UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Charles Henrique Delage Moura

Misturas Finitas de Modelos de Regressão Multivariados Assimétricos

Juiz de Fora
2022

Charles Henrique Delage Moura

Misturas Finitas de Modelos de Regressão Multivariados Assimétricos

Trabalho de conclusão de curso apresentado ao Departamento de Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientadora: Prof[a]. Dr[a]. Camila Borelli Zeller

Juiz de Fora

2022

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Moura, Charles Henrique Delage.
Misturas Finitas de Modelos de Regressão Multivariados Assimétricos / Charles Henrique Delage Moura. – 2022.
66 f. : il.

Orientadora: Camila Borelli Zeller
Trabalho de Conclusão de Curso (graduação) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Departamento de Estatística, 2022.

1. algoritmo EM. 2. misturas finitas. 3. modelos de regressão multivariados. 4. distribuições assimétricas. 5.misturas de especialistas. 6. classificação. I. Zeller, Camila Borelli, orient. II. Misturas Finitas de Modelos de Regressão Multivariados Assimétricos.

Charles Henrique Delage Moura

Misturas Finitas de Modelos de Regressão Multivariados Assimétricos

Trabalho de conclusão de curso apresentado ao Departamento de Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de bacharel em Estatística.

Aprovada em 25 de fevereiro de 2022.

BANCA EXAMINADORA

Prof[a]. Dr[a]. Camila Borelli Zeller - Orientadora
Universidade Federal de Juiz de Fora

Prof. Dr. Clécio da Silva Ferreira - Coorientador
Universidade Federal de Juiz de Fora

Prof. Dr. Marcel de Toledo Vieira - Coorientador
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Agradeço a Deus por todos que contribuíram para eu concluir mais essa etapa na vida.

RESUMO

A estimativa usual de modelos de regressão no contexto de misturas finitas é baseada na suposição de normalidade dos erros e, portanto, é sensível a valores atípicos, erros de cauda pesada e/ou erros assimétricos. Neste trabalho são apresentadas duas propostas para lidar com essas questões simultaneamente considerando misturas finitas de modelos de regressão sob distribuições multivariadas de Misturas de Escala *Skew*-Normal (MESN). Essas abordagens permitem modelar dados com grande flexibilidade, acomodando simultaneamente assimetria e caudas pesadas. A principal virtude de considerar modelos de regressão sob a classe MESN é que eles têm uma boa representação hierárquica que permite uma fácil implementação de inferências. As propostas de modelos de regressão estudadas foram o método clássico Misturas Finitas de Modelos de Regressão sob as distribuições MESN (MR-MF-MESN) e o método de Misturas Finitas de Especialistas de Modelos de Regressão sob as distribuições MESN (MoE-MF-MESN). Empregou-se um algoritmo simples do tipo EM para realizar inferência de máxima verossimilhança dos parâmetros dos modelos propostos. Estudos de simulação são apresentados para comparar os dois modelos em relação à classificação das observações e para analisar a convergência das estimativas assintoticamente. Por fim, um conjunto de dados reais é analisado, ilustrando a utilidade dos métodos propostos.

Palavras-chave: algoritmo EM, misturas finitas, modelos de regressão multivariados, distribuições assimétricas, misturas de especialistas, classificação.

ABSTRACT

The usual estimation of regression models in the context of finite mixtures is based on the assumption of normality of errors and, therefore, is sensitive to outliers, heavy tail errors and/or asymmetric errors. In this work two proposals are presented to deal with these issues simultaneously considering finite mixtures of regression models under multivariate distributions of Skew-Normal Mixture Scales (MESN). These approaches allow you to model data with great flexibility while accommodating asymmetry and heavy tails. The main virtue of considering regression models under the MESN class is that they have a good hierarchical representation that allows an easy implementation of inferences. The proposed regression models studied were the classical method Finite Mixtures of Regression Models under the MESN distributions (MR-MF-MESN) and the Finite Mixtures method of Regression Models Experts under the MESN distributions (MoE-MF-MESN). A simple EM-type algorithm was used to perform maximum likelihood inference of the parameters of the proposed models. Simulation studies are presented to compare the two models with respect to the classification of observations and to analyze the convergence of estimates asymptotically. Finally, a real dataset is analyzed, illustrating the usefulness of the proposed methods.

Keywords:EM algorithm, finite mixtures, multivariate regression models, asymmetric distributions, expert mixtures, classification.

Lista de Figuras

Heterogeneidade: no histograma à esquerda, as proporções entre os grupos são desconhecidas; no histograma à direita, as cores indicam as proporções estimadas para cada grupo. 11
Boxplots das estimativas de $\boldsymbol{\beta}$ para o modelo MR-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro. 46
Boxplots das estimativas de $\boldsymbol{\beta}$ para o modelo MoE-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro. 47
Boxplots das estimativas de $\boldsymbol{\sigma}$ para o modelo MR-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro. 47
Boxplots das estimativas de $\boldsymbol{\sigma}$ para o modelo MoE-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro. 48
Boxplots das estimativas de $\boldsymbol{\lambda}$ para o modelo MR-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro. 48
Boxplots das estimativas de $\boldsymbol{\lambda}$ para o modelo MoE-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro. 49
Boxplots das estimativas de $\boldsymbol{p}$ para o modelo MR-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro. 49
Boxplots das estimativas de $\boldsymbol{\alpha}$ para o modelo MoE-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro. 49
EQM-R das médias das estimativas de $\boldsymbol{\beta}_1$ para o modelo MR-MF-MESN. 50
EQM-R das médias das estimativas de $\boldsymbol{\beta}_1$ para o modelo MoE-MF-MESN. 50
EQM-R das médias das estimativas de $\boldsymbol{\beta}_2$ para o modelo MR-MF-MESN. 50
EQM-R das médias das estimativas de $\boldsymbol{\beta}_2$ para o modelo MoE-MF-MESN. 50
EQM-R das médias das estimativas de $\boldsymbol{\sigma}_1^2$ para o modelo MR-MF-MESN. 51
EQM-R das médias das estimativas de $\boldsymbol{\sigma}_1^2$ para o modelo MoE-MF-MESN. 51
EQM-R das médias das estimativas de $\boldsymbol{\sigma}_2^2$ para o modelo MR-MF-MESN. 51
EQM-R das médias das estimativas de $\boldsymbol{\sigma}_2^2$ para o modelo MoE-MF-MESN. 51
EQM-R das médias das estimativas de $\boldsymbol{\lambda}_1$ para o modelo MR-MF-MESN. 52
EQM-R das médias das estimativas de $\boldsymbol{\lambda}_1$ para o modelo MoE-MF-MESN. 52
EQM-R das médias das estimativas de $\boldsymbol{\lambda}_2$ para o modelo MR-MF-MESN. 52
EQM-R das médias das estimativas de $\boldsymbol{\lambda}_2$ para o modelo MoE-MF-MESN. 52
EQM-R das médias das estimativas de $\boldsymbol{p}$ para o modelo MR-MF-MESN. 53
EQM-R das médias das estimativas de $\boldsymbol{\alpha}$ para o modelo MoE-MF-MESN. 53
Aplicação: densidades por Kernel de *science* (y1) e *write* (y2). 54

Aplicação: Gráficos de dispersão dos dados hsb2 (y1 e y2), as linhas de regressão de mistura ajustadas e a classificação pelo ajuste da distribuição *Skew*-normal para MoE-MF-MESN de 2 componentes (grupo 1 = triângulo, grupo 2 = ponto). 56

Lista de Tabelas

Tabela 1 – Recuperação dos Parâmetros: os valores verdadeiros dos parâmetros (Verd) estão entre parênteses. Média e DPA são as respectivas médias e desvios padrão amostrais das estimativas dos parâmetros provenientes do ajuste de uma distribuição *Skew-t* para MR-MF-MESN, com diferentes configurações de tamanho de amostra (n) com base em 500 replicações. 44
Tabela 2 – Recuperação dos Parâmetros: os valores verdadeiros dos parâmetros (Verd) estão entre parênteses. Média e DPA são as respectivas médias e desvios padrão amostrais das estimativas dos parâmetros provenientes do ajuste de uma distribuição *Skew-t* para MoE-MF-MESN, com diferentes configurações de tamanho de amostra (n) com base em 500 replicações. 45
Tabela 3 – Aplicação: Estimativas e erros padrão (EP) das distribuições *Skew*-normal (SN) e *Skew*-t (ST) para os modelos MR-MF-MESN (MR) e MoE-MF-MESN (MoE). Os erros padrão foram calculados via *bootstrap* paramétrico. . . 55
Tabela 4 – Aplicação: Seleção do modelo: log-verossimilhança e BIC 55
Tabela 5 – Aplicação: SN-MoE-MF-MESN: estimativas, erros padrão (EP), p-valor e significância dos coeficientes de regressão $\boldsymbol{\beta}$. Códigos das significâncias, dado o p-valor: $0 = ***$; $0,001 = **$; $0,01 = *$; $0,05 = \cdot$; $0,1 = \text{' '}$; $1 = 1$. . 55

Sumário

1 INTRODUÇÃO 11
1.1 PROPOSTA DO TRABALHO 12
1.2 ORGANIZAÇÃO DO TRABALHO 13
2 A CLASSE DE DISTRIBUIÇÕES DE MISTURA DE ESCALA *SKEW*-NORMAL 15
2.1 DISTRIBUIÇÕES MESN MULTIVARIADAS 15
2.1.1 Definição 15
2.1.2 Representação estocástica 15
2.1.3 Propriedades 16
2.1.4 Distribuição marginal e independência 17
2.2 EXEMPLO 18
2.2.1 Distribuição *Skew*-t multivariada 18
2.3 INFERÊNCIA PELO MÉTODO DA MÁXIMA VEROSSIMILHANÇA 19
2.3.1 Representação hierárquica 19
2.3.2 O algoritmo EM em modelos MESN 20
3 MISTURA FINITA DE DENSIDADES 23
3.1 MISTURAS FINITAS DE DENSIDADES 23
3.1.1 Definição 23
3.1.2 Identificabilidade 24
3.2 A ESTRUTURA DE DADOS INCOMPLETOS PARA O PROBLEMA DE MISTURAS 24
3.3 O ALGORITMO EM NOS MODELOS DE MISTURAS 26
4 MISTURAS FINITAS DE DENSIDADES MESN 28
4.1 O MODELO MF-MESN 28
4.1.1 Definição 28
4.1.2 A representação hierárquica 29
4.1.3 O algoritmo EM em modelos MF-MESN 29
5 MODELAGEM DE MISTURAS FINITAS DE MODELOS DE REGRESSÃO SOB AS DISTRIBUIÇÕES MESN 33
5.1 INTRODUÇÃO 33
5.2 O MODELO ESTUDADO 33
5.2.1 Estimação da Máxima Verossimilhança pelo Algoritmo EM . . 35
5.2.2 Definindo os valores iniciais 37
5.2.3 Critério de Parada de Aceleração de Aitken 38
6 MISTURAS FINITAS DE ESPECIALISTAS DE MODELOS DE REGRESSÃO SOB AS DISTRIBUIÇÕES MESN 39
6.1 O MODELO PROPOSTO 40

6.1.1 Estimação da Máxima Verossimilhança pelo Algoritmo EM . . 40
6.1.2 Definindo os valores iniciais e Critério de Parada 41
7 EXEMPLOS NUMÉRICOS 43
7.1 ESTUDOS DE SIMULAÇÃO 43
7.1.1 Cenários Simulados 43
7.1.2 Recuperação dos Parâmetros 43
7.2 APLICAÇÃO (pesquisa *High School and Beyond*) 53
8 CONCLUSÃO 57
Referências Bibliográficas 58
APÊNDICE A – Lemas 64
APÊNDICE B – Demonstração dos estimadores 65

1 INTRODUÇÃO

Modelos estatísticos baseados em distribuições de misturas finitas capturam muitas propriedades específicas de dados reais, como multimodalidade, assimetria, curtose e heterogeneidade não observada. Devido a essa abrangência, eles podem ser aplicados em diversas áreas do conhecimento, como podemos observar algumas contextualizadas por Yuksel et al. (2012) (72): previsão de demanda de eletricidade, previsão climática, reconhecimento de escrita, reconhecimento facial, classificação de sinal de eletroencefalograma, classificador de batimentos cardíacos, classificação de texto, detecção de minas terrestres, previsão financeira e estimativa de risco de retornos de ativos.

Tomando como exemplo os dados representados na Figura 1, no histograma da esquerda podemos inferir que há três subpopulações presentes na mistura, porém não conseguimos definir as proporções correspondentes, nem especificar a qual desses grupos uma determinada observação pertence, caracterizando uma heterogeneidade sobre a qual não dispomos de informações para discriminar os dados. Porém, ao ajustarmos um modelo de misturas finitas a esses dados, poderemos estimar as proporções (representadas pelas diferentes cores no histograma da direita) e os parâmetros das distribuições pertinentes a cada grupo, bem como será possível estimar o grupo específico de cada observação, classificando-a.

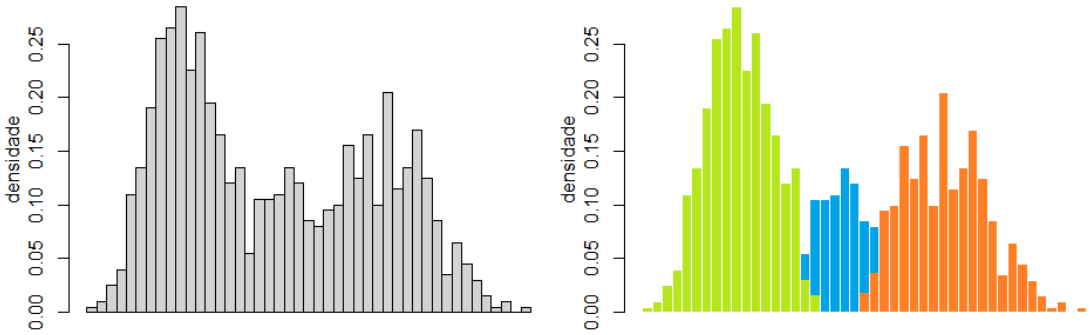

Figura 1 – Heterogeneidade: no histograma à esquerda, as proporções entre os grupos são desconhecidas; no histograma à direita, as cores indicam as proporções estimadas para cada grupo.

Em decorrência da capacidade dos modelos de misturas finitas ajustarem dados com quantidade variada de grupos e com distribuições que podem pertencer a classes paramétricas diferentes, há um amplo campo de estudo e aplicações devido essa flexibilidade; o que se repercute numa vasta literatura em artigos e um grande número de livros sobre o assunto, como Lindsay (1995) (40) , Böhning (2000) (12) , McLachlan e Peel (2000) (45) , Frühwirth-Schnatter (2006) (23) e Mengersen et al. (2011) (51) .

O objetivo deste trabalho consiste em abordar dois tipos de misturas finitas, ambos modelos de regressão linear, sendo o primeiro, o mais usual, chamado em alguns artigos de Modelo Ingênuo, e o segundo conhecido como Modelo Linear de Especialistas (MoE)

(Jacobs et al., 1991) (26). Relacionada a este último modelo, uma proposta de inovação é apresentada, a qual consiste em estender os estudos do MoE utilizando variáveis respostas multivariadas e no contexto assimétrico.

1.1 PROPOSTA DO TRABALHO

Desde a segunda metade do século passado, há uma crescente produção de livros e trabalhos acadêmicos na modelagem de dados por meio de misturas finitas, como nos artigos de Day (1969) (18) e Wolfe (1965, 1967, 1970) (69) (70) (71). Os desafios para os pesquisadores são muitos, pois numa população heterogênea deseja-se conhecer quantas e quais distribuições atuam no processo e a qual delas determinada observação está relacionada, independentemente se o contexto é uni ou multivariado. Com a evolução computacional e a implementação de algoritmos e estruturas matemáticas eficientes, os obstáculos tem sido superados e novas oportunidades surgem numa busca contínua por melhoria. A recompensa pelo esforço na otimização dos modelos de ajuste é facilmente visível na solução de problemas práticos do cotidiano em inúmeras áreas do interesse humano como a biologia e medicina, encontradas em Schlattmann (2009) (61); física, agricultura, economia, marketing, citadas por Leisch (2004) (33), e outras áreas do conhecimento, que podem ser vistas em McLachlan et. al. (2004) (46). Em síntese ao abordado por Basso (2009) (8), é pertinente dizer que, nos primórdios dos estudos das misturas finitas, as densidades normais eram preponderantes, pois podiam ser empregadas para representar densidades de qualquer complexidade, como comentam McLachlan e Peel (2000, seção 6.1), (45) e, devido à sua simplicidade algébrica, foram muito úteis quando os meios computacionais ainda não faziam parte do cotidiano. No final da década de 1970, o algoritmo EM, de Dempster, Laird e Rubin (1977) (19), ampliou significativamente o uso de misturas quando se observava heterogeneidade populacional, devido à sua capacidade de simplificar os ajustes de modelos baseados na máxima verossimilhança aumentada, proposta por Liu (1998) (42). Em decorrência, a expansão do conhecimento e das capacidades computacionais ampliaram o espectro de possibilidades para distribuições com caudas mais pesadas, empregando o modelo de misturas de t-*Student*, que por meio de seu parâmetro adicional consegue acomodar valores extremos, conforme o trabalhos de Peel e McLachlan (2000) (45), Shoham (2002) (64), Shoham et al. (2003) (65), Lin et al. (2004) (35) e Wang (2004) (68); ou para dados de comportamento assimétrico, quando se destaca o modelo de mistura baseado na distribuição *Skew*-normal univariada, proposta por Azzalini (1985) (4). A junção dessas duas vertentes, inclusive no contexto multivariado, levou ao emprego da classe de mistura de escala *Skew*-normal (MESN), apresentada por Lin (2009) (38), bem como às distribuições *Skew*-t e *Skew*-slash, que podem ser vistas em Branco e Dey (2001) (13).

Então, no contexto das misturas finitas de misturas de escala *Skew*-normal (MF-

MESN), visando ampliar os conhecimentos, neste estudo será buscada uma junção de inovações: a formulação do modelo Misturas Finitas de Modelos de Regressão sob as distribuições MESN (MR-MF-MESN), comparando-os com a variação proposta pelo método de Misturas Finitas de Especialistas de Modelos de Regressão sob as distribuições MESN (MoE-MF-MESN) para obtenção das proporções das densidades, que passam a ser modeladas por uma regressão multinomial logística com uso de covariáveis em substituição ao método clássico G-componente (DeSarbo e Cron, 1988 (20); Jones e McLachlan, 1992)) (28). Assim, será proposto um modelo de regressão de mistura robusto baseado nas distribuições multivariadas misturas de escala *Skew*-normal (MR-MF-MESN), estendendo a mistura de regressão univariada proposta por Zeller et al. (2016) (74), tendo como objetivos (i) propor um modelo de regressão multivariado baseado na mistura finita de MESN. (ii) implementar o método de Misturas Lineares de Especialistas (MoE). (iii) implementar e avaliar a proposta computacionalmente. (iv) aplicar esses resultados a uma análise de um conjunto de dados reais.

1.2 ORGANIZAÇÃO DO TRABALHO

O objetivo deste trabalho consiste em estender os conhecimentos aprendidos ao longo do curso de graduação em Estatística da Universidade Federal de Juiz de Fora (UFJF), motivado pela abrangência do tema escolhido em termos de teorias estatísticas e pelo significativo aumento na solução de problemas em aplicações práticas. A intenção principal é estudar o que já é do domínio acadêmico relacionado ao modelo de regressão multivariado baseado na mistura finita de misturas de escala de normais assimétricas multivariadas e também o modelo de misturas lineares de especialistas, apresentando e estendendo as principais ideias com base nos trabalhos desenvolvidos por Basso (2009) (8), Lachos et al. (2010) (31), Lachos et al. (2018) (32), Benites (2018) (15) e Mirfarah et al.(2021) (52), onde as demonstrações podem ser encontradas ou referenciadas.

Para que o trabalho de conclusão de curso não se estenda muito, o leitor deste texto poderá consultar no Capítulo 2 de Basso (2009) (8) definições, propriedades, caracterizações, provas e demais informações adicionais sobre as seguintes distribuições: distribuição *Skew*-normal padrão univariada, Azzalini (1985) (4); a representação univariada com três parâmetros; e a distribuição *Skew*-normal multivariada, considerada uma definição unificada das definições encontradas em Arellano-Valle, Bolfarine e Lachos (2005) (3), bem como suas funções densidade de probabilidade e funções de distribuição acumulada.

A partir do entendimento dos conceitos relacionados às distribuições normais assimétricas, o segundo capítulo aborda as distribuições de misturas de escala *Skew*-normal (MESN) multivariadas, proposta por Branco e Dey (2001) (13), ampliando as possibilidades de englobar dados que apresentam caudas pesadas.

No terceiro capítulo, é introduzido um novo assunto, fundamental para viabilizar

a mudança da visão unimodal da MESN para o contexto multimodal. Nessa etapa, são mostradas explicações detalhadas e pormenorizadas sobre misturas finitas de densidades (definição, distribuição marginal, identificabilidade, a estrutura de dados incompletos para o problema de misturas, o algoritmo EM em modelos de misturas, encontradas na dissertação de mestrado de Basso (Basso, 2009) (8).

A junção dos dois principais aspectos estudados anteriormente, a MESN e as misturas finitas, se dá no quarto capítulo: a multimodalidade. Nesta parte do texto, aborda-se o modelo, a definição, a representação hierárquica e o algoritmo EM para as misturas finitas de densidades de mistura de escala *Skew*-normal (MF-MESN) multivariadas, o que possibilita condições para prosseguir com a obtenção do modelo de regressão.

Sendo um aspecto crucial do trabalho, as Misturas Finitas de Modelos de Regressão sob as distribuições MESN (MR-MF-MESN), vistas no capítulo cinco, conectam todo o conteúdo estudado nos capítulos anteriores; apresentando, além das definições pertinentes, as expressões relacionadas aos estimadores dos parâmetros e erros padrão, bem como aspectos fundamentais relacionados ao algoritmo EM como os valores iniciais e critérios de parada. Cabe destacar ainda que muitas expressões mostradas aqui são as mesmas utilizadas pelo Modelo Linear de Especialistas, estudado logo a seguir.

A busca da inovação neste trabalho ocorre no capítulo seis, fazendo o estudo de Misturas Finitas de Especialistas de Modelos de Regressão sob as distribuições MESN (MoE-MF-MESN) e colocando em evidência as diferenças em relação ao modelo do capítulo anterior (MR-MF-MESN), a fim de criar as condições necessárias para o desfecho do estudo no capítulo sete, com as comparações entre os dois modelos, onde são mostradas simulações e uma aplicação com dados reais.

2 A CLASSE DE DISTRIBUIÇÕES DE MISTURA DE ESCALA *SKEW*-NORMAL

Este capítulo inicia a parte teórica do trabalho apresentando a classe de distribuições mistura de escala *Skew*-normal (MESN) multivariada (Branco & Dey, 2001) (13), sua representação estocástica e algumas de suas propriedades, além de um exemplo de distribuição pertencente a essa classe (*Skew*-t multivariada) (Branco & Dey, 2001) (13). Ao final será apresentado o algoritmo EM para estimação de máxima verossimilhança dos parâmetros em modelos pertencentes a essa classe.

2.1 DISTRIBUIÇÕES MESN MULTIVARIADAS

2.1.1 Definição

Definição 2.1.1 Seja $\mathbf{Y}$ um vetor aleatório p-dimensional, com parâmetro de locação $\boldsymbol{\mu} \in \mathbb{R}$, matriz de escala $\boldsymbol{\Sigma}_{p\times p}$ (positiva definida) e parâmetro de assimetria $\boldsymbol{\lambda} \in \mathbb{R}$, então sua distribuição será da classe MESN, Branco & Dey (2001) (13), se sua função densidade de probabilidade é dada por

$$\psi(\mathbf{y}) = 2\int_0^\infty \phi_p(\mathbf{y};\boldsymbol{\mu},\kappa(U)\boldsymbol{\Sigma})\Phi_1(\kappa^{-1/2}(U)A)dH(U), \tag{2.1.1}$$

em que $A = \boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})$; U é um fator de escala (variável aleatória positiva) com função de distribuição acumulada (fda) $H(\boldsymbol{\nu})$ e função densidade de probabilidade (fdp) $h(\boldsymbol{\nu})$, e $\kappa(\cdot)$ uma função peso bem definida. A notação da distribuição cuja fdp está representada em (2.1.1) é $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},H)$, sendo que se $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_p$ tem-se a distribuição MESN padrão denotada por $\text{MESN}_p(\boldsymbol{\lambda},H)$.

O parâmetro $\boldsymbol{\nu}$ (escalar ou vetorial) que controla as caudas da distribuição indexa a distribuição de U. Se $\kappa(U) = 1$, a densidade (2.1.1) torna-se a fdp de uma distribuição *Skew*-normal, (Azzalini & Dalla-Vale, 1996) (5); e, se além disso $\boldsymbol{\lambda} = 0$, obtém-se a classe de distribuições de mistura de escala normal (MEN) Andrews & Mallows (1974) (2).

2.1.2 Representação estocástica

O estudo das propriedades de uma MESN e a geração de números pseudo-aleatórios de um vetor aleatório $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},H)$ podem ser facilitados pela representação estocástica.

Proposição 2.1.1 Se $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},H)$, sua representação estocástica é dada por

$$\mathbf{Y} \overset{\text{d}}{=} \boldsymbol{\mu} + \kappa(U)^{1/2}\mathbf{Z}, \tag{2.1.2}$$

tal que $\mathbf{Z} \sim \text{SN}_p(\mathbf{0},\boldsymbol{\Sigma},\boldsymbol{\lambda})$ e U é uma variável aleatória positiva (com fda H) independente de $\mathbf{Z}$.

Demonstração. A prova segue do fato que $\mathbf{Y}|U = u \sim SN_p(\boldsymbol{\mu},\kappa(U)\boldsymbol{\Sigma},\boldsymbol{\lambda})$. □

Há também outra forma de representação estocástica das distribuições MESN, vista a seguir, muito útil para implementar o algoritmo EM e viabilizar inferência estatística. As proposições e propriedades enunciadas a partir deste ponto mostram passo-a-passo quais são as expressões matemáticas que irão compor o algoritmo EM.

Proposição 2.1.2 Se $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$, sua representação estocástica tem a seguinte forma

$$\mathbf{Y} \overset{\text{d}}{=} \boldsymbol{\mu} + \boldsymbol{\Delta} T + \kappa^{1/2}(U)\boldsymbol{\Gamma}^{1/2}\mathbf{T}_1, \tag{2.1.3}$$

tal que $T = \kappa^{1/2}(U)|T_0|$, $T_0 \sim N(0,1)$, independente de $\mathbf{T}_1 \sim N_p(\mathbf{0}, \mathbf{I}_p)$, $\boldsymbol{\delta} = \boldsymbol{\lambda}/\left(1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda}\right)^{1/2}$, $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ e $\boldsymbol{\Gamma} = \boldsymbol{\Sigma} - \boldsymbol{\Delta}\boldsymbol{\Delta}^\top$.

Demonstração. A prova segue da Proposição 2.1.1 e da representação estocástica de um vetor aleatório com distribuição *Skew*-normal, vide Azzalini e Dalla-Valle (1996) (5). □

2.1.3 Propriedades

Proposição 2.1.3 Seja $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$, então sua função geradora de momentos (fgm) é dada por

$$M_{\mathbf{y}}(\mathbf{s}) = E[e^{\mathbf{s}^\top\mathbf{Y}}] = \int_0^\infty 2e^{\mathbf{s}^\top\boldsymbol{\mu} + \frac{1}{2}\kappa(u)\mathbf{s}^\top\boldsymbol{\Sigma}\mathbf{s}}\Phi_1(\kappa^{1/2}(u)\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{1/2}\mathbf{s})dH(u), \mathbf{s} \in \mathbb{R}. \tag{2.1.4}$$

Demonstração. Da prova da Proposição 2.1.1 tem-se $\mathbf{Y}|U = u \sim SN_p(\boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma}, \boldsymbol{\lambda})$. Agora, de propriedades conhecidas de esperança condicional, segue que $M_{\mathbf{y}}(\mathbf{s}) = E_U[E[e^{\mathbf{s}^\top\mathbf{Y}}|U]]$ e conclui-se a prova utilizando o Corolário 2.3.1, página 17, em Basso (2009) (8). □

A seguir, será derivado o vetor de médias e a matriz de covariâncias de um vetor aleatório com distribuição MESN. A prova segue da representação estocástica (2.1.2).

Proposição 2.1.4 Seja $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$, então

a) Se $E[\kappa^{1/2}(U)] < \infty$, então $E[\mathbf{Y}] = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}k_1\boldsymbol{\Delta}$,

b) Se $E[\kappa(U)] < \infty$, então $Var[\mathbf{Y}] = k_2\boldsymbol{\Sigma} + \frac{2}{\pi}k_1^2\boldsymbol{\Delta}\boldsymbol{\Delta}^\top$,

com $k_m = E[\kappa^{m/2}(U)]$ e $\boldsymbol{\Delta}$ como definido anteriormente.

Tendo em vista a forma como as quantidades mostradas nesta subseção serão empregadas, torna-se fundamental que sejam especificadas as terminologias usuais nas aplicações do algoritmo EM de Dempster, Laird e Rubin (1977) (19) cuja lógica consiste em obter estimativas dos parâmetros do modelo pelo método de máxima verossimilhança, por meio de cálculos iterativos. Porém, não da forma usual, empregando somente os dados observados. É necessário definir o conceito de variáveis latentes relacionadas a dados não observados. Essas variáveis são incluídas na função da log-verossimilhança através

de uma transformação que a simplifica, deixando-a tratável em termos computacionais e possibilitando eventualmente a obtenção de equações fechadas para os estimadores dos parâmetros. Entretanto, as variáveis latentes formam distribuições conjuntas com a variável observável, exigindo o cálculo de novas quantidades, as esperanças, as quais são importantes para os cálculos computacionais. Assim, o processo pode ser melhor interpretado como um problema de estimação a partir de dados incompletos, aumentando o vetor de dados observados ($\mathbf{y}_{\text{obs}}$) com a inclusão de variáveis latentes ($\mathbf{y}_{\text{mis}}$), sendo *mis* do inglês *missing*, as quais não são observadas diretamente. Obtém-se, deste modo, o vetor de dados completos $\mathbf{y}_c = (\mathbf{y}_{\text{obs}},\ \mathbf{y}_{\text{mis}})$, de tal forma que a função de log-verossimilhança completa é representada por $\boldsymbol{\ell}_c\,(\boldsymbol{\theta}|\mathbf{y}_c)$.

O algoritmo funciona em duas etapas, E (do inglês, *Expectation*) e M (do inglês, *Maximization*). Na primeira, são obtidas as esperanças relacionadas aos dados não observados, a qual consiste em tomar a esperança da log-verossimilhança completa condicional ao vetor de dados observados. Na segunda, realiza-se a substituição dos valores esperados calculados na log-verossimilhança completa, maximizando-a com o propósito de estimação dos parâmetros.

Na sequência, alguns momentos condicionais fundamentais para a implementação do algoritmo EM são:

$$\kappa_r = E[\kappa^{-r}(U)|\mathbf{y}]\ \ e$$
$$\tau_r = E[\kappa^{-r/2}(U)W_\Phi(\kappa^{-1/2}(U)A)|\mathbf{y}],\ \textit{com}\ W_\Phi(x) = \phi_1(x)/\Phi_1(x),\ x \in \mathbb{R}.$$

Proposição 2.1.5 Seja $\boldsymbol{Y} \sim \text{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$ e $\mathbf{U} \sim H(\boldsymbol{\nu})$ o fator de mistura de escala, então

$$\kappa_r = \frac{2\psi_0(\mathbf{y})}{\psi(\mathbf{y})} E[\kappa^{-r}(U_y)\Phi(\kappa^{-1/2}(U_y)A)]\ e\ \tau_r = \frac{2\psi_0(\mathbf{y})}{\psi(\mathbf{y})} E[\kappa^{-r/2}(U_y)\phi(\kappa^{-1/2}(U_y)A)] \quad (2.1.5)$$

com ψ_0 a fda de $Y_0 \sim MEN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, H)$ e $U_y \overset{\text{d}}{=} U|Y_0$.

Demonstração. Veja Proposição 1 em Lachos et al. (2009) (30). □

2.1.4 Distribuição marginal e independência

A seguir, é mostrado que um vetor aleatório com distribuição MESN é invariante quanto a transformações lineares, implicando que a distribuição marginal desse vetor também possui distribuição MESN.

Proposição 2.1.6 Sendo $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$, então para qualquer vetor fixado $\mathbf{b} \in \mathbb{R}^m$ e uma matriz $\mathbf{A} \in \mathbb{R}^{m\times p}$ de posto completo nas linhas,

$$\mathbf{V} = \mathbf{b} + \mathbf{A}\mathbf{Y} \sim \text{MESN}_m(\mathbf{b} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top, \boldsymbol{\lambda}^*, H), \quad (2.1.6)$$

com $\boldsymbol{\lambda}^* = \boldsymbol{\delta}^*/(1-\boldsymbol{\delta}^{*\top}\boldsymbol{\delta}^*)^{1/2}$, $\boldsymbol{\delta}^* = (\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)^{-1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$. Além disso, se $m = p$ a matriz $\mathbf{A}$ é não-singular, então $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}$. Adicionalmente, para qualquer $\mathbf{a} \in \mathbb{R}^p$,

$$\mathbf{a}^\top\mathbf{Y} \sim \text{MESN}(\mathbf{a}^\top\boldsymbol{\mu}, \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}, \lambda^*, H),$$

com $\lambda^* = \alpha/(1-\alpha^2)^{1/2}$, $\alpha = \left\{\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}(1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda})\right\}^{-1/2}\ \mathbf{a}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda}$.

Demonstração. A prova desse resultado é obtida diretamente da Proposição 2.1.3, já que $M_{\mathbf{b}+\mathbf{A}\mathbf{Y}}(\mathbf{s}) = e^{\mathbf{s}^\top\mathbf{b}}M_{\mathbf{Y}}(\mathbf{A}^\top\mathbf{s})$. Quando $\mathbf{A}$ é uma matriz não singular, é fácil ver que $\boldsymbol{\delta}^* = \boldsymbol{\delta}$. □

A partir da Proposição 2.1.6, com $\mathbf{A} = [\mathbf{I}_{p1},\ \mathbf{0}_{p2}]$, $p_1 + p_2 = p$, tem-se o seguinte resultado para um vetor aleatório MESN, relacionado com sua distribuição marginal.

Corolário 2.1.1 Seja $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$ e $\mathbf{Y}$ particionado como $\mathbf{Y}^\top = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top)^\top$ de dimensões p_1 e p_2, respectivamente. Seja

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},\ \boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top,$$

as correspondentes partições de $\boldsymbol{\Sigma}$ e $\boldsymbol{\mu}$. Então, $\mathbf{Y}_1 \sim MESN_{p1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{11}^{1/2}\tilde{\boldsymbol{\upsilon}}, H)$ com

$$\tilde{\boldsymbol{\upsilon}} = \frac{\boldsymbol{\upsilon}_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\upsilon}_2}{\sqrt{1+\boldsymbol{\upsilon}_2^\top\boldsymbol{\Sigma}_{22.1}\boldsymbol{\upsilon}_2}},$$

$\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\upsilon} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\lambda} = (\boldsymbol{\upsilon}_1^\top,\ \boldsymbol{\upsilon}_2^\top)^\top$.

2.2 EXEMPLO

Tendo em vista haver citações relacionadas à distribuição *Skew*-normal nas seções anteriores - Azzalini (1985) (4) e Arellano-Valle, Bolfarine e Lachos (2005) (3)), a seguir é apresentado um exemplo da classe de distribuições MESN, cabendo ressaltar que há outros casos particulares como a *Skew-Slash* e a *Skew*-normal contaminada.

2.2.1 Distribuição *Skew*-t multivariada

Sendo $\boldsymbol{Y} \sim ST_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ uma distribuição *Skew*-t multivariada p-dimensional (obtida do modelo de misturas de escala *Skew*-normal (2.1.1), com $U \sim Gama(\nu/2, \nu/2)$, $\nu > 0$ e $\kappa(u) = 1/u$, sua fdp é dada por

$$\psi(\mathbf{y}) = 2t_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)T\left(\sqrt{\frac{\nu+p}{\nu+d}}; \nu+p\right),\ \mathbf{y} \in \mathbb{R}^p, \tag{2.2.1}$$

em que $t_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ é a fdp da t multivariada e $T(\cdot; \nu)$ é a fda da t univariada, e $d = (\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ a distância de Mahalanobis. Com $\nu = 1$, temos a distribuição *Skew*-Cauchy e, quando $\nu \uparrow \infty$, obtém-se a distribuição *Skew*-normal no limite.

A partir da Proposição 2.1.4, a média e a matriz de covariâncias de $\mathbf{Y}$ são dadas por,

$$E[\mathbf{Y}] = \boldsymbol{\mu} + \sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\boldsymbol{\Delta},\ \nu > 1,$$

$$Var[\mathbf{Y}] = \frac{\nu}{\nu-2}\boldsymbol{\Sigma} - \frac{\nu}{\pi}\left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)\boldsymbol{\Delta}\boldsymbol{\Delta}^\top,\ \nu > 2.$$

Pela Proposição 2.1.5, $\mathbf{Y}_0 \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, i.e. $\mathbf{Y}_0|U = u \sim N_p(\boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma})$ e $U \sim Gama(\nu/2, \nu/2)$. Sendo que $U_{\mathbf{y}} \overset{\mathrm{d}}{=} U|Y_0 = \mathbf{y} \sim Gama((\nu+p)/2, (\nu+d)/2)$, tem-se os seguintes resultados para as esperanças condicionais κ_r e τ_r.

Corolário 2.2.1 Seja $\mathbf{Y} \sim \mathrm{ST}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$. Então,

$$\kappa_r = \frac{\psi_0(\mathbf{y})}{\psi(\mathbf{y})}\frac{2^{r+1}\Gamma\left(\frac{\nu+p+2r}{2}\right)(\nu+d)^{-r}}{\Gamma\left(\frac{\nu+p}{2}\right)}T\left(\sqrt{\frac{\nu+p+2r}{\nu+d}}A;\ \nu+p+2r\right),$$

$$\tau_r = \frac{\psi_0(\mathbf{y})}{\psi(\mathbf{y})}\frac{2^{(r+1)/2}\Gamma\left(\frac{\nu+p+r}{2}\right)}{\pi^{1/2}\Gamma\left(\frac{\nu+p}{2}\right)}\frac{(\nu+d)^{(\nu+p)/2}}{(\nu+d+A^2)^{(\nu+p+r)/2}}.$$

2.3 INFERÊNCIA PELO MÉTODO DA MÁXIMA VEROSSIMILHANÇA

A resolução de problemas de estimação de parâmetros via algoritmo EM, numa abordagem para dados completos, é fundamentada pela utilização da representação hierárquica dos modelos na classe MESN. O entendimento das expressões apresentadas, conforme visto a seguir, será muito útil por ocasião do processo de estimação dos parâmetros em modelos de misturas finitas de distribuições da classe MESN, apresentado no Capítulo 4.

2.3.1 Representação hierárquica

De acordo com a representação estocástica (2.1.3), o modelo MESN pode ser apresentado sob um ponto de vista para dados incompletos.

Proposição 2.3.1 Seja a amostra $\mathbf{Y}_i \sim \mathrm{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$ com $i = 1, \ldots, n$, então o modelo hierárquico para cada vetor aleatório $\mathbf{Y}_i$ é dado por

$$\mathbf{Y}_i|u_i, t_i \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Delta}t_i, \kappa(u_i)\boldsymbol{\Gamma}), \tag{2.3.1}$$

$$T_i|u_i \sim HN(0, \kappa(u_i)), \tag{2.3.2}$$

$$U_i \sim H(u_i; \boldsymbol{\nu}), \tag{2.3.3}$$

sendo que $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$, $\boldsymbol{\Gamma} = \boldsymbol{\Sigma} - \boldsymbol{\Delta}\boldsymbol{\Delta}^\top$ e $HN(0, \kappa(u_i))$ a distribuição *half*-normal com média zero e variância $\kappa(u_i)$.

Considerando $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{u} = (u_1, \ldots, u_n)^\top$, $\mathbf{t} = (t_1, \ldots, t_n)^\top$ e $\mathbf{y}_c = (\mathbf{y}, \mathbf{u}, \mathbf{t})$ como vetor de dados completos, a função de log-verossimilhança completa de

$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, com $\boldsymbol{\sigma}$ denotando o vetor com os elementos da matriz triangular superior $\boldsymbol{\Sigma}$, é dada por

$$\begin{aligned}\ell_c(\boldsymbol{\theta}) &= C - \frac{n}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\sum_{i=1}^{n}\kappa^{-1}(u_i)(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}t_i)^\top\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}t_i) + \sum_{i=1}^{n}\log(h(u_i;\boldsymbol{\nu})) \\ &= C - \frac{n}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\sum_{i=1}^{n}[\kappa^{-1}(u_i)(\mathbf{y}_i - \boldsymbol{\mu})^\top\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) \\ &\quad - 2\kappa^{-1}(u_i)t_i\boldsymbol{\Delta}^\top\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) + \kappa^{-1}(u_i)t_i^2\boldsymbol{\Delta}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}] + \sum_{i=1}^{n}\log(h(u_i;\boldsymbol{\nu}))\end{aligned} \tag{2.3.4}$$

com C uma constante independente do vetor de parâmetro $\boldsymbol{\theta}$.

2.3.2 O algoritmo EM em modelos MESN

Para a estimação dos parâmetros na classe MESN, usa-se a abordagem de dados incompletos por meio da representação hierárquica do modelo. O vetor de dados observados é dado por $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, e os dados aumentados por $\mathbf{u} = (u_1, \ldots, u_n)^\top$, $\mathbf{t} = (t_1, \ldots, t_n)^\top$. Portanto, na etapa E do algoritmo, toma-se o valor esperado da log-verossimilhança completa (2.3.4) condicional à $\mathbf{y}$ e a $\boldsymbol{\theta}$ no seu estado corrente, obtendo-se as seguintes quantidades

$$\widehat{\kappa}_i = E\left\{\kappa^{-1}(U_i)|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \boldsymbol{y}_i\right\},$$

$$\widehat{s_{2i}} = E\left\{\kappa^{-1}(U_i)T_i|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \boldsymbol{y}_i\right\},$$

$$\widehat{s_{3i}} = E\left\{\kappa^{-1}(U_i)T_i^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \boldsymbol{y}_i\right\}.$$

A partir da Proposição 2.1.5, obtém-se $\widehat{\kappa}_i$. Para $\widehat{s_{2i}}$ e $\widehat{s_{3i}}$, inicialmente se descobre qual é a distribuição condicional $T_i|\mathbf{y}_i, u_i$, por meio das equações (2.3.1) e (2.3.2) em associação ao resultado conhecido no Lema A.2 (dado no Apêndice A), temos que:

$$2\phi_p(\mathbf{y}_i;\boldsymbol{\mu}+\boldsymbol{\Delta}t_i,\kappa(u_i)\boldsymbol{\Gamma})\times\phi_1(t_i;0,\kappa(u_i)) = 2\phi_p(\mathbf{y};\boldsymbol{\mu},\kappa(u_i)(\boldsymbol{\Gamma}+\boldsymbol{\Delta}\boldsymbol{\Delta}^\top))\times\phi_1(t_i;\mu_{T_i},\kappa(u_i)M_{T_i}^2)$$

em que $t_i > 0$.

Portanto, $T_i|\mathbf{y}_i, u_i \sim HN(\mu_{T_i}, \kappa(u_i)M_{T_i}^2)$ com $M_T^2 = 1/(1 + \boldsymbol{\Delta}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta})$ e $\mu_{T_i} = M_T^2\boldsymbol{\Delta}^\top\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$.

Em seguida, com as definições de $\widehat{\kappa}_i$ e $\widehat{\tau}_i$, também da Proposição 2.1.5, e usando

propriedades de esperança condicional, encontram-se os seguintes resultados

$$\begin{aligned}
\widehat{s_{2i}} &= E\left\{\kappa^{-1}(U_i)T_i|\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}},\mathbf{y}_i\right\}\\
&= E_{U_i|\mathbf{y}_i}\left\{E_{T_i|u_i,\mathbf{y}_i}[\kappa^{-1}(U_i)T_i]\right\}\\
&= E_{U_i|\mathbf{y}_i}\left\{\kappa^{-1}(U_i)\left[\mu_{T_i}+W_\Phi\left(\frac{\kappa^{-1/2}(U_i)\mu_{T_i}}{M_{T_i}}\right)\kappa^{1/2}(U_i)M_{T_i}\right]\right\}\\
&= \mu_{T_i}E_{U_i|\mathbf{y}_i}\left\{\kappa^{-1}(U_i)\right\}+M_{T_i}E_{U_i|\mathbf{y}_i}\left\{W_\Phi\left(\frac{\kappa^{-1/2}(U_i)\mu_{T_i}}{M_{T_i}}\right)\kappa^{-1/2}(U_i)\right\}\\
&= \widehat{\kappa}_i\widehat{\mu}_{T_i}+\widehat{M}_{T_i}\widehat{\tau}_i,\ i=1,\ldots,n,
\end{aligned}$$

de onde a última igualdade utiliza o Lema A.3 para momentos de uma distribuição *half*-normal. Analogamente, tem-se que

$$\widehat{s_{3i}} = \widehat{\kappa}_i\widehat{\mu}_{T_i}^2+\widehat{M}_{T_i}^2+\widehat{M}_{T_i}\widehat{\mu}_{T_i}\widehat{\tau}_i.$$

Desta forma, o valor esperado condicional da log-verossimilhança completa é

$$\begin{aligned}
Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= E[\ell_c(\boldsymbol{\theta})|\mathbf{y},\widehat{\boldsymbol{\theta}}]\\
&= C-\frac{n}{2}\log|\boldsymbol{\Gamma}|\\
&\quad-\frac{1}{2}\sum_{i=1}^{n}\left\{(\mathbf{y}_i-\boldsymbol{\mu})^\top\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})\widehat{\kappa}_i-2\boldsymbol{\Delta}^\top\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})\widehat{s_{2i}}+\boldsymbol{\Delta}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}\widehat{s_{3i}}\right\}\\
&\quad+\sum_{i=1}^{n}E\left[\log(h(u_i;\boldsymbol{\nu}))|\mathbf{y}_i,\widehat{\boldsymbol{\theta}}\right].
\end{aligned}$$

A etapa M do algoritmo consiste na maximização de $\boldsymbol{\theta}$ em $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$. Será utilizada, inicialmente, uma sequência de passos de maximizações condicionais (CM - *conditional maximization*), referente ao algoritmo ECM (Meng e Rubin, 1993) (50), finalizando com o algoritmo ECME (Liu e Rubin, 1994) (41), uma extensão do EM e do ECM, maximizando a função log-verossimilhança dos dados observados, conhecido como passo CML. Portanto, o algoritmo ECME fica dado por:

Etapa E: Dado $\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(k)}$, obter $\widehat{\kappa}_i^{(k)}$, $\widehat{s_{2i}}^{(k)}$ e $\widehat{s_{3i}}^{(k)}$ para $i=1,\ldots,n$.

Etapa CM: Atualizar $\widehat{\boldsymbol{\theta}}^{(k)}$ maximizando $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})=E[\ell_c(\boldsymbol{\theta})|\mathbf{y},\widehat{\boldsymbol{\theta}}^{(k)}]$ sobre $\boldsymbol{\theta}$, que resulta nas seguintes formas fechadas para os parâmetros transformados $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$ e $\boldsymbol{\Delta}$, respectivamente:

$$\widehat{\boldsymbol{\mu}}^{(k+1)} = \sum_{i=1}^{n}(\widehat{\kappa}_i^{(k)}\mathbf{y}_i-\widehat{s_{2i}}^{(k)}\widehat{\boldsymbol{\Delta}}^{(k)})/(\sum_{i=1}^{n}\widehat{\kappa}_i^{(k)}), \tag{2.3.5}$$

$$\begin{aligned}
\widehat{\boldsymbol{\Gamma}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\Big[&\widehat{\kappa}_i^{(k)}(\mathbf{y}_i-\widehat{\boldsymbol{\mu}}^{(k)})(\mathbf{y}_i-\widehat{\boldsymbol{\mu}}^{(k)})^\top-\widehat{s_{2i}}^{(k)}\widehat{\boldsymbol{\Delta}}^{(k)}(\mathbf{y}_i-\widehat{\boldsymbol{\mu}}^{(k)})^\top\\
&-\widehat{s_{2i}}^{(k)}(\mathbf{y}_i\widehat{\boldsymbol{\mu}}^{(k)})\widehat{\boldsymbol{\Delta}}^{(k)\top}+\widehat{s_{3i}}^{(k)}\widehat{\boldsymbol{\Delta}}^{(k)}\widehat{\boldsymbol{\Delta}}^{(k)\top}\Big],
\end{aligned} \tag{2.3.6}$$

$$\widehat{\boldsymbol{\Delta}}^{(k+1)} = \sum_{i=1}^{n}\widehat{s_{2i}}^{(k)}(\mathbf{y}_i-\widehat{\boldsymbol{\mu}}^{(k)})/\sum_{i=1}^{n}\widehat{s_{3i}}^{(k)}, \tag{2.3.7}$$

$$\boldsymbol{\Sigma}_j^{(k+1)} = \boldsymbol{\Gamma}_j^{(k+1)} + \boldsymbol{\Delta}_j^{(k+1)} \left(\boldsymbol{\Delta}_j^{(k+1)}\right)^\top,$$

$$\boldsymbol{\lambda}_j^{(k+1)} = \left(\boldsymbol{\Sigma}_j^{(k+1)}\right)^{1/2} \boldsymbol{\Delta}_j^{(k+1)} / \left(1 - \left(\boldsymbol{\Delta}_j^{(k+1)}\right)^\top \left(\boldsymbol{\Sigma}_j^{(k+1)}\right)^{1/2} \boldsymbol{\Delta}_j^{(k+1)}\right).$$

Etapa CML: Atualizar $\boldsymbol{\nu}^{(k+1)}$ maximizando a função de verossimilhança marginal, obtendo

$$\widehat{\boldsymbol{\nu}}^{(k+1)} = \arg\max_{\boldsymbol{\nu}} \sum_{i=1}^{n} \log(\psi(y_i; \boldsymbol{\mu}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\nu})), \tag{2.3.8}$$

com $\psi(\mathbf{y}; \boldsymbol{\theta})$ dado em (2.1.1).

Repetem-se iterações até que as diferenças entre os parâmetros estimados ou as diferenças entre as log-verossimilhanças sejam muito pequenas em valores absolutos, ou seja, se $||\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}||$ ou $||\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})||$ forem suficientemente pequenos.

Antes de iniciar as iterações no algoritmo, podem ser utilizados o vetor de média amostral e a matriz de covariâncias amostral para estimar os valores iniciais dos parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, respectivamente. Para a j-ésima coordenada do vetor de assimetria, considere $\widehat{\rho}_j$ a assimetria amostral para a variável j. Então, $\lambda_j^{(0)} = 3 \times sign(\widehat{\rho}_j)$.

3 MISTURA FINITA DE DENSIDADES

Os modelos de misturas têm sido objeto de estudo frequente em situações de heterogeneidade populacional por sua extrema flexibilidade de ajuste, possibilitando aplicações em diversas áreas da estatística, como análise de agrupamento (McLachlan & Chang, 2004) (47), análise discriminante (Rausch & Kelley, 2009) (59), análise de sobrevivência (McLachlan & McGiffin, 1994) (44), métodos não-paramétricos (Kasahara & Shimotsu, 2013) (29) ou semi-paramétricos e em processamento de imagens (Sfikas et al., 2007) (63).

Na literatura disponível, eles são empregados para aproximar densidades complexas, multimodais e/ou totalmente assimétricas, e são preferíveis quando uma única família paramétrica de distribuições não produz uma modelagem satisfatória.

A seguir será apresentado o modelo de misturas de distribuições e algumas propriedades, para na sequência do trabalho ser considerado no âmbito de dados incompletos, os quais serão utilizados na estimação de máxima verossimilhança via algoritmo EM.

3.1 MISTURAS FINITAS DE DENSIDADES

A literatura estatística possui vários livros sobre misturas finitas, sendo os principais: Everitt & Hand (1981) (22), Titterington et al. (1985) (67), McLachlan & Basford (1988) (43), Lindsay (1995) (40), Böhning (1999) (11), McLachlan & Peel (2000a) (45), Frühwirth-Schnatter (2006) (23), Mengersen et al. (2011) (51), e McNicholas (2017) (49).

3.1.1 Definição

Definição 3.1.1 Se a função de densidade de $\mathbf{Y} \in \mathbb{R}^p$ é dada por

$$f(\mathbf{y}) = \sum_{j=1}^{G} p_j \psi_j(\mathbf{y}), \ e \ p_j \geq 0 \ e \ \sum_{j=1}^{G} p_j = 1, \tag{3.1.1}$$

sua distribuição é uma mistura de densidades. A função $f(\cdot)$ é denominada de mistura finita de densidades com G componentes, com os parâmetros $p_1, \ldots, p_G$ denominados proporções de misturas e as densidades $\psi_1, \ldots, \psi_G$ as componentes de misturas.

A fdp (3.1.1) pode ser definida também de outra forma, quando as componentes $\psi_j(\cdot)$ pertencem à famílias paramétricas de distribuições:

$$f(\mathbf{y}; \boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \psi_j(\mathbf{y}; \boldsymbol{\theta_j}), \tag{3.1.2}$$

com $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)$ e $\boldsymbol{\theta}_j$ os parâmetros que definem cada uma das componentes ψ_j, não necessariamente definidos no mesmo espaço paramétrico. Entretanto, caso as componentes de mistura ψ_j pertençam à mesma família paramétrica de distribuições, os parâmetros $\boldsymbol{\theta}_j$ serão do mesmo espaço paramétrico.

3.1.2 Identificabilidade

Se para cada membro distinto de uma família paramétrica $\mathcal{F}$ de funções de distribuição de probabilidade $\psi(\cdot;\boldsymbol{\theta})$ existe um único $\boldsymbol{\theta}$ específico, então essa família é dita identificável. A identificabilidade dos parâmetros em um modelo garante que os mesmos podem ser estimados de maneira única. Para famílias de misturas finitas de densidades é necessário complementar essa definição com mais uma condição, a de permutabilidade.

Definição 3.1.2 Seja $\mathcal{F} = \psi(\mathbf{y};\boldsymbol{\theta}) : \mathbf{y} \in \mathbb{R}^p$ uma família paramétrica de densidades e

$$\mathcal{P} = \left\{ f(\mathbf{y};\boldsymbol{\theta}) : f(\mathbf{y};\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \psi(\mathbf{y};\boldsymbol{\theta}),\ p_j \geq 0,\ \sum_{j=1}^{G} p_j = 1,\ \psi(\mathbf{y};\boldsymbol{\theta}) \in \mathcal{F},\ \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G) \right\}$$

uma classe de misturas finitas de densidades. A classe $\mathcal{P}$ é dita identificável se, para quaisquer dois membros

$$f(\mathbf{y};\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \psi(\mathbf{y};\boldsymbol{\theta}_j) \ e \ f(\mathbf{y};\boldsymbol{\theta}') = \sum_{j=1}^{G'} p'_j \psi(\mathbf{y};\boldsymbol{\theta}'_j)$$

tem-se que $f(\mathbf{y};\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta}')$ se, e somente se, $G = G'$ e ainda se pode permutar os índices das componentes de forma que $p_j = p'_j$ e $\psi(\mathbf{y};\boldsymbol{\theta}_j) = \psi(\mathbf{y};\boldsymbol{\theta}'_j)$ com $j = 1, \ldots, G$.

No exemplo seguinte, Duda e Hart (1973) (21) ilustram essa questão com uma mistura de duas densidades normais de variância unitária,

$$f(y;\boldsymbol{\theta}) = \frac{p_1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}(y-\mu_1)^2\right] + \frac{p_2}{\sqrt{2\pi}} exp\left[-\frac{1}{2}(y-\mu_2)^2\right].$$

Tendo os valores dos parâmetros fixados, se permutarmos os índices em $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\boldsymbol{\theta}_j = (p_j, \mu_j)$, $j = 1, 2$, a densidade $f(\cdot;\boldsymbol{\theta})$ terá o mesmo valor em cada $y \in \mathbb{R}$. Desta forma, a identificabilidade no contexto de misturas de distribuições deve ser também permutável.

Essa questão da identificabilidade é importante porque ocorrem dificuldades quando as componentes de uma mistura pertencem à mesma família de distribuições, conforme apresentado por McLachlan e Basford (1988, seção 1.5) (43): o valor da densidade de mistura terá o mesmo valor independentemente da ordem dos índices dos parâmetros. Então, como os índices de $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ podem ser permutados $G!$ Assim, a densidade de mistura será invariante para todas as permutações dos índices em $\boldsymbol{\theta}$.

3.2 A ESTRUTURA DE DADOS INCOMPLETOS PARA O PROBLEMA DE MISTURAS

Para introduzir a abordagem de dados incompletos no contexto de misturas, inicialmente vejamos como criar vetores pseudo aleatórios de uma mistura de densidades. Para gerar um vetor aleatório $\mathbf{Y}_i$ de uma densidade $f(\mathbf{y}_i)$, conforme (3.1.1), considere $Z_i,\ i = 1, \ldots, G$, uma variável aleatória categórica com probabilidades $p_1, \ldots, p_G$, e suponha que a densidade de $\mathbf{Y}_i$ dado $Z_i = j$ é $\psi_j(\mathbf{y}_i), j = 1, \ldots, G$, então, a densidade

marginal de $\mathbf{Y}_i$ é dada por (3.1.1). A variável Z_i pode ser interpretada como uma variável latente, indicando a componente do qual o vetor $\mathbf{Y}_i$ é proveniente. Por conveniência de notações, algébricas e interpretações, será utilizado um vetor aleatório G-dimensional $\mathbf{Z}_i$ ao invés da variável aleatória Z_i, sendo o j-ésimo elemento de $\mathbf{Z}_i$, Z_{ij}, igual a um, se essa observação é proveniente da componente j, ou zero, caso contrário, i.e, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})$ e

$$Z_{ij} = \begin{cases} 1, & \text{se a } \mathbf{Y}_i \text{ observação é proveniente do } j\text{-ésimo componente} \\ 0, & \text{caso contrário} \end{cases} \tag{3.2.1}$$

Deste modo, a variável $\mathbf{Z}_i$ tem distribuição multinomial considerando uma retirada em G categorias, com probabilidades $p_1, \ldots, p_G$, isto é,

$$P(\mathbf{Z}_i = \mathbf{z}_i) = p_1^{z_{i1}} p_2^{z_{i2}} \ldots p_G^{z_{ij}}, \text{ ou } \mathbf{Z}_i \sim Multi_g(1, p_1, \ldots, p_G).$$

Assim, consegue-se gerar vetores aleatórios de uma distribuição de misturas de densidades com a inclusão do vetor $\mathbf{Z}_i$, que regula as proporções de mistura das componentes.

Uma situação onde o modelo de misturas de distribuições pode ser diretamente aplicável é supor que $\mathbf{Y}_i$ pertence a uma população constituída de G grupos em proporções $p_1, \ldots, p_G$. É importante frisar que $\mathbf{Z}_i$, o vetor latente, é um artifício criado. Conhecemos apenas $\mathbf{Y}_i = \mathbf{y}_i$, para cada $i = 1, \ldots, n$, pertencente à amostra. Então, para cada observação existirá um vetor latente associado, indicando a qual componente ela pertence. Essa ideia é que viabiliza a estimação de máxima verossimilhança através do algoritmo EM.

Se $\mathbf{y}_1, \ldots, \mathbf{y}_n$ são as n realizações dos vetores aleatórios independentes e identicamente distribuídos (iid) $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ com densidade comum $f(\mathbf{y})$ dada por (3.1.1). Então

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{\text{iid}}{\sim} F,$$

sendo $F(\cdot)$ a $f.d.a$ correspondente a densidade de mistura $f(\cdot)$. No algoritmo EM, os dados $\mathbf{y}_1, \ldots, \mathbf{y}_n$ são vistos como incompletos, pois os vetores $\mathbf{z}_1, \ldots, \mathbf{z}_n$ indicadores de componentes não são observáveis, sendo o vetor de dados completos definido por

$$\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{z}^\top),$$

com $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)$ e $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)$. Os vetores $\mathbf{z}_1, \ldots, \mathbf{z}_n$ são realizações dos vetores aleatórios $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, supostamente independentes do vetor de observações e com distribuição dada por

$$\mathbf{Z}_1, \ldots, \mathbf{Z}_n \overset{\text{iid}}{\sim} Multi_G(1, p_1, \ldots, p_G).$$

A *priori*, a j-ésima proporção de mistura pode ser vista como a probabilidade de que uma entidade pertença a j-ésima componente de mistura, enquanto que a probabilidade a

posteriori de que a entidade pertença a j-ésima componente com $\mathbf{y}_i$ já observado, é dada por

$$\begin{aligned}\widehat{z}_{ij} &= P(\text{entidade} \in j\text{-ésima componente}|\mathbf{y}_i) \\ &= P(Z_{ij} = 1|\mathbf{y}_i) \\ &= \frac{p_j\psi_j(\mathbf{y}_i)}{f(\mathbf{y}_i)},\end{aligned} \tag{3.2.2}$$

em que a ultima igualdade vêm do teorema de Bayes.

Uma das grandes vantagens do modelo de misturas finitas é que ele permite a classificação das observações por meio da estimativa de $\widehat{z}_{ij}$. O problema consiste em classificar $\mathbf{Y} \in \mathbb{R}^p$ (ou o vetor de dados observados) a um dos g grupos, ou componentes de mistura. Sabendo que Z assume valores no conjunto $\mathcal{A} = \{1, \ldots, G\}$ com probabilidades a *priori* $p_1, \ldots, p_G$ de que $\mathbf{Y} = \mathbf{y}$ pertença ao seu grupo correspondente; então a solução é utilizar a probabilidade a *posteriori* definida em 3.2.2, na construção de um classificador.

Considerando a independência entre $\mathbf{Z}_i$ e $\mathbf{Y}_i$, $i = 1, \ldots, n$, pode-se derivar a verossimilhança completa dos dados,

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{j=1}^{G}[p_j\psi_j(\mathbf{y}_i;\boldsymbol{\theta}_j)]^{z_{ij}},$$

e portanto, a log-verossimilhança completa fica dada por

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}[\log p_j + \log\psi_j(\mathbf{y}_i;\boldsymbol{\theta}_j)]. \tag{3.2.3}$$

3.3 O ALGORITMO EM NOS MODELOS DE MISTURAS

Etapa E: Estima-se a variável Z_{ij} por meio do valor esperado da log-verossimilhança completa $\ell_c(\boldsymbol{\theta})$ condicional aos dados observados $\mathbf{y}$. Então, se $\boldsymbol{\theta}^{(k)}$ é o valor estimado de $\boldsymbol{\theta}$ na (k)-ésima iteração do algoritmo, na etapa E da $(k+1)$-ésima iteração do algoritmo, deve-se obter

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = E[\log L_c(\boldsymbol{\theta})|\mathbf{y}].$$

Como a log-verossimilhança completa dos dados é linear na variável não observada z_{ij}, na etapa E se obtém a esperança condicional de Z_{ij} dado o vetor de dados observados $\mathbf{y}$, sendo Z_{ij} a variável aleatória correspondente a z_{ij}. Desta forma,

$$E[Z_{ij}|\mathbf{y}] = P(Z_{ij} = 1|\mathbf{y}_i) = \widehat{z}_{ij},$$

isto é,

$$\begin{aligned}\widehat{z}_{ij} &= \frac{p_j^{(k)}\psi(\mathbf{y}_i;\boldsymbol{\theta}_j^{(k)})}{f(\mathbf{y}_i;\boldsymbol{\theta}^{(k)})} \\ &= \frac{p_j^{(k)}\psi(\mathbf{y}_i;\boldsymbol{\theta}_j^{(k)})}{\sum_{j=1}^{G} p_j^{(k)}\psi(\mathbf{y}_i;\boldsymbol{\theta}_j^{(k)})},\end{aligned} \tag{3.3.1}$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, G$. Portanto, usando (3.3.1), a esperança de (3.2.3) condicional a $\mathbf{y}$ fica dada por

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{G} \widehat{z}_{ij} [\log p_j^{(k)} + \log \psi_j(\mathbf{y}_i; \boldsymbol{\theta}_j^{(k)})]. \tag{3.3.2}$$

É importante destacar que a abordagem do algoritmo EM feita nesta seção tem como foco somente os dados aumentados da variável Z_{ij}. Na sequência do estudo, outras quantidades relativas ao vetor de dados aumentados serão incluídas, com mais resultados a serem calculados na etapa E.

Etapa CM: A etapa M do algoritmo na $(k+1)$-ésima iteração requer a maximização de $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ com respeito à $\boldsymbol{\theta}$, para resultar na estimativa atualizada $\boldsymbol{\theta}^{(k+1)}$. No contexto de misturas, a estimativa de $p_j^{(k+1)}$ depende exclusivamente da esperança condicional $\widehat{z}_{ij}$ atualizada na etapa E da $(k+1)$-ésima iteração, conforme visto a seguir:

$$\widehat{p}_j^{(k+1)} = \frac{1}{n} \sum_{j=1}^{G} \widehat{z}_{ij}^{(k+1)}, \quad (j = 1, \ldots, G).$$

A demonstração do estimador de p_j pode ser vista no Apêndice Demonstração de estimadores.

Assim, na estimativa de $\widehat{p}_j$ na $(k+1)$-ésima iteração do algoritmo, há uma contribuição de cada observação $\mathbf{y}_i$ igual a sua probabilidade *a posteriori* de pertencer a j-ésima componente de mistura do modelo. Os estimadores dos demais parâmetros do modelo também são obtidos a partir da maximização de (3.3.2).

4 MISTURAS FINITAS DE DENSIDADES MESN

Neste capítulo ocorre a convergência da estrutura sistematicamente montada até aqui para esse momento, o primeiro objetivo do trabalho, a composição do modelo de mistura finita de distribuições da classe MESN (MF-MESN) multivariada, o qual propõe uma metodologia robusta para dados de distribuições complexas, fora da normalidade, acomodando simultaneamente multimodalidade, assimetria e valores extremos.

Citando alguns trabalhos relacionados a MF-MESN, Lin et al. (2007a) (36), para lidar de forma eficiente com a heterogeneidade e assimetria populacional, propôs uma estrutura de mistura baseada na distribuição *Skew*-normal (SN) (Azzalini 1985) (4). Estendendo este trabalho, Lin et al. (2007b) (37) consideraram robustez para observações discrepantes, usando misturas das distribuições t-*Student* assimétricas (ST) definidas em Azzalini e Capitanio (2003) (6). Basso et al.(2010) (9) consideraram as misturas finitas univariadas onde os componentes são membros da classe flexível de misturas de escala de distribuições *Skew*-normal (MESN) (Lachos et al. 2010) (31). Lin (2009) (38) propôs modelos de mistura SN multivariados, e Pyne et al. (2009) (57) e Lin (2010) (39) propuseram modelos multivariados de mistura de ST. Os modelos de mistura em que os componentes são membros da classe MESN são propostos por Cabral et al. (2012) (16) e Prates et al. (2013) (56).

4.1 O MODELO MF-MESN

4.1.1 Definição

O modelo de mistura de distribuições MESN pode ser obtido juntando as definições Definição 2.1.1 e Definição 3.1.1.

Definição 4.1.1 Denota-se por MF-MESN, com G-componentes de misturas, o modelo com $\mathbf{Y}_i \in \mathbb{R}^p,\ i = 1, \ldots, n$, independentes e identicamente distribuídos, com fdp dada por

$$f(\mathbf{y}_i; \boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \psi_j(\mathbf{y}_i), \quad p_j \geq 0 \ \ e \ \ \sum_{j=1}^{G} p_j = 1, \tag{4.1.1}$$

e $\psi_j(\cdot) = \psi(\cdot; \boldsymbol{\theta}_j)$ uma densidade MESN$(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, H)$.

A família de densidades MESN é paramétrica, com vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$, sendo $\boldsymbol{\theta}_j = (p_j, \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu}),\ \ j = 1, \ldots, G$, o vetor de parâmetros específicos das componentes e p_j as probabilidades de mistura. Como se observa, por conveniência computacional, o vetor de parâmetros relativos ao fator de mistura de escala *Skew*-normal possui os mesmos valores para cada componente, isto é, $\boldsymbol{\nu}_1 = \ldots = \boldsymbol{\nu}_G = \boldsymbol{\nu}$.

4.1.2 A representação hierárquica

Os dados latentes $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})$ tem por finalidade associar a i-ésima observação da amostra a uma das G componentes de misturas. Assim, como na Seção 3.2, esse vetor compõe o modelo hierárquico, sendo $\mathbf{Y}_i|Z_{ij} = 1$ é MESN$(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, H)$ e $\mathbf{Z}_i \sim Multi(1; p_1, \ldots, p_G)$. Então, de acordo com a representação hierárquica de um vetor aleatório MESN dada em (2.3.1) a (2.3.3), obtemos o resultado da seguinte proposição.

Proposição 4.1.1 Sendo a amostra $\mathbf{Y}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, com fdp da Definição 4.1.1, o modelo hierárquico para cada vetor aleatório MESN é dado por

$$\mathbf{Y}_i|u_i, t_i, Z_{ij} = 1 \sim N_p(\boldsymbol{\mu}_j + \boldsymbol{\Delta}_j t_i, \kappa(u_i)\boldsymbol{\Gamma}_j), \tag{4.1.2}$$

$$T_i|u_i, Z_{ij} = 1 \sim HN(0, \kappa(u_i)), \tag{4.1.3}$$

$$U_i|Z_{ij} = 1 \sim H(u_i; \boldsymbol{\nu}), \tag{4.1.4}$$

$$\mathbf{Z}_i \sim Multi(1; p_1, \ldots, p_G), \tag{4.1.5}$$

com $\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_j^{1/2}\boldsymbol{\delta}_j$, $\boldsymbol{\delta}_j = \frac{\boldsymbol{\lambda}_j}{\sqrt{1+\boldsymbol{\lambda}_j^\top \boldsymbol{\lambda}_j}}$, e $\boldsymbol{\Gamma}_j = \boldsymbol{\Sigma}_j - \boldsymbol{\Delta}_j\boldsymbol{\Delta}_j^\top$.

Considerando $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{u} = (u_1, \ldots, u_n)^\top$, $\mathbf{t} = (t_1, \ldots, t_n)^\top$, $\mathbf{z} = (z_1, \ldots, z_n)^\top$ e $\mathbf{y}_c = (\mathbf{y}, \mathbf{u}, \mathbf{t}, \mathbf{z})$ como vetor de dados completos, a função de log-verossimilhança completa de $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_n^\top)^\top$, com $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu})^\top$, e $\boldsymbol{\sigma}_j$ denotando o vetor com os elementos da matriz triangular superior $\boldsymbol{\Sigma}_j$, é dada por

$$\begin{aligned}
\ell_c(\boldsymbol{\theta}) &= C + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\left[\log(p_j) - \frac{1}{2}|\log\boldsymbol{\Gamma}_j| - \kappa^{-1}(u_i)(\mathbf{y}_i - \boldsymbol{\mu}_j - \boldsymbol{\Delta}_j t_i)^\top \boldsymbol{\Gamma}_j^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_j - \boldsymbol{\Delta}_j t_i)\right. \\
&\quad \left. + \log(h(u_i; \boldsymbol{\nu}))\right], \\
&= C + \sum_{i=1}^{n}\sum_{j=1}^{G}\left[z_{ij}\log(p_j) - \frac{1}{2}z_{ij}|\log\boldsymbol{\Gamma}_j| - \frac{1}{2}z_{ij}\kappa^{-1}(u_i)(\mathbf{y}_i - \boldsymbol{\mu}_j)^\top\boldsymbol{\Gamma}_j^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_j)\right. \\
&\quad \left. + z_{ij}\kappa^{-1}(u_i)T_i(\mathbf{y}_i - \boldsymbol{\mu}_j)^\top\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j + \frac{1}{2}z_{ij}\kappa^{-1}(u_i)T_i^2\boldsymbol{\Delta}_j^\top\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j + \log(h(u_i; \boldsymbol{\nu}))\right].
\end{aligned} \tag{4.1.6}$$

com C uma constante independente de $\boldsymbol{\theta}$.

4.1.3 O algoritmo EM em modelos MF-MESN

A implementação do algoritmo EM para estimação de máxima verossimilhança dos parâmetros do modelo $MF\text{-}MESN$ é semelhante àquela mostrada na Subseção **2.3.2**, pois há apenas um nível a mais na hierarquia do modelo em relação àquele proposto fora do contexto de misturas.

A representação do modelo de misturas de distribuições $MESN$ via dados aumentados pode ser obtida através do modelo hierárquico proposto de (4.1.2) a (4.1.5), sendo os dados observados definidos por $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$; e os dados aumentados pela inclusão de

$\mathbf{u} = (u_1, \ldots, u_n)^\top$, $\mathbf{t} = (t_1, \ldots, t_n)^\top$ e $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)$. Por meio dos resultados obtidos na Subseção **2.3.2** para distribuições condicionais e propriedades da distribuição *half*-normal, calculam-se as quantidades a seguir:

$$\begin{aligned}
\widehat{z}_{ij} &= \mathrm{E}\left\{Z_{ij}|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\right\}, \\
\widehat{s}_{1ij} &= \mathrm{E}\left\{Z_{ij}\kappa^{-1}(U_i)|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\right\}, \\
\widehat{s}_{2ij} &= \mathrm{E}\left\{Z_{ij}\kappa^{-1}(U_i)T_i|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\right\}, \\
\widehat{s}_{3ij} &= \mathrm{E}\left\{Z_{ij}\kappa^{-1}(U_i)T_i^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\right\},
\end{aligned}$$

Da Subseção 3.3, tem-se que

$$\widehat{z}_{ij} = \frac{\widehat{p}_j\psi(\mathbf{y}_i; \widehat{\boldsymbol{\theta}}_j)}{\sum_{j=1}^{G}\widehat{p}_j\psi(\mathbf{y}_i; \widehat{\boldsymbol{\theta}}_j)}.$$

Utilizando propriedades de esperança condicional, tem-se que

$$\begin{aligned}
\widehat{s}_{1ij} &= \mathrm{E}\left\{Z_{ij}\kappa^{-1}(U_i)|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\right\} \\
&= \mathrm{E}_{Z_{ij}|\mathbf{y}_i}\left\{\mathrm{E}_{U|\mathbf{y},Z_{ij}}\left[Z_{ij}\kappa^{-1}(U_i)\right]\right\} \\
&= \mathrm{E}_{Z_{ij}|\mathbf{y}_i}\left\{Z_{ij}\mathrm{E}_{U|\mathbf{y},Z_{ij}}\left[\kappa^{-1}(U_i)\right]\right\} \\
&= \widehat{z}_{ij}\widehat{\kappa}_{ij},
\end{aligned}$$

e

$$\begin{aligned}
\widehat{s}_{2ij} &= \mathrm{E}\left\{Z_{ij}\kappa^{-1}(U_i)T_i|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\right\} \\
&= \mathrm{E}_{Z_{ij}|\mathbf{y}_i}\left\{\mathrm{E}_{U_i|\mathbf{y}_i,Z_{ij}}\left[\mathrm{E}_{T_i|u_i,\mathbf{y}_i,Z_{ij}}(Z_{ij}\kappa^{-1}(U_i)T_i)\right]\right\} \\
&= \mathrm{E}_{Z_{ij}|\mathbf{y}_i}\left\{\mathrm{E}_{U_i|\mathbf{y}_i,Z_{ij}}\left[Z_{ij}\kappa^{-1}(U_i)\left(\mu_{T_i} + W_\Phi\left(\frac{\kappa^{-1/2}(U_i)\mu_{T_i}}{M_{T_i}}\right)\kappa^{1/2}(U_i)M_{T_i}\right)\right]\right\} \\
&= \mathrm{E}_{Z_{ij}|\mathbf{y}_i}\left\{Z_{ij}\left[\mu_{T_i}\mathrm{E}_{U_i|\mathbf{y}_i}\left(\kappa^{-1}(U_i)\right) + M_{T_i}\mathrm{E}_{U_i|\mathbf{y}_i}\left(W_\Phi\left(\frac{\kappa^{-1/2}(U_i)\mu_{T_i}}{M_{T_i}}\right)\kappa^{-1/2}(U_i)\right)\right]\right\} \\
&= \widehat{z}_{ij}\left(\widehat{\kappa}_{ij}\widehat{\mu}_{T_{ij}} + \widehat{M}_{T_j}\widehat{\tau}_{ij}\right),
\end{aligned}$$

Além disso, por analogia, segue que

$$\widehat{s}_{3ij} = \widehat{z}_{ij}\left(\widehat{\kappa}_{ij}\widehat{\mu}_{T_{ij}}^2 + \widehat{M}_{T_j}^2 + \widehat{M}_{T_j}\widehat{\mu}_{T_{ij}}\widehat{\tau}_{ij}\right),$$

lembrando que

$$\begin{aligned}
\widehat{\tau}_{ij} &= \mathrm{E}\left\{\kappa^{-1/2}(U_i)W_\Phi\left(\frac{\kappa^{-1/2}(U_i)\widehat{\mu}_{T_i}}{\widehat{M}_{T_i}}\right)|\widehat{\boldsymbol{\theta}}, \mathbf{y}_i, Z_{ij} = 1\right\}, \\
\widehat{M}_{T_j}^2 &= 1/(1 + \widehat{\boldsymbol{\Delta}}_j^\top\widehat{\boldsymbol{\Gamma}}_j^{-1}\widehat{\boldsymbol{\Delta}}_j), \\
\widehat{\mu}_{T_{ij}} &= \widehat{M}_{T_j}^2\widehat{\boldsymbol{\Delta}}_j^\top\widehat{\boldsymbol{\Gamma}}_j^{-1}(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_j), \\
\widehat{\kappa}_{ij} &= \mathrm{E}\left\{\kappa^{-1}(U_j)|\widehat{\boldsymbol{\theta}}, \mathbf{y}_i, Z_{ij} = 1\right\},
\end{aligned}$$

Da verossimilhança (4.1.6), mostra-se que a esperança condicional da log-verossimilhança dos dados completos, dado os dados observados e $\widehat{\boldsymbol{\theta}}$ na atual iteração, é dada por

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= \mathrm{E}\left[\ell_c(\boldsymbol{\theta})|\mathbf{y}, \widehat{\boldsymbol{\theta}}\right] \\ &= C + \sum_{i=1}^{n}\sum_{j=1}^{G}\left[\widehat{z}_{ij}\log(p_j) - \frac{1}{2}\widehat{z}_{ij}\log|\boldsymbol{\Gamma}_j| - \frac{1}{2}\widehat{s}_{1ij}(\mathbf{y}_i - \boldsymbol{\mu}_j)^\top\boldsymbol{\Gamma}_j^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_j)\right. \\ &\quad \left. + \widehat{s}_{2ij}(\mathbf{y}_i - \boldsymbol{\mu}_j)^\top\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j + \frac{1}{2}\widehat{s}_{3ij}\boldsymbol{\Delta}_j^\top\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j + E\left[\log(h(u_i;\boldsymbol{\nu}))|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\right]\right]. \end{aligned}$$

Sendo uma extensão direta do algoritmo apresentado na Subseção **2.3.2**, o algoritmo ECME para a estimação de máxima verossimilhança do parâmetro $\boldsymbol{\theta}$ pode ser descrito como:

Etapa E: Dado $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, obter $\widehat{z}_{ij}$, $\widehat{s}_{1ij}$, $\widehat{s}_{2ij}$, $\widehat{s}_{3ij}$, para $i = 1, \ldots, n$ e $j = 1, \ldots, G$.

Etapa CM: Para $j = 1, \ldots, G$, atualizar $\widehat{p}_j^{(k)}$, $\widehat{\boldsymbol{\mu}}_j^{(k)}$, $\widehat{\boldsymbol{\Gamma}}_j^{(k)}$, $\widehat{\boldsymbol{\Delta}}_j^{(k)}$ usando as seguintes expressões fechadas para os parâmetros transformados

$$\begin{aligned} \widehat{p}_j^{(k+1)} &= n^{-1}\sum_{i=1}^{n}\widehat{z}_{ij}^{(k)}, \\ \widehat{\boldsymbol{\mu}}_j^{(k+1)} &= \sum_{i=1}^{n}\left(\widehat{s}_{1ij}^{(k)}\mathbf{y}_i - \widehat{\boldsymbol{\Delta}}_j^{(k)}\widehat{s}_{2ij}^{(k)}\right) / \sum_{i=1}^{n}\widehat{s}_{1ij}^{(k)}, \\ \widehat{\boldsymbol{\Delta}}_j^{(k+1)} &= \left[\sum_{i=1}^{n}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_j^{(k+1)}\right)\widehat{s}_{2ij}^{(k)}\right] / \sum_{i=1}^{n}\widehat{s}_{3ij}^{(k)}, \\ \widehat{\boldsymbol{\Gamma}}_j^{(k+1)} &= \left(\sum_{i=1}^{n}\widehat{z}_{ij}^{(k)}\right)^{-1}\sum_{i=1}^{n}\left\{\widehat{s}_{1ij}^{(k)}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_j^{(k+1)}\right)\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_j^{(k+1)}\right)^\top\right. \\ &\quad - \left[\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_j^{(k+1)}\right)\left(\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\right)^\top + \widehat{\boldsymbol{\Delta}}_j^{(k+1)}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_j^{(k+1)}\right)^\top\right]\widehat{s}_{2ij}^{(k)} \\ &\quad \left. + \widehat{\boldsymbol{\Delta}}_j^{(k+1)}\left(\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\right)^\top\widehat{s}_{3ij}^{(k)}\right\}, \\ \widehat{\boldsymbol{\Sigma}}_j^{(k+1)} &= \widehat{\boldsymbol{\Gamma}}_j^{(k+1)} + \widehat{\boldsymbol{\Delta}}_j^{(k+1)}\left(\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\right)^\top, \\ \widehat{\boldsymbol{\lambda}}_j^{(k+1)} &= \left(\widehat{\boldsymbol{\Sigma}}_j^{(k+1)}\right)^{1/2}\widehat{\boldsymbol{\Delta}}_j^{(k+1)} / \left(1 - \left(\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\right)^\top\left(\widehat{\boldsymbol{\Sigma}}_j^{(k+1)}\right)^{1/2}\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\right). \end{aligned}$$

Etapa CML: Atualizar $\boldsymbol{\nu}^{(k)}$ maximizando a função de log-verossimilhança marginal, obtendo

$$\widehat{\boldsymbol{\nu}}^{(k+1)} = \arg\max_{\boldsymbol{\nu}}\sum_{i=1}^{n}\log\left(\sum_{j=1}^{G}p_j\psi\left(\mathbf{y}_i; \widehat{\boldsymbol{\mu}}_j^{(k+1)}, \widehat{\boldsymbol{\Sigma}}_j^{(k+1)}, \widehat{\boldsymbol{\lambda}}_j^{(k+1)}, \boldsymbol{\nu}\right)\right).$$

As iterações ocorrem repetidamente até que alguma regra de convergência adequada seja satisfeita como, por exemplo, se o módulo da diferença entre duas estimativas seguidas dos parâmetros, $||\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}||$, ou entre duas estimativas seguidas da log-verossimilhança, $||\ell\left(\boldsymbol{\theta}^{(k+1)}\right) - \ell\left(\boldsymbol{\theta}^{(k)}\right)||$, forem suficientemente pequenas.

De forma semelhante ao explicado no final do Capítulo 2, devem ser estabelecidos valores iniciais para os parâmetros a serem estimados ($\widehat{p}_j^{(0)}$, $\widehat{\boldsymbol{\mu}}_j^{(0)}$, $\widehat{\boldsymbol{\Sigma}}_j^{(0)}$, $\widehat{\boldsymbol{\lambda}}_j^{(0)}$) antes da

primeira iteração do algoritmo. Para $\widehat{p}_j^{(0)}$, uma proposta consiste em particionar os dados em G grupos usando o algoritmo de agrupamento K-*means*, calculando a proporção de pontos de dados pertencentes ao mesmo *cluster* j, $j = 1, \ldots, G$, a fim de atribuir a cada $\widehat{p}_j^{(0)}$ a sua proporção correspondente. Podem ser utilizados o vetor de média amostral e a matriz de covariâncias amostral para estimar os valores iniciais dos parâmetros $\boldsymbol{\mu}_j^{(0)}$ e $\boldsymbol{\Sigma}_j^{(0)}$, respectivamente. Para a j-ésima coordenada do vetor de assimetria, considere $\widehat{\rho}_j$ a assimetria amostral para a variável j. Então, $\lambda_j^{(0)} = 3 \times sign(\widehat{\rho}_j)$.

A estimação dos erros padrão é importante para que possam ser avaliadas as significâncias dos parâmetros do modelo. Neste contexto, os erros padrão podem ser obtidos através de técnicas computacionais como o procedimento *bootstrap* paramétrico ou por meio da aproximação da matriz de covariância assintótica de $\widehat{\boldsymbol{\theta}}$ pelo inverso da matriz de informação observada, o que será detalhado nos Capítulos 5 e 6, respectivamente.

Os modelos para MF-MESN estão implementados no pacote *mixsmsn*, de Prates et. al. (2013) (56) do Programa R (R Core Team, 2017) (60).

5 MODELAGEM DE MISTURAS FINITAS DE MODELOS DE REGRESSÃO SOB AS DISTRIBUIÇÕES MESN

Estendendo os conceitos abordados nos capítulos anteriores, apresentado as misturas finitas de modelos de regressão multivariados assimétricos, no contexto da classe de distribuições MESN, proposta por Benites (2018) (15).

5.1 INTRODUÇÃO

Assumindo que $\mathbf{Y}_i$ (variável aleatória) é dependente de $\mathbf{x}_i = (1,\ x_{i1}, \ldots, x_{id-1})^\top$, i-ésima linha da matriz Xi, por meio de $E[\mathbf{Y}_i|\boldsymbol{\beta}, \mathbf{x}_i] = \mathbf{X}_i\boldsymbol{\beta}$, sendo $\boldsymbol{\beta}$ um vetor de coeficientes de regressão desconhecidos de dimensão d; e considerando que em estatística aplicada nem sempre o coeficiente de regressão é fixo sobre todas as realizações possíveis de $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, o ajuste de misturas finitas de modelos de regressão tendem a capturar melhor as eventuais mudanças neste vetor de parâmetros e assim são bastante utilizados para investigar a relação entre variáveis provenientes de vários grupos homogêneos latentes desconhecidos, os quais se originam de uma heterogeneidade presente, em que não é possível discriminar os dados de acordo com uma determinada característica de interesse, devido esta não ter sido observada.

São apresentadas neste capítulo as misturas finitas de modelos de regressão considerando a classe de distribuições MESN, propostas por Benites (2018) (15) e, portanto, estendendo o trabalho de Basso et al. (2010) para a configuração de regressão. O uso desta classe de distribuições assimétricas no contexto de misturas finitas facilita a implementação do algoritmo EM devido sua representação estocástica, permite robustez no processo de estimação uma vez que podem atribuir diferentes pesos a cada observação, o que possibilita controlar a influência das observações nas estimativas dos parâmetros, conforme visto em Lachos et al. (2018) (32).

5.2 O MODELO ESTUDADO

Nesta seção, estuda-se as misturas finitas de modelos de regressão sendo que os erros aleatórios seguem uma distribuição MESN (MR-MF-MESN). As misturas finitas de modelos de regressão sob normalidade (MR-MF-N) (Quandt, 1972) (58) podem ser definidas como:

Seja Z uma variável latente tal que, dado $Z = j$, a resposta $\mathbf{Y}$ depende do preditor p-dimensional $\mathbf{X}$ de forma linear

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j,\ j = 1, \ldots, G, \tag{5.2.1}$$

em que G é o número de grupos (também chamados de componentes no modelo de misturas finitas) na população e $\boldsymbol{\epsilon}_j \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_j)$, então a densidade condicional de $\mathbf{Y}$ dado $\mathbf{X}$, sem

observar Z, é

$$f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \boldsymbol{\psi}_p(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j), \tag{5.2.2}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$, com $\boldsymbol{\theta}_j = (p_j, \boldsymbol{\beta}_j^\top, \boldsymbol{\sigma}_j)$, $\boldsymbol{\sigma}_j$ denotando o vetor com os elementos da matriz triangular superior que compõem $\boldsymbol{\Sigma}_j$. Seguindo os trabalhos de Lachos et al. (2018) (32) e Benites (2018) (15), a definição de MR-MF-N pode ser estendida considerando a seguinte suposição na relação linear em (5.2.1):

$$\boldsymbol{\epsilon}_j \sim \text{MESN}_p(b\boldsymbol{\Delta}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu}_j), \tag{5.2.3}$$

em que $\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_j^{1/2}\boldsymbol{\delta}_j$, $\boldsymbol{\delta}_j = \frac{\boldsymbol{\lambda}_j}{\sqrt{1+\boldsymbol{\lambda}_j^2}}$, $b = -\sqrt{\frac{2}{\pi}}K_1$, com $K_r = E[U^{-r/2}]$, $r = 1, 2, \ldots$, que corresponde ao modelo de regressão onde a distribuição do erro tem média zero e, portanto, os parâmetros de regressão são todos comparáveis.

As misturas finitas de modelos de regressão sob a a classe de distribuições MESN podem ser definidas, de modo semelhante a (5.2.2), da seguinte forma:

$$f(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \boldsymbol{\psi}_p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}_j), \quad p_j \geq 0, \quad \sum_{j=1}^{G} p_j = 1, \tag{5.2.4}$$

em que $\boldsymbol{\psi}_p(\cdot|\mathbf{X},\boldsymbol{\theta}_j)$ é a função densidade da $\text{MESN}(\mathbf{X}^\top\boldsymbol{\beta}_j + b\boldsymbol{\Delta}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu})$ e $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, p_1, \ldots, p_G)^\top$, sendo $\boldsymbol{\theta}_j = (p_j, \boldsymbol{\beta}_j, \boldsymbol{\sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu})$, $j = 1, \ldots, G$, tal que assume-se $\boldsymbol{\nu}_1 = \ldots = \boldsymbol{\nu}_G = \boldsymbol{\nu}$, conforme dito na Definição 4.1.1. O modelo considera que o coeficiente de regressão e a variância do erro não são homogêneos em todos os pares independentes possíveis $(\mathbf{y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$. De fato, eles mudam entre os subgrupos de observações.

No contexto da inferência clássica, o parâmetro desconhecido $\boldsymbol{\theta}$, dadas as observações $(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_n, \mathbf{y}_n)$, é tradicionalmente estimado pela log-verossimilhança.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log(f(\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\theta})). \tag{5.2.5}$$

Portanto, como a expressão de 5.2.5 não possui uma solução explícita, envolve integrais e o emprego de métodos numéricos que geralmente não são estáveis, emprega-se o algoritmo EM (Dempster et al., 1977) (19) que é mais estável e possibilita a obtenção de expressões fechadas para estimar a maioria dos parâmetros do modelo de interesse.

Assumindo que o número de componentes G é conhecido e fixo, enquanto $p_1, \ldots, p_G$ é um vetor desconhecido de pesos de mistura e são restritos a serem positivos, é importante a identificabilidade no MR-MF-MESN. Nesse contexto, Hennig (2000) (25) e Grün & Leisch (2008) (24) provaram que misturas de regressão linear univariada ou multivariada são identificáveis; e Otiniano et al. (2015) (55) provaram a identificabilidade da mistura finita de distribuições *Skew*-normal e *Skew*-t.

5.2.1 Estimação da Máxima Verossimilhança pelo Algoritmo EM

Conforme já visto na Seção 3.2, é mais apropriado lidar com um vetor aleatório $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})^\top$ no lugar de uma variável aleatória Z_i, assim como está definido em 3.2.1.

Consequentemente, sabendo que o vetor aleatório $\mathbf{Z}_i \overset{\text{iid}}{\sim} Multinomial(1; p_1, \ldots, p_G)$, temos que

$$\mathbf{Y}_i | Z_{ij} = 1 \overset{\text{iid}}{\sim} \text{MESN}_p(\mathbf{X}_i\boldsymbol{\beta}_j + \boldsymbol{\Delta}_j t_i, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu}).$$

Observe que $Z_{ij} = 1$ se e somente se $Z_i = j$. Assim, a configuração definida acima, junto com (2.3.1), (2.3.2) e (2.3.3), pode ser escrita hierarquicamente como

$$\boldsymbol{Y}_i | T_i = t_i, U_i = u_i, Z_{ij} = 1 \overset{\text{ind}}{\sim} N_p(\mathbf{X}_i\boldsymbol{\beta}_j + \boldsymbol{\Delta}_j t_i, \kappa(u_i)\boldsymbol{\Gamma}_j), \tag{5.2.6}$$

$$T_i | U_i = u_i, Z_{ij} = 1 \overset{\text{iid}}{\sim} TN(b, \kappa(u_i); (b, \infty)), \tag{5.2.7}$$

$$U_i | Z_{ij} = 1 \overset{\text{iid}}{\sim} H(\boldsymbol{\nu}), \tag{5.2.8}$$

$$\mathbf{Z}_i \overset{\text{iid}}{\sim} Multinomial(1; p_1, \ldots, p_G), \tag{5.2.9}$$

para $i = 1, \ldots, n$, todos independentes, em que $\boldsymbol{\delta}_j = \boldsymbol{\lambda}_j / \sqrt{1 + \boldsymbol{\lambda}_j^\top \boldsymbol{\lambda}_j}$, $\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_j^{1/2}\boldsymbol{\delta}_j$, $\boldsymbol{\Gamma}_j = \boldsymbol{\Sigma}_j - \boldsymbol{\Delta}_j\boldsymbol{\Delta}_j^\top$ e $TN_1(r, s; (a, b))$ denota a distribuição normal univariada, truncada no intervalo (a, b). Fazendo $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)$, $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{t} = (t_1, \ldots, t_n)$ e $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)$. Então, sob a representação hierárquica (5.2.6) a (5.2.9), segue a função log-verossimilhança associada com $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{u}^\top, \mathbf{t}^\top, \mathbf{z}^\top)^\top$ é

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) = C + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij} \Bigg[&\log(p_j) - \frac{1}{2}\log|\boldsymbol{\Gamma}_j| \\ &-\frac{1}{2}u_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j - \boldsymbol{\Delta}_j t_i)^\top \boldsymbol{\Gamma}_j^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j - \boldsymbol{\Delta}_j t_i) + \log(h(u_i;\boldsymbol{\nu}))\Bigg],\end{aligned}$$

em que C é uma constante que é independente do vetor de parâmetros $\boldsymbol{\theta}$ e $h(\cdot;\boldsymbol{\nu})$ é a densidade de U_i. Considerando $\widehat{\boldsymbol{\theta}}_j^{(k)} = (\widehat{p}_j^{(k)}, \widehat{\boldsymbol{\beta}}_j^{(k)\top}, \boldsymbol{\sigma}_j^{(k)}, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{j}}^{(\boldsymbol{k})}, \boldsymbol{\nu}^{(k)})^\top$, $\boldsymbol{\sigma}_j^{(k)}$ denotando o vetor com os elementos da matriz triangular superior que compõem $\boldsymbol{\Sigma}_j$, os estimadores de $\boldsymbol{\theta}$ na k-ésima iteração, segue que $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ é dada por

$$\begin{aligned}Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = C &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij}^{(k)} \log p_j \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij}^{(k)} \log\boldsymbol{\Gamma}_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{1ij}^{(k)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^\top\boldsymbol{\Gamma}_j^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{2ij}^{(k)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^\top\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{3ij}^{(k)}\boldsymbol{\Delta}_j^\top\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j,\end{aligned}$$

em que $\widehat{z}_{ij}^{(k)} = E[Z_{ij}|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}]$, $\widehat{s}_{1ij}^{(k)} = \widehat{zu}_{ij}^{(k)} = E[Z_{ij}U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}]$, $\widehat{s}_{2ij}^{(k)} = \widehat{zut}_{ij}^{(k)} = E[Z_{ij}U_iT_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}]$ e $\widehat{s}_{3ij}^{(k)} = \widehat{zut^2}_{ij}^{(k)} = E[Z_{ij}U_iT_i^2|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}]$. Usando propriedades conhecidas de esperança condicional, obtemos

$$\widehat{z}_{ij}^{(k)} = \frac{\widehat{p}_j^{(k)} \psi_{\text{MESN}_p}(\mathbf{y}_i | \widehat{\boldsymbol{\theta}}_j^{(k)})}{\sum_{l=1}^{G} \widehat{p}_l^{(k)} \psi_{\text{MESN}_p}(\mathbf{y}_i | \widehat{\boldsymbol{\theta}}_l^{(k)})}, \tag{5.2.10}$$

$\widehat{zu}_{ij}^{(k)} = \widehat{z}_{ij}^{(k)} \widehat{u}_{ij}^{(k)}$, $\widehat{zut}_{ij}^{(k)} = \widehat{z}_{ij}^{(k)} \widehat{ut}_{ij}^{(k)}$ e $\widehat{zut^2}_{ij}^{(k)} = \widehat{z}_{ij}^{(k)} \widehat{ut^2}_{ij}^{(k)}$, com

$$\widehat{ut}_{ij}^{(k)} = \widehat{\kappa}_{ij}^{(k)} (\widehat{\mu}_{T_{ij}}^{(k)} + b) + \widehat{M}_{T_j}^{2(k)} \widehat{\tau}_{ij}^{(k)}, \tag{5.2.11}$$

$$\widehat{ut^2}_{ij}^{(k)} = \widehat{\kappa}_{ij}^{(k)} (\widehat{\mu}_{T_{ij}}^{(k)} + b)^2 + \widehat{M}_{T_j}^{2(k)} + \widehat{M}_{T_j}^{2(k)} (\widehat{\mu}_{T_{ij}}^{(k)} + 2b) \widehat{\tau}_{ij}^{(k)}, \tag{5.2.12}$$

em que

$$\begin{aligned}
\widehat{\tau}_{ij}^{(k)} &= \text{E}\left\{ \kappa^{-1/2}(U_i) W_\Phi \left(\frac{\kappa^{-1/2}(U_i) \widehat{\mu}_{T_{ij}}^{(k)}}{\widehat{M}_{T_j}^{(k)}} \right) | \widehat{\boldsymbol{\theta}}, \mathbf{y}_i, Z_{ij} = 1 \right\}, \\
\widehat{M}_{T_j}^{2(k)} &= 1/(1 + \widehat{\boldsymbol{\Delta}}_j^\top \widehat{\boldsymbol{\Gamma}}_j^{-1} \widehat{\boldsymbol{\Delta}}_j), \\
\widehat{\mu}_{T_{ij}}^{(k)} &= \widehat{M}_{T_j}^{2(k)} \widehat{\boldsymbol{\Delta}}_j^\top \widehat{\boldsymbol{\Gamma}}_j^{-1} (\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}}_j - b \widehat{\boldsymbol{\Delta}}_j), \\
\widehat{\kappa}_{ij}^{(k)} &= \text{E}\left\{ \kappa^{-1}(U_j) | \widehat{\boldsymbol{\theta}}, \mathbf{y}_i, Z_{ij} = 1 \right\},
\end{aligned}$$

com $i = 1, \ldots, n$, sendo todos estas quantidades avaliados em $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$. Considerando $\mathbf{a}_{ij} = \frac{\widehat{\mu}_{T_{ij}}}{\widehat{M}_{T_j}} = \boldsymbol{\Gamma}_j \boldsymbol{\Sigma}_j^{1/2} (\mathbf{y}_i^\top - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j - b \widehat{\boldsymbol{\Delta}}_j)$, a esperança condicional obtida em (5.2.11) e (5.2.12), especificamente $\widehat{\kappa}_{ij}^{(k)}$ e $\widehat{\tau}_{ij}^{(k)}$, podem ser facilmente derivadas a partir do resultado dado em Seção 2.2. Assim, pelo menos para as distribuições *Skew*-normal e *Skew*-t, temos uma expressão de forma fechada para as quantidades $\widehat{\kappa}_{ij}^{(k)}$ e $\widehat{\tau}_{ij}^{(k)}$, como pode ser encontrado em Zeller et al. (2011) (73) e Basso et al. (2010) (9).

Novamente, para atualizar a estimativa de $\boldsymbol{\nu}$ foi realizada a maximização direta da log-verossimilhança marginal. Assim, o algoritmo ECME para a estimativa de máxima verossimilhança de $\boldsymbol{\theta}$ é definido como segue:

Etapa E: Dado $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, obter $\widehat{z}_{ij}$, $\widehat{s}_{1ij}$, $\widehat{s}_{2ij}$, $\widehat{s}_{3ij}$, para $i = 1, \ldots, n$ e $j = 1, \ldots, G$, usando (5.2.11) e (5.2.12).

Etapa CM: Para $j = 1, \ldots, G$, atualizar $\widehat{p}_j^{(k)}$, $\widehat{\boldsymbol{\beta}}_j^{(k)}$, $\widehat{\boldsymbol{\Gamma}}_j^{(k)}$, $\widehat{\boldsymbol{\Delta}}_j^{(k)}$ usando as seguintes expressões fechadas para os parâmetros transformados

$$
\begin{aligned}
\widehat{p}_j^{(k+1)} &= n^{-1}\sum_{i=1}^{n}\widehat{z}_{ij}^{(k)},\\
\widehat{\boldsymbol{\beta}}_j^{(k+1)} &= \left(\sum_{i=1}^{n}\widehat{s}_{1ij}^{(k)}\mathbf{X}_i^\top\mathbf{X}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{X}_i^\top\left(\widehat{s}_{1ij}^{(k)}\mathbf{y}_i^\top-\widehat{s}_{2ij}^{(k)}\widehat{\boldsymbol{\Delta}}_j^{(k)}\right),\\
\widehat{\boldsymbol{\Delta}}_j^{(k+1)} &= \left[\sum_{i=1}^{n}\widehat{s}_{2ij}^{(k)}\left(\mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}_j^{(k)}\right)^\top\right]/\sum_{i=1}^{n}\widehat{s}_{3ij}^{(k)},\\
\widehat{\boldsymbol{\Gamma}}_j^{(k+1)} &= \left(\sum_{i=1}^{n}\widehat{z}_{ij}^{(k)}\right)^{-1}\sum_{i=1}^{n}\left\{\widehat{s}_{1ij}^{(k)}\left(\mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}_j^{(k)}\right)^\top\left(\mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}_j^{(k)}\right)\right.\\
&\quad-\left[\widehat{\boldsymbol{\Delta}}_j^{(k)}\left(\mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}_j^{(k)}\right)+\left(\mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}_j^{(k)}\right)^\top\left(\widehat{\boldsymbol{\Delta}}_j^{(k)}\right)^\top\right]\widehat{s}_{2ij}^{(k)}\\
&\quad\left.+\widehat{s}_{3ij}^{(k)}\widehat{\boldsymbol{\Delta}}_j^{(k)}\left(\widehat{\boldsymbol{\Delta}}_j^{(k)}\right)^\top\right\},\\
\widehat{\boldsymbol{\Sigma}}_j^{(k+1)} &= \widehat{\boldsymbol{\Gamma}}_j^{(k+1)}+\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\left(\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\right)^\top,\\
\widehat{\boldsymbol{\lambda}}_j^{(k+1)} &= \left(\widehat{\boldsymbol{\Sigma}}_j^{(k+1)}\right)^{1/2}\widehat{\boldsymbol{\Delta}}_j^{(k+1)}/\left(1-\left(\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\right)^\top\left(\widehat{\boldsymbol{\Sigma}}_j^{(k+1)}\right)^{1/2}\widehat{\boldsymbol{\Delta}}_j^{(k+1)}\right).
\end{aligned}
$$

vide demonstração do estimador de $\widehat{p}_j^{(k+1)}$ no Apêndice Demonstração dos estimadores.

Etapa CML: Atualizar $\boldsymbol{\nu}^{(k)}$ maximizando a função de log-verossimilhança marginal, obtendo

$$
\widehat{\boldsymbol{\nu}}^{(k+1)} = \arg\max_{\boldsymbol{\nu}}\sum_{i=1}^{n}\log\left[\sum_{j=1}^{G}\widehat{p}_j^{(k)}\psi_{\text{MESN}_p}\left(\mathbf{y}_i|\mathbf{X}_i\widehat{\boldsymbol{\beta}}_j^{(k+1)},\widehat{\boldsymbol{\Sigma}}_j^{(k+1)},\widehat{\boldsymbol{\lambda}}_j^{(k+1)},\boldsymbol{\nu}\right)\right],
$$

em que $\psi_{\text{MESN}_p}(\cdot|\mathbf{X}_i,\widehat{\boldsymbol{\theta}}_j)$ é definido em (5.2.4).

Neste trabalho foi adotado o critério de informação Bayesiano (BIC), Schwarz (1978) (62), para selecionar o modelo da classe MR-MF-MESN, particularmente se *Skew*-normal ou *Skew*-t. Aqui, a forma de BIC é dada por

$$
\text{BIC} = -2\ell\left(\widehat{\boldsymbol{\theta}}\right)+\rho\log n,
$$

em que $\ell\left(\widehat{\boldsymbol{\theta}}\right)$ é a log-verossimilhança dos dados observados e ρ é o número de parâmetros no modelo e n o tamanho da amostra. Consequentemente, modelos com pontuações BIC pequenas são preferidos. Recentemente McNicholas & Murphy (2008) (48) mostraram a eficácia do BIC na seleção do número de componentes para modelos de mistura gaussiana e Zeller et al. (2018) (75) observam que BIC apresenta melhor performance na seleção de modelos - na classe MEN, quando comparado ao critério de informação Akaike (AIC) e ao critério de determinação eficiente (EDC). Dessa forma, neste trabalho, iremos considerar o BIC para selecionar os modelos na classe MESN.

5.2.2 Definindo os valores iniciais

É bem conhecido que EM e algoritmos relacionados podem ser sensíveis aos valores iniciais escolhidos, veja a discussão em McLachlan & Peel (2000) (45) por exemplo.

A escolha dos valores iniciais para o algoritmo EM no contexto de misturas finitas desempenha um grande papel nas estimativas dos parâmetros, pois modelos dessa natureza podem fornecer uma função de log-verossimilhança multimodal, podendo fazer o método de estimativa de máxima verossimilhança por meio do algoritmo EM não fornecer soluções globais máximas se os valores iniciais estiverem distantes dos valores reais dos parâmetros.

Observe que o número de componentes G é considerado conhecido a priori e, portanto, eles não são parâmetros.

Para esse estudo, os valores iniciais para o modelo MR-MF-MESN estão descritos a seguir e fazem parte dos estudos teóricos e computacionais de Benites (2018) (15).

- Particionar os resíduos em G grupos usando o algoritmo de agrupamento K-means;
- Calcule a proporção de pontos de dados pertencentes ao mesmo *cluster* j, digamos $p_j^{(0)},\ j = 1, \ldots, G$. Este é o valor inicial para p_j;
- Para $\boldsymbol{\beta}_j^{(0)}$, usar a estimativa de mínimos quadrados ordinários no modelo de regressão definido em (5.2.4);
- Para a j-ésima coordenada do vetor de assimetria, considere $\widehat{\rho}_j$ a assimetria amostral para a variável j. Então, $\lambda_j^{(0)} = 3 \times sign(\widehat{\rho}_j)$;
- Para cada grupo j, calcule os valores iniciais $\boldsymbol{\Sigma}_j^{(0)}$ usando a variância amostral das variáveis respostas de acordo com o respectivo componente j;
- Os valores iniciais para ν são considerados iguais a 3 nos casos de t-*Student*.

5.2.3 Critério de Parada de Aceleração de Aitken

Para avaliar a convergência do algoritmo EM são alternados repetidamente até que um critério de parada adequado seja satisfeito. Adotamos o método de aceleração de Aitken (Aitken, 1926) (1). Suponha que o algoritmo converge a uma taxa linear, exigindo muitas iterações para atingir a convergência, o critério utilizado acelera a convergência, como o próprio nome diz. A aceleração de Aitken estima o máximo assintótico da log-verossimilhança em cada iteração k de modo que

$$\ell_\infty^{(k+1)} = \ell^{(k)} + \frac{\ell^{(k+1)} - \ell^{(k)}}{1 - a^{(k)}}, \quad k > 1. \tag{5.2.13}$$

em que $\ell^{(k+1)}$ é a verossimilhança observada avaliada em $\boldsymbol{\theta}^{(k+1)}$ e $a^{(k)} = \frac{\ell^{(k+1)} - \ell^{(k)}}{\ell^{(k)} - \ell^{(k-1)}}$, denotando a aceleração de Aitken na k-ésima iteração, que é muito próxima de 1 na convergência. Conforme recomendado por Böhning et al. (1994) (10), considera-se que o algoritmo atingiu a convergência se $\ell_\infty^{(k+1)} - \ell^{(k+1)} < \varepsilon$ em que ε é a tolerância desejada. Nos exemplos numéricos do Capítulo 7, um valor padrão de $\varepsilon = 10^{-6}$ foi usado para encerrar as iterações. Observe que o procedimento acima também é aplicável para o caso simples $G = 1$.

6 MISTURAS FINITAS DE ESPECIALISTAS DE MODELOS DE REGRESSÃO SOB AS DISTRIBUIÇÕES MESN

No contexto do uso de Modelos Lineares de Especialistas (MoE) (Jacobs et al., 1991) (26), Nguyen & McLachlan (2016) (53) comentam que a maioria das aplicações desses modelos utiliza uma distribuição Gaussiana para erro aleatório, conhecidas por serem sensíveis a valores discrepantes. Abordando aspectos sobre a limitação do modelo G-componente clássico (DeSarbo e Cron, 1988 (20); Jones e McLachlan, 1992) (28), onde há suposição de proporções de misturas constantes, Xue & Yao (2021), trabalhando com MoE univariado, argumentam que em muitas aplicações as covariáveis carregam informações importantes sobre as proporções de misturas. Mirfarah et al.(2021) (52), empregando o algoritmo EM, propõem um modelo robusto, baseado em mistura de escala da classe de distribuições normais (MEN) no contexto univariado, para resolver o desafio da sensibilidade do MoE clássico a observações atípicas e discrepantes. Em Chamroukhi (2017) (17) vemos uma proposta de modelo MoE univariado para lidar com dados assimétricos, caudas pesadas e observações atípicas introduzindo a distribuição *Skew*-t na modelagem de especialistas.

Um aspecto favorável à modelagem empregando MoE está na possibilidade de estimar em melhores condições a classificação das observações. De acordo com Mengersen et al. (2011) (51), a terminologia usada na literatura sobre mistura de modelos de especialistas chama as densidades ($\boldsymbol{\psi}$) de 'especialistas' e as proporções ($\boldsymbol{p}$) de 'portas', enquanto na formulação original em Jacobs et al. (1991) (26), o modelo para as densidades é um modelo linear geral e para as proporções é um modelo de regressão logística multinomial. A classificação das observações é um aspecto relevante no modelo MoE-MF-MESN, segundo Mengersen et al. (2011) (51) a inclusão de covariáveis por meio de mistura de especialistas pode fornercer resultados de agrupamento diferentes e muitas vezes com uma estrutura mais clara devido ao uso das covariáveis em duas fontes de informação, tanto nas densidades (especialistas) como nas proporções (portas). Esses autores reforçam a importância de considerar que a escolha das covariáveis precisa ser orientada pela interpretação da estrutura latente no contexto da aplicação, de como as covariáveis entram no modelo. Abordando estudos sobre a relação da concentração de antioxidantes no sangue e nos tecidos do corpo e sua potencial proteção contra danos de oxidativos às células e tecidos, Schlattmann (2009) (61) ilustra a questão da importância da identificação da estrutura latente usando os dados sobre os níveis plasmáticos de betacaroteno de Nierenberg et al. (1989) (54) e Stukel (2008) (66), a fim de verificar se existem subgrupos latentes presentes e, em caso afirmativo, se essa heterogeneidade pode ser explicada por covariáveis como idade, sexo ou tabagismo. Como conclusão, empregando um modelo de mistura finita ajustado por covariáveis foram encontradas quatro subpopulações, para as quais o efeito do betacaroteno na dieta foi considerado diferente em cada uma delas, enquanto que, para as demais covariáveis não

foram identificados comportamentos diferentes nessas quatro populações.

A proposta apresentada neste capítulo consiste em estender os estudos do MoE, trabalhando com variáveis respostas multivariadas, mantendo-se o parâmetro ($\boldsymbol{\nu}$) para modelar caudas pesadas e o parâmetro ($\boldsymbol{\lambda}$) para modelar dados assimétricos no contexto de MESN.

6.1 O MODELO PROPOSTO

No MoE, p_j é modelado como a função logística multinomial da entrada $\mathbf{r}$, sendo uma função de passagem. Assim, para o caso de MF-MESN, pode-se obter fdp do MoE redefinindo (5.2.4) da seguinte forma

$$f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j\left(\mathbf{r};\boldsymbol{\alpha}\right)\boldsymbol{\psi}(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}_j), \quad p_j \geq 0, \quad \sum_{j=1}^{G} p_j = 1, \tag{6.1.1}$$

em que $\mathbf{r}$ e $\mathbf{x}$ podem ser diferentes, $\mathbf{r} = (1, r_1, \ldots, r_{q-1})^\top \in \mathbb{R}^q$, $\mathbf{x} \in \mathbb{R}^p$ e $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_{G-1}^\top\right)^\top$, sendo $\boldsymbol{\alpha}_l = \left(\alpha_{l0}, \ldots, \alpha_{l(q-1)}\right)^\top$, tal que

$$p_j\left(\mathbf{r};\boldsymbol{\alpha}\right) = P(Z = j|\mathbf{r}) = \frac{\exp\left\{\boldsymbol{\alpha}_j^\top \mathbf{r}\right\}}{1 + \sum_{l=1}^{G-1}\exp\left\{\boldsymbol{\alpha}_l^\top \mathbf{r}\right\}} \tag{6.1.2}$$

e o conjunto de parâmetros do modelo é $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{G-1})^\top$, sendo $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu})$, $j = 1, \ldots, G$, tal que assume-se $\boldsymbol{\nu}_1 = \ldots = \boldsymbol{\nu}_G = \boldsymbol{\nu}$, conforme dito na Definição 4.1.1.

6.1.1 Estimação da Máxima Verossimilhança pelo Algoritmo EM

Como as diferenças entre o modelo definido no Capítulo 5 e o modelo proposto neste Capítulo ocorre na modelagem da proporção através da expressão (6.1.2), temos que a representação hierárquica dada em (5.2.6) - (5.2.9) sofre alteração apenas em (5.2.9). Assim, a função de log-verossimilhança dos dados completos, $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{u}^\top, \mathbf{t}^\top, \mathbf{z}^\top)^\top$, é dada por

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) = C + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\Big[&\log p_j\left(\mathbf{r};\boldsymbol{\alpha}\right) - \frac{1}{2}\log|\boldsymbol{\Gamma}_j| \\ &-\frac{1}{2}u_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j - \boldsymbol{\Delta}_j t_i)^\top\boldsymbol{\Gamma}_j^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j - \boldsymbol{\Delta}_j t_i) + \log(h(u_i;\boldsymbol{\nu}))\Big],\end{aligned}$$

Fazendo $\widehat{\boldsymbol{\theta}}_j^{(k)} = (\widehat{\boldsymbol{\alpha}}_j^{(k)}, \widehat{\boldsymbol{\beta}}_j^{(k)\top}, \widehat{\boldsymbol{\sigma}}_j^{(k)}, \widehat{\boldsymbol{\lambda}}_j^{(k)}, \boldsymbol{\nu}^{(k)})^\top$ os estimadores de $\boldsymbol{\theta}$ na k-ésima iteração. O parâmetro $\boldsymbol{\nu}$ assume os mesmos valores conforme descrito na Definição 4.1.1.

Dessa forma, a função $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ é

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = C &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij}^{(k)} \log p_j\left(\mathbf{r};\boldsymbol{\alpha}\right) \\
&- \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij}^{(k)} \log \boldsymbol{\Gamma}_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{1ij}^{(k)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^\top \boldsymbol{\Gamma}_j^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j) \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{2ij}^{(k)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^\top \boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{3ij}^{(k)} \boldsymbol{\Delta}_j^\top \boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j.
\end{aligned}
$$

em que $\widehat{s}_{1ij}^{(k)}$, $\widehat{s}_{2ij}^{(k)}$ e $\widehat{s}_{3ij}^{(k)}$ são calculadas conforme Subseção **5.2.1** e

$$
\widehat{z}_{ij}^{(k)} = \frac{p_j\left(\mathbf{r};\widehat{\boldsymbol{\alpha}}_j^{(k)}\right)\psi_{\text{MESN}_p}(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}_j^{(k)})}{\sum_{l=1}^{G} p_j\left(\mathbf{r};\widehat{\boldsymbol{\alpha}}_j^{(k)}\right)\psi_{\text{MESN}_p}(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}_l^{(k)})}. \tag{6.1.3}
$$

Assim, o algoritmo ECME para estimativa de máxima verossimilhança de $\boldsymbol{\theta}$ é definido como segue:

Etapa E: Dado $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, obter $\widehat{z}_{ij}$ (6.1.3), $\widehat{s}_{1ij}$, $\widehat{s}_{2ij}$, $\widehat{s}_{3ij}$ (Subseção **5.2.1**), para $i = 1,\ldots,n$ e $j = 1,\ldots,G$.

Etapa CM: Atualizar $\widehat{\boldsymbol{\theta}}^{(k+1)}$ maximizando $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ sobre $\boldsymbol{\theta}$, empregando as expressões da Subseção **5.2.1** para os estimadores $\widehat{\boldsymbol{\beta}}_j^{(k+1)}$, $\widehat{\boldsymbol{\Gamma}}_j^{(k+1)}$, $\widehat{\boldsymbol{\Delta}}_j^{(k+1)}$, $\widehat{\boldsymbol{\Sigma}}_j^{(k+1)}$ *e* $\widehat{\boldsymbol{\lambda}}_j^{(k+1)}$, sendo para $\widehat{\boldsymbol{\alpha}}_j^{(k+1)}$, $j = 1,\ldots,G$

$$
\widehat{\boldsymbol{\alpha}}_j^{(k+1)} = 4\left(\sum_{i=1}^{n}\mathbf{r}_i\mathbf{r}_i^\top\right)^{-1}\left(\sum_{i=1}^{n}\widehat{z}_{ij}^{(k+1)}\left[1 - p_j\left(\mathbf{r}_i;\widehat{\boldsymbol{\alpha}}^{(k)}\right)\right]\mathbf{r}_i\right) + \widehat{\boldsymbol{\alpha}}_j^{(k)}. \tag{6.1.4}
$$

O parâmetro $\boldsymbol{\nu}$ é estimado conforme Etapa CML do Capítulo 5.

Para a estimação de 6.1.4 utiliza-se aproximação de primeira ordem numa série de Taylor, conforme demonstração do estimador de $\boldsymbol{\alpha}_j^{(k+1)}$ vista no Apêndice Demonstração de estimadores.

6.1.2 Definindo os valores iniciais e Critério de Parada

A diferença para os valores iniciais entre o modelo MR-MF-MESN e o modelo MoE-MF-MESN consiste no acréscimo de $\boldsymbol{\alpha}^{(0)}$ e na transformação do vetor de proporções numa matriz de proporções com n linhas iguais a $p_1,\ldots,p_G$, podendo os valores iniciais para $\boldsymbol{\beta}_j^{(0)}$, $\boldsymbol{\Sigma}_j^{(0)}$, $\lambda_j^{(0)}$ e $\boldsymbol{\nu}_j^{(0)}$ serem aqueles definidos na Subseção **5.2.2**.

Assim, os valores iniciais para a matriz de proporções e para $\boldsymbol{\alpha}^{(0)}$ estão descritos a seguir:

- Particionar os resíduos em G grupos usando o algoritmo de agrupamento K-means;
- Calcule a proporção de pontos de dados pertencentes ao mesmo *cluster* j, digamos $p_j^{(0)}$, $j = 1,\ldots,G$. Este é o valor inicial para p_j. Porém, como o resultado da

expressão 6.1.2 é uma matriz $n \times G$, cria-se uma matriz $p_{n,G}$ em que $p_{ij} = p_j$, $i = 1, \ldots, n$;

- Para inicializar $\boldsymbol{\alpha}$, uma das formas consiste em definir $\boldsymbol{\alpha}^{(0)} = 0$.

Desta forma, com essas configurações iniciais para os estimadores dos parâmetros, tendo as mesmas proporções para todas as observações e o vetor $\boldsymbol{\alpha}$ como zero, no início da primeira iteração o modelo MoE-MF-MESN será reduzido ao modelo MR-MF-MESN.

Quanto ao critério de parada, pode ser usado o de Aceleração de Aitken, conforme visto na Subseção **5.2.3**.

7 EXEMPLOS NUMÉRICOS

Exemplos numéricos considerando dados simulados e reais são apresentados para ilustrar o modelo e os resultados inferenciais desenvolvidos.

7.1 ESTUDOS DE SIMULAÇÃO

Os estudos de simulação avaliaram o desempenho e as propriedades dos estimadores de máxima verossimilhança das distribuições *Skew*-normal e *Skew*-t nos modelos MR-MF-MESN e MoE-MF-MESN. São analisados conjuntos de dados artificiais e reais, sendo as estimativas obtidas por meio do algoritmo EM apresentado nas seções 3.3, **4.1.3**, **5.2.1** e **6.1.1**.

7.1.1 Cenários Simulados

Foram realizadas 500 replicações dos modelos a seguir, respectivamente MR-MF-MESN e MoE-MF-MESN, empregando a distribuição *Skew*-t.

$$\begin{aligned} f\left(\mathbf{y}_i|\boldsymbol{\theta}\right) = p\psi_{\text{MESN}_2}\left(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_1 + b(\boldsymbol{\nu})\boldsymbol{\Delta}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\lambda}_1, \boldsymbol{\nu}\right) \\ + (1-p)\psi_{\text{MESN}_2}\left(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_2 + b(\boldsymbol{\nu})\boldsymbol{\Delta}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\lambda}_2, \boldsymbol{\nu}\right) \end{aligned} \tag{7.1.1}$$

$$\begin{aligned} f\left(\mathbf{y}_i|\boldsymbol{\theta}\right) = p(\mathbf{r}_i, \boldsymbol{\alpha})\psi_{\text{MESN}_2}\left(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_1 + b(\boldsymbol{\nu})\boldsymbol{\Delta}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\lambda}_1, \boldsymbol{\nu}\right) \\ + \left[1 - p(\mathbf{r}_i, \boldsymbol{\alpha})\right]\psi_{\text{MESN}_2}\left(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_2 + b(\boldsymbol{\nu})\boldsymbol{\Delta}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\lambda}_2, \boldsymbol{\nu}\right) \end{aligned} \tag{7.1.2}$$

em que

$$\begin{aligned} \boldsymbol{\beta}_1 &= \left(\beta_{11}, \beta_{21}, \beta_{31}\right)^\top, \\ \boldsymbol{\beta}_2 &= \left(\beta_{12}, \beta_{22}, \beta_{32}\right)^\top, \\ \boldsymbol{\Sigma}_j &= \begin{bmatrix} \sigma_{j.11} & \sigma_{j.12} \\ \sigma_{j.21} & \sigma_{j.22} \end{bmatrix}, \\ \boldsymbol{\lambda}_j &= \left(\lambda_{1j}, \lambda_{2j}\right)^\top, \quad j = 1, 2. \end{aligned}$$

com $\boldsymbol{\beta}_1 = (-1, -4, -3)^\top$, $\boldsymbol{\beta}_2 = (3, 7, 9)^\top$, $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $\boldsymbol{\lambda}_1 = (3, 5)^\top$, $\boldsymbol{\lambda}_2 = (1, 4)^\top$, $p = 0,3$, $\boldsymbol{\alpha} = (0.7, 1, 2)^\top$, $\boldsymbol{\nu} = 3$ e $\mathbf{X}_i = [1,\ X_{i.1},\ X_{i.2}]$, de modo que $X_{i.1} \sim U(0, 1)$ e $X_{i.2} \sim U(2, 4)$, $\mathbf{r}_i = [1,\ r_{i.1},\ r_{i.2}]$, de modo que $r_{i.1} \sim U(-2, 1)$ e $r_{i.2} \sim U(-1, 1)$, $i = 1, \ldots, n$ (n = 50, 100, 250, 500, 1000 e 2500).

7.1.2 Recuperação dos Parâmetros

Para cada amostra gerada artificialmente dos modelos definidos acima, estimam-se os parâmetros via algoritmo EM e registram-se as estimativas provenientes das 500

replicações consideradas. Os resultados estão apresentados nas Tabelas 1 e 2, onde, para os diferentes tamanhos de amostra (n), são mostradas as médias amostrais e o desvios padrão amostrais (DPA) das 500 estimativas de cada um dos parâmetros relacionados ao ajuste de uma distribuição *Skew*-t para MR-MF-MESN e MoE-MF-MESN, respectivamente.

j	Medidas	Parâmetros									
		β_{1j}	β_{2j}	β_{3j}	$\sigma_{j,11}$	$\sigma_{j,12}$	$\sigma_{j,22}$	λ_{1j}	λ_{2j}	p_1	ν
n=50											
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0.3)	(3)
1	Média	-1,2072	-4,0309	-2,9297	3,7069	0,5974	2,9125	3,7643	7,5847	0,2999	6,4563
	DPA	2,1605	1,3552	0,6391	2,9385	2,1001	2,5222	9,4629	9,7661	0,0676	15,4134
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
2	Média	3,0308	6,9900	8,9966	2,4054	1,0862	1,9556	1,8271	6,9078		
	DPA	0,8999	0,4911	0,2583	1,3314	1,1555	1,2607	3,4840	6,0135		
n=100											
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,3)	(3)
1	Média	-0,9392	-3,9561	-3,0133	3,3959	0,8507	2,9028	4,3100	7,0893	0,3031	3,3893
	DPA	1,1805	0,7035	0,3518	1,6663	1,3260	1,5153	4,3408	5,4782	0,0469	1,2330
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
2	Média	3,0335	7,0032	8,9971	2,1900	1,0923	2,0460	1,4022	5,0577		
	DPA	0,5698	0,3132	0,1608	0,9352	0,8384	0,9647	1,7256	3,8303		
n=250											
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,3)	(3)
1	Média	-0,8656	-4,0188	-3,0077	3,3539	1,1709	3,2924	3,6435	5,8468	0,2998	3,1988
	DPA	0,6690	0,3570	0,1902	1,1423	0,8389	1,0996	2,2425	3,0050	0,0289	0,5710
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
2	Média	3,0338	7,0002	9,0004	2,1523	1,0841	2,0935	1,0912	4,0885		
	DPA	0,3742	0,1947	0,0963	0,5925	0,5713	0,6883	0,8890	1,6113		
n=500											
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,3)	(3)
1	Média	-0,9039	-4,0072	-2,9912	3,3801	1,2510	3,4169	3,4438	5,5951	0,3010	3,185
	DPA	0,4374	0,2551	0,1241	0,8292	0,5890	0,7903	1,4086	1,9718	0,0220	0,4296
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
2	Média	3,0694	6,9963	8,9971	2,1755	1,1430	2,1726	1,1213	4,0832		
	DPA	0,2733	0,1314	0,0674	0,4538	0,4477	0,5755	0,6489	1,3065		
n=1000											
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,3)	(3)
1	Média	-0,8851	-3,9962	-2,9989	3,3737	1,2961	3,4033	3,4362	5,5342	0,2988	3,1979
	DPA	0,3006	0,1646	0,0845	0,5672	0,4392	0,5457	0,9760	1,3169	0,0144	0,2750
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
2	Média	3,0695	6,9944	8,9988	2,1686	1,1544	2,1925	1,1424	4,1337		
	DPA	0,1909	0,0898	0,0465	0,3244	0,3338	0,4416	0,4918	1,0104		
n=2500											
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,3)	(3)
1	Média	-0,8831	-4,0011	-3,0029	3,3368	1,2657	3,3908	3,3599	5,4507	0,3003	3,1830
	DPA	0,1795	0,1059	0,0510	0,3513	0,2660	0,3508	0,6186	0,8737	0,009	0,2017
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
2	Média	3,0656	7,0030	9,0010	2,1718	1,1629	2,2330	1,1763	4,2833		
	DPA	0,1213	0,0557	0,0273	0,2236	0,2256	0,3046	0,3313	0,6804		

Tabela 1 – Recuperação dos Parâmetros: os valores verdadeiros dos parâmetros (Verd) estão entre parênteses. Média e DPA são as respectivas médias e desvios padrão amostrais das estimativas dos parâmetros provenientes do ajuste de uma distribuição *Skew-t* para MR-MF-MESN, com diferentes configurações de tamanho de amostra (n) com base em 500 replicações.

j	Medidas	Parâmetros											
		β_{1j}	β_{2j}	β_{3j}	$\sigma_{j,11}$	$\sigma_{j,12}$	$\sigma_{j,22}$	λ_{1j}	λ_{2j}	α_1	α_2	α_3	ν
n=50													
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,7)	(1)	(2)	(3)
1	Média	-0,9992	-3,9577	-2,9987	3,4043	0,7604	2,8922	5,1293	8,6557	0,7764	1,1056	2,2498	4,7164
	DPA	1,3404	0,7763	0,3956	1,9607	1,3261	1,6493	6,0045	7,4385	0,5161	0,5407	0,7984	10,0056
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)				
2	Média	2,9331	7,0472	9,0258	2,6028	1,1608	1,9782	2,5122	8,7304				
	DPA	1,1349	0,6556	0,3394	1,9125	1,5787	1,7427	6,1963	8,5490				
n=100													
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,7)	(1)	(2)	(3)
1	Média	-0,9841	-3,9622	-2,9908	3,3104	1,0435	3,1212	4,0181	6,5049	0,0,7257	1,0375	2,0724	3,8831
	DPA	0,8215	0,4699	0,2416	1,2768	0,9358	1,2402	3,1124	4,4694	0,3061	0,3108	0,4965	6,3613
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)				
2	Média	3,0325	7,0190	9,0072	2,3345	1,1791	2,1299	1,6519	5,7756				
	DPA	0,7034	0,4041	0,1950	1,1502	1,0603	1,2577	2,3974	4,7392				
n=250													
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,7)	(1)	(2)	(3)
1	Média	-0,8622	-3,9920	-3,0114	3,3752	1,2161	3,3516	3,3842	5,4760	0,7046	1,0235	2,0425	3,2200
	DPA	0,4424	0,2782	0,1252	0,8779	0,6490	0,8390	1,5167	2,0809	0,1883	0,1950	0,3120	0,5655
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)				
2	Média	3,0666	6,9765	8,9975	2,2025	1,1321	2,0951	1,1666	4,1613				
	DPA	0,4538	0,2311	0,1208	0,7523	0,7254	0,8542	1,2012	2,1705				
n=500													
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,7)	(1)	(2)	(3)
1	Média	-0,8878	-3,9907	-2,9994	3,3728	1,2691	3,4041	3,2785	5,3298	0,7094	1,0033	2,0110	3,2190
	DPA	0,3311	0,1780	0,0942	0,5844	0,4407	0,5833	0,9198	1,3185	0,1372	0,1395	0,2207	0,3923
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)				
2	Média	3,0405	7,0017	9,0009	2,1704	1,1143	2,1087	1,0830	3,9822				
	DPA	0,3187	0,1609	0,0794	0,4984	0,4831	0,6099	0,7439	1,4122				
n=1000													
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,7)	(1)	(2)	(3)
1	Média	-0,8813	-3,9946	-3,0008	3,3813	1,2877	3,4127	3,3203	5,3479	0,7025	1,0076	2,0150	3,2079
	DPA	0,2283	0,1329	0,0637	0,4265	0,3177	0,4099	0,6668	0,9078	0,0903	0,0940	0,1551	0,2880
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)				
2	Média	3,0605	6,9972	8,9996	2,1803	1,1382	2,1713	1,1159	4,0504				
	DPA	0,2255	0,1113	0,0555	0,3772	0,3885	0,4825	0,5693	1,0756				
n=2500													
	Verd	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0,7)	(1)	(2)	(3)
1	Média	-0,8880	-4,0055	-3,0000	3,3510	1,2684	3,3810	3,3006	5,3414	0,7018	1,0036	2,0127	3,1920
	DPA	0,1391	0,0795	0,0371	0,2644	0,1886	0,2557	0,4314	0,6296	0,0575	0,0598	0,0933	0,1769
	Verd	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)				
2	Média	3,0648	6,9968	9,0001	2,1689	1,1511	2,2001	1,1624	4,2295				
	DPA	0,1486	0,0708	0,0341	0,2545	0,2647	0,3557	0,4088	0,8573				

Tabela 2 – Recuperação dos Parâmetros: os valores verdadeiros dos parâmetros (Verd) estão entre parênteses. Média e DPA são as respectivas médias e desvios padrão amostrais das estimativas dos parâmetros provenientes do ajuste de uma distribuição *Skew-t* para MoE-MF-MESN, com diferentes configurações de tamanho de amostra (n) com base em 500 replicações.

Em geral, nota-se que à medida que o tamanho amostra aumenta o desvio padrão tende a diminuir e a média torna-se mais próxima do valor verdadeiro do parâmetro. Estes resultados podem ser visualizados através de *boxplots*.

Os *boxplots* relativos às estimativas dos parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\sigma^2}$, $\boldsymbol{\lambda}$ e $\boldsymbol{p}$ para o modelo MR-MF-MESN são mostrados respectivamente nas Figuras 2, 4, 6 e 8, enquanto os *boxplots* relativos às estimativas dos parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\sigma^2}$, $\boldsymbol{\lambda}$ e $\boldsymbol{\alpha}$ para o modelo MoE-MF-MESN são mostrados respectivamente nas Figuras 3, 5, 7 e 9. Como esperado, estes gráficos indicam que o viés e a variabilidade das estimativas dos parâmetros diminuem quando o tamanho da amostra aumenta, o que está essencialmente de acordo com as propriedades assintóticas dos estimadores de máxima verossimilhança, ou seja, consistência.

Além disso, a consistência dos estimadores de máxima verossimilhança pode ser estudada através do cálculo do EQM-R, tal que

$$\text{EQM-R}\left(\theta_j\right) = \sqrt{\frac{1}{500}\sum_{i=1}^{500}\left(\widehat{\theta}_{ij} - \theta_j\right)^2}$$

em que θ_j é o j-ésimo parâmetro, definido na Subseção **7.1.1**. Os resultados para os estimadores dos parâmetros do MR-MF-MESN são mostrados nas Figuras 9, 11, 13, 15, 17, 19 e Figuras 10, 12, 14, 16, 18, 20 e 22 e do MoE-MF-MESN são apresentados nas Figuras 11, 13, 15, 17, 19, 21 e 23. Observa-se que os EQM-R tendem a se aproximar de zero quando o tamanho da amostra aumenta, indicando que os estimadores provenientes dos algoritmos EM propostos, nas Subseções **5.2.1** e **6.1.1**, são consistentes.

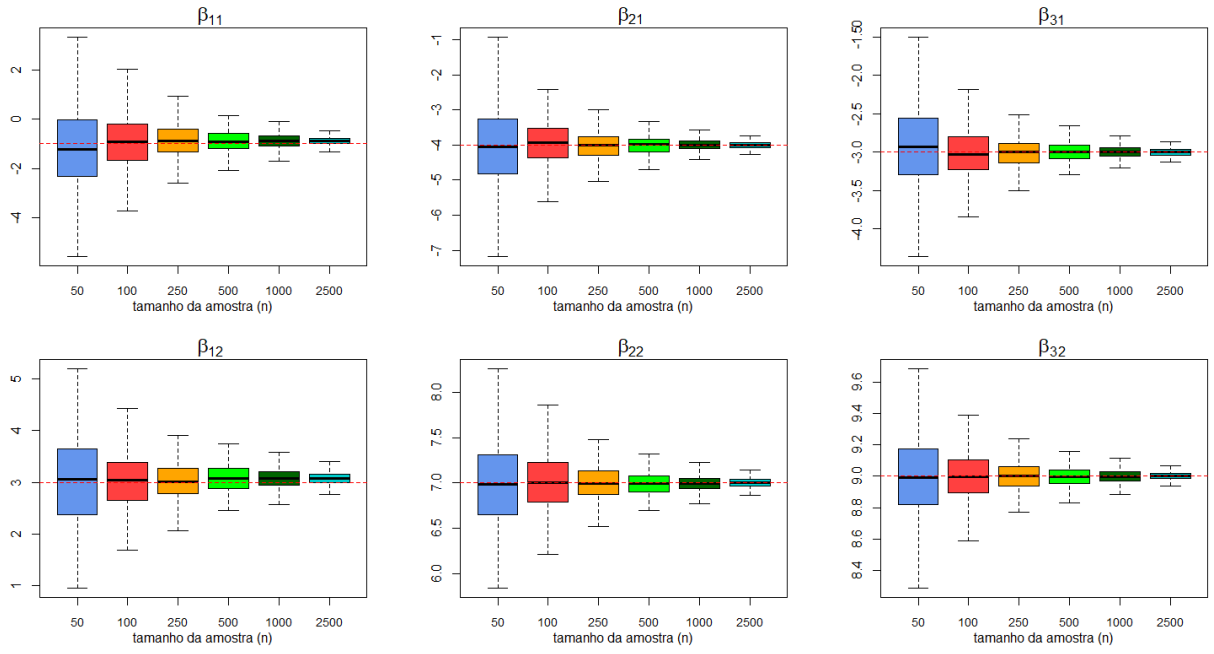

Figura 2 – *Boxplots* das estimativas de $\boldsymbol{\beta}$ para o modelo MR-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro.

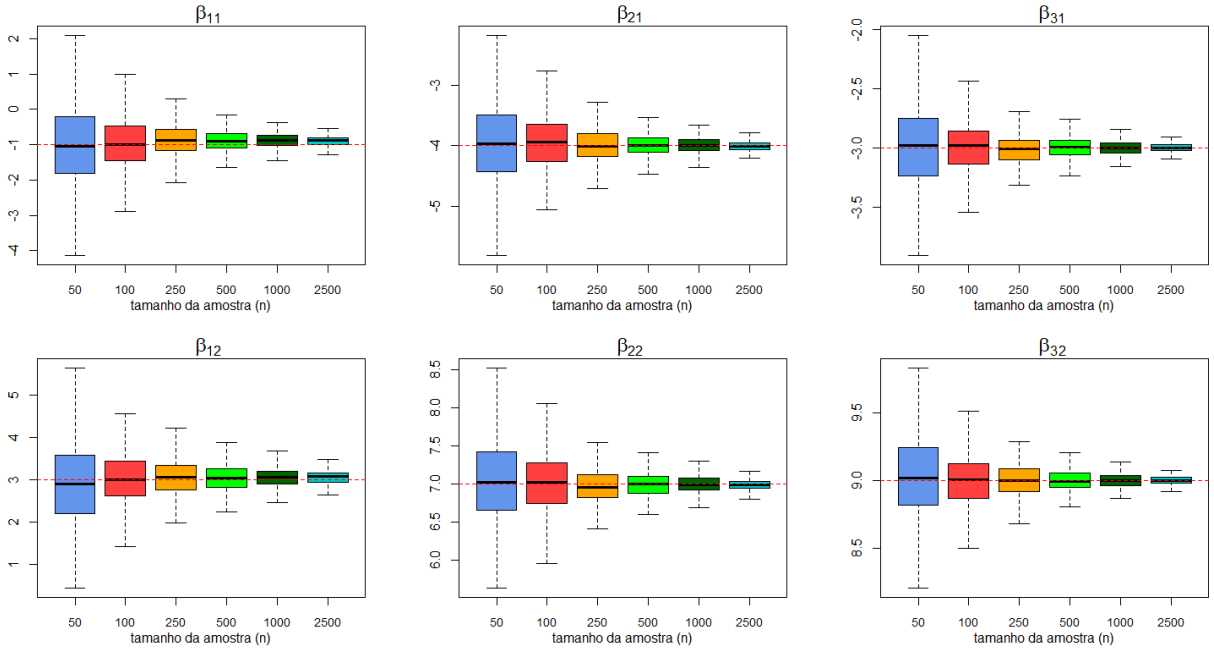

Figura 3 – *Boxplots* das estimativas de $\boldsymbol{\beta}$ para o modelo MoE-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro.

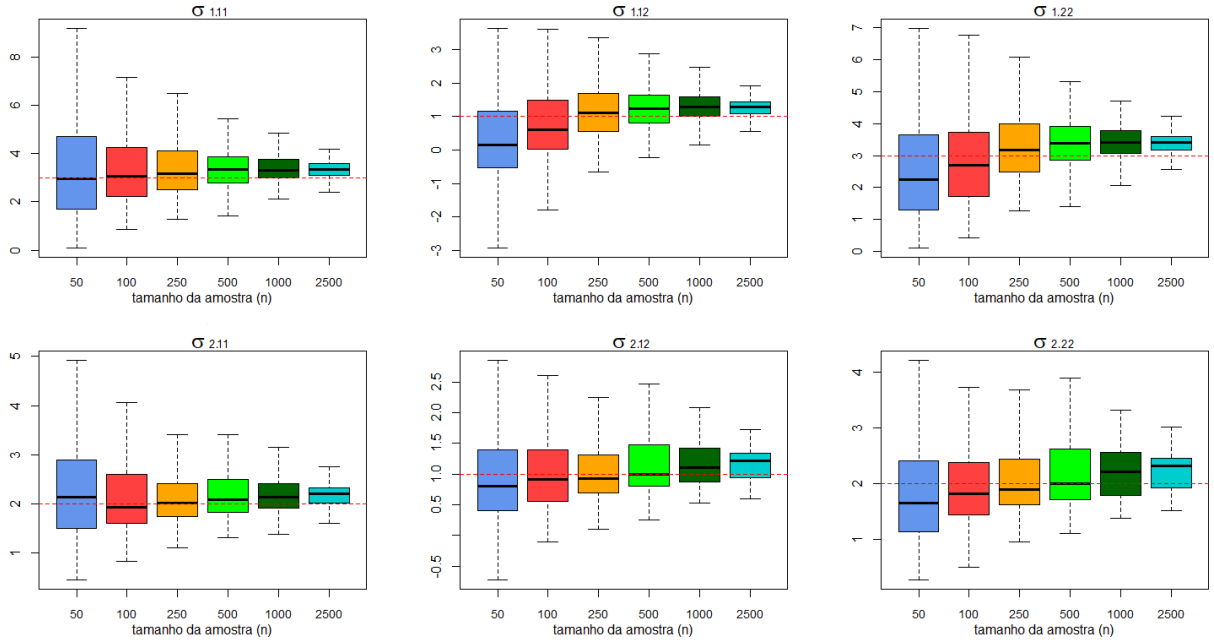

Figura 4 – *Boxplots* das estimativas de $\boldsymbol{\sigma}$ para o modelo MR-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro.

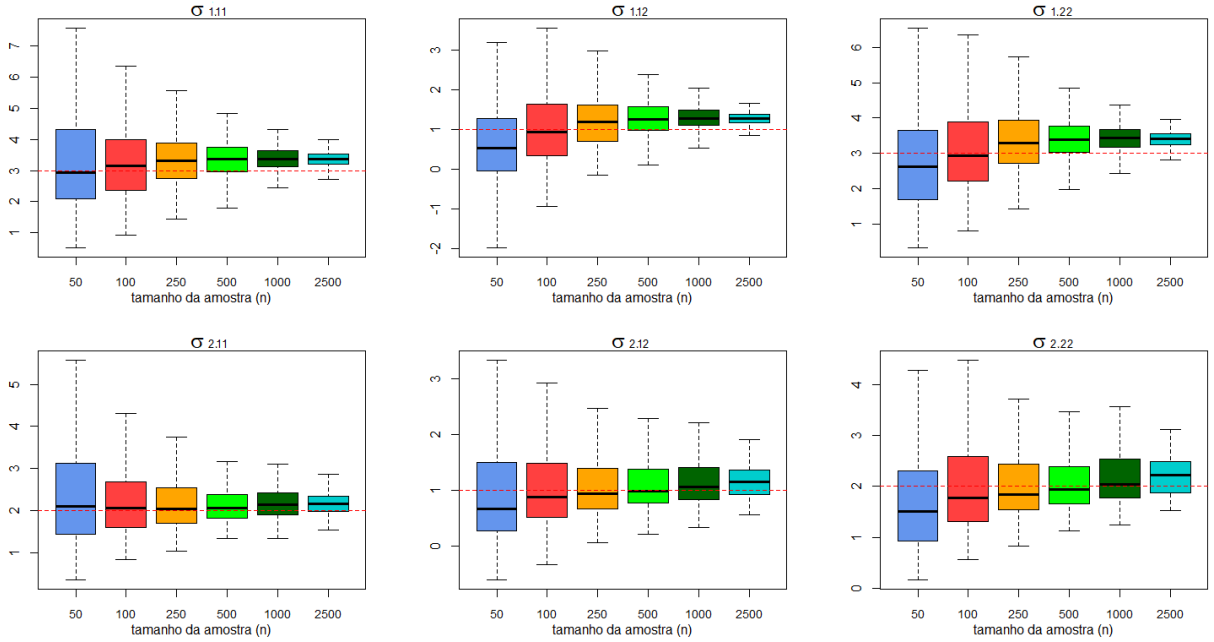

Figura 5 – *Boxplots* das estimativas de $\boldsymbol{\sigma}$ para o modelo MoE-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro.

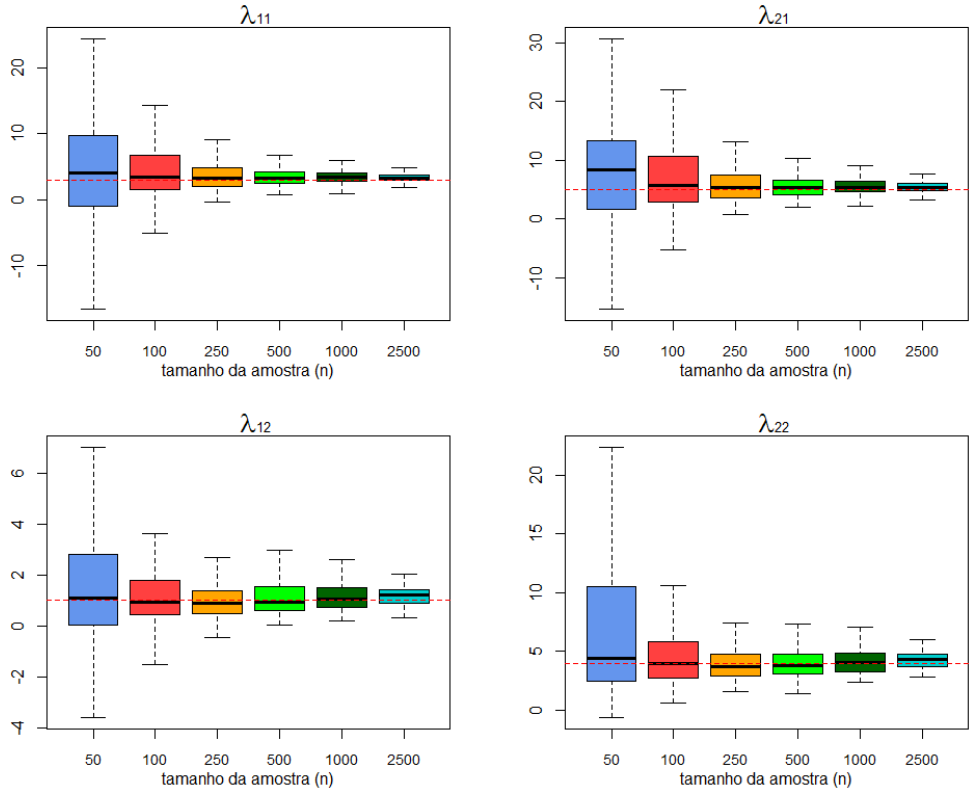

Figura 6 – *Boxplots* das estimativas de $\boldsymbol{\lambda}$ para o modelo MR-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro.

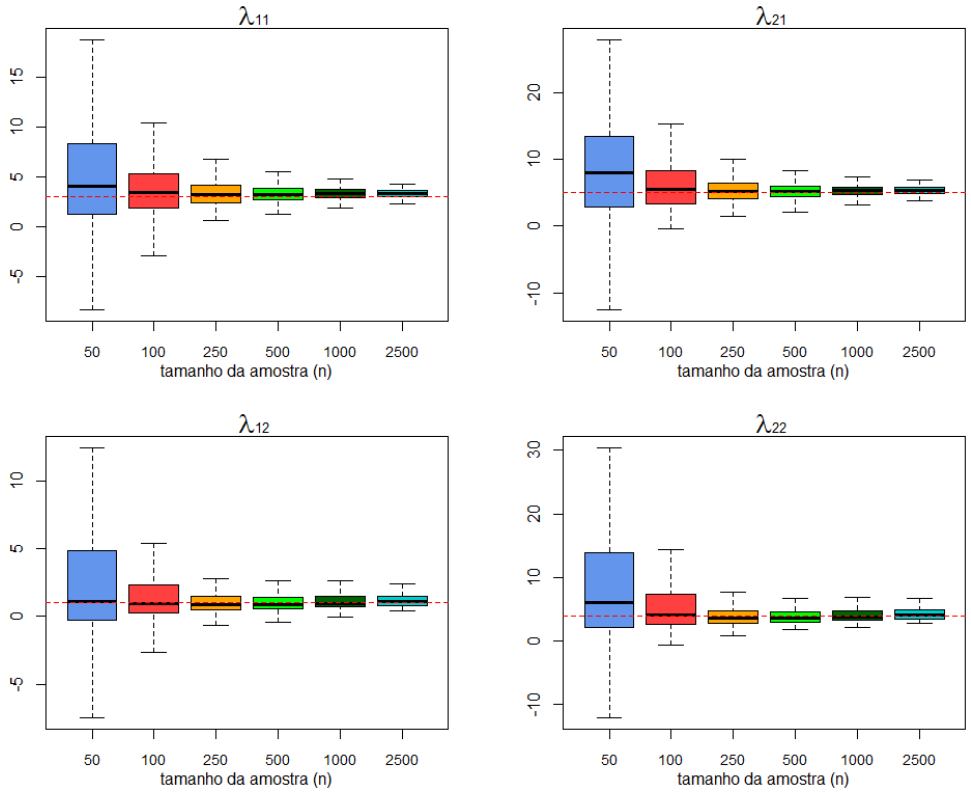

Figura 7 – *Boxplots* das estimativas de $\boldsymbol{\lambda}$ para o modelo MoE-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro.

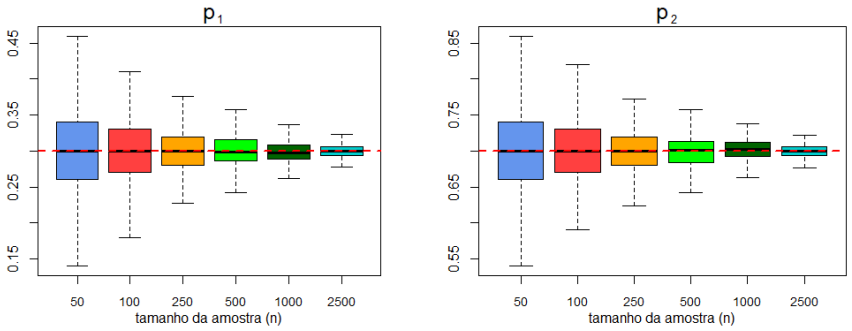

Figura 8 – *Boxplots* das estimativas de $\boldsymbol{p}$ para o modelo MR-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro.

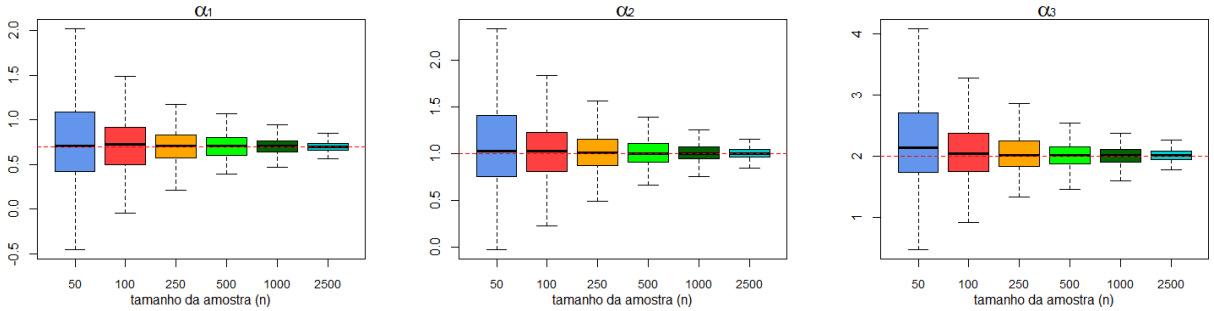

Figura 9 – *Boxplots* das estimativas de $\boldsymbol{\alpha}$ para o modelo MoE-MF-MESN. A linha horizontal tracejada indica o valor real do parâmetro.

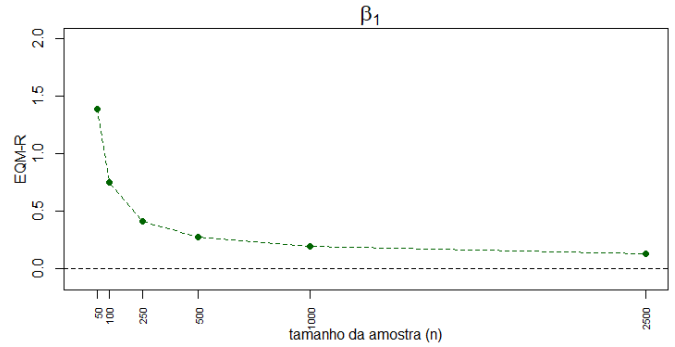

Figura 10 – EQM-R das médias das estimativas de $\boldsymbol{\beta}_1$ para o modelo MR-MF-MESN.

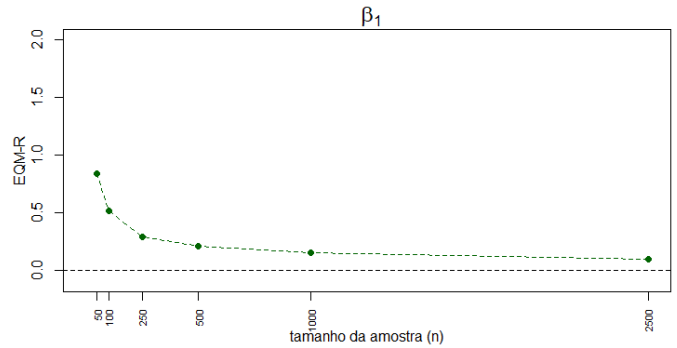

Figura 11 – EQM-R das médias das estimativas de $\boldsymbol{\beta}_1$ para o modelo MoE-MF-MESN.

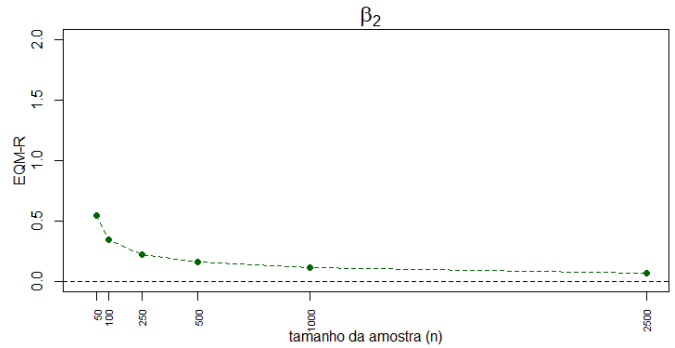

Figura 12 – EQM-R das médias das estimativas de $\boldsymbol{\beta}_2$ para o modelo MR-MF-MESN.

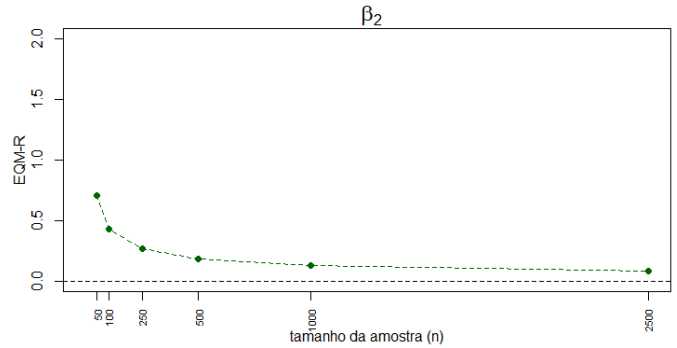

Figura 13 – EQM-R das médias das estimativas de $\boldsymbol{\beta}_2$ para o modelo MoE-MF-MESN.

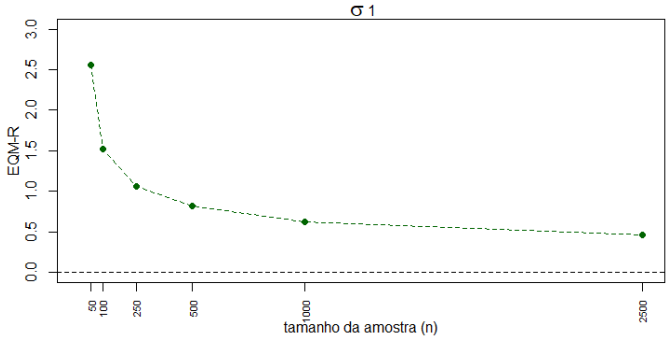

Figura 14 – EQM-R das médias das estimativas de $\boldsymbol{\sigma_1^2}$ para o modelo MR-MF-MESN.

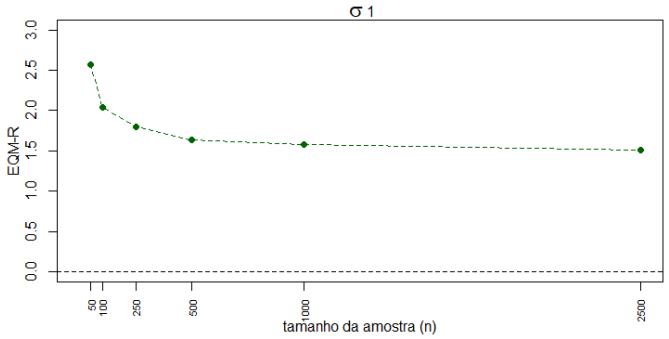

Figura 15 – EQM-R das médias das estimativas de $\boldsymbol{\sigma_1^2}$ para o modelo MoE-MF-MESN.

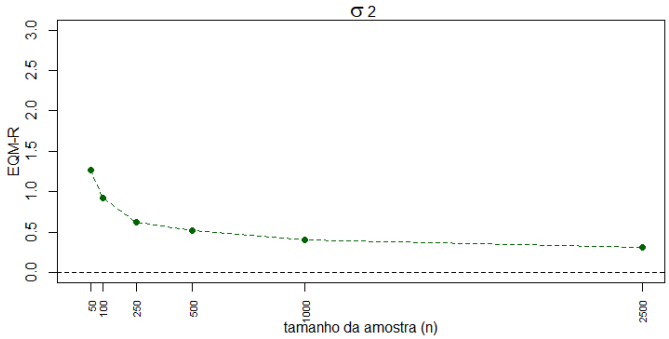

Figura 16 – EQM-R das médias das estimativas de $\boldsymbol{\sigma_2^2}$ para o modelo MR-MF-MESN.

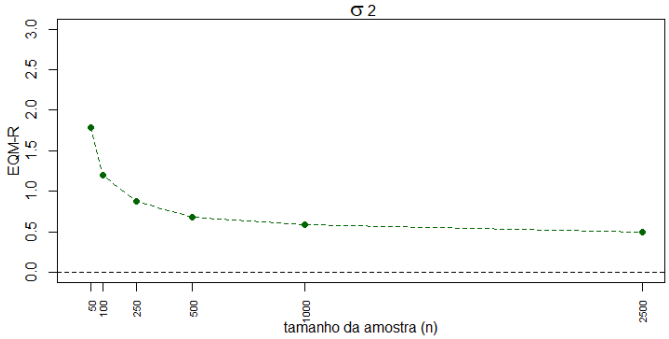

Figura 17 – EQM-R das médias das estimativas de $\boldsymbol{\sigma_2^2}$ para o modelo MoE-MF-MESN.

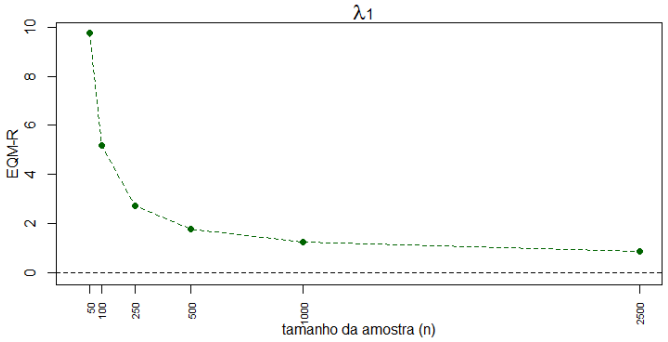

Figura 18 – EQM-R das médias das estimativas de $\boldsymbol{\lambda}_1$ para o modelo MR-MF-MESN.

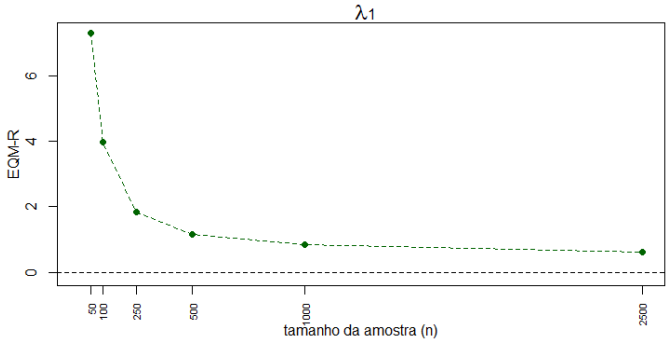

Figura 19 – EQM-R das médias das estimativas de $\boldsymbol{\lambda}_1$ para o modelo MoE-MF-MESN.

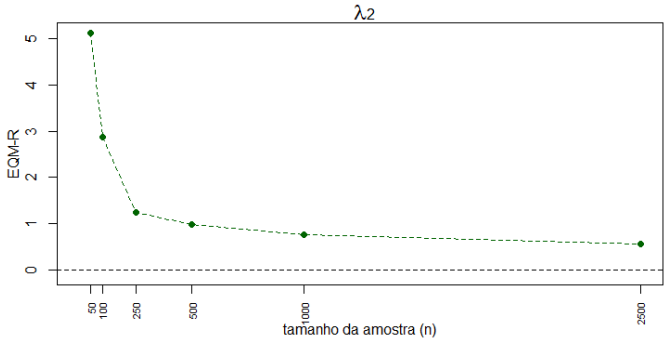

Figura 20 – EQM-R das médias das estimativas de $\boldsymbol{\lambda}_2$ para o modelo MR-MF-MESN.

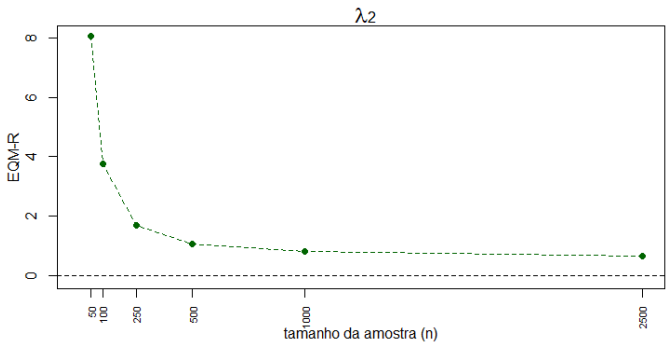

Figura 21 – EQM-R das médias das estimativas de $\boldsymbol{\lambda}_2$ para o modelo MoE-MF-MESN.

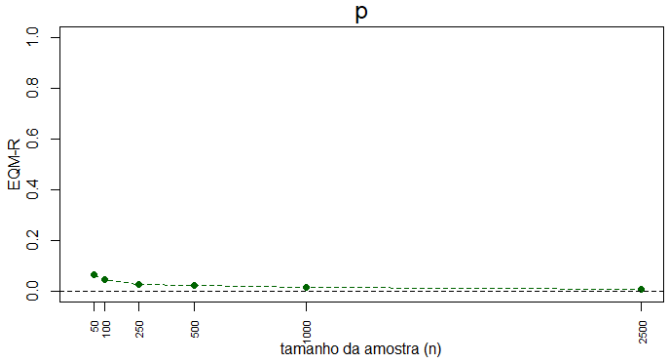

Figura 22 – EQM-R das médias das estimativas de $\boldsymbol{p}$ para o modelo MR-MF-MESN.

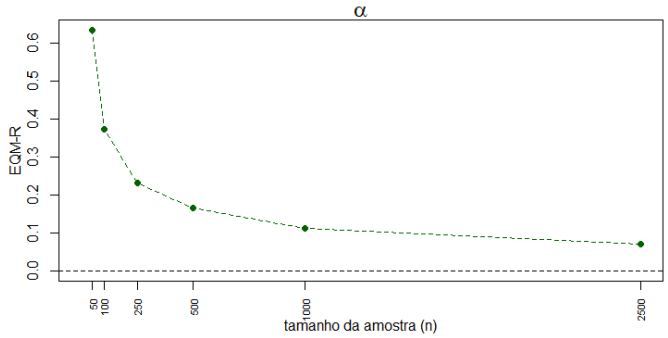

Figura 23 – EQM-R das médias das estimativas de $\boldsymbol{\alpha}$ para o modelo MoE-MF-MESN.

7.2 APLICAÇÃO (pesquisa *High School and Beyond*)

O *National Center for Education Statistics* (NCES) é a principal entidade federal para coleta e análise de dados relacionados à educação nos Estados Unidos da América. O conjunto de dados em estudo nesta seção consiste numa amostra aleatória de 200 observações e 11 variáveis obtida da pesquisa *High School and Beyond* (hsb2), realizada pelo NCES com estudantes do ensino médio. Numa adaptação do trabalho de Lim et al. (2016) (34), foram aplicados os modelos MR-MF-MESN e MoE-MF-MESN considerando *read* (x1) como variável explicativa e *science* (y1) e *write* (y2) como variável resposta bivariada, sendo que as três variáveis correspondem às pontuações obtidas pelos estudantes, respectivamente nas habilidades de leitura, ciências e redação. Os dados podem ser encontrados em http://www.ats.ucla.edu/stat/data/hsb2.csv ou no pacote R FMsmsnReg Benites (2016) (14).

Na Figura 24 observam-se as estimativas de densidade por Kernel de y1 e y2, indicando que os dados têm padrões de distribuições bimodais. Então, foram ajustadas as distribuições *Skew*-normal (SN) e *Skew*-t (ST) no contexto de misturas finitas de modelos de regressão para G = 2, considerando $\mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{1i}^\top \\ \mathbf{x}_{2i}^\top \end{bmatrix}$. Nesta seção, iremos considerar a seguinte notação MR-MF-SN e MR-MF-ST para os modelos de misturas finitas definidos

no Capítulo 5 ou MoE-MF-SN e MoE-MF-ST para os modelos de misturas finitas de especialistas, definidos no Capítulo 6.

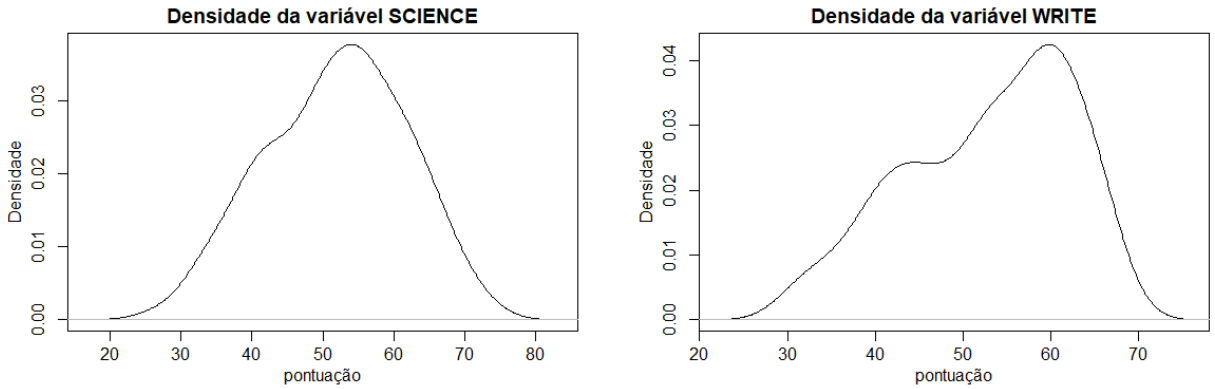

Figura 24 – Aplicação: densidades por Kernel de *science* (y1) e *write* (y2).

A Tabela 3 apresenta o quadro geral dos quatro modelos ajustados com as estimativas e erros padrão de todos os parâmetros, onde nota-se não haver discrepância acentuada nas estimativas dos coeficientes de regressão. Além disso, observe o alto valor estimado para os graus de liberdade da ST, já fornecendo indícios de que um modelo SN se ajuste melhor a este conjunto de dados.

A Tabela 4 resume os resultados dos ajustes dos modelos, em termos da log-verossimilhança e do critério BIC. Percebe-se que os modelos MoE-MF-SN e MoE-MF-ST ajustam-se melhor aos dados, com uma pequena diferença a favor do MoE-MF-SN. Assim, chegamos ao seguinte modelo para os dados hsb2:

$$f(\mathbf{y}_i|\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j(\mathbf{r}_i, \boldsymbol{\alpha})\boldsymbol{\psi}\text{MESN}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_j + b(\boldsymbol{\nu})\boldsymbol{\Delta}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu}), \tag{7.2.1}$$

em que $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j})^\top$, $\boldsymbol{\Sigma}_j = \begin{bmatrix} \sigma_{j.11} & \sigma_{j.12} \\ \sigma_{j.21} & \sigma_{j.22} \end{bmatrix}$, $\boldsymbol{\lambda}_j = (\lambda_{1j}, \lambda_{2j})^\top$, $\boldsymbol{\alpha}$, $j = 1, 2$.

Complementando a análise dos dados, na Tabela 5 são observadas as estimativas dos coeficientes de regressão ($\boldsymbol{\beta}$) no modelo MoE-MF-SN, seus correspondentes erros padrão e p-valor, sendo todos os parâmetros significativos, considerando o nível de significância de 5%.

Outro objetivo no ajuste do modelo foi classificar os dados y1 e y2 em G = 2 grupos. Considerando a estrutura de dados incompletos descrita na Seção 3.2, o objetivo foi inferir $\mathbf{z}_i$ em termos dos dados observados $\mathbf{y}_i$. A cada iteração no algoritmo EM, o vetor $\widehat{\boldsymbol{\theta}}$ de parâmetros desconhecidos do modelo é utilizado para se obter uma classificação das observações em termos probabilísticos, baseada em suas probabilidades a posteriori. Assim, para cada elemento da amostra, as duas probabilidades $\widehat{z}_{i1}$ e $\widehat{z}_{i2}$ dão as probabilidades a posteriori da i-ésima observação pertencer ao primeiro e segundo grupo, respectivamente.

Parâmetro	SN-MR		SN-MoE		ST-MR		ST-MoE	
	Estimativa	EP	Estimativa	EP	Estimativa	EP	Estimativa	EP
β_{10}	38.8958	11.927	43.3904	10.765	38.1606	12.661	42.8421	10.953
β_{11}	0.3636	0.152	0.2610	0.131	0.3739	0.158	0.2664	0.144
β_{20}	14.6316	9.041	24.7419	10.766	14.7409	12.194	22.6040	10.894
β_{21}	0.6601	0.116	0.4283	0.131	0.6633	0.148	0.5059	0.144
$\sigma_{1.11}$	63.8828	24.470	47.2570	20.312	50.1321	27.980	36.2612	19.942
$\sigma_{1.12}$	-17.6831	15.360	-17.4464	17.490	-5.5128	13.844	-4.3797	11.109
$\sigma_{1.22}$	32.0114	22.839	58.9199	28.795	19.4368	20.833	33.9833	21.325
$\sigma_{2.11}$	46.8970	16.355	46.5791	19.463	43.2257	17.608	53.0951	18.645
$\sigma_{2.12}$	-14.5352	17.358	-14.4349	17.649	-2.0919	11.019	4.9494	11.66
$\sigma_{2.22}$	81.1810	23.117	102.5317	29.021	45.3552	21.286	53.9373	21.057
λ_{11}	3.0619	2.894	1.7338	2.016	1.6625	2.170	0.8614	1.475
λ_{12}	-3.3535	2.563	-3.8449	2.841	-1.6931	1.942	-2.5036	1.612
λ_{21}	-0.5105	1.994	-0.4817	2.177	0.0918	1.975	0.4471	1.592
λ_{22}	1.2610	2.090	1.9072	2.814	0.2988	2.063	0.4876	1.805
p_1	0.3894	0.085	-	-	0.4039	0.114	-	-
p_2	0.6106	0.085	-	-	0.5961	0.114	-	-
α_1	-	-	-7.0001	2.091	-	-	5.9229	6.164
α_2	-	-	0.1425	0.048	-	-	-0.1156	0.123
ν	-	-	-	-	19.4224	38.264	18.5312	38.848

Tabela 3 – Aplicação: Estimativas e erros padrão (EP) das distribuições *Skew*-normal (SN) e *Skew*-t (ST) para os modelos MR-MF-MESN (MR) e MoE-MF-MESN (MoE). Os erros padrão foram calculados via *bootstrap* paramétrico.

Distribuição	Log-verossimilhança	BIC
SN-MR-MF-MESN	-1352,615	2784,705
ST-MR-MF-MESN	-1455,980	3017,925
SN-MoE-MF-MESN	**-1345,689**	**2776,152**
ST-MoE-MF-MESN	-1346,232	2777,237

Tabela 4 – Aplicação: Seleção do modelo: log-verossimilhança e BIC

Parâmetro	SN-MoE			
	Estimativa	EP	p-valor	
β_{10}	43,3904	10.765	6e-05	***
β_{11}	0,2610	0.131	0.04637	·
β_{20}	24,7419	10.766	0.02155	·
β_{21}	0,4283	0.131	0.00108	*

Tabela 5 – Aplicação: SN-MoE-MF-MESN: estimativas, erros padrão (EP), p-valor e significância dos coeficientes de regressão $\boldsymbol{\beta}$. Códigos das significâncias, dado o p-valor: $0 = ***$; $0,001 = **$; $0,01 = *$; $0,05 = \cdot$; $0,1 =$ ' '; $1 = 1$.

Segundo a regra de Bayes determinada, classifica-se a observação para aquele grupo correspondente ao maior valor observado das probabilidades a posteriori. Considerando que as densidades das variáveis respostas mostradas na Figura 24 caracterizam de modo sutil (em curvas suaves) a mistura em dois grupos distintos, por meio da análise exploratória dos gráficos da Figura 25 pode-se concluir que o modelo obteve êxito satisfatório em estimar a classificação das observações entre os dois grupos distintos.

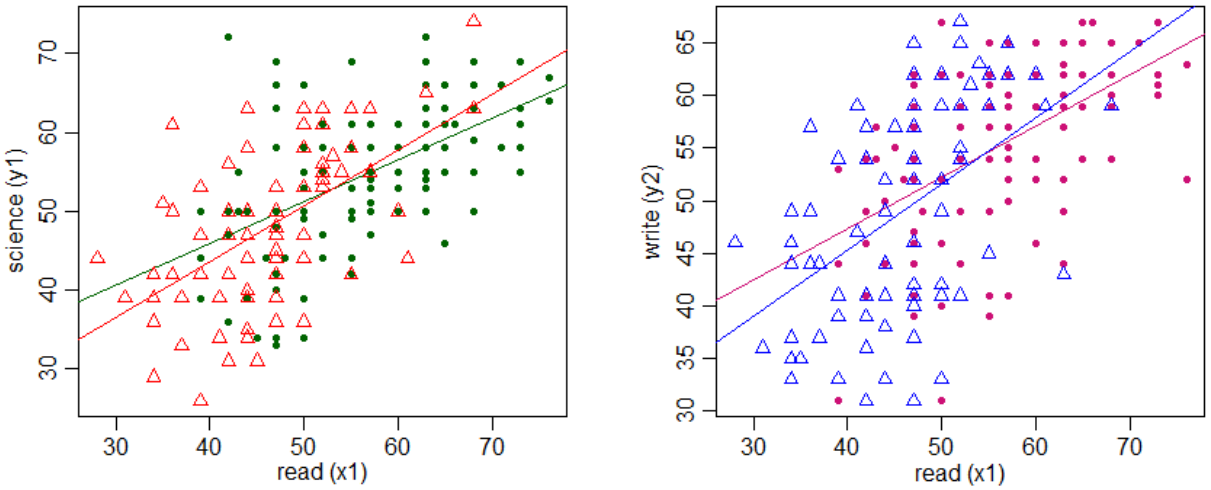

Figura 25 – Aplicação: Gráficos de dispersão dos dados hsb2 (y1 e y2), as linhas de regressão de mistura ajustadas e a classificação pelo ajuste da distribuição *Skew*-normal para MoE-MF-MESN de 2 componentes (grupo 1 = triângulo, grupo 2 = ponto).

Interpretando os resultados, observamos pelas retas nos gráficos da Figura 25 que as variáveis respostas y1 e y2 são correlacionadas positivamente com a variável explicativa x1, corroborando com os coeficientes de regressão $\widehat{\boldsymbol{\beta}}_{11} = 0,2610$ e $\widehat{\boldsymbol{\beta}}_{22} = 0,4283$ da Tabela 5. Outra evidência é que nesses dois gráficos a maioria das observações do grupo 1 se concentra na região de menor pontuação (o quadrante inferior esquerdo), sendo que as médias amostrais das variáveis nos grupos 1 e 2 correspondem respectivamente a: (1) 46,057 e 57,080 para *read* (x1); (2) 44,295 e 57,786 para *science* (y1); e (3) 44,273 e 59,455 para *write* (y2), ou seja, as pontuações dos alunos do grupo 1 são, em media, mais baixas. Desta forma, de acordo com esses resultados, há fortes evidências para se afirmar, a um nível de 5% de confiança, que a capacidade de leitura dos alunos influi significativamente nas pontuações de ciências e redação.

8 CONCLUSÃO

Neste trabalho, foram apresentadas duas propostas úteis na modelagem de dados com a presença de multimodalidade, assimetria, curtose e heterogeneidade não observada. Dois conceitos essenciais para o trabalho são o de mistura finita de densidades e a classe de distribuições MESN. Algoritmos do tipo EM foram obtidos, via uma representação estocástica adequada dos modelos propostos. Também foram consideradas as questões relativas à classificação de observações, no sentido que cada indivíduo na população pertence a um de G grupos, que existem devido a uma característica não observada dos dados. O principal objetivo foi estudar os modelos de misturas finitas de modelos de regressão multivariados assimétricos, em particular, no contexto da classe de distribuições misturas de escala *Skew*-normal (MESN). As propostas estudadas foram o método clássico Misturas Finitas de Modelos de Regressão sob as distribuições MESN (MR-MF-MESN) e o método de Misturas Finitas de Especialistas de Modelos de Regressão sob as distribuições MESN (MoE-MF-MESN). Finalmente, exemplos numéricos considerando dados simulados e reais foram apresentados para ilustrar os modelos e os resultados inferenciais desenvolvidos.

Espera-se que este trabalho seja útil para despertar o interesse de estudantes, pesquisadores e profissionais pelo tema, que acreditamos ser de grande aplicabilidade.

Devido a parte final do trabalho (Capítulo 6) se tratar de uma inovação, como trabalho futuro e para fins de publicação, poderão ser realizadas outras simulações com intuito de refinar os resultados encontrados, analisar conjuntos de dados reais onde a característica de interesse relacionada aos grupos seja observada e empregar as derivadas de segunda ordem com a finalidade de utilizar a Matriz de Informação Observada, ao invés da técnica computacional *bootstrap* paramétrico, para obter as estimativas dos erros padrão dos estimadores dos parâmetros.

Referências Bibliográficas

1 AITKEN, A. (1926). On bernoulli's numerical solution of algebraic equations. Proc R Soc Edinburgh, 46, 289–305. Seç.5.2

2 ANDREWS, D. F., & MALLOWS, C. L. (1974). Scale Mixtures of Normal Distributions. Journal of the Royal Statistical Society: Series B (Methodological), 36(1), 99–102. doi:10.1111/j.2517-6161.1974.tb00989.x Seç.2.1

3 ARELLANO-VALLE, R. B.; BOLFARINE, H.; LACHOS, V. H. Skew-normal linear mixed models. Journal of Data Science, v. 3, p. 415–438, 2005. Seç.1.2 - Seç.2.2

4 AZZALINI, A. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, v. 12, p. 171–178, 1985. Seç.1.1 - Seç.1.2 - Seç.2.2 - Cap.4

5 AZZALINI, A., & DALLA-VALLE, A. (1996). The multivariate skew-normal distribution. Biometrika, 83, 715–726. Seç.2.1 - Seç.2.1

6 AZZALINI, A., & CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society, Series B, 65, 367–389. Cap.4

7 BASFORD, K. E., GREENWAY, D. R., MCLACHLAN, G. J. & PEEL, D. (1997). Standard errors of fitted component means of normal mixtures. Computational Statistics, 12, 1–17. ??

8 BASSO, R. M.. Misturas finitas de misturas de escala skew-normal. 2009. 83f. Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matematica, Estatistica e Computação Cientifica, Campinas, SP. Seç.1.1 - Seç.1.2 - Seç.1.2 - Seç.1.2 - Seç.2.1 - Apêndice A

9 BASSO, R. M., LACHOS, V. H., CABRAL, C. R. B., & GHOSH, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions. Computational Statistics and Data Analysis, 54, 2926–2941. Cap.4 - Seç.5.2

10 BOHNING, D., DIETZ, E., SCHAUB, R., SCHLATTMANN, P. & LINDSAY, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Annals of the Institute of Statistical Mathematics, 46, 373–388. Seç.5.2

11 BOHNING D. 1999. Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others. New York: Chapman & Hall/CRC. Seç.3.1

12 BÖHNING, D. (2000). Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others. Boca Raton: ChapmanHall/CRC. Cap.1

13 BRANCO, M. D.; DEY, D. K. A general class of multivariate skew-elliptical distributions. Journal of Multivariate Analysis, v. 79, p. 99–113, 2001. Seç.1.1 - Seç.1.2 - Cap.2 - Seç.2.1

14 BENITES, L., MAEHARA, R. P., & LACHOS, V. H. (2016). FMsmsnReg: Regression Models with Finite Mixtures of Skew Heavy-Tailed Errors. Pacote no R Core Team. https://cran.r-project.org/web/packages/FMsmsnReg/index.html Seç.7.2

15 BENITES, L. E. Mistura de modelos de regressão. 2018. 121 f. Tese (Doutorado) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018. Seç.1.2 - Cap.5 - Seç.5.1 - Seç.5.2 - Seç.5.2 - **??**

16 CABRAL, C. R. B., LACHOS, V. H., & PRATES, M. O. (2012). Multivariate mixture modeling using skewnormal independent distributions. Computational Statistics and Data Analysis, 56, 126–142. Cap.4

17 CHAMROUKHI, F. (2017). Skew t mixture of experts. Neurocomputing, 266, 390–408. doi:10.1016/j.neucom.2017.05.044. Cap.6

18 DAY, N. Estimating the components of a mixture of a two normal distribution. Biometrika, v. 56, p. 463–474, 1969. Seç.1.1

19 DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, v. 39, p. 1–38, 1977. Seç.1.1 - Seç.2.1 - Seç.5.2

20 DESARBO, W. S., & CRON, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. Journal of Classification, 5(2), 249–282. doi:10.1007/bf01897167. Seç.1.1 - Cap.6

21 DUDA R.O. & HART, P. Pattern Classification and Scene Analysis. New York: Wiley, 1988. Seç.3.1

22 EVERITT B, HAND D. 1981. Finite Mixture Distributions. New York: Chapman & Hall Seç.3.1

23 FRÜHWIRTH-SCHNATTER S. 2006. Finite Mixture and Markov Switching Models. New York: Springer Mengersen K, Robert C, Titterington D, eds. 2011. Mixtures: Estimation and Applications. New York: Wiley. Cap.1 - Seç.3.1

24 GRÜN, B. & LEISCH, F. (2008). Finite Mixtures of Generalized Linear Regression Models, pages 205–230. PhysicaVerlag HD, Heidelberg. Seç.5.2

25 HENNIG, C. (2000). Identifiablity of models for clusterwise linear regression. Journal of Classification, 17(2), 273–296. Seç.5.2

26 JACOBS, R.A., JORDAN, M.I., NOWLAN, S.J., Hinton, G.E., et al., 1991. Adaptive mixtures of local experts. Neural Comput. 3 (1), 79–87. Cap.1 - Cap.6 - Cap.6

27 JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1994). Continuous Univariate Distributions, Vol. 1. Wiley, New York. Apêndice A

28 JONES, P.N. and MCLACHLAN, G.J. (1992). Fitting finite mixture models in a regression context. Australian Journal of Statistics 34, 233-240. Seç.1.1 - Cap.6

29 KASAHARA, H., & SHIMOTSU, K. (2013). Non-parametric identification and estimation of the number of components in multivariate mixtures. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), 97–111. doi:10.1111/rssb.12022 Cap.3

30 LACHOS, V. H.; GHOSH, P.; ARELLANO-VALLE, R. B. Likelihood based inference for skew normal independent linear mixed models. Statistica Sinica, 2009. Seç.2.1

31 LACHOS, V. H., GHOSH, P. & ARELLANO-VALLE, R. B. (2010). Likelihood based inference for skew normal independentlinear mixed models. S tatistica Sinica, 20, 303–322. Seç.1.2 - Cap.4

32 LACHOS, V. H.; CABRAL, C. R. B.; ZELLER, C. B. (2018). Finite Mixture of Skewed Distributions. Springer. Seç.1.2 - Seç.5.1 - Seç.5.2

33 LEISCH, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. Journal of Statistical Software, 11(8), 1–18. https://doi.org/10.18637/jss.v011.i08. Seç.1.1

34 LIM, H. K., NARISETTY, N. N., & CHEON, S. (2016). Robust multivariate mixture regression models with incomplete data. Journal of Statistical Computation and Simulation, 87(2), 328–347. doi:10.1080/00949655.2016.1209198. Seç.7.2

35 LIN, T. I.; LEE, J. C.; NI, H. F. Bayesian analysis of mixture modelling using the multivariate t distribution. Statistics and Computing, v. 14, p. 119–130, 2004. Seç.1.1

36 LIN, T. I., LEE, J. C., & YEN, S. Y. (2007a). Finite mixture modelling using the skew normal distribution. Statistica Sinica, 17, 909–927. Cap.4

37 LIN, T. I., LEE, J. C., & HSIEH, W. J. (2007b). Robust mixture modelling using the skew t distribution. Statistics and Computing, 17, 81–92. Cap.4

38 LIN, T. I. Maximum likelihood estimation for multivariate skew normal mixture models. Journal of Multivariate Analysis, v. 100, p. 257–265, 2009. Seç.1.1 - Cap.4

39 LIN, T. I. (2010). Robust mixture modeling using multivariate skew t distributions. Statistics and Computing, 20, 343–356. Cap.4

40 LINDSAY BG. 1995. Mixture Models: Theory, Geometry and Applications. Hayward, CA: Inst. Math. Stat. Cap.1 - Seç.3.1

41 LIU, C.; RUBIN, D. B. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika, v. 81, p. 633–648, 1994. Seç.2.3

42 LIU, C (1998). "Parameter expansion to accelerate EM: The PX-EM algorithm". Biometrika. 85 (4): 755–770. Seç.1.1

43 MCLACHLAN G.J. & BASFORD, K. Mixture Models: Inference and applications to clustering. New York: Marcel Dekker, 1988. Seç.3.1 - Seç.3.1

44 MCLACHLAN, G., & MCGIFFIN, D. (1994). On the role of finite mixture models in survival analysis. Statistical Methods in Medical Research, 3(3), 211–226. doi:10.1177/096228029400300302 Cap.3

45 MCLACHLAN, G. J.; PEEL, G. J. Finite Mixture Models. [S.l.]: John Wiley and Sons, 2000. Cap.1 - Seç.1.1 - Seç.1.1 - Seç.3.1 - Seç.5.2

46 MCLACHLAN, Geoffrey J.; PEEL, David, Finite Mixture Models Wiley Series in Probability and Statistics John Wiley & Sons, 2004. Seç.1.1

47 MCLACHLAN, G. J., & CHANG, S. U. (2004). Mixture modelling for cluster analysis. Statistical Methods in Medical Research, 13(5), 347–361. doi:10.1191/0962280204sm372ra Cap.3

48 MCNICHOLAS, P. & MURPHY, T. (2008). Parsimonious Gaussian mixture models. Statistics and Computing, 18, 285–296. Seç.5.2

49 MCNICHOLAS PD. 2017. Mixture Model-Based Classification. Boca Raton, FL: CRC. Seç.3.1

50 MENG, X. L.; RUBIN, D. B. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika, v. 80, p. 267–278, 1993. Seç.2.3

51 MENGERSEN, Kerrie. and ROBERT, Christian P. and TITTERINGTON, D. M. Mixtures: estimation and applications / edited by Kerrie L. Mengersen, Christian P. Robert, D. Michael TitteringtonWiley; John Wiley [distributor] Hoboken, N.J. : Chichester 2011. Cap.1 - Seç.3.1 - Cap.6 - Cap.6

52 MIRFARAH, Elham; NADERI, Mehrdad; CHEN, Ding-Geng; Mixture of linear experts model for censored data: A novel approach with scale-mixture of normal distributions, Computational Statistics & Data Analysis, Volume 158, 2021, 107182, ISSN 0167-9473, https://doi.org/10.1016/j.csda.2021.107182. (https://www.sciencedirect.com/science/article/pii/S0167947321000165) Seç.1.2 - Cap.6

53 NGUYEN, H. D., & MCLACHLAN, G. J. (2016). Laplace mixture of linear experts. Computational Statistics & Data Analysis, 93, 177–191. doi:10.1016/j.csda.2014.10.016. Cap.6

54 NIERENBERG, D., T. STUKEL, J. BARON, B. DAIN, and E. GREENBERG (1989). Determinants of plasma levels of beta-carotene and retinol. Skin Cancer Prevention Study Group. Am J Epidemiol 130(3), 511–21. Cap.6

55 OTINIANO, C., RATHIE, P. & OZELIM, L. (2015). On the identifiability of finite mixture of skew-normal and skew-t distributions. Statistics & Probability Letters, 106, 103–108. Seç.5.2

56 PRATES, M. O., CABRAL, C. R. B., & LACHOS, V. H. (2013). mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions. Journal of Statistical Software, 54(12), 1–20. https://doi.org/10.18637/jss.v054.i12 Cap.4 - Seç.4.1

57 PYNE, S., HU, X., WANG, K., ROSSIN, E., LIN, T., BAECHER-ALLAN, L. M. M. C., MCLACHLAN, G. J. P., Tamayo, D. A. H., De Jager, P. L., & Mesirov, J. P. (2009). Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences USA, 106, 8519–8524. Cap.4

58 QUANDT, R. E. (1972). A new approach to estimating switching regressions. Journal of the American Statistical Association, 67, 306–310. Seç.5.2

59 RAUSCH, J. R., & KELLEY, K. (2009). A comparison of linear and mixture models for discriminant analysis under nonnormality. Behavior Research Methods, 41(1), 85–98. doi:10.3758/brm.41.1.85 Cap.3

60 R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Seç.4.1

61 SCHLATTMANN, P., 2009: Medical Applications of Finite Mixture Models. Statistics for Biology and Health, Springer, 246 pp. ,https://doi.org/10.1007/978-3-540-68651-4. Seç.1.1 - Cap.6

62 SCHWARZ, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461–464. Seç.5.2

63 SFIKAS, G., NIKOU, C., & GALATSANOS, N. (2007). Robust image segmentation with mixtures of Student's t-distributions. IEEE International Conference on Image Processing, 1. ICIP 2007. Cap.3

64 SHOHAM, S. Robust clustering by deterministic agglomeration em of mixtures of multivariate t-distributions. Pattern Recognition, v. 35, p. 1127–1142, 2002. Seç.1.1

65 SHOHAM, S.; FELLOWS, M. R.; NORMANN, R. A. Robust, automatic spike sorting using mixtures of multivariate t-distributions. Journal of Neuroscience Methods, v. 127, p. 111–122, 2003. Seç.1.1

66 STUKEL, T. (2008). Determinants of Plasma Retinol and Beta-Carotene Levels. http://lib.stat.cmu.edu/datasets/Plasma Retinol. Cap.6

67 TITTERINGTON DM, SMITH A, MAKOV U. 1985. Statistical Analysis of Finite Mixture Distributions. New York: Wiley. Seç.3.1

68 WANG, H. X. et al. Robust mixture modelling using multivariate t-distribution with missing information. Pattern Recognition Letters, v. 25, p. 701–710, 2004. Seç.1.1

69 WOLFE, J. A computer program for the computation of maximum likelihood analysis of type. Research memo. SRM 65-12 San Diego: U.S. Naval Personnel Research Activity, 1965. Seç.1.1

70 WOLFE, J. NORMIX: Computational methods for estimating the parameters of multivariate normal mixtures of distributions. Research memo. SRM 68-2 San Diego: U.S. Naval Personnel Research Activity, 1967. Seç.1.1

71 WOLFE, J. Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, v. 5, p. 329–350, 1970. Seç.1.1

72 YUKSEL, S. E., WILSON, J. N., & GADER, P. D. (2012). Twenty Years of Mixture of Experts. IEEE Transactions on Neural Networks and Learning Systems, 23(8), 1177–1193. doi:10.1109/tnnls.2012.2200299. Cap.1

73 ZELLER, C. B., LACHOS, V. H., & VILCA-LABRA, F. E. (2011). Local influence analysis for regression models with scale mixtures of skew-normal distributions. Journal of Applied Statistics, 38, 348–363. Seç.5.2

74 ZELLER, C. B., CABRAL, C. R. B. & LACHOS, V. H. (2016). Robust mixture regression modeling based on scale mixtures of skew-normal distributions. TEST, 25(2), 375–396. Seç.1.1

75 ZELLER, C. B., CABRAL, C. R. B., LACHOS, V. H., & BENITES, L. (2018). Finite mixture of regression models for censored data based on scale mixtures of normal distributions. Advances in Data Analysis and Classification. doi:10.1007/s11634-018-0337-y. Seç.5.2

APÊNDICE A – Lemas

Os Lemas descritos a seguir foram obtidos no trabalho de Basso (2009) (8).

Lema A.1 Seja $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Então, para algum vetor fixo $\mathbf{a}$ de dimensão k e alguma matriz fixa $\mathbf{B}_{k\times n}$, temos

$$E[\Phi_k(\mathbf{a}+\mathbf{B}\mathbf{Y}|\boldsymbol{\eta},\boldsymbol{\Omega})] = \Phi_k(\mathbf{a}|\boldsymbol{\eta}-\mathbf{B}\boldsymbol{\mu}, \boldsymbol{\Omega}+\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top),$$
$$E[\phi_k(\mathbf{a}+\mathbf{B}\mathbf{Y}|\boldsymbol{\eta},\boldsymbol{\Omega})] = \phi_k(\mathbf{a}|\boldsymbol{\eta}-\mathbf{B}\boldsymbol{\mu}, \boldsymbol{\Omega}+\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top).$$

Lema A.2 Seja $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e $\mathbf{X} \sim N_q(\boldsymbol{\eta}, \boldsymbol{\Omega})$. Então

$$\phi_p(\mathbf{y}|\boldsymbol{\mu}+\mathbf{A}\mathbf{x}, \boldsymbol{\Sigma})\phi_p(\mathbf{x}|\boldsymbol{\eta}, \boldsymbol{\Omega}) = \phi_p(\boldsymbol{\mu}+\mathbf{A}\boldsymbol{\eta}, \boldsymbol{\Sigma}+\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top) \times \phi_p(\mathbf{x}|\boldsymbol{\eta}+\boldsymbol{\Lambda}\mathbf{A}^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}-\mathbf{A}\boldsymbol{\eta}), \boldsymbol{\Lambda}),$$

em que $\lambda = (\boldsymbol{\Omega}^{-1}+\mathbf{A}^\top\boldsymbol{\Sigma}^{-1}\mathbf{A})$.

Demonstração. Fazendo $\mathbf{z} = \mathbf{y}-\boldsymbol{\mu}-\mathbf{A}\boldsymbol{\eta}$ e $\mathbf{W} = \mathbf{x}-\boldsymbol{\eta}$, a prova segue do fato que

$$(\mathbf{z}-\mathbf{A}\mathbf{W})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\mathbf{A}\mathbf{W}) + \mathbf{W}^\top\boldsymbol{\Omega}^{-1}\mathbf{W} = \mathbf{z}(\boldsymbol{\Sigma}+\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top)^{-1}\mathbf{z} + (\mathbf{W}-\Lambda\mathbf{A}^\top\boldsymbol{\Sigma}^{-1}\mathbf{z})^\top\Lambda^{-1}(\mathbf{W}-\Lambda\mathbf{A}^\top\boldsymbol{\Sigma}^{-1}\mathbf{z}),$$

a prova segue também por notar que $|\boldsymbol{\Sigma}+\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top||\Lambda| = |\boldsymbol{\Sigma}||\boldsymbol{\Omega}|$. □

O seguinte lema é uma propriedade da distribuição da normal truncada (veja Johnson et al. 1994, Seção 10.1) (27).

Lema A.3 Seja $X \sim N(\eta, \tau^2)$. Então, para para qualquer constante real a segue que

$$E[X|X>a] = \eta + \frac{\phi_1\left(\frac{a-\eta}{\tau}\right)}{1-\Phi_1\left(\frac{a-\eta}{\tau}\right)}\tau,$$
$$E[X^2|X>a] = \eta^2+\tau^2 + \frac{\phi_1\left(\frac{a-\eta}{\tau}\right)}{1-\Phi_1\left(\frac{a-\eta}{\tau}\right)}(\eta+a)\tau.$$

APÊNDICE B – Demonstração dos estimadores

Demonstração do estimador de p_j:

Para a demonstração do estimador p_j, deve ser considerada a expressão vista na Subseção **5.2.1** dada por:

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = C &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij}^{(k)} \log p_j \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij}^{(k)} \log \boldsymbol{\Gamma}_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{1ij}^{(k)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^\top \boldsymbol{\Gamma}_j^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{2ij}^{(k)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^\top \boldsymbol{\Gamma}_j^{-1} \boldsymbol{\Delta}_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{3ij}^{(k)} \boldsymbol{\Delta}_j^\top \boldsymbol{\Gamma}_j^{-1} \boldsymbol{\Delta}_j. \end{aligned}$$

Temos interesse em maximizar a função

$$\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij} \log p_j,$$

com a restrição $\sum_{j=1}^{G} p_j = 1$, utilizando multiplicadores de Lagrange.

A função lagrangeana é dada por:

$\mathrm{L}\,(p_1, \ldots, p_G, \lambda) = \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij} \log p_j - \lambda\left(\sum_{j=1}^{G} p_j - 1\right)$,

$$\frac{\partial L}{\lambda} = -\left(\sum_{j=1}^{G} p_j - 1\right) = 0 \Rightarrow \sum_{j=1}^{G} p_j = 1, \text{a restrição dada.}$$

Sabendo que $\sum_{j=1}^{G}\sum_{i=1}^{n} \widehat{z}_{ij} = n$ e $\sum_{j=1}^{G} p_j = 1$, temos

$$\begin{aligned} \frac{\partial L}{\partial p_j} = \sum_{i=1}^{n} \widehat{z}_{ij}\frac{1}{p_j} - \lambda = \frac{\sum_{i=1}^{n} \widehat{z}_{ij} - \lambda p_j}{p_j} = 0 \\ \sum_{i=1}^{n} \widehat{z}_{ij} = \lambda p_j \\ \sum_{j=1}^{G}\sum_{i=1}^{n} \widehat{z}_{ij} = \lambda \sum_{j=1}^{G} p_j, \Rightarrow \lambda = n \\ \sum_{i=1}^{n} \widehat{z}_{ij} = n\widehat{p}_j \\ \widehat{p}_j = \frac{\sum_{i=1}^{n} \widehat{z}_{ij}}{n}, \text{ como se queria demonstrar.} \end{aligned}$$

Demonstração do estimador de $\boldsymbol{\alpha}$:

Para a demonstração do estimador $\boldsymbol{\alpha}$, deve ser considerada a expressão vista na Subseção **6.1.1** dada por:

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = C &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij}^{(k)} \log p_j\left(\mathbf{r};\boldsymbol{\alpha}\right) \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z}_{ij}^{(k)} \log \boldsymbol{\Gamma}_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{1ij}^{(k)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^\top \boldsymbol{\Gamma}_j^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{2ij}^{(k)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^\top \boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{s}_{3ij}^{(k)} \boldsymbol{\Delta}_j^\top \boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Delta}_j. \end{aligned}$$

Sendo

$$p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right) = \frac{\exp(\mathbf{r}_i^\top \boldsymbol{\alpha}_j)}{1 + \sum_{l=1}^{g-1}\exp(\mathbf{r}_i^\top \boldsymbol{\alpha}_l)}$$

e

$$\frac{d(p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right))}{d(\boldsymbol{\alpha})} = p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\left[1 - p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\right]\mathbf{r}_i,$$

temos que

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}} &= \frac{\partial}{\partial \boldsymbol{\alpha}}\left\{\sum_{i=1}^{n}\sum_{j=1}^{g} \widehat{z}_{ij} \log\left[p_j\left(\mathbf{r};\boldsymbol{\alpha}\right)\right]\right\} \\ &= \sum_{i=1}^{n} \widehat{z}_{ij} \frac{1}{p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)} p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\left[1 - p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\right]\mathbf{r}_i \\ &= \sum_{i=1}^{n} \widehat{z}_{ij}\left[1 - p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\right]\mathbf{r}_i. \end{aligned}$$

Assim, considerando que $p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\left[1 - p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\right] \leq \frac{1}{4}$, calcula-se a segunda derivada

$$\begin{aligned} \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j} &= \partial\left(\sum_{i=1}^{n} \widehat{z}_{ij}\left[1 - p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\right]\mathbf{r}_i\right)/\partial \boldsymbol{\alpha}_j \\ &= -\sum_{i=1}^{n} p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\left[1 - p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\right]\mathbf{r}_i\mathbf{r}_i^\top \\ &= -\frac{1}{4}\sum_{i=1}^{n} \mathbf{r}_i\mathbf{r}_i^\top. \end{aligned}$$

Fazendo aproximação de primeira ordem numa série de Taylor com os resultados acima encontrados, obtém-se o resultado

$$\begin{aligned} \boldsymbol{\alpha}_j^{(k+1)} &= \boldsymbol{\alpha}_j^{(k)} - \left[\frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}}\right]^{-1}\left[\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}}\right] \\ &= 4\left(\sum_{i=1}^{n} \mathbf{r}_i\mathbf{r}_i^\top\right)^{-1}\sum_{i=1}^{n} \widehat{z}_{ij}\left[1 - p_j\left(\mathbf{r}_i;\boldsymbol{\alpha}\right)\right]\mathbf{r}_i + \boldsymbol{\alpha}_j^{(k)}, \text{ como se queria demonstrar.} \end{aligned}$$