigma}_\phi = (\sum_{i=1}^n \mathbf{W}_i' \boldsymbol{\Gamma}^{-1} \mathbf{W}_i + \mathbf{B}^{-1})^{-1}$, $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$, $\mathbf{t} = \{t_1, \dots, t_n\}$ e $\mathbf{u} = \{u_1, \dots, u_n\}$.

Em relação à inversa da matriz de dispersão $\boldsymbol{\Psi} = \boldsymbol{\Gamma}^{-1}$, a condicional completa é

dada por

$$\begin{aligned}
\pi(\Psi|\mathbf{y}, \mathbf{t}, \mathbf{D}, \mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\Delta}) &\propto \pi(\mathbf{y}|\Psi, \mathbf{t}, \mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\Delta})\pi(\Psi|\mathbf{D}) \\
&= \det(\Psi)^{n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n u_i (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}t_i)' \Psi (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}t_i) \right\} \\
&\quad \times \det(\Psi)^{c-(p+1)/2} \exp \{ -\text{tr}(\mathbf{D}\Psi) \} \\
&= \det(\Psi)^{(n/2)+c-(p+1)/2} \\
&\quad \times \exp \left\{ -\text{tr} \left(\frac{1}{2} \sum_{i=1}^n u_i (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}t_i)(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}t_i)' + \mathbf{D} \right) \Psi \right\},
\end{aligned}$$

implicando em

$$\Psi|\mathbf{y}, \mathbf{t}, \mathbf{D}, \mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\Delta} \sim \text{Wish}_p \left(\frac{n}{2} + c, \frac{1}{2} \sum_{i=1}^n u_i (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}t_i)(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}t_i)' + \mathbf{D} \right).$$

A distribuição condicional completa de \mathbf{D} é dada por

$$\mathbf{D} \sim \text{Wish}_p(c + e, \mathbf{F} + \Psi).$$

A distribuição condicional completa de λ é TGama($2, \nu, (\lambda_0, \lambda_1)$), que é uma distribuição Gamma truncada no intervalo (λ_0, λ_1) .

A distribuição condicional completa de ν , $\pi(\nu|\mathbf{u})$ não tem uma forma conhecida, e uma alternativa é gerar amostras a partir desta distribuição usando passos de Metropolis-Hastings paths. No entanto, como comentado anteriormente, uma alternativa mais eficiente é utilizar o *collapsing principle*, gerando amostras a partir da distribuição condicional marginalizada $\pi(\nu|\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma})$, que é obtida integrando a verossimilhança aumentada (3.1)–(3.3) em \mathbf{u} e \mathbf{t} . Este procedimento produz, em geral, convergência mais rápida e amostras menos correlacionadas do que obter amostras a partir de $\pi(\nu|\mathbf{u})$. Assim, geramos amostras de ν a partir da distribuição de

$$\pi(\nu|\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}) \propto \exp \{ -\lambda \nu \} \prod_{i=1}^n \text{St}_p(\mathbf{y}_i|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \nu), \quad (4.2)$$

onde $\text{St}_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \nu)$ denota a densidade t de Student assimétrica. Dado um valor amostral $\nu^{(t-1)}$ obtido no estágio $t - 1$ do algoritmo, geramos uma observação candidata ν^c a partir da distribuição lognormal $\text{LN}(\nu^{(t-1)}, \sigma_\nu^2)$. Esta proposta é aceita com probabilidade

$$Q = \min \left\{ \frac{\pi(v^c | \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}) v^c}{\pi(v^{(t-1)} | \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}) v^{(t-1)}}, 1 \right\}. \quad (4.3)$$

Agora vamos considerar o caso slash assimétrico. Então, $\pi(u_i | v)$ é uma densidade beta com parâmetros v e 1, de tal maneira que

$$\begin{aligned} \pi(u_i | \mathbf{y}_i^o, t_i, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}, v) &\propto \pi(\mathbf{y}_i^o | u_i, t_i, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}) \pi(t_i | u_i) \pi(u_i | v) \\ &\propto u_i^{p_i^o/2} \exp \left\{ -\frac{u_i}{2} (\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i))' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i)) \right\} \\ &\quad \times u_i^{1/2} \exp \left\{ -\frac{u_i}{2} t_i^2 \right\} \\ &\quad \times u_i^{v-1} \mathbb{I}_{(0,1)}(u_i) \\ &= u_i^{(p_i^o + 2v + 1)/2 - 1} \\ &\quad \times \exp \left\{ -\frac{u_i}{2} [(\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i))' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i)) + t_i^2] \right\} \\ &\quad \times \mathbb{I}_{(0,1)}(u_i). \end{aligned}$$

Assim, temos a seguinte distribuição condicional completa

$$U_i | \mathbf{y}_i^o, t_i, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}, v \sim \text{TGamma} \left(\frac{p_i^o + 2v + 1}{2}, \frac{(\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i))' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i)) + t_i^2}{2}, (0, 1) \right).$$

A distribuição condicional completa de λ é, como anteriormente, $\text{TGamma}(2, v, (\lambda_0, \lambda_1))$. Amostras da posteriori de v são obtidas a partir da distribuição

$$\text{Gamma} \left(n + 1, \lambda - \sum_{i=1}^n \log u_i \right),$$

que é obtida por conjugação usando o fato que $u_i \sim \text{Beta}(v, 1)$ em conjunto com a especificação a priori para v .

Agora vamos considerar o caso normal contaminado assimétrico. Então, é imediato provar que a distribuição condicional completa de u_i é discreta tomando os valores τ ,

com probabilidade $\eta_i/(\eta_i + \zeta_i)$, e 1, com probabilidade $\zeta_i/(\eta_i + \zeta_i)$, onde

$$\eta_i = \rho \tau^{(p_i^o+1)/2} \exp \left\{ -\frac{\tau}{2} [(\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i))' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i)) + t_i^2] \right\}, \text{ e}$$

$$\zeta_i = (1 - \rho) \exp \left\{ -\frac{1}{2} [(\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i))' (\boldsymbol{\Gamma}_i^{oo})^{-1} (\mathbf{y}_i^o - (\boldsymbol{\mu}_i^o + \boldsymbol{\Delta}_i^o t_i)) + t_i^2] \right\}.$$

Para obter a distribuição condicional completa de ρ , observemos primeiramente que

$$\pi(u_i | \rho, \tau) = (1 - \rho)^{(u_i - \tau)/(1 - \tau)} \rho^{(1 - u_i)/(1 - \tau)}.$$

Lembrando que a distribuição a priori de ρ é Beta(ρ_0, ρ_1), temos

$$\begin{aligned} \pi(\rho | \mathbf{u}, \tau) &\propto \pi(\mathbf{u} | \rho) \pi(\rho) \\ &\propto (1 - \rho)^{(\sum_{i=1}^n u_i - n\tau)/(1 - \tau) + \rho_1 - 1} \rho^{(n - \sum_{i=1}^n u_i)/(1 - \tau) + \rho_0 - 1}, \end{aligned}$$

implicando que a distribuição condicional completa de ρ é Beta com parâmetros $(n - \sum_{i=1}^n u_i)/(1 - \tau) + \rho_0$ e $(\sum_{i=1}^n u_i - n\tau)/(1 - \tau) + \rho_1$. É possível mostrar que o primeiro parâmetro é igual a ρ_0 somado com o número de u_i s iguais a τ .

Em relação ao parâmetro τ , Liu (1996) notou que, se tomarmos amostras a partir da sua condicional completa, então τ tipicamente ficará concentrado em torno de seu valor inicial indefinidamente. Para evitar isto, integramos a verossimilhança aumentada o máximo possível e consideramos a distribuição condicional completa marginal de τ , dada por

$$g(\tau | \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \rho, \mathbf{y}) \propto \tau^{\tau_0 - 1} (1 - \tau)^{\tau_1 - 1} \prod_{i=1}^n \text{SCN}_p(\mathbf{y}_i | \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \tau, \rho).$$

Ao invés de considermos amostras a partir desta distribuição, consideramos a transformação $\bar{\tau} = \tau/(1 - \tau)$, que é uma função 1-1 entre o intervalo $(0, 1)$ e os reais positivos \mathbb{R}^+ . Então, tomamos amostras a partir da distribuição de $\bar{\tau}$, dada por

$$\pi(\bar{\tau} | \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \rho, \mathbf{y}) = g(\bar{\tau}/(1 + \bar{\tau}) | \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \rho, \mathbf{y}) / (1 + \bar{\tau})^2.$$

Para fazer isto, utilizamos um passo de Metropolis-Hastings com propostas a partir de uma distribuição lognormal, seguindo os mesmos passos utilizados no caso t de Student assimétrico, veja (4.3). Então, para reverter a transformação, usamos $\tau = \bar{\tau}/(1 + \bar{\tau})$, obtendo a amostra desejada para τ .

Capítulo 5

Comparação de modelos

Considere o problema de comparar vários modelos da família SMSN, com diferentes proporções de dados faltantes. Como mensurar qual o melhor modelo para um problema real ou simulado? Na literatura existem muitos critérios de comparação de modelos, tanto no contexto frequentista, como o AIC (Akaike, 1974) e o BIC (Schwarz, 1978) quanto no contexto Bayesiano, como o DIC (Spiegelhalter *et al.*, 2002) e o WAIC (Watanabe, 2010), que são as opções que utilizaremos neste trabalho.

5.1 DIC

Como primeira proposta de critério de seleção de modelos temos o DIC (deviance information criterion) observado, que vamos denotar por DIC_{obs} . Ao contrário do DIC tradicional, proposto por (Spiegelhalter *et al.*, 2002), o DIC_{obs} utiliza em sua formulação a verossimilhança integrada, ou seja, a verossimilhança obtida integrando as variáveis latentes. Como comentado por Li *et al.* (2012), critérios como o DIC usual não devem ser utilizados no contexto de dados aumentados, pois a verossimilhança dos dados completos não é regular, o que invalida os argumentos assintóticos que são necessários para justificá-lo. O cálculo do DIC_{obs} é baseado nas amostras a posteriori MCMC, sendo uma versão modificada do DIC usual. Veja os comentários na Seção 3.1 de Celeux *et al.* (2006), onde o DIC_{obs} é denotado por DIC_3 .

Para a amostra observada $\mathbf{y}^o = \{\mathbf{y}_1^o, \dots, \mathbf{y}_n^o\}$, DIC observado é dado por

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

onde Θ é o vetor com os parâmetros em foco,

$$\overline{D(\Theta)} = -2E\{\log[L(\mathbf{y}^o|\Theta)]|\mathbf{y}\} \text{ e } L(\mathbf{y}^o|\Theta) = \prod_{i=1}^n \text{SMSN}_{p_i^o}(\mathbf{y}_i^o|\Theta_i^o), \quad (5.1)$$

$\Theta_i^o = (\boldsymbol{\mu}_i^o, \boldsymbol{\Delta}_i^o, \boldsymbol{\Gamma}_i^{oo}, \mathbf{v})$ contém as correspondentes partições do vetor de locação $\boldsymbol{\mu}$, do vetor de forma $\boldsymbol{\Delta}$, da matriz de dispersão $\boldsymbol{\Gamma}$, e o parâmetro do fator de escala \mathbf{v} . A função de verossimilhança para o indivíduo i , $L(\mathbf{y}_i^o|\Theta_i^o) = \text{SMSN}_{p_i^o}(\mathbf{y}_i^o|\Theta_i^o)$, pode ser calculada utilizando a Proposição 3. Observe que

$$E\{\log[L(\mathbf{y}^o|\Theta)]|\mathbf{y}\} = \int \log[L(\mathbf{y}^o|\Theta)]\pi(\Theta|\mathbf{y})d\Theta$$

é a esperança posteriori de $\log[L(\mathbf{y}^o|\Theta)]$. τ_D é uma medida do número efetivo de parâmetros no modelo, dada por

$$\tau_D = \overline{D(\Theta)} + 2p(\mathbf{y}^o),$$

onde

$$p(\mathbf{y}^o) = \log(E[L(\mathbf{y}^o|\Theta)|\mathbf{y}]) = \sum_{i=1}^n \log E[L(\mathbf{y}_i^o|\Theta_i^o)|\mathbf{y}] = \sum_{i=1}^n \log \int L(\mathbf{y}_i^o|\Theta_i^o)\pi(\Theta|\mathbf{y})d\Theta$$

é o logaritmo da densidade preditiva a posteriori.

Assim, podemos escrever

$$\text{DIC}_{\text{obs}} = 2\overline{D(\Theta)} + 2p(\mathbf{y}^o). \quad (5.2)$$

Seja $\Theta^{(l)}$ a amostra MCMC gerada na iteração l do algoritmo, $l = 1, \dots, m$. O primeiro termo nesta expressão pode ser aproximado por

$$\overline{D} = -\frac{2}{m} \sum_{l=1}^m \sum_{i=1}^n \log [L(\mathbf{y}_i^o|\Theta^{(l)})]. \quad (5.3)$$

Como recomendado por Celeux *et al.* (2006), uma estimativa MCMC para o valor da densidade preditiva a posteriori $p(\mathbf{y}^o)$ em (5.2) é dada por

$$\widehat{p(\mathbf{y}^o)} = 2 \sum_{i=1}^n \log \widehat{p(\mathbf{y}_i^o)}, \quad \text{com} \quad \widehat{p(\mathbf{y}_i^o)} = \frac{1}{m} \sum_{l=1}^m L(\mathbf{y}_i^o|\Theta^{(l)}). \quad (5.4)$$

Com o propósito de comparar diferentes modelos, aquele com menor DIC_{obs} será mais

adequado.

5.2 WAIC

O WAIC, do inglês *Watanabe-Akaike information criterion*, é um critério para a seleção de modelos introduzido por Watanabe (2010) considera a qualidade e complexidade do ajuste, além de ser assintoticamente equivalente à perda por validação cruzada de Bayes.

Basicamente, o WAIC é $p(\mathbf{y}^o)$ mais uma correção para o número efetivo de parâmetros. Existem duas formas diferentes para calcular esta correção, e ambas podem ser vistas como uma aproximação ao método de validação cruzada, como discutido em Gelman *et al.* (2014). Elas são dadas por

$$\rho_{\text{WAIC}_1} = 2p(\mathbf{y}^o) + \overline{D(\Theta)} \text{ e } \rho_{\text{WAIC}_2} = \sum_{i=1}^n \text{Var}[\log L(\mathbf{y}_i^o | \Theta_i^o) | \mathbf{y}],$$

onde $\overline{D(\Theta)}$ é dado em (5.1). Temos então as seguintes versões para o WAIC:

$$\text{WAIC}_k = 2\rho_{\text{WAIC}_k} - 2p(\mathbf{y}^o), \quad k = 1, 2.$$

Observe que o valor de WAIC_1 é igual ao valor do DIC_{obs} . Então, neste trabalho, consideraremos somente o WAIC_2 . Para a comparação de modelos, aquele que apresentar menor WAIC é o escolhido como melhor modelo.

Capítulo 6

Aplicação com Dados Simulados e Reais

Com o propósito de avaliar a performance da metodologia proposta neste trabalho, dois estudos de simulação foram delineados. Todos os procedimentos computacionais foram implementados utilizando o software R (R Core Team, 2016). O primeiro estudo analisa o comportamento do vício e do erro quadrático médio das estimativas MCMC para os diferentes modelos SMSN como função da taxa de observações faltantes. O segundo estudo compara os diferentes modelos SMSN, via critérios de seleção, também levando-se em consideração a taxa de observações faltantes.

6.1 Estudo de Simulação I

O objetivo deste estudo de simulação é verificar o comportamento do vício e o do erro quadrático médio (EQM) das estimativas MCMC obtidas via o algoritmo proposto no Capítulo 4 como função das taxas de observações faltantes. Com este objetivo geramos amostras de tamanho $n = 100$ a partir das distribuições normal, normal assimétrica, t de Student e t de Student assimétrica. Os parâmetros foram fixados em $\boldsymbol{\mu} = (1, 2, 3)$, $\boldsymbol{\Gamma} = \text{diag}\{0.1, 0.2, 0.3\}$, $\boldsymbol{\Delta} = (-2, 0, 2)$ (para as distribuições normal assimétrica e t assimétrica) e $\nu = 3$ (para as distribuições t e t assimétrica). As taxas de observações faltantes foram fixadas em 3% (leve), 10% (moderada) e 30% (severa). 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 proporção de observações faltantes, foram simulados 100 conjuntos de

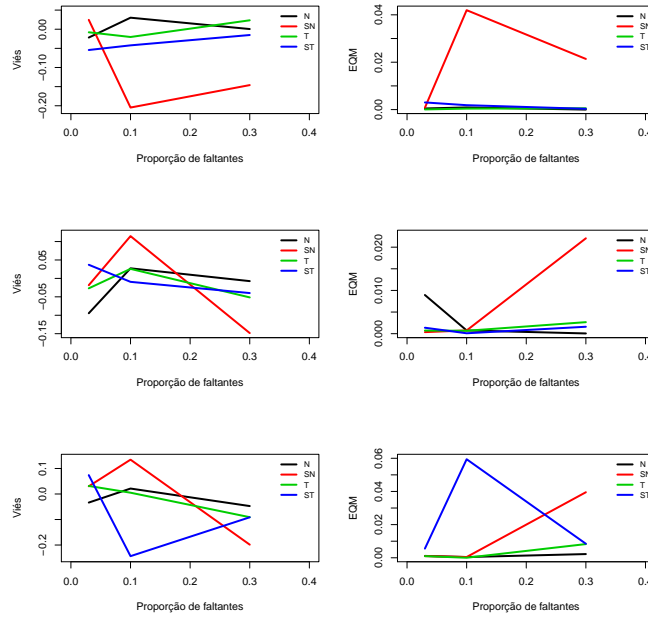


Figura 6.1: Viés e EQM das estimativas de μ_1 , μ_2 e μ_3 para os dados simulados. Primeira coluna, viés. Segunda coluna, EQM.

dados. Para cada um deles, ajustamos os modelos citados anteriormente e armazenamos as estimativas MCMC dos parâmetros. Então, calculamos o vício e o EQM para as estimativas em cada modelo. Para obter as estimativas partir de um determinado conjunto de dados fixado, foram consideradas 5000 amostras MCMC, descontando-se as primeiras 1000 como um período de *burn-in*. Assim, para o parâmetro μ_j , $j = 1, 2, 3$, definimos o vício e o EQM estimados por

$$\text{Vício} = \frac{1}{100} \sum_{i=1}^{100} (\hat{\mu}_j^{(i)} - \mu_j), \quad \text{MSE} = \frac{1}{100} \sum_{i=1}^{100} (\hat{\mu}_j^{(i)} - \mu_j)^2$$

onde $\hat{\mu}_j^{(i)}$ é a estimativa MCMC de μ_j para o i -ésimo conjunto de dados simulado, $j = 1, 2, 3$, $i = 1, 2, \dots, 100$. Definições análogas de vício e EQM valem para os outros parâmetros.

A Tabela 6.1 e as Figuras 6.1, 6.2, 6.3 e 6.4 contêm os resultados de viés e EQM para os dados simulados. Observe que, em geral, os valores são pequenos para todas as combinações de parâmetros e modelos, e que não há um padrão de variação bem definido nem para o viés, e nem para o EQM como função da proporção dos dados faltantes.

Tabela 6.1: Viés e EQM para as estimativas para diferentes percentuais de dados faltantes e diferentes modelos.

Parâmetro	%	Viés				EQM			
		<i>N</i>	<i>SN</i>	<i>T</i>	<i>ST</i>	<i>N</i>	<i>SN</i>	<i>T</i>	<i>ST</i>
μ_1	3%	-0.0217340	0.0244784	-0.0080207	-0.0542924	0.0004730	0.0007941	0.0000660	0.0030172
	10%	0.0301004	-0.2046788	-0.0203018	-0.0421789	0.0009069	0.0419651	0.0004139	0.0018564
	30%	0.0004602	-0.1460148	0.0233225	-0.0151663	0.0000015	0.0213815	0.0005472	0.0003312
μ_2	3%	-0.0945613	-0.0184041	-0.0269431	0.0369741	0.0089423	0.0003597	0.0007269	0.0013773
	10%	0.0273515	0.1147328	0.0258900	-0.0093459	0.0007485	0.0007485	0.0006712	0.0000971
	30%	-0.0073824	-0.1483527	-0.0515794	-0.0397948	0.0000553	0.0220466	0.0026642	0.0016042
μ_3	3%	-0.0335563	0.0302460	0.0315613	0.0739337	0.0011262	0.0010181	0.0009972	0.0055326
	10%	0.0213240	0.1346062	0.0050781	-0.2436013	0.0004550	0.0004550	0.0000271	0.0594191
	30%	-0.0473052	-0.1985765	-0.0908761	-0.0912534	0.0022389	0.0395229	0.0082608	0.0084526
Δ_1	3%	-	-0.0506307	-	-0.2902104	-	0.0034006	-	0.0868588
	10%	-	0.1402186	-	0.1247990	-	0.0198803	-	0.1247990
	30%	-	0.1834540	-	-0.1516041	-	0.0342481	-	0.0249360
Δ_2	3%	-	-0.0275607	-	-0.0005581	-	0.0007945	-	0.0000134
	10%	-	-0.0302051	-	0.0272431	-	0.0009357	-	0.0272431
	30%	-	0.2051649	-	0.0464010	-	0.0421727	-	0.0021754
Δ_3	3%	-	0.0753997	-	0.2291213	-	0.0064808	-	0.0548850
	10%	-	0.0158893	-	0.2947251	-	0.0006406	-	0.2947251
	30%	-	0.2125993	-	0.3328450	-	0.0460559	-	0.1132737
ν	3%	-	-	0.1717294	-0.6735358	-	-	0.0310339	0.4553955
	10%	-	-	2.9852195	0.6769982	-	-	8.9201325	0.4644901
	30%	-	-	0.0786169	2.2502669	-	-	0.0069760	5.0881793
Γ_{11}	3%	-0.0077612	0.0935716	0.0166499	0.0991184	0.0000603	0.0089511	0.0002784	0.0099219
	10%	-0.0020849	0.1137680	0.0246432	0.2738700	0.0000044	0.0129756	0.0006079	0.0751853
	30%	0.0167103	0.0939310	0.0549450	0.3607427	0.0002794	0.0088549	0.0030217	0.1304940
Γ_{12}	3%	0.0008696	0.0043586	-0.0208228	-0.0646031	0.0000008	0.0000342	0.0004338	0.0041889
	10%	0.0219387	-0.0046974	0.0190298	-0.0176999	0.0004814	0.0000327	0.0003623	0.0003225
	30%	-0.0227561	0.0222016	0.0011762	-0.0926485	0.0005183	0.0005185	0.0000021	0.0086069
Γ_{13}	3%	0.0192430	-0.0012594	-0.0326425	-0.0823008	0.0003704	0.0000947	0.0010658	0.0068768
	10%	-0.0041918	-0.0574592	-0.0240747	-0.0676453	0.0000177	0.0033588	0.0005800	0.0046509
	30%	0.0215532	-0.0387317	0.0254187	-0.0276274	0.0004656	0.0015274	0.0006472	0.0009981
Γ_{22}	3%	-0.0008625	-0.0428921	0.0312512	0.0258223	0.0000009	0.0018399	0.0009811	0.0006816
	10%	0.0012194	0.0189730	0.0675318	0.0077312	0.0000017	0.0003605	0.0045637	0.0000664
	30%	0.0191587	-0.0028385	0.0625465	0.0757823	0.0003674	0.0000093	0.0039177	0.0057592
Γ_{23}	3%	0.0090013	-0.0353056	0.0208020	0.0455016	0.0000812	0.0012621	0.0004331	0.0020851
	10%	0.0109003	0.0513786	0.0522612	-0.0223145	0.0001190	0.0026497	0.0027321	0.0005101
	30%	-0.0036242	-0.0297612	0.0520818	-0.0828143	0.0000136	0.0009231	0.0027148	0.0068758
Γ_{33}	3%	0.0111860	-0.1348016	0.0000518	-0.0443161	0.0001257	0.0182253	0.0000066	0.0021095
	10%	-0.0526070	0.0187728	0.0767872	-0.0209646	0.0027678	0.0005821	0.0059008	0.0005518
	30%	-0.0541423	-0.1533618	-0.0243476	0.0974018	0.0029321	0.0235544	0.0005995	0.0097160

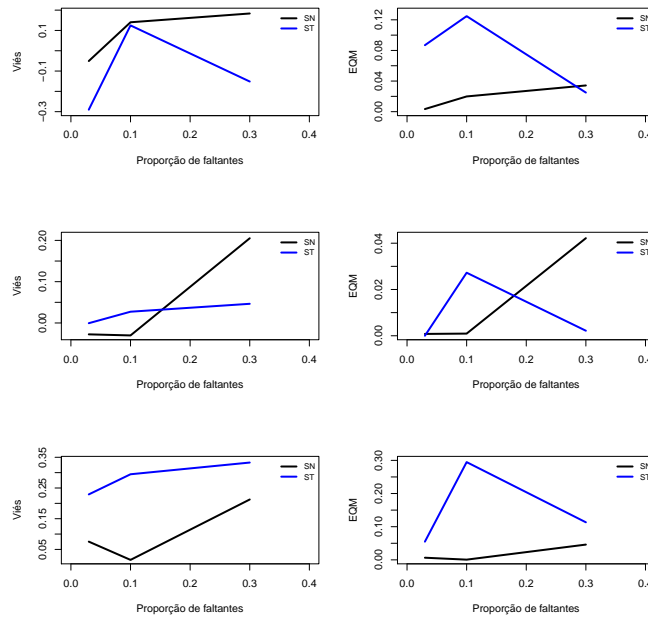


Figura 6.2: Viés e EQM das estimativas de Δ_1 , Δ_2 e Δ_3 para os dados simulados. Primeira linha, viés. Segunda linha, EQM.

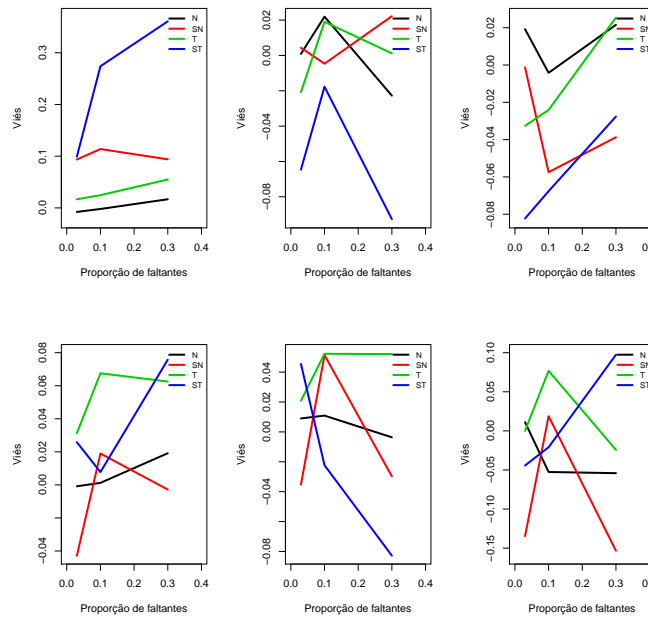


Figura 6.3: Viés para os dados simulados. Primeira linha, Γ_{11} , Γ_{12} e Γ_{13} . Segunda linha, Γ_{22} , Γ_{23} e Γ_{33} .

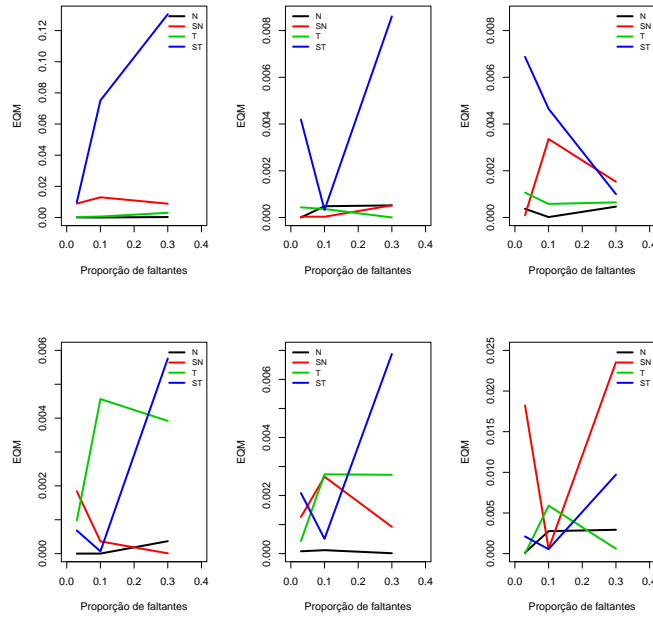


Figura 6.4: EQM para os dados simulados. Primeira linha, Γ_{11} , Γ_{12} e Γ_{13} . Segunda linha, Γ_{22} , Γ_{23} e Γ_{33} .

6.2 Estudo de Simulação II

Neste estudo de simulação o objetivo é comparar a performance de alguns modelos na classe SMSN sob a presença de dados faltantes utilizando os critérios de seleção DIC_{obs} e WAIC definidos no Capítulo 5. Os modelos considerados para este estudo foram: normal, t de Student, slash, normal contaminada, normal assimétrica, t de Student assimétrica, slash assimétrica e normal contaminada assimétrica.

Foram geradas 100 amostras de $n = 60$ de uma distribuição normal contaminada assimétrica com os seguintes parâmetros fixados ao longo de todas as simulações: $\boldsymbol{\mu} = (2, 6, 8)$, $\boldsymbol{\Gamma} = \text{diag}\{0.1, 0.1, 0.1\}$, $\boldsymbol{\Delta} = (2, 2, 2)$, $\tau = 0.1$ e $\rho = 0.8$. Estes valores foram escolhidos com o propósito de garantir uma assimetria acentuada e um elevado grau de contaminação. Consideramos novamente as taxas de observações faltantes 3% (leve), 10% (moderada) e 30% (severa) e uma taxa adicional de 0%. Para obter amostras de τ e ν nos passos de Metropolis-Hastings, 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 conjunto de dados ajustamos os modelos citados anteriormente e estimativas MCMC dos parâmetros. Para os modelos slash e slash assimétrico, por questões de tempo computacional, fixamos o valor do parâmetro do fator de escala em $\nu = 1$. En-

tão, para cada modelo, calculamos o DIC_{obs} e WAIC. A proporção de vezes (dentre 100) em que cada modelo é escolhido por cada critério é apresentada na Tabela 6.2. Nesta tabela podemos observar que ambos os critérios favorecem, em sua maioria, os modelos assimétricos ou os modelos assimétricos com caudas pesadas, não parecendo haver uma influência significativa da proporção de observações faltantes na variação das proporções apresentadas na tabela. O grande valor para a proporção referente à distribuição normal assimétrica pode ser explicado pela definição da densidade da distribuição normal contaminada assimétrica, dada na Seção 2.2.3. Observe a densidade SCN é uma mistura de duas distribuições normais assimétricas, sendo uma com peso $\rho = 0.8$. Então é natural que, para determinadas amostras geradas, uma componente SN seja suficiente para obter um bom ajuste.

Tabela 6.2: Proporção de vezes em que um dos modelos foi melhor do que os demais através dos métodos de comparação DIC e WAIC, para diferentes percentuais de faltantes.

Modelos	0%		3%		10%		30%	
	DIC	WAIC	DIC	WAIC	DIC	WAIC	DIC	WAIC
Normal	0.02	0.02	0.01	0.03	0.02	0.03	0.06	0.05
Normal Assimétrica	0.38	0.32	0.48	0.38	0.26	0.19	0.15	0.16
T	0.04	0.05	0	0	0	0	0	0.06
T Assimétrica	0.21	0.25	0.17	0.2	0.07	0.1	0.06	0.12
Slash	0.02	0	0	0	0	0	0	0
Slash Assimétrica	0.13	0.16	0.13	0.18	0.14	0.19	0.14	0.08
Normal Contaminada	0	0.02	0	0	0.03	0.06	0	0
Normal Contaminada Assimétrica	0.2	0.18	0.21	0.21	0.48	0.43	0.31	0.25

6.3 Análise de Dados Reais – Australian Institute of Sport Data

Cook & Weisberg (1994) consideram observações referentes a 202 atletas Australianos de elite, que faziam os seus treinamentos no Australian Institute of Sport (AIS). Os indivíduos observados eram de várias modalidades esportivas diferentes. Além disso, 100 eram do sexo feminino e 102 do sexo masculino. Várias características consideradas importantes para o estudo foram observadas mas, para efeito desta análise, vamos considerar somente as variáveis LBM (lean body mass, a massa magra corporal) e BMI (body mass index, o índice de massa corporal). Esta análise também foi considerada em trabalhos anteriores na literatura como Azzalini & Dalla Valle (1996), Wang & Genton (2006) e Tan

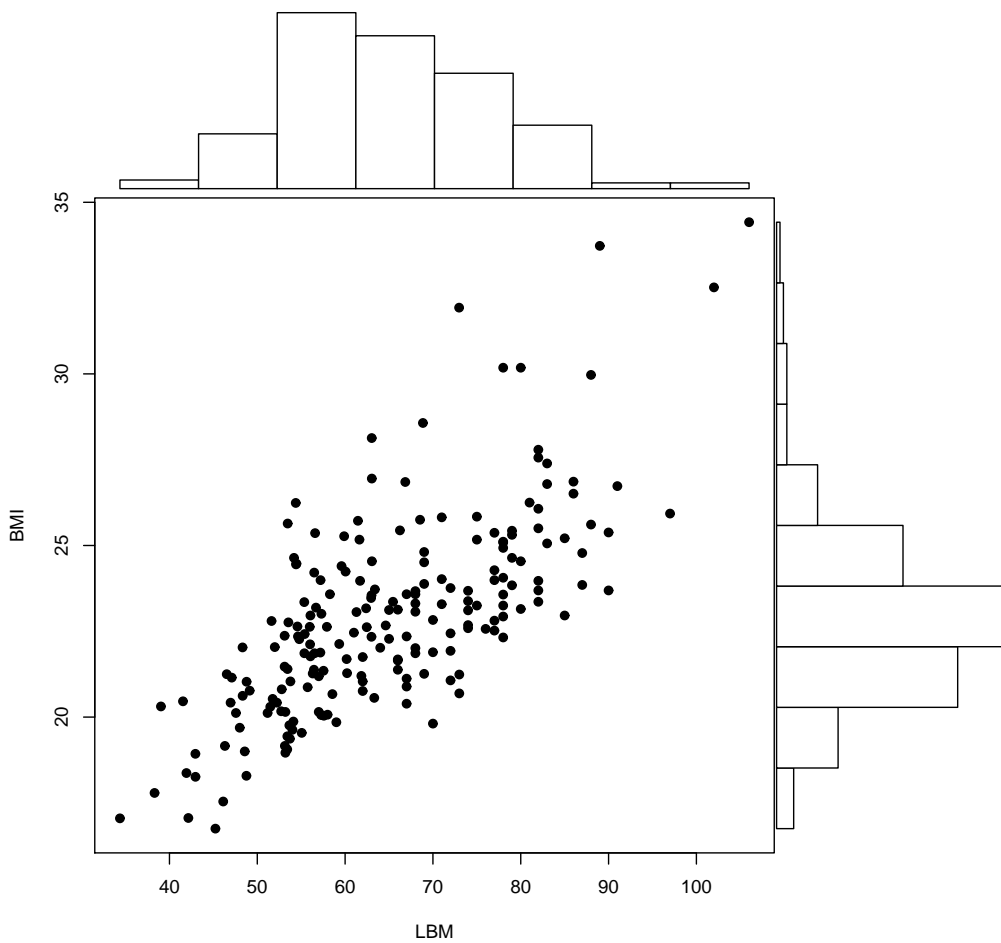


Figura 6.5: Diagrama de dispersão e histogramas das variáveis BMI e LBM.

et al. (2015). A Figura 6.5 apresenta um diagrama de dispersão para estas duas variáveis, mostrando um padrão de assimetria e também observações discrepantes. Os dados originais não contém observações faltantes. Então, para explorar a modelagem proposta neste trabalho, utilizamos o mecanismo gerador de dados faltantes MAR (*missing at random*) com a fim de criar artificialmente e de forma aleatória observações faltantes no conjunto de dados, totalizando 20 observações faltantes que corresponde a aproximadamente 10% do tamanho amostral. Seguindo o que foi feito em Tan *et al.* (2015), antes de ajustarmos os modelos, padronizamos as variáveis.

Na Tabela 6.3 temos os valores dos critérios de comparação para os diversos modelos na família SMSN utilizados para o ajuste dos dados AIS. Dois modelos assimétricos e com caudas pesadas são nitidamente favorecidos pelos critérios em relação aos demais: o modelo slash assimétrico e o modelo t de Student assimétrico. Isso evidencia que estes

modelos são uma alternativa viável aos modelos tradicionais simétricos quando existe a presença de observações faltantes. Na Tabela 6.4 temos as médias a posteriori para os parâmetros sob o modelo t de Student assimétrico, bem como os limites dos intervalos de credibilidade a posteriori. Uma exceção é o parâmetro de graus de liberdade ν : o valor que aparece na Tabela 6.4 é a mediana a posteriori (a distribuição a posteriori de ν aparenta ter um forte grau de assimetria, como veremos a seguir). Observe que o valor para a estimativa de ν aponta para a viabilidade do modelo com caudas pesadas neste caso. Nas Figuras 6.6, 6.7 e 6.8 temos os gráficos para as amostras MCMC das distribuições a posteriori para cada parâmetro a partir do ajuste do modelo t de Student assimétrico.

Tabela 6.3: Critérios de seleção para os modelos da família SMNSN aplicados aos dados AIS.

Modelos	DIC	WAIC
Normal	1027.049	1027.397
Normal Assimétrica	1005.557	1006.788
t de Student	1025.915	1026.190
t de Student Assimétrica	1000.255	1001.129
Slash	1025.525	1062.019
Slash Assimétrica	999.664	1001.472
Normal Contaminada	1025.861	1026.431
Normal Contaminada Assimétrica	1005.282	1013.761

Tabela 6.4: Estimativas dos parâmetros para o conjunto de dados AIS. A sigla dp denota o desvio padrão, LI e LS representam os limites inferior e superior do intervalo de credibilidade a posteriori.

Parâmetro	Estimativa	dp	LI	LS
μ_1	-0.2702	0.1711	-0.5898	0.0850
μ_2	-1.1992	0.1287	-1.4598	-0.9552
Δ_1	0.2898	0.1721	-0.0303	0.6488
Δ_2	1.4526	0.2054	1.0892	1.8612
Γ_{11}	0.7813	0.1218	0.5387	1.0152
Γ_{12}	-0.0655	288.7149	-0.2333	0.0841
Γ_{22}	0.1261	0.0524	0.0400	0.2314
ν	13.991	17.8293	3.8744	47.8684

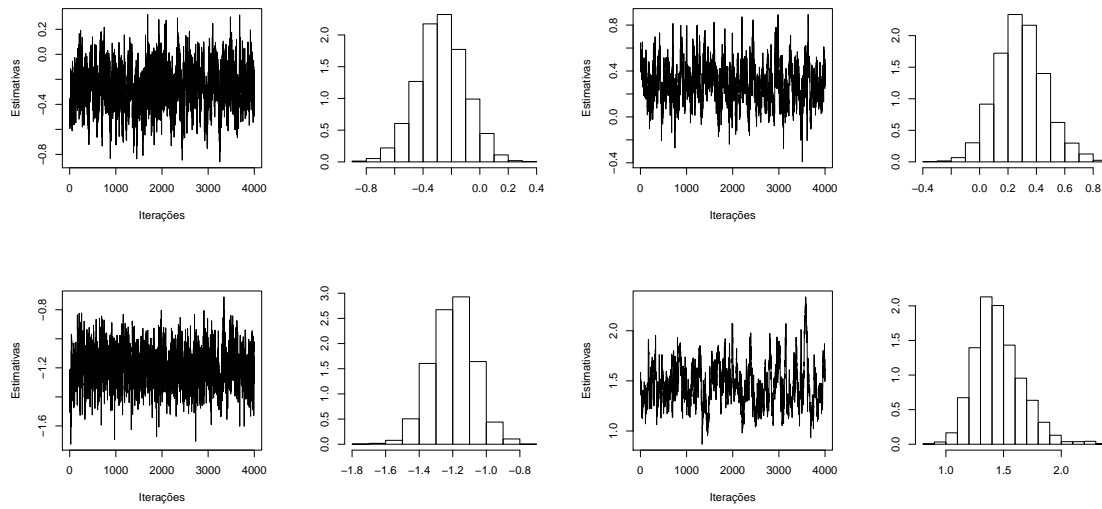


Figura 6.6: Primeira e segunda colunas: traceplot e histograma de μ_1 e μ_2 . Terceira e quarta colunas traceplot e histograma de Δ_1 e Δ_2 .

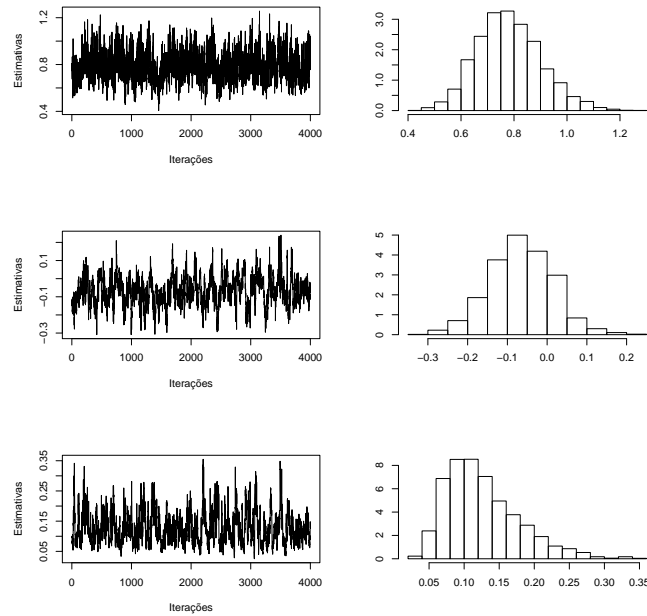


Figura 6.7: Traceplot e histograma de Γ_{11} , Γ_{12} e Γ_{22} .

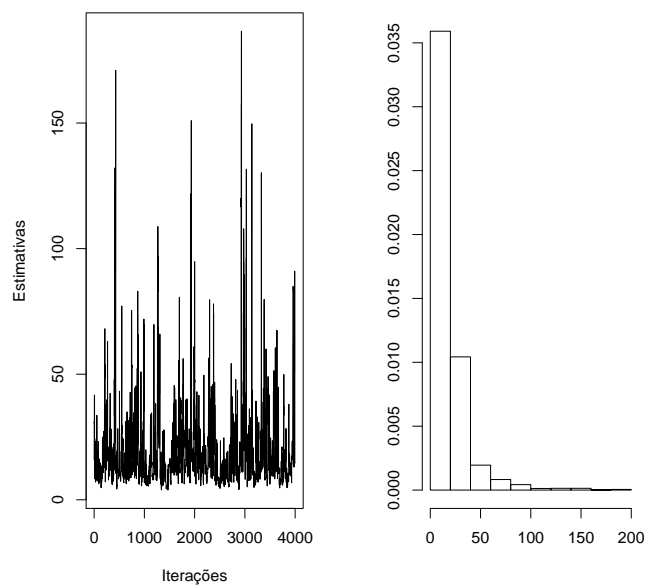


Figura 6.8: Traceplot e histograma de v .

Capítulo 7

Conclusões e Pesquisas Futuras

Neste trabalho propusemos uma metodologia Bayesiana para estimação dos parâmetros e imputação para modelos na família das misturas de escala da normal assimétrica com dados faltantes. Um algoritmo do tipo MCMC foi proposto, e um estudo de simulação foi conduzido com o objetivo de avaliar a variação o vício e o erro quadrático médio das estimativas como função da proporção de observações faltantes. Observou-se que, neste quesito, o método é relativamente robusto em relação a mudanças na proporção de observações faltantes. Também foi conduzido um estudo de simulação para avaliar o ajuste dos diferentes modelos SMSN como função das taxas de observações faltantes, novamente não sendo notadas variações significativas. Uma análise de dados reais foi realizada, e os critérios de seleção considerados apontaram que os modelos que acomodam ao mesmo tempo assimetria e caudas pesadas são os mais adequados em comparação aos seus equivalentes simétricos.

Como proposta para pesquisas futuras, pretendemos estender este trabalho para o caso de misturas finitas de distribuições SMSN com dados faltantes, possivelmente incorporando censuras nas observações.

Referências Bibliográficas

- (2012). The infinite student's t-factor mixture analyzer for robust clustering and classification. *Pattern Recognition*, **45**(12), 4346 – 4357.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.
- Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200–203.
- Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, **36**, 99–102.
- Arellano-Valle, R. B. & Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, **33**, 561–574.
- Arellano-Valle, R. B. & Genton, M. G. (2005). On fundamental skew distributions. *Journal of Multivariate Analysis*, **96**, 93–116.
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.
- Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, **83**(4), 715–726.
- Bayes, C. L. & Branco, M. D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics*, **21**, 141–163.
- Branco, M. D. & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99–113.
- 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.

- Cancho, V. G., Dey, D. K., Lachos, V. H. & Andrade, M. G. (2010). Bayesian non-linear regression models with scale mixtures of skew normal distributions: Estimation and case influence diagnostics. *Computational Statistics and Data Analysis*, **55**, 588–602.
- Celeux, G., Forbes, F., Robert, C. P. & Titterton, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651–674.
- Cook, R. D. & Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.
- Demirtas, H., Freelsa, S. A. & Yucelb, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, **78**, 69–84.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.
- Fonseca, T. C. O., Ferreira, M. A. R. & Migon, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika*, **95**, 325–333.
- Garay, A. M., Bolfarine, H., Lachos, V. H. & Cabral, C. R. B. (2015). Bayesian analysis of censored linear regression models with scale mixtures of normal distributions. *Journal of Applied Statistics*, **42**, 2694–2714.
- Gelman, A., Hwang, J. & A. Vehtari (2014). Understanding predictive information criteria for Bayesian models. *Statistic and Computing*, **24**, 997–1016.
- Gentle, J. E. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer.
- Gupta, A. K. (2003). Multivariate skew t - distribution. *Statistics*, **37**, 359–363.
- Hocking, R. R. & Smith, W. B. (1968). Estimation of parameters in the multivariate normal distribution with missing observations. *Journal of the American Statistical Association*, **63**, 159–173.
- Kotz, S. & Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge.
- Lachos, V. H. & Labra, F. V. (2014). Multivariate skew-normal/independent distributions: properties and inference. *Pro Mathematica*, **28**(56), 11–53.

- Lachos, V. H., Dey, D. K. & Cancho, V. G. (2009a). Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective. *Journal of Statistical Planning and Inference*, **139**, 4098–4110.
- Lachos, V. H., Garibay, V., Labra, F. V. & Aoki, R. (2009b). Robust multivariate measurement error model with skew-normal/independent distributions and Bayesian MCMC implementation. *Statistical Methodology*, **6**, 527–541.
- Lange, K. & Sinsheim, J. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, **2**, 175–198.
- Li, Y., Zeng, T. & Yu, J. (2012). Robust deviance information criterion for latent variable models. Economics and statistics working paper series, SMU.
- Lin, T. I. & Chy, L. T. (2011). Robust statistical modelling using the multivariate skew t distribution with complete and incomplete data. *Statistical Modelling*, **11**, 253–277.
- Lin, T. I., Ho, H. J. & Chen, C. L. (2009). Analysis of multivariate skew normal models with incomplete data. *Journal of Multivariate Analysis*, **100**, 2337–2351.
- Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis*, **53**, 139–158.
- Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association*, **91**, 1219–1227.
- Liu, C. (1999). Efficient ML estimation of the multivariate normal distribution from incomplete data. *Journal of Multivariate Analysis*, **69**, 206–217.
- Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958–966.
- Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.
- Nguyen, T. M. & Wu, Q. J. (2012). Robust student's-t mixture model with spatial constraints and its application in medical image segmentation. *IEEE Transactions on Medical Imaging*, **31**(1), 103–116.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York, second edition.
- Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
- Sahu, S. K., Dey, D. K. & Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *The Canadian Journal of Statistics*, **31**, 129–150.
- 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, Series B*, **64**, 583–639.
- Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. Ph.D. thesis, Magdalen College, Oxford.
- Sun, D. & Sun, X. (2006). Estimation of multivariate normal covariance and precision matrices in a star-shape model with missing data. *Journal of Multivariate Analysis*, **97**, 698–719.
- Tan, F., Tang, Y., H. & Peng (2015). The multivariate slash and skew-slash student t distributions. *Journal of Statistical Distributions and Applications*, **2**, 1–22.
- Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.
- van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton.
- Vidal, I. & Castro, L. M. (2010). Influential observations in the independent student-t measurement error model with weak nondifferential error. *Chilean Journal of Statistics*, **1**(2), 17–34.
- Wang, J. & Genton, M. G. (2006). The multivariate skew-slash distribution. *Journal of Statistical Planning and Inference*, **136**, 209–220.

- 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.
- Yang, F. & Yuan, H. (2016). A non-iterative bayesian sampling algorithm for linear regression models with scale mixtures of normal distributions. *Computational Economics*, pages 1–19.