

UNIVERSIDADE ESTADUAL DE CAMPINAS  
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA  
DEPARTAMENTO DE ESTATÍSTICA

# Misturas Finitas de Misturas de Escala Skew-Normal

Rodrigo Marreiro Basso

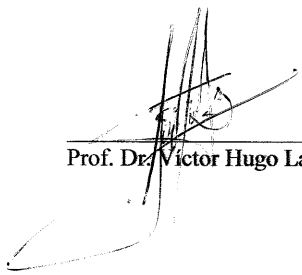
Dissertação de Mestrado orientada pelo

Prof. Dr. Victor Hugo Lachos Dávila

## MISTURAS FINITAS DE MISTURAS DE ESCALA SKEW-NORMAL

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por **Rodrigo Marreiro Basso** e aprovada pela comissão julgadora.

Campinas, 12 de março de 2009



Prof. Dr. Victor Hugo Lachos Dávila

Banca Examinadora:

1. Prof. Dr. Victor Hugo Lachos Dávila IMECC - UNICAMP
2. Prof. Dr. Celso Rômulo Barbosa Cabral UFAM
3. Prof. Dr. Aluísio de Souza Pinheiro IMECC - UNICAMP
4. Prof. Dr. Filidor Edilson Vilca Labra IMECC - UNICAMP
5. Prof. Dr. Rolando de la Cruz Messia PUC-CHILE

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, **UNICAMP**, como requisito parcial para obtenção do Título de **MESTRE em ESTATÍSTICA**

**FICHA CATALOGRÁFICA ELABORADA PELA  
BIBLIOTECA DO IMECC DA UNICAMP**

Bibliotecária: Maria Fabiana Bezerra Müller – CRB8 / 6162

Basso, Rodrigo Marreiro  
B295m Misturas finitas de misturas de escala skew-normal/ Rodrigo  
Marreiro Basso -- Campinas, [S.P. : s.n.], 2009.

Orientador : Victor Hugo Lachos Dávila  
Dissertação (mestrado) - Universidade Estadual de Campinas,  
Instituto de Matemática, Estatística e Computação Científica.

1.Algoritmos de expectativa de maximização. 2.Distribuição  
(Probabilidades). 3.Misturas finitas. 4.Distribuição normal assimétrica.  
5.Misturas de escala. . I. Lachos Dávila, Victor Hugo. II. Universidade  
Estadual de Campinas. Instituto de Matemática, Estatística e  
Computação Científica. III. Título.

(mfbm/imecc)

Título em inglês: Mixtures modelling using scale mixtures of skew-normal distributions

Palavras-chave em inglês (Keywords): 1. EM algorithms. 2. Distribution (Probability theory).  
3. Finite mixtures. 4. Skew normal distribution. 5. Scale mixtures.

Área de concentração: Estatística Computacional

Titulação: Mestre em Estatística

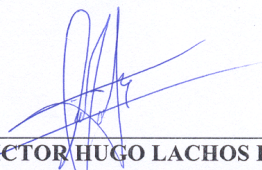
Banca examinadora: Prof. Dr. Victor Hugo Lachos Dávila (IMECC – Unicamp)  
Prof. Dr. Celso Rômulo Barbosa Cabral (UFAM)  
Prof. Dr. Aluísio de Sousa Pinheiro (IMECC – Unicamp)  
Prof. Dr. Filidor Edilson Vilca Labra (IMECC - Unicamp)

Data da defesa: 12/03/2009

Programa de Pós-Graduação: Mestrado em Estatística

Dissertação de Mestrado defendida em 12 de março de 2009 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.



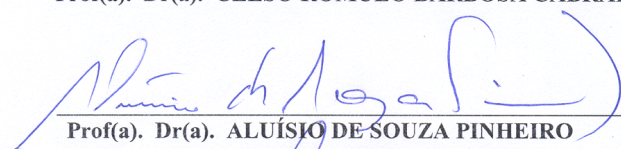
---

Prof(a). Dr(a). VICTOR HUGO LACHOS DÁVILA



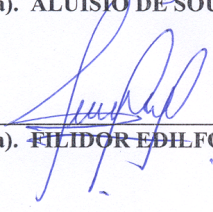
---

Prof(a). Dr(a). CELSO RÔMULO BARBOSA CABRAL



---

Prof(a). Dr(a). ALUÍSIO DE SOUZA PINHEIRO



---

Prof(a). Dr(a). FILIDOR EDLFONSO VILCA LABRA

*A todos aqueles que  
contribuíam para que aqui  
estivesse.*

## *Agradecimentos*

- Aos meus familiares, por serem complacentes quanto a minha ausência nesses anos voltados à vida acadêmica.
- Ao meu orientador Prof. Dr. Victor Hugo Lachos Dávila, por sua dedicação, incentivo e amizade nesses anos que se passaram. Em especial, por muito ter contribuído em ensinar-me os caminhos da pesquisa científica, incitado por muita motivação e perseverança.
- Ao Prof. Dr. Celso Romulo Barbosa Cabral, por sua enorme contribuição e prestabilidade para que esse trabalho pudesse ser realizado.
- Aos amigos Marley Saraiva e Eduardo Botelho, pelo companherismo e em especial por ajudarem na formatação dessa dissertação.
- À toda banca examinadora, agradeço por ter aceitado o convite, pela dedicação para com esse trabalho e por suas sugestões e correções.
- À CNPq, pelo apoio financeiro

## *Resumo*

Nesse trabalho será considerada uma classe flexível de modelos usando misturas finitas de distribuições da classe de misturas de escala *skew-normal*. O algoritmo EM é empregado para se obter estimativas de máxima verossimilhança de maneira iterativa, sendo discutido com maior ênfase para misturas de distribuições *skew-normal*, *skew-t*, *skew-slash* e *skew-normal* contaminada. Também será apresentado um método geral para aproximar a matrix de covariância assintótica das estimativas de máxima verossimilhança. Resultados obtidos da análise de quatro conjuntos de dados reais ilustram a aplicabilidade da metodologia proposta.

## *Abstract*

In this work we consider a flexible class of models using finite mixtures of multivariate scale mixtures of skew-normal distributions. An EM-type algorithm is employed for iteratively computing maximum likelihood estimates and this is discussed with emphasis on finite mixtures of skew-normal, skew-t, skew-slash and skew-contaminated normal distributions. A general information-based method for approximating the asymptotic covariance matrix of the maximum likelihood estimates is also presented. Results obtained from the analysis of four real data sets are reported illustrating the usefulness of the proposed methodology.



# *Sumário*

<b>Lista de Figuras</b>	<b>xi</b>
<b>Lista de Tabelas</b>	<b>xii</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Proposta do Trabalho . . . . .	1
1.2 Organização do Trabalho . . . . .	3
<b>2 Distribuição <i>Skew</i>-Normal</b>	<b>5</b>
2.1 A Distribuição <i>Skew</i> -Normal Padrão Univariada . . . . .	6
2.1.1 Função densidade de probabilidade e função de distribuição acumulada . . . . .	6
2.1.2 Propriedades . . . . .	8
2.1.3 Caracterizações . . . . .	10
2.2 A Representação Univariada com Três Parâmetros . . . . .	12
2.3 A Distribuição <i>Skew</i> -Normal Multivariada . . . . .	13
2.3.1 Função densidade de probabilidade e função de distribuição acumulada . . . . .	14
2.3.2 Propriedades . . . . .	16
2.3.3 Caracterizações . . . . .	18

<b>3</b>	<b>A classe de distribuições de Mistura de Escala <i>Skew</i>-Normal</b>	<b>20</b>
3.1	Distribuições MESN Multivariada . . . . .	21
3.1.1	Definição . . . . .	21
3.1.2	Representação estocática . . . . .	22
3.1.3	Propriedades . . . . .	22
3.1.4	Distribuição marginal e independência . . . . .	24
3.2	Exemplos . . . . .	25
3.2.1	Distribuição <i>skew</i> -t multivariada . . . . .	25
3.2.2	Distribuição <i>skew</i> -slash multivariada . . . . .	26
3.2.3	Distribuição <i>skew</i> -normal contaminada multivariada . . . . .	27
3.3	Inferência pelo Método da Máxima Verossimilhança . . . . .	28
3.3.1	Representação hierárquica . . . . .	28
3.3.2	O algoritmo EM em modelos MESN . . . . .	29
<b>4</b>	<b>Mistura Finita de Densidades</b>	<b>33</b>
4.1	Misturas finitas de densidades . . . . .	34
4.1.1	Definição . . . . .	34
4.1.2	Distribuição marginal . . . . .	34
4.1.3	Identificabilidade . . . . .	35
4.2	A Estrutura de Dados Incompletos para o Problema de Misturas . . . . .	37
4.3	O algoritmo EM em Modelos de Misturas . . . . .	39
4.4	Matriz de Informação Observada . . . . .	40
4.5	Métodos de Seleção de Modelos . . . . .	42
4.5.1	CrITÉrio de informação de <i>Akaike</i> - AIC . . . . .	42

---

4.5.2	Critério de informação bayesiano - BIC . . . . .	44
4.5.3	Critério de informação por validação cruzada - CVIC . . . . .	46
4.6	<i>Clusterização</i> com Modelos de Misturas . . . . .	47
<b>5</b>	<b>Misturas Finitas de Densidades <i>MESN</i></b>	<b>51</b>
5.1	O modelo <i>MF-MESN</i> . . . . .	52
5.1.1	Definição . . . . .	52
5.1.2	A representação hierárquica . . . . .	52
5.1.3	O algoritmo EM em modelos <i>MF-MESN</i> . . . . .	53
5.1.4	Matriz de informação observada . . . . .	56
5.1.5	Notas para implementação do algoritmo EM . . . . .	60
5.2	Aplicações a Dados Reais - Caso Univariado . . . . .	62
5.2.1	<i>Body Mass Index</i> data . . . . .	62
5.2.2	<i>Old Faithful</i> data . . . . .	63
5.3	Aplicações a Dados Reais - Caso Multivariado . . . . .	65
5.3.1	<i>Swiss Bank</i> data . . . . .	65
5.3.2	<i>Old Faithful</i> data . . . . .	69
<b>6</b>	<b>Conclusões e Perspectivas</b>	<b>73</b>
	<b>Apêndice A – Lemas</b>	<b>75</b>
	<b>Referências</b>	<b>77</b>

# *Lista de Figuras*

2.1.1 Função Densidade de Probabilidade da <i>Skew</i> -Normal . . . . .	7
2.3.1 Contornos da <i>skew</i> -normal bivariada . . . . .	15
5.2.1 Histograma dos dados de IMC com as curvas ajustadas pelos modelos MF-NOR, MF-ST and MF-SS . . . . .	64
5.2.2 Histograma dos dados <i>Old Faithful</i> com as curvas ajustadas pelos modelos MF-NOR, MF-SN, MF-ST . . . . .	66
5.3.1 Densidades de contorno e pontos classificados pelos modelos MF-NOR (esquerda) e MF-SS para os dados <i>Swiss bank</i> - pontos vermelhos classificados como notas verdadeiras	70
5.3.2 Densidades de contorno e pontos classificados pelos modelos MF-NOR (esquerda) e MF-SS para os dados <i>Old Faithful</i> . . . . .	72

## *Lista de Tabelas*

5.2.1 Estimativas de máxima verossimilhança e desvios padrão para os dados <i>IMC</i> . . . . .	64
5.2.2 Estimativas de máxima verossimilhança e desvios padrão para os dados <i>Old Faithful</i> . . . . .	65
5.3.1 Estimativas de máxima verossimilhança e desvios padrão para os dados <i>Swiss bank</i> . . . . .	68
5.3.2 Critérios de seleção de modelos para os dados <i>Swiss bank</i> . . . . .	69
5.3.3 Critérios de seleção de modelos para os dados <i>Old Faithful</i> . . . . .	71
5.3.4 Critérios de seleção de modelos para os dados perturbados . . . . .	71

# 1 *Introdução*

## 1.1 Proposta do Trabalho

A importância dada a modelos de misturas finitas em análise estatística de dados vem aumentando substancialmente ao longo dos anos. O número crescente de trabalhos publicados, tanto na literatura estatística quanto em outras áreas do conhecimento, evidenciam a grande aplicabilidade atribuída à esses modelos. Depois da publicação do trabalho de McLachlan e Basford (1988) em misturas finitas, que mostra aplicações de modelos de misturas à vários problemas outrora tratados sob outra perspectiva, muitos outros autores se engajaram nesse tema propondo novas soluções e extensões das ferramentas já utilizadas nesse contexto. Nesses trabalhos encontram-se problemas de identificabilidade, o ajuste de modelos de mistura via algoritmo EM, propriedades dos estimadores de máxima verossimilhança, a determinação do número de componentes a ser usada na mistura, resultados em teoria assintótica para inferência estatística em modelos de mistura, e muitos outros.

O modelo de mistura finita de densidades normais é sem dúvida o mais empregado nas aplicações que aparecem na literatura. Isso por que modelos de misturas finitas podem ser empregados para representar densidades de qualquer complexidade, e em particular, qualquer distribuição multivariada pode ser aproximada por uma mistura finita de densidades normais, como comentam McLachlan e Peel (2000, seção 6.1). Outro motivo para isso está no fato da simplicidade algébrica envolvida na distribuição normal, que no passado se fazia necessária devido a falta de métodos computacionalmente atrativos.

Somente nos últimos vinte anos que consideráveis avanços foram feitos em modelos de misturas finitas, em particular pelo método da máxima verossimilhança, em virtude do avanço computacional. Na década de 60, o ajuste de modelos de mistura foi estudado em uma grande quantidade de trabalhos, incluindo os artigos de Day (1969) e Wolfe (1965, 1967, 1970). Entretanto, foi com a publicação do artigo de Dempster, Laird e Rubin (1977), sobre o algoritmo EM, que se estimulou o interesse no uso de misturas de distribuições para modelar dados na presença de heterogeneidade populacional. Isso por que o ajuste de modelos de mistura por máxima verossimilhança é um exemplo clássico de um problema que é consideravelmente simplificado pelo conceito de estimação EM em dados que podem ser vistos como sendo incompletos. Como notado por Aitkin e Aitkin (1994), praticamente todas as aplicações de modelos de misturas após 1978 utilizam o algoritmo EM, como no texto de Titterington, Smith e Makov (1985) e McLachlan e Basford (1988).

Embora esses modelos sejam atrativos, há ainda a necessidade de se checar as suposições distribucionais das componentes de mistura, pois além da heterogeneidade, os dados podem apresentar tanto um comportamento assimétrico, como o de caudas mais pesadas. Neste último caso, uma alternativa ao modelo de misturas de normais seria o modelo de misturas de *t-student*, sendo esse preferível por envolver um parâmetro adicional na análise, responsável pela acomodação de valores extremos. Trabalhos considerando misturas de *t-student* incluem Peel e McLachlan (2000), Shoham (2002), Shoham et al. (2003), Lin et al. (2004) e Wang et al. (2004). Para lidar com assimetria e heterogeneidade, Lin et al. (2007b) propõem um modelo de misturas baseado na distribuição *skew-normal* univariada, proposta por Azzalini (1985). Este trabalho foi estendido por Lin et al. (2007a), que considerou também robustez à valores extremos, usando misturas de distribuições *skew-t*, proposta por Azzalini e Capitanio (2003). Modelos de misturas baseados em normais, *t-student* e *skew-normal* podem ser vistos como caso particular de misturas de *skew-t*. Para um ponto de vista utilizando abordagem bayesiana, ver Cabral et al. (2008).

Recentemente, Lin (2009) apresentou um modelo de misturas de *skew-normal* mul-

tivariada, utilizando a representação proposta por Sahu et al. (2003), mostrando sua grande flexibilidade em modelar dados heterogêneos com comportamento assimétrico. A proposta desta dissertação de mestrado é estender os trabalhos de Lin assumindo uma classe de distribuições multivariada mais flexível, a classe de mistura de escala *skew-normal* (*MESN*); Branco e Dey - 2001) já que essa contém como casos particulares versões univariadas e multivariadas da distribuição *skew-normal*, *skew-t*, a *skew-slash* e a *skew-normal* contaminada, as quais têm caudas mais pesadas que a normal (e a *skew-normal*), e conseqüentemente são mais robustas para acomodar valores extremos. Maior ênfase será dada ao problema de estimação dos parâmetros do modelo via algoritmo EM e também ao cálculo da matriz de informação observada. Esse trabalho é motivado pelo fato de muitos conjuntos de dados considerados na literatura apresentarem significativa multimodalidade e comportamento não gaussiano, tais como assimetria e caudas pesadas.

## 1.2 Organização do Trabalho

Essa dissertação de mestrado está dividida em cinco capítulos. No segundo, será apresentada a distribuição *skew-normal* proposta por Azzalini (1985). Definições e propriedades dessa distribuição serão discutidas, inicialmente para o caso univariado e em seguida para o caso multivariado. Neste capítulo, destacam-se os resultados referentes aos momentos dessa distribuição e também quanto a sua representação estocástica, a qual será utilizada de forma recorrente ao decorrer dos demais capítulos.

No terceiro capítulo, será apresentada a classe de distribuições *MESN* proposta por Branco e Dey (2001). O capítulo se inicia com a definição dessa classe em um contexto multivariado. Aqui destacam-se três resultados: o primeiro é referente à representação estocástica de um vetor aleatório com distribuição *MESN*. Então, o modelo hierárquico é apresentado e, em seguida, mostra-se como derivar o algoritmo EM para estimação dos parâmetros envolvidos no modelo.

No quarto capítulo serão discutidos resultados referentes à modelos de mistura em um



contexto geral. Este capítulo inicia-se com a definição de misturas finitas de densidades e trata também o problema de identificabilidade desses modelos. Em seguida, é apresentada a estrutura de dados incompletos para problemas de mistura, como derivar o algoritmo EM e a matriz de informação observada, nesse contexto. Por final, alguns métodos de seleção de modelos são discutidos.

O quinto capítulo pode ser considerado o objetivo desse trabalho. Aqui será apresentado o modelo de misturas de distribuições da classe *MESN*. Serão discutidos sua representação hierárquica, algoritmo EM para estimação dos parâmetros e também como obter a matriz de informação de uma forma bastante geral. Por fim, quatro aplicações desses modelos são consideradas, tanto no contexto univariado como no multivariado. Nessas aplicações serão considerados os problemas de estimação de densidades complexas e também o problema de *clusterização* de observações.

No sexto e último capítulo serão apresentadas as conclusões desse trabalho e perspectivas futuras.

## 2 *Distribuição Skew-Normal*

Existe uma tendência geral na literatura estatística para encontrar métodos flexíveis que representem da forma mais verossímil possível as características de fenômenos encontrados nas mais diversas áreas da ciência. Em muitos métodos propostos se faz presente a suposição de normalidade dos dados, essa que nem sempre pode ser a mais adequada. Para modelar desvios desta suposição, muitos enfoques podem ser encontrados na literatura. Possivelmente, o método mais comum adotado para alcançar a simetria é a transformação de variáveis. Embora essa estratégia possa trazer resultados razoáveis Azzalini e Capitanio (1999) apresentam inúmeras razões para se evitar esse procedimento.

Portanto, estudar modelos paramétricos que sejam mais robustos que a distribuição Gaussiana é de grande interesse na literatura atual, e a construção de novas famílias de distribuições que sejam capazes de incorporar assimetria, de forma analiticamente tratável, tornaram-se uma grande motivação para pesquisadores nos últimos anos. Sob essa motivação é que será proposto neste capítulo um estudo centrado na distribuição *skew-normal*.

Neste capítulo será proposta uma revisão sobre algumas das propriedades e resultados referentes à distribuição *skew-normal* univariada e multivariada. Essa distribuição será utilizada nos demais capítulos para compor uma nova classe de distribuições de mistura de escala *skew-normal* (*MESN*), classe essa menos sensível a valores extremos. Por praticidade, o capítulo é iniciado tratando da distribuição *skew-normal* univariada e uniparamétrica, mostrando algumas de suas propriedades e caracterizações. Em seguida, se faz a relação entre a forma uniparamétrica com o modelo de três parâmetros. A versão

multivariada é então apresentada, com ênfase em algumas de suas propriedades e caracterizações.

Para uma discussão mais detalhada sobre os resultados propostos nesse capítulo, e muitos outros adicionais, ver Azzalini (1985, 1986, 2006). É vasta a quantidade de trabalhos relevantes sobre esse assunto na literatura, muitos deles encontram-se citados neste trabalho.

## 2.1 A Distribuição *Skew-Normal* Padrão Univariada

### 2.1.1 Função densidade de probabilidade e função de distribuição acumulada

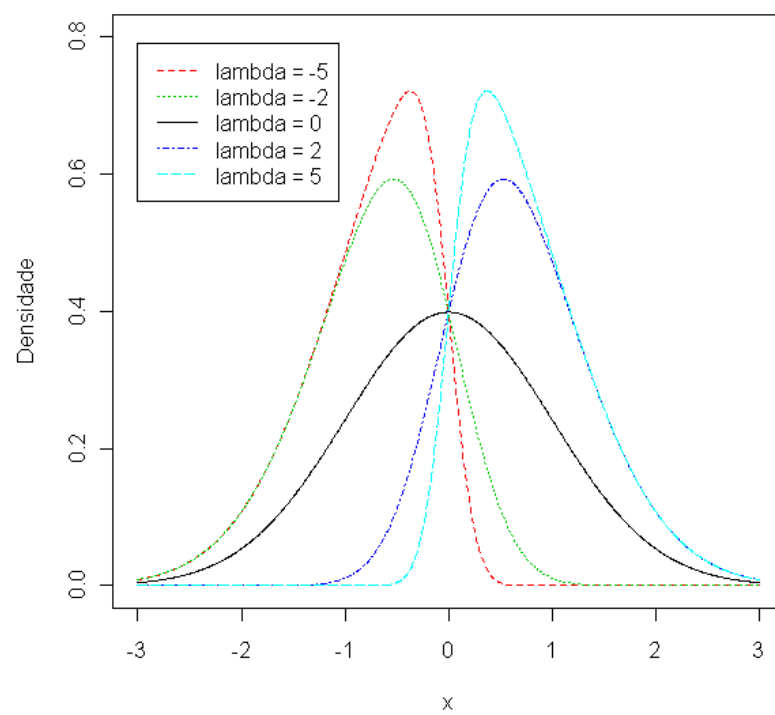
**Definição 2.1.1.** *Uma variável aleatória  $Z$  tem distribuição skew-normal padrão se sua função de densidade de probabilidade é dada por*

$$f_Z(z) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R}, \quad (2.1.1)$$

onde  $\phi(\cdot)$  e  $\Phi(\cdot)$  denotam a função densidade de probabilidade (fdp) e a função de distribuição acumulada (fda) de uma normal padrão, respectivamente, com  $\lambda \in \mathbb{R}$ . O parâmetro  $\lambda$  está associado com a forma da distribuição e se é identicamente nulo determina em (2.1.1) a densidade da normal padrão. Valores positivos (negativos) de  $\lambda$  indicam assimetria positiva (negativa) na densidade acima. Aos seguintes capítulos será utilizada a notação  $Z \sim SN(\lambda)$  para representar uma variável aleatória com densidade skew-normal de parâmetro  $\lambda$ . A Figura 2.1.1 apresenta alguns exemplos de densidades skew-normal para diferentes valores do parâmetro de assimetria.

**Proposição 2.1.1.** *A função de distribuição associada a densidade (2.1.1), denotada por  $F_Z(z; \lambda)$  é da forma*

$$F_Z(z; \lambda) = 2\Phi_2((z, 0)^\top; \mathbf{0}, \Omega), \quad \text{com } \Omega = \begin{bmatrix} 1 & -\delta \\ -\delta & 1 \end{bmatrix}, \quad \delta = \frac{\lambda}{\sqrt{1 + \lambda^2}}. \quad (2.1.2)$$

Figura 2.1.1: Função Densidade de Probabilidade da *Skew-Normal*

onde  $\Phi_2(\cdot; \mathbf{0}, \Omega)$  denota a função de distribuição acumulada de uma normal bivariada com vetor de médias zero e matriz de variância-covariância  $\Omega$ .

*Demonstração.*

$$\begin{aligned}
F_Z(z; \lambda) &= 2 \int_{-\infty}^z \phi(t) \Phi(\lambda z) dt = 2 \int_{-\infty}^z \int_{-\infty}^{\lambda z} \phi(t) \phi(u) du dt \\
&\stackrel{(a)}{=} 2\sqrt{1+\lambda^2} \int_{-\infty}^z \int_{-\infty}^0 \phi(t) \phi(v\sqrt{1+\lambda^2} + \lambda t) dv dt \\
&= 2 \int_{-\infty}^z \int_{-\infty}^0 \frac{\sqrt{1+\lambda^2}}{2\pi} \exp \left\{ \frac{1}{2} (v^2(1+\lambda^2) + 2\lambda\sqrt{1+\lambda^2}vt + \lambda^2t^2) \right\} dv dt \\
&= 2 \int_{-\infty}^z \int_{-\infty}^0 \frac{\sqrt{1+\lambda^2}}{2\pi} \exp \left\{ \frac{1}{2} \left[ \begin{pmatrix} t \\ v \end{pmatrix}^t \begin{pmatrix} 1+\lambda^2 & \lambda\sqrt{1+\lambda^2} \\ \lambda\sqrt{1+\lambda^2} & 1+\lambda^2 \end{pmatrix} \begin{pmatrix} t \\ v \end{pmatrix} \right] \right\} dv dt \\
&= 2 \int_{-\infty}^z \int_{-\infty}^0 \frac{\sqrt{1+\lambda^2}}{2\pi} \exp \left\{ \frac{1}{2} \left[ \begin{pmatrix} t \\ v \end{pmatrix}^t \begin{pmatrix} 1 & \frac{\lambda}{\sqrt{1+\lambda^2}} \\ \frac{\lambda}{\sqrt{1+\lambda^2}} & 1 \end{pmatrix}^{-1} \begin{pmatrix} t \\ v \end{pmatrix} \right] \right\} dv dt \\
&= 2 \int_{-\infty}^z \int_{-\infty}^0 \frac{1}{2\pi\sqrt{1-\delta^2}} \exp \left\{ \frac{1}{2} \left[ \begin{pmatrix} t \\ v \end{pmatrix}^t \begin{pmatrix} 1 & -\delta \\ -\delta & 1 \end{pmatrix}^{-1} \begin{pmatrix} t \\ v \end{pmatrix} \right] \right\} dv dt,
\end{aligned}$$

onde (a) segue da transformação  $v = \frac{u-\lambda t}{\sqrt{1+\lambda^2}}$  □

### 2.1.2 Propriedades

As propriedades listadas a seguir são dadas para uma variável aleatória  $Z$ , com distribuição *skew-normal* de parâmetro  $\lambda$ . As demonstrações são aqui omitidas por não serem de principal foco deste trabalho. Para uma discussão mais detalhada quanto as propriedades seguintes ver Azzalini (1985), Bayes (2004) e Bazán (2005).

#### Propriedades

1. Se  $Z \sim SN(\lambda)$ , então  $|Z| \sim HN(0, 1)$  (a distribuição *Half-Normal* padrão)
2. Quando  $\lambda \rightarrow \infty$  a densidade 2.1.1 converge para uma  $HN(0, 1)$

3. Se  $Z \sim SN(\lambda)$ , então  $-Z \sim SN(-\lambda)$
4. A densidade 2.1.1 é log-côncava
5.  $1 - F_Z(z; \lambda) = F_Z(z; -\lambda)$
6.  $F_Z(z; 1) = [\Phi(z)]^2$
7. Se  $Z \sim SN(\lambda)$ , então  $Z^2 \sim \chi_1^2$

A seguir, será derivada a função geradora de momentos da distribuição *skew-normal*, da qual pode-se obter medidas de interesse como média, variância, coeficientes de assimetria e curtoses.

**Proposição 2.1.2** (Função geratriz de momentos). *Seja  $Z \sim SN(\lambda)$ , então sua fgm é dada por*

$$M_Z(t) = 2e^{\frac{t^2}{2}} \Phi(\delta t), \quad t \in \mathbb{R}$$

com  $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$ .

*Demonstração.*

$$\begin{aligned}
 M_Z(t) &= E[e^{zt}] = 2 \int_{-\infty}^{\infty} e^{zt} \phi(z) \Phi(\lambda z) dz \\
 &= 2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2 - 2zt)} \Phi(\lambda z) dz \\
 &= 2e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} \Phi(\lambda z) dz \\
 &\stackrel{x=z-t}{=} 2e^{t^2/2} \int_{-\infty}^{\infty} \Phi(\lambda x + \lambda t) \phi(x) dx \\
 &= 2e^{t^2/2} E[\Phi(\lambda x + \lambda t | 0, 1)].
 \end{aligned}$$

Utilizando o Lema A.1 com  $\mathbf{a} = \lambda x$ ,  $\mathbf{B} = \lambda$ ,  $\boldsymbol{\mu} = \boldsymbol{\eta} = 0$  e  $\boldsymbol{\Sigma} = \boldsymbol{\Omega} = 1$ , segue o resultado.  $\square$

A partir dessa proposição pode-se mostrar que

$$E[Z] = \sqrt{\frac{2}{\pi}} \delta, \quad Var[Z] = 1 - \frac{2}{\pi} \delta^2.$$

Ainda, os momentos de ordem par e de ordem ímpar (ver Henze, 1986) ficam dados por

$$E[Z^{2k}] = 2^{-k} \frac{(2k)!}{k!}, \quad (2.1.3)$$

$$E[Z^{2k+1}] = \sqrt{\frac{2}{\pi}} \lambda (1 + \lambda^2)^{-(k+1/2)} 2^{-k} [(2k+1)!] \sum_{j=0}^k \frac{j! (2\lambda)^{2j}}{(2j+1)! (k-j)!}. \quad (2.1.4)$$

Os coeficientes de assimetria e curtose, definidos por

$$\gamma = \frac{E[Z - E(Z)]^3}{(Var[Z])^{3/2}} \quad \text{e} \quad \kappa = \frac{E[Z - E(Z)]^4}{(Var[Z])^2} - 3$$

podem ser obtidos utilizando as relações (2.1.3) e (2.1.4) e ficam dados por

$$\gamma = \sqrt{\frac{2}{\pi}} \left( \frac{4}{\pi} - 1 \right) \delta^3 \left( 1 - \frac{2}{\pi} \delta^2 \right)^{-\frac{3}{2}}, \quad (2.1.5)$$

$$\kappa = \frac{8}{\pi^2} (\pi - 3) \delta^4 \left( 1 - \frac{2}{\pi} \delta^2 \right)^{-2}. \quad (2.1.6)$$

O coeficiente de assimetria é uma medida que caracteriza o quanto a distribuição se afasta da condição de simetria. Note que  $\gamma$  é crescente em  $\lambda$ , e se  $\lambda = 0$  então  $\gamma = 0$  indicando uma distribuição simétrica. O coeficiente de curtose é uma medida que caracteriza a distribuição quanto ao seu achatamento.

### 2.1.3 Caracterizações

**Proposição 2.1.3.** *Seja  $Y_1, Y_2 \stackrel{iid}{\sim} N(0, 1)$  e definimos a variável aleatória  $Z = [Y_1 | \lambda Y_1 > Y_2]$ . Então  $Z \sim SN(\lambda)$*

*Demonstração.*

$$\begin{aligned}
 P(Z < z) &= P(Y_1 < z | \lambda Y_1 > Y_2) = \frac{P(Y_1 < z, \lambda Y_1 > Y_2)}{P(\lambda Y_1 > Y_2)} \\
 &= \frac{1}{P(\lambda Y_1 > Y_2)} \int_{-\infty}^z \int_{-\infty}^{\lambda y} \phi(y) \phi(w) dw dy \\
 &= \frac{1}{P(\lambda Y_1 > Y_2)} \int_{-\infty}^z \phi(y) \Phi(\lambda y) dy,
 \end{aligned}$$

basta notar agora que  $P(\lambda Y_1 > Y_2) = \frac{1}{2}$  por simetria.  $\square$

**Proposição 2.1.4.** *Seja o vetor  $(Y_1, Y_2)$  com distribuição normal bivariada com marginais  $N(0, 1)$  e correlação  $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$  e definimos a variável aleatória  $Z = [Y_1 | Y_2 > 0]$ . Então  $Z \sim SN(\lambda)$ .*

*Demonstração.* Podemos escrever, sem perda de generalidade,  $Y_2 = \delta Y_1 - (1 - \delta^2)^{1/2} W$  com  $Y_1$  e  $W$  i.i.d.  $N(0, 1)$ . Então o evento  $[Y_2 > 0]$  é equivalente a  $[\frac{\lambda}{\sqrt{1+\lambda^2}} Y_1 > W]$ . A prova segue da Proposição 2.1.3.  $\square$

A seguir será apresentada a representação estocástica de uma variável aleatória *skew-normal* padronizada. Tal representação é de suma importância no decorrer desse trabalho visto que inúmeras propriedades e resultados podem ser derivados desta.

**Proposição 2.1.5** (Representação Estocástica). *Seja  $X_0, X_1 \stackrel{iid}{\sim} N(0, 1)$  então*

$$Z = \delta |X_0| + (1 - \delta^2)^{1/2} X_1, \quad \delta = \frac{\lambda}{\sqrt{1 + \lambda^2}}, \quad (2.1.7)$$

*tem distribuição  $SN(\lambda)$*

*Demonstração.* Note que  $Z | |X_0| = t \sim N(\delta t, (1 - \delta^2))$  com  $|X_0| \sim HN(0, 1)$  (a distribuição *half-normal* padronizada). Então, pelo lema A.2 temos que

$$\begin{aligned}
 f_Z(z) &= \int_0^\infty \phi(z | \delta t, (1 - \delta^2)) 2\phi(t) dt \\
 &= 2 \int_0^\infty \phi(z | 0, 1) \phi(t | \delta z, (1 - \delta^2)) dt \\
 &= 2\phi(z | 0, 1) \Phi\left(\frac{\delta z}{\sqrt{1 - \delta^2}}\right),
 \end{aligned}$$



isto é,  $Z \sim SN(\lambda)$  com  $\lambda = \frac{\delta}{\sqrt{1-\delta^2}}$  □

As três proposições apresentadas anteriormente podem ser utilizadas para gerar valores pseudo-aleatórios da distribuição *skew-normal* padrão.

## 2.2 A Representação Univariada com Três Parâmetros

É natural querer estender o modelo (2.1.1) introduzindo parâmetros de locação e escala. Neste caso diz-se que  $Y \sim SN(\mu, \sigma^2, \lambda)$ .

**Definição 2.2.1.**  *$Y$  é uma variável aleatória skew-normal com parâmetro de locação  $\mu \in \mathbb{R}$  e escala  $\sigma^2 \in \mathbb{R}_+^*$  se sua função densidade de probabilidade é da forma*

$$f_Y(y) = \frac{2}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) \Phi\left(\lambda \frac{y-\mu}{\sigma}\right), \quad y \in \mathbb{R} \quad (2.2.1)$$

A seguinte proposição pode ser utilizada para gerar numeros pseudo-aleatórios a partir de uma distribuição *skew-normal* com três parâmetros. A demonstração desse resultado é direta.

**Proposição 2.2.1.** *Se  $Z \sim SN(\lambda)$  e  $Y = \mu + \sigma Z$ , então  $Y \sim SN(\mu, \sigma^2, \lambda)$*

**Proposição 2.2.2.** *A função de distribuição associada a densidade 2.2.1, denotada por  $F_Y(y; \mu, \sigma^2, \lambda)$  é da forma*

$$F_Y(y; \mu, \sigma^2, \lambda) = 2\Phi_2\left(\frac{y-\mu}{\sigma}, 0; \mathbf{0}, \Omega\right), \quad (2.2.2)$$

com

$$\Omega = \begin{bmatrix} 1 & -\delta \\ -\delta & 1 \end{bmatrix}, \quad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}.$$

onde  $\Phi_2(\cdot; \mathbf{0}, \Omega)$  denota a função de distribuição acumulada de uma normal bivariada com vetor de médias zero e matriz de variância-covariância  $\Omega$ .

*Demonstração.* Sem perda de generalidade, considere  $Y = \mu + \sigma Z$  com  $Z \sim SN(\lambda)$ , então

$$F_Y(y; \mu, \sigma^2, \lambda) = P(Y \leq y) = P(\mu + \sigma Z \leq y) = P(Z \leq \frac{y - \mu}{\sigma})$$

o que segue e consequente da Proposição 2.1.1.  $\square$

A próxima proposição, cuja prova é direta, dá a função geratriz de momentos de uma distribuição *skew-normal* de três parâmetros

**Proposição 2.2.3.** *A função geratriz de momentos da distribuição skew-normal com três parâmetro é dada por*

$$M_Y(t) = 2e^{t\mu + \frac{t^2\sigma^2}{2}} \Phi(\delta\sigma t).$$

Utilizando a Proposição 2.2.1 (ou a fgm para o caso de três parâmetros) pode-se mostrar que a média e variância de uma variável aleatória  $Y \sim SN(\mu, \sigma^2, \lambda)$  são expressas por

$$E[Y] = \mu + \sigma\delta\sqrt{\frac{2}{\pi}}, \quad Var[Y] = \sigma^2 \left(1 - \frac{2}{\pi}\delta^2\right).$$

Os coeficientes de assimetria e de curtose são os mesmos do caso uniparamétrico.

## 2.3 A Distribuição *Skew-Normal* Multivariada

Existem muitas definições de distribuições *skew-normal* multivariada, com diferentes parametrizações e interpretações. Nesta seção, será considerada uma definição unificada das definições encontradas em Arellano-Valle, Bolfarine e Lachos (2005). A escolha dessa definição para esse trabalho se deve ao fato de que muitas propriedades, caracterizações e manipulações algébricas são obtidas facilmente a partir desta e, além disso, com essa definição o caso univariado pode ser visto como um caso particular derivado da representação multivariada, o que pode não ocorrer com outras definições dessa distribuição, fazendo-se necessária algumas reparametrizações. A seguir, será introduzida uma notação que será utilizada ao decorrer desse trabalho.

Denota-se por  $\phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  e  $\Phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  a fdp e a fda, respectivamente, da distribuição  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  avaliada em  $\mathbf{x}$ . Quando  $\boldsymbol{\mu} = \mathbf{0}$  e  $\boldsymbol{\Sigma} = \mathbf{I}_p$  (a matriz identidade  $p \times p$ ), denotam-se essas funções como  $\phi_p(\mathbf{x})$  e  $\Phi_p(\mathbf{x})$ .

### 2.3.1 Função densidade de probabilidade e função de distribuição acumulada

**Definição 2.3.1.** *Um vetor aleatório  $p$ -dimensional  $\mathbf{Y}$  segue uma distribuição skew-normal com vetor de locação  $\boldsymbol{\mu} \in \mathbb{R}^p$ , matriz de variância-covariância  $\boldsymbol{\Sigma}$  (definida positiva) e vetor de assimetria  $\boldsymbol{\lambda} \in \mathbb{R}^p$ , se sua fdp é dada por*

$$f_Y(\mathbf{y}) = 2\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(A), \quad \mathbf{y} \in \mathbb{R}^p, \quad (2.3.1)$$

tal que  $A = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$ . Denota-se por  $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ . Quando  $\boldsymbol{\mu} = \mathbf{0}$  e  $\boldsymbol{\Sigma} = \mathbf{I}_p$  temos

$$f_Y(\mathbf{y}) = 2\phi_p(\mathbf{y})\Phi_1(\boldsymbol{\lambda}^\top \mathbf{y}), \quad \mathbf{y} \in \mathbb{R}^p, \quad (2.3.2)$$

e neste caso denota-se  $\mathbf{Y} \sim SN_p(\boldsymbol{\lambda})$

A Figura 2.3.1 apresenta os contornos de uma distribuição skew-normal bivariada com  $\boldsymbol{\mu} = (0, 0)^\top$ ,  $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$  e  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$  para diferentes valores de  $\lambda_1$ ,  $\lambda_2$  e  $\rho$ .

**Proposição 2.3.1.** *A função de distribuição acumulada associada a uma variável aleatória  $\mathbf{Z} \sim SN_p(\boldsymbol{\lambda})$  é dada por*

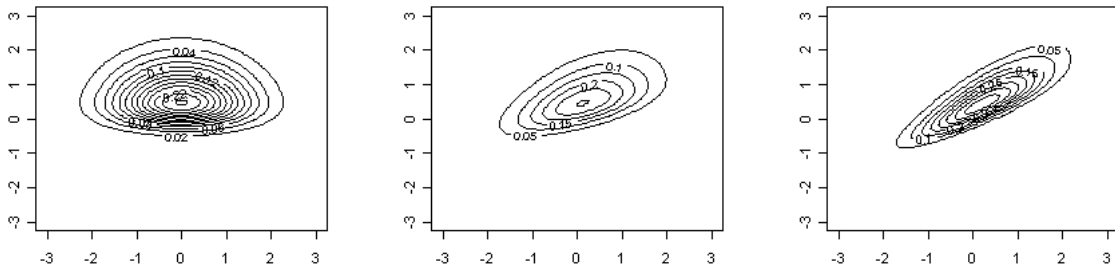
$$F_Z(\mathbf{z}) = 2\Phi_{p+1}((\mathbf{z}^\top, 0)^\top; \mathbf{0}, \boldsymbol{\Omega}), \text{ com } \boldsymbol{\Omega} = \begin{bmatrix} \mathbf{I}_p & -\boldsymbol{\lambda} \\ -\boldsymbol{\lambda} & 1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda} \end{bmatrix}. \quad (2.3.3)$$

*Demonstração.*

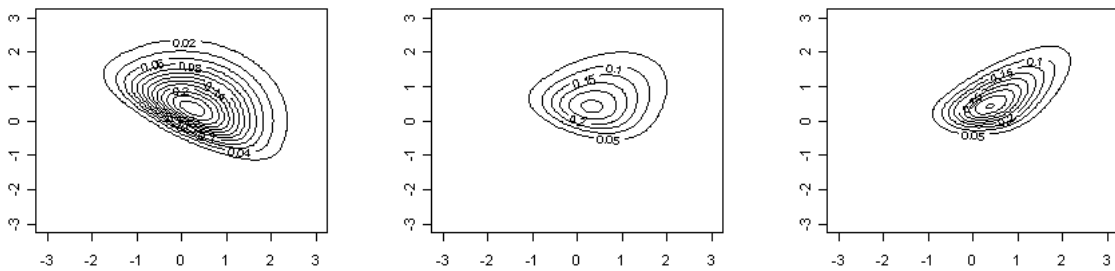
$$\begin{aligned} F_Z(\mathbf{z}) &= P(\mathbf{Z} \leq \mathbf{z}) = 2 \int_{\mathbf{u} \leq \mathbf{0}} \phi_p(\mathbf{u} + \mathbf{z}) \Phi_1(\boldsymbol{\lambda}^\top (\mathbf{u} + \mathbf{z})) d\mathbf{u} \\ &= 2 \int_{\mathbf{u} \leq \mathbf{0}} \int_{v \leq 0} \phi_p(\mathbf{u} + \mathbf{z}) \phi_1(v + \boldsymbol{\lambda}^\top (\mathbf{u} + \mathbf{z})) dv d\mathbf{u} \\ &= 2P(\mathbf{U} \leq \mathbf{0}, V \leq 0) \end{aligned}$$

Figura 2.3.1: Contornos da *skew*-normal bivariada

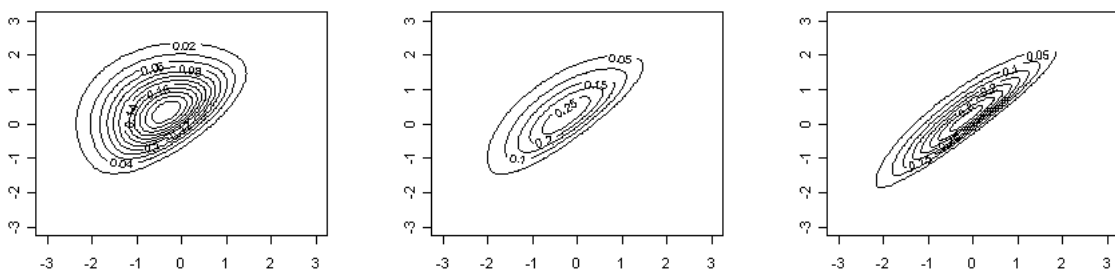
(a) Para  $\boldsymbol{\lambda} = (0, 3)^\top$  e  $\rho = 0, 0.5$  e  $0.9$ , respectivamente



(b) Para  $\boldsymbol{\lambda} = (2, 3)^\top$  e  $\rho = 0, 0.5$  e  $0.9$ , respectivamente



(c) Para  $\boldsymbol{\lambda} = (-2, 2)^\top$  e  $\rho = 0, 0.5$  e  $0.9$ , respectivamente



onde  $V|\mathbf{U} = \mathbf{u} \sim N_1(-\boldsymbol{\lambda}^\top(\mathbf{u} + \mathbf{z}), 1)$  e  $\mathbf{U} \sim N_p(-\mathbf{z}, \mathbf{I}_p)$ . Assim, a prova segue do fato de

$$\begin{pmatrix} \mathbf{U} \\ V \end{pmatrix} \sim N_{p+1} \left( \begin{pmatrix} -\mathbf{z} \\ 0 \end{pmatrix}, \begin{bmatrix} \mathbf{I}_p & -\boldsymbol{\lambda} \\ -\boldsymbol{\lambda} & 1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda} \end{bmatrix} \right).$$

□

**Observação 2.3.1.** *A fda de uma variável aleatória skew-normal de três parâmetros pode ser obtida utilizando a Proposição 2.3.2.*

### 2.3.2 Propriedades

A primeira propriedade apresentada nessa seção é a relação entre a distribuição *skew-normal* multivariada de três parâmetros com a distribuição *skew-normal* multivariada de apenas um parâmetro.

**Proposição 2.3.2.** *Seja  $\mathbf{Z} \sim SN_p(\boldsymbol{\lambda})$  e considere a transformação linear  $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Z}$ , onde  $\boldsymbol{\Sigma}$  é definida Positiva. Então*

$$\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}).$$

*Demonstração.* A prova segue do fato que  $f_Y(\mathbf{y}) = |\boldsymbol{\Sigma}|^{-1/2} f_Z(\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}))$ . □

As propriedades listadas a seguir são dadas para uma variável aleatória  $\mathbf{Z}$  com distribuição *skew-normal* multivariada de parâmetro  $\boldsymbol{\lambda}$ . Novamente, as demonstrações estão omitidas por não serem de foco principal neste trabalho. Para uma discussão mais detalhada quanto as propriedades e demonstrações, ver Lachos (2004).

#### Propriedades

1.  $-\mathbf{Z} \sim SN_p(-\boldsymbol{\lambda})$
2.  $\mathbf{a}^\top \mathbf{Z} \sim SN_p(\mathbf{a}^\top \boldsymbol{\lambda})$ , para algum vetor unitário  $\mathbf{a} \in \mathbb{R}$
3.  $\mathbf{AZ} \sim SN_p(\mathbf{A}\boldsymbol{\lambda})$ , para alguma matriz ortogonal  $\mathbf{A}_{n \times n}$

No próximo resultado será apresentada a função geratriz de momentos da distribuição skew-normal multivariada padronizada. Propriedades adicionais para essa distribuição serão obtidas a partir dessa.

**Proposição 2.3.3.** *Seja  $\mathbf{Z} \sim SN_p(\boldsymbol{\lambda})$ . Então sua fgm é dada por*

$$M_Z(s) = 2e^{\frac{1}{2}s^\top s} \Phi_1(\delta s), \quad s \in \mathbb{R}^p.$$

*Demonstração.* Note que  $e^{s^\top \mathbf{Z}} \phi_p(z) = e^{\frac{1}{2}s^\top s} \phi_p(\mathbf{z} - s)$ , então

$$\begin{aligned} M_Z(s) &= E(e^{s^\top \mathbf{Z}}) = 2 \int_{\mathbb{R}} e^{\frac{1}{2}s^\top s} \phi_p(\mathbf{u} - s) \Phi_1(\boldsymbol{\lambda} \mathbf{u}) d\mathbf{u} \\ &= 2e^{\frac{1}{2}s^\top s} \int_{\mathbb{R}} \phi_p(\mathbf{y}) \Phi_1(\boldsymbol{\lambda}^\top (\mathbf{y} + s)) d\mathbf{y} = 2e^{\frac{1}{2}s^\top s} \int_{\mathbb{R}} \int_{u \leq \boldsymbol{\lambda}^\top s} \phi_p(\mathbf{y}) \phi_1(u + \boldsymbol{\lambda}^\top \mathbf{y}) d\mathbf{y} du \\ &= 2e^{\frac{1}{2}s^\top s} \int_{u \leq \boldsymbol{\lambda}^\top s} E[\phi_1(u + \boldsymbol{\lambda}^\top \mathbf{Y})] du, \quad \text{com } \mathbf{Y} \sim N_p(0, \mathbf{I}_p) \end{aligned}$$

o resultado segue usando o Lema A.1 □

Como consequência imediata da proposição anterior temos o seguinte corolário

**Corolário 2.3.1.** *Seja  $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ . Então sua fgm é dada por*

$$M_Y(s) = 2e^{s^\top \boldsymbol{\mu} + \frac{1}{2}s^\top \boldsymbol{\Sigma} s} \Phi_1(\delta^\top \boldsymbol{\Sigma}^{1/2} s), \quad s \in \mathbb{R}^p. \quad (2.3.4)$$

*Demonstração.* A prova segue de  $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \mathbf{Z}$  e  $M_{\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \mathbf{Z}}(s) = e^{s^\top \boldsymbol{\mu}} M_Z(\boldsymbol{\Sigma}^{1/2} s)$  □

As proposições apresentadas a seguir são resultados relacionados com os momentos de um vetor aleatório skew-normal, cujas demonstrações podem ser obtidas a partir de sua fgm.

**Proposição 2.3.4.** *Seja  $\mathbf{Z} \sim SN_p(\boldsymbol{\lambda})$ , então*

$$a) \quad \boldsymbol{\mu}_1 = E[\mathbf{Z}] = \sqrt{\frac{2}{\pi}} \boldsymbol{\delta},$$

$$b) \quad \boldsymbol{\mu}_2 = E[\mathbf{Z} \mathbf{Z}^\top] = \mathbf{I}_p, \quad \text{Var}(\mathbf{Z}) = \mathbf{I}_p - \frac{2}{\pi} \boldsymbol{\delta} \boldsymbol{\delta}^\top,$$

$$c) \boldsymbol{\mu}_3 = E[\text{vec}(\mathbf{Z}\mathbf{Z}^\top)]\mathbf{Z}^\top = \sqrt{\frac{2}{\pi}}[\boldsymbol{\delta} \otimes \mathbf{I}_p + \text{vec}(\mathbf{I}_p)\boldsymbol{\delta}^\top + \mathbf{I}_p \otimes \boldsymbol{\delta} - \boldsymbol{\delta} \otimes \boldsymbol{\delta}\boldsymbol{\delta}^\top]$$

$$d) \boldsymbol{\mu}_4 = E[\text{vec}(\mathbf{Z}\mathbf{Z}^\top) \otimes \text{vec}(\mathbf{Z}\mathbf{Z}^\top)^\top] = \mathbf{I}_{p^2} + \mathbf{K}_{pm} + \text{vec}(\mathbf{I}_p)\text{vec}^\top(\mathbf{I}_p)$$

**Proposição 2.3.5.** *Seja  $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ , então*

$$a) E[\mathbf{Y}] = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta},$$

$$b) E[\mathbf{Y}\mathbf{Y}^\top] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top,$$

$$c) \text{Var}(\mathbf{Y}) = \boldsymbol{\Sigma} - \frac{2}{\pi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{1/2}$$

### 2.3.3 Caracterizações

Os resultados apresentados nessa seção são de suma importância ao decorrer desse trabalho e serão utilizados consistentemente nas seguintes seções. A representação estocástica de um vetor aleatório *skew-normal* pode ser utilizada também para a geração de vetores pseudo aleatórios.

**Proposição 2.3.6.** *Seja  $\mathbf{Z} \sim SN_p(\boldsymbol{\lambda})$ , então*

$$\mathbf{Z} \stackrel{d}{=} \boldsymbol{\delta}|X_0| + (\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^\top)^{1/2}\mathbf{X}_1, \quad \text{com } \boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{(1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{1/2}}$$

e  $X_0 \sim N(0, 1)$  e  $\mathbf{X}_1 \sim N_p(\mathbf{0}, \mathbf{I}_p)$  independentes.

*Demonstração.* Seja  $\mathbf{Z} = \boldsymbol{\delta}|X_0| + (\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^\top)^{1/2}\mathbf{X}_1$ . Note ainda que  $\mathbf{Z}||X_0| = t \sim N_p(\boldsymbol{\delta}t, \mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^\top)$ , onde  $|X_0| \sim HN(0, 1)$  (a distribuição *half-normal* padronizada). Desta forma temos

que

$$\begin{aligned}
 f_{\mathbf{Z}}(\mathbf{z}) &= \int_0^\infty \phi_p(\mathbf{z}|\boldsymbol{\delta}t, \mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^\top) 2\phi(t) dt \\
 &= \int_0^\infty \phi_p(\mathbf{z}|\mathbf{0}, \mathbf{I}_p) 2\phi(t|\boldsymbol{\delta}^\top \mathbf{z}, 1 - \boldsymbol{\delta}\boldsymbol{\delta}^\top) dt \\
 &= 2\phi_p(\mathbf{z}) \int_0^\infty \phi(t|\boldsymbol{\delta}^\top \mathbf{z}, 1 - \boldsymbol{\delta}\boldsymbol{\delta}^\top) dt \\
 &= 2\phi_p(\mathbf{z}) \Phi_1 \left( \frac{\boldsymbol{\delta}^\top \mathbf{z}}{\sqrt{1 - \boldsymbol{\delta}\boldsymbol{\delta}^\top}} \right),
 \end{aligned}$$

de onde a primeira passagem segue do Lema A.2 e  $\boldsymbol{\lambda} = \frac{\boldsymbol{\lambda}^\top}{\sqrt{1 - \boldsymbol{\lambda}\boldsymbol{\lambda}^\top}}$ . □

**Observação 2.3.2.** Não é difícil ver que para  $X_0 \sim N(0, 1)$ ,  $\mathbf{X}_1 \sim N_n(\mathbf{0}, \mathbf{I}_n)$  com correlação igual a  $\boldsymbol{\delta}$ , o vetor aleatório  $\mathbf{Z} = [\mathbf{X}_1 | X_0 > 0]$  tem distribuição  $SN_n(\boldsymbol{\lambda})$ . Assim, alternativamente à representação estocástica, pode-se gerar vetores pseudo aleatórios  $SN_p(\boldsymbol{\lambda})$  através do seguinte mecanismo

$$\mathbf{Z} = \begin{cases} \mathbf{X}_1, & \text{Se } X_0 > 0 \\ -\mathbf{X}_1, & \text{c.c.} \end{cases} \quad (2.3.5)$$

Veja Wang et al. (2004) e Azzalini e Capitanio (1999) para uma discussão mais detalhada a respeito da representação estocástica condicional em 2.3.5.



### 3 *A classe de distribuições de Mistura de Escala Skew-Normal*

No capítulo anterior foi apresentada a família de distribuições *skew-normal* que contém como caso particular a distribuição normal. Esse modelo tem particular importância já que pode se adaptar às distribuições que estão em uma vizinhança de uma normal. Mesmo sendo atrativa, a *skew-normal* ainda não parece ser adequada para análise de dados com valores extremos, já que pode ser mostrado que a estimação dos parâmetros pode estar comprometida em virtude disso.

Ainda no contexto de modelos simétricos, a distribuição *student-t* torna-se uma alternativa à distribuição normal para lidar com valores extremos, já que apresenta caudas mais pesadas que a normal, podendo proporcionar ajustes mais robustos. Isso por que a *student-t* apresenta um atrativo adicional: um parâmetro extra que pode ser entendido como o responsável pela acomodação de valores extremos.

Neste sentido, a classe de distribuições de mistura de escala normal (*MEN*) pode ser vista como mais abrangente que o modelo *t-student*, propondo uma nova classe de distribuições mais robusta quanto a valores extremos. Portanto, assim como a classe *MEN* procura tratar dados com valores extremos no contexto simétrico, a classe de distribuições propostas neste capítulo procura tratar dados com valores extremos no caso assimétrico. Sob essa motivação é que será apresentada a classe de modelos de mistura de escala *skew-normal* (*MESN*).

Neste capítulo será proposta a classe de distribuições *MESN* multivariada, sua representação estocástica e algumas de suas propriedades. Em seguida, serão apresentados alguns exemplos de distribuições pertencentes a essa classe. Ao final do capítulo será apresentado o algoritmo EM para estimação de máxima verossimilhança dos parâmetros em modelos pertencentes a essa classe.

## 3.1 Distribuições MESN Multivariada

### 3.1.1 Definição

**Definição 3.1.1.** *Um vetor aleatório  $p$ -dimensional  $\mathbf{Y}$  é dito ter distribuição na classe MESN com parâmetro de locação  $\boldsymbol{\mu} \in \mathbb{R}^p$ , matriz de escala  $\boldsymbol{\Sigma}_{p \times p}$  (positiva definida) e parâmetro de assimetria  $\boldsymbol{\lambda} \in \mathbb{R}^p$  se sua função densidade de probabilidade é dada por*

$$\psi(\mathbf{y}) = 2 \int_0^\infty \phi_p(\mathbf{y}; \boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma}) \Phi_1(\kappa^{-1/2}(u)A) dH(u), \quad (3.1.1)$$

tal que  $A = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$ ,  $U$  uma variável aleatória positiva com fda  $H(\cdot; \boldsymbol{\nu})$  e fdp  $h(\cdot; \boldsymbol{\nu})$ , e  $\kappa(\cdot)$  uma função peso bem definida. Será utilizada a notação  $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$  para representar um vetor aleatório de fdp dada em (3.1.1). Quando  $\boldsymbol{\mu} = \mathbf{0}$  and  $\boldsymbol{\Sigma} = \mathbf{I}_p$  tem-se a distribuição MESN padrão denotada por  $\text{SNSM}_p(\boldsymbol{\lambda}, H)$ .

Aqui,  $\boldsymbol{\nu}$  é um parâmetro escalar ou vetorial indexando a distribuição do fator de escala  $U$ . Esse parâmetro adicional pode ser entendido por um fator de acomodação de valores extremos. Se a fda  $H(u; \boldsymbol{\nu})$  converge fracamente para a distribuição de massa pontual em 1, então pode-se afirmar que a densidade 3.1.1 converge para a densidade de um vetor aleatório com distribuição *skew-normal*. A prova desse resultado é similar àquela encontrada em Lange e Sinsheimer (1993). Quando  $\boldsymbol{\lambda} = \mathbf{0}$ , obtem-se a classe de distribuições de mistura de escala normal *MEN* representada pela fdp  $\psi_0(\mathbf{y}) = \int_0^\infty \phi_p(\mathbf{y} | \boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma}) dH(u; \boldsymbol{\nu})$ .

### 3.1.2 Representação estocástica

A representação estocástica dada abaixo pode ser utilizada para gerar números pseudo aleatórios de um vetor aleatório  $\mathbf{Y} \sim MESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}; H)$  e também para estudar muitas de suas propriedades.

**Proposição 3.1.1.** *Um vetor aleatório  $\mathbf{Y} \sim MESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$  tem representação estocástica dada por*

$$\mathbf{Y} \stackrel{d}{=} \boldsymbol{\mu} + \kappa(U)^{1/2} \mathbf{Z}, \quad (3.1.2)$$

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

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

A seguinte proposição mostra outra forma de se representar estocasticamente um vetor aleatório com distribuição na classe  $MESN$ . Esse resultado será utilizado posteriormente para representar o modelo de misturas de escala *skew*-normal hierarquicamente. Tal representação é de extrema importância para se fazer inferência estatística sob essa classe já que a partir desta pode-se implementar o algoritmo EM de forma bastante geral.

**Proposição 3.1.2.** *Seja  $\mathbf{Y} \sim MESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}; H)$ . Então sua representação estocástica pode ser dada por*

$$\mathbf{Y} \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{\Delta}T + \kappa^{1/2}(u)\boldsymbol{\Gamma}^{1/2}\mathbf{T}_1, \quad (3.1.3)$$

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

*Demonstração.* A prova segue da Proposição 3.1.1 e da representação estocástica de um vetor aleatório com distribuição *skew*-normal.  $\square$

### 3.1.3 Propriedades

A próxima proposição apresenta a função geradora de momentos de um vetor aleatório com distribuição  $MESN$ .

**Proposição 3.1.3.** *Seja  $\mathbf{Y} \sim MESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$ . Então sua fgm é dada por*

$$M_{\mathbf{Y}}(\mathbf{s}) = E[e^{\mathbf{s}^\top \mathbf{Y}}] = \int_0^\infty 2e^{\mathbf{s}^\top \boldsymbol{\mu} + \frac{1}{2}\kappa(u)\mathbf{s}^\top \boldsymbol{\Sigma} \mathbf{s}} \Phi_1(\kappa^{1/2}(u)\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{1/2} \mathbf{s}) dH(u), \quad \mathbf{s} \in \mathbb{R}^p. \quad (3.1.4)$$

*Demonstração.* Da prova da Proposição 3.1.1 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.  $\square$

Na proposição seguinte, será derivado o vetor de médias e a matriz de covariância de um vetor aleatório com distribuição MESN. A prova segue da representação estocastica (3.1.2).

**Proposição 3.1.4.** *Seja  $\mathbf{Y} \sim MESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$ , 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.

O próximo resultado apresenta o cálculo de alguns momentos condicionais importantes para a implementação do algoritmo EM. A seguinte notação será utilizada:  $\kappa_r = E[\kappa^{-r}(U)|\mathbf{y}]$  e  $\tau_r = E[\kappa^{-r/2}(U)W_\Phi(\kappa^{-1/2}(U)A)|\mathbf{y}]$ , com  $W_\Phi(x) = \phi_1(x)/\Phi(x)$ ,  $x \in \mathbb{R}$ .

**Proposição 3.1.5.** *Seja  $\mathbf{Y} \sim MESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$  e  $\mathbf{U} \sim H$  o fator de mistura de escala, então*

$$\kappa_r = \frac{2\psi_0(\mathbf{y})}{\psi(\mathbf{y})} E[\kappa^{-r}(U_y)\Phi(\kappa^{-1/2}(U_y)A)] \quad e \quad \tau_r = \frac{2\psi_0(\mathbf{y})}{\psi(\mathbf{y})} E[\kappa^{-r/2}(U_y)\phi(\kappa^{-1/2}(U_y)A)] \quad (3.1.5)$$

com  $\psi_0$  a fdp de  $Y_0 \sim MEN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, H)$  e  $U_y \stackrel{d}{=} U|Y_0$ .

*Demonstração.* Veja proposição 1 em Lachos et al. (2009).  $\square$

### 3.1.4 Distribuição marginal e independência

A proposição a seguir mostra que qualquer vetor aleatório com distribuição na classe *MESN* é invariante quanto a transformações lineares, implicando que a distribuição marginal desse vetor permanece na classe *MESN*.

**Proposição 3.1.6.** *Seja  $\mathbf{Y} \sim \text{MESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$ . 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,*

$$\mathbf{V} = \mathbf{b} + \mathbf{A}\mathbf{Y} \sim \text{MESN}_p(\mathbf{b} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top, \boldsymbol{\lambda}^*, H), \quad (3.1.6)$$

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  $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}$ . Também, para qualquer  $\mathbf{a} \in \mathbb{R}^p$ ,

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

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

*Demonstração.* A prova desse resultado é obtida diretamente da proposição 3.1.3, 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  $\boldsymbol{\delta}^* = \boldsymbol{\delta}$ .  $\square$

Pela Proposição 3.1.6, com  $\mathbf{A} = [\mathbf{I}_{p_1}, \mathbf{0}_{p_2}]$ ,  $p_1 + p_2 = p$ , tem-se o seguinte resultado para um vetor aleatório *MESN*, relacionado com sua distribuição marginal.

**Corolário 3.1.1.** *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  $p_1$  and  $p_2$ , respectivamente; Seja*

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

as correspondente partições de  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\mu}$ . Então,  $\mathbf{Y}_1 \sim \text{MESN}_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{11}^{1/2} \tilde{\mathbf{v}}; H)$ , com

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

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

## 3.2 Exemplos

### 3.2.1 Distribuição *skew-t* multivariada

A Distribuição *skew-t* multivariada (Branco and Dey, 2001; Azzalini and Capitanio, 2003) com  $\nu$  graus de liberdade,  $ST_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$  digamos, pode ser obtida do modelo de misturas de escala *skew-normal* (3.1.1), tomando  $U$  distribuído como  $Gamma(\nu/2, \nu/2)$ ,  $\nu > 0$  e  $\kappa(u) = 1/u$ . A função densidade de probabilidade de  $\mathbf{Y}$  é

$$\psi(\mathbf{y}) = 2t_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) T \left( \sqrt{\frac{\nu+p}{\nu+d}} A; \nu+p \right), \quad \mathbf{y} \in \mathbb{R}^p, \quad (3.2.1)$$

sendo  $t_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$  e  $T(\cdot; \nu)$  a fdp e a fda, respectivamente, da distribuição *t*-Student  $p$ -variada, e  $d = (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$  a distância de Mahalanobis. Um caso particular da distribuição *skew-t* é a distribuição *skew-Cauchy*, quando  $\nu = 1$ . Também, quando  $\nu \uparrow \infty$ , tem-se a distribuição *skew-normal* no limite. Aplicações da distribuição *skew-t* em estimações robustas podem ser encontradas em Lin et al. (2007b) e Azzalini and Genton (2007). Neste caso, da proposição 3.1.4, a média e a matriz de covariâncias de  $\mathbf{Y}$  são dadas por,

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

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

Além disso, para esse modelo, temos pela proposição 3.1.5 que  $\mathbf{Y}_0 \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ , i.e.  $\mathbf{Y}_0|U = u \sim N_p(\boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma})$  e  $U \sim Gamma(\nu/2, \nu/2)$ . Considerando então o fato de que  $U_{\mathbf{y}} \stackrel{d}{=} U|Y_0 = \mathbf{y} \sim Gamma((\nu+p)/2, (\nu+d)/2)$ , tem-se, depois de alguma algebra, os seguintes resultados para as esperanças condicionais  $\kappa_r$  and  $\tau_r$ .

**Corolário 3.2.1.** *Seja  $\mathbf{Y} \sim ST_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ . Então,*

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

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

### 3.2.2 Distribuição *skew-slash* multivariada

Outra distribuição da classe *MESN*, chamada como *skew-slash* multivariada e denotada por  $SSL_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ , é obtida quando a distribuição de  $U$  é  $Beta(\nu, 1)$ ,  $\nu > 0$  e  $\kappa(u) = 1/u$ . Sua fdp é dada por

$$\psi(\mathbf{y}) = 2\nu \int_0^1 u^{\nu-1} \phi_p(\mathbf{y}; \boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma}) \Phi(u^{1/2}A) du, \quad \mathbf{y} \in \mathbb{R}^p. \quad (3.2.2)$$

A distribuição *skew-slash* se reduz à *skew-normal* quando  $\nu \uparrow \infty$ . Da proposição 3.1.4, pode-se mostrar que

$$\begin{aligned} E[\mathbf{Y}] &= \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}} \frac{2\nu}{2\nu-1} \boldsymbol{\Delta}, \quad \nu > 1/2, \\ Var[\mathbf{Y}] &= \frac{\nu}{\nu-1} \boldsymbol{\Sigma} - \frac{2}{\pi} \left( \frac{2\nu}{2\nu-1} \right)^2 \boldsymbol{\Delta} \boldsymbol{\Delta}^\top, \quad \nu > 1 \end{aligned}$$

Os momentos condicionais  $\kappa_r$  e  $\tau_r$  para a distribuição *skew-slash* (3.2.2) são dados a seguir, cuja prova segue por considerar na proposição 3.1.5 que  $U_{\mathbf{Y}} \sim \text{Gamma}((2\nu + p + 2r)/2, d/2) \mathbb{I}_{(0,1)}$ . uma distribuição gamma truncada no intervalo (0,1). Aplicações da distribuição *skew-slash* podem ser encontradas em Wang e Genton (2006).

**Corolário 3.2.2.** *Seja  $\mathbf{Y} \sim SSL_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ , então*

$$\begin{aligned} \kappa_r &= \frac{\psi_0(\mathbf{y})}{\psi(\mathbf{y})} \frac{2\Gamma(\frac{2\nu+p+2r}{2})}{\Gamma(\frac{2\nu+p}{2})} \left(\frac{2}{d}\right)^r \frac{P_1\left(\frac{2\nu+p+2r}{2}, \frac{d}{2}\right)}{P_1\left(\frac{p+2\nu}{2}, \frac{d}{2}\right)} E[\Phi(S^{1/2}A)], \\ \tau_r &= \frac{\psi_0(\mathbf{y})}{\psi(\mathbf{y})} \frac{2^{r/2+1/2} \Gamma(\frac{2\nu+p+r}{2})}{\Gamma(\frac{2\nu+p}{2}) \pi^{1/2}} \frac{d^{(2\nu+p)/2}}{(d+A^2)^{(2\nu+p+r)/2}} \frac{P_1\left(\frac{2\nu+p+r}{2}, \frac{d+A^2}{2}\right)}{P_1\left(\frac{p+2\nu}{2}, \frac{d}{2}\right)}, \end{aligned}$$

sendo  $P_x(a, b)$  a fda da distribuição  $\text{Gamma}(a, b)$  em  $x$  e  $S \sim \text{Gamma}((2\nu+p+2r)/2, d/2) \mathbb{I}_{(0,1)}$ .

Note que no corolário anterior a expressão  $E[\Phi(S^{1/2}A)]$  pode ser computada por integração Monte Carlo, gerando observações independentes  $S_1, \dots, S_L$  de  $S$  e aproximar então a esperança por  $\frac{1}{L} \sum_{i=1}^L \Phi(S_i A)$ . Observações de gamma truncada podem ser geradas

usando o pacote R RUNRAN (R Development Core Team, 2008) com a função URGAMMA

### 3.2.3 Distribuição *skew-normal* contaminada multivariada

A distribuição *skew-normal* contaminada é obtida quando o fator de mistura de escala  $U$  é uma variável aleatória discreta tomando um de dois estados. A função de probabilidade de  $U$ , dado o vetor de parâmetros  $\boldsymbol{\nu} = (\nu_1, \nu_2)^\top$ , é denotada por

$$h(u; \boldsymbol{\nu}) = \nu_1 \mathbb{I}_{(u=\nu_2)} + (1 - \nu_1) \mathbb{I}_{(u=1)}, \quad 0 < \nu_1 < 1, \quad 0 < \nu_2 \leq 1. \quad (3.2.3)$$

Para  $\kappa(u) = 1/u$ , segue que

$$\psi(\mathbf{y}) = 2 \left\{ \nu_1 \phi_p(\mathbf{y}; \boldsymbol{\mu}, \nu_2^{-1} \boldsymbol{\Sigma}) \Phi(\nu_2^{1/2} A) + (1 - \nu_1) \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi(A) \right\}.$$

Essa distribuição é denotada por  $SCN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ . O parâmetro  $\nu_1$  pode ser interpretado como a proporção de valores extremos, ou *outliers*, enquanto  $\nu_2$  pode ser interpretado como um fator de escala. A distribuição *skew-normal* contaminada se reduz a distribuição *skew-normal* quando  $\nu_1 = \nu_2 = 1$ . Da proposição 3.1.4, pode-se mostrar que

$$\begin{aligned} E[\mathbf{Y}] &= \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}} \left( \frac{\nu_1}{\nu_2^{1/2}} + 1 - \nu_1 \right) \boldsymbol{\Delta}, \\ \text{Var}[\mathbf{Y}] &= \left( \frac{\nu_1}{\nu_2} + 1 - \nu_1 \right) \boldsymbol{\Sigma} - \frac{2}{\pi} \left( \frac{\nu_1}{\nu_2^{1/2}} + 1 - \nu_1 \right)^2 \boldsymbol{\Delta} \boldsymbol{\Delta}^\top. \end{aligned}$$

Considerando que  $U_{\mathbf{y}}$  é uma variável aleatória com função de probabilidade condicional  $h_0(u|\mathbf{y}) = (1/f_0(\mathbf{y})) \{ \nu_1 \phi_p(\mathbf{y}; \boldsymbol{\mu}, \nu_2^{-1} \boldsymbol{\Sigma}) \mathbb{I}_{(u=\nu_2)} + (1 - \nu_1) \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathbb{I}_{(u=1)} \}$ , para  $f_0(\mathbf{y}) = \nu_1 \phi_p(\mathbf{y}; \boldsymbol{\mu}, \nu_2^{-1} \boldsymbol{\Sigma}) + (1 - \nu_1) \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Da Proposição 3.1.5 seguem os seguintes resultados para os momentos condicionais  $u_r$  and  $\tau_r$ .

**Corolário 3.2.3.** *Seja  $\mathbf{Y} \sim SCN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu_1, \nu_2)$ . Então,*

$$\begin{aligned} \kappa_r &= \frac{2}{\psi(\mathbf{y})} \left\{ \nu_1 \nu_2^r \phi_p(\mathbf{y}; \boldsymbol{\mu}, \nu_2^{-1} \boldsymbol{\Sigma}) \Phi(\nu_2^{1/2} A) + (1 - \nu_1) \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi(A) \right\}, \\ \tau_r &= \frac{2}{\psi(\mathbf{y})} \left\{ \nu_1 \nu_2^{r/2} \phi_p(\mathbf{y}; \boldsymbol{\mu}, \nu_2^{-1} \boldsymbol{\Sigma}) \phi_1(\nu_2^{1/2} A) + (1 - \nu_1) \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \phi_1(A) \right\}. \end{aligned}$$



### 3.3 Inferência pelo Método da Máxima Verossimilhança

Essa é uma das seções mais importantes para o desenvolvimento desse trabalho. Aqui será apresentada a representação hierárquica dos modelos na classe  $MESN$ , e a partir desta pode-se tratar o problema de estimação dos parâmetros via o algoritmo EM, considerando a abordagem por dados aumentados. Os Resultados aqui apresentados serão utilizados principalmente na Seção 5.1. Uma vez compreendidos nesta seção, torna-se mais fácil o entendimento do processo de estimação dos parâmetros em modelos de misturas finitas de distribuições da classe MESN.

#### 3.3.1 Representação hierárquica

Utilizando a representação estocástica 3.1.3, o modelo  $MESN$  pode ser apresentado sob um ponto de vista com dados incompletos.

**Proposição 3.3.1.** *Considere a amostra  $\mathbf{Y}_i \sim MESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$  com  $i = 1, \dots, n$ . Então o modelo hierárquico para cada vetor aleatório MESN dessa amostra é dado por*

$$\mathbf{Y}_i | u_i, t_i \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Delta} t_i, \kappa(u_i) \boldsymbol{\Gamma}), \quad (3.3.1)$$

$$T_i | u_i \sim HN(0, \kappa(u_i)), \quad (3.3.2)$$

$$U_i \sim H(u_i; \boldsymbol{\nu}), \quad (3.3.3)$$

lembrando 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 variancia  $\kappa(u_i)$  (antes de truncar em  $(0, \infty)$ ). Além disso, considerando  $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$ ,  $\mathbf{u} = (u_1, \dots, u_n)^\top$ ,  $\mathbf{t} = (t_1, \dots, t_n)^\top$ , a função de log-verossimilhança completa de  $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ , com  $\boldsymbol{\alpha}$  denotando o vetor com os elementos da matriz

triangular superior  $\Sigma$ , é dada por

$$\begin{aligned}
 \ell_c(\boldsymbol{\theta}) &= C - \frac{n}{2} \log |\Gamma| - \frac{1}{2} \sum_{i=1}^n \kappa^{-1}(u_i) (\mathbf{y}_i - \boldsymbol{\mu} - \Delta t_i)^\top \Gamma^{-1} (\mathbf{y}_i - \boldsymbol{\mu} - \Delta t_i) + \sum_{i=1}^n \log(h(u_i; \boldsymbol{\nu})) \\
 &= C - \frac{n}{2} \log |\Gamma| - \frac{1}{2} \sum_{i=1}^n [\kappa^{-1}(u_i) (\mathbf{y}_i - \boldsymbol{\mu})^\top \Gamma^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \\
 &\quad - 2\kappa^{-1}(u_i) t_i \Delta^\top \Gamma^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) + \kappa^{-1}(u_i) t_i^2 \Delta^\top \Gamma^{-1} \Delta] + \sum_{i=1}^n \log(h(u_i; \boldsymbol{\nu})), \tag{3.3.4}
 \end{aligned}$$

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

### 3.3.2 O algoritmo EM em modelos MESN

A estimação dos parâmetros do modelo pelo método da máxima verossimilhança, pode ser realizada via algoritmo EM (Dempester, Laird e Rubin - 1977). O algoritmo é aplicado ao problema de estimação a partir de dados incompletos, aumentando o vetor de dados observados com a inclusão de variáveis latentes, não observáveis diretamente, de modo que a verossimilhança do vetor de dados completos simplifique as análises a serem envolvidas. O algoritmo procede iterativamente em duas etapas, E (do inglês, *Expectation*) e M (do inglês, *Maximization*). A inclusão desses dados não observáveis ao problema é tratada na etapa E, a qual consiste em tomar a esperança da log-verossimilhança completa condicional ao vetor de dados observados. Em seguida, no passo M, é realizada a maximização da log-verossimilhança completa em relação aos parâmetros do modelo, substituindo os dados latentes por seus valores esperados condicionais obtidos na etapa E.

Para a estimação dos parâmetros na classe *MESN*, pode-se obter a abordagem de dados incompletos através da representação hierárquica do modelo. O vetor de dados observados é dado por  $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$ , e os dados aumentados por  $\mathbf{u} = (u_1, \dots, u_n)^\top$ ,  $\mathbf{t} = (t_1, \dots, t_n)^\top$ . Portanto, na etapa E do algoritmo, toma-se o valor esperado da log-verossimilhança completa 3.3.4 condicional à  $\mathbf{y}$  e a  $\boldsymbol{\theta}$  no seu estado corrente. Note que as

seguintes quantidades deve ser obtidas

$$\begin{aligned}\widehat{\kappa}_i &= E\{\kappa^{-1}(U_i)|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\}, \\ \widehat{s}_{2i} &= E\{\kappa^{-1}(U_i)T_i|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\}, \\ \widehat{s}_{3i} &= E\{\kappa^{-1}(U_i)T_i^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i\}.\end{aligned}$$

A primeira dessas quantidades pode ser obtida utilizando a Proposição 3.1.5. Para as demais quantidades, deve ser obter a distribuição condicional  $T_i|\mathbf{y}_i, u_i$ . Das equações (3.3.1) e (3.3.2) da representação hierárquica do modelo *MESN* mais o Lema A.2 tem-se que

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

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  $\mu_{T_i} = M_{T_i}^2 \boldsymbol{\Delta}^\top \boldsymbol{\Gamma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$ .

Agora, definindo  $\widehat{\kappa}_i$  e  $\widehat{\tau}_i$  como na Proposição 3.1.5, e usando propriedades de esperança condicional, obtêm-se as seguintes expressões

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

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

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

Assim, o valor esperado condicional da log-verossimilhança completa é

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

A etapa M do algoritmo envolve a maximização em  $\boldsymbol{\theta}$  da esperança condicional da log-verossimilhança completa acima. Assim como será mostrado posteriormente, obter as quantidades na etapa E para as distribuições *skew-t* e *skew-normal* contaminada não é uma tarefa árdua, já que das proposições 3.2.1 e 3.2.3 têm-se expressões fechadas, matematicamente atrativas e fáceis de serem implementadas. Entretanto esse não é o caso para a distribuição *skew-slash*. Neste caso, a expressão a ser maximizada envolve integrais complexas, sendo assim necessária a utilização de métodos alternativos para computá-los. Por exemplo, a integração Monte Carlo pode ser empregada na etapa M do algoritmo, o que resulta no chamado algoritmo MCEM. Quando se tem expressões intratáveis na etapa M do algoritmo, é comum utilizar uma sequência de passos de maximizações condicionais (CM - *conditional maximization*). Essa modificação resulta no algoritmo ECM (Meng e Rubin, 1993). O algoritmo ECME (Liu e Rubin, 1994), uma extensão do EM e do ECM computacionalmente mais rápida, é obtida por maximizar a esperança da função de verossimilhança completa restrita, com alguns passos CM que maximizam a função de verossimilhança marginal restrita, chamados de passos CML. Portanto, o algoritmo ECME fica dado por:

**Etapa E:** Dado  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$ , obter  $\hat{\kappa}_i^{(k)}$ ,  $\hat{s}_{2i}^{(k)}$  e  $\hat{s}_{3i}^{(k)}$  para  $i = 1, \dots, n$ .

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

$$\hat{\boldsymbol{\mu}}^{(k+1)} = \sum_{i=1}^n (\hat{\kappa}_i^{(k)} \mathbf{y}_i - \hat{s}_{2i}^{(k)} \hat{\boldsymbol{\Delta}}^{(k)}) / (\sum_{i=1}^n \hat{\kappa}_i^{(k)}), \quad (3.3.5)$$

$$\begin{aligned} \hat{\mathbf{\Gamma}}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^n \left[ \hat{\kappa}_i^{(k)} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}^{(k)}) (\mathbf{y}_i - \hat{\boldsymbol{\mu}}^{(k)})^\top - \hat{s}_{2i}^{(k)} \hat{\boldsymbol{\Delta}}^{(k)} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}^{(k)})^\top \right. \\ &\quad \left. - \hat{s}_{2i}^{(k)} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}^{(k)}) \hat{\boldsymbol{\Delta}}^{(k)\top} + \hat{s}_{3i}^{(k)} \hat{\boldsymbol{\Delta}}^{(k)} \hat{\boldsymbol{\Delta}}^{(k)\top} \right], \end{aligned} \quad (3.3.6)$$

$$\hat{\boldsymbol{\Delta}}^{(k+1)} = \sum_{i=1}^n \hat{s}_{2i}^{(k)} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}^{(k)}) / \sum_{i=1}^n \hat{s}_{3i}^{(k)},$$

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

$$\boldsymbol{\nu}^{(k+1)} = \arg \max_{\boldsymbol{\nu}} \sum_{i=1}^n \log(\psi(y_i; \boldsymbol{\mu}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\nu})), \quad (3.3.7)$$

com  $\psi(\mathbf{y}; \boldsymbol{\theta})$  dado em (3.1.1).

As iterações são repetidas até que alguma regra de convergência adequada seja satisfeita, por exemplo, se  $\|\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Theta}^{(k)}\|$  for suficientemente pequeno ou alguma distância envolvendo a log-verossimilhança, como  $\|\ell(\boldsymbol{\Theta}^{(k+1)}) - \ell(\boldsymbol{\Theta}^{(k)})\|$ . Veja Dias e Wedel (2004) no contexto de misturas de normais univariadas.

Os valores iniciais para o algoritmo podem ser obtidos utilizando o seguinte esquema: Para o caso *skew-normal*, utiliza-se como valores iniciais para o vetor de médias e matriz de covariâncias, o vetor de média amostral e matriz de covariâncias amostral, respectivamente. Para a  $j$ -ésima cordenada do vetor de assimetria, considere  $\hat{\rho}_j$  a assimetria amostral para a variável  $j$ . Então,  $\lambda_j^{(0)} = 3 \times \text{sign}(\hat{\rho}_j)$ . As estimativas EM encontradas para esse caso são então passadas como valores iniciais para o algoritmo considerando os outros elementos da classe MESN. Na pratica, recomenda-se a utilização de diferentes valores iniciais para o algoritmo EM. Isto por que se existir mais de um máximo, pode-se determinar o global comparando seus valores de verossimilhança e comprovar a estabilidade da estimativa obtida.

## 4 *Mistura Finita de Densidades*

Por ser um método extremamente flexível de modelagem, os modelos de misturas tem recebido muita atenção nos últimos anos para aquelas situações onde há presença de heterogeneidade populacional. Aplicações desses modelos podem ser encontradas em várias áreas da estatística, tais como análise de agrupamento, análise discriminante, análise de sobrevivência, métodos não-paramétricos ou semi-paramétricos e até em processamento de imagens.

Muitos trabalhos podem ser encontrados na literatura utilizando modelos de misturas para aproximar densidades complexas, desde aquelas com aspectos multimodais à outras totalmente assimétricas, sendo preferíveis em situações em que uma única família paramétrica de distribuições não produz uma modelagem satisfatória.

Neste capítulo, será introduzido de uma forma geral o modelo de misturas de distribuições. Algumas propriedades desses modelos serão aqui discutidas. Em seguida, o modelo de misturas será tratado sob o ponto de vista de dados incompletos utilizando na estimação de máxima verossimilhança via algoritmo EM. Uma aproximação da matriz de informação observada é em seguida derivada. Por fim, alguns métodos de seleção de modelos são apresentados e também como aplicar modelos de misturas em problemas de classificações de observações.

## 4.1 Misturas finitas de densidades

### 4.1.1 Definição

**Definição 4.1.1.** Um vetor aleatório  $\mathbf{Y} \in \mathbb{R}^p$  com função de densidade dada por

$$f(\mathbf{y}) = \sum_{j=1}^g p_j \psi_j(\mathbf{y}), \quad p_j \geq 0 \quad e \quad \sum_{j=1}^g p_j = 1, \quad (4.1.1)$$

é dito ter uma distribuição de mistura de densidades. A função  $f(\cdot)$  é denominada de mistura finita de densidades com  $g$  componentes, com os parâmetros  $p_1, \dots, p_g$  denominados proporções de misturas e as densidades  $\psi_1, \dots, \psi_g$  as componentes de misturas.

Quando as componentes  $\psi_j(\cdot)$  pertencem à famílias paramétricas de distribuições, pode-se reescrever o modelo (4.1.1) por

$$f(\mathbf{y}; \Theta) = \sum_{j=1}^g p_j \psi_j(\mathbf{y}; \theta_j), \quad (4.1.2)$$

com  $\Theta = (\theta_1^\top, \dots, \theta_g^\top)$  e  $\theta_j$  os parâmetros que definem cada uma das componentes  $\psi_j$ , não necessariamente definidos no mesmo espaço paramétrico.

Na maioria das aplicações encontradas na literatura e as que são aqui apresentadas, as componentes de mistura  $\psi_j$  pertencem a mesma família paramétrica de distribuições e assim, a mistura finita de densidades será denotada por

$$f(\mathbf{y}; \Theta) = \sum_{j=1}^g p_j \psi(\mathbf{y}; \theta_j), \quad \mathbf{y} \in \mathbb{R}^p. \quad (4.1.3)$$

É interessante ressaltar que sob essas suposições os parâmetros  $\theta_j$  agora pertencem a um mesmo espaço paramétrico.

### 4.1.2 Distribuição marginal

O seguinte teorema mostra que as distribuições marginais em misturas finitas de densidades (com um dado número de componentes) são também misturas finitas (de

mesmo número de componentes).

**Teorema 4.1.1.** *Seja  $\mathbf{Y}$  um vetor aleatório em  $\mathbb{R}^p$  cuja distribuição é uma mistura finita de densidades com  $g$  componentes. Então, a distribuição de qualquer vetor aleatório formado com elementos de  $\mathbf{Y}$  em  $\mathbb{R}^q$ ,  $q < p$ , é também uma mistura finita de densidades com  $g$  componentes.*

*Demonstração.* Seja  $\mathbf{X}$  um vetor aleatório cujas variáveis compoentes formam um subconjunto das variáveis de  $\mathbf{Y}$ . Então, a densidade de  $\mathbf{X}$  é dada por

$$f_{\mathbf{X}}(\mathbf{x}) = \int f(\mathbf{y}) d\mathbf{y}_{(\mathbf{x})},$$

sendo que notação acima representa a integral tomada naqueles compoentes de  $\mathbf{Y}$  que não estão em  $\mathbf{X}$ . Desta forma

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \int \sum_{j=1}^g p_j \psi_j(\mathbf{y}) d\mathbf{y}_{(\mathbf{x})} \\ &= \sum_{j=1}^g p_j \int \psi_j(\mathbf{y}) d\mathbf{y}_{(\mathbf{x})}, \end{aligned}$$

e o resultado está demonstrado já que  $\int \psi_j(\mathbf{y}) d\mathbf{y}_{(\mathbf{x})}$  é densidade.  $\square$

### 4.1.3 Identificabilidade

A condição de indentificabilidade dos parâmetros em um modelo é de considerável importância na estatística por garantir que esses possam ser estimados de maneira única. Em geral, uma família paramétrica  $\mathcal{F}$  de fdp's  $\psi(\cdot; \boldsymbol{\theta})$  é dita ser identificável se valores distintos de  $\boldsymbol{\theta}$  determinam membros distintos da família. Entretanto, para famílias de misturas finitas de densidades é necessária uma definição mais específica. Como exemplo, Duda e Hart (1973) ilustram essa questão com uma mistura de duas densidades normais de variância unitária,

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



Na mistura acima, com os valores dos parâmetros fixados, se permutarmos os índices em  $\Theta = (\theta_1, \theta_2)$ ,  $\theta_j = (p_j, \mu_j)$ ,  $j = 1, \dots, 2$ , a densidade  $f(\cdot; \Theta)$  terá o mesmo valor em cada  $y \in \mathbb{R}$ . Desta forma, a identificabilidade no contexto de misturas de distribuições deve ter ainda mais uma condição: a de permutabilidade, como a seguir.

**Definição 4.1.2.** *Seja  $\mathcal{F} = \{\psi(\mathbf{y}; \theta) : \mathbf{y} \in \mathbb{R}^p\}$  uma família paramétrica de densidades e*

$$\mathcal{P} = \left\{ f(\mathbf{y}; \Theta) : f(\mathbf{y}; \Theta) = \sum_{j=1}^g p_j \psi(\mathbf{y}; \theta_j), \quad p_j \geq 0, \quad \sum_{j=1}^g p_j = 1, \right. \\ \left. \psi(\mathbf{y}; \theta) \in \mathcal{F}, \quad \Theta = (\theta_1, \dots, \theta_g) \right\}$$

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

$$f(\mathbf{y}; \Theta) = \sum_{j=1}^g p_j \psi(\mathbf{y}; \theta_j) \quad e \quad f(\mathbf{y}; \Theta') = \sum_{j=1}^{g'} p'_j \psi(\mathbf{y}; \theta'_j),$$

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

Em Titterton et al. (1985), são discutidas as questões teóricas sobre as condições para uma mistura finita de densidades ser identificável, e ainda comentam que a maioria das misturas de distribuições contínuas são identificáveis. Algumas dificuldades, entretanto, aparecem quando as componentes de uma mistura pertencem à mesma família de distribuições, fato apresentado por McLachlan e Basford (1988, seção 1.5). Nesse caso, o valor da densidade de mistura terá o mesmo valor se os índices  $j$  forem permutados em  $\Theta = (\theta_1, \dots, \theta_g)$ . Embora a mistura seja identificável, o vetor  $\Theta$  não é. De fato, se cada componente for da mesma família de distribuições, então a densidade de mistura será invariante para as  $g!$  permutações dos índices em  $\Theta$ .

## 4.2 A Estrutura de Dados Incompletos para o Problema de Misturas

Para introduzir a abordagem de dados incompletos para o problema de misturas, por ora será tratada a questão de como gerar vetores pseudo aleatórios de uma mistura de densidades. Considere então o problema proposto. Uma maneira de se gerar um vetor aleatório  $\mathbf{Y}_i$  de uma densidade  $f(\mathbf{y}_i)$  dada em 4.1.1, segue como: Considere  $Z_i$  uma variável aleatória categórica tomando os valores  $1, \dots, g$  com probabilidades  $p_1, \dots, p_g$ , respectivamente, e suponha que a densidade de  $\mathbf{Y}_i$  condicional a  $Z_i = j$  é  $\psi_j(\mathbf{y}_i)$ , ( $j = 1, \dots, g$ ). Então, a densidade marginal de  $\mathbf{Y}_i$  é dada por 4.1.1. Nesse contexto, a variável  $Z_i$  pode ser interpretada como uma variável de latente, indicando a componente do qual o vetor  $\mathbf{Y}_i$  é proveniente. A título de simplificar notações, algebra e futuras interpretações, é conveniente lidar com 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,  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ig})$  e

$$Z_{ij} = \begin{cases} 1, & \text{Se } \mathbf{Y}_i \text{ pertence ao grupo } j \\ 0, & \text{caso contrário} \end{cases} \quad (4.2.1)$$

Consequentemente, sob essa abordagem a variável  $\mathbf{Z}_i$  tem distribuição multinomial considerando uma retirada em  $g$  categorias, com probabilidades  $p_1, \dots, p_g$ , isto é

$$P(\mathbf{Z}_i = \mathbf{z}_i) = p_1^{z_{i1}} p_2^{z_{i2}} \dots p_g^{z_{ig}}, \quad \text{ou} \quad \mathbf{Z}_i \sim \text{Multi}_g(1, p_1, \dots, p_g).$$

Desta forma, foi visto como tratar o problema de se gerar vetores aleatórios de uma distribuição de misturas de densidades com a inclusão do vetor  $\mathbf{Z}_i$ , o qual regula as proporções de mistura das componentes.

Neste sentido, uma situação obvia onde o modelo de misturas de distribuições é diretamente aplicável é quando  $\mathbf{Y}_i$  é suposto pertencer a uma população composta de  $g$  grupos em proporções  $p_1, \dots, p_g$ . Note que agora só é conhecido o vetor  $\mathbf{Y}_i = \mathbf{y}_i$ , para cada  $i = 1, \dots, n$  pertencente a amostra, e não o vetor  $\mathbf{Z}_i$  associado a essa observação. O conceito de existir esse vetor latente associando a observação e a componente a qual

ela pertence é muito útil, embora não pareça ser intuitivo num sentido físico. Ao decorrer desse trabalho, será visto que esse conceito é o que permite a estimação de máxima verossimilhança através do algoritmo EM.

Considere então  $\mathbf{y}_1, \dots, \mathbf{y}_n$  as  $n$  realizações dos vetores aleatórios independentes e identicamente distribuídos (iid)  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  com densidade comum  $f(\mathbf{y}_i)$  dada por 4.1.1. Então

$$\mathbf{Y}_1, \dots, \mathbf{Y}_n \stackrel{iid}{\sim} F,$$

com  $F(\cdot)$  a f.d.a correspondente a densidade de mistura  $f(\cdot)$ . Sob a perspectiva do algoritmo EM, os dados  $\mathbf{y}_1, \dots, \mathbf{y}_n$  são vistos como incompletos já que os vetores  $\mathbf{z}_1, \dots, \mathbf{z}_n$  indicadores de componentes não são observáveis. Desta forma, o vetor de dados completos é definido por

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

com  $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$  e  $\mathbf{z} = (\mathbf{z}_1^\top, \dots, \mathbf{z}_n^\top)^\top$ . Os vetores  $\mathbf{z}_1, \dots, \mathbf{z}_n$  são realizações dos vetores aleatórios  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ , os quais supostos independentes do vetor de observações, distribuídos por

$$\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{iid}{\sim} \text{Multi}_g(1, p_1, \dots, p_g).$$

A  $j$ -ésima proporção de mistura pode ser vista como a probabilidade *a priori* 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

$$\begin{aligned} \hat{z}_{ij} &= P(\text{entidade} \in j - \text{ésima componente} | \mathbf{y}_i) \\ &= P(Z_{ij} = 1 | \mathbf{y}_i) \\ &= \frac{p_j \psi_j(\mathbf{y}_i)}{f(\mathbf{y}_i)}, \end{aligned} \tag{4.2.2}$$

de onde a ultima igualdade vêm do teorema de bayes. Considerando então a independencia entre  $\mathbf{Z}_i$  e  $\mathbf{Y}_i$ ,  $i = 1, \dots, n$ , pode-se derivar a verossimilhança completa dos dados,

$$L_c(\Theta) = \prod_{i=1}^n \prod_{j=1}^g [p_j \psi_j(\mathbf{y}_i; \theta_j)]^{z_{ij}},$$

e portanto, a log-verossimilhança completa fica dada por

$$l_c(\Theta) = \sum_{i=1}^n \sum_{j=1}^g z_{ij} [\log p_j + \log \psi(\mathbf{y}_i; \theta_j)]. \quad (4.2.3)$$

### 4.3 O algoritmo EM em Modelos de Misturas

**Etapa E:** A inclusão dos dados não observáveis ao problema (aqui a variável  $z_{ij}$ ) é tratada na etapa E do algoritmo tomando o valor esperado da log-verossimilhança completa  $l_c(\Theta)$  condicional aos dados observados  $\mathbf{y}$ . Seja então  $\Theta^{(k)}$  o valor estimado de  $\Theta$  na  $(k)$ -ésima iteração do algoritmo. Então, na etapa E da  $(k+1)$ -ésima iteração do algoritmo, deve-se obter

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

Como a log-verossimilhança completa dos dados é linear na variável não observada  $z_{ij}$ , a etapa E se resume simplesmente em obter 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,

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

isto é,

$$\begin{aligned} \hat{z}_{ij} &= \frac{p_j^{(k)} \psi(\mathbf{y}_i; \theta_j^{(k)})}{f(\mathbf{y}_i; \Theta^{(k)})} \\ &= \frac{p_j^{(k)} \psi(\mathbf{y}_i; \theta_j^{(k)})}{\sum_{j=1}^g p_j^{(k)} \psi(\mathbf{y}_i; \theta_j^{(k)})}, \end{aligned} \quad (4.3.1)$$

para  $i = 1, \dots, n$  e  $j = 1, \dots, g$ . Portanto, usando (4.3.1), a esperança de (4.2.3) condicional a  $\mathbf{y}$  fica dada por

$$Q(\Theta; \Theta^{(k)}) = \sum_{i=1}^n \sum_{j=1}^g \hat{z}_{ij} [\log p_j^{(k)} + \log \psi(\mathbf{y}_i; \theta_j^{(k)})]. \quad (4.3.2)$$

Cabe ressaltar nesse ponto, que o algoritmo EM aqui derivado somente considera como dados aumentados a variável  $z_{ij}$ . No capítulo seguinte será derivado o algoritmo EM para

misturas de densidades MESN, e nesse caso, o vetor de dados aumentados não é composto somente pela variável de rótulo, o que modifica substancialmente a etapa E aqui proposta.

**Etapa M:** A etapa M do algoritmo na  $(k + 1)$ -ésima iteração requer a maximização de  $Q(\Theta; \Theta^{(k)})$  com respeito à  $\Theta$ , para resultar na estimativa atualizada  $\Theta^{(k+1)}$ . No contexto de misturas, a estimativa de  $p_j^{(k)}$  das proporções de mistura  $p_j$  são calculadas independentemente dos demais parâmetros do modelo. Se  $z_{ij}$  fosse observável, então o estimador de máxima verossimilhança de  $p_j$ , considerando os dados completos, seria dado por

$$\hat{p}_j = \frac{1}{n} \sum_{j=1}^g z_{ij} \quad (j = 1, \dots, g).$$

Como na etapa E simplesmente substituiu-se  $z_{ij}$  pela sua correspondente esperança condicional  $\hat{z}_{ij}$  na log-verossimilhança completa, então a estimativa atualizada de  $p_j$  é dada por substituir  $z_{ij}$  na equação acima

$$p_j^{(k+1)} = \frac{1}{n} \sum_{j=1}^g \hat{z}_{ij} \quad (j = 1, \dots, g).$$

Assim, na estimativa de  $p_j$  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. A estimativa dos demais parâmetros do modelo também são obtidas a partir da maximização de (4.3.2) em relação aos mesmos, e será derivada no capítulo seguinte para o caso de misturas de densidades MESN.

## 4.4 Matriz de Informação Observada

Uma grande variedade de métodos tem sido proposta por diferentes autores para se obter a matriz de covariâncias dos estimadores de máxima verossimilhança  $\hat{\Theta}$  do vetor de parâmetro  $\Theta$ , obtidos a partir do algoritmo EM. A maioria desses métodos são baseados na matriz de informação de Fisher. Para o caso de dados independentes e identicamente distribuídos, essa matriz pode ser aproximada sem muito trabalho além daquele utilizado para se obter as estimativas propriamente ditas.

Considere uma amostra aleatória  $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$  e sua log-verossimilhança  $\log L(\boldsymbol{\Theta})$ , expressada na seguinte forma

$$\log L(\boldsymbol{\Theta}) = \sum_{i=1}^n \log L_i(\boldsymbol{\Theta}),$$

tal que  $L_i(\boldsymbol{\Theta}) = f(\mathbf{y}_i; \boldsymbol{\Theta})$  é a verossimilhança formada a partir de uma única observação  $\mathbf{y}_i$  ( $i = 1, \dots, n$ ). Pode-se denotar o vetor de escore  $\mathbf{S}(\mathbf{y}; \boldsymbol{\Theta})$  como

$$\mathbf{S}(\mathbf{y}; \boldsymbol{\Theta}) = \sum_{i=1}^n \mathbf{s}(\mathbf{y}_i; \boldsymbol{\Theta}),$$

tal que  $\mathbf{s}(\mathbf{y}_i; \boldsymbol{\Theta}) = \partial \log L_i(\boldsymbol{\Theta}) / \partial \boldsymbol{\Theta}$ . A matriz de informação esperada  $\mathcal{I}(\boldsymbol{\Theta})$  é dada por

$$\mathcal{I}(\boldsymbol{\Theta}) = n\mathbf{i}(\boldsymbol{\Theta}), \quad (4.4.1)$$

com

$$\begin{aligned} \mathbf{i}(\boldsymbol{\Theta}) &= E[\mathbf{s}(\mathbf{Y}_i; \boldsymbol{\Theta}) \mathbf{s}^\top(\mathbf{Y}_i; \boldsymbol{\Theta})] \\ &= \text{cov}[\mathbf{s}(\mathbf{Y}_i; \boldsymbol{\Theta})], \end{aligned} \quad (4.4.2)$$

a informação contida em uma única observação. Correspondente a (4.4.2), a matriz de informação empírica (em uma única observação) pode ser definida como

$$\begin{aligned} \bar{\mathbf{i}}(\boldsymbol{\Theta}) &= n^{-1} \sum_{i=1}^n \mathbf{s}(\mathbf{y}_i; \boldsymbol{\Theta}) \mathbf{s}^\top(\mathbf{y}_i; \boldsymbol{\Theta}) - \bar{\mathbf{s}} \bar{\mathbf{s}}^\top \\ &= n^{-1} \sum_{i=1}^n \mathbf{s}(\mathbf{y}_i; \boldsymbol{\Theta}) \mathbf{s}^\top(\mathbf{y}_i; \boldsymbol{\Theta}) \\ &\quad - n^{-2} \mathbf{S}(\mathbf{y}; \boldsymbol{\Theta}) \mathbf{S}^\top(\mathbf{y}; \boldsymbol{\Theta}), \end{aligned} \quad (4.4.3)$$

de onde a ultima igualdade se da por notar que  $\bar{\mathbf{s}} = n^{-1} \sum_{i=1}^n \mathbf{s}(\mathbf{y}_i; \boldsymbol{\Theta})$ .

Desta forma, correspondente a essa forma empírica (4.4.3) para  $\mathbf{i}(\boldsymbol{\Theta})$ ,  $\mathcal{I}(\boldsymbol{\Theta})$  é estimado

por

$$\begin{aligned}
 I_e(\boldsymbol{\Theta}) &= n \bar{\mathbf{i}}(\boldsymbol{\Theta}) \\
 &= \sum_{i=1}^n \mathbf{s}(\mathbf{y}_i; \boldsymbol{\Theta}) \mathbf{s}^\top(\mathbf{y}_i; \boldsymbol{\Theta}) \\
 &\quad - n^{-1} \mathbf{S}(\mathbf{y}; \boldsymbol{\Theta}) \mathbf{S}^\top(\mathbf{y}; \boldsymbol{\Theta}).
 \end{aligned} \tag{4.4.4}$$

No ponto  $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}$ , a matriz de informação empírica se reduz a

$$I_e(\boldsymbol{\Theta}) = \sum_{i=1}^n \mathbf{s}(\mathbf{y}_i; \hat{\boldsymbol{\Theta}}) \mathbf{s}^\top(\mathbf{y}_i; \hat{\boldsymbol{\Theta}}). \tag{4.4.5}$$

Na prática, essa matriz de informação empírica é muito utilizada como uma aproximação à matriz de informação observada  $\mathcal{I}(\boldsymbol{\Theta})$ . Uma discussão mais detalhada a respeito da matriz de informação empírica pode ser encontrada em McLachlan e Krishnan (2008) e Basford et al. (1997).

## 4.5 Métodos de Seleção de Modelos

Os critérios de seleção de modelos propostos nessa seção podem ser utilizados em vários aspectos, desde a comparação entre modelos, como também na determinação do número  $g$  de componentes no contexto de mistura. Cabe ressaltar que esses critérios não podem ser utilizados como regras de decisões, mas sim como uma ferramenta que dá evidências de qual modelo pode ser preferível em detrimento de outros. Para uma discussão mais detalhada a respeito desses métodos e outros, ver McLachlan e Peel (2000, seção 6.8).

### 4.5.1 Critério de informação de *Akaike* - AIC

O problema de seleção de modelos pode ser visto em termos da informação de Kullback-Leiber (1951) do verdadeiro modelo com respeito ao modelo estimado. Se  $f(\mathbf{x}|\boldsymbol{\Theta}^*)$  denota a densidade do modelo verdadeiro e  $f(\mathbf{x}|\hat{\boldsymbol{\Theta}})$  denota a densidade do modelo estimado, a

informação de Kullback-Leiber é dada por

$$\begin{aligned} I_{KL}(\Theta^*, \hat{\Theta}) &= \int f(\mathbf{x}|\Theta^*) \log \left( \frac{f(\mathbf{x}|\Theta^*)}{f(\mathbf{x}|\hat{\Theta})} \right) d\mathbf{x} \\ &= \int f(\mathbf{x}|\Theta^*) \log f(\mathbf{x}|\Theta^*) d\mathbf{x} - \int f(\mathbf{x}|\Theta^*) \log f(\mathbf{x}|\hat{\Theta}) d\mathbf{x}, \end{aligned}$$

que é uma medida de divergência entre o modelo verdadeiro e o modelo estimado, sendo o objetivo, portanto, minimizar essa divergência. Como o primeiro termo na expressão acima não depende do modelo estimado, vemos que somente o segundo termo é relevante à minimização. Denominando de log-verossimilhança esperada, temos

$$\begin{aligned} \eta(\mathbf{y}; F) &= \int f(\mathbf{x}|\Theta^*) \log f(\mathbf{x}|\hat{\Theta}) d\mathbf{x} \\ &= \int \log f(\mathbf{x}|\hat{\Theta}) dF(\mathbf{x}), \end{aligned} \tag{4.5.1}$$

sendo  $F$  a verdadeira função de distribuição e  $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)$  os dados observados. Um estimador de  $\eta(\mathbf{y}; F)$  é dado por

$$\eta(\mathbf{y}; \hat{F}_n) = \frac{1}{n} \sum_{i=1}^n \log f(\mathbf{y}_i|\hat{\Theta}),$$

obtido por substituir  $F$  em (4.5.1) pela sua função de distribuição empírica, a qual atribui massa  $1/n$  em cada observação  $\mathbf{y}_i$ , ( $i = 1, \dots, n$ ). O que ocorre em geral é que isso produz uma superestimativa da log-verossimilhança esperada, já que a função de distribuição empírica é geralmente mais próxima à distribuição estimada do que da verdadeira distribuição desconhecida. Portanto, define-se então o viés do estimador  $\eta(\mathbf{y}; \hat{F}_n)$  como o funcional

$$\begin{aligned} b(F) &= E_F[\eta(\mathbf{Y}; \hat{F}_n) - \eta(\mathbf{Y}; F)] \\ &= E_F\left[\frac{1}{n} \sum_{i=1}^n \log f(\mathbf{y}_i|\hat{\Theta}) - \int \log f(\mathbf{x}|\hat{\Theta}) dF(\mathbf{x})\right]. \end{aligned} \tag{4.5.2}$$

Um critério de informação para seleção de modelos pode ser construído baseado na log-verossimilhança com essa correção de viés, e é dado por

$$\log L(\hat{\Theta}) - b(F),$$



utilizando uma estimativa apropriada do termo  $b(F)$ . A ideia é escolher o modelo mais adequado segundo aquele que maximize a relação acima. Na literatura, entretanto, o critério de informação é geralmente formado pelo dobro do negativo dessa diferença, ou seja

$$-2\log L(\hat{\Theta}) + C,$$

sendo que o primeiro termo dessa relação mensura a falta de ajuste do modelo e o segundo termo  $C$  é uma penalização que mensura a complexidade deste. Desta forma, o objetivo agora é escolher o modelo que minimize esse critério.

Akaike (1973, 1974) mostrou que  $b(F)$  é assintoticamente igual a  $d$ , que representa o número total de parâmetros do modelo. Desta forma, o critério de informação de Akaike seleciona o modelo que minimiza

$$AIC = -2\log L(\hat{\Theta}) + 2d.$$

Algumas considerações devem ser feitas a respeito do AIC. A validade para essas aproximações de  $b(F)$  depende fortemente das mesmas condições de regularidades conhecidas da teoria assintótica para a distribuição da estatística de razão de verossimilhança. Como se sabe, essas condições não são válidas para testes quanto ao número de componentes em um modelo de misturas. Apesar disso, esse critério tem sido muito utilizado na prática para determinar a ordem de uma mistura. Muitos autores (por exemplo, Koehler e Murphee (1998) e Celeux e Soromenho (1996)) comentam que o AIC é inconsistente em ordem, e tende a superestimar a dimensão do modelo, e no contexto de misturas, isso significa uma tendência em selecionar modelos com um número de componentes superior ao verdadeiro.

#### 4.5.2 Critério de informação bayesiano - BIC

O critério de informação Bayesiano está baseado na teoria Bayesiana de seleção de modelos. A ideia básica desse método é considerar alguns possíveis modelos, com suas probabilidades *a priori*, e selecionar aquele que apresenta a maior probabilidade *a posteriori*, dadas as observações. Sendo  $M_1, M_2, \dots, M_K$  os modelos considerados e  $p(M_k)$ , ( $k = 1, \dots, K$ ), as respectivas probabilidades *a priori*, pelo teorema de Bayes, a

probabilidade *a posteriori* de  $M_k$  dado as observações é

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k)p(M_k)}{\sum_{t=1}^K p(\mathbf{y}|M_t)p(M_t)}. \quad (4.5.3)$$

Da expressão acima, vemos que é necessário obter  $p(\mathbf{y}|M_k)$ , conhecida como *verossimilhança integrada*. Agora, se as probabilidades *a priori* forem todas iguais, então o procedimento seleciona o modelo com a maior verossimilhança integrada. Quando existem parâmetros nos modelos, essa verossimilhança é obtida por integração no espaço de parâmetros, ou seja, considerando um modelo  $M$

$$\begin{aligned} p(\mathbf{y}|M) &= \int p(\mathbf{y}|M, \Theta)p(\Theta|M) d\Theta \\ &= \int \exp[\log(p(\Theta; \mathbf{y}))] d\Theta, \end{aligned} \quad (4.5.4)$$

com  $p(\Theta; \mathbf{y}) = p(\mathbf{y}|M, \Theta)p(\Theta|M)$ . Note que  $p(\mathbf{y}|M, \Theta)$  representa a verossimilhança do modelo  $M$  e  $p(\Theta|M)$  a probabilidade *a priori* de  $\Theta$ . A ideia agora é encontrar uma aproximação para 4.5.4. Seja, então,  $\tilde{\Theta}$  uma moda *a posteriori* satisfazendo  $\partial \log p(\tilde{\Theta}; \mathbf{y})/\partial \Theta = 0$ . Seja  $H(\tilde{\Theta})$  a matriz hessiana de  $\log p(\Theta; \mathbf{y})$  calculada em  $\tilde{\Theta}$ . Para aproximar a integral 4.5.4, o integrando é expandido em uma série de Taylor até segunda ordem em torno do ponto  $\Theta = \tilde{\Theta}$  resultando em

$$\log p(\Theta; \mathbf{y}) \approx \log p(\tilde{\Theta}; \mathbf{y}) - \frac{1}{2}(\Theta - \tilde{\Theta})^\top H(\tilde{\Theta})(\Theta - \tilde{\Theta}), \quad (4.5.5)$$

por notar que os termos de primeira ordem da série se anulam, já que  $\tilde{\Theta}$  é moda. Substituindo a relação acima no integrando de 4.5.4, pode-se perceber que, a menos de uma constante normalizadora, tem-se uma densidade de uma normal com média  $\tilde{\Theta}$  e matriz de covariâncias  $H(\tilde{\Theta})$ , ou seja

$$\begin{aligned} p(\mathbf{y}|M_k) &= \exp[\log p(\tilde{\Theta}; \mathbf{y})] \int \exp\left(-\frac{1}{2}(\Theta - \tilde{\Theta})^\top H(\tilde{\Theta})(\Theta - \tilde{\Theta})\right) d\Theta \\ &= p(\tilde{\Theta}; \mathbf{y})(2\pi)^{d/2}|H(\tilde{\Theta})|^{-1/2}, \end{aligned} \quad (4.5.6)$$

então a log-verossimilhança integrada é aproximada por

$$\log p(\mathbf{y}|M) \approx \log L(\tilde{\Theta}) + \log p(\tilde{\Theta}) - \frac{1}{2} \log |H(\tilde{\Theta})| + \frac{1}{2} d \log(2\pi). \quad (4.5.7)$$

Uma variação importante da equação acima é quando  $\tilde{\Theta}$  é substituído pelo estimador de máxima verossimilhança  $\hat{\Theta}$  e a matriz hessiana pela matriz de informação de Fisher

$$\log p(\mathbf{y}|M) \approx \log L(\tilde{\Theta}) + \log p(\tilde{\Theta}) - \frac{1}{2} \log |I(\tilde{\Theta}, \mathbf{y})| + \frac{1}{2} d \log(2\pi). \quad (4.5.8)$$

Essa aproximação assume que a priori é muito difusa e seu efeito pode ser ignorado. Assim, o critério de informação Bayesiano de Schwarz (1978) é obtido ignorando os termos de  $O(1)$  em (4.5.8) e notando que  $|I(\hat{\Theta}; \mathbf{y})| = O(n^d)$ , resultando

$$BIC = -2 \log L(\hat{\Theta}) + d \log n. \quad (4.5.9)$$

O procedimento seleciona então aquele modelo que apresenta o menor valor de BIC. A forma com que o BIC foi construído sugere que esse critério tende a favorecer modelos mais simples, visto que o seu termo penalizador depende do tamanho amostral, que geralmente é maior que o número de parâmetros, considerado pelo critério AIC.

Diferente do AIC, o BIC é consistente em ordem, o que implica que assintoticamente tende a selecionar o modelo de dimensão correta. Esse critério, entretanto, foi desenvolvido sob as mesmas condições de regularidade, que não se verificam para modelos de misturas finitas. Entretanto, como Fraley e Raftery (1998) apontam, há considerável suporte para utilizar o BIC nesse contexto. Estudos de simulações nos quais o uso do BIC é empregado na seleção de modelos podem ser encontrados em Roeder e Wasserman (1997).

### 4.5.3 Critério de informação por validação cruzada - CVIC

O critério de informação por validação cruzada, proposto por Smith (2000), escolhe o melhor modelo baseado na verossimilhança cruzada, isto é

$$\sum_{i=1}^n \log f(\mathbf{y}_i; \hat{\Theta}_{(i)}), \quad (4.5.10)$$

sendo  $\hat{\Theta}_{(i)}$  o estimador de máxima verossimilhança de  $\Theta$  formado da amostra observada  $\mathbf{y}_1, \dots, \mathbf{y}_n$  depois de deletada a  $i$ -ésima observação ( $i = 1, \dots, n$ ). O uso da validação cruzada nesse contexto, pode ser visto como um método de avaliação do modelo ajustado em uma amostra de teste de mesmo tamanho da amostra original (ou amostra de

treinamento). O modelo selecionado por esse procedimento é aquele que apresenta o maior valor de verossimilhança cruzada. Este procedimento para seleção de modelos, entretanto, pode ser muito custoso computacionalmente visto que apenas uma observação é deletada por vez. Outras alternativas podem ser encontradas em McLachlan e Peel (2000) para diminuir o custo desse processo.

## 4.6 Clusterização com Modelos de Misturas

A teoria da decisão pode ser utilizada como uma ferramenta eficiente para construir regras de discriminação em situações onde a alocação de uma entidade em categorias é desejada. Em um contexto de modelos de misturas, a alocação é com respeito à componente de mistura.

Para isso, considere então o vetor de características  $\mathbf{Y} \in \mathbb{R}^p$  (ou o vetor de dados observados), e o problema de classifica-lo a um dos  $g$  grupos, ou componentes de mistura. Assim como definido na Seção 4.2, considere a variável aleatória  $Z$ , assumindo valores no conjunto  $\mathcal{A} = \{1, \dots, g\}$  com probabilidades  $p_1, \dots, p_g$ , lembrando que essas são as probabilidades *a priori* de que a entidade  $\mathbf{Y} = \mathbf{y}$  pertença ao seu correspondente grupo. A ideia, então, é utilizar a probabilidade *a posteriori* definida em (4.2.2), na construção de um classificador. Para formalizar essa ideia, será apresentado em seguida uma série de resultados de teoria da decisão para construir um classificador ótimo, sob algum aspecto estatístico.

**Definição 4.6.1.** *Uma regra de decisão (ou classificador) é qualquer função  $r(\cdot)$  para qual se tem  $r : \mathbb{R}^p \rightarrow \mathcal{A} = \{1, \dots, g\}$ , sendo  $\mathcal{A}$  o espaço de decisões.*

Da definição anterior, dado um classificador  $r$  e uma observação  $\mathbf{Y} = \mathbf{y}$ ,  $r(\mathbf{y}) = j$  significa que a observação é alocada para o grupo  $j$ . Note que, pela definição de regra de decisão, existem uma infinidade de classificadores, e portanto, surge a necessidade de avaliá-los com relação a algum critério.

A maneira mais comum de formar uma regra de decisão é através de uma função perda. Essa função, denominada por  $\lambda : (\mathcal{A}, \mathcal{A}) \rightarrow \mathbb{R}$ , dá a perda decorrente do processo

de alocação. Uma função de perda comumente utilizada é a perda 0 – 1, que atribui perda 1 para uma observação alocada incorretamente e 0, caso contrário. Formammente, a perda 0 – 1 é definida por

$$\lambda(Z, r(\mathbf{Y})) = \begin{cases} 0, & \text{Se } Z = r(\mathbf{Y}) \\ 1, & \text{Se } Z \neq r(\mathbf{Y}) \end{cases} \quad (4.6.1)$$

É importante ressaltar que a função perda  $\lambda(\cdot, \cdot)$  é uma função do vetor aleatório  $\mathbf{Y}$  e da variável aleatória  $Z$ , e com isso também é aleatória. Desta forma, a construção da regra de decisão ótima deve ser feita em termos de seu valor esperado.

**Definição 4.6.2.** *Para um classificador  $r$ , com uma função de perda  $\lambda$ , temos que:*

1. *A função de Risco é a perda esperada condicional a  $Z$ , ou seja*

$$\begin{aligned} R(i, r) &= E[\lambda(i, r(\mathbf{Y})) | Z = i] \\ &= \sum_{j=1}^g \lambda(i, j) P[r(\mathbf{Y}) = j | Z = i]. \end{aligned} \quad (4.6.2)$$

2. *O risco total, ou risco de  $r$ , é a perda total esperada como função de  $\mathbf{Y}$  e  $Z$ , ou seja*

$$\begin{aligned} R(r) &= E[R(Z, r)] \\ &= \sum_{i=1}^g R(i, r) P[Z = i] \\ &= \sum_{i=1}^g \sum_{j=1}^g \lambda(i, j) P[r(\mathbf{Y}) = j | Z = i] P[Z = i]. \end{aligned} \quad (4.6.3)$$

Pela definição acima,  $R(i, r)$  é a perda esperada por alocar as observações pertencentes ao grupo  $i$  enquanto que  $R(r)$  é a perda esperada em todo o processo de alocação

empregando o classificador  $r$ . Considerando a perda  $0 - 1$ , obtem-se

$$R(i, r) = \sum_{\substack{j=1 \\ j \neq i}}^g P[r(\mathbf{Y}) = j | Z = i], \quad (4.6.4)$$

$$R(r) = \sum_{i=1}^g \sum_{\substack{j=1 \\ j \neq i}}^g P[r(\mathbf{Y}) = j | Z = i] P[Z = i]. \quad (4.6.5)$$

De (4.6.4) e (4.6.5), nota-se que  $R(i, r)$  é a probabilidade de má classificação dos objetos pertencentes a grupo  $i$  e  $R(r)$  é a probabilidade total de má classificação considerando a regra de decisão  $r$ .

Lembrando que  $P[Z = j] = p_j$  e que  $\hat{z}_j = \frac{p_j \psi_j(\mathbf{Y})}{f(\mathbf{Y})}$ , o seguinte resultado pode ser enunciado

**Teorema 4.6.1.** *Considere a regra de classificação*

$$r^*(x) = k, \quad \text{Se } \hat{z}_k = \max_j \hat{z}_j \quad (4.6.6)$$

então, considerando a função de perda  $\lambda$  como a  $0 - 1$ ,  $r^*$  é a regra que minimiza o risco total.

*Demonstração.* Note que, da equação (4.6.3) temos

$$\begin{aligned} R(r) &= \sum_{i=1}^g \sum_{j=1}^g \lambda(i, j) P[r(\mathbf{Y}) = j | Z = i] P[Z = i] \\ &= \sum_{j=1}^g \sum_{i=1}^g \lambda(i, j) P[Z = i | \mathbf{Y} = \mathbf{y}] P[r(\mathbf{Y}) = j]. \end{aligned}$$

Portanto, para que o risco total seja minimizado, a regra  $r$  deve tomar uma classe  $k$  de forma a minimizar  $\sum_{i=1}^g \lambda(i, k) P[Z = i | \mathbf{Y} = \mathbf{y}]$ . Considerando a perda  $0 - 1$ , temos que

$$\sum_{i=1}^g \lambda(i, k) P[Z = i | \mathbf{Y} = \mathbf{y}] = \sum_{\substack{i=1 \\ i \neq k}}^g P[Z = i | \mathbf{Y} = \mathbf{y}] = 1 - P[Z = k | \mathbf{Y} = \mathbf{y}] = 1 - \hat{z}_k$$

Assim,  $\hat{z}_k$  deve ser máximo dentre as probabilidades *a posteriori*  $\hat{z}_j$ ,  $j = 1, \dots, g$ , o que

equivale a tomar a decisão (4.6.6). □

Se for considerada uma família paramétrica de distribuições para as componentes de mistura, o verdadeiro valor de  $\hat{z}_j$  não pode ser calculado por não se conhecer o verdadeiro valor dos parâmetros envolvidos na análise. Uma maneira de contornar esse problema é utilizar a estimativa desse vetor de parâmetros, obtida via algoritmo EM. Portanto, a regra de decisão (4.6.6) pode ser estimada pela chamada *plug-in-rule*  $r^*(\mathbf{y}; \hat{\Theta})$ .

Suponha, portanto, que o objetivo de se ajustar um modelo de misturas definido em 4.1.1 seja classificar uma amostra aleatória  $\mathbf{y}_1, \dots, \mathbf{y}_n$  em  $g$  grupos. Considerando a estrutura de dados incompletos descrita na Seção 4.2, o objetivo é então inferir  $\mathbf{z}_i$  em termos dos dados observados  $\mathbf{y}_i$ . Depois de se ajustar um modelo de mistura de  $g$ -componentes, estima-se o vetor  $\hat{\Theta}$  de parâmetros desconhecidos do modelo, o qual 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  $g$  probabilidades  $\hat{z}_{i1}, \hat{z}_{i2}, \dots, \hat{z}_{ig}$  dão as probabilidades *a posteriori* da  $i$ -ésima observação pertencer ao primeiro, segundo, ..., e  $g$ -ésimo grupo, respectivamente. Segundo a regra de bayes determinada, classifica-se a observação para aquele grupo correspondente ao maior valor observado das probabilidades *a posteriori*.

## 5 *Misturas Finitas de Densidades MESN*

Este capítulo pode ser visto como resultado final desse trabalho, pois é aqui que toda metodologia proposta nos capítulos anteriores é utilizada de maneira a compor o modelo de mistura finita de distribuições da classe *MESN* (*MF-MESN*). Esse modelo propõe uma metodologia robusta para lidar com dados tomados de distribuições complexas, longe da normalidade, sendo capaz de acomodar simultaneamente multimodalidade, assimetria e valores extremos.

O capítulo se inicia com a definição do modelo *MF-MESN*, mostrando em seguida a sua representação hierárquica, a qual é utilizada para derivar o algoritmo EM para estimação dos parâmetros envolvidos no modelo. Um aproximação da matriz de informação observada é então derivada para elementos dessa classe. Uma breve discussão a respeito dos valores iniciais do algoritmo EM também é tratada nesse capítulo, visto sua grande importância no contexto de estimação em modelos de misturas. Por fim, algumas aplicações mostram a grande flexibilidade desses modelos e sua necessidade em situações onde a suposição de normalidade das componentes não é satisfeita.



## 5.1 O modelo MF-MESN

### 5.1.1 Definição

O modelo de mistura de distribuições MESN pode ser obtido juntando as definições 3.1.1 e 4.1.1.

**Definição 5.1.1.** *Considere os vetores aleatórios  $\mathbf{Y}_i \in \mathbb{R}^p$ ,  $i = 1, \dots, n$ , independentes e identicamente distribuídos, com densidade dada por*

$$f(\mathbf{y}_i; \boldsymbol{\Theta}) = \sum_{j=1}^g p_j \psi_j(\mathbf{y}_i), \quad p_j \geq 0 \quad e \quad \sum_{j=1}^g p_j = 1, \quad (5.1.1)$$

e  $\psi_j(\cdot) = \psi(\cdot; \boldsymbol{\theta}_j)$  uma densidade MESN( $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, H$ ). Sob esta representação, denota-se este modelo por MF-MESN com  $g$ -componentes de mistura.

Note que a família de densidades MESN é paramétrica, e portanto o vetor de parâmetros de interesse fica dado por  $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_g^\top)^\top$ , com  $\boldsymbol{\theta}_j = (p_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu})$ ,  $j = 1, \dots, g$ , o vetor de parâmetros específicos das componentes e  $p_j$  as probabilidades de mistura. Ainda, foi assumido que os vetores de parâmetros indexando a distribuição do fator de mistura de escala skew-normal sejam iguais para cada componente, isto é,  $\boldsymbol{\nu}_1 = \dots = \boldsymbol{\nu}_g = \boldsymbol{\nu}$ .

### 5.1.2 A representação hierárquica

Como na Seção 4.2, será considerado para determinação do modelo hierárquico, o vetor de dados latentes  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ig})$ . Cabe lembrar que esse vetor de dados não é observável e tem por finalidade associar a  $i$ -ésima observação da amostra a uma das  $g$  componentes de misturas consideradas. Neste sentido, a distribuição de  $\mathbf{Y}_i | Z_{ij} = 1$  é MESN( $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, H$ ) e  $\mathbf{Z}_i \sim \text{Multi}(1; p_1, \dots, p_g)$ . Considerando ainda a representação hierárquica de um vetor aleatório MESN dada em (3.3.1)-(3.3.3), tem-se o seguinte resultado

**Proposição 5.1.1.** *Considere a amostra  $\mathbf{Y}_i \in \mathbb{R}^p$ ,  $i = 1, \dots, n$ , com densidade dada na definição 5.1.1. Então o modelo hierárquico para cada vetor aleatório MESN dessa*

amostra é dado por

$$\mathbf{Y}_i|u_i, t_i, Z_{ij} = 1 \sim N_p(\boldsymbol{\mu}_j + \boldsymbol{\Delta}_j t_i, \kappa(u_i)\boldsymbol{\Gamma}_j), \quad (5.1.2)$$

$$T_i|u_i, Z_{ij} = 1 \sim HN(0, \kappa(u_i)), \quad (5.1.3)$$

$$U_i|Z_{ij} = 1 \sim H(u_i; \boldsymbol{\nu}), \quad (5.1.4)$$

$$\mathbf{Z}_i \sim \text{Multi}(1; p_1, \dots, p_g), \quad (5.1.5)$$

com

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

Além disso, considerando  $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$ ,  $\mathbf{u} = (u_1, \dots, u_n)^\top$ ,  $\mathbf{t} = (t_1, \dots, t_n)^\top$ , a função de log-verossimilhança completa de  $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_g^\top)^\top$ , com  $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\alpha}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu})^\top$ , e  $\boldsymbol{\alpha}_j$  denotando o vetor com os elementos da matriz triangular superior  $\boldsymbol{\Sigma}_j$ , é dada por

$$\begin{aligned} \ell_c(\boldsymbol{\Theta}) &= C + \sum_{i=1}^n \sum_{j=1}^g Z_{ij} \left[ \log(p_j) - \frac{1}{2} \log |\boldsymbol{\Gamma}_j| \right. \\ &\quad \left. - \kappa^{-1}(u_i) (\mathbf{y}_i - \boldsymbol{\mu}_j - \boldsymbol{\Delta}_j t_i)^\top \boldsymbol{\Gamma}_j^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_j - \boldsymbol{\Delta}_j t_i) \right. \\ &\quad \left. + \log(h(u_i; \boldsymbol{\nu})) \right], \\ &= C + \sum_{i=1}^n \sum_{j=1}^g \left[ Z_{ij} \log(p_j) - \frac{1}{2} Z_{ij} \log |\boldsymbol{\Gamma}_j| - \frac{1}{2} Z_{ij} \kappa^{-1}(u_i) (\mathbf{y}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Gamma}_j^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_j) \right. \\ &\quad \left. + Z_{ij} \kappa^{-1}(u_i) T_i (\mathbf{y}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Gamma}_j^{-1} \boldsymbol{\Delta}_j + \frac{1}{2} Z_{ij} \kappa^{-1}(u_i) T_i^2 \boldsymbol{\Delta}_j^\top \boldsymbol{\Gamma}_j^{-1} \boldsymbol{\Delta}_j \right. \\ &\quad \left. + \log(h(u_i; \boldsymbol{\nu})) \right] \end{aligned} \quad (5.1.6)$$

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

### 5.1.3 O algoritmo EM em modelos MF-MESN

Nesta seção, será implementado o algoritmo EM para estimação de máxima verossimilhança dos parâmetros do modelo MF-MESN. Cabe ressaltar que o modo como o algoritmo é derivado é muito semelhante àquele apresentado na Seção 3.3.2. Isto se dá pelo fato de se ter apenas um nível a mais na hierarquia do modelo em relação àquele

proposto fora do contexto de misturas. Em detrimento disso, algumas passagens podem ser omitidas.

No contexto de misturas de distribuições *MESN*, representação do modelo via dados aumentados pode ser obtida através do modelo hierarquico proposto na seção anterior. O vetor de dados observados é dado por  $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$ , e os dados aumentados por  $\mathbf{u} = (u_1, \dots, u_n)^\top$ ,  $\mathbf{t} = (t_1, \dots, t_n)^\top$  e  $\mathbf{z} = (\mathbf{z}_1^\top, \dots, \mathbf{z}_n^\top)^\top$ . Utilizando os mesmos resultados obtidos na Seção 3.3.2 para distribuições condicionais e propriedades da distribuição *half-normal*, as seguintes quantidades devem ser obtidas

$$\begin{aligned}\hat{z}_{ij} &= E\{Z_{ij} | \Theta = \hat{\Theta}, \mathbf{y}_i\}, \\ \hat{s}_{1ij} &= E\{Z_{ij}\kappa^{-1}(U_i) | \Theta = \hat{\Theta}, \mathbf{y}_i\}, \\ \hat{s}_{2ij} &= E\{Z_{ij}\kappa^{-1}(U_i)T_i | \Theta = \hat{\Theta}, \mathbf{y}_i\}, \\ \hat{s}_{3ij} &= E\{Z_{ij}\kappa^{-1}(U_i)T_i^2 | \Theta = \hat{\Theta}, \mathbf{y}_i\},\end{aligned}$$

Da Seção 4.3, tem-se que

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

Utilizando propriedades de esperança condicional, tem-se que

$$\begin{aligned}\hat{s}_{1ij} &= E\{Z_{ij}\kappa^{-1}(U_i) | \Theta = \hat{\Theta}, \mathbf{y}_i\} \\ &= E_{Z_{ij}|\mathbf{y}_i}\{E_{U|\mathbf{y}, Z_{ij}}[Z_{ij}\kappa^{-1}(U_i)]\} \\ &= E_{Z_{ij}|\mathbf{y}_i}\{Z_{ij} E_{U|\mathbf{y}, Z_{ij}}[\kappa^{-1}(U_i)]\} \\ &= \hat{z}_{ij}\hat{\kappa}_{ij},\end{aligned}$$

também

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

analogamente, obtem-se

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

lembrando que

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

da verossimilhança (5.1.6) e mais alguma algebra, pode-se mostrar que a log-verossimilhança esperada é dada por

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

O algoritmo ECME para a estimação de máxima verossimilhança do parâmetro  $\boldsymbol{\Theta}$  é uma extensão direta do algoritmo apresentado na Seção 3.3.2 e pode ser resumido como o seguinte:

**Etapa E:** Dado  $\Theta = \hat{\Theta}^{(k)}$ , obter  $\hat{z}_{ij}$ ,  $\hat{s}_{1ij}$ ,  $\hat{s}_{2ij}$ ,  $\hat{s}_{3ij}$ , for  $i = 1, \dots, n$  e  $j = 1, \dots, g$ .

**Etapa CM:** Para  $j = 1, \dots, g$ , atualizar  $\hat{\mu}_j^{(k)}$ ,  $\hat{\Gamma}_j^{(k)}$  e  $\hat{\Delta}_j^{(k)}$  usando as seguintes expressões fechadas para os parâmetros transformados

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

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

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

As iterações são repetidas até que alguma regra de convergência adequada seja satisfeita, por exemplo, se  $\|\Theta^{(k+1)} - \Theta^{(k)}\|$  for suficientemente pequeno ou alguma distância envolvendo a log-verossimilhança, como  $\|\ell(\Theta^{(k+1)}) - \ell(\Theta^{(k)})\|$

### 5.1.4 Matriz de informação observada

Nesta seção será derivada a matriz de informação observada para as estimativas de máxima verossimilhança obtidas anteriormente. Considerando o proposto na seção 4.4, a

matriz de covariâncias pode ser aproximada por  $I_e(\hat{\Theta})$ , sendo

$$I_e(\Theta) = \sum_{i=1}^n \mathbf{s}(\mathbf{y}_i; \Theta) \mathbf{s}^\top(\mathbf{y}_i; \Theta), \quad (5.1.7)$$

com

$$\mathbf{s}(\mathbf{y}_i; \Theta) = \frac{\partial}{\partial \Theta} \left( \log \sum_{j=1}^g p_j \psi_j(\mathbf{y}_i) \right),$$

isto é, o vetor gradiente da log-verossimilhança do modelo MF-MESN associada a  $\mathbf{y}_i$ .

Seja  $f(\cdot; \Theta)$  a função densidade de probabilidade do modelo MF-MESN em (5.1.1). Definindo

$$s_{i, \mu_r} = \frac{\partial}{\partial \mu_r} \log f(\mathbf{y}_i; \Theta),$$

e

$$D_{\mu_r}(\psi_r(\mathbf{y}_i)) = \frac{\partial \psi_r(\mathbf{y}_i)}{\partial \mu_r}.$$

As derivadas parciais com respeito aos outros parâmetros, específicos das componentes de misturas, são denotadas de maneira análoga. Então,

$$\begin{aligned} s_{i, p_r} &= \frac{\psi_r(\mathbf{y}_i) - \psi_g(\mathbf{y}_i)}{f(\mathbf{y}_i; \Theta)}, \quad i = 1, \dots, n, \quad r = 1, \dots, g-1, \\ s_{i, \mu_r} &= \frac{p_r D_{\mu_r}(\psi_r(\mathbf{y}_i))}{f(\mathbf{y}_i; \Theta)}, \quad s_{i, \alpha_{rk}} = \frac{p_r D_{\alpha_{rk}}(\psi_r(\mathbf{y}_i))}{f(\mathbf{y}_i; \Theta)}, \\ s_{i, \lambda_r} &= \frac{p_r D_{\lambda_r}(\psi_r(\mathbf{y}_i))}{f(\mathbf{y}_i; \Theta)}, \quad s_{i, \nu} = \frac{\sum_{j=1}^g p_j D_{\nu}(\psi_j(\mathbf{y}_i))}{f(\mathbf{y}_i; \Theta)}, \\ &i = 1, \dots, n, \quad r = 1, \dots, g, \end{aligned}$$

com  $\alpha_{rk}$  denotando o  $k$ -ésimo elemento de  $\alpha_r$ . Note que  $D_{\nu}(\psi_j(\mathbf{y}_i))$  deve ser obtido para cada caso particular mencionado na seção 3.2. Depois de alguma manipulação algébrica de  $\psi_r(\mathbf{y}_i)$  dado por (3.1.1), e usando a notação

$$\begin{aligned} I_{ir}^\Phi(w) &= \int_0^\infty \kappa^{-w}(u_i) \exp \left\{ -\frac{1}{2} \kappa^{-1}(u_i) d_{ir} \right\} \times \Phi(\kappa^{-1/2}(u_i) A_{ir}) dH(u_i), \\ I_{ir}^\phi(w) &= \int_0^\infty \kappa^{-w}(u_i) \exp \left\{ -\frac{1}{2} \kappa^{-1}(u_i) d_{ir} \right\} \times \phi_1(\kappa^{-1/2}(u_i) A_{ir}; 0, 1) dH(u_i), \end{aligned}$$

com

$$d_{ir} = (\mathbf{y}_i - \boldsymbol{\mu}_r)^\top \boldsymbol{\Sigma}_r^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_r), \quad \mathbf{A}_{ir} = \boldsymbol{\lambda}_r^\top \boldsymbol{\Sigma}_r^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_r),$$

$i = 1, \dots, n$ ,  $r = 1, \dots, g$ , pode-se mostrar que

$$\begin{aligned} D_{\boldsymbol{\mu}_r}(\psi_r(\mathbf{y}_i)) &= \frac{2|\boldsymbol{\Sigma}_r|^{-1/2}}{(2\pi)^{p/2}} \left[ \left( \frac{\partial A_{ir}}{\partial \boldsymbol{\mu}_r} \right) I_{ir}^\phi \left( \frac{p+1}{2} \right) \right. \\ &\quad \left. - \frac{1}{2} \left( \frac{\partial d_{ir}}{\partial \boldsymbol{\mu}_r} \right) I_{ir}^\Phi \left( \frac{p}{2} + 1 \right) \right], \end{aligned}$$

$$\begin{aligned} D_{\alpha_{rk}}(\psi_r(\mathbf{y}_i)) &= \frac{2}{(2\pi)^{p/2}} \left[ \left( \frac{\partial |\boldsymbol{\Sigma}_r|^{-1/2}}{\partial \alpha_{rk}} \right) I_{ir}^\Phi \left( \frac{p}{2} \right) \right. \\ &\quad \left. - \frac{1}{2} \left( \frac{\partial d_{ir}}{\partial \alpha_{rk}} \right) |\boldsymbol{\Sigma}_r|^{-1/2} I_{ir}^\Phi \left( \frac{p}{2} + 1 \right) \right. \\ &\quad \left. + |\boldsymbol{\Sigma}_r|^{-1/2} \left( \frac{\partial A_{ir}}{\partial \alpha_{rk}} \right) I_{ir}^\phi \left( \frac{p+1}{2} \right) \right], \end{aligned}$$

$$D_{\boldsymbol{\lambda}_r}(\psi_r(\mathbf{y}_i)) = \frac{2|\boldsymbol{\Sigma}_r|^{-1/2}}{(2\pi)^{p/2}} \left( \frac{\partial A_{ir}}{\partial \boldsymbol{\lambda}_r} \right) I_{ir}^\phi \left( \frac{p+1}{2} \right).$$

Expressões para as derivadas são dadas no apêndice. A substituição de  $H$  nas integrais acima resulta nos seguintes resultados para cada distribuição considerada nesse trabalho, a saber

- *Distribuição skew-t*

$$\begin{aligned}
I_{ir}^{\Phi}(w) &= \frac{2^w \nu^{\nu/2} \Gamma(w + \nu/2)}{\sqrt{2\pi} \Gamma(\nu/2) (\nu + d_{ir})^{\nu/2+w}} \\
&\quad \times T\left(\frac{\mathbf{A}_{ir}}{(d_{ir} + \nu)^{1/2}} \sqrt{2w + \nu}; 2w + \nu\right), \\
I_{ir}^{\phi}(w) &= \frac{2^w \nu^{\nu/2}}{\sqrt{2\pi} \Gamma(\nu/2)} \left(\frac{1}{d_{ir} + \mathbf{A}_{ir}^2 + \nu}\right)^{\frac{\nu+2w}{2}} \\
&\quad \times \Gamma\left(\frac{\nu + 2w}{2}\right),
\end{aligned}$$

$$\begin{aligned}
D_{\nu}(\psi_j(\mathbf{y}_i)) &= (2\pi)^{-p/2} |\Sigma_j|^{-1/2} (I_{ij}^{\Phi}(p/2)(1 + \log(\nu/2) \\
&\quad - DG(\nu/2)) - I_{ij}^{\Phi}(1 + p/2) \\
&\quad + \int_0^{\infty} u_i^{p/2} \log(u_i) \exp(-u_i d_{ij}/2) \\
&\quad \times \Phi(u_i^{1/2} \mathbf{A}_{ij}) h(u_i; \nu) du_i),
\end{aligned}$$

sendo  $DG$  a função digamma e  $h(u; \nu)$  a densidade gamma com parametros  $\nu/2$ .

- *Distribuição skew-slash*

$$\begin{aligned}
I_{ir}^{\Phi}(w) &= \frac{2^{2+\nu} \Gamma(w + \nu)}{d_{ir}^{w+\nu}} P_1(w + \nu, d_{ir}/2) E[\Phi(S_{ir}^{1/2}) \mathbf{A}_{ir}], \\
I_{ir}^{\phi}(w) &= \frac{\nu 2^{w+\nu} \Gamma(w + \nu)}{\sqrt{2\pi} (d_{ir} + \mathbf{A}_{ir}^2)^{w+\nu}} P_1\left(w + \nu, \frac{d_{ir} + \mathbf{A}_{ir}^2}{2}\right),
\end{aligned}$$

com  $S_{ir} \sim \text{Gamma}(w + \nu, d_{ir}/2) I_{(0,1)}$ , e

$$\begin{aligned}
D_{\nu}(\psi_j(\mathbf{y}_i)) &= 2(2\pi)^{-p/2} |\Sigma_j|^{-1/2} \left\{ I^{\Phi}(p/2 + \nu - 1) \right. \\
&\quad + \nu \int_0^1 u_i^{\frac{p}{2} + \nu - 1} \log(u_i) \exp\{-u_i d_{ij}/2\} \\
&\quad \times \Phi(u_i^{1/2} \mathbf{A}_{ij}) du_i \left. \right\}.
\end{aligned}$$



- *Distribuição skew-normal contaminada*

$$\begin{aligned}
I_{ir}^\Phi(w) &= \sqrt{2\pi} \{ \nu_1 \nu_2^{w-1/2} \phi(d_{ir}|0, 1/\nu_2) \\
&\quad \times \Phi(\nu_2^{1/2} \mathbf{A}_{ir}) + (1 - \nu_1) \phi(d_{ir}|0, 1) \\
&\quad \times \Phi(\mathbf{A}_{ir}) \}, \\
I_{ir}^\phi(w) &= \{ \nu_1 \nu_2^{w-1/2} \phi(d_{ir} + \mathbf{A}_{ir}^2|0, 1/\nu_2) \\
&\quad + (1 - \nu_1) \phi(d_{ir} + \mathbf{A}_{ir}^2|0, 1) \}, \\
D_{\nu_1}(\psi_j(\mathbf{y}_i)) &= 2(\phi(y_i|\mu_j, \nu_2^{-1}\sigma_j^2)\Phi(\nu_2^{1/2} \mathbf{A}_{ij}) \\
&\quad - \phi(y_i|\mu_j, \sigma_j^2)\Phi(\mathbf{A}_{ij})), \\
D_{\nu_2}(\psi_j(\mathbf{y}_i)) &= (2\pi)^{-p/2} |\Sigma_j|^{-1/2} \nu_1 \nu_2^{p/2} \\
&\quad \times \exp\{-\nu_2 d_{ij}/2\} \left\{ \mathbf{A}_{ij} \nu_2^{-1/2} \phi(\nu_2^{1/2} \mathbf{A}_{ij}) \right. \\
&\quad \left. - d_{ij} \Phi(\nu_2^{1/2} \mathbf{A}_{ij}) + p \nu_2^{-1} \Phi(\nu_2^{1/2} \mathbf{A}_{ij}) \right\}.
\end{aligned}$$

A aproximação (5.1.7) é assintoticamente aplicável. Entretanto, pode não ser confiável quando o tamanho amostral não é suficientemente grande. Na prática, é comum utilizar a aproximação por bootstrap (Efron e Tibshirani, 1986) para obter uma estimativa da matriz de covariância para  $\hat{\Theta}$ . Esse método pode aproximar os desvios padrão para as estimativas de maneira mais precisa que (5.1.7). Neste trabalho não será utilizada o método bootstrap já que requer um custo computacional muito elevado.

### 5.1.5 Notas para implementação do algoritmo EM

É sabido que modelos de mistura podem apresentar log-verossimilhança com mais de uma moda. Neste sentido, a estimação de máxima verossimilhança via algoritmo EM pode não resultar em máximos globais se os valores iniciais estiverem muito afastados do verdadeiro valor dos parâmetros. Consequentemente, a escolha dos valores iniciais para o algoritmo EM é muito importante no contexto de estimação em modelos de misturas.

Seguindo Lin (2009), esta dificuldade pode ser evitada tomando o seguinte procedimento:

- (i) Performa-se a clusterização pelo método K-Means (Hartigan e Wong, 1979);
- (ii) Especifica-se as indicadoras de componentes  $Z_{ij}^{(0)}$  de acordo com os resultados obtidos pela clusterização K-Means;
- (iii) Os valores iniciais para as probabilidades de mistura, locação e matriz de escala das componentes, são obtidas, respectivamente, por

$$\begin{aligned}
 p_j^{(0)} &= \frac{1}{n} \sum_{i=1}^n Z_{ij}^{(0)}, \quad \boldsymbol{\mu}_j^{(0)} = \frac{\sum_{i=1}^n Z_{ij}^{(0)} \mathbf{y}_i}{\sum_{i=1}^n Z_{ij}^{(0)}}, \\
 \boldsymbol{\Sigma}_j^{(0)} &= \sum_{i=1}^n Z_{ij}^{(0)} (\mathbf{y}_i - \boldsymbol{\mu}_j^{(0)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(0)})^\top / \sum_{i=1}^n Z_{ij}^{(0)}.
 \end{aligned}$$

Nas aplicações com conjunto de dados reais consideradas nas seções 5.2 e 5.3 será utilizada a seguinte notação: MF-NOR, MF-SN, MF-ST, MF-SNC e MF-SS, para representar misturas de normais, *skew*-normal, *skew*-*t*, *skew*-normal contaminada e *skew*-slash, respectivamente. O procedimento apresentado para se obter os valores iniciais foi utilizado apenas para o modelo MF-NOR. Então, as estimativas EM para os vetores de locação, matrizes de covariancia e proporções de mistura, foram utilizados como valores iniciais no algoritmo EM para a estimação dos parâmetros correspondentes no o modelo MF-SN. Seja  $\lambda_{ij}$  a  $j$ -ésima cordenada do vetor de parâmetros de assimetria associado com a  $i$ -ésima componente. Seu valor inicial foi obtido da seguinte forma: para a sub-amostra alocada à  $i$ -ésima componente determinada pelo método K-Means, seja  $\hat{\rho}_{ij}$  o coeficiente de assimetria amostral para variável  $j$ . Então,  $\lambda_{ij}^{(0)} = 3 \times \text{sign}(\hat{\rho}_{ij})$ . As estimativas EM obtidas no o modelo MF-SN foram passadas como valores iniciais para o algoritmo EM nos modelos MF-ST e MF-SCN. Os valores iniciais para  $\nu$  foram tomados próximos de 10 e 0.5 para os modelos MF-ST e MF-SCN, respectivamente. Para o modelo MF-SSL, os valores iniciais utilizados foram os resultados obtidos na estimativa EM considerando o modelo MF-ST.

Mesmo que esse procedimento pareça razoável para determinar os valores iniciais, a tradição na prática é tentar diferentes valores iniciais para o algoritmo EM. Isto é recomendado pois, se existir mais de uma moda, pode-se determinar a global comparando seus valores de log-verossimilhança. Também, com esse método pode-se comprovar a

estabilidade das estimativas resultantes. A escolha dos valores iniciais poderia ser feita, por exemplo, aplicando métodos bootstrap, seguindo o esquema anterior. Outra sugestão - válida para misturas de modelos assimétricos - é: depois da clusterização por K-means e dados os valores iniciais para assimetria e para o parâmetro de graus de liberdade (talvés utilizando as estratégias consideradas anteriormente), pode-se usar as expressões da esperança populacional e matriz de covariâncias dados nas seções 3.2.1, 3.2.2 e 3.2.3 para obter as estimativas pelo método dos momentos, o qual seriam utilizados como valores iniciais.

## 5.2 Aplicações a Dados Reais - Caso Univariado

### 5.2.1 *Body Mass Index* data

A primeira aplicação no contexto univariado considera o índice de massa corporal para homens com idade entre 18 e 80 anos no exame nacional de saúde e nutrição, realizado pelo Centro Nacional para Estatísticas de Saúde (National Center for Health Statistics - NCHS), do Centro para Controle de Doenças (Center for Disease Control - CDC) nos EUA. O problema de obesidade tem chamado muita atenção nos últimos anos devido a sua forte relação com muitas doenças crônicas recorrentes. O índice de massa corporal (IMC,  $kg/m^2$ ) se tornou uma media padrão quando se trata de sobrepeso e obesidade. Esse índice é dado pela razão do peso corporal em kilogramas e o quadrado da altura em metros.

Esse conjunto de dados foi analisado por Lin et al. (2007b), que considerou apenas os relatórios datados de 1999–2000 e 2001–2002. Originalmente, o conjunto de dados continha 4579 participantes com registros de seus IMC. Entretanto, para explorar um comportamento de misturas, Lin et al. (2007b) considerou somente os participantes que tinham seus pesos entre [39.50 kg, 70.00 kg] e [95.01 kg, 196.80 kg]. O conjunto de dados resultante consite de 1069 participantes no primeiro subgrupo e 1054 no segundo. Lin ajustou e comparou os modelos de misturas de duas componentes normais (MF-NOR), duas componentes *t-Student* (MF-T), duas componentes *skew-normal* (MF-SN) e duas

componentes *skew-t* (MF-ST).

Além desses modelos propostos, serão aqui ajustados também os modelos de mistura de duas componentes *skew-normal* contaminada (MF-SNC) e o de duas componentes *skew-slash* (MF-SS). Cada uma dessas distribuições pode ser considerada como um caso particular da classe MESN. Mais especificamente, foi considerado o seguinte modelo para ajustar os dados

$$f(y_i; \Theta) = p\psi(y_i; \mu_1, \sigma_1^2, \lambda_1, \nu) + (1 - p)\psi(y_i; \mu_2, \sigma_2^2, \lambda_2, \nu), \quad (5.2.1)$$

com  $\psi(y; \Theta)$  dado por 3.1.1. Note que, quando  $\psi(y; \Theta)$  é a densidade *skew-t* com  $\lambda_1 = \lambda_2 = 0$  e  $\nu \rightarrow \infty$  obtem-se o modelo MF-NOR, só com  $\lambda_1 = \lambda_2 = 0$  obtem-se o modelo MF-T e só com  $\nu \rightarrow \infty$  o modelo MF-SN. Os modelos MF-NOR e MF-SN também podem ser obtidos como casos especiais de outros elementos da classe MESN.

Na Tabela 5.2.1 são apresentadas as estimativas de máxima verossimilhança e os desvios padrão de todos os parâmetros, para cada modelo mencionando anteriormente. Os valores iniciais passados para o algoritmo EM foram determinados seguindo o esquema proposto na Seção 5.1.5. Para comparação dos modelos, foi utilizado o AIC e o BIC.

Por fim, é interessante comparar os ajustes das densidades do modelo MF-NOR, MF-SS e MF-ST (o melhor ajustado). A Figura 5.2.1 mostra o histograma dos dados sobreposto com as curvas ajustadas das densidades dos três modelos que foram aqui considerados. Considerando o gráfico, os ajustes resultantes dos modelos MF-ST, bem como o MF-SS, podem ser considerados melhores que o ajuste MF-NOR, fato também evidenciado pelos critérios AIC e BIC.

### 5.2.2 *Old Faithful* data

A segunda aplicação considerada aqui é baseada no conjunto de dados *Old Faithful Geyser* encontrado em Silverman (1986). Este conjunto de dados consiste de 272 medições

<i>Parameter</i>	MF-NOR		MF-SN		MF-ST		MF-SNC		MF-SS	
	Mle	Se	Mle	Se	Mle	Se	Mle	Se	Mle	Se
$p$	0.391	0.0188	0.528	0.0125	0.538	0.0142	0.538	0.0140	0.536	0.0135
$\mu_1$	21.412	0.0936	19.500	0.2429	19.572	0.2432	19.487	0.2363	19.512	0.2372
$\mu_2$	32.548	0.3681	28.760	0.1456	29.100	0.1652	29.023	0.1651	28.972	0.1544
$\sigma_1^2$	4.071	0.0873	14.365	0.2841	12.916	0.3072	12.864	0.3022	9.896	0.2636
$\sigma_2^2$	41.191	0.1578	63.217	0.1580	45.841	0.3100	44.543	0.4440	36.115	0.3414
$\lambda_1$	-	-	1.902	0.3446	1.900	0.3723	2.003	0.3973	1.955	0.3636
$\lambda_2$	-	-	10.588	2.7408	7.131	1.8474	7.656	2.1135	8.330	2.1402
$\nu$	-	-	-	-	8.759	2.1238	0.141	0.061	2.421	0.4169
$\gamma$	-	-	-	-	-	-	0.284	0.069	-	-
AIC	13833.35		13750.89		13726.67		13726.73		13726.56	
BIC	13961.61		13790.46		13771.89		13777.61		13771.78	

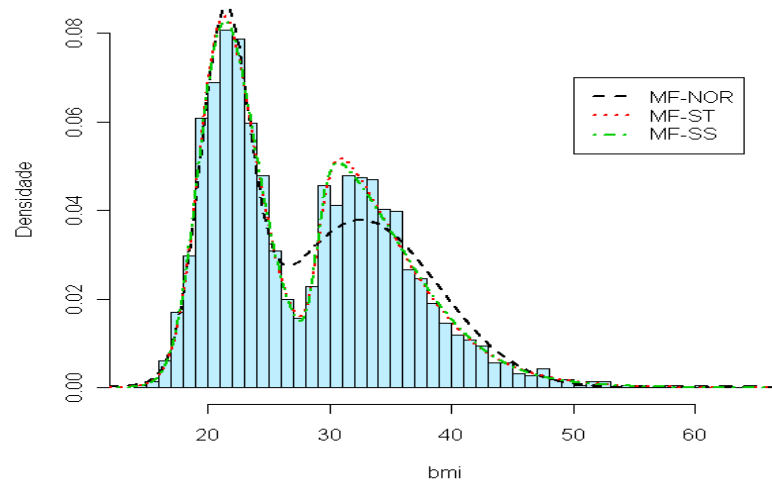
Tabela 5.2.1: Estimativas de máxima verossimilhança e desvios padrão para os dados *IMC*

Figura 5.2.1: Histograma dos dados de IMC com as curvas ajustadas pelos modelos MF-NOR, MF-ST and MF-SS

de tempos de erupção (em minutos) do géiser conhecido por *the Old Faithful*, localizado no Parque Nacional de Yellowstone, Wyoming, EUA. No contexto univariado, esse conjunto de dados foi analisado por Lin et al. (2007a) que ajustou um modelo de mistura de duas componentes *skew-normal*. Aqui serão considerados os mesmos modelos propostos na seção anterior, a saber, MF-NOR, MF-ST, MF-SNC e MF-SS.

Na Tabela 5.2.2 são apresentadas as estimativas de máxima verossimilhança e os desvios padrão de todos os parâmetros para cada modelo considerado. A figura 5.2.2 mostra o histograma dos dados *Old Faithful* sobreposto com as curvas ajustadas das densidades dos modelos MF-NOR, MF-SN e MF-ST. Neste caso, pode-se ver que os valores de AIC e BIC favorecem o modelo MF-SN.

Parameter	MF-NOR		MF-SN		MF-ST		MF-SNC		MF-SS	
	Mle	Se	Mle	Se	Mle	Se	Mle	Se	Mle	Se
$p$	0.348	0.0291	0.348	0.0293	0.348	0.0294	0.348	0.0438	0.348	0.0297
$\mu_1$	2.018	0.0221	1.726	0.0290	1.728	0.0287	1.727	0.0309	1.728	0.0298
$\mu_2$	4.273	0.0367	4.797	0.0514	4.793	0.0524	4.795	0.0719	4.797	0.0553
$\sigma_1^2$	0.055	0.0221	0.145	0.0415	0.137	0.0430	0.129	0.1165	0.119	0.0852
$\sigma_2^2$	0.190	0.0252	0.465	0.0620	0.448	0.0667	0.419	0.4796	0.396	0.1474
$\lambda_1$	-	-	5.811	2.1462	5.707	2.1269	5.780	2.2347	5.730	2.1891
$\lambda_2$	-	-	-3.438	1.1329	-3.366	1.1329	-3.372	1.1543	-3.396	1.1970
$\nu$	-	-	-	-	51.520	16.376	0.172	1.4654	5.685	11.6222
$\gamma$	-	-	-	-	-	-	0.5964	4.0236	-	-
AIC	562.720		529.135		531.067		533.058		531.073	
BIC	580.749		554.376		559.913		565.511		559.960	

Tabela 5.2.2: Estimativas de máxima verossimilhança e desvios padrão para os dados *Old Faithful*

## 5.3 Aplicações a Dados Reais - Caso Multivariado

### 5.3.1 *Swiss Bank* data

Como primeira aplicação no contexto multivariado, será empregada a metodologia proposta aqui no famoso conjunto de dados *Swiss bank*, primeiramente analisado por Flury e Riedwyl (1988) e posteriormente por Ma e Genton (2004) com uma distribuição flexível *skew-simétrica*. Lin (2009) também analisou este mesmo conjunto de dados com um modelo *skew-normal* ligeiramente diferente daquele proposto aqui.

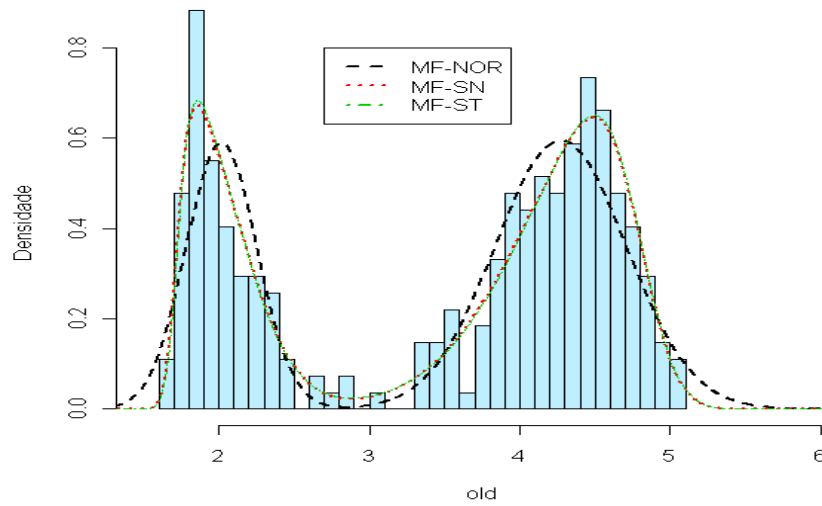


Figura 5.2.2: Histograma dos dados *Old Faithful* com as curvas ajustadas pelos modelos MF-NOR, MF-SN, MF-ST

Os dados consistem de seis medidas de dimensões tomadas em 100 notas verdadeiras e 100 notas falsas de 1000 francos suíços. Para explorar o contexto de misturas, a atenção foi focada em somente duas das seis medidas tomadas nas notas,  $X_1$ : a largura da borda direita e  $X_2$ : o comprimento diagonal da imagem central. Neste exemplo, os modelos de mistura estão sendo utilizados como um meio de classificar as observações em dois grupos, tratando o problema como sendo de classificação supervisionada em grupos de notas falsas e verdadeiras, evidentemente, ignorando esse conhecimento a priori.

O objetivo aqui foi checar qual dos modelos de misturas *MF-MESN* de duas componentes produz resultados mais satisfatórios quando ajustados aos dados. Mais especificamente, o modelo pode ser escrito por

$$f(\mathbf{y}_i; \Theta) = p \psi(\mathbf{y}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\lambda}_1, \nu) + (1 - p) \psi(\mathbf{y}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\lambda}_2, \nu),$$

com  $\psi(y; \Theta)$  dado pela equação (3.1.1) e

$$\begin{aligned} \boldsymbol{\mu}_j &= (\mu_{j1}, \mu_{j2})^\top, \quad \boldsymbol{\Sigma}_j = \begin{bmatrix} \sigma_{j,11} & \sigma_{j,12} \\ \sigma_{j,21} & \sigma_{j,22} \end{bmatrix}, \quad \boldsymbol{\lambda}_j = (\lambda_{j1}, \lambda_{j2})^\top, \\ j &= 1, 2. \end{aligned}$$

A Tabela 5.3.1 mostra as estimativas de máxima verossimilhança, a log-verossimilhança avaliada nessas estimativas e os desvios padrão de todos os parâmetros considerados em cada modelo. Como a convergência do algoritmo EM pode ficar comprometida em virtude da escolha dos valores iniciais, foi seguido o esquema sugerido na Seção 5.1.5 para determiná-los. A Tabela 5.3.2 apresenta os resultados para seleção de modelos baseado no critério de informação de Akaike (AIC), o critério de informação bayesiano (BIC), o critério de informação por validação cruzada (CVIC) e o critério de determinação eficiente (EDC) proposto por Bai et al. (1989) e Zhao et al. (2001). Assim como os critérios AIC e BIC, o EDC tem a forma

$$-2\ell(\hat{\boldsymbol{\theta}}) + \gamma c_n,$$

com  $\gamma$  o número de parâmetros a ser estimado no modelo e  $c_n$  um termo de penalização. Aqui, foi utilizado  $c_n = 0.2\sqrt{n}$ . Por esses métodos não terem sido originalmente propostos para o contexto de misturas, sua utilização deve ser tomada apenas para uma indicação de melhor modelo, e é claro que é preciso uma maior investigação a respeito da eficiência dos métodos de seleção para o contexto *MF-MESN*.



Tabela 5.3.1: Estimativas de máxima verossimilhança e desvios padrão para os dados *Swiss bank*

<i>Parameter</i>	MF-NOR		MF-SN		MF-ST		MF-SCN		MF-SS	
	Mle	Se	Mle	Se	Mle	Se	Mle	Se	Mle	Se
$\mu_{11}$	130.2050	0.0327	130.1140	0.0656	130.1160	0.0660	130.1200	0.0655	130.1140	0.0652
$\mu_{12}$	139.4960	0.0844	140.0110	0.0780	139.9860	0.7758	139.9760	0.0747	139.9940	0.0763
$\sigma_{1,11}$	0.0934	0.0186	0.0855	0.0256	0.0704	0.0302	0.0485	0.0609	0.0540	0.0228
$\sigma_{1,12}$	0.0441	0.0452	-0.0152	0.0894	-0.0097	0.0867	-0.0048	0.0715	-0.0094	0.0742
$\sigma_{1,22}$	0.3597	0.0452	0.6245	0.0726	0.4967	0.0855	0.3327	0.1757	0.3945	0.0633
$\lambda_{11}$	-	-	1.4528	0.9394	1.2679	0.7844	1.1835	0.7287	1.3183	0.8311
$\lambda_{12}$	-	-	-5.0930	2.0816	-4.4342	1.7948	-4.2199	1.7187	-4.5983	1.8575
$\mu_{21}$	129.6910	0.0339	129.3310	0.0678	129.3760	0.0828	129.3810	0.0881	129.3690	0.0785
$\mu_{22}$	141.5570	0.0413	141.7720	0.1211	141.7910	0.1273	141.8020	0.1235	141.7870	0.1285
$\sigma_{2,11}$	0.1023	0.0280	0.2993	0.0607	0.2250	0.0833	0.1606	0.1245	0.1746	0.0610
$\sigma_{2,12}$	-0.0004	0.0376	-0.1298	0.1067	-0.1082	0.1081	-0.0824	0.1023	-0.0820	0.8945
$\sigma_{2,22}$	0.1525	0.0345	0.2395	0.0749	0.2253	0.0855	0.1684	0.1305	0.1704	0.0709
$\lambda_{21}$	-	-	2.7130	1.1659	2.1254	1.0739	2.1092	1.1385	2.1821	1.0255
$\lambda_{22}$	-	-	-1.4263	1.0470	-1.3905	1.0820	-1.4715	1.0988	-1.3728	1.4000
$\nu$	-	-	-	-	12.7664	9.8835	0.4325 ( $\nu_1$ )	0.5201	2.7973	0.9958
	-	-	-	-	-	-	0.3851 ( $\nu_2$ )	0.1896	-	-
$p$	0.5200	0.0362	0.4985	0.0360	0.4959	0.0365	0.4938	0.3780	0.4964	0.3620

Tabela 5.3.2: Critérios de seleção de modelos para os dados *Swiss bank*

	MF-NOR	MF-SN	MF-ST	MF-SCN	MF-SS
log-likelihood	-322.1643	-307.9436	-306.2159	-305.5799	-306.2195
AIC	666.3285	645.8872	644.4339	645.1598	645.4388
BIC	702.6100	695.3620	697.2070	701.2312	698.2119
EDC	675.4413	658.3136	657.6866	659.2430	657.6938
CVIC	-335.1804	-328.8056	-323.8958	-324.3175	-320.0183

Neste ponto, alguns comentários devem ser feitos. Primeiro, nota-se que os critérios de seleção de modelos considerados aqui favorecem os modelos da classe *MF-MESN*. Segundo, o teste univariado assintótico de Wald mostra alguma evidência contra a hipótese  $\lambda_{ij} = 0$  for  $(i, j) = (1, 2)$  and  $(i, j) = (2, 1)$ , mostrando que a suposição de simetria não é adequada para esse caso. Além disso, o valor estimado para o parâmetro de acomodação de valores extremos ( $\nu$ ) mostra a necessidade de se utilizar componentes com as caudas mais pesadas do que a Normal.

Na Figura 5.3.1 encontram-se os contornos das densidades com o valor dos parâmetros substituídos por aqueles estimados na tabela 5.3.1, e os pontos já classificados pelos modelos MF-NOR e MF-SS, respectivamente. Essa classificação foi obtida através das probabilidades a posteriori  $\hat{z}_{ij}$ .

### 5.3.2 *Old Faithful* data

A segunda aplicação considerada aqui é baseada no conjunto de dados *Old Faithful Geyser* encontrado em Silverman (1986). Este conjunto de dados consiste de 272 medições de tempos de erupção (em minutos) do géiser conhecido por *the Old Faithful*, localizado no Parque Nacional de Yellowstone, Wyoming, EUA.

Esse famoso conjunto de dados tem sido analisado por muitos autores com diferentes perspectivas e metodologias, sob ponto de vista frequentista e bayesiano. Como exemplo, Lin et al. (2007b) ajustou modelos MF-SN e MF-NOR com duas componentes e comparou os resultados obtidos. Para considerar uma versão bivariada desses dados, Garcia-

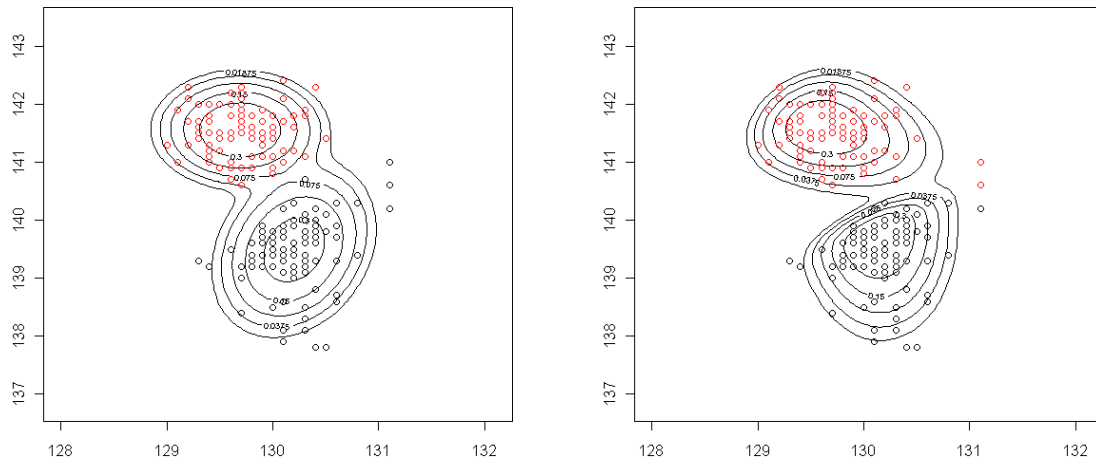


Figura 5.3.1: Densidades de contorno e pontos classificados pelos modelos MF-NOR (esquerda) e MF-SS para os dados *Swiss bank* - pontos vermelhos classificados como notas verdadeiras

Escudero e Gordaliza (1999) e, posteriormente, McLachlan e Peel (2000), defasaram os dados para explorar uma mistura de três componentes no contexto multivariado. Entretanto, para explorar a mistura de duas componentes bivariadas, foi considerado nessa aplicação os 272 tempos de erupção (em minutos) e o tempo de espera (em minutos) para a próxima erupção.

Primeiramente, foram ajustados os mesmos modelos considerados na seção anterior e seus resultados foram comparados em termos dos critérios AIC e BIC. Os resultados podem ser encontrados na Tabela 5.3.3. Note que ambos os critérios mostram o modelo MF-SN como melhor ajuste para dados. A Figura 5.3.2 mostra as densidades de contorno para os modelos MF-NOR e MF-SN.

Finalmente, alguns valores extremos foram introduzidos no conjunto de dados original. Isto foi feito por adicionar algumas constantes (colona 1 da Tabela 5.3.4) à 149ª observação. O objetivo aqui é mostrar a necessidade de um modelo mais robusto para lidar com valores extremos. Na Tabela 5.3.4 encontram-se os critérios AIC e BIC para os

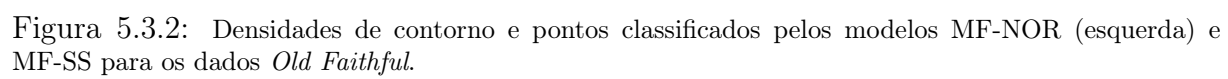
Tabela 5.3.3: Critérios de seleção de modelos para os dados *Old Faithful*

	MF-NOR	MF-SN	MF-ST	MF-SNC	MF-SS
log-likelihood	-1130.264	-1110.228	-1110.271	-1110.228	-1110.228
AIC	228.528	2250.456	2252.542	2254.546	2252.456
BIC	2322.192	2304.543	2310.234	2315.755	2310.149

Tabela 5.3.4: Critérios de seleção de modelos para os dados perturbados

Const	MF-NOR		MF-SN		MF-ST		$\hat{\nu}$
	AIC	BIC	AIC	BIC	AIC	BIC	
-10	2531.681	2571.344	2415.273	2469.360	2287.393	2345.086	6.645
-5	2357.391	2397.055	2302.452	2356.545	2275.493	2333.186	8.925
-2	2294.833	2334.497	2259.005	2313.093	2259.186	2316.179	23.258
0	2282.528	2322.192	2250.456	2304.543	2252.542	2310.234	>100
2	2316.083	2355.747	2294.799	2348.886	2278.547	2336.240	8.767
5	2407.290	2446.954	2365.781	2419.868	2290.356	2348.049	6.365
10	2553.117	2592.781	2436.404	2490.491	2298.270	2355.962	5.620

modelos MF-NOR, MF-SN e MF-ST para cada versão perturbada do conjunto de dados original. Também, é mostrado o valor  $\hat{\nu}$  para o modelo MF-ST, que pode ser entendido como parâmetro de acomodação de valores extremos. Claramente, pode-se ver que o modelo MF-ST tem melhor ajuste quanto mais atípica é a observação.



## 6 *Conclusões e Perspectivas*

Este trabalho teve por objetivo propor um modelo abrangente baseado em misturas finitas de distribuições da classe *MESN*, capaz de acomodar simultaneamente multimodalidade, assimetria e caudas pesadas. Por conter como casos particulares os modelos de misturas de normais, *t-student*, *skew-t* e outras, a metodologia proposta nessa tese mostra ter grande aplicabilidade em inúmeras situações encontradas na natureza. Foi contemplado nas aplicações que aqui se encontram o problema de se estimar densidades complexas e também o problema de classificação de observações.

O esforço principal desse trabalho foi dado no processo de estimação dos parâmetros do modelo via algoritmo EM. A preocupação que se teve foi em propor um algoritmo simples de ser implementado em qualquer ambiente de programação, com boas propriedades de convergência e bons aspectos computacionais, isso em virtude do grande número de formulas fechadas, tanto na etapa E quanto na etapa M do algoritmo. Cabe ainda ressaltar que a representação hierárquica do modelo não só aplica se a estimação por máxima verossimilhança, e pode ser empregada também sob o ponto de vista bayesiano (ver Cabral et al. - 2008), podendo ser implementada sem maiores dificuldades com o software WinBUGS.

Como perspectiva de trabalho futuro, pode-se propor a aplicação desses modelos na teoria de estimação em modelos lineares mistos. Esse tema tem recebido bastante interesse de pesquisa visto que a má especificação da distribuição dos efeitos aleatórios do modelo pode comprometer a estimação dos efeitos fixos, que geralmente são os parâmetros de

interesse em situações práticas. É sob essa motivação que se deseja propor uma classe de distribuições robusta e flexível para a distribuição dos efeitos aleatórios no modelo linear misto. Neste sentido, considere

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i + \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, n,$$

sendo  $\mathbf{Y}_i$  um vetor de dimensão  $n_i \times 1$  de respostas observadas da  $i$ -ésima unidade amostral,  $\mathbf{X}_i$  a matriz  $n_i \times p$  de desenho, correspondente aos efeitos fixos,  $\boldsymbol{\beta}$  o vetor de dimensão  $p \times 1$  de efeitos fixos,  $\mathbf{Z}_i$  a matriz de desenho  $n_i \times q$  associada ao vetor de efeitos aleatórios  $\mathbf{b}_i$  de dimensão  $q \times 1$  e  $\boldsymbol{\epsilon}_i$  o vetor de erros aleatórios ( $n_i \times 1$ ). Comunmente na literatura, é assumido que os efeitos aleatórios  $\mathbf{b}_i$  e as componentes residuais  $\boldsymbol{\epsilon}_i$  são independentes com  $\mathbf{b}_i \stackrel{iid}{\sim} N_q(\mathbf{0}, \mathbf{D})$  e  $\boldsymbol{\epsilon}_i \stackrel{iid}{\sim} N_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$ .

Como extensão, pode-se especificar um modelo linear misto de dois estágios que incorpora a distribuição *MF-MESN* da seguinte forma:

$$\begin{aligned} \mathbf{Y}_i | \mathbf{b}_i &\stackrel{ind}{\sim} N_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i + \mathbf{b}_i, \sigma_e^2 \mathbf{I}_{n_i}) \\ \mathbf{b}_i &\stackrel{ind}{\sim} \sum_{k=1}^G p_k \text{MESN}_q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\lambda}_k). \end{aligned} \quad (6.0.1)$$

Outras trabalhos podem ser desenvolvidos nesse tema, a saber,

- Estudos de simulação para determinar qual critério de seleção de modelos pode ser melhor empregado à classe de modelos *MF-MESN*.
- Intencificar estudos em teoria assintótica para se fazer inferência estatística quanto aos parâmetros.
- Implementar o algoritmo PX-EM (Liu, Rubin e Wu - 1998, e Liu - 2003b) para acelerar a convergência do algoritmo EM.

## APÊNDICE A – Lemas

**Lemma 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*

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

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

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

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

*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

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

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

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



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

$$\begin{aligned} E[X|X > a] &= \eta + \frac{\phi_1(\frac{a-\eta}{\tau})}{1 - \Phi_1(\frac{a-\eta}{\tau})} \tau \\ E[X^2|X > a] &= \eta^2 + \tau^2 + \frac{\phi_1(\frac{a-\eta}{\tau})}{1 - \Phi_1(\frac{a-\eta}{\tau})} (\eta + a) \tau \end{aligned}$$

## Matriz de informação

Neste apêndice, será utilizada a reparametrização  $\Sigma = \mathbf{D}^2$  para facilitar os calculos das derivadas, lembrando que  $A_{ir} = \boldsymbol{\lambda}_r^\top \Sigma_r^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_r)$  and  $d_{ir} = (\mathbf{y}_i - \boldsymbol{\mu}_r)^\top \Sigma_r^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_r)$ :

- $d_{ir}$

$$\begin{aligned} \frac{\partial d_{ir}}{\partial \boldsymbol{\mu}_r} &= -2\mathbf{D}^{-2}(\mathbf{y}_i - \boldsymbol{\mu}_r) \\ \frac{\partial d_{ir}}{\partial \alpha_{rk}} &= -(\mathbf{y}_i - \boldsymbol{\mu}_r)^\top \mathbf{D}_r^{-1} (\dot{D}_{rk} \mathbf{D}_r^{-1} + \mathbf{D}_r^{-1} \dot{D}_{rk}) \mathbf{D}_r^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_r) \end{aligned}$$

- $A_{ir}$

$$\begin{aligned} \frac{\partial A_{ir}}{\partial \boldsymbol{\mu}_r} &= -\mathbf{D}_r^{-1} \boldsymbol{\lambda}_r \\ \frac{\partial A_{ir}}{\partial \boldsymbol{\lambda}_r} &= \mathbf{D}_r^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_r) \\ \frac{\partial A_{ir}}{\partial \alpha_{rk}} &= -\boldsymbol{\lambda}_r \mathbf{D}_r^{-1} \dot{D}_{rk} \mathbf{D}_r^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_r) \end{aligned}$$

- $\mathbf{D}_r$

$$\frac{\partial |\mathbf{D}_r^2|^{-1/2}}{\partial \alpha_{rk}} = \frac{-1}{|\mathbf{D}_r|^2} \frac{\partial |\mathbf{D}_r|}{\partial \alpha_{rk}},$$

where  $\dot{D}_{rk} = \frac{\partial \mathbf{D}_r}{\partial \alpha_{rk}}$ .

## *Referências*

- AKAIKE, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, v. 19, p. 716–723, 1974.
- ANDREWS, D. F.; MALLOWS, C. L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, v. 36, p. 99–102, 1974.
- ARELLANO-VALLE, R.; GENTON, M. G. Fundamental skew distributions. *Institute of Statistics Mimeo Series*, 2003.
- ARELLANO-VALLE, R. B.; BOLFARINE, H.; LACHOS, V. H. Skew-normal linear mixed models. *Journal of Data Science*, v. 3, p. 415–438, 2005.
- ARELLANO-VALLE, R. B.; BOLFARINE, H.; VILCA-LABRA, F. Ultrastructural elliptical models. *The Canadian Journal of Statistics*, v. 24, n. 2, p. 207–216, 1996.
- ARELLANO-VALLE, R. B.; GENTON, M. G. On fundamental skew distributions. *Journal of Multivariate Analysis*, v. 96, p. 93–116, 2005.
- AZZALINI, A. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, v. 12, p. 171–178, 1985.
- AZZALINI, A. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, v. 32, p. 159–188, 2005.
- AZZALINI, A.; CAPITANIO, A. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B*, v. 61, p. 579–602, 1999.
- AZZALINI, A.; CAPITANIO, A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society, Series B*, v. 65, p. 367–389, 2003.
- AZZALINI, A.; DALLA VALLE, A. The multivariate skew-normal distribution. *Biometrika*, v. 83, n. 4, p. 715–726, 1996.
- AZZALINI, A.; GENTON, M. G. Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, v. 76, p. 106–129, 2008.
- BASFORD, K. E. et al. Standard errors of fitted component means of normal mixtures. *Computational Statistics*, v. 12, p. 1–17, 1997.

- BAYES, C. L. *Inferência Bayesiana no modelo normal assimétrico*. Dissertação — Instituto de Matemática e Estatística da Universidade de São Paulo, 2005.
- BAYES, C. L.; BRANCO, M. D. Bayesian inference for the skewness parameter. *Brazilian Journal of Probability and Statistics*, To appear, 2007.
- BENSMAIL, H. et al. Inference in model-based cluster analysis. *Statistics and Computing*, v. 7, p. 1–10, 1997.
- BICKEL, P. J.; DOKSUM, K. *Mathematical Statistics – Basic Ideas and Selected Topics*. 2. ed. [S.l.]: Prentice Hall, 2000.
- BIERNACKI, C.; CELEUX, G.; GOVAER, G. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, v. 41, p. 561–575, 2003.
- BÖHNING, D. *Computer-Assisted Analysis of Mixtures and Applications*. [S.l.]: Chapman & Hall/CRC, 2000.
- BRANCO, M. D. et al. Bayesian calibration under a Student-t model. *Computational Statistics*, v. 13, p. 319–338, 1998.
- BRANCO, M. D.; DEY, D. K. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, v. 79, p. 99–113, 2001.
- CABRAL, C. R. B. *Testes de Distância em Modelos de Regressão com Erros nas Variáveis*. Tese (Doutorado) — IME-USP, São Paulo, 2000.
- CABRAL, C. R. B.; BOLFARINE, H. *Bayesian Inference for the Skew Student-t-Normal Model*. [S.l.], 2008.
- CABRAL, C. R. B.; BOLFARINE, H.; PEREIRA, J. R. G. *Bayesian density estimation using skew Student-t-Normal mixtures*. [S.l.], 2008.
- CABRAL, C. R. B.; BOLFARINE, H.; PEREIRA, J. R. G. Bayesian density estimation using skew Student-t-normal mixtures. *Computational Statistics & Data Analysis*, v. 52, p. 5075–5090, 2008.
- CELEUX, G. Bayesian inference for mixtures: the label switching problem. In: PAYNE, R.; GREEN, P. (Ed.). *COMPSTAT 98*. [S.l.]: Physica-Verlag, 1998. p. 227–232.
- CELEUX, G.; HURN, M.; ROBERT, C. P. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, v. 95, n. 451, p. 957–970, 2000.
- CELEUX G. & SOROMENHO, G. An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal*, v. 13, p. 195–212, 1996.

- DAY, N. Estimating the components of a mixture of a two normal distribution. *Biometrika*, v. 56, p. 463–474, 1969.
- 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.
- DIAS, J. G.; WEDEL, M. An empirical comparison of EM, SEM and MCMC performance for problematic gaussian mixture likelihoods. *Statistics and Computing*, v. 14, p. 323–332, 2004.
- DUDA R.O. & HART, P. *Pattern Classification and Scene Analysis*. New York: Wiley, 1988.
- EFRON, B.; TIBSHIRANI, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, v. 1, p. 54–75, 1986.
- FLURY, B.; RIEDWYL, H. *Multivariate statistics, a practical approach*. Cambridge: Cambridge University Press, 1988.
- FRALEY C. & RAFTERY, A. How many clusters? Which clustering method? Answers via model based cluster analysis. *Technical Report of Departament of Statistics: University of Washington*, v. 329, 1998.
- FRANKLIN, J. N. *Matrix Theory*. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1968.
- GARCIA-ESCUADERO, L. A.; GORDALIZA, A. Robustness properties of k means and trimmed means. *Journal o the American Statistical Association*, v. 94, p. 956–969, 1999.
- GENTON, M. G. (Ed.). *Skew-Elliptical Distributions and Their Applications*. [S.l.]: Chapman and Hall, 2004.
- GEWEKE, J. Interpretation and inference in mixture models: simple MCMC works. *Computational Statistics and Data Analysis*, v. 51, p. 3529–3550, 2007.
- GÓMEZ, H. W.; VENEGAS, O.; BOLFARINE, H. Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics*, v. 18, p. 395–407, 2007.
- GUPTA, A. K.; CHANG, F. C.; HUANG, W. J. Some skew-symmetric models. *Random Operators and Stochastic Equations*, v. 10, p. 133–140, 2002.
- HENZE, N. A probabilistic representation of the skew-normal distribution. *Scand. J. Statist.*, v. 13, p. 271–275, 1986.

- HENZE, N. A probabilistic representation of the skew-normal distribution. *Scandinavian Journal of Statistics*, v. 13, p. 271–275, 1986.
- JAMES, B. R. *Probabilidade: um Curso Intermediário*. Rio de Janeiro: IMPA, 1981.
- KARLIS, D.; XEKALAKI, E. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, v. 41, p. 577–590, 2003.
- KULLBACK S. & LEIBER, R. On information and sufficiency. *Annals of Mathematical Statistics*, v. 22, p. 79–86, 1951.
- LACHOS, V. H. et al. Likelihood-based inference for multivariate skew-normal regression models. *Communications in Statistics - Theory and Methods*, v. 36, p. 1769–1786, 2007.
- LACHOS, V. H.; GHOSH, P.; ARELLANO-VALLE, R. B. Likelihood based inference for skew normal independent linear mixed models. *Statistica Sinica*, 2009. To appear.
- LACHOS, V. H.; LABRA, F. V.; GHOSH, P. Multivariate skew-normal/independent distributions: properties and inference. Mimeo. (Departamento de Estatística, Universidade Estadual de Campinas, São Paulo, Brazil). 2008.
- LACHOS, V. H. et al. Robust multivariate measurement error models with scale mixtures of skew-normal distribution. *Statistics*, 2008.
- LANGE, K.; SINSHEIM, J. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, v. 2, p. 175–198, 1993.
- LANGE, K.; SINSHEIMER, J. S. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, v. 2, p. 175–198, 1993.
- LEHMANN, E. L. *Theory of Point Estimation*. New York: John Wiley and Sons, 1983.
- LEHMANN, E. L. *Testing Statistical Hypotheses*. 2. ed. New York: John Wiley and Sons, 1986.
- LIN, T. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 2008. Under Revision.
- LIN, T. I. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, v. 100, p. 257–265, 2009.
- LIN, T. I.; LEE, J. C.; HSIEH, W. J. Robust mixture modelling using the skew t distribution. *Statistics and Computing*, v. 17, p. 81–92, 2007.

- 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.
- LIN, T. I.; LEE, J. C.; YEN, S. Y. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, v. 17, p. 909–927, 2007.
- 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.
- MCLACHLAN, G. J.; KRISHNAN, T. *The EM Algorithm and Extensions*. 2. ed. [S.l.]: John Wiley, 2008.
- MCLACHLAN, G. J.; PEEL, G. J. *Finite Mixture Models*. [S.l.]: John Wiley and Sons, 2000.
- MCLACHLAN G.J. & BASFORD, K. *Mixture Models: Inference and applications to clustering*. New York: Marcel Dekker, 1988.
- MENG, X. L.; RUBIN, D. B. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, v. 80, p. 267–278, 1993.
- NADARAJAH, S.; KOTZ, S. Skewed distributions generated by the normal kernel. *Statistics and Probability Letters*, v. 65, p. 269–277, 2003.
- NADARAJAH, S.; KOTZ, S. Skew distributions generated from different families. *Acta Applicandae Mathematicae*, v. 91, p. 1–37, 2006.
- NELDER, J.; WEDDERBURN, R. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, v. 135, p. 370–384, 1972.
- NITYASUDDHI, D.; BÖHNING, D. Asymptotic properties of the em algorithm estimate for normal mixture models with component specific variances. *Computational Statistics & Data Analysis*, v. 41, p. 591–601, 2003.
- PEEL, D.; MCLACHLAN, G. J. Robust mixture modelling using the t distribution. *Statistics and Computing*, v. 10, p. 339–348, 2000.
- PEREIRA, J. R. G. *Misturas Finitas de Densidades com Aplicações em Reconhecimento Estatístico de Padrões*. Tese — UNICAMP, 2001.
- PROUST, C.; JACQMIN-GADDA, H. Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer Methods and Programs in Biomedicine*, v. 78, p. 165–173, 2005.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008. ISBN 3-900051-07-0. Disponível em: <<http://www.R-project.org>>.

- ROBERT, C. P. Mixtures of distributions: Inference and estimation. In: GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. J. (Ed.). *Markov Chain Monte Carlo in Practice*. [S.l.]: Chapman & Hall, 1996. p. 441–464.
- ROBERT, C. P.; CASELLA, G. *Monte Carlo Statistical Methods*. 2. ed. [S.l.]: Springer, 2004.
- ROEDER K. & WASSERMAN, L. Practical bayesian density estimation using mixture of normals. *Journal of the American Statistical Association*, v. 92(439), p. 894–902, 1997.
- SAHU, S. K.; DEY, D. K.; BRANCO, M. D. A new class of multivariate skew distributions with applications to Bayesian regression models. *The Canadian Journal of Statistics*, v. 31, p. 129–150, 2003.
- SCHWARZ, G. Estimating the dimension of a model. *Annals of Statistics*, v. 6, p. 461–464, 1978.
- SHOHAM, S. Robust clustering by deterministic agglomeration em of mixtures of multivariate t-distributions. *Pattern Recognition*, v. 35, p. 1127–1142, 2002.
- 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.
- SILVA, E. A.; NETO, J. C.; CABRAL, C. Perfil do usuário da vila olímpica. 1993.
- SILVERMAN, B. W. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.
- SMYTH, P. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, v. 10, p. 63–72, 2000.
- STEPHENS, M. *Bayesian methods for mixtures of normal distributions*. Tese — Magdalen College, Oxford, 1997.
- TITTERINGTON D.M., S. A. . M. U. *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1988.
- WANG, H. X. et al. Robust mixture modelling using multivariate t-distribution with missing information. *Pattern Recognition Letters*, v. 25, p. 701–710, 2004.
- WANG, J.; GENTON, M. G. The multivariate skew-slash distribution. *Journal of Statistical Planning and Inference*, v. 136, p. 209–220, 2006.
- 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.

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.

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

ZHAO, L. C.; DOREA, C. C. Y.; GONÇALVES, C. R. On determination of the order of a markov chain. *Statistical Inference for Stochastic Processes*, v. 4, p. 273–282, 2001.