UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

**Ítalo Santos Monteiro**

# Misturas Finitas de Distribuições Hiperbólica Generalizada Normal Assimétrica

Juiz de Fora
2020

**Ítalo Santos Monteiro**

# Misturas Finitas de Distribuições Hiperbólica Generalizada Normal Assimétrica

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Matemática. Área de concentração: Matemática Aplicada

Orientadora: Prof$^a$. Dr$^a$. Camila Borelli Zeller

Juiz de Fora

2020

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

Monteiro, Ítalo Santos.
Misturas Finitas de Distribuições Hiperbólica Generalizada Normal Assimétrica / Ítalo Santos Monteiro. – 2020.
95 f. : il.

Orientadora: Camila Borelli Zeller
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Matemática, 2020.

1. Misturas finitas. 2. Hiperbólica Generalizada Normal Assimétrica. 3. Algoritmo EM. I. Zeller, Camila Borelli, orient. II. Título.

**Ítalo Santos Monteiro**

**Misturas Finitas de Distribuições Hiperbólica Generalizada Normal Assimétrica**

Dissertação apresentada ao Programa de Pós-graduação em Matemática da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Matemática. Área de concentração: Matemática Aplicada

Aprovada em 21 de dezembro de 2020

BANCA EXAMINADORA

Profª. Drª. Camila Borelli Zeller - Orientadora
Universidade Federal de Juiz de Fora

Prof. Dr. Celso Rômulo Barbosa Cabral
Universidade Fed. do Amazonas

Prof. Dr. Luis Benites Sánchez
Pontifica Universidad Católica del Peru

Dedico este trabalho a minha mãe.

## AGRADECIMENTOS

Primeiramente à Deus, por me iluminar e dar forças durante todas as etapas da minha vida.

Agradeço em especial à minha mãe, Rosimeri Castro, que sempre esteve comigo nessa caminhada, me inspirando a crescer tanto como pessoa quanto profissional, nunca deixando me abalar com as dificuldades enfrentadas no decorrer dessa trajetória.

À minha irmã e meu padrasto, Isabela Monteiro e Manoel Carlos Audizio, que estiveram presentes a todo momento, me incentivando e dando forças para continuar.

Aos meus familiares, por serem complacentes quanto a minha ausência nesses anos voltados à vida acadêmica.

Ao meu companheiro e amigo, José Luiz Ernandes, que foi paciente comigo me dando conselhos muitos importantes nessa jornada acadêmica.

As minhas amigas Joseana Frango, Melise de Souza e Clarisse Xavier, que estiveram me dando apoio, sempre acreditando no meu potencial.

Aos meus amigos do mestrado, Mariane Bispo, Rodrigo Barbosa, Jéssica Correia, Julio Lanazca, Juan Sebastian, Paula dos Reis, Marcelo Oliveira, Wilian Rocha, Daniel Moraes, Iris Nascimento e Jhonnatan Carvalho, por me acompanharem nesse momento tão importante da minha vida.

À todos que colaboraram de forma direta ou indireta na realização desta dissertação.

À toda banca examinadora, por terem aceito o convite e pela dedicação para com este trabalho com suas sugestões e correções.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CNPq), pelo apoio financeiro.

E principalmente à minha orientadora, professora Camila Borelli, pela confiança, dedicação, incentivo constantes e amizade ao longo desses anos. Obrigado pelos ensinamentos, disposição e prestatividade, pois sem sua contribuição esse trabalho não poderia ser realizado.

“Se o porco inteiro fosse perfeito, não haveria cachorro quente.” (STEVEN & GREG UNIVERSO).

## RESUMO

Neste trabalho é apresentado um modelo flexível de misturas finitas de densidades, que tem como base a classe de distribuições hiperbólica generalizada normal assimétrica, no qual é determinado algumas propriedades e resultados importantes, para que posteriormente seja realizado a estimação por máxima verossimilhança dos parâmetros presentes no modelo proposto. Para isso, foi utilizado o algoritmo EM obtendo as estimativas dos parâmetros do modelo de maneira iterativa. Além disso, é discutido alguns casos particulares do modelo proposto, incluindo comentários adicionais sobre a implementação desse algoritmo. Também são projetados estudos de simulação, bem como quatro aplicações à dados reais, que ilustram o comportamento do modelo proposto e os resultados inferenciais desenvolvidos. Os resultados obtidos mostraram uma forte possibilidade no uso desse modelo em análises estatísticas envolvendo dados que apresentam significante multimodalidade e comportamentos não gaussianos, como assimetria e caudas mais pesadas.



Palavras-chave: Misturas finitas. Hiperbólica Generalizada Normal Assimétrica. Algoritmo EM.

**ABSTRACT**

In this work, a flexible model of finite density mixtures is presented, based on the class of asymmetric generalized hyperbolic distributions, in which some important properties and results are determined, so that an estimation study by maximum likelihood of the parameters is subsequently carried out. present in the proposed model. For this, the EM algorithm was used, obtaining the estimates of the model parameters in an iterative way. In addition, some particular cases are discussed, including additional comments on the implementation of this algorithm. A simulation study is also designed to evaluate its performance, as well as four applications to real data, which illustrate the behavior of the model. The results obtained showed a strong possibility in the use of this model in statistical analyzes involving data that present significant multimodality and non-Gaussian behaviors, such as asymmetry and heavier tails.

Keywords: Finite mixtures. *Skew*-Normal Generalized Hyperbolic. EM algorithm.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## SUMÁRIO

# 1 INTRODUÇÃO

Ao longo dos anos, o uso de ferramentas flexíveis capazes de modelar dados, em análise estatística, vem recebendo atenção devido sua gama crescente de aplicações em diversas áreas do conhecimento. Dentre os principais modelos nesse âmbito, destacam-se os modelos de misturas finitas de densidades. Em consequência disso tem-se o aumento no número de produções científicas utilizando modelos de misturas finitas de densidades, tanto na literatura estatística quanto em outras áreas das ciências, evidenciando o uso dessa técnica na modelagem de dados. Esse crescimento ocorreu após a publicação do trabalho sobre misturas finitas de McLachlan e Basford (1988) [60], onde apresentaram aplicações de modelos de misturas à vários problemas que anteriormente foram tratados sob outras perspectivas [17, 28].

Com isso, muitos autores ficaram interessados nas pesquisas envolvendo misturas finitas de densidades, originando diversos trabalhos como os de Aitkin & Aitkin (1996) [3], Shoham (2002) [73], Böhning et al. (2007) [22], entre outros, com temas propondo métodos de estimação dos parâmetros do modelo de interesse, extensões das ferramentas adotadas, discussões sobre o problemas de identificabilidade, ajustes com auxílio de algoritmos e métodos computacionais, propriedades relacionadas aos estimadores de máxima verossimilhança, determinação do número de componentes a serem usados nas misturas de distribuições, e muitos outros [17].

Do ponto de vista prático, o modelo de misturas finitas de densidades normais é sem dúvidas o mais adotado nas aplicações que aparecem na literatura. Isso porque os modelos de misturas finitas possuem a capacidade de representar densidades de alta complexidade, em particular, para qualquer distribuição multivariada é possível construir aproximações por meio de uma mistura finita de densidades normais [62]. Além dessa característica, considera-se também a simplicidade algébrica envolvida na distribuição normal, que no passado era necessária por causa da falta de alternativas computacionalmente acessíveis.

No que diz respeito ao desenvolvimento de métodos computacionais para modelos de misturas finitas, somente nos últimos vinte anos que consideráveis avanços foram alcançados, especialmente no contexto do método de estimação por máxima verossimilhança. Na literatura, durante a década de 60, autores como Cohen (1967) [27], Wolfe (1967) [80] e Day (1969) [32] apresentaram trabalhos que discutiram formas e teorias com relação aos ajustes dos modelos de misturas finitas. Posteriormente, o trabalho de Dempster et al. (1977) [33], sobre o algoritmo EM, simplificou os ajustes de modelos de misturas finitas, propondo a estimação por máxima verossimilhança no contexto de dados incompletos, incentivando e disseminando o interesse em usar tais modelos para analisar dados com presença de heterogeneidade populacional [61].

Os modelos de misturas finitas de densidades são bastantes úteis na modelagem de

dados com a presença de multimodalidade, quando tem-se conhecimento que as observações pertencem a subpopulações distintas, porém não sabe-se como discrimina-lás. Embora seu uso nesse contexto seja bastante atrativo, existe ainda a necessidade de investigar as suposições distribucionais das componentes de mistura, já que os dados também podem apresentar um comportamento assimétrico e/ou caudas mais pesadas [17, 31].

Uma abordagem alternativa às misturas finitas de gaussianas é a utilização de modelos de misturas finitas capazes de modelar dados heterogêneos preservando a estrutura assimétrica e de caudas mais pesadas, assim como possibilitar a redução da influência de *outliers* e permitir a concentração de observações em torno do valor central. Como exemplo, tem-se os modelos de misturas finitas baseados na classe de distribuições mistura escala normal assimétrica. Visto que, tal classe contém casos particulares como as distribuições $t$ assimétrica, *slash* assimétrica e normal assimétrica contaminada, as quais possuem caudas mais pesadas que os modelos de misturas finitas de normais e normal assimétrica, que consequentemente são mais robustas na presença de valores extremos [17, 23, 31].

Assim, neste trabalho, devido à considerável pesquisa que tem sido feita para introduzir modelos paramétricos flexíveis que acomodem multimodalidade e desvios da suposição de normalidade e então, amenizem a necessidade de transformações dos dados, considera-se uma outra classe de distribuições assimétricas, a distribuição Hiperbólica Generalizada Normal Assimétrica; veja Vilca et al. (2014) [78] para mais detalhes. Dessa forma, propõe-se as misturas finitas de distribuições Hiperbólica Generalizada Normal Assimétrica.

## 1.1 MOTIVAÇÃO

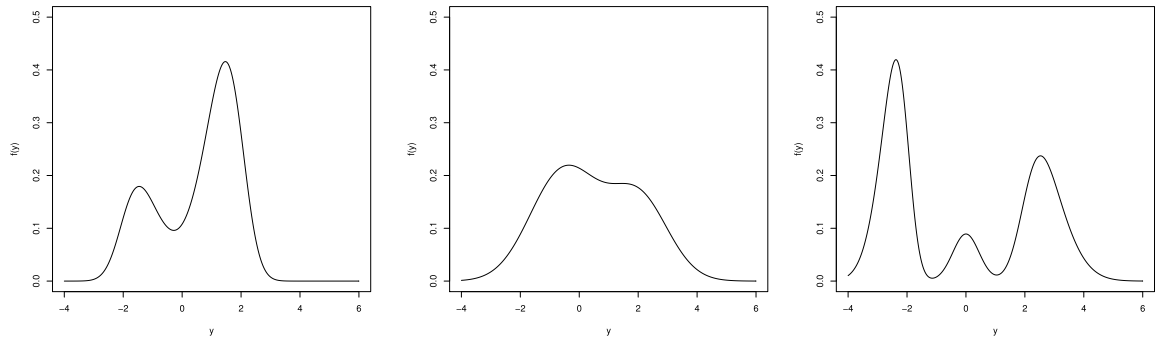
A modelagem de dados baseada em modelos de misturas finitas de densidades apresenta uma estrutura que permite essencialmente aplicações interdisciplinares em diversas áreas do conhecimento, como biometria, agricultura, biologia, ciências médicas, genética, economia, entre outras, pois são capazes de capturar características importantes presentes em alguns conjuntos de dados reais, como multimodalidade, assimetria, curtose e heterogeneidade não observada [31].

Dessa maneira, os modelos de misturas finitas podem ser considerados extremamente flexíveis. Por exemplo, na Figura 1 pode-se observar diferentes comportamentos das curvas de densidades provenientes do modelo de misturas finitas de normal assimétrica (veja Dávila (2018) [31]), ao assumir os seguintes valores para os parâmetros: $\mu_1 =$ -2, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 = 1$, $\lambda_1 = 2$, $\lambda_2 =$ -2 e $\rho_1 =$ 0,3 à esquerda; $\mu_1 =$ -1, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, $\lambda_1 = \lambda_2 = \lambda_3 = 0$ e $\rho_1 = \rho_2 =$ 0,3 ao centro; $\mu_1 =$ -2, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 =$ 0,5, $\sigma_2^2 =$ 0,2, $\sigma_3^2 = 1$, $\lambda_1 =$ -2, $\lambda_2 = 0$, $\lambda_3 = 2$, $\rho_1 =$ 0,5 e $\rho_2 =$ 0,1 à direita.

Recentemente, Basso (2010) [17], Browne & McNicholas (2015) [24], Wang et al. (2018) [79] e Zeller et al. (2018) [84] apresentaram trabalhos utilizando diferentes

modelos de misturas finitas, mostrando sua grande flexibilidade e aplicabilidade em distintos cenários. Sendo assim, esta dissertação de mestrado é motivada pelo fato de existir vários conjuntos de dados que são considerados na literatura, que apresentam características de significante multimodalidade e comportamentos não gaussiano, tais como assimetria e caudas pesadas. Nesta linha de pesquisa, serão desenvolvidos nesse trabalho alguns resultados adicionais para a classe de distribuições hiperbólica generalizada normal assimétrica [78], acopladas ao modelo de misturas finitas.

Figura 1 – Curvas densidades provenientes do modelo de misturas finitas de normais assimétricas univariada.




Fonte: O autor (2020).

## 1.2 OBJETIVOS

Conforme exposto neste trabalho, observa-se que a literatura indica resultados satisfatórios quando se trata da utilização de modelos de misturas finitas para análise estatística de dados que apresentam significante multimodalidade e comportamento não gaussiano, como assimetria e caudas pesadas, mostrando também robustez na acomodação de valores extremos. Assim, esta dissertação de mestrado se propõe em apresentar um modelo de misturas finitas flexível e capaz de analisar dados com tais características, bem como estender trabalhos que utilizam de classes de distribuições nesse sentido.

### Objetivos gerais

O objetivo geral proposto para este trabalho é apresentar a construção de um novo modelo de misturas finitas de densidades, que tem como base a classe de distribuições hiperbólica generalizada normal assimétrica, que foi proposta inicialmente por Vilca et al. (2014) [78] e estendida por Calegari (2020) [25]. Desta maneira, apresenta-se a definição do modelo de misturas finitas quando as densidades seguem uma distribuição hiperbólica generalizada normal assimétrica, uma análise gráfica ilustrando sua flexibilidade, tanto no contexto univariado como multivariado, e por fim a representação hierárquica que facilita a implementação do algoritmo EM, útil para obter os estimadores de máxima verossimilhança dos parâmetros presentes no modelo proposto.

**Objetivos específicos**

Como objetivos específicos, neste trabalho propõe-se o desenvolvimento e implementação do algoritmo EM no contexto do modelo de misturas finitas de distribuições hiperbólica generalizada normal assimétrica. Por fim, são apresentados os estudos de simulação e quatro aplicações a conjuntos de dados reais conhecidos na literatura, com a finalidade de ilustrar o modelo proposto e os resultados inferenciais desenvolvidos.

## 1.3 ORGANIZAÇÃO DO TRABALHO

Tendo em vista alcançar os objetivos definidos anteriormente, o presente trabalho é distribuído em cinco capítulos. Este capítulo, Introdução, teve como principal função realizar uma breve revisão da literatura sobre modelos de misturas finitas de uma maneira geral, bem como uma motivação para utilizar esses modelos nesta dissertação de mestrado.

No segundo capítulo são descritos os conceitos fundamentais para construção deste trabalho, onde serão apresentados definições e propriedades importantes das classes de distribuições mistura escala normal, mistura escala normal assimétrica e a classe de distribuições hiperbólica generalizada normal assimétrica, assim como as distribuições normal assimétrica e gaussiana inversa generalizada, que foram utilizadas para construção do modelo proposto.

O terceiro capítulo pode ser considerado como o objetivo principal desse trabalho. Este capítulo inicia-se com uma discussão sobre resultados referentes aos modelos de misturas finitas em um contexto geral, onde são apresentados definições e propriedades acerca desses modelos, bem como o problema de identificabilidade, a estrutura de dados incompletos e a construção do algoritmo EM nesse contexto. Em seguida, apresenta-se o modelo fundamental proposto para esse trabalho, isto é, o de misturas finitas de distribuições hiperbólica generalizada normal assimétrica, juntamente com uma análise gráfica, sua representação hierárquica e algoritmo EM para estimação dos parâmetros, no caso univariado e multivariado. Por fim, também serão discutidos alguns casos particulares, incluindo comentários adicionais sobre o algoritmo EM.

Na sequência, no quarto capítulo, os resultados obtidos foram aplicados em conjuntos de dados reais ou simulados. Por fim, quatro aplicações desse modelo são realizadas, tanto no contexto univariado como no multivariado, comparando os resultados gerados, em termos de ajuste e classificação, com outros modelos conhecidos na literatura estatística de misturas finitas. Verifica-se também a adequação dos modelos aos dados inspecionando alguns critérios de informação.

No quinto e último capítulo serão feitas considerações acerca da utilidade do modelo de misturas finitas proposto, com base nos resultados obtidos no decorrer do desenvolvimento deste trabalho, assim como algumas diretrizes para estudos futuros.

# 2 CONCEITOS PRELIMINARES

No presente capítulo, antes de iniciar os estudos envolvendo misturas finitas de distribuições, será necessário destacar formalmente alguns conceitos fundamentais que devem ser estabelecidos, bem como algumas distribuições de interesse que serão utilizadas ao longo do desenvolvimento deste trabalho.

Portanto, na Seção 2.1 será feita uma revisão sucinta a respeito da classe de distribuições de Misturas de Escala Normal (SMN). Em seguida, a Seção 2.2 traz um estudo direcionado sobre algumas das propriedades e resultados referentes à classe de distribuições Misturas de Escala Normal Assimétrica (SMSN), bem como algumas características com relação a distribuição Normal Assimétrica (SN), ambas no contexto multivariado. Na Seção 2.3 são descritas informações importantes sobre a classe de distribuições Hiperbólica Generalizada Normal Assimétrica (SNGH), principalmente acerca da distribuição Gaussiana Inversa Generalizada (GIG), pois será utilizada nos demais capítulos para compor o modelo de misturas finitas de densidades SNGH.

## 2.1 CLASSE DE DISTRIBUIÇÕES DE MISTURAS DE ESCALA NORMAL

A classe de distribuições de misturas de escala normal foi proposta por Andrews & Mallows (1974) [5] como uma extensão paramétrica do modelo normal, resultando uma nova classe de distribuições que procura tratar dados com valores extremos preservando sua estrutura simétrica [58]. Sendo assim, a representação estocástica da classe de distribuições SMN é definida da seguinte forma.

**Definição 2.1.1.** *Considere* $\mathbf{Y}$ *um vetor aleatório contínuo. Tal vetor tem distribuição na família de mistura de escala normal (SMN), se ela admite a representação estocástica a seguir*

$$\mathbf{Y} = \boldsymbol{\mu} + [\kappa(U)]^{1/2}\mathbf{Z}, \tag{2.1}$$

*onde* $\boldsymbol{\mu}$ *é o parâmetro de locação,* $U$ *uma variável aleatória não-negativa com função densidade de probabilidade* $h(u;\boldsymbol{\nu})$ *e função de distribuição acumulada* $H(u;\boldsymbol{\nu})$*, indexadas pelo parâmetro* $\boldsymbol{\nu}$ *que controla as caudas da distribuição, podendo ser conhecido (fixo) ou não, independente de* $\mathbf{Z}$*,* $\kappa(U)$ *uma função positiva de* $U$ *e* $\mathbf{Z}$ *é um vetor aleatório que segue a distribuição normal com vetor de média* $\mathbf{0}$ *e matriz de variância e covariância* $\boldsymbol{\Sigma}$*.*

Assim, utilizando da representação estocástica dada em (2.1), pode-se obter a função densidade de probabilidade para a variável aleatória $\mathbf{Y}$ da seguinte maneira.

**Definição 2.1.2.** *Seja* $\mathbf{Y}$ *um vetor aleatório p-dimensional. Se* $\mathbf{Y}$ *tem distribuição na classe SMN, então sua função densidade de probabilidade é dada por*

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_0^\infty \phi_p(\mathbf{y};\boldsymbol{\mu},\kappa(u)\boldsymbol{\Sigma})dH(u;\boldsymbol{\nu}), \tag{2.2}$$

*em que* $\phi_p(\cdot;\boldsymbol{\mu},\kappa(u)\boldsymbol{\Sigma})$ *é a densidade da normal p-variada,* $U = u$ *é o fator de escala e* $H(u;\boldsymbol{\nu})$ *distribuição de mistura. Em termos de notação, o vetor aleatório p-dimensional* $\boldsymbol{Y}$ *com distribuição SMN e parâmetros definidos acima, é denotado por* $\boldsymbol{Y} \sim \text{SMN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\nu})$.

Vale ressaltar que, se assumir $\kappa(U) = U^{-1}$, a distribuição de $\mathbf{Y}$ ficará restrita à classe de distribuições Normal Independente (NI), como discutido em Lange & Sinsheimer (1993) [50]. Distribuições pertencentes à classe NI possuem a capacidade de acomodar caudas pesadas, ou seja, amplitudes que gradativamente decrescem assintoticamente. Com essa característica, ao considerar distintas distribuições para $U$, tal classe inclui casos particulares como, as versões multivariadas das distribuições *t-Student*, Slash, Pearson Tipo VII, Normal Contaminada, entre outras [25, 29].

Por ser uma classe de distribuições capaz de acondicionar *outliers*, isto é, observações que apresentam um excessivo afastamento em comparação aos outros dados, a classe SMN vem sendo amplamente utilizada ao longo dos anos na literatura. Peel e McLachlan (2000) [65], por exemplo, usaram esta classe de distribuições para descrever uma modelagem mais robusta, com relação aos *outliers*, em misturas finitas de distribuições *t-Student* multivariadas.

Sendo assim, da mesma maneira que a classe SMN que procura tratar dados com valores extremos no contexto simétrico, a classe de distribuições proposta na seção a seguir, se propõe em tratar dados extremos também, porém num contexto assimétrico.

## 2.2 CLASSE DE DISTRIBUIÇÕES DE MISTURAS DE ESCALA NORMAL ASSIMÉTRICA

Conforme mencionado anteriormente, nesta seção serão apresentadas algumas propriedades e resultados interessantes que auxiliam na definição das distribuições Misturas de Escala Normal Assimétrica. Entretanto, inicialmente, será necessário realizar uma revisão sobre alguns conceitos fundamentais referentes à distribuição normal assimétrica multivariada, uma vez que, tal distribuição é usada para a construção da classe de distribuições SMSN.

### 2.2.1 **Distribuição normal assimétrica**

Os estudos envolvendo a construção de novas famílias de distribuições que sejam capazes de incorporar assimetria são de amplo interesse na literatura atual de tal maneira que tornaram-se uma grande motivação para pesquisadores nos últimos anos. Sob essa motivação, Azzalini (1985) [10] propôs a distribuição normal assimétrica [29].

A distribuição normal assimétrica é uma extensão da distribuição Normal, onde é permitido incorporar um parâmetro de assimetria. Proposta, inicialmente de forma univariada, na qual mais tarde foi generalizada para o caso multivariado por Azzalini e

Dalla-Valle (1996) [13], fez com que uma vasta quantidade de trabalhos surgissem sobre esse assunto na literatura. Para uma discussão mais ampla e detalhada sobra tal distribuição e seu resultados, veja Azzalini (2005) [11].

De acordo com alguns trabalhos presentes na literatura, como os de Arellano-Valle & Del Pino (2004) [7], Arellano-Valle & Genton (2005) [8] e Arellano-Valle, Bolfarine e Lachos (2005) [6], existem inúmeras definições de distribuições normal assimétrica, com diferentes parametrizações e interpretações [31, 83]. Dessa forma, neste trabalho será considerada uma definição unificada das definições apresentadas nos trabalhos citados anteriormente. Tal decisão se deve ao fato que, ao realizar manipulações algébricas acerca das propriedades e caracterização poderão ser obtidas de maneira mais simples. Além disso, o caso univariado pode ser visto como um caso particular derivado da representação multivariada [17].

Posto isto, denota-se por $\phi_p(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ a função densidade de probabilidade da normal $p$-variada, com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, e $\Phi_p(\cdot)$ a função de distribuição acumulada da normal padrão. No caso em que $\boldsymbol{\mu}=\mathbf{0}$ e $\boldsymbol{\Sigma}=\mathbf{I}_p$, onde $\mathbf{I}_p$ é a matriz identidade $p\times p$, essas funções, avaliadas em $\mathbf{y}$, são reescritas como $\phi_p(\mathbf{y})$ e $\Phi_p(\mathbf{y})$, respectivamente.

**Definição 2.2.1.** *Seja* $\mathbf{Y}$ *um vetor aleatório p-dimensional. É dito que* $\mathbf{Y}$ *segue uma distribuição normal assimétrica (SN), onde* $\boldsymbol{\mu}\in\mathbb{R}^p$ *é o vetor de locação,* $\boldsymbol{\Sigma}$ *a matriz de variância-covariância, definida positiva e* $\boldsymbol{\lambda}\in\mathbb{R}^p$ *o vetor de assimetria, quando a sua função densidade de probabilidade é dada por*

$$f_{\mathbf{Y}}(\mathbf{y})=2\phi_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})),\quad \mathbf{y}\in\mathbb{R}^p. \tag{2.3}$$

*em que* $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}=\boldsymbol{\Sigma}$ *e* $\boldsymbol{\Sigma}^{1/2}=\mathbf{SDS^{-1}}$*, com* $\mathbf{S}$ *sendo uma matriz não-singular e* $\mathbf{D}$ *uma matriz diagonal com elementos não-negativos na diagonal [52]. E será denotada por* $\mathbf{Y}\sim \mathrm{SN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda})$*. Quando* $\boldsymbol{\mu}=\mathbf{0}$ *e* $\boldsymbol{\Sigma}=\mathbf{I}_p$*, tem-se que*

$$f_{\mathbf{Y}}(\mathbf{y})=2\phi_p(\mathbf{y})\Phi(\boldsymbol{\lambda}^\top\mathbf{y}),\quad \mathbf{y}\in\mathbb{R}^p, \tag{2.4}$$

*e nesse caso, denota-se como* $\mathbf{Y}\sim \mathrm{SN}_p(\boldsymbol{\lambda})$*.*

A seguir é denotada a representação estocástica de um vetor aleatório que segue distribuição normal assimétrica $p$-dimensional.

**Definição 2.2.2.** *Seja* $\mathbf{Y}$ *um vetor aleatório com* $\mathbf{Y}\sim \mathrm{SN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda})$*, então admite-se a seguinte representação estocástica*

$$\mathbf{Y}=\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\delta}|T_0|+(\mathbf{I}_p-\boldsymbol{\delta}\boldsymbol{\delta}^\top)^{1/2}\mathbf{T}_1), \tag{2.5}$$

*onde* $\boldsymbol{\delta}=\dfrac{\boldsymbol{\lambda}}{(1+\boldsymbol{\lambda}^T\boldsymbol{\lambda})^{1/2}}$*,* $|T_0|$ *é o valor absoluto de uma variável aleatória, em que* $T_0\sim \mathrm{N}_1(0,1)$ *e* $\mathbf{T}_1\sim \mathrm{N}_1(\mathbf{0},\mathbf{I}_p)$ *sejam independentes.*

Após definir a distribuição normal assimétrica, assim como sua representação estocástica, o próximo resultado apresenta o valor esperado e a variância de $\mathbf{Y} \sim \mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$.

**Proposição 2.2.1.** *Seja* $\mathbf{Y} \sim \mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. *Então, os momentos de* $\mathbf{Y}$ *são dado por*

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

*b)* $Var[\mathbf{Y}] = \boldsymbol{\Sigma} + \frac{2}{\pi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^{\top}\boldsymbol{\Sigma}^{1/2}$,

*em que* $\boldsymbol{\delta} = \dfrac{\boldsymbol{\lambda}}{\left(1 + \boldsymbol{\lambda}^{\top}\boldsymbol{\lambda}\right)^{1/2}}$.

*Demonstração.* A prova segue da Proposição 2.14 que encontra-se no trabalho de Dávila (2004) [30], onde fazendo a transformação linear $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Z}$ com $\mathbf{Z} \sim \mathrm{SN}_p(\boldsymbol{\lambda})$, conclui-se os resultados. □

Portanto, com esta breve discussão sobre a distribuição normal assimétrica e algumas propriedades de interesse para o trabalho, na seção seguinte serão utilizados esses resultados para apresentar uma revisão sobre a distribuição SMSN.

### 2.2.2 Distribuições misturas de escala normal assimétrica

Na seção anterior, foi apresentada a distribuição normal assimétrica no contexto multivariado, tal distribuição, apesar de bastante utilizada, não se mostra adequada para análises de dados com valores extremos, uma vez que pode comprometer as estimações dos parâmetros em virtude disso. Sendo assim, para contornar esse problema, utiliza-se da classe de distribuições de Mistura de Escala Normal Assimétrica, que foi proposta por Branco & Dey (2001) [23] como uma extensão da classe de mistura escala normal.

Neste sentido, a classe de distribuições SMSN pode ser vista como mais abrangente que outros modelos, propondo ser uma classe de distribuições mais robusta que procura tratar na prática dados com valores extremos, evidências amostrais de assimetria e/ou caudas pesadas [17]. Desta forma, a seguir será apresentada a definição das distribuições mistura de escala normal assimétrica no contexto multivariado, junto de alguns resultados, como sua representação estocástica e algumas de suas propriedades que foram importantes no desenvolvimento de todo esse trabalho.

**Definição 2.2.3.** *Seja* $\mathbf{Y}$ *um vetor aleatório contínuo. É dito que* $\mathbf{Y}$ *tem distribuição na família de mistura de escala normal assimétrica (SMSN), se admite a seguinte representação estocástica*

$$\mathbf{Y} = \boldsymbol{\mu} + [\kappa(U)]^{1/2}\mathbf{Z}, \tag{2.6}$$

*em que $\boldsymbol{\mu}$ é o parâmetro de locação, $U$ uma variável aleatória não-negativa com função densidade de probabilidade $h(u;\boldsymbol{\nu})$ e função de distribuição acumulada $H(u;\boldsymbol{\nu})$, indexadas pelo parâmetro $\boldsymbol{\nu}$ que controla as caudas da distribuição, podendo ser conhecido (fixo) ou não, e independente de $\mathbf{Z}$, $\kappa(U)$ é uma função positiva de $U$ e $\mathbf{Z}$ é um vetor aleatório que segue uma distribuição normal assimétrica com vetor de média $\mathbf{0}$, matriz de variância e covariância $\boldsymbol{\Sigma}$ e vetor de assimetria $\boldsymbol{\lambda}$. E será denotada por $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$.*

Note que, quando $\boldsymbol{\lambda} = \mathbf{0}$, obtém-se a classe de distribuições SMN, definida na Seção 2.1. Vale ressaltar que, se assumir $\kappa(U) = U^{-1}$, a distribuição de $\mathbf{Y}$, ficará restrita à classe de distribuições Normal Assimétrica Independente (SNI), como discutido em Lachos & Vilca (2007) [48].

Como consequência, utiliza-se da representação estocástica dada em (2.6), para definir a função densidade de probabilidade para a vetor aleatório $\mathbf{Y}$ da seguinte maneira.

**Definição 2.2.4.** *Considere $\mathbf{Y}$ um vetor aleatório p-dimensional. Se $\mathbf{Y}$ tem distribuição na classe SMSN, então sua função densidade de probabilidade é dada por*

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_0^\infty \phi_p(\mathbf{y}; \boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma}) \Phi\left( [\kappa(u)]^{-1/2} C \right) dH(u; \boldsymbol{\nu}), \tag{2.7}$$

*onde $U = u$ é o fator de escala, $C = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})$ e $H(u;\boldsymbol{\nu})$ distribuição de mistura. Em termos de notação, o vetor aleatório p-dimensional $\boldsymbol{Y}$ com distribuição* SMSN *e parâmetros definidos acima, é denotado por $\boldsymbol{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, e quando $\boldsymbol{\mu} = \mathbf{0}$ e $\boldsymbol{\Sigma} = \mathbf{I}_p$, tem-se a distribuição* SMSN *padrão e denotada por $\text{SMSN}_p(\boldsymbol{\lambda}, \boldsymbol{\nu})$.*

A seguir são apresentadas algumas propriedades importantes da classe de distribuições SMSN, onde tais resultados podem ser derivados facilmente da representação estocástica dada em (2.6), com auxílio de transformações de variáveis aleatórias [48]. Assim, pode-se obter as formas gerais do vetor de médias e da matriz de covariância, como será mostrado na proposição a seguir.

**Proposição 2.2.2.** *Seja $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$. Então tem-se que:*

*a) Se $E\left[(\kappa(U))^{1/2}\right] < \infty$, então $E[\mathbf{Y}] = \boldsymbol{\mu} + \sqrt{\dfrac{2}{\pi}} \kappa_1 \boldsymbol{\Delta}$,*

*b) Se $E\left[\kappa(U)\right] < \infty$, então $Var[\mathbf{Y}] = \kappa_2 \boldsymbol{\Sigma} + \dfrac{2}{\pi} {\kappa^2}_1 \boldsymbol{\Delta}\boldsymbol{\Delta}^\top$,*

*com $\kappa_m = E\left[(\kappa(U))^{m/2}\right]$, $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ e $\boldsymbol{\delta} = \dfrac{\boldsymbol{\lambda}}{\left(1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda}\right)^{1/2}}$.*

*Demonstração.* A prova segue pela representação estocástica definida em (2.6) junto das propriedades condicionais de esperança e variância. Para mais detalhes sobre a demonstração, veja em Basso (2009) [17] e Calegari (2020) [25]. □

O resultado seguinte será útil para a obtenção do valor esperado de uma transformação da variável mistura $U$ condicionado aos dados observados.

**Lema 2.2.1.** *Considere* $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$*, com representação dada em (2.6) e uma função mensurável* $g : \mathbb{R} \to \mathbb{R}$*. Então,*

$$E\left[g(U)|\mathbf{Y} = \mathbf{y}\right] = 2\frac{f_0(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}\int_0^\infty g(u)\Phi\left(\left[\kappa(u)\right]^{-1/2} C\right) h_0(u|\mathbf{y_0})du, \tag{2.8}$$

*onde* $C = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$ *e* $f_0(\cdot)$ *sendo a função densidade de probabilidade de* $\mathbf{Y}_0 \sim \text{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$*.*

*Demonstração.* A prova segue do Lema A.1.1 em Zeller (2009) [83]. □

A proposição a seguir é importante para a implementação da metodologia que será utilizada no próximo capítulo, pois apresenta o cálculo de alguns momentos que serão úteis na estimação dos parâmetros do modelo de interesse via algoritmo EM.

**Proposição 2.2.3.** *Seja* $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ *e* $\mathbf{U} \sim H$ *o fator de mistura de escala. Então, os seguintes valores esperados são obtidos*

$$\begin{aligned} a_r &= E\left[(\kappa(U))^{-r}|\mathbf{y}\right] \\ &= 2\frac{f_0(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}E\left[(\kappa(U_{\mathbf{y}}))^{-r}\,\Phi\left((\kappa(U_{\mathbf{y}}))^{-1/2} C\right)\right], \qquad (2.9) \\ b_r &= E\left[(\kappa(U_{\mathbf{y}}))^{-r/2}\, S\left((\kappa(U_{\mathbf{y}}))^{-1/2} C\right)|\mathbf{y}\right] \\ &= 2\frac{f_0(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}E\left[(\kappa(U_{\mathbf{y}}))^{-r/2}\,\phi\left((\kappa(U_{\mathbf{y}}))^{-1/2} C\right)\right], \qquad (2.10) \end{aligned}$$

*com* $C$ *definido anteriormente,* $S(s) = \dfrac{\phi(s)}{\Phi(s)}$*, tal que* $s \in \mathbb{R}$*, em que* $U_{\mathbf{y}} = U/\mathbf{Y}_0 = \mathbf{y}$ *e* $f_0(\mathbf{y})$ *sendo a função densidade de probabilidade de* $\mathbf{Y}_0 \sim \text{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$*.*

*Demonstração.* A prova segue como consequência direta da aplicação do Lema 2.2.1, com $g(u) = (\kappa(u))^{-r}$ para $a_r$ e $g(u) = (\kappa(u))^{-r/2} S\left((\kappa(u))^{-1/2} C\right)$ para $b_r$. □

Na próxima seção é apresentada uma outra classe de distribuições assimétricas, onde se baseia o objetivo deste trabalho.

## 2.3 CLASSE DE DISTRIBUIÇÕES HIPERBÓLICA GENERALIZADA NORMAL ASSIMÉTRICA

Nas Seções 2.1 e 2.2, foram apresentadas as representações estocásticas de um vetor aleatório $\mathbf{Y}$ com distribuição pertencente à classe de distribuições SMN e SMSN, como mostrado em (2.1) e (2.6), respectivamente. Desta forma, muitos resultados para a classe de distribuições hiperbólica generalizada normal assimétrica podem ser obtidos através das

propriedades da distribuição normal, membro particular da classe SMN e da distribuição normal assimétrica, da classe SMSN. Note que, na literatura é comum encontrar trabalhos em que consideram $\kappa(U) = {}^1\!/\!_U$ com $\mathbf{Z} \sim \mathrm{SN}_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ ou $\mathbf{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ e $\kappa(U) = U$ com $\mathbf{Z} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. No primeiro e segundo caso, quando $U \sim \mathrm{Gamma}({}^\nu\!/\!_2, {}^\nu\!/\!_2)$, obtém-se a *skew-t* ou sua versão simétrica *t-Student* e por exemplo, se $U \sim \mathrm{Beta}(\nu, 1)$, tem-se a *skew-slash* e a slash, todas membros da classe de distribuições SMSN, definida na Seção 2.2. Já no terceiro caso, ou seja, $\kappa(U) = U$ com $\mathbf{Z} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, quando $U$ segue uma distribuição Gaussiana Inversa Generalizada (GIG), tem-se as distribuições hiperbólicas simétricas. Em especial, neste contexto, se $U$ tem distribuição Inversa Gaussiana (IG), são obtidas as distribuições Normal Inversa Gaussiana (NIG) simétricas. A família de distribuições hiperbólicas generalizadas tem encontrado muitas aplicações nas áreas de finanças e economia, como apresentado em Prause (1999) [67] e Eberlein & Prause (2002) [35], por ser uma distribuição flexível e com características leptocúrticas.

Contudo, existe um grande desafio quando utiliza-se de extensões em (2.6) no contexto em que $U$ segue uma GIG, uma vez que a densidade de $\mathbf{Y}$ é bastante complexa e sua aplicação num contexto prático é bem limitada. Por esse motivo, ao levar em consideração $\kappa(U)$ como ${}^1\!/\!_U$ ou $U$ se torna uma boa alternativa, pois ao considerar $\kappa(U)$ da forma

$$\kappa(U) = U^a,$$

com $a \in \mathbb{R}$, pode-se obter uma função densidade de probabilidade facilmente para $\mathbf{Y}$, desde que $U$ seja uma distribuição com $U = V^{1/a}$, em que $V$ segue uma distribuição GIG. Sob esta motivação descrita acima, Vilca et al. (2014) [78] propõe a classe de distribuições Hiperbólica Generalizada Normal Assimétrica e posteriormente, Calegari (2020) [25] deriva algumas propriedades desta nova classe de distribuições e desenvolve um método de estimação dos parâmetros do modelo.

Desta forma, antes de apresentar as propriedades fundamentais sobre a distribuição SNGH que serão utilizados na construção do objetivo principal desse trabalho, faz-se necessário realizar uma revisão teórica de alguns conceitos importantes envolvendo a distribuição gaussiana inversa generalizada.

### 2.3.1 **Distribuição gaussiana inversa generalizada**

A distribuição gaussiana inversa generalizada, atualmente, amplamente utilizada em estudos com aplicações nas áreas de finanças, linguística estatística, geoestatística e dentre outras, foi proposta pela primeira vez por Halphen (1941) [39] para descrever a frequência de fluxos de rios. Contudo, devido à sua complexidade envolvendo a função de Bessel, trabalhos com essa distribuição ficaram estagnados por anos. Apesar disso, tal distribuição foi redescoberta por Barndorff-Nielsen (1978) [15] e Jørgensen (1982)

[43] discutiu sobre as propriedades estatísticas desse modelo em suas notas de aula [25]. Informações sobre a função Bessel podem ser encontradas no Apêndice A desta dissertação.

Em teoria de probabilidade e estatística, a distribuição gaussiana inversa generalizada, comumente denominada como GIG, é uma família de distribuições contínuas com três parâmetros e sua função densidade é definida a seguir.

**Definição 2.3.1.** *Seja U uma variável aleatória positiva com distribuição gaussiana inversa generalizada, então sua função densidade de probabilidade é definida por*

$$f_U(\mathrm{u};\eta,\psi,\gamma)=\frac{\left(\gamma/\psi\right)^{\eta/2}}{2K_\eta\left(\sqrt{\psi\gamma}\right)}\mathrm{u}^{(\eta-1)}e^{\left(-\frac{1}{2}\left(\frac{\psi}{\mathrm{u}}+\gamma\mathrm{u}\right)\right)},\ \text{com u}>0, \tag{2.11}$$

*em que $K_\eta(\psi\gamma)$ denota a função de Bessel modificada do terceiro tipo de índice $\eta$ e avaliada em $\psi\gamma$. E será denotada por $U\sim GIG(\eta,\psi,\gamma)$, em que seu espaço paramétrico pode ser definido por*

$$\eta\in\mathbb{R}\quad\text{e}\quad(\psi,\gamma)\in\boldsymbol{\Theta}_\eta$$

*onde*

$$\boldsymbol{\Theta}_\eta=\begin{cases}\{(\psi,\gamma):\ \psi\geq 0,\quad \gamma>0\} & \text{se}\quad \eta>0,\\ \{(\psi,\gamma):\ \psi>0,\quad \gamma>0\} & \text{se}\quad \eta=0,\\ \{(\psi,\gamma):\ \psi>0,\quad \gamma\geq 0\} & \text{se}\quad \eta<0.\end{cases}$$

Para os casos que $\psi=0$ e $\gamma=0$, a constante normativa pode ser encontrada utilizando de alguns resultados importantes sobre a relação assintótica da função de Bessel modificada de terceiro tipo, veja Abramowitz & Stegun (1948) [1] para mais detalhes [43].

Com essas características, a família de distribuições GIG$(\eta,\psi,\gamma)$ contém várias distribuições como casos particulares. Por exemplo, as distribuições Gama ($\psi=0$, $\eta>0$), Gama Inversa (I*Gamma*) ($\gamma=0$, $\eta<0$), Gaussiana Inversa (IG) ($\eta=-1/2$), entre outras [25, 78, 83]. De acordo com Calegari (2020) [25], uma distribuição de particular interesse que é utilizada na construção da distribuição SNGH, é a GIG$(\eta,\omega,\omega)$, onde sua função densidade de probabilidade é obtida por meio da reparametrização de (2.11) da seguinte maneira

$$f_U(\mathrm{u};\eta,\omega)=\frac{\mathrm{u}^{(\eta-1)}}{2K_\eta(\omega)}e^{\left(-\frac{1}{2}\left(\frac{\omega}{\mathrm{u}}+\omega\mathrm{u}\right)\right)},\ \text{com u}>0. \tag{2.12}$$

Seguindo esta mesma ideia, Browne & McNicholas (2015) [24] também utilizaram dessa reparametrização para propor os modelos de misturas finitas de Hiperbólica Generalizada. Desta forma, neste trabalho, utiliza-se desta distribuição reparametrizada para a construção do modelo de misturas finitas proposto no Capítulo 3.

Os resultados apresentados a seguir serão importantes para a obtenção de algumas propriedades que são discutidos na próxima seção, pois determinam os momentos de uma variável aleatória com distribuição GIG, como mostra a proposição a seguir.

**Proposição 2.3.1.** *Seja* $U \sim \text{GIG}(\eta, \psi, \gamma)$. *Então, os momentos de* $U$ *são dados por*

$$a)\ E\left[U^m\right] = \frac{\left(\frac{\gamma}{\psi}\right)^m K_{\eta+m}\left(\sqrt{\psi\gamma}\right)}{K_\eta\left(\sqrt{\psi\gamma}\right)},$$

$$b)\ E\left[\log\left(U\right)\right] = \log\left(\sqrt{\frac{\gamma}{\psi}}\right) + \frac{\partial}{\partial\eta}K_\eta\left(\sqrt{\psi\gamma}\right),$$

$$c)\ Var[U] = \left(\frac{\psi}{\gamma}\right)\left[\frac{K_{\eta+2}(\sqrt{\psi\gamma})}{K_\eta(\sqrt{\psi\gamma})} + \left(\frac{K_{\eta+1}(\sqrt{\psi\gamma})}{K_\eta(\sqrt{\psi\gamma})}\right)^2\right],$$

*em que* $\psi\gamma > 0$ *e* $m \in \mathbb{R}$.

*Demonstração.* A prova segue da definição do valor esperado e da variância de uma variável aleatória para o caso contínuo, bem como algumas propriedades da função Bessel que encontram-se descritas no Apêndice A. □

Após essa breve descrição dos resultados da distribuição gaussiana inversa generalizada, na próxima seção, considera-se esta distribuição para construir a classe de distribuições hiperbólica generalizada normal assimétrica que será estendida no Capítulo 3 no contexto de misturas finitas.

### 2.3.2 Distribuições hiperbólica generalizada normal assimétrica

Nesta seção serão discutidas algumas propriedades relevantes da distribuição SNGH que foram importantes para o desenvolvimento do capítulo seguinte, com a finalidade de continuar a revisão sobre os conceitos preliminares que são utilizados neste trabalho.

A distribuição hiperbólica generalizada normal assimétrica (SNGH), mencionada anteriormente como uma proposta de Vilca et al. (2014) [78] e estendida por Calegari (2020) [25], pode ser considerada como um caso especial das distribuições SMSN quando $\kappa(U) = U$ e $U$ segue uma distribuição GIG na representação estocástica (2.6). Isto é, pode-se definir uma representação estocástica de um vetor aleatório SNGH da seguinte maneira. É importante ressaltar que, a definição a seguir é diferente das utilizadas por Protassov (2004, equação 5) [68] e Browne & McNicholas (2015, equação 4) [24]. Esses dois trabalhos, se baseiam na distribuição de mistura de média-variância normal.

**Definição 2.3.2.** *Seja* $\mathbf{Y}$ *uma vetor aleatório contínuo. É dito que* $\mathbf{Y}$ *segue uma distribuição hiperbólica generalizada normal assimétrica (SNGH), se admite a seguinte representação estocástica*

$$\mathbf{Y} = \boldsymbol{\mu} + \sqrt{U}\mathbf{Z}, \tag{2.13}$$

*em que* $\boldsymbol{\mu}$ *é o parâmetro de locação,* $U$ *uma variável aleatória que segue uma distribuição* $GIG(\eta, \psi, \gamma)$ *e* $\mathbf{Z}$ *é um vetor aleatório com distribuição normal assimétrica de vetor de*

*média* $\mathbf{0}$*, matriz variância e covariância* $\boldsymbol{\Sigma}$ *e vetor de assimetria* $\boldsymbol{\lambda}$*, independente de* $U$*. Em termos de notação, o vetor aleatório p-dimensional* $\mathbf{Y}$ *com distribuição* SNGH *é denotado por* $\mathbf{Y} \sim \text{SNGH}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$*, com* $\boldsymbol{\nu} = (\eta, \psi, \gamma)^\top$ *o vetor de parâmetros associado à distribuição GIG. No caso que* $\boldsymbol{\mu} = \mathbf{0}$ *e* $\boldsymbol{\Sigma} = \mathbf{I}_p$*, tem-se a distribuição* SNGH *p-variada padrão.*

Consequentemente, por meio da representação estocástica apresentada anteriormente, pode-se então definir a função densidade de probabilidade da distribuição SNGH. No entanto, antes de mostrar esse resultado, é necessário apresentar a definição da função densidade de probabilidade de um vetor aleatório que segue uma distribuição Hiperbólica Generalizada, que pode ser vista como um caso particular da distribuição SNGH quando $\boldsymbol{\lambda} = \mathbf{0}$.

**Definição 2.3.3.** *Seja* $\mathbf{Y}$ *um vetor aleatório com distribuição* $\text{SNGH}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$*. Então, para* $\boldsymbol{\lambda} = \mathbf{0}$ *obtém-se a distribuição Hiperbólica Generalizada (GH) simétrica e sua função densidade de probabilidade é dada por*

$$f_{\mathbf{Y}}(\mathbf{y}) = 2a(\boldsymbol{\nu}) \left|\boldsymbol{\Sigma}\right|^{-1/2} \left(q\left(\mathbf{y}\right)\right)^{\eta - p/2} K_{\eta - p/2}\left(q\left(\mathbf{y}\right)\gamma^{1/2}\right), \tag{2.14}$$

*onde* $a(\boldsymbol{\nu}) = \left(\dfrac{\gamma^{1/2}}{2\pi}\right)^{p/2} \left(\psi^{\eta/2} K_\eta\left(\sqrt{\psi\gamma}\right)\right)^{-1}$*, em que* $q(\mathbf{y}) = \sqrt{\psi + d(\mathbf{y})}$*, com* $d(\mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ *e* $\boldsymbol{\nu} = (\eta, \psi, \gamma)$*.*

Assim, no próximo resultado é definida a função densidade de probabilidade para um vetor aleatório $p$-dimensional $\mathbf{Y}$ que segue uma distribuição $\text{SNGH}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ da seguinte forma.

**Definição 2.3.4.** *Considere* $\mathbf{Y}$ *um vetor aleatório contínuo p-dimensional. Seja* $\mathbf{Y}$ *com a distribuição SNGH, como mostrado em (2.13). Então, sua função densidade de probabilidade é dada por*

$$f_{\mathbf{Y}}(\mathbf{y}) = 2a(\boldsymbol{\nu}) \left|\boldsymbol{\Sigma}\right|^{-1/2} \left(q\left(\mathbf{y}\right)\right)^{\eta - p/2} K_{\eta - p/2}\left(q\left(\mathbf{y}\right)\gamma^{1/2}\right) \mathbf{F}_X\left(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2} \left(\mathbf{y} - \boldsymbol{\mu}\right); \boldsymbol{\nu}_{\mathbf{y}}\right), \tag{2.15}$$

*onde* $\mathbf{y} \in \mathbb{R}^p$ *e* $a(\boldsymbol{\nu}) = \left(\dfrac{\gamma^{1/2}}{2\pi}\right)^{p/2} \left(\psi^{\eta/2} K_\eta\left(\sqrt{\psi\gamma}\right)\right)^{-1}$*, em que* $q(\mathbf{y}) = \sqrt{\psi + d(\mathbf{y})}$*, com* $d(\mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$*. Além disso,* $\mathbf{F}_X$ *é uma função acumulada de* $X \sim \text{GH}(0, 1; \boldsymbol{\nu}_{\mathbf{y}})$ *simétrica, na qual* $\boldsymbol{\nu}_{\mathbf{y}} = \left(\eta - p/2, q\left(\mathbf{y}\right)^2, \gamma\right)^\top$*.*

Dessa maneira, a função densidade de probabilidade no contexto univariado para a distribuição SNGH, será definida como mostra o resultado a seguir. Para isto, basta substituir o valor de $p$ por 1 e considerar os parâmetros unidimensionais na expressão dada por (2.15).

**Definição 2.3.5.** *Seja* Y *um vetor aleatório contínuo com distribuição* $SNGH_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$. *Se para* $p = 1$ *tem-se o caso univariado de (2.15), então sua função densidade de probabilidade é dada por*

$$f_{\mathrm{Y}}(\mathrm{y}) = 2a(\boldsymbol{\nu})\sigma^{-1}\left(q\left(\mathrm{y}\right)\right)^{\eta-1/2} K_{\eta-1/2}\left(q\left(\mathrm{y}\right)\gamma^{1/2}\right) F_X\left(\lambda\sigma^{-1}\left(\mathrm{y}-\mu\right); \boldsymbol{\nu}_{\mathrm{y}}\right), \tag{2.16}$$

*em que* $\mathrm{y} \in \mathbb{R}^p$ *e* $a(\boldsymbol{\nu}) = \left(\frac{\gamma^{1/2}}{2\pi}\right)^{1/2}\left(\psi^{\eta/2}K_{\eta}\left(\sqrt{\psi\gamma}\right)\right)^{-1}$, *onde* $q(\mathbf{y}) = \sqrt{\psi + d(\mathbf{y})}$, *com* $d(\mathrm{y}) = \frac{(\mathrm{y}-\mu)}{\sigma^2}$. *Além disso,* $X \sim \mathrm{GH}(0, 1, \boldsymbol{\nu}_{\mathrm{y}})$ *simétrica, na qual* $\boldsymbol{\nu}_{\mathrm{y}} = \left(\eta - 1/2, q\left(\mathrm{y}\right)^2, \gamma\right)^\top$. *É denotado por* $Y \sim \mathrm{SNGH}(\mu, \sigma^2, \lambda, \boldsymbol{\nu})$, *com* $\boldsymbol{\nu} = (\eta, \psi, \gamma)^\top$ *o vetor de parâmetros associado à distribuição GIG.*

A determinação dos momentos de um vetor aleatório, que segue uma distribuição SNGH, serão de fundamental importância para o estudo desenvolvido no capítulo seguinte. Uma vez que, ter o conhecimento do seu valor esperado e da sua variância, é uma consequência essencial para problemas como de identificabilidade [25]. Assim, utilizando da Proposição 2.3.1 no resultado a seguir, são apresentadas tais propriedades.

**Proposição 2.3.2.** *Considere* $\mathbf{Y} \sim \mathrm{SNGH}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, *com com* $\boldsymbol{\nu} = (\eta, \psi, \gamma)^\top$, *sendo o vetor de parâmetros associado à distribuição GIG. Então, tem-se que:*

*a)* $E\left[\mathbf{Y}\right] = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}E\left[U^{1/2}\right]\boldsymbol{\Delta}$,

*b)* $Var\left[\mathbf{Y}\right] = E\left[U\right]\boldsymbol{\Sigma} + \frac{2}{\pi}\left(E\left[U^{1/2}\right]\right)^2\boldsymbol{\Delta}\boldsymbol{\Delta}^\top$,

*em que* $E\left[U^{1/2}\right] = \frac{\left(\frac{\gamma}{\psi}\right)^{1/4}K_{\eta+1/2}\left(\sqrt{\psi\gamma}\right)}{K_{\eta}\left(\sqrt{\psi\gamma}\right)}$ *e* $E\left[U\right] = \frac{\left(\frac{\gamma}{\psi}\right)^{1/2}K_{\eta+1}\left(\sqrt{\psi\gamma}\right)}{K_{\eta}\left(\sqrt{\psi\gamma}\right)}$, *com* $\psi\gamma > 0$, $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ *e* $\boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{\left(1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda}\right)^{1/2}}$.

*Demonstração.* A prova segue pelas Proposições 2.2.2 e 2.3.1. Para mais detalhes sobre a demonstração, veja Calegari (2020) [25]. □

Os resultados determinados a seguir serão constantemente utilizados na elaboração do trabalho, principalmente na obtenção das expressões das esperanças condicionais, úteis na implementação do algoritmo EM para obtenção dos estimadores de máxima verossimilhança dos parâmetros do modelo de misturas finitas de SNGH, descritos no próximo capítulo.

**Lema 2.3.1.** *Considere* $V \sim \mathrm{GIG}\left(\eta, \psi, \gamma\right)$ *e* $\Phi_p(\cdot)$ *a função de distribuição acumulada da normal padrão p-variada. Sejam* $\mathbf{c} \in \mathbb{R}^p$ *e* $a \in \mathbb{R}$. *Então, tem-se que*

$$E\left[\Phi_p\left(V^a\mathbf{c}\right)\right] = F_{\mathbf{W}}\left(\mathbf{c}; \boldsymbol{\nu}_a\right), \tag{2.17}$$

*onde* $F_{\mathbf{W}}(\cdot;\boldsymbol{\nu}_a)$ *é a função distribuição acumulada de* $\mathrm{SMN}_p(\mathbf{0},\mathbf{I}_p,\boldsymbol{\nu}_a)$*, com a representação estocástica dada por* $\mathbf{W} = V^{-a}\mathbf{Z}$*, em que* $\mathbf{Z} \sim \mathrm{N}_p(\mathbf{0},\mathbf{I}_p)$ *e independente de* $V$*.*

*Demonstração.* Veja Calegari (2020) [25]. □

**Proposição 2.3.3.** *Considere* $\mathbf{Y} \sim \mathrm{SNGH}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\boldsymbol{\nu})$*, com* $\boldsymbol{\nu} = (\eta,\psi,\gamma)^\top$ *e representação estocástica dada em (2.13). Então, a função densidade de probabilidade de* $U|\mathbf{Y} = \mathbf{y}$ *é dada por*

$$h(u|\mathbf{Y}=\mathbf{y}) = 2\frac{f_0(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}\Phi\left(u^{-1/2}\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})\right)h_0(u|\mathbf{y}), \tag{2.18}$$

*em que* $h_0(\cdot|\mathbf{y})$ *é a função densidade de probabilidade condicional de* $U|\mathbf{Y} = \mathbf{y}$,$\boldsymbol{\lambda} = \mathbf{0} \sim \mathrm{GIG}\left(\eta - p/2, q(\mathbf{y})^2, \gamma\right)$*, com a constante* $q(\mathbf{y})$ *definidas em (2.3.3) e* $f_0(\mathbf{y})$ *a função densidade de probabilidade da GH simétrica.*

*Demonstração.* A prova segue pelas propriedades da distribuição condicional. Veja Calegari (2020) [25]. □

Posto isto, os resultados expressos neste capítulo, como uma revisão dos conceitos preliminares que serão utilizados no capítulo seguinte, é de fundamental importância na construção e desenvolvimento da estimação dos parâmetros do modelo de misturas finitas de SNGH proposto neste trabalho.

# 3 MISTURAS FINITAS DE DENSIDADES HIPERBÓLICA GENERALIZADA NORMAL ASSIMÉTRICA

Com intuito de modelar dados constituídos por aproximações de densidades complexas, desde aquelas com aspectos multimodais à totalmente assimétricas, utilizam-se de modelos de probabilidade, já consolidados, como as misturas finitas de densidades. Por se tratar de um método extremamente flexível, que vem recebendo bastante atenção nos últimos anos por sua eficiência na modelagem, pode-se dizer que os modelos de misturas finitas são interdisciplinares em sua aplicação, principalmente em casos onde há presença de heterogeneidade populacional.

Alguns exemplos de aplicações de misturas finitas podem ser encontrados em inúmeras áreas da estatística, tais como nas análises de agrupamento, discriminantes e sobrevivência, métodos não-paramétricos ou semi-paramétricos e até mesmo em processamento de imagens. Existem diversas áreas do conhecimento que se utilizam desses modelos, como a astronomia, biologia, genética, medicina, psiquiatria, economia e engenharia [28]. Na literatura, autores como Titterington et al. (1985) [77], McLachan e Basford (1988) [60], Lindsay (1995) [57], Böhning (2000) [21], McLachan e Peel (2004) [62] e Zeller et al (2018) [84] apresentam um amplo espectro de suas teorias e aplicações distintas de misturas finitas de densidades.

Sendo assim, neste trabalho será apresentado um modelo de misturas finitas de densidades para a classe de distribuições SNGH, discutidas em Vilca et al. (2014) [78] e Calegari (2020) [25], juntamente com uma análise gráfica, sua representação hierárquica e algoritmo EM para estimação dos parâmetros, no caso univariado e multivariado. Por fim, também serão discutidos alguns casos particulares, incluindo comentários adicionais sobre o algoritmo EM. Entretanto, antes de iniciar a obtenção do modelo proposto, será feita uma revisão teórica sobre os modelos de misturas finitas de modo geral.

## 3.1 MODELO DE MISTURAS FINITAS

Nesta seção, o modelo de mistura finita de densidades será introduzido de forma geral, bem como algumas propriedades importantes que serão discutidas. Posteriormente, apresenta-se uma abordagem de dados incompletos para o modelo que em seguida será utilizado na estimação de máxima verossimilhança via algoritmo EM.

### 3.1.1 **Definição**

De maneira geral, um modelo de misturas finitas de densidades é uma combinação convexa das funções de distribuição acumuladas e pode ser representado através expressão determinada na definição a seguir [28].

**Definição 3.1.1.** *Seja* $\mathbf{Y} \in \mathbb{R}^p$ *um vetor aleatório com função densidade dada por*

$$f(\mathbf{y}) = \sum_{j=1}^{G} \rho_j g_j(\mathbf{y}), \tag{3.1}$$

*é dito ter uma distribuição de mistura de densidade, com* $\rho_j \geq 0$ *e* $\sum_{j=1}^{G} \rho_j = 1$. *Assim, tem-se que, a função* $f(\cdot)$ *é chamada de misturas finitas de densidades, tal que* $g_1, \ldots, g_G$ *são as componentes de densidades da misturas e* $\rho_1, \ldots, \rho_G$ *são denominados proporções ou pesos de misturas.*

No caso em que as componentes $g_j(\cdot)$ são pertencentes à famílias paramétricas de distribuições, pode-se reescrever o modelo dado por (3.1), como

$$f(\mathbf{y}; \boldsymbol{\Theta}) = \sum_{j=1}^{G} \rho_j g_j(\mathbf{y}; \boldsymbol{\theta}_j), \tag{3.2}$$

com $\boldsymbol{\Theta} = ((\boldsymbol{\theta}_1{}^\top, \rho_1), \ldots, (\boldsymbol{\theta}_G{}^\top, \rho_G))^\top$ os parâmetros que são responsáveis por definir cada uma das componentes de $g_j$, não necessariamente estão definidos no mesmo espaço paramétrico. Ainda assim, em sua maioria das vezes, as aplicações que são encontradas na literatura, bem como as que serão apresentadas neste trabalho, consideram que as componentes de mistura $g_j$ sejam pertencentes à mesma família paramétrica de distribuições. Sendo assim, a mistura finita de densidades será denotada por

$$f(\mathbf{y}; \boldsymbol{\Theta}) = \sum_{j=1}^{G} \rho_j g(\mathbf{y}; \boldsymbol{\theta}_j), \ \mathbf{y} \in \mathbb{R}^p. \tag{3.3}$$

Convém ressaltar que, sob tais suposições, os parâmetros $\boldsymbol{\theta}_j$ agora pertencem a um mesmo espaço paramétrico.

### 3.1.2 Identificabilidade

De acordo com McLachlan e Peel (2004) [62], a estimação de um parâmetro $\boldsymbol{\Theta}$, para uma determinada distribuição $f(\mathbf{y}, \boldsymbol{\theta})$, com base nas observações do vetor aleatório $\mathbf{Y}$, só irá fazer sentido se $\boldsymbol{\theta}$ for identificável, isto é, se o parâmetro possui uma caracterização única, então fica viável sua identificação.

Neste contexto, antes de prosseguir com as discussões sobre problemas de estimação, é importante discutir a identificabilidade da mistura, pois tal condição relativa ao modelo garante a unicidade de seus parâmetros, ou seja, que esses possam ser estimados de forma única. De modo geral, uma família paramétrica de funções densidades de probabilidades é dita ser identificável, se valores distintos de seus parâmetros determinam membros diferentes da família, em termos matemáticos, se a classe de funções $\mathcal{F} = \{g(\mathbf{y}; \boldsymbol{\Theta}) : \boldsymbol{\Theta} \in \Omega\}$, onde $g$ é distribuição de probabilidade e $\Omega$ o espaço paramétrico especificado, é identificável se, e somente se para cada $g \in \mathcal{F}$ e $\boldsymbol{\Theta} \neq \boldsymbol{\Theta}^*$, tem-se que $\forall\ \mathbf{y} \in \mathbb{R}^p$, então $g(\mathbf{y}; \boldsymbol{\Theta}) \neq g(\mathbf{y}; \boldsymbol{\Theta}^*)$.

Assim, segue abaixo a definição de identificabilidade no contexto de misturas finitas de densidades [17].

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

$$\mathcal{H} = \left\{ f(\mathbf{y};\boldsymbol{\Theta}) : f(\mathbf{y};\boldsymbol{\Theta}) = \sum_{j=1}^{G} \rho_j g(\mathbf{y};\boldsymbol{\theta}_j), \quad \rho_j \geq 0 \text{ e } \sum_{j=1}^{G} \rho_j = 1, \right.$$
$$\left. g(\mathbf{y};\boldsymbol{\theta}_j) \in \mathcal{F}, \quad \boldsymbol{\Theta} = ((\boldsymbol{\theta}_1{}^\top, \rho_1), \ \ldots, (\boldsymbol{\theta}_G{}^\top, \rho_G))^\top \right\},$$

*uma classe de misturas finitas de densidades. Suponha*

$$f(\mathbf{y};\boldsymbol{\Theta}) = \sum_{j=1}^{G} \rho_j g(\mathbf{y};\boldsymbol{\theta}_j) \quad \text{e} \quad f(\mathbf{y};\boldsymbol{\Theta}^*) = \sum_{j=1}^{G^*} \rho_j{}^* g(\mathbf{y};\boldsymbol{\theta}_j^*),$$

*dois membros quaisquer da classe* $\mathcal{H}$*. Se* $\mathcal{H}$ *tem-se que* $f(\mathbf{y};\boldsymbol{\Theta}) = f(\mathbf{y};\boldsymbol{\Theta}^*)$ *se, e somente se,* $G = G^*$ *e ainda se pode permutar os índices das componentes de forma que* $\rho_j = \rho_j{}^*$ *e* $g(\mathbf{y};\boldsymbol{\theta}_j) = g(\mathbf{y};\boldsymbol{\theta}_j^*)$ *com* $j = 1, ..., G$*. Então, a classe* $\mathcal{H}$ *é dita identificável.*

Nas últimas décadas, estudos envolvendo o problema de identificabilidade de modelos de misturas finitas vem recebendo uma certa importância. O artigo de Teicher (1963) [76] é o ponto de partida para tais investigações. No entanto, o processo de identificabilidade é declarado sob algumas condições restritivas que não são aplicáveis a algumas famílias multiparamétricas. Yakowitz & Spragins (1968) [82] propõem uma condição suficiente e necessária para identificabilidade de uma classe de misturas finitas de distribuições pertencentes a mesma família paramétrica, é que este conjunto seja linearmente independente sobre o corpo dos números reais. Em Titterington et al. (1985) [77] discute questões teóricas sobre as condições para que uma mistura seja identificável, e ainda comenta que a maioria das misturas finitas de densidades, das quais suas distribuições são contínuas, são identificáveis. Já em Atienza et al. (2006) [9] foi proposto um novo resultado que fornece uma condição necessária e suficiente para que uma classe de misturas finitas cujas componentes pertencem a diferentes famílias de distribuições seja identificável, com menos requisitos apresentados no trabalho de Teicher (1963) [76].

Segundo McLachlan e Basford (1988) [60], existem algumas dificuldades quando as componentes de uma mistura pertencem à mesma família de distribuições. Nesse caso, quando os índices $j$ forem permutados em $\boldsymbol{\Theta} = ((\boldsymbol{\theta}_1{}^\top, \rho_1), \ \ldots, (\boldsymbol{\theta}_G{}^\top, \rho_G))^\top$, o valor da densidade de mistura será o mesmo independente do índice. Com isso, apesar da mistura ser identificável, o vetor de parâmetros $\boldsymbol{\Theta}$ não será. 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 $\boldsymbol{\Theta}$ [17].

Além dessas dificuldades encontradas, a falta de identificabilidade pode gerar alguns problemas em certas situações, como o caso em que a função de verossimilhança atinge o

máximo local para distintos valores de $\boldsymbol{\Theta}$, com isso, as equações de verossimilhança terão várias raízes. De acordo com McLachlan e Krishnan (2007) [61], esse problema pode ser grave quando o interesse é estimar valores específicos para cada parâmetro envolvido [26].

Neste âmbito, o problema de identificabilidade no contexto de misturas finitas de SMSN, ainda não foi discutido de forma teórica, deixando em aberto para discussões sobre esse tema. Do mesmo modo acontece com as misturas finitas de distribuições SNGH, uma vez que, existe uma certa dificuldade em encontrar soluções teóricas para o problema de identificabilidade nesses modelos de misturas de finitas, se tratando assim de algo não tão trivial de solucionar. Alguns comentários sobre os métodos utilizados para contorna tal problema, podem ser encontrados na Seção 3.2.4. Contudo, vale ressaltar que não são provas teóricas relacionadas a este assunto, podendo então ser considerado como trabalhos futuros.

### 3.1.3 O Algoritmo EM em modelos de misturas finitas

Nesta seção, será descrito o algoritmo EM no contexto de misturas finitas de densidades de forma geral. Entretanto, como de costume na estrutura do algoritmo, inicialmente é necessário introduzir a estrutura de dados incompletos para o problema de misturas [62]. Para isso, primeiramente será abordada a questão de como gerar vetores pseudo aleatórios provenientes das mistura de densidades.

Sendo assim, considere o problema de se gerar vetores aleatórios que seguem uma distribuição de misturas finitas de densidades. Seja $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ uma amostra aleatória de tamanho $n$ de um vetor aleatório $\mathbf{Y}_i$, $p$-dimensional, com distribuição dada por (3.1) e suponha $\mathbf{Z}_i$ um vetor aleatório para $i = 1, \ldots, n$, com $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})^\top$, onde suas componentes são definidas como variáveis indicadoras da seguinte forma

$$Z_{ij} = \begin{cases} 1, & \text{se } \mathbf{Y}_i \sim f_j(\mathbf{y}; \boldsymbol{\theta}_j); \\ 0, & \text{se } \mathbf{Y}_i \sim f_k(\mathbf{y}; \boldsymbol{\theta}_k),\ j \neq k. \end{cases} \tag{3.4}$$

Nesse contexto, a variável $Z_{ij}$ pode ser interpretada como uma variável latente, não observável, associada ao vetor $\mathbf{Y}_i$, indicando de qual componente da mistura é proveniente. Dessa maneira, os valores observados $\mathbf{y}_1, \ldots, \mathbf{y}_n$ de $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ são considerados como *dados incompletos* e $\mathbf{z}_1, \ldots, \mathbf{z}_n$ sendo os valores para $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, segue então que o vetor de *dados completos* é definidos por $\mathbf{y}_c = \left(\mathbf{y}^\top, \mathbf{z}^\top\right)^\top$, com $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$ e $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$. Consequentemente, sob essa abordagem, assume-se que o vetor $\mathbf{Z}_i$ tem distribuição multinomial, considerando uma retirada em $G$ categorias, com probabilidades $\rho_1, \ldots, \rho_G$, isto é, formam uma amostra aleatória

$$\mathbf{Z}_1, \ldots, \mathbf{Z}_n \sim Multi_G(1, \rho_1, \ldots, \rho_G),$$

desse modo a distribuição de $\mathbf{Z}_i$ será da forma

$$f(\mathbf{z}_i;(\rho_1,\ldots,\rho_G)) = \prod_{j=1}^{G} {\rho_j}^{z_{ij}}.$$

Assim, pôde-se contornar o problema de gerar vetores aleatórios de uma distribuição de misturas de densidades com a inclusão do vetor $\mathbf{Z}_i$, vetor esse responsável por regular as proporções de mistura das componentes. O conceito de existir esse vetor latente, associando a observação e a componente em que ele provém, é muito útil, embora não pareça ser intuitivo num sentido físico. Ao decorrer deste trabalho, será visto que esse conceito é o que permite a Estimação de Máxima Verossimilhança (EMV) através do algoritmo EM [17].

Note que a função de verossimilhança dos dados observados é dada por

$$\begin{aligned} L(\boldsymbol{\Theta}) &= L(\boldsymbol{\Theta}|\mathbf{y}_i) \\ &= \prod_{i=1}^{n} f(\mathbf{y}_i;\boldsymbol{\Theta}) \\ &= \prod_{i=1}^{n}\sum_{j=1}^{G} \rho_j g_j(\mathbf{y}_i;\boldsymbol{\theta}_j). \end{aligned} \tag{3.5}$$

A maximização direta da função de verossimilhança dada por (3.5) pode levar a um problema numérico difícil e instável. Para contornar tal problema, a melhor opção é utilizar o algoritmo EM de Dempster et al. (1977) [33]. Assim, pode-se obter a função de verossimilhança de dados completos da seguinte forma

$$\begin{aligned} L_c(\boldsymbol{\Theta}) &= L_c(\boldsymbol{\Theta}|\mathbf{y}_c) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{G} f(\mathbf{y}_i;\boldsymbol{\theta}_j)^{z_{ij}} f(\mathbf{z}_i;(\rho_1,\ldots,\rho_G)) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{G} \left[\rho_j g(\mathbf{y}_i;\boldsymbol{\theta}_j)\right]^{z_{ij}}, \end{aligned} \tag{3.6}$$

e sua log-verossimilhança completa é dada por

$$\begin{aligned} l_c(\boldsymbol{\Theta}) &= \log L_c(\boldsymbol{\Theta}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\left[\log(\rho_j) + \log(g(\mathbf{y}_i;\boldsymbol{\theta}_j))\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\log(\rho_j) + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\log(g(\mathbf{y}_i;\boldsymbol{\theta}_j)). \end{aligned} \tag{3.7}$$

Então, o algoritmo EM, será implementado em duas etapas, a primeira é constituída pela etapa E (Esperança) e logo após, a etapa M (Maximização) finaliza o algoritmo [17, 31].

**Etapa E**: Nesse passo, calcula-se o valor esperado da log-verossimilhança $l_c(\boldsymbol{\Theta})$ condicional aos dados observados $\mathbf{y}$ e $\widehat{\boldsymbol{\Theta}}^{(k)}$, o valor estimado de $\boldsymbol{\Theta}$ na $(k)$-ésima iteração. Dessa forma, obtém-se a função $Q$ dada por

$$\begin{aligned} Q(\boldsymbol{\Theta}; \widehat{\boldsymbol{\Theta}}^{(k)}) &= E\left[l_c(\boldsymbol{\Theta})|\mathbf{y}, \widehat{\boldsymbol{\Theta}}^{(k)}\right] \\ &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij}\left(\log(\rho_j) + \log(g(\mathbf{y}_i; \boldsymbol{\theta}_j))\right)|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{G} E\left[Z_{ij}|\mathbf{y}; \widehat{\boldsymbol{\theta}}^{(k)}\right]\left(\log(\rho_j) + \log(g(\mathbf{y}_i; \boldsymbol{\theta}_j))\right), \end{aligned} \tag{3.8}$$

onde $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$. Como a log-verossimilhança completa é linear na variável não observada $Z_{ij}$, a etapa E consiste em obter a esperança condicional de $Z_{ij}$, dado o vetor de dados observados $\mathbf{y}$ e $\widehat{\boldsymbol{\Theta}}^{(k)}$. Observe que

$$E\left[Z_{ij}|\mathbf{y}; \widehat{\boldsymbol{\theta}}^{(k)}\right] = P(Z_{ij} = 1|\mathbf{y}; \widehat{\boldsymbol{\theta}}^{(k)}) = \widehat{z_{ij}}.$$

Note que $E\left[Z_{ij}|\mathbf{y}\right]$ é fácil de calcular computacionalmente, pois

$$\begin{aligned} \widehat{z_{ij}}^{(k)} &= \frac{\widehat{\rho_j}^{(k)} g\left(\mathbf{y}_i; \widehat{\boldsymbol{\theta}}_j^{(k)}\right)}{f(\mathbf{y}_i; \widehat{\boldsymbol{\Theta}}^{(k)})} \\ &= \frac{\widehat{\rho_j}^{(k)} g\left(\mathbf{y}_i; \widehat{\boldsymbol{\theta}}_j^{(k)}\right)}{\sum_{l=1}^{G} \widehat{\rho_l}^{(k)} g(\mathbf{y}_i; \widehat{\boldsymbol{\theta}}_l^{(k)})}, \end{aligned} \tag{3.9}$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, G$. Portanto, usando (3.9) e o resultado obtido em (3.8), tem-se que a função $Q$ é dada por

$$\begin{aligned} Q(\boldsymbol{\Theta}; \widehat{\boldsymbol{\Theta}}^{(k)}) &= \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z_{ij}}^{(k)}\left[\log(\rho_j) + \log(g(\mathbf{y}_i; \boldsymbol{\theta}_j))\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z_{ij}}^{(k)} \log(\rho_j) + \sum_{i=1}^{n}\sum_{j=1}^{G} \widehat{z_{ij}} \log(g(\mathbf{y}_i; \boldsymbol{\theta}_j)). \end{aligned} \tag{3.10}$$

**Etapa M**: A etapa M do algoritmo consiste em maximizar a função $Q$ (3.10) com respeito à $\boldsymbol{\Theta}$, resultando na estimativa atualizada, na $(k+1)$-ésima iteração, do parâmetro, ou seja, $\widehat{\boldsymbol{\Theta}}^{(k+1)}$. Caso $z_{ij}$ fosse observável, então o estimador de máxima verossimilhança de $\rho_j$, para $j = 1, \ldots, G$, considerando a estrutura de dados completos, seria determinado por

$$\widehat{\rho_j} = \frac{1}{n}\sum_{j=1}^{G} z_{ij}, \tag{3.11}$$

conforme feito na etapa anterior, a etapa E, substitui-se $z_{ij}$ pela sua esperança condicional $\widehat{z_{ij}}$ na log-verossimilhança completa, então a estimativa atualizada de $\rho_j$, para $j = 1, \ldots, G$,

é obtida após substituir $\widehat{z_{ij}}$ em (3.11) da seguinte forma

$$\widehat{\rho_j}^{(k+1)} = \frac{1}{n}\sum_{j=1}^{G}\widehat{z_{ij}}^{(k+1)}, \tag{3.12}$$

Cabe ressaltar que o algoritmo EM descrito nesta seção é da forma geral, portanto considera-se somente como dados aumentados a variável $Z_{ij}$. Vale lembrar também que, como o algoritmo EM é um processo iterativo, as iterações são repetidas até que cumpra com o critério de parada estipulado. Nesse trabalho, nas aplicações numéricas, o critério adotado é baseado no erro relativo absoluto das log-verossimilhanças dos dados observados nas iterações $(k)$ e $(k+1)$, isto é,

$$\left|\left|\frac{l(\widehat{\boldsymbol{\Theta}}^{(k)}|\mathbf{y}) - l(\widehat{\boldsymbol{\Theta}}^{(k+1)}|\mathbf{y})}{l(\widehat{\boldsymbol{\Theta}}^{(k+1)}|\mathbf{y})}\right|\right| < \epsilon. \tag{3.13}$$

Os valores iniciais são necessários para a implementação desse algoritmo. Tais valores serão obtidos e descritos com mais detalhes no próximo capitulo, na Seção 4.1.

Na Seção 3.2 será construído o algoritmo EM para misturas finitas de densidades SNGH, em ambos os contextos univariado e multivariado, e nesse caso, as etapas E e M serão modificadas substancialmente em relação as etapas aqui apresentadas em um contexto geral.

### 3.1.4 **Métodos de seleção de modelos**

A escolha de um critério de seleção de modelos é adotada em vários aspectos, que vão desde a comparação entre modelos, até a determinação do número de componentes $G$ no contexto de misturas finitas. Essa ferramenta dá evidências de qual modelo pode se adequar melhor aos dados com diferentes tipos de metodologias empregadas nessa avaliação, conforme descritas abaixo. Desta maneira, neste trabalho serão utilizados: o Critério de Informação de Akaike (AIC) e Bayesiano (BIC).

#### 3.1.4.1 *Critério de informação de Akaike*

O critério de informação Akaike (AIC) tem como objetivo encontrar um melhor modelo que se ajuste adequadamente aos dados, a partir de um conjunto finito de modelos contendo poucos parâmetros. Deste modo, o modelo selecionado é aquele que minimiza a distância de Kullback-Leibler, isto é, o modelo é escolhido quando a medida de divergência entre o modelo verdadeiro e o modelo estimado é minimizada. Contudo, de acordo com Akaike (1974) [4], ao definir $d$ como o número total de parâmetros a serem estimados do modelo, o critério de informação de Akaike seleciona o modelo que minimiza a expressão escrita como

$$AIC = -2\log L\left(\widehat{\boldsymbol{\Theta}}\right) + 2d,$$

onde $L\left(\widehat{\boldsymbol{\Theta}}\right)$ é a função de verossimilhança em $\widehat{\boldsymbol{\Theta}}$.

#### 3.1.4.2 *Critério de informação Bayesiano*

Este critério de informação tem como base a teoria Bayesiana para seleção de modelos, onde a ideia básica desse método é considerar alguns possíveis modelos num conjunto finito de modelos, com suas probabilidades a *priori*, e selecionar aquele que apresenta a maior probabilidade a *posteriori*, dadas as observações. O critério de informação Bayesiano (BIC) também pode ser encontrado na literatura como critério de informação de Schawarz (SIC). Tal critério foi proposto por Schwarz (1978) [72] da seguinte maneira

$$BIC = -2\log L\left(\widehat{\boldsymbol{\Theta}}\right) + d\log\left(n\right),$$

onde $L\left(\widehat{\boldsymbol{\Theta}}\right)$ é a função de verossimilhança em $\widehat{\boldsymbol{\Theta}}$, $n$ é o tamanho da amostra e $d$ o número total de parâmetros.

### 3.1.5 **Classificação em modelos de misturas finitas**

A teoria da classificação é uma ferramenta muito importante e eficiente utilizada para construir uma regra de distinção em situações onde é desejado alocar uma entidade em categorias. Nesse sentido, no contexto de misturas finitas, tal procedimento é feito pelas proporções ou pesos de misturas. Portanto, suponha que a intenção seja classificar uma amostra aleatória $\mathbf{y}_1, \ldots, \mathbf{y}_n$ em $G$ grupos.

Então, ao considerar a estrutura de dados incompletos, conforme descrita no início da Seção 3.1.3, o objetivo é deduzir $\mathbf{z}_i$ em termos dos dados observados $\mathbf{y}_i$. Assim, após ajustar o modelo de mistura de $G$-componentes, o vetor de estimativas $\widehat{\boldsymbol{\Theta}}$ dos parâmetros do modelo é utilizado para obter uma classificação das observações em termos probabilísticos, com base em probabilidades *a posteriori*. Dessa forma, para cada elemento da amostra, as $G$ probabilidades de $\widehat{z_{ij}}$, para $j = 1, \ldots, G$, definem as probabilidades *a posteriori* da $i$-ésima observação que pertence ao primeiro grupo até o $G$-ésimo grupo, respectivamente. Então, classifica-se a observação para aquele grupo correspondente ao maior valor observado das probabilidades *a posteriori*. Conforme a regra de Bayes, tem-se que a probabilidade *a posteriori* de que a entidade pertença a $j$-ésima componente com $\mathbf{y}_i$ já observado, é

$$\begin{aligned} \widehat{z_{ij}} &= P(Z_{ij} = 1|\mathbf{y}_i) \\ &= \frac{\rho_j g_j\left(\mathbf{y}_i\right)}{f(\mathbf{y}_i)}. \end{aligned} \tag{3.14}$$

onde a $j$-ésima proporção $\rho_j$ da mistura, pode ser interpretada como a probabilidade *a priori* da $i$-ésima observação proveniente da $j$-ésima componente da mistura com $i = 1, \ldots, n$.

Na literatura, é comum encontrar aplicações de modelos envolvendo a teoria de classificação, que são frequentemente utilizadas para análise de dados nos segmentos dos negócios para resolver problemas contemporâneos. Assim, o modelo de misturas escolhido tem a capacidade de interpretar simultâneas características distintas de cada grupo dos segmentos com base nos fenômenos específicos de cada agrupamento [44].

## 3.2 MODELO DE MISTURAS FINITAS DE DENSIDADES HIPERBÓLICA GENERALIZADA NORMAL ASSIMÉTRICA

Nesta seção, toda metodologia abordada ao longo desse trabalho será utilizada para a elaboração e desenvolvimento do modelo de misturas finitas de distribuições da classe SNGH. Inicialmente é definido o modelo de misturas finitas de SNGH (FM-SNGH) para o caso univariado e multivariado. Em seguida, sua representação hierárquica é apresentada com o intuito de auxiliar na construção do algoritmo EM a fim de estimar os parâmetros envolvidos no modelo.

### 3.2.1 Definição

Para a construção do modelo de misturas finitas de SNGH, utiliza-se das Definições 2.3.4 e 3.1.1, especificamente da expressão (3.3).

**Definição 3.2.1.** *Considere* $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$*, uma amostra aleatória de tamanho* $n$*, com os vetores aleatórios* $\mathbf{Y}_i \in \mathbb{R}^p$*, p-dimensional, retirados de uma mistura de* $G$ *componentes de densidades* SNGH *dada como*

$$f(\mathbf{y}_i; \boldsymbol{\Theta}) = \sum_{j=1}^{G} \rho_j f_{\text{SNGH}}(\mathbf{y}_i; \boldsymbol{\theta}_j),\ \rho_j \geq 0 \text{ e } \sum_{j=1}^{G} \rho_j = 1, \tag{3.15}$$

*onde* $\rho_j$ *são as probabilidades de mistura,* $f_{\text{SNGH}}$ *é a notação referente à densidade da distribuição SNGH definida em (2.3.4),* $\boldsymbol{\Theta} = ((\boldsymbol{\theta}_1{}^\top, \rho_1), \ldots, (\boldsymbol{\theta}_G{}^\top, \rho_G))^\top$*, com o vetor de parâmetros específicos das componentes* $\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j)^\top$ *no contexto univariado e para o caso multivariado, tem-se* $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\alpha}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu}_j)^\top$*, na qual* $\boldsymbol{\alpha}$ *é definido como o conjunto minimal de parâmetros de* $\boldsymbol{\Sigma}$*, de modo que essa matriz seja uma matriz de escala bem definida, para* $j = 1, \ldots, G$*, e* $\boldsymbol{\nu}_j = (\eta_j, \psi_j, \gamma_j)^\top$ *é o vetor de parâmetros associado à distribuição GIG.*

No que se refere ao parâmetro $\boldsymbol{\nu}_j$ associado à distribuição GIG, é importante notar que para a estimação desse parâmetro via algoritmo EM e por conveniência computacional, assume-se que $\boldsymbol{\nu}_1 = \ldots = \boldsymbol{\nu}_G = \boldsymbol{\nu}$. Tal estratégia funciona muito bem na realização de estudos empíricos, simplificando bastante o problema de otimização [31].

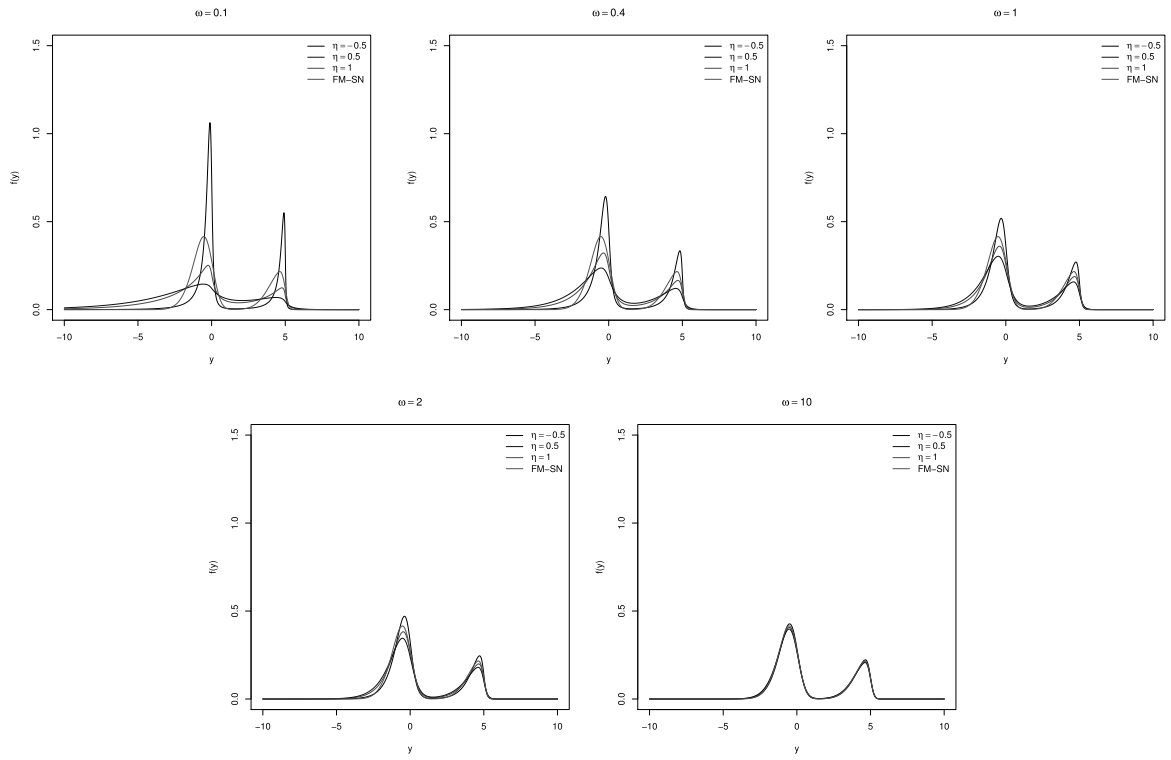
### 3.2.2 Análise gráfica do modelo de misturas finitas de densidades hiperbólica generalizada normal assimétrica

A fim de exemplificar a flexibilidade do modelo apresentado na Definição 3.2.1, foram construídos alguns gráficos que descrevem o comportamento do modelo ao variar alguns elementos do vetor de parâmetros específicos $\boldsymbol{\theta}_j$ da expressão dada por (3.15), em particular do parâmetro $\boldsymbol{\nu}$ associado à distribuição GIG, no contexto em que $\psi = \gamma = \omega$.

Sendo assim, o objetivo para construção dos gráficos a seguir, é avaliar o comportamento do modelo de misturas finitas de SNGH (FM-SNGH), no caso univariado, com $G = 2$, sem perda de generalidade, alterando os valores dos parâmetros $\omega$ e $\eta$. Assim, adota-se a seguinte configuração para os parâmetros do modelo de interesse: $\omega$ = 0,1; 0,4; 1; 2 e 10, e $\eta$ = -0,5; 0; 0,5 e 1, quando $\rho_1 = 0{,}4$; $\mu_1 = 0$, $\mu_2 = 5$, $\sigma_1^2 = \sigma_2^2 = 1$, $\lambda_1$ = -2 e $\lambda_2$ = -5. A análise gráfica foi dividida em dois casos, onde no primeiro considera-se fixo o valor de $\omega$ variando $\eta$ e no segundo, fixa-se $\omega$ e varia $\eta$. Adicionalmente, com o intuito de comparar as curvas obtidas, foi considerado o modelo de misturas finitas de normais assimétricas (FM-SN) usando a mesma configuração adotada para o modelo FM-SNGH.

Na Figura 2, quando $\omega$ é fixo, à medida que o valor escolhido de $\eta$ cresce, foi percebido que as modas dos grupos tenderam a diminuir e consequentemente surgirem caudas mais pesadas. Pôde ser observado também que quanto maior o valor adotado para $\omega$, independente do valor de $\eta$, o modelo FM-SNGH se aproxima do modelo FM-SN.

Figura 2 – Curvas de densidades provenientes do modelo de misturas finitas de SNGH univariado para $\omega$ fixo.

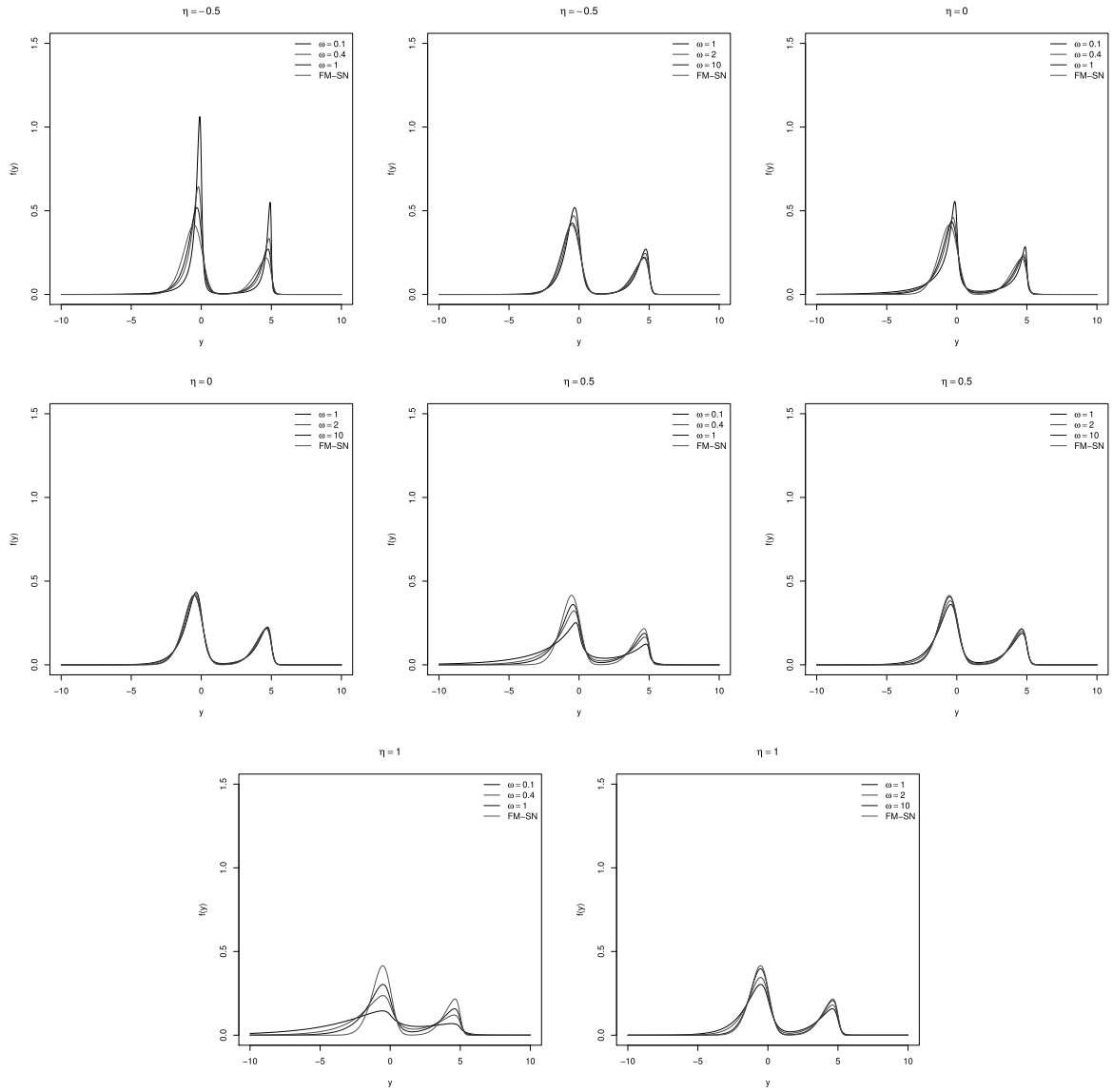



Fonte: O autor (2020).

Agora para $\eta$ fixo e $\omega$ variando entre os valores estabelecidos, pôde-se observar na Figura 3, que à medida que o valor de $\omega$ aumenta, quando $\eta \leq 0$, as modas tendem a diminuir e como consequência disso as caudas ficam mais pesadas. Já para os casos em que $\eta > 0$, ocorreu o contrário do que aconteceu com $\eta \leq 0$, isto é, houve um aumento na tendência das modas e consequentemente as caudas ficaram menos pesadas. Além disso,

em geral, o modelo FM-SNGH aproximou-se bem do modelo FM-SN quando $\omega \geq 1$ tende a aumentar.

Figura 3 – Curvas de densidades provenientes do modelo de misturas finitas de SNGH univariado para $\eta$ fixo.




Fonte: O autor (2020).

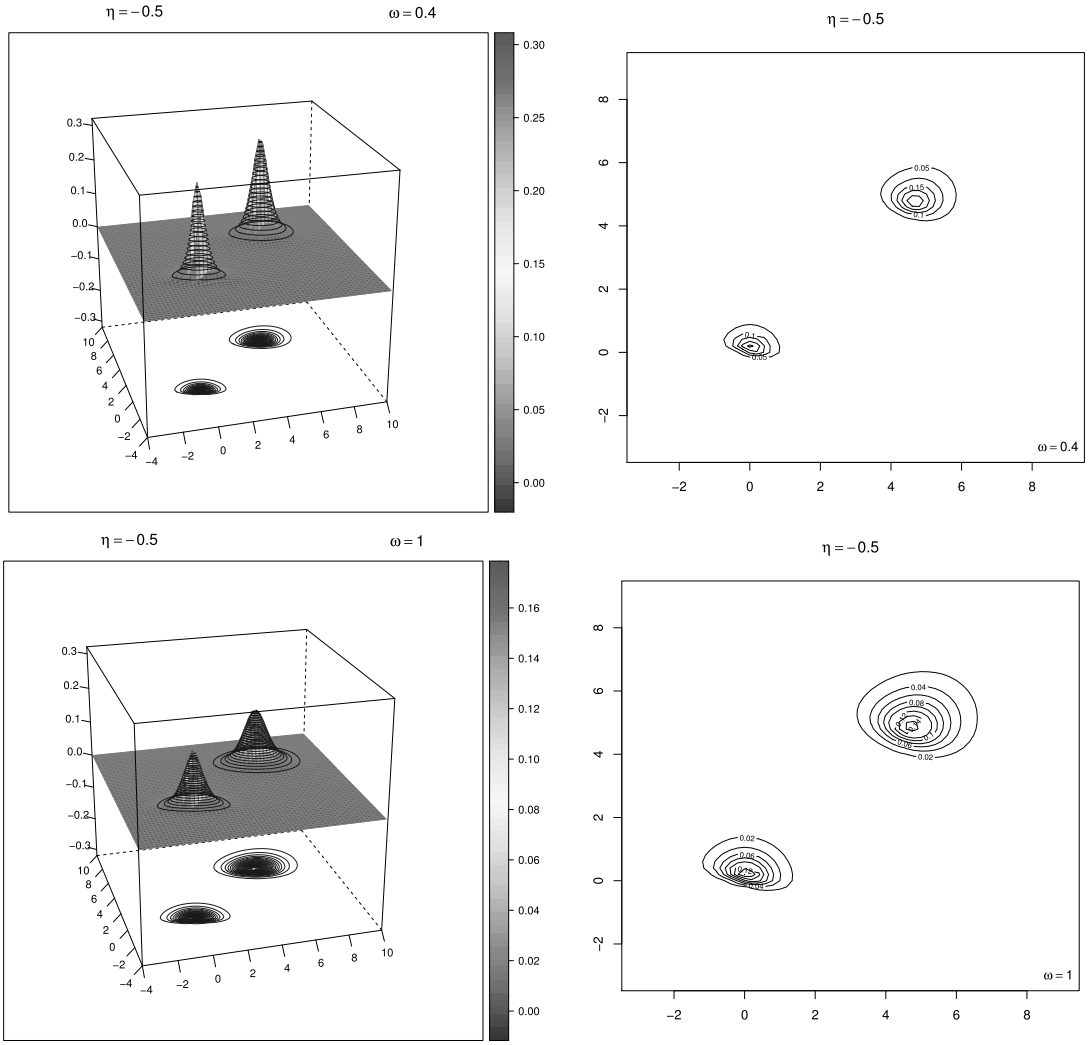
Com o mesmo objetivo definido no início dessa seção, agora para o contexto multivariado do modelo FM-SNGH, com $G = 2$, são apresentados os gráficos de superfície e suas curvas de níveis correspondentes, quando $\rho_1 = 0{,}3$; $\boldsymbol{\mu}_1 = (0{,}0)^\top$, $\boldsymbol{\mu}_2 = (5{,}5)^\top$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_2$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0{,}5 \\ 0{,}5 & 2 \end{pmatrix}$, $\boldsymbol{\lambda}_1 = (1{,}4)^\top$ e $\boldsymbol{\lambda}_2 = (1{,}2)^\top$, considerando a mesma configuração adotada anteriormente para os parâmetros $\omega$ e $\eta$. Desta maneira, nas Figuras 4 à 6 são apresentadas na coluna à esquerda as superfícies obtidas conforme varia-se os valores do parâmetro $\omega$ e à direita suas respectivas curvas de níveis. Ao final encontram-se as

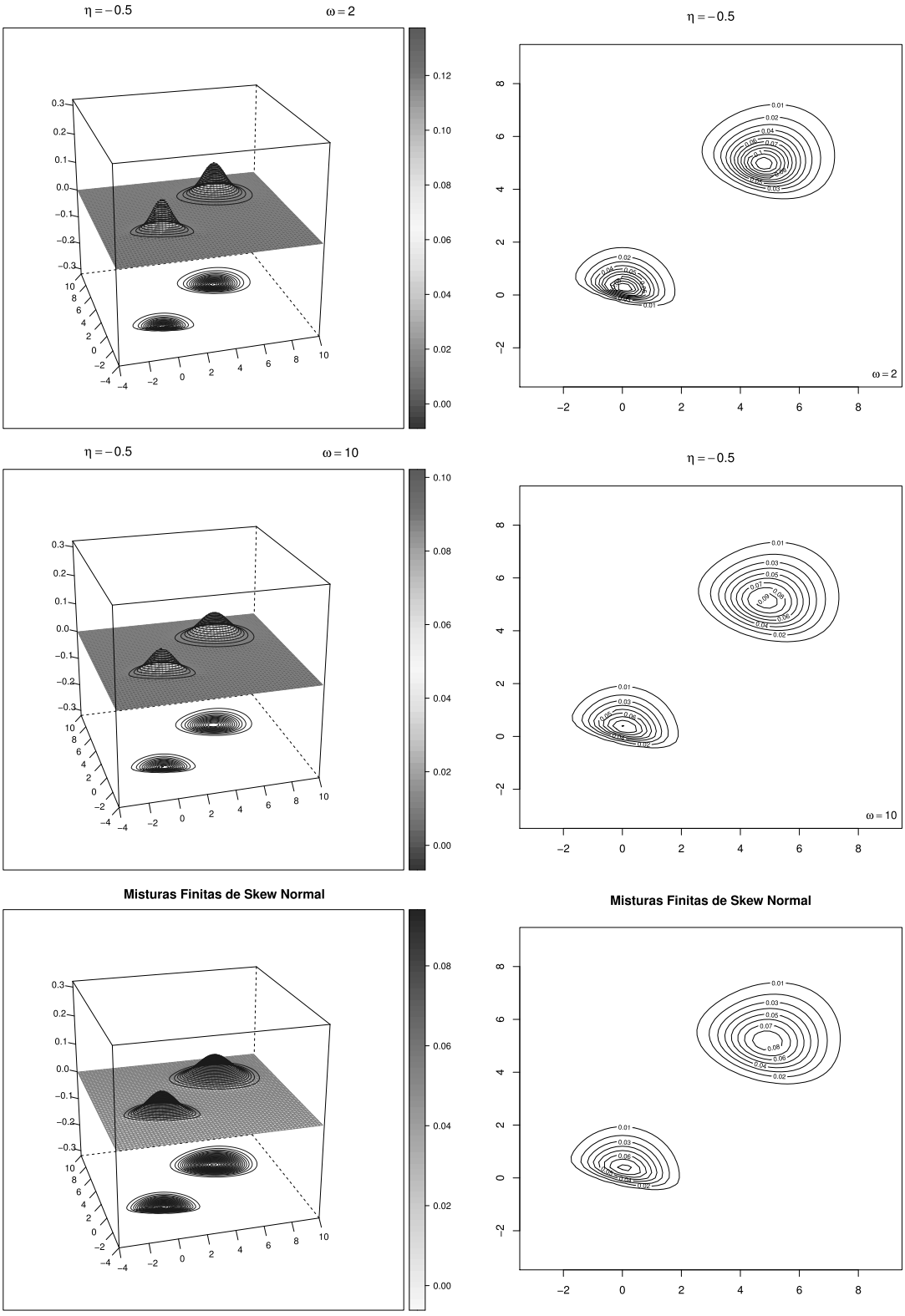
superfícies geradas para o modelo FM-SN usando da mesma configuração adotada para o modelo FM-SNGH.

Na Figura 4, quando $\eta$ = -0,5 fixo, foi observado que quanto maior o valor de $\omega$, maior o achatamento das superfícies e suas curvas de níveis tendem a ser mais largas, consequentemente surgem as caudas pesadas.

Por fim, para $\eta = 0{,}5$ e $\eta = 1$ fixos, como mostram as Figuras 5 e 6, pode-se observar que, quanto maior é o valor de $\omega$, mais elevados ficam os picos das curvas de superfícies e suas curvas de níveis tendem a ficar mais concentradas. Além do mais, com $\omega = 10$, percebe-se a aproximação do modelo FM-SNGH ao modelo FM-SN.

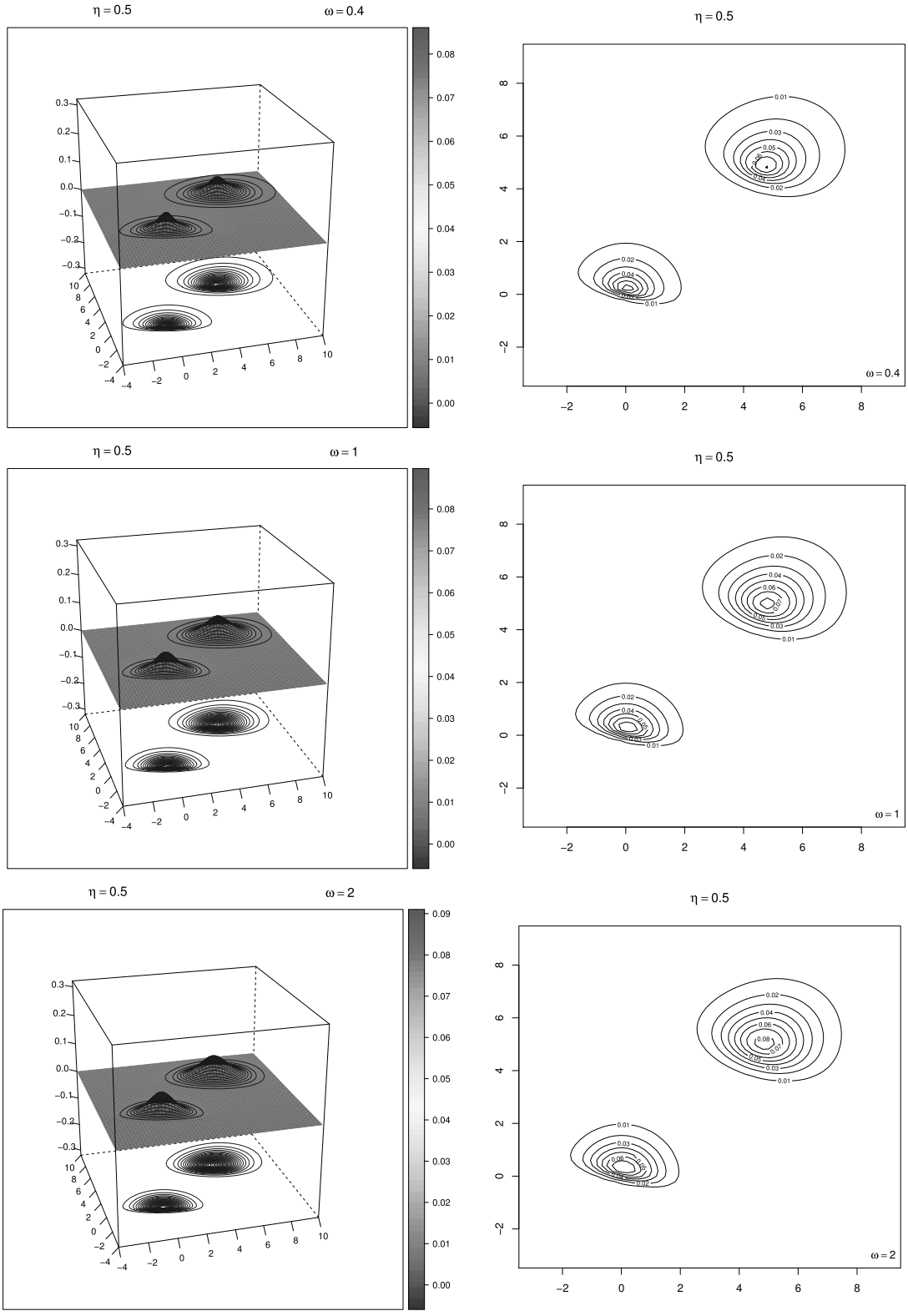
Figura 4 – Gráficos das superfícies (à esquerda) e suas respectivas curvas de nível (à direita) do modelo de misturas finitas de SNGH multivariado para $\eta$ = -0,5 fixo.

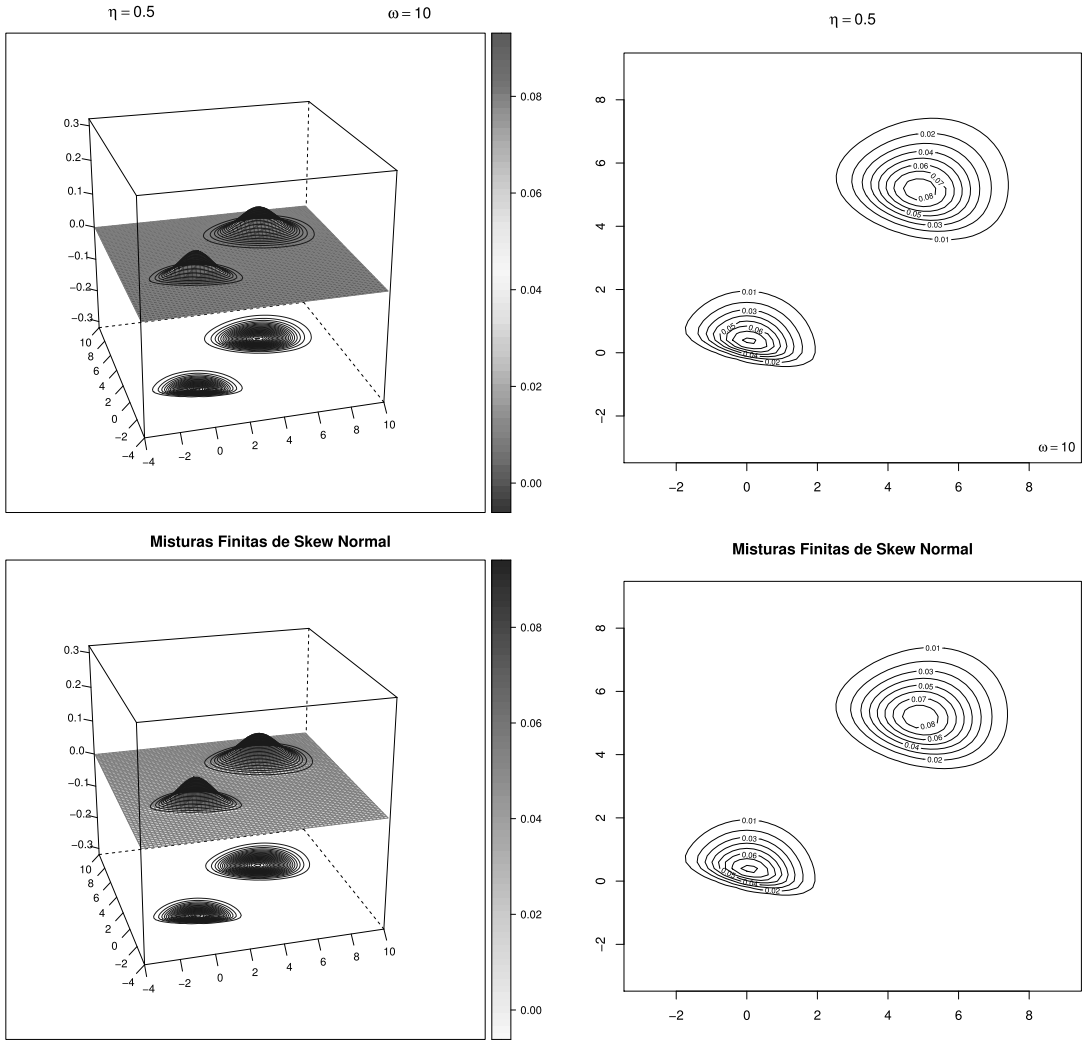





Fonte: O autor (2020).

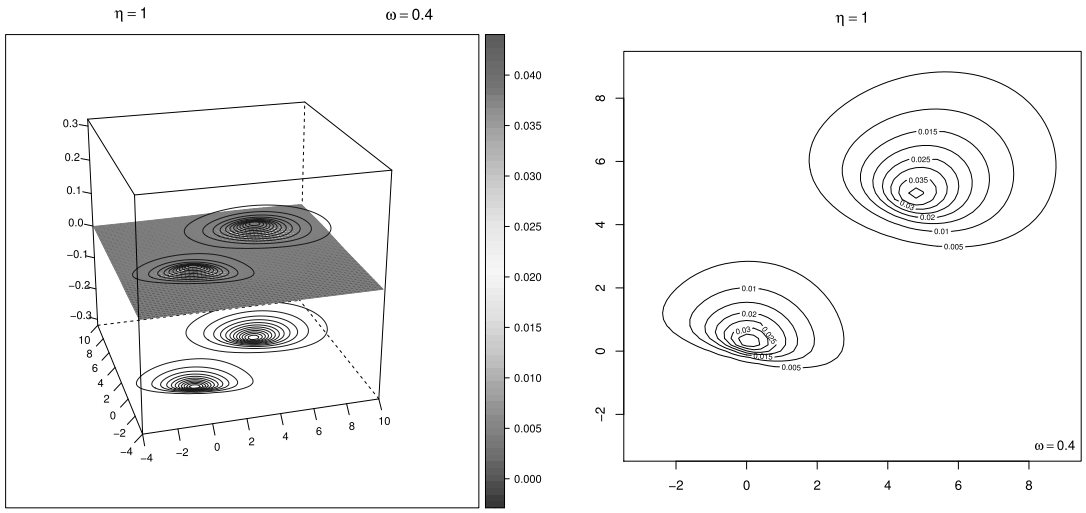
Figura 5 – Gráficos das superfícies (à esquerda) e suas respectivas curvas de nível (à direita) do modelo de misturas finitas de SNGH multivariado para $\eta = 0{,}5$ fixo.

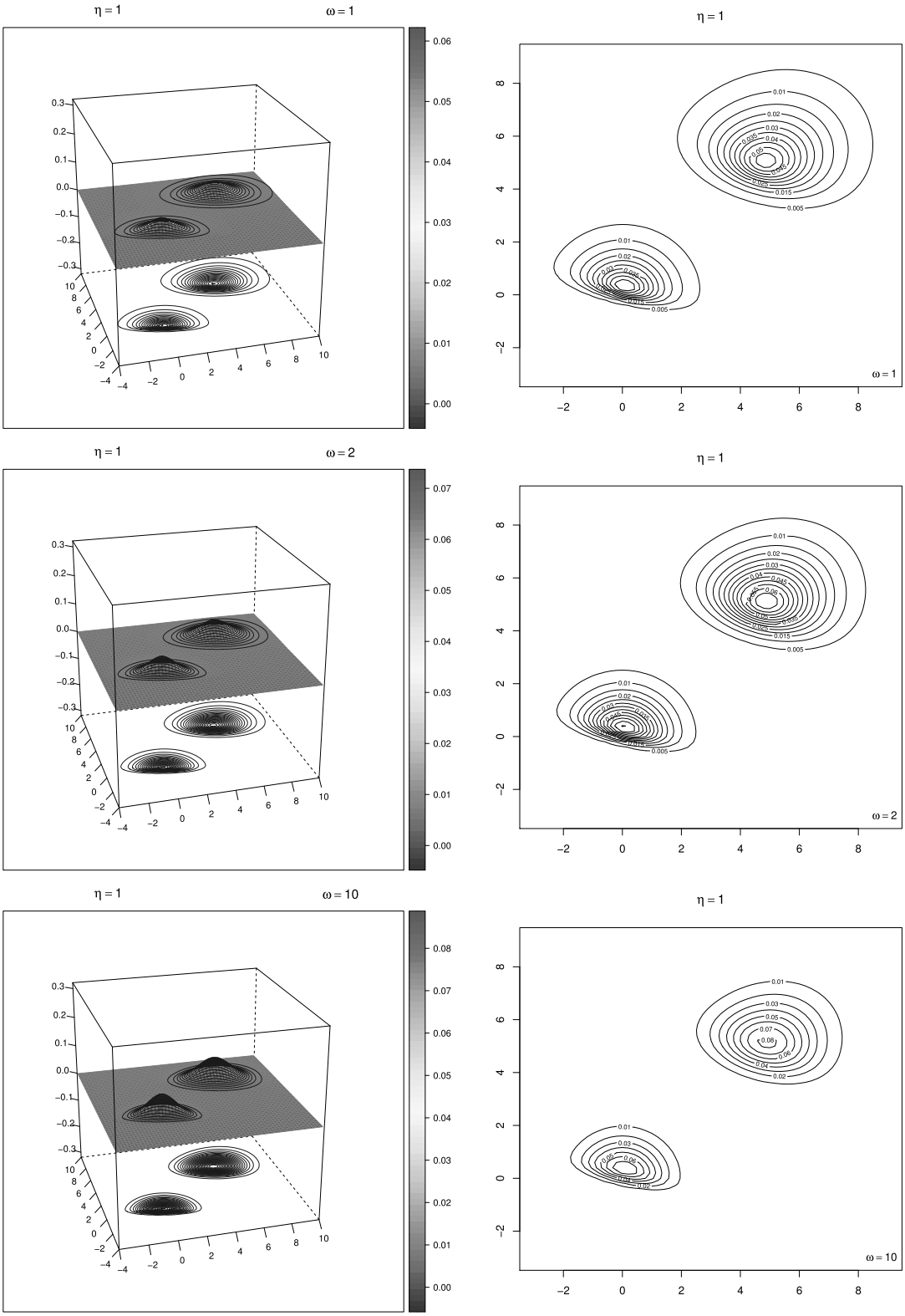


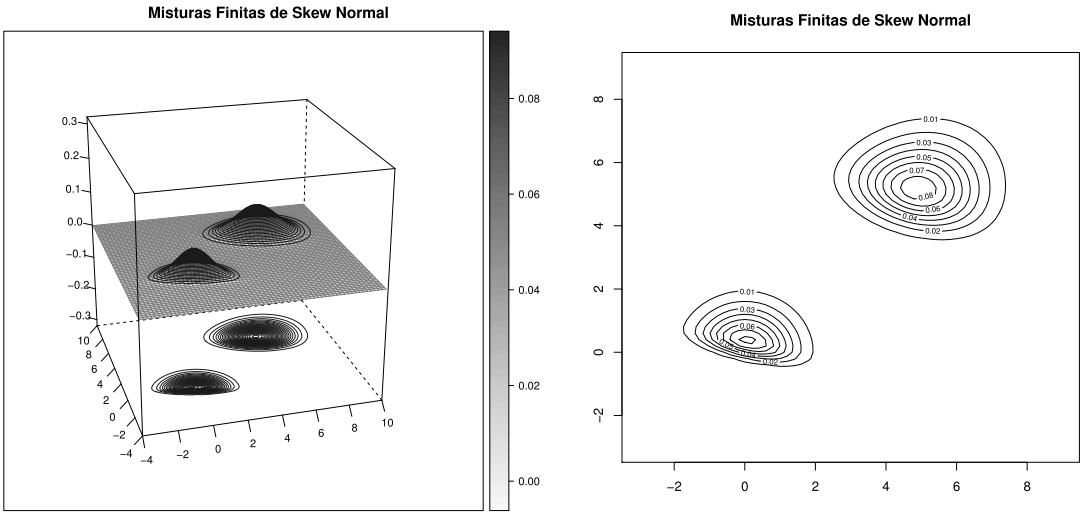



Fonte: O autor (2020).

Figura 6 – Gráficos das superfícies (à esquerda) e suas respectivas curvas de nível (à direita) do modelo de misturas finitas de SNGH multivariado para $\eta = 1$ fixo.




η = 1
ω = 1
η = 1
ω = 1
η = 1
ω = 2
η = 1
ω = 2
η = 1
ω = 10
η = 1
ω = 10




Fonte: O autor (2020).

Será realizada uma comparação entre os modelos de misturas finitas de SNGH com os modelos de misturas finitas de SMSN (a saber, FM-SN e FM-ST) no contexto de alguns conjuntos de dados reais. Tais resultados podem ser encontrados no próximo capítulo, na Seção 4.2.

### 3.2.3 **Representação hierárquica**

Como na Seção 3.1.3, será introduzido um vetor de dados latentes $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})^\top$, $i = 1, \ldots, n$, onde tal vetor é o de dados não observáveis e tem por finalidade associar a $i$-ésima observação da amostra a uma das $G$ componentes de misturas consideradas. Desta maneira, a distribuição de

$$\mathbf{Y}_i | Z_{ij} = 1 \sim \text{SNGH}_p(\boldsymbol{\theta}_j) \quad \text{e} \quad \mathbf{Z}_i \sim \text{Multi}(1; \rho_1, \ldots, \rho_G). \tag{3.16}$$

De (3.16) e da representação hierárquica de um vetor aleatório com distribuição SNGH apresentado por Calegari (2020) [25], tem-se o resultado a seguir.

**Proposição 3.2.1.** *Seja $\mathbf{Y}_i \in \mathbb{R}^n$ uma amostra aleatória, para $i = 1 \ldots, n$, que tem densidade dada por (3.15). Então, para cada vetor aleatório* SNGH *dessa amostra, admite-se a seguinte representação hierárquica:*

$$\begin{aligned}
\mathbf{Y}_i | T_i = t_i, U_i = u_i, Z_{ij} = 1 &\sim \text{N}_p\left(\boldsymbol{\mu}_j + \boldsymbol{\Delta}_j t_i, u_i \boldsymbol{\Gamma}_j\right), &(3.17)\\
T_i | U_i = u_i, Z_{ij} = 1 &\sim \text{HN}_1\left(0, u_i\right), &(3.18)\\
U_i | Z_{ij} = 1 &\sim \text{GIG}\left(u_i; \boldsymbol{\nu}\right), &(3.19)\\
\mathbf{Z}_i &\sim \text{Multi}(1; \rho_1, \ldots, \rho_G), &(3.20)
\end{aligned}$$

*onde no contexto univariado, tem-se $\Delta_j = \sigma_j \delta_j$, $\Gamma_j = \sigma_j^2\left(1 - {\delta_j}^2\right)$ e $\delta_j = \lambda_j\left(1 + \lambda_j^2\right)^{-1/2}$, e no multivariado é dado por $\boldsymbol{\Delta}_j = {\boldsymbol{\Sigma}_j}^{1/2}\delta_j$, $\boldsymbol{\Gamma}_j = \boldsymbol{\Sigma}_j - \boldsymbol{\Delta}_j\boldsymbol{\Delta}_j^\top$ e $\delta_j = \boldsymbol{\lambda}_j\left(1 + {\boldsymbol{\lambda}_j}^\top\boldsymbol{\lambda}_j\right)^{-1/2}$, a notação $HN_1(\cdot;\cdot)$ denota a distribuição half-normal.*

*Demonstração.* A prova é dada por meio das esperanças condicionais da representação estocástica da SNGH apresentada em (2.13). □

### 3.2.4 O Algoritmo EM em modelos de misturas finitas de densidades hiperbólica generalizada normal assimétrica

Nesta subseção, é apresentado o algoritmo EM, conforme descrito na Seção 3.1.3, para as estimativas dos parâmetros do modelo de misturas finitas de SNGH, no caso univariado e multivariado. Contudo, antes de iniciar sua implementação, são feitas umas considerações sobre algumas condições relacionadas a classe de distribuições SNGH, como identificabilidade e a complexidade na estimação do parâmetro $\eta$ associado a distribuição de mistura $U$.

O problema de identificabilidade da distribuição SNGH também ocorre na distribuição Hiperbólica Generalizada (GH). A distribuição GH foi proposta por Barndorff-Nielsen (1997) [16], onde para resolver esta questão, foi assumido que $|\boldsymbol{\Sigma}| = 1$. Essa condição pode ser considerada nas distribuições SNGH para solucionar o problema de identificabilidade. No entanto, de acordo com Browne & McNicholas (2015) [24], esta imposição restringe muito o uso da distribuição, por exemplo, nas aplicações com interesses em classificação. Uma alternativa é considerar algumas restrições nos parâmetros da distribuição de mistura $U$, na qual viabiliza relaxar a condição de que $|\boldsymbol{\Sigma}| = 1$, fazendo considerações sobre os parâmetros da distribuição GIG [25].

Neste trabalho, considera-se $U \sim \text{GIG}(\eta, \omega, \omega)$, ou seja, quando estabeleceu-se que $\psi = \gamma = \omega$. Além disso, assume-se $\eta$ fixo (conhecido) a fim de evitar as dificuldades relacionadas à estimação de tal parâmetro, conforme relatadas em Barndorff-Nielsen (1994) [15], Protassov (2004) [68], Snoussi & Idier (2006) [75], Browne & McNicholas (2015) [24] e Wraith & Forbes (2015) [81]. Esta será a ideia aplicada para a obtenção do algoritmo EM. Desta forma, o vetor de parâmetros univariado será $\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2, \lambda_j, \omega)^\top$ e multivariado $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j^\top, \boldsymbol{\alpha}_j^\top, \boldsymbol{\lambda}_j^\top, \omega)^\top$, para $j = 1, \ \ldots, G$.

Da representação hierárquica (3.17) - (3.20) e considerando o vetor de dados observados $\mathbf{y} = \left(\mathbf{y}_1{}^\top, \ \ldots, \mathbf{y}_n{}^\top\right)^\top$, bem como os vetores de dados não observáveis $\mathbf{u} = (u_1, \ \ldots, u_n)^\top$, $\mathbf{t} = (t_1, \ \ldots, t_n)^\top$ e $\mathbf{z} = \left(\mathbf{z}_1{}^\top, \ \ldots, \mathbf{z}_n{}^\top\right)^\top$, tem-se que o vetor de dados completos é dado por $\mathbf{y}_c = \left(\mathbf{y}^\top, \mathbf{u}^\top, \mathbf{t}^\top, \mathbf{z}^\top\right)^\top$. Assim, determina-se a função de verossimilhança dos dados completos $\mathbf{y}_c$ da seguinte forma

$$\begin{aligned} L_c(\boldsymbol{\Theta}) &= L_c(\boldsymbol{\Theta}|\mathbf{y}_c) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{G}\left[\rho_j \phi_p(\mathbf{y}; \boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma}) \phi_1(t_i; 0, u_i) f_{\text{GIG}}(u_i; \boldsymbol{\nu})\right]^{z_{ij}}, \end{aligned} \tag{3.21}$$

onde $\phi_p$ é o mesmo da Definição 2.2 e $f_{GIG}$ está definida em (2.12).

Desta maneira, nas seções a seguir, serão apresentados os Estimadores de Máxima Verossimilhança (EMV) via Algoritmo EM para os modelos de misturas finitas de SNGH no contexto univariado e multivariado, respectivamente.

#### 3.2.4.1 *Estimador de máxima verossimilhança via algoritmo EM no contexto univariado*

A partir deste momento, será mostrado a construção do algoritmo EM para a estimação por máxima verossimilhança dos parâmetros do modelo de misturas de SNGH univariado. Sendo assim, usando do resultado apresentado em (3.21), a função de log-verossimilhança completa de $\boldsymbol{\Theta} = ((\boldsymbol{\theta}_1{}^\top, \rho_1),\ \ldots, (\boldsymbol{\theta}_G{}^\top, \rho_G))^\top$, considerando $p = 1$, onde $\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2, \lambda_j, \omega)^\top$ é dada por

$$\begin{aligned} l_c(\boldsymbol{\Theta}|\mathbf{y}_c) = \quad & c \quad + \sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij}\left[\log(\rho_j) - \frac{1}{2}\log|\Gamma_j| - \frac{u_i^{-1}}{2\Gamma_j}(\mathbf{y}_i - \mu_j - \Delta_j t_i)^2\right. \\ & + \quad \left.\log(\eta-1)\log(u_i) - \log K_\eta(\omega) - \frac{\omega}{2}\left(u_i^{-1} + u_i\right)\right], \\ = \quad & c \quad + \sum_{i=1}^{n}\sum_{j=1}^{G}\left[Z_{ij}\log(\rho_j) - \frac{1}{2}Z_{ij}\log|\Gamma_j| - Z_{ij}\frac{u_i^{-1}}{2\Gamma_j}(\mathbf{y}_i - \mu_j - \Delta_j t_i)^2\right. \\ & + \quad \left. Z_{ij}(\eta-1)\log(u_i) - Z_{ij}\log K_\eta(\omega) - Z_{ij}\frac{\omega}{2}\left(u_i^{-1} + u_i\right)\right], \end{aligned} \tag{3.22}$$

sendo $c$ uma constante independente do parâmetro $\boldsymbol{\Theta}$.

Assim, tem-se que a função $Q$ é da seguinte maneira

$$\begin{aligned} Q(\boldsymbol{\Theta}; \widehat{\boldsymbol{\Theta}}^{(k)}) = \quad & E \quad \left[l_c(\boldsymbol{\Theta}|\mathbf{y}_c)|\mathbf{y}, \widehat{\boldsymbol{\Theta}^{(k)}}\right] \\ = \quad & c \quad + \sum_{i=1}^{n}\sum_{j=1}^{G}\left[\widehat{z_{ij}}^{(k)}\log(\rho_j) - \frac{1}{2}\widehat{z_{ij}}^{(k)}\log|\Gamma_j| - \frac{1}{2\Gamma_j}\widehat{zu_{ij}}^{(k)}(\mathbf{y}_i - \mu_j)^2\right. \\ & + \quad \widehat{zut_{ij}}^{(k)}(\mathbf{y}_i - \mu_j){\Gamma_j}^{-1}\Delta_j + \frac{1}{2\Gamma_j}\widehat{zut^2{}_{ij}}^{(k)}\Delta_j^2 \\ & + \quad \left.(\eta-1)\widehat{zs_{ij}}^{(k)} - \widehat{z_{ij}}^{(k)}\log K_\eta(\omega) - \frac{\omega}{2}\left(\widehat{zu_{ij}}^{(k)} + \widehat{zv_{ij}}^{(k)}\right)\right], \end{aligned} \tag{3.23}$$

em que os seguintes valores esperados são

$$\begin{aligned} \widehat{z_{ij}}^{(k)} &= E\left[Z_{ij}|\mathbf{y}_i, \widehat{\boldsymbol{\Theta}}^{(k)}\right], \quad \widehat{zu_{ij}}^{(k)} = E\left[Z_{ij}{U_i}^{-1}|\mathbf{y}_i, \widehat{\boldsymbol{\Theta}}^{(k)}\right], \\ \widehat{zut_{ij}}^{(k)} &= E\left[Z_{ij}{U_i}^{-1}T_i|\mathbf{y}_i, \widehat{\boldsymbol{\Theta}}^{(k)}\right], \quad \widehat{zut^2{}_{ij}}^{(k)} = E\left[Z_{ij}{U_i}^{-1}{T_i}^2|\mathbf{y}_i, \widehat{\boldsymbol{\Theta}}^{(k)}\right], \\ \widehat{zs_{ij}}^{(k)} &= E\left[Z_{ij}\log(U_i)|\mathbf{y}_i, \widehat{\boldsymbol{\Theta}}^{(k)}\right] \quad \text{e} \quad \widehat{zv_{ij}}^{(k)} = E\left[Z_{ij}U_i|\mathbf{y}_i\widehat{\boldsymbol{\Theta}}^{(k)}\right]. \end{aligned}$$

Consequentemente, usando do resultado (3.9) e assumindo $g$ como a densidade da SNGH dada em (2.16), no caso univariado, tem-se que

$$\widehat{z_{ij}}^{(k)} = \frac{\widehat{\rho_j}^{(k)} f_{\text{SNGH}}(\mathbf{y}_i; \widehat{\boldsymbol{\theta}_j}^{(k)})}{\sum_{j=1}^{G}\widehat{\rho_j}^{(k)} f_{\text{SNGH}}(\mathbf{y}_i; \widehat{\boldsymbol{\theta}_j}^{(k)})}. \tag{3.24}$$

Em seguida, ao fazer uso de algumas propriedades já conhecidas de esperança condicional, pode-se obter a expressão para $\widehat{zu_{ij}}$ através da seguinte relação

$$E\left[Z_{ij}{U_i}^{-1}|\mathbf{Y}_i\right] = E\left[E\left[Z_{ij}{U_i}^{-1}|\mathbf{Y}_i, Z_{ij}\right]|\mathbf{Y}_i\right] = E\left[Z_{ij}E\left[{U_i}^{-1}|\mathbf{Y}_i, Z_{ij}\right]|\mathbf{Y}_i\right], \quad (3.25)$$

e então é possível reescrever $\widehat{zu_{ij}}$ como

$$\widehat{zu_{ij}}^{(k)} = \widehat{z_{ij}}^{(k)}\widehat{u_{ij}}^{(k)}. \quad (3.26)$$

De modo análogo e com algumas manipulações algébricas, as outras esperanças condicionais são dadas por

$$\begin{array}{rcl} \widehat{zut_{ij}}^{(k)} & = & \widehat{z_{ij}}^{(k)}\widehat{ut_{ij}}^{(k)}, \quad \widehat{zut^2{}_{ij}}^{(k)} = \widehat{z_{ij}}^{(k)}\widehat{ut^2{}_{ij}}^{(k)}, \\ \widehat{zs_{ij}}^{(k)} & = & \widehat{z_{ij}}^{(k)}\widehat{s_{ij}}^{(k)} \quad \text{e} \quad \widehat{zv_{ij}}^{(k)} = \widehat{z_{ij}}^{(k)}\widehat{v_{ij}}^{(k)}, \end{array} \quad (3.27)$$

onde $\widehat{u_{ij}}^{(k)} = E\left[{U_i}^{-1}|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right]$, $\widehat{ut_{ij}}^{(k)} = E\left[{U_i}^{-1}T_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right]$, $\widehat{ut^2{}_{ij}}^{(k)} = E\left[{U_i}^{-1}{T_i}^2|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\right]$, $\widehat{s_{ij}}^{(k)} = E\left[\log U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\right]$ e $\widehat{v_{ij}}^{(k)} = E\left[U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\right]$ são apresentados com mais detalhes no Apêndice B e para maiores detalhes relacionados aos cálculos, veja Calegari (2020) [25].

Portanto, obtém-se o seguinte algoritmo EM:

**Etapa 1**: Dado $\eta$ fixo, forneça o chute inicial $\boldsymbol{\Theta}^{(0)} = \left((\widehat{\boldsymbol{\theta}_1}^{(0)}, \widehat{\rho_1}^{(0)}), \ \ldots, (\widehat{\boldsymbol{\theta}_G}^{(0)}, \widehat{\rho_G}^{(0)})\right)$.

**Etapa E**: Dado $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}^{(k)}$, calcule $\widehat{z_{ij}}^{(k)}$, $\widehat{zu_{ij}}^{(k)}$, $\widehat{zut_{ij}}^{(k)}$, $\widehat{zut^2{}_{ij}}^{(k)}$, $\widehat{zs_{ij}}^{(k)}$ e $\widehat{zv_{ij}}^{(k)}$, para $i = 1, \ \ldots, n$ e $j = 1, \ \ldots, G$.

**Etapa CM**: Para $j = 1, \ \ldots, G$, atualizar $\widehat{\rho_j}^{(k)}$, $\widehat{\mu_j}^{(k)}$, $\widehat{\Delta_j}^{(k)}$, $\widehat{\sigma_j^2}^{(k)}$ e $\widehat{\lambda_j}^{(k)}$ utilizando as seguintes expressões fechadas para os estimadores dos parâmetros de interesse.

$$\begin{array}{rcl} \widehat{\rho_j}^{(k+1)} & = & \dfrac{1}{n}\displaystyle\sum_{j=1}^{n}\widehat{z_{ij}}^{(k)}, \\ \widehat{\mu_j}^{(k+1)} & = & \left(\displaystyle\sum_{i=1}^{n}\widehat{zu_{ij}}^{(k)}\right)^{-1}\displaystyle\sum_{i=1}^{n}\left(\widehat{zu_{ij}}^{(k)}\mathbf{y}_i - \widehat{zut_{ij}}^{(k)}\widehat{\Delta_j}^{(k)}\right), \\ \widehat{\Delta_j}^{(k+1)} & = & \left(\displaystyle\sum_{i=1}^{n}\widehat{zut^2{}_{ij}}^{(k)}\right)^{-1}\left[\displaystyle\sum_{i=1}^{n}\widehat{zut_{ij}}^{(k)}\left(\mathbf{y}_i - \widehat{\mu_j}^{(k+1)}\right)\right], \\ \widehat{\Gamma_j}^{(k+1)} & = & \left(\displaystyle\sum_{i=1}^{n}\widehat{z_{ij}}^{(k)}\right)^{-1}\displaystyle\sum_{i=1}^{n}\Big(\widehat{zu_{ij}}^{(k)}\left(\mathbf{y}_i - \widehat{\mu_j}^{(k+1)}\right)^2 \\ & - & 2\left(\mathbf{y}_i - \widehat{\mu_j}^{(k+1)}\right)\widehat{\Delta_j}^{(k+1)}\widehat{zut_{ij}}^{(k)} + \widehat{\Delta_j^2}^{(k+1)}\widehat{zut^2{}_{ij}}^{(k)}\Big), \\ \widehat{\sigma_j^2}^{(k+1)} & = & \widehat{\Gamma_j}^{(k+1)} + \widehat{\Delta_j^2}^{(k+1)}, \\ \widehat{\lambda_j}^{(k+1)} & = & \left[\left(\widehat{\sigma_j^2}^{(k+1)}\right)^{-1/2}\widehat{\Delta_j}^{(k+1)}\right]\left(1 - \left(\widehat{\sigma_j^2}^{(k+1)}\right)^{-1}\widehat{\Delta_j^2}^{(k+1)}\right)^{-1/2}. \end{array}$$

Conforme foi mencionado no início da Seção 3.2.4, os estimadores acima são obtidos ao assumir o vetor $\boldsymbol{\nu} = (\eta, \omega, \omega)^\top$, com $\eta$ fixo. Esta estrategia é adotada pois existe uma

certa complexidade em estimar o parâmetro $\eta$. Apesar de considerar um tamanho amostral suficientemente grande, a log-verossimilhança em relação a $\eta$ acaba se tornando *flat*. Nesse caso, devido essa dificuldade, considera-se $\eta$ fixo, assim pode-se concentrar na estimação de $\omega$ por meio de uma modificação do algoritmo EM, como mostra a etapa a seguir.

**Etapa CML**: Para $j = 1, \ldots, G$, atualizar $\widehat{\omega}^{(k)}$ maximizando a função $Q$ (3.23) sobre $\omega$, obtém-se

$$\widehat{\omega}^{(k+1)} = \text{argmax } Q_\omega \left( \omega \left| (\widehat{\boldsymbol{\theta}_j}^{(k)}, \widehat{\rho_j}^{(k)}), \ \ldots, (\widehat{\boldsymbol{\theta}_G}^{(k)}, \widehat{\rho_G}^{(k)}) \right. \right).$$

3.2.4.2 *Estimador de máxima verossimilhança via algoritmo EM no contexto multivariado*

Agora para o contexto multivariado, serão obtidos os estimadores de máxima verossimilhança via algoritmo EM do modelo de misturas finitas de SNGH. Para isso, define-se a função de log-verossimilhança completa de $\boldsymbol{\Theta} = ((\boldsymbol{\theta}_1{}^\top, \rho_1), \ \ldots, (\boldsymbol{\theta}_G{}^\top, \rho_G))^\top$, onde $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\alpha}_j, \boldsymbol{\lambda}_j, \omega)^\top$, com $\boldsymbol{\alpha}_j$ o vetor de elementos da matriz triangular superior $\boldsymbol{\Sigma}$, da seguinte maneira

$$\begin{aligned}
l_c(\boldsymbol{\Theta}|\mathbf{y}_c) = \ & C \ + \sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij} \Big[ \log(\rho_j) - \frac{1}{2}\log|\boldsymbol{\Gamma}_j| \\
& - \ \frac{u_i^{-1}}{2}\left(\mathbf{y}_i - \boldsymbol{\mu}_j - \boldsymbol{\Delta}_j t_i\right)^\top \boldsymbol{\Gamma}_j^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}_j - \boldsymbol{\Delta}_j t_i\right) + \\
& + \ \log(\eta-1)\log(u_i) - \log K_\eta(\omega) - \frac{\omega}{2}\left(u_i^{-1} + u_i\right) \Big] \\
= \ & C \ + \sum_{i=1}^{n}\sum_{j=1}^{G} \Big[ Z_{ij}\log(\rho_j) - \frac{1}{2} Z_{ij}\log|\boldsymbol{\Gamma}_j| \\
& - \ Z_{ij}\frac{(u_i^{-1})}{2}\left(\mathbf{y}_i - \boldsymbol{\mu}_j\right)^\top \boldsymbol{\Gamma}_j^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}_j\right) \\
& + \ Z_{ij}(u_i{}^{-1}) t_i \left(\mathbf{y}_i - \boldsymbol{\mu}_j\right)^\top \boldsymbol{\Gamma}_j{}^{-1}\boldsymbol{\Delta}_j + Z_{ij}\frac{(u_i^{-1})t_i^2}{2}\boldsymbol{\Delta}_j^\top \boldsymbol{\Gamma}_j{}^{-1}\boldsymbol{\Delta}_j \\
& + \ Z_{ij}(\eta-1)\log(u_i) - Z_{ij}\log K_\eta(\omega) - Z_{ij}\frac{\omega}{2}\left(u_i^{-1} + u_i\right) \Big], \qquad (3.28)
\end{aligned}$$

sendo $C$ uma constante independente do parâmetro $\boldsymbol{\Theta}$.

Em seguida a função $Q$ é obtida da seguinte maneira

$$\begin{aligned}
Q(\boldsymbol{\Theta}; \widehat{\boldsymbol{\Theta}}^{(k)}) \ &= \ E\left[ l_c(\boldsymbol{\Theta}|\mathbf{y}_c) | \mathbf{y}, \widehat{\boldsymbol{\Theta}}^{(k)} \right] \\
&= \ C + \sum_{i=1}^{n}\sum_{j=1}^{G} \Big[ \widehat{z_{ij}}^{(k)} \log(\rho_j) - \frac{1}{2}\widehat{z_{ij}}^{(k)} \log|\boldsymbol{\Gamma}_j| \\
&- \ \frac{1}{2}\widehat{zu_{ij}}^{(k)}\left(\mathbf{y}_i - \boldsymbol{\mu}_j\right)^\top \boldsymbol{\Gamma}_j^{-1}\left(y_i - \boldsymbol{\mu}_j\right) \\
&+ \ \widehat{zut_{ij}}^{(k)}\left(\mathbf{y}_i - \boldsymbol{\mu}_j\right)^\top \boldsymbol{\Gamma}_j{}^{-1}\boldsymbol{\Delta}_j + \frac{1}{2}\widehat{zut^2_{ij}}^{(k)}\boldsymbol{\Delta}_j^\top \boldsymbol{\Gamma}_j{}^{-1}\boldsymbol{\Delta}_j \\
&+ \ (\eta-1)\widehat{zs_{ij}}^{(k)} - \widehat{z_{ij}}^{(k)} \log K_\eta(\omega) - \frac{\omega}{2}\left(\widehat{zu_{ij}}^{(k)} + \widehat{zv_{ij}}^{(k)}\right) \Big], \qquad (3.29)
\end{aligned}$$

em que as quantidades, $\widehat{z_{ij}}^{(k)}$, $\widehat{zu_{ij}}^{(k)}$, $\widehat{zut_{ij}}^{(k)}$, $\widehat{zut^2{}_{ij}}^{(k)}$, $\widehat{zs_{ij}}^{(k)}$ e $\widehat{zv_{ij}}^{(k)}$ serão obtidas de forma análoga aos resultados demonstrados para o algoritmo EM no contexto univariado (3.2.4.1).

Portanto, a expressão para $\widehat{z_{ij}}^{(k)}$ é determinada usando do resultado (3.9) junto a distribuição SNGH definida em (2.15). Já as demais esperanças condicionais, $\widehat{zu_{ij}}^{(k)}$, $\widehat{zut_{ij}}^{(k)}$, $\widehat{zut^2{}_{ij}}^{(k)}$, $\widehat{zs_{ij}}^{(k)}$ e $\widehat{zv_{ij}}^{(k)}$ podem ser calculadas através das fórmulas apresentadas em (3.26) e (3.27), na qual seus resultados finais, com maiores detalhes, encontram-se no Apêndice B. Assim, com as esperanças condicionais determinadas, dá-se continuidade ao desenvolvimento do algoritmo EM para a estimação por máxima verossimilhança do parâmetro $\boldsymbol{\Theta}$ como segue.

**Etapa 1**: Dado $\eta$ fixo, forneça o chute inicial $\boldsymbol{\Theta}^{(0)} = \left((\widehat{\boldsymbol{\theta}_1}^{(0)}, \widehat{\rho_1}^{(0)}),\ \ldots, (\widehat{\boldsymbol{\theta}_G}^{(0)}, \widehat{\rho_G}^{(0)})\right)$.

**Etapa E**: Dado $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}^{(k)}$, calcule $\widehat{z_{ij}}^{(k)}$, $\widehat{zu_{ij}}^{(k)}$, $\widehat{zut_{ij}}^{(k)}$, $\widehat{zut^2{}_{ij}}^{(k)}$, $\widehat{zs_{ij}}^{(k)}$ e $\widehat{zv_{ij}}^{(k)}$, para $i = 1,\ \ldots, n$ e $j = 1,\ \ldots, G$.

**Etapa CM**: Para $j = 1,\ \ldots, G$, atualizar $\widehat{\boldsymbol{\Theta}}^{(k+1)}$ maximizando a função (3.29) em relação à $\boldsymbol{\Theta}$, obtendo as seguintes expressões fechadas.

$$
\begin{aligned}
\widehat{\rho_j}^{(k+1)} &= \frac{1}{n}\sum_{j=1}^{n}\widehat{z_{ij}},\\
\widehat{\boldsymbol{\mu}_j}^{(k+1)} &= \sum_{i=1}^{n}\left(\widehat{zu_{ij}}^{(k)}\mathbf{y}_i - \widehat{zut_{ij}}^{(k)}\widehat{\boldsymbol{\Delta}_j}^{(k)}\right)\left(\sum_{i=1}^{n}\widehat{zu_{ij}}^{(k)}\right)^{-1},\\
\widehat{\boldsymbol{\Delta}_j}^{(k+1)} &= \left[\sum_{i=1}^{n}\widehat{zut_{ij}}^{(k)}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}_j}^{(k+1)}\right)\right]\left(\sum_{i=1}^{n}\widehat{zut^2{}_{ij}}^{(k)}\right)^{-1},\\
\widehat{\boldsymbol{\Gamma}_j}^{(k+1)} &= \left(\sum_{i=1}^{n}\widehat{z}_{ij}^{(k)}\right)^{-1}\sum_{i=1}^{n}\left(\widehat{zu_{ij}}^{(k)}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}_j}^{(k+1)}\right)\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}_j}^{(k+1)}\right)^{\top}\right.\\
&- \left[\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}_j}^{(k+1)}\right)\left(\widehat{\boldsymbol{\Delta}_j}^{(k+1)}\right)^{\top} + \widehat{\boldsymbol{\Delta}_j}^{(k+1)}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}_j}^{(k+1)}\right)^{\top}\right]\widehat{zut_{ij}}^{(k)}\\
&+ \left.\widehat{\boldsymbol{\Delta}_j}^{(k+1)}\left(\widehat{\boldsymbol{\Delta}_j}^{(k+1)}\right)^{\top}\widehat{zut^2{}_{ij}}^{(k)}\right),\\
\widehat{\boldsymbol{\Sigma}_j}^{(k+1)} &= \widehat{\boldsymbol{\Gamma}_j}^{(k+1)} + \widehat{\boldsymbol{\Delta}_j}^{(k+1)}\widehat{\boldsymbol{\Delta}_j^{\top}}^{(k+1)},\\
\widehat{\boldsymbol{\lambda}_j}^{(k+1)} &= \left[\left(\widehat{\boldsymbol{\Sigma}_j}^{(k+1)}\right)^{-1/2}\widehat{\boldsymbol{\Delta}_j}^{(k+1)}\right]\left(1 - \widehat{\boldsymbol{\Delta}_j^{\top}}^{(k+1)}\left(\widehat{\boldsymbol{\Sigma}_j}^{(k+1)}\right)^{-1}\widehat{\boldsymbol{\Delta}_j}^{(k+1)}\right)^{-1/2}.
\end{aligned}
$$

Da mesma maneira que foi descrita na seção anterior para o caso univariado, com relação ao parâmetro $\eta$, também é considerado $\eta$ fixo no contexto multivariado.

**Etapa CML**: Atualizar $\widehat{\omega}^{(k)}$ maximizando a função $Q$ (3.29) em relação à $\omega$, o que leva à seguinte expressão.

$$
\widehat{\omega}^{(k+1)} = \operatorname{argmax}\, Q_{\omega}\left(\omega \,\middle|\, (\widehat{\boldsymbol{\theta}_1}^{(k)}, \widehat{\rho_1}^{(k)}),\ \ldots, (\widehat{\boldsymbol{\theta}_G}^{(k)}, \widehat{\rho_G}^{(k)})\right).
$$

#### 3.2.4.3 *Casos particulares com $\eta$ fixo*

Essas condições específicas são derivadas através dos casos particulares da distribuição GIG, os quais foram citados brevemente na Seção 2.3.1. Calegari (2020) [25] apresenta uma discussão mais detalhada sobre o assunto, onde são determinadas as funções densidades de probabilidade de cada caso, assim como suas diferentes formulações para o algoritmo EM considerando $\eta$ fixo e $\omega$ desconhecido para as variáveis misturas obtidas. Neste trabalho, estas discussões são expandidas para o contexto de misturas finitas de SNGH.

Sendo assim, no primeiro caso, Protassov (2004) [68] e Browne & McNicholas (2015) [24] comentam sobre o motivo da escolha de $\eta = {}^{-}1/2$, pois ao usar desse valor para tal parâmetro, quando $\psi = \gamma = \omega > 0$, a variável aleatória $U \sim \text{GIG}(\eta, \omega, \omega)$ converge para uma distribuição $\text{IG}(\omega, \omega)$. Assim, a estimação de $\omega$, para $i = 1, \ldots, n$ e $j = 1, \ldots, G$, na etapa M do algoritmo EM será da seguinte maneira

$$\widehat{\omega}^{(k+1)} = \left( \frac{\widehat{zu}_{ij}^{(k)} + \widehat{zv}_{ij}^{(k)}}{\widehat{z}_{ij}^{(k)}} - 2 \right)^{-1}. \tag{3.30}$$

Já para o segundo caso, em que $\eta < 0$, $\psi \longrightarrow 0$ e $\gamma = \omega$, Calegari (2020) [25] mostrou que a variável aleatória $U$ com distribuição GIG converge para uma distribuição inversa gama e nessa situação, se $U \sim \text{I}Gamma({}^{\omega}/_{2}, {}^{\omega}/_{2})$, então tem-se que $\mathbf{Y} \sim \text{ST}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \omega)$, onde algumas de suas propriedades são discutidas em Basso et al. (2010) [18], Lin (2010) [54], Ho & Lin (2010) [41] e Lee & McLachlan (2014) [51], no contexto de misturas finitas de densidades.

O último dos casos particulares para a variável aleatória $U \sim \text{GIG}(\eta, \psi, \gamma)$ é quando $\eta > 0$, $\psi \longrightarrow 0$ e $\gamma = \omega$, ou seja, $U \sim \text{Gamma}(\eta, {}^{\omega}/_{2})$ e então, tem-se a generalização da distribuição variância-gama. Agora se $\eta = 1$ e $\omega = 2$, isto é, $U \sim \text{Exp}(1)$ encontra-se a generalização da distribuição assimétrica de Laplace [25]. Na literatura podem ser encontrados estudos sobre a distribuição Laplace em Kozubowski et al. (2000) [47], Kotz et al. (2001) [46], entre outros.

Cabe ressaltar que, as situações comentadas anteriormente, ao considerar $\eta$ fixo e $\omega$ desconhecido, com $\eta = {}^{-}1/2$ e $\eta = 1$ foram utilizadas nas aplicações numéricas que serão apresentadas no próximo capítulo. Além do mais, vale lembrar que os resultados mostrados neste capítulo generalizam os trabalhos de Vilca et al. (2014) [78] e Calegari (2020) [25], uma vez que, ao considerar $G = 1$, os resultados são retornados aos apresentados por Calegari (2020); veja [25], para mais detalhes.

# 4 APLICAÇÕES NUMÉRICAS

Neste capítulo, estudos de simulações e quatro aplicações em dados reais serão apresentados com o propósito de ilustrar o modelo e os resultados inferenciais aqui desenvolvidos; veja as Seções 4.1 e 4.2. Nesse trabalho, foram considerados os seguintes conjuntos de dados: *Body Mass Index*, para o contexto univariado, *Swiss Bank*, para o contexto multivariado e *Old Faitful*, em ambos os casos. Assim, pode-se analisar a flexibilidade do modelo FM-SNGH, comparando-o com outros modelos já conhecidos na literatura, por exemplo, os modelos de misturas finitas de normal assimétrico e $t$ assimétrica [18]. O procedimento de estimação por máxima verossimilhança, via algoritmo EM, no modelo de misturas finitas sob a classe de distribuições SMSN foi feito usando o pacote *mixsmsn* do programa R [66].

Para obtenção dos resultados apresentados neste trabalho, utilizou-se do *software* R na versão 4.0.2 instalado em um notebook com processador Intel Core i5-7200 de 2.5 GHz com *Turbo Boost up* para 3.1 GHz, 8 GB de memoria RAM DDR4, placa de video NVIDIA GeForce 940MX com 2 GB de RAM GDDR5 e sistema operacional de 64 bits - Windows 10 *Home Single Language*. Ressalta-se que, para o procedimento de estimação por máxima verossimilhança, via algoritmo EM, no modelo de misturas finitas sob a classe de distribuições SNGH foi programado no R [70] e usou-se (3.13) como critério de parada do algoritmo EM com $\epsilon = 10^{-4}$.

## 4.1 ESTUDOS DE SIMULAÇÃO

Esse estudo de simulação foi projetado com intuito de investigar o desempenho e as mudanças que ocorrem nas estimativas de máxima verossimilhança dos parâmetros do modelo FM-SNGH, nos contextos univariado e multivariado, com $G = 2$, ao variar os tamanhos amostrais, $n$, em 100, 300, 500 e 1000. De modo que, para cada valor escolhido de $n$, são geradas artificialmente 250 e 500 amostras, no contexto univariado e multivariado, respectivamente, através da representação estocástica apresentada em (2.13). Além disso, com intenção de descrever e entender melhor o comportamento dessas estimativas, o estudo de simulação é dividido em dois cenários, onde essa divisão é realizada de acordo com os valores adotados para o parâmetro $\eta$ (fixo). A escolha dos valores para tal parâmetro, foi realizada de acordo com os casos particulares conhecidos da distribuição GIG apresentados na Subseção 3.2.4.3. Dessa maneira, quando $\eta = {}^{-1}/_{2}$, tem-se uma distribuição Inversa Gaussiana (IG), onde pôde-se estimar $\omega$ através da fórmula fechada dada em (3.30). E para $\eta = {}^{1}/_{2}$, também pode-se obter $\omega$ a partir dessa expressão, uma vez que, ao usar da propriedade (i) da função modificada de Bessel do terceiro tipo, que encontra-se no Apêndice A, obtém-se a mesma fórmula fechada. No entanto, quando $\eta = 1$, é aplicada uma outra abordagem na estimação de $\omega$, na qual usou-se da função *optim* programada e

presente no código do programa R.

No Cenário 1, considera-se $\eta = {}^{-1}/_{2}$ para os contextos univariado e multivariado do modelo de FM-SNGH. Já para o Cenário 2, é considerado $\eta = 1$ para o caso univariado e $\eta = {}^{1}/_{2}$ para o multivariado. Na Tabela 1, apresentam-se as configurações assumidas para os parâmetros do modelo FM-SNGH em ambos contextos e cenários.

Tabela 1 – Configurações dos valores verdadeiros adotados para os parâmetros do modelo FM-SNGH.

| **Univariado** | | **Multivariado** | |
|---|---|---|---|
| **Parâmetros** | **Valores adotados** | **Parâmetros** | **Valores adotados** |
| $\rho_1$ | 0,4 | $\rho_1$ | 0,7 |
| $\rho_2$ | 0,6 | $\rho_2$ | 0,3 |
| $\mu_1$ | 15 | $\boldsymbol{\mu}_1$ | (0, 0) |
| $\mu_2$ | 20 | $\boldsymbol{\mu}_2$ | (5, 5) |
| ${\sigma^2}_1$ | 1 | $\boldsymbol{\Sigma}_1$ | $\mathbf{I}_2$ |
| ${\sigma^2}_2$ | 1 | $\boldsymbol{\Sigma}_2$ | $\begin{pmatrix} 2 & 0,5 \\ 0,5 & 2 \end{pmatrix}$ |
| $\lambda_1$ | -4 | $\boldsymbol{\lambda}_1$ | (1, 4) |
| $\lambda_2$ | -1 | $\boldsymbol{\lambda}_2$ | (1, 2) |
| $\omega$ | 1 | $\omega$ | 2 |

Fonte: Elaborada pelo autor (2020).

Com a primeira etapa do estudo de simulação definida, o próxima passo é discutir sobre os valores iniciais que são introduzidos ao algoritmo. No entanto, sabe-se que modelos de misturas finitas podem apresentar log-verossimilhança com mais de uma moda. Assim, a estimação de máxima verossimilhança via algoritmo EM, pode não resultar em máximos globais se os valores iniciais estiverem distantes dos verdadeiros valores dos parâmetros do modelo. Sendo assim, escolher valores iniciais para o algoritmo EM de forma correta, possui uma grande importância no contexto de estimação em modelos de misturas.

Posto isto, para que a proposta de chute inicial do algoritmo EM fosse bem sucedida, foi utilizada uma abordagem que coletasse informações dos parâmetros de proporção, locação, escala e assimetria através de uma estratégia similar à adotada por Biernacki (2000) [20] e Browne et al. (2015) [24]. Por exemplo, [24] usaram uma abordagem EM para inicialização. Especificamente, para cada um dos 100 valores iniciais aleatórios, o algoritmo EM foi executado por 50 iterações. O algoritmo com a log-verossimilhança dos dados observados mais alta (finita) após 50 iterações é então iterado até a convergência. Nesse sentido, por meio do software R, foi feito um pré-processamento ou pré-estimação para obter os valores inicias dos parâmetros, $\mu$, $\sigma^2$, $\lambda$ do modelo univariado, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\lambda}$ do modelo multivariado e $\rho$ em ambos casos, em cada tamanho de amostra $n$ gerada de

acordo com os cenários estipulados anteriormente, conforme os passos mostrados a seguir:

**Passo 1**: Estimar os parâmetros de locação, escala e assimetria, através do algoritmo EM, no contexto do modelo de misturas de normais assimétricas. Neste passo, usa-se o pacote *mixsmsn*;

**Passo 2**: Gerar um valor inicial de $\omega$ através da amostra aleatória uniforme no intervalo [0,20] de tamanho 100;

**Passo 3**: Maximizar a log-verossimilhança do modelo proposto com relação ao $\omega$, usando como ponto de partida o valor obtido do Passo 2, avaliada nas estimativas dos demais parâmetros, calculadas no Passo 1;

**Passo 4**: Executar o algoritmo EM, proposto neste trabalho, utilizando as estimavas iniciais dos parâmetros obtidas nos passos anteriores, com um número máximo de iterações igual a 25.

Os resultados obtidos no Passo 4 serão os valores inicias para o algoritmo EM do modelo FM-SNGH para G=2, nos casos univariado e multivariado. Para o critério de parada, foi adotado o mesmo em ambos os casos, no qual encontra-se descrito na Seção 3.1.3, dado pela expressão (3.13), com $\epsilon = 10^{-4}$.

De acordo cada cenário estabelecido, avalia-se o desempenho dos estimadores de máxima verossimilhança propostos em termos das medidas viés médio (BIAS) e erro quadrático médio (MSE), definidas a seguir. Considere $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$, tal que

$$\text{BIAS}(\phi_i) = \frac{1}{n_p}\sum_{j=1}^{n_p}\left|\widehat{\phi}_{ij}^{(k)} - \phi_i\right| \quad \text{e} \quad \text{MSE}(\phi_i) = \frac{1}{n_p}\sum_{j=1}^{n_p}\left(\widehat{\phi}_{ij}^{(k)} - \phi_i\right)^2,$$

em que $n_p$ é o número de simulações realizadas e $i = 1, \ldots, p$. No caso multivariado, é feito uma média dessas medidas de cada parâmetro da seguinte maneira

$$\text{MBIAS}(\boldsymbol{\phi}) = \frac{1}{p}\sum_{i=1}^{p}\text{BIAS}(\phi_i) \quad \text{e} \quad \text{MMSE}(\boldsymbol{\phi}) = \frac{1}{p}\sum_{i=1}^{p}\text{MSE}(\phi_i).$$

Cabe ressaltar que, no contexto de misturas finitas, a função de verossimilhança é invariante sob a permutação das proporções nos vetores de parâmetros. Então, pode ocorrer o problema de troca de *labels* quando algumas proporções da mistura permutam. Pode-se contornar esse problema escolhendo os *labels* minimizando a distância das estimativas, obtidas via algoritmo EM, aos valores verdadeiros do parâmetro $\rho$ [62, 85].

Sendo assim, nas próximas seções, com auxilio dos boxplots e gráficos serão descritos os resultados obtidos, de acordo com os cenários estabelecidos, deste estudo de simulação.

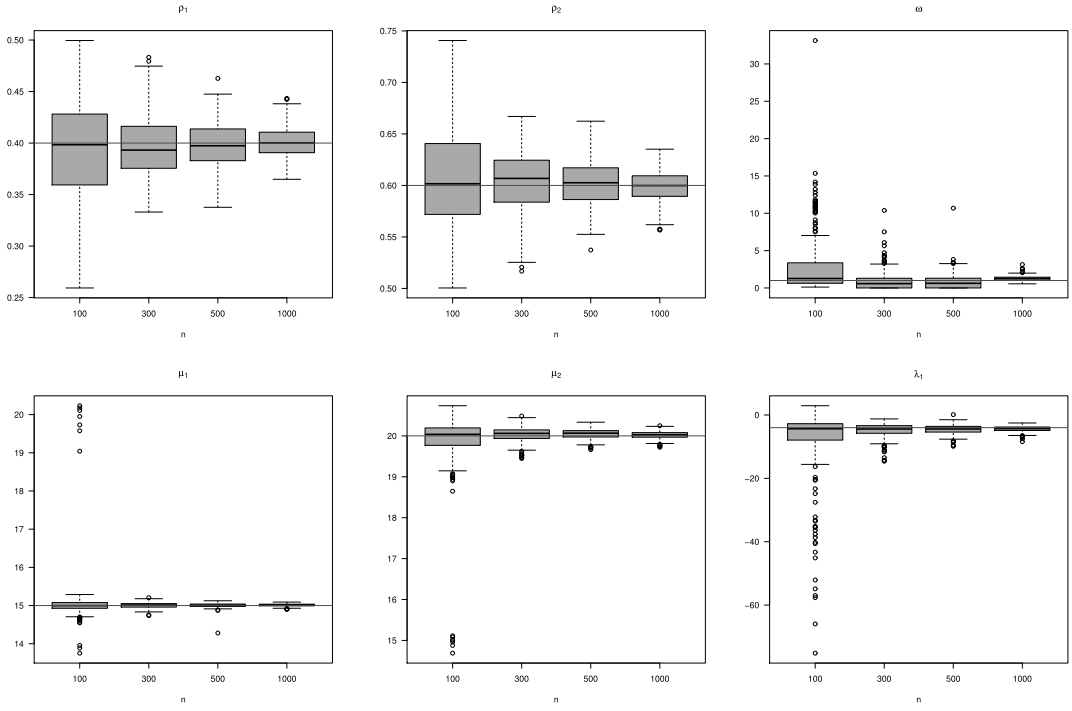
### 4.1.1 **Desempenho das estimativas de máxima verossimilhança no contexto univariado**

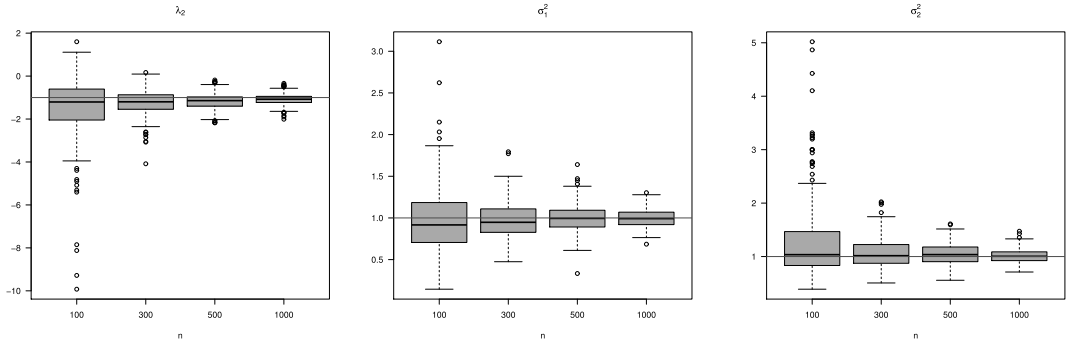
Como foi comentando anteriormente, nesta seção, serão descritos os comportamentos das estimativas de máxima verossimilhança dos parâmetros no modelo FM-SNGH univariado com dois grupos de mistura, nos dois cenários considerados de acordo com o valor adotado para o parâmetro $\eta$. É interessante ressaltar que o desempenho do algoritmo EM na recuperação dos valores verdadeiros dos parâmetros é mais satisfatório no cenário 1, como será observado nas Subseções 4.1.1.1 e 4.1.1.2. Além disso, as Tabelas 8 à 12 (consulte o Apêndice C) apresentam as estatísticas de resumo para as estimativas dos parâmetros do modelo FM-SNGH nesses cenários.

#### 4.1.1.1 *Cenário 1*

Após finalizada as simulações no contexto univariado para $\eta = {}^{-1}/_{2}$ (fixo). Observou-se que o algoritmo EM proposto foi capaz de recuperar os verdadeiros valores dos parâmetros. Quando o tamanho amostral aumenta, nota-se que as estimativas dos parâmetros de interesse se aproximam dos seus respectivos valores verdadeiros com redução da variabilidade. Desta maneira, na Figura 7 a seguir, encontram-se os boxplots das estimativas para cada parâmetro do modelo FM-SNGH univariado, com $G = 2$, de acordo com as configurações mostradas na Tabela 1.

Figura 7 – Boxplots das estimativas de $\rho_1$, $\rho_2$, $\omega$, $\mu_1$, $\mu_2$, $\lambda_1$, $\lambda_2$, $\sigma_1^2$ e $\sigma_2^2$ (linha vermelha indica o valor real dos parâmetros), do modelo FM-SNGH univariado, quando $\eta = {}^{-1}/_{2}$.



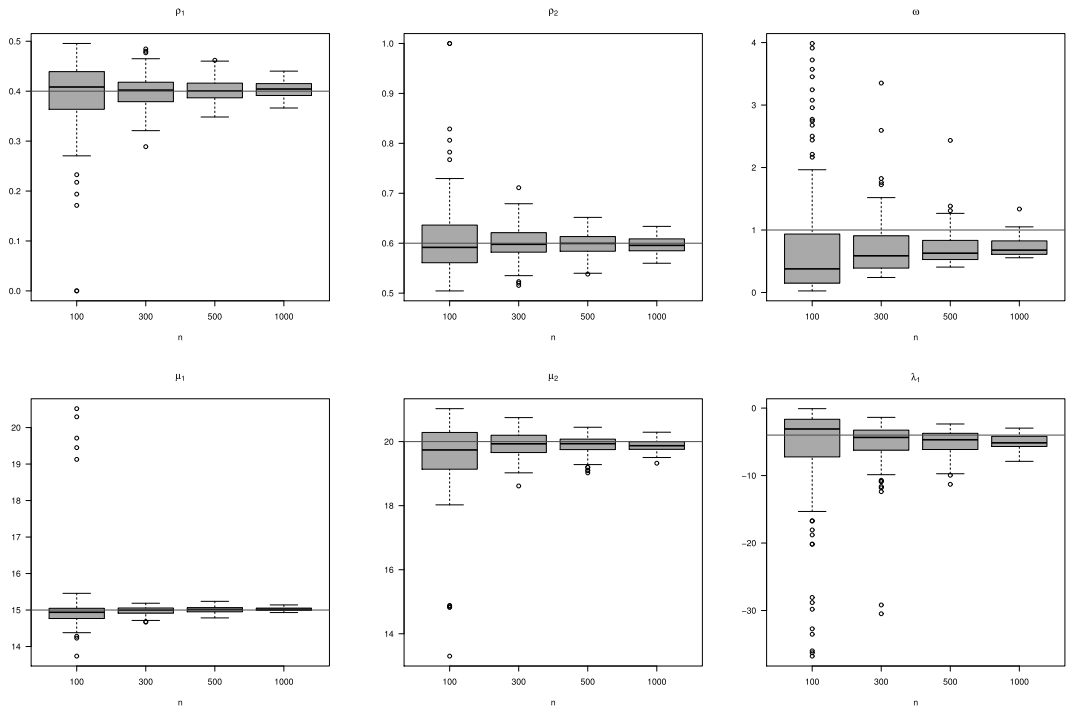


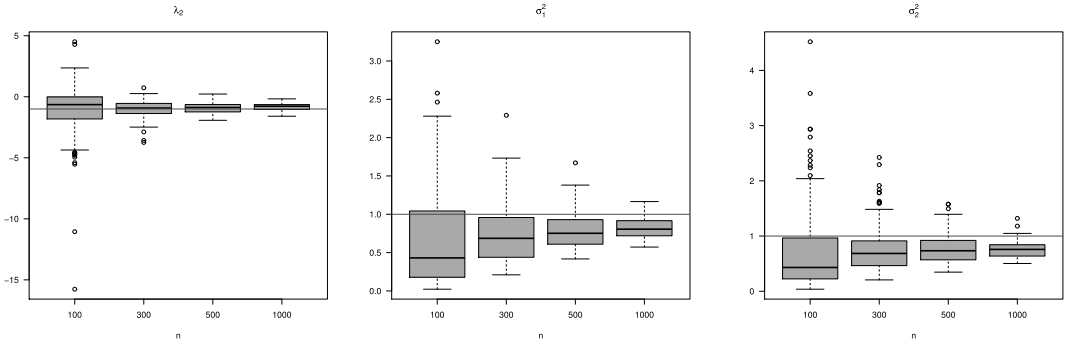

Fonte: O autor (2020).

#### 4.1.1.2 *Cenário 2*

Com as simulações no contexto univariado concluídas, quando $\eta = 1$ (fixo), foi observado que os resultados não foram tão satisfatórios como no Cenário 1. Embora tenha ocorrido uma melhora significante das estimativas, em termos de viés e principalmente variabilidade, com a variação do tamanho amostral $n$, ainda assim o processo de estimação de $\omega$ via *optim* do R parece instável e consequentemente afetando as estimativas dos parâmetros $\sigma_1^2$ e $\sigma_2^2$ (veja Figura 8). Além disso, verificou-se que tais resultados apresentavam característica de subestimação.

Figura 8 – Boxplots das estimativas de $\rho_1$, $\rho_2$, $\omega$, $\mu_1$, $\mu_2$, $\lambda_1$, $\lambda_2$, $\sigma_1^2$ e $\sigma_2^2$ (linha vermelha indica o valor real dos parâmetros), do modelo FM-SNGH univariado, quando $\eta = 1$.






Fonte: O autor (2020).

### 4.1.2 **Desempenho das estimativas de máxima verossimilhança no contexto multivariado**

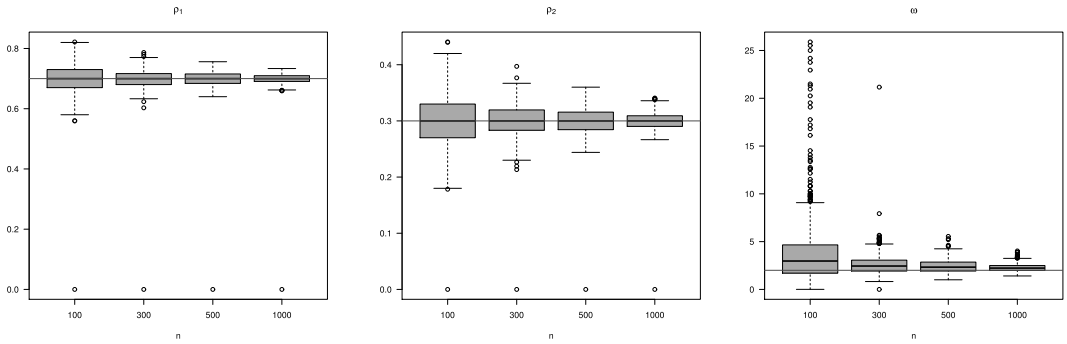
Assim como foi realizado na seção anterior, porém no contexto univariado, nesta seção usa-se da mesma ideia para analisar o desempenho das estimativas de máxima verossimilhança dos parâmetros no modelo FM-SNGH multivariado, com dois grupos de mistura, em ambos cenários estabelecidos na Tabela 1. Embora o modelo tenha produzido estimativas satisfatórias, em termos de viés e variabilidade, sob os diferentes valores de $\eta$ assumidos, os resultados para os elementos do primeiro grupo de mistura ficaram mais aproximados de seus valores reais que no segundo grupo, como podem ser vistos nos cenários descritos a seguir. Além disso, nas Tabelas 13 à 20 (consulte o Apêndice C) encontram-se as estatísticas de resumo das estimativas dos parâmetros do modelo FM-SNGH de acordo com os cenários considerados.

#### 4.1.2.1 *Cenário 1*

Agora com as simulações finalizadas e os resultados gerados para o contexto multivariado, em que também utiliza-se $\eta = {}^{-1}/_{2}$ (fixo), pôde-se notar, de maneira geral, que enquanto aumentava-se o tamanho amostral, as estimativas dos parâmetros de interesse se aproximavam dos seus respectivos valores verdadeiros (veja Tabela 1) com redução da variabilidade. As Figuras 9 a 11, mostram os boxplots das estimativas dos parâmetros $\rho$, $\omega$, $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ e $\boldsymbol{\Sigma}$.

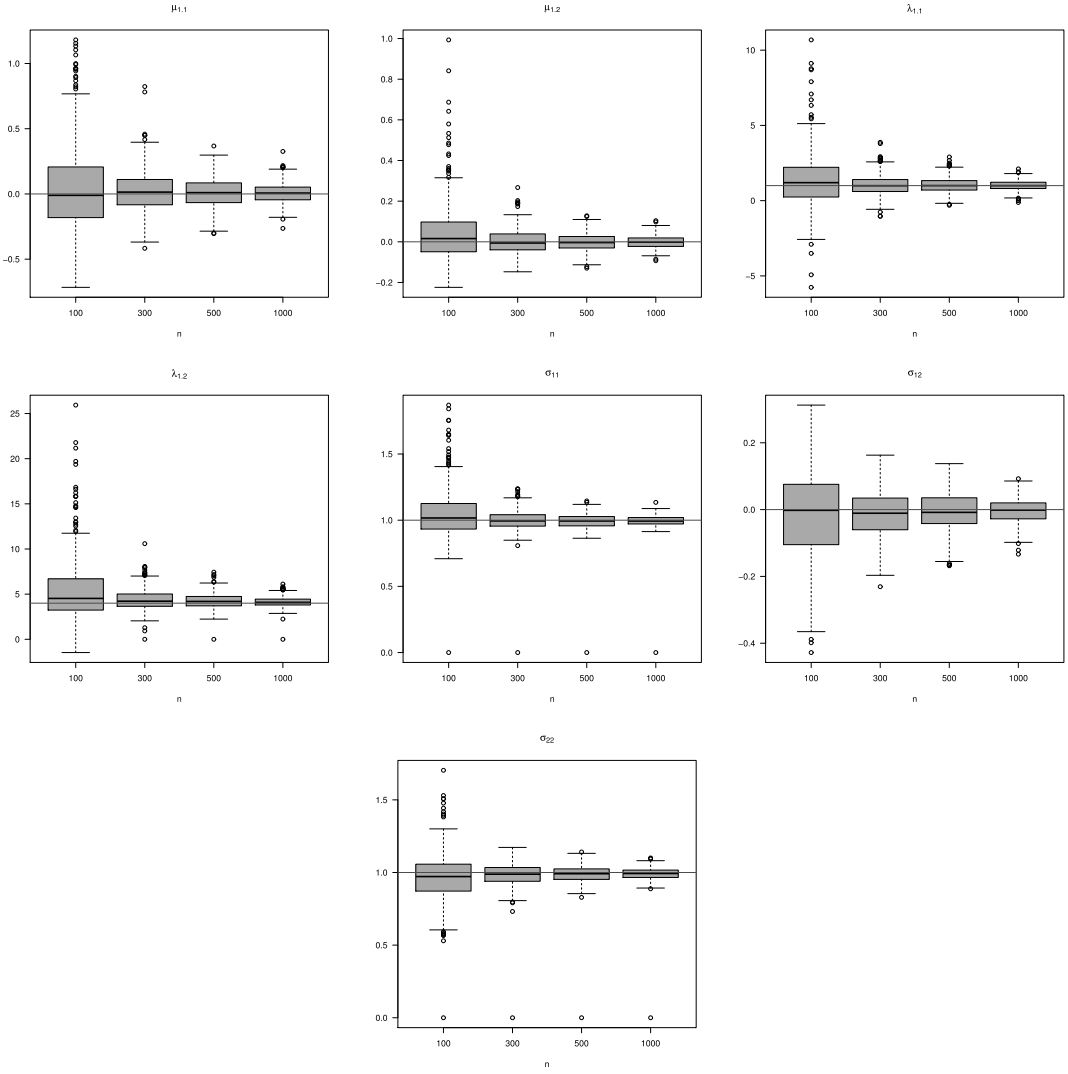
Conforme mostra a Figura 10, o processo de estimação dos parâmetros do modelo proposto foram satisfatórios no primeiro grupo de mistura, porém na Figura 11, observou-se que o desempenho do algoritmo EM na recuperação dos valores verdadeiros dos parâmetros, relativos ao segundo grupo, é menos satisfatório, principalmente em relação ao parâmetro de covariância ($\boldsymbol{\Sigma}$), em particular para o elemento $\boldsymbol{\Sigma}_2$. Além disso, notou-se que tais resultados apresentaram uma característica de subestimação.

Figura 9 – Boxplots das estimativas de $\rho_1$, $\rho_2$ e $\omega$ (linha vermelha indica o valor real dos parâmetros), do modelo FM-SNGH multivariado, quando $\eta = {}^{-1}/_{2}$.



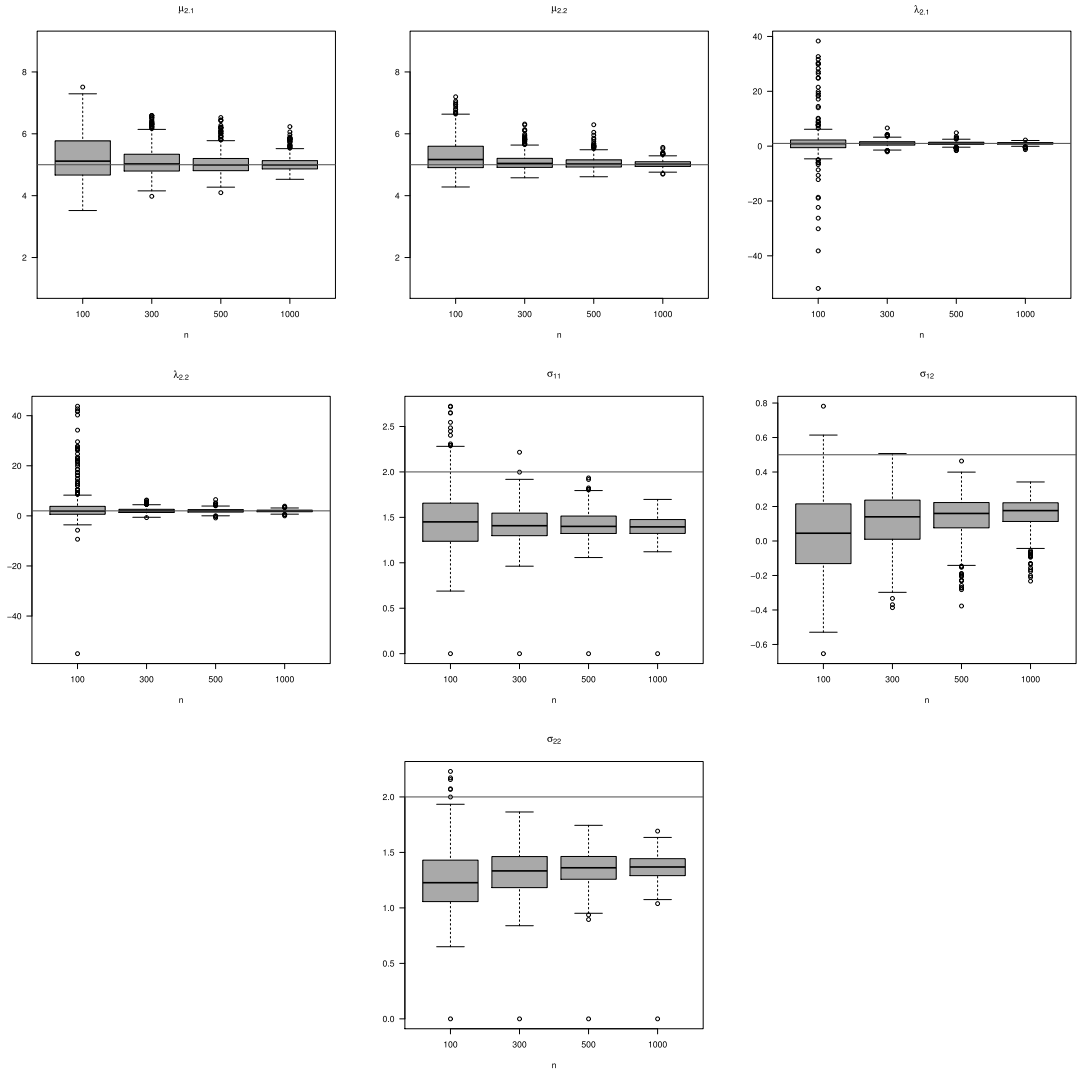

Fonte: O autor (2020).

Figura 10 – Boxplots das estimativas de $\boldsymbol{\mu}_1$, $\boldsymbol{\lambda}_1$ e $\boldsymbol{\Sigma}_1$ (linha vermelha indica o valor real dos parâmetros), do modelo FM-SNGH multivariado, quando $\eta = {}^{-1}/_{2}$.




Fonte: O autor (2020).

Figura 11 – Boxplots das estimativas de $\boldsymbol{\mu}_2$, $\boldsymbol{\lambda}_2$ e $\boldsymbol{\Sigma}_2$ (linha vermelha indica o valor real dos parâmetros), do modelo FM-SNGH multivariado, quando $\eta = {}^{-1}/_{2}$.

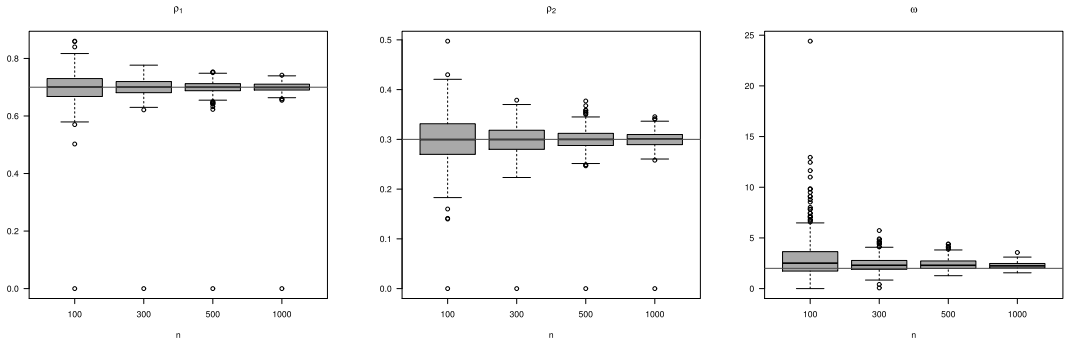



Fonte: O autor (2020).

#### 4.1.2.2 *Cenário 2*

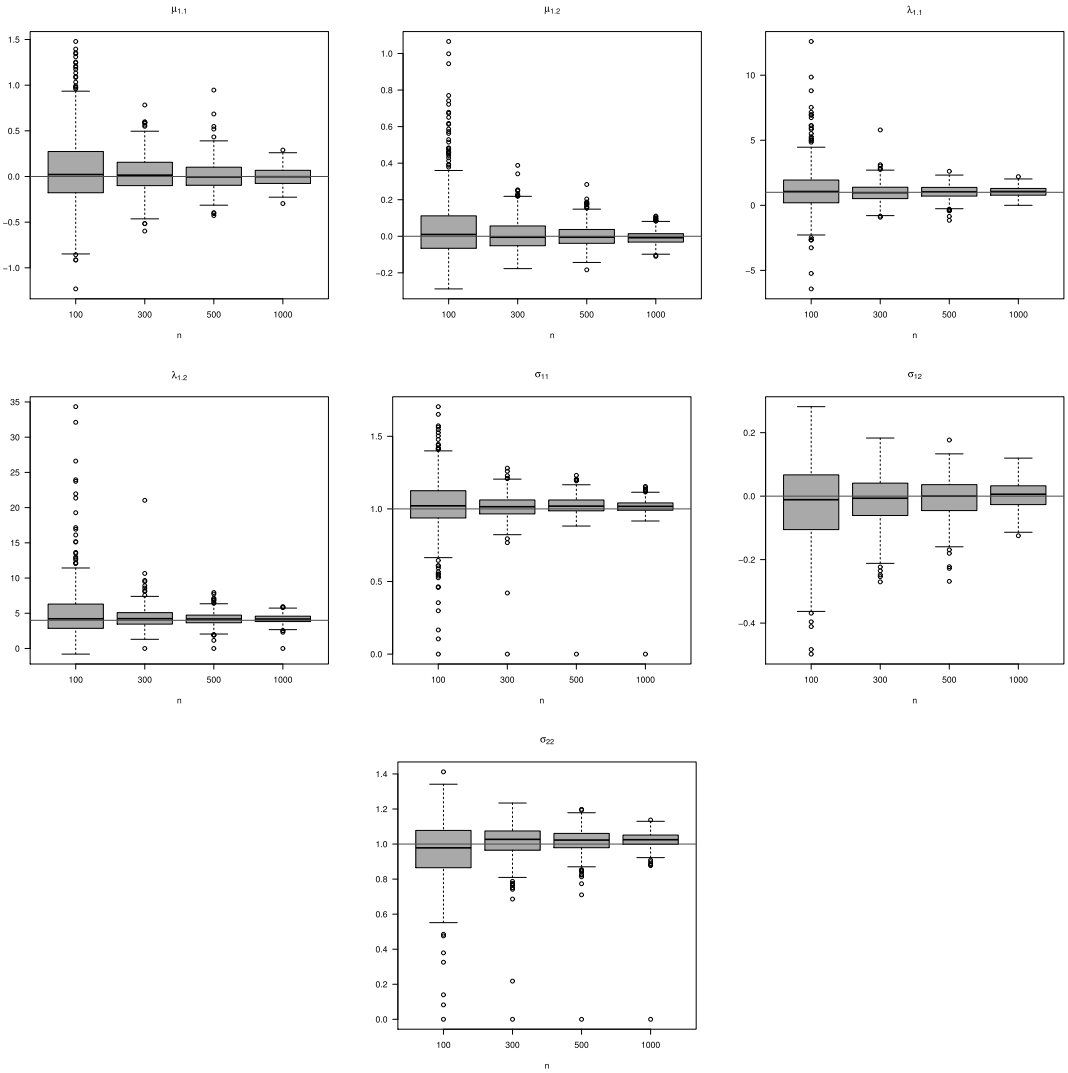
Já para o Cenário 2, após concluir as simulações com $\eta = {}^{1}/_{2}$ (fixo), as Figuras 12 à 14 mostram os boxplots das estimativas dos parâmetros do modelo FM-SNGH, com $G = 2$. Assim como ocorreu no Cenário 1, o desempenho do algoritmo EM na recuperação dos valores verdadeiros dos parâmetros, relativos ao segundo grupo, é menos satisfatório, com destaque para o parâmetro $\boldsymbol{\Sigma}$ (veja Figura 14).

Figura 12 – Boxplots das estimativas de $\rho_1$, $\rho_2$ e $\omega$ (linha vermelha indica o valor real dos parâmetros), do modelo FM-SNGH multivariado, quando $\eta = 1/2$.



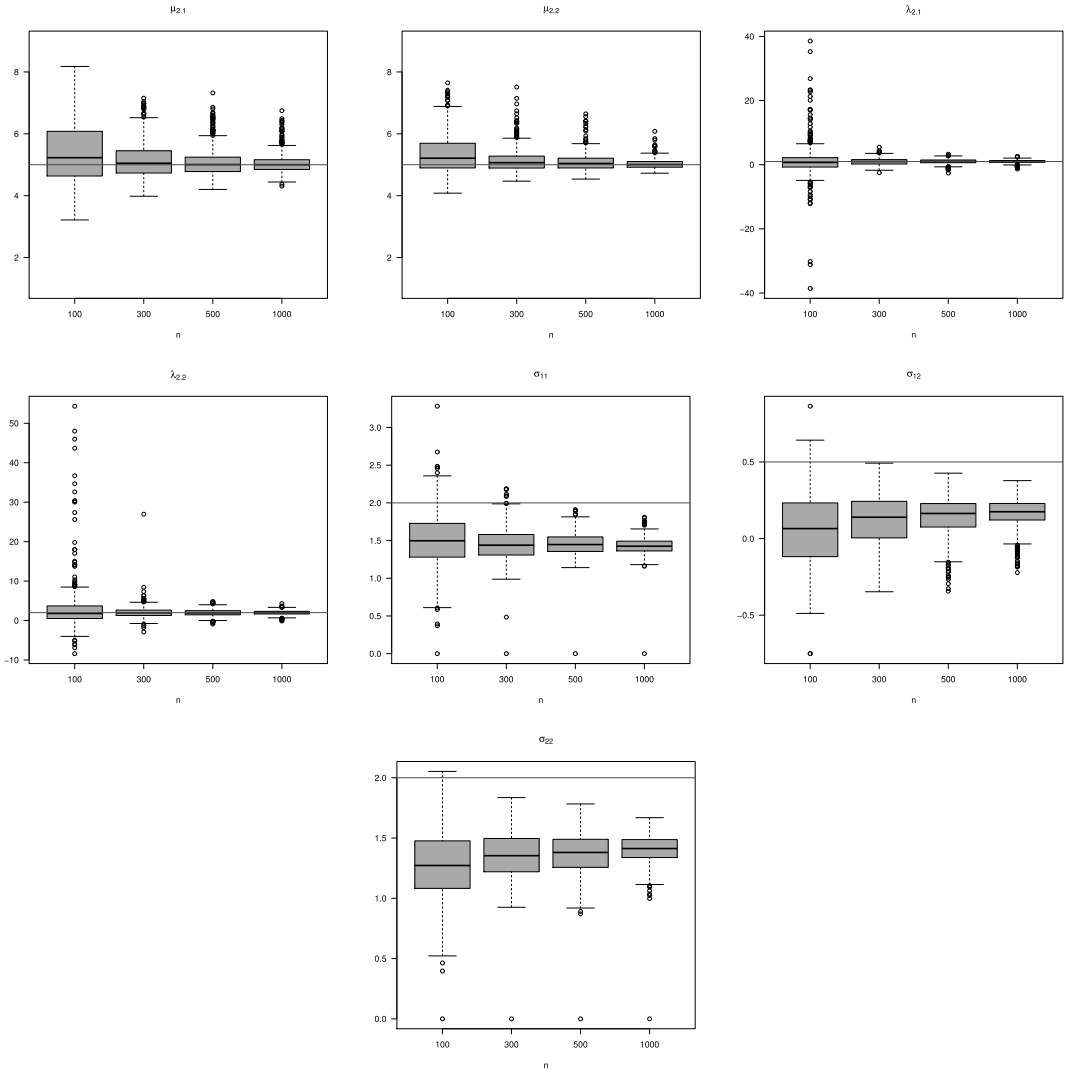

Fonte: O autor (2020).

Figura 13 – Boxplots das estimativas de $\boldsymbol{\mu}_1$, $\boldsymbol{\lambda}_1$ e $\boldsymbol{\Sigma}_1$ (linha vermelha indica o valor real dos parâmetros), do modelo FM-SNGH multivariado, quando $\eta = 1/2$.




Fonte: O autor (2020).

Figura 14 – Boxplots das estimativas de $\boldsymbol{\mu}_2$, $\boldsymbol{\lambda}_2$ e $\boldsymbol{\Sigma}_2$ (linha vermelha indica o valor real dos parâmetros), do modelo FM-SNGH multivariado, quando $\eta = 1/2$.




Fonte: O autor (2020).

Para completar o estudo de simulação projetado, na seção a seguir, serão apresentados alguns gráficos das medidas BIAS e MSE que irão auxiliar no entendimento do comportamento dos estimadores de máxima verossimilhança para os parâmetros do modelo proposto nesse trabalho.

### 4.1.3 **Análise dos cenários**

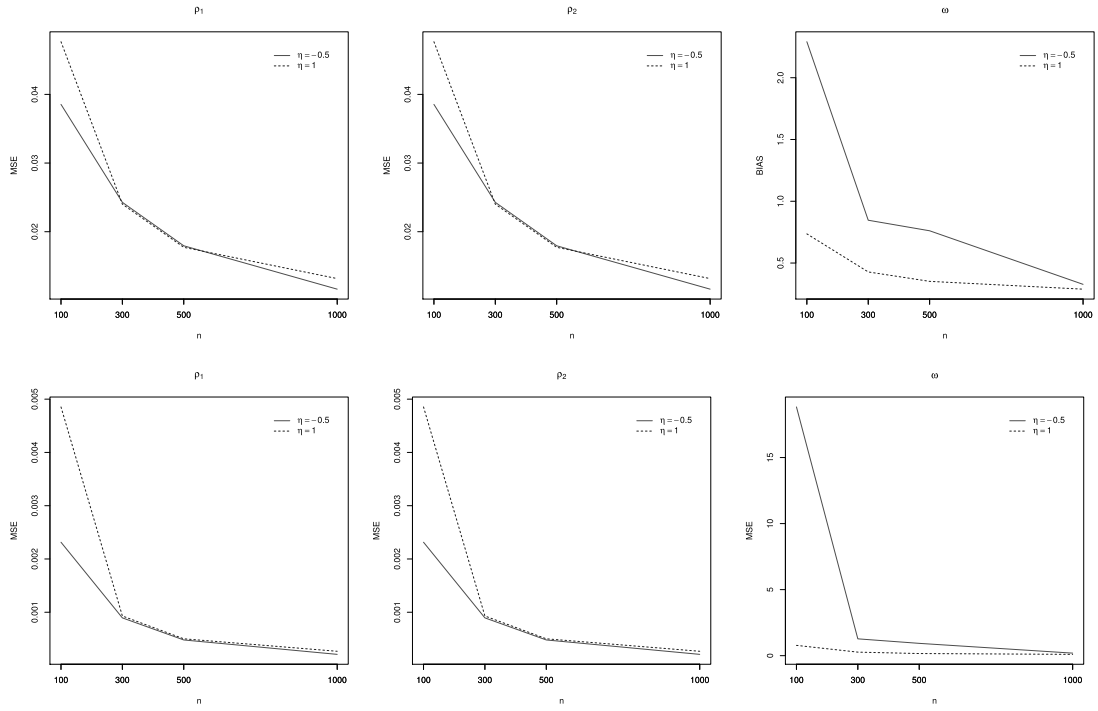
Adicionalmente, nesta seção, com intuito de elucidar e complementar o resultados inferenciais descritos anteriormente, realiza-se a análise dos cenários, definidos na Tabela 1, em termos do viés médio e do erro quadrático médio.

Desta maneira, pode-se observar que para cada um dos cenários, os valores de BIAS e do MSE das estimativas dos parâmetros diminuem, quando o tamanho da amostra aumenta.

Isso concorda essencialmente com as propriedades assintóticas dos estimadores de máxima verossimilhança. Portanto, conclui-se que os estimadores de máxima verossimilhança propostos, neste trabalho, são consistentes.

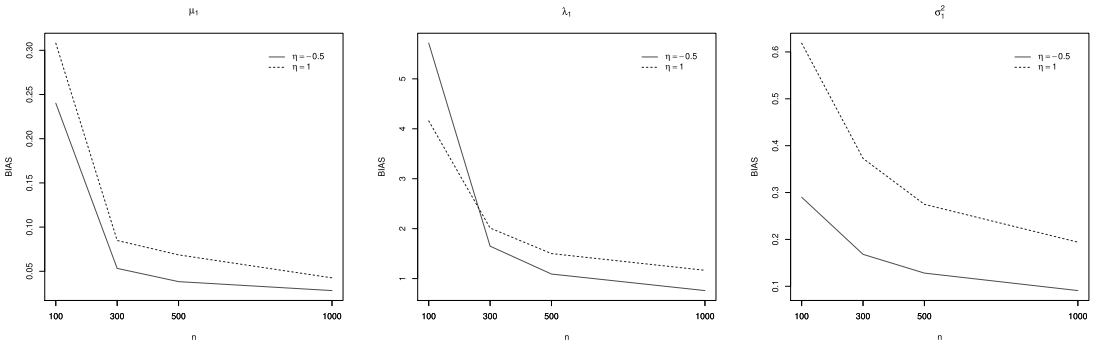
As Figuras 15 à 20, apresentadas a seguir, ilustram o viés absoluto (BIAS) e erro quadrático médio (MSE) baseados nas estimativas dos parâmetros no modelo FM-SNGH, nos contextos univariado e multivariado. Além disso, nas Figuras 19 e 20 são mostradas um caso especial das medidas BIAS e MSE, que descrevem as médias do viés absoluto (MBIAS) e do erro quadrático médio (MMSE) das estimativas dos parâmetros $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ e $\boldsymbol{\Sigma}$ no modelo FM-SNGH multivariado.

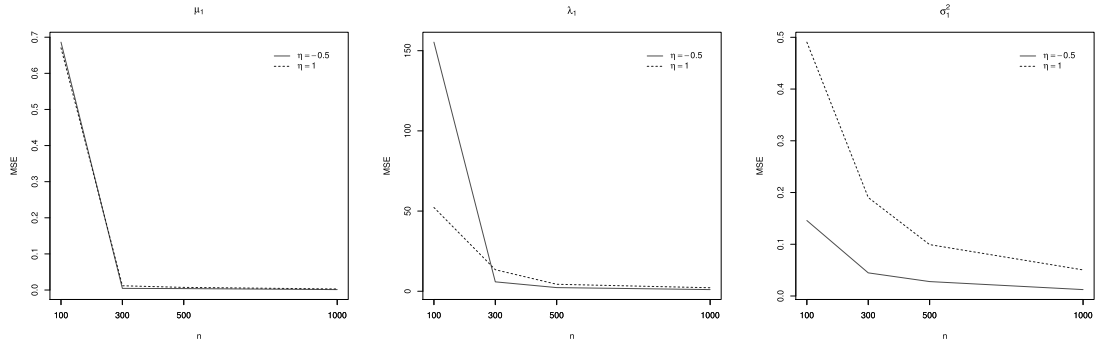
Figura 15 – BIAS e MSE dos elementos $\rho_1$, $\rho_2$ e $\omega$, nos cenários univariado.




Fonte: O autor (2020).

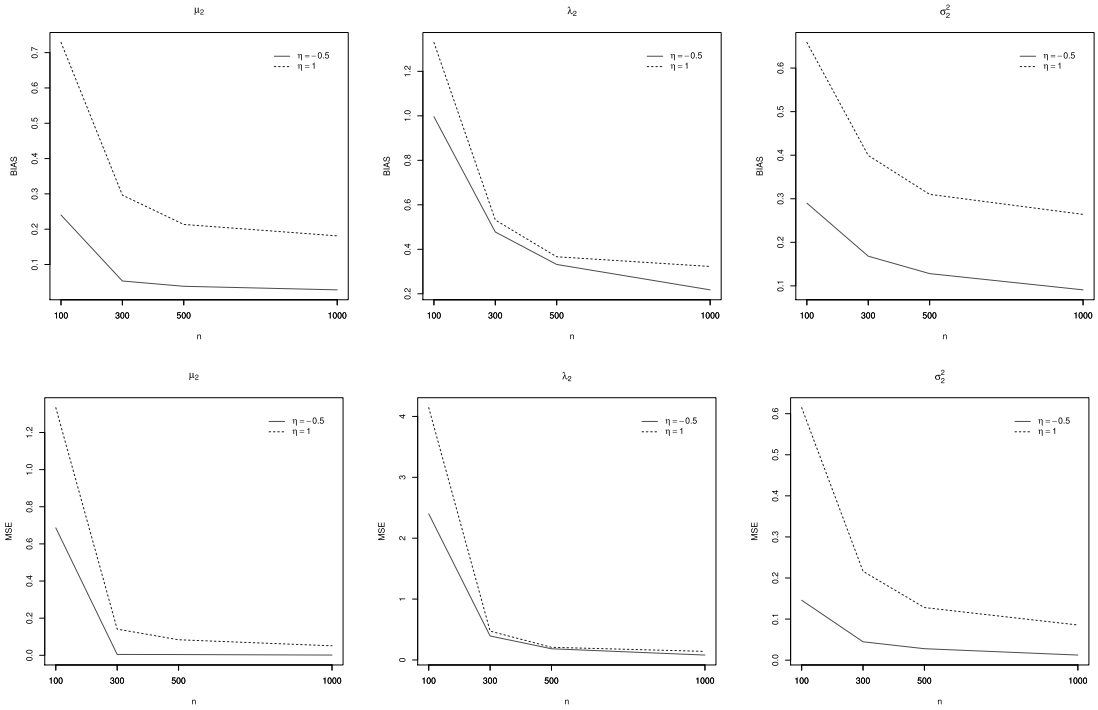
Figura 16 – BIAS e MSE dos elementos $\mu_1$, $\lambda_1$ e $\sigma^2{}_1$, nos cenários univariado.





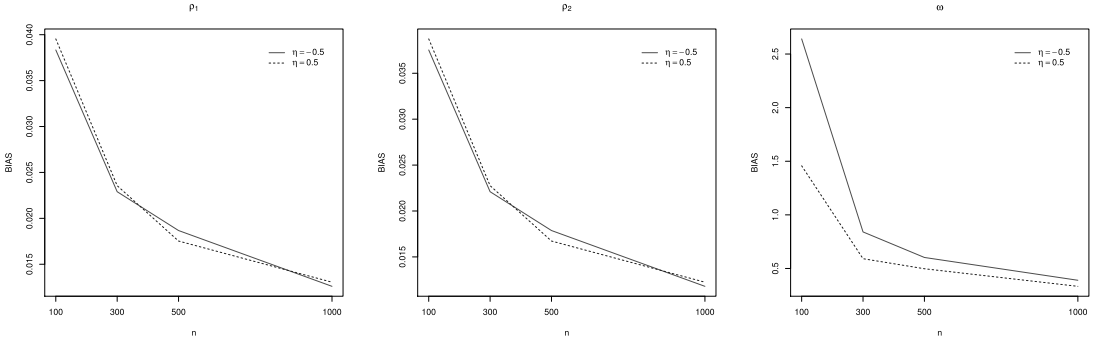

Fonte: O autor (2020).

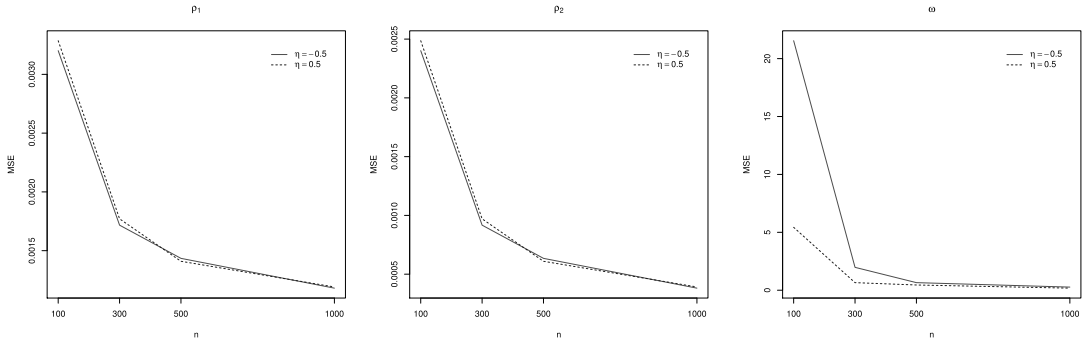
Figura 17 – BIAS e MSE dos elementos $\mu_2$, $\lambda_2$ e $\sigma^2{}_2$, nos cenários univariado.




Fonte: O autor (2020).

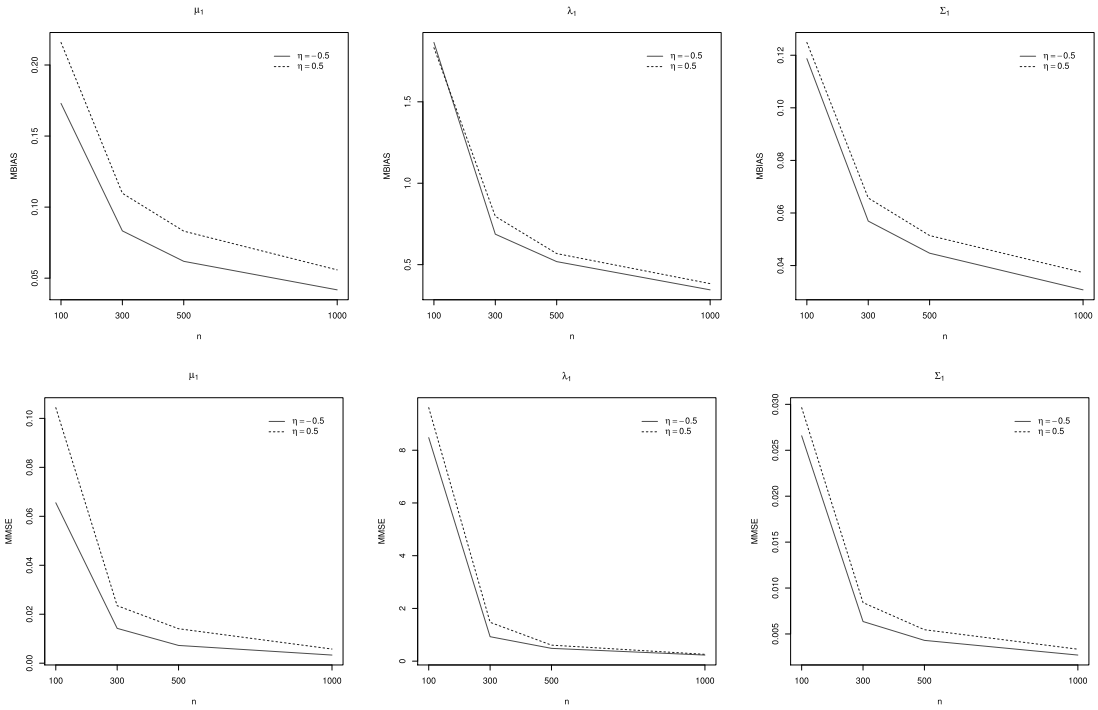
Figura 18 – BIAS e MSE dos elementos $\rho_1$, $\rho_2$ e $\omega$, nos cenários multivariado.





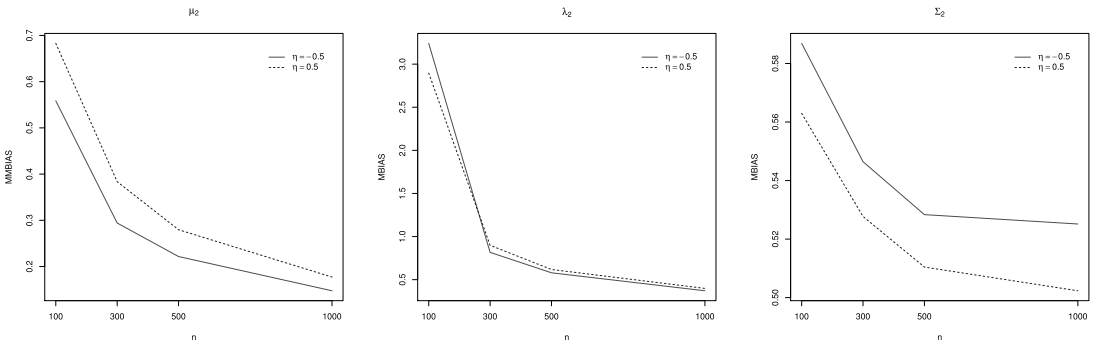

Fonte: O autor (2020).

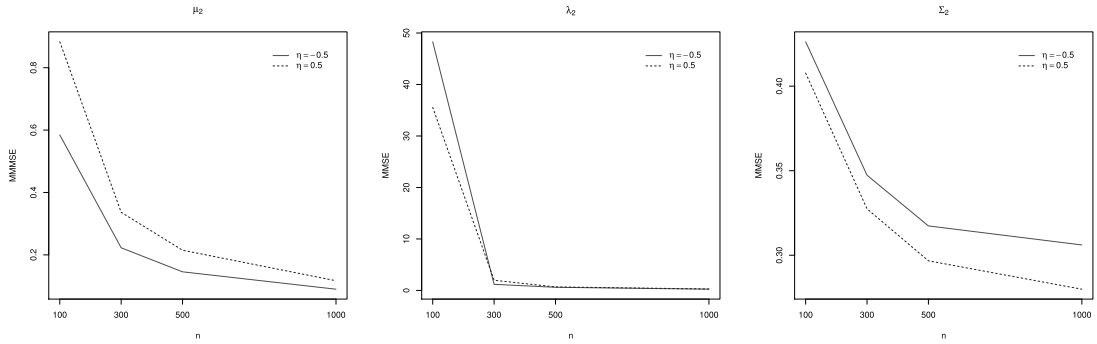
Figura 19 – MBIAS e MMSE dos elementos $\boldsymbol{\mu}_1$, $\boldsymbol{\lambda}_1$ e $\boldsymbol{\Sigma}_1$, nos cenários multivariado.




Fonte: O autor (2020).

Figura 20 – MBIAS e MMSE dos elementos $\boldsymbol{\mu}_2$, $\boldsymbol{\lambda}_2$ e $\boldsymbol{\Sigma}_2$, nos cenários multivariado.






Fonte: O autor (2020).

### 4.1.4 **Estudos de seleção dos critérios**

Nesta seção, compara-se a capacidade de alguns critérios de informação clássicos em selecionar o modelo apropriado aos dados, desde a comparação entre as distribuições assumidas até a determinação do número de componentes $G$ no contexto de misturas finitas. Como não existe um critério universal para esse tipo de seleção, foram escolhidos os critérios mais populares, o AIC e BIC, mostrados na Seção 3.1.4.

Posto isto, para realização desse estudo, utiliza-se de 100 amostras geradas, artificialmente, de tamanho $n = 300$ com duas e três componentes. Os valores adotados nos parâmetros do modelo univariado e multivariado são os mesmos considerados no estudo de simulação da Seção 4.1 e podem ser encontrados na Tabela 1. Adicionalmente, para o parâmetro $\eta$, optou-se pelos valores de $\eta =$ -1/2, 1 e 1/2 em ambos os casos do modelo proposto. Além disso, para cada réplica gerada, são ajustados os modelos FM-SNGH, FM-SN e FM-ST para $G = 1, \cdots, 5$ e então registrados valores dos critérios AIC e BIC.

Desta maneira, nas próximas seções serão descritos os resultados obtidos desse estudo de comparação entre os critérios AIC e BIC no contexto do modelo FM-SNGH univariado e multivariado, com $G = 2$ e 3. Também são apresentadas as diferenças entre o número indicado de componentes pelos critérios de seleção e o verdadeiro número de componentes adotado, ao variar o parâmetro $\eta$. Para valores maiores que zero, tem-se sobreajuste, isto é, o número de componentes escolhido é maior que o verdadeiro, e menores que zero, ocorre subajuste, ou seja, o número de componentes selecionado é menor que o real. Quando este valor é igual a zero, representa a seleção correta do número de componentes.

#### 4.1.4.1 *Critério de seleção - Caso univariado*

Com as simulações finalizadas para o contexto univariado do modelo FM-SNGH com duas e três componentes, inicialmente, pôde-se notar que ao comparar tal modelo com os modelos FM-SN e FM-ST com vários números de componentes, os critério de seleção AIC e BIC, favoreciam o modelo proposto neste trabalho. As porcentagens de acordo com

a preferência de cada critério podem ser encontradas na Tabela 2, descritas de acordo com os valores adotados para o parâmetro $\eta$.

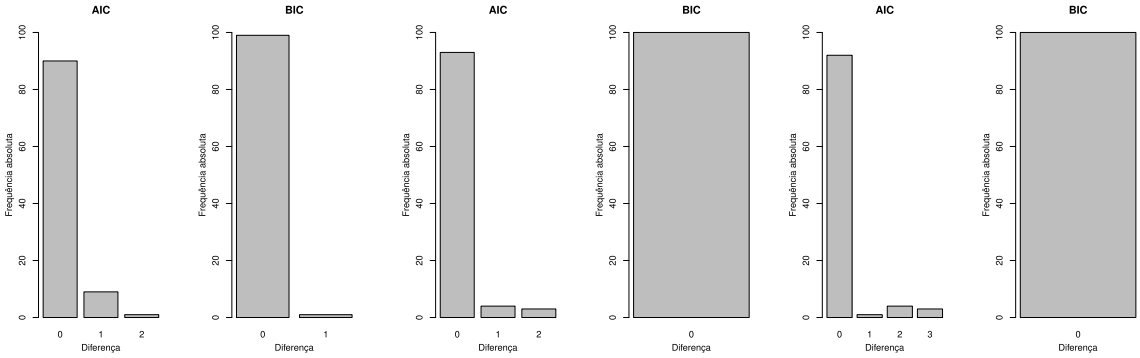
Tabela 2 – Porcentagens que o modelo FM-SNGH univariado com duas e três componentes é escolhido sobre os modelos FM-SN e FM-ST ajustados.

| | **Valor de $\eta$** | **Critérios** | **Número de componentes** | | | | |
|---|---|---|---|---|---|---|---|
| | | | **1** | **2** | **3** | **4** | **5** |
| **Modelo FM-SNGH com duas componentes** | | | | | | | |
| **FM-SN** | $\eta = -1/2$ | AIC | 100 | 97 | 92 | 95 | 99 |
| | | BIC | 100 | 97 | 98 | 100 | 100 |
| | $\eta = 1/2$ | AIC | 100 | 91 | 94 | 100 | 99 |
| | | BIC | 100 | 91 | 99 | 100 | 100 |
| | $\eta = 1$ | AIC | 100 | 92 | 96 | 97 | 100 |
| | | BIC | 100 | 92 | 100 | 100 | 100 |
| **FM-ST** | $\eta = -1/2$ | AIC | 100 | 60 | 92 | 96 | 100 |
| | | BIC | 100 | 60 | 100 | 100 | 100 |
| | $\eta = 1/2$ | AIC | 100 | 60 | 99 | 100 | 100 |
| | | BIC | 100 | 60 | 100 | 100 | 100 |
| | $\eta = 1$ | AIC | 100 | 89 | 100 | 100 | 100 |
| | | BIC | 100 | 89 | 100 | 100 | 100 |
| **Modelo FM-SNGH com três componentes** | | | | | | | |
| **FM-SN** | $\eta = -1/2$ | AIC | 100 | 99 | 97 | 94 | 84 |
| | | BIC | 100 | 99 | 97 | 97 | 97 |
| | $\eta = 1/2$ | AIC | 100 | 100 | 99 | 98 | 88 |
| | | BIC | 100 | 99 | 99 | 98 | 99 |
| | $\eta = 1$ | AIC | 100 | 100 | 100 | 92 | 85 |
| | | BIC | 100 | 100 | 100 | 100 | 100 |
| **FM-ST** | $\eta = -1/2$ | AIC | 100 | 99 | 91 | 51 | 83 |
| | | BIC | 100 | 98 | 91 | 94 | 97 |
| | $\eta = 1/2$ | AIC | 100 | 100 | 97 | 98 | 99 |
| | | BIC | 100 | 99 | 97 | 99 | 99 |
| | $\eta = 1$ | AIC | 100 | 100 | 99 | 100 | 100 |
| | | BIC | 100 | 99 | 100 | 100 | 100 |

Fonte: Elaborada pelo autor (2020).

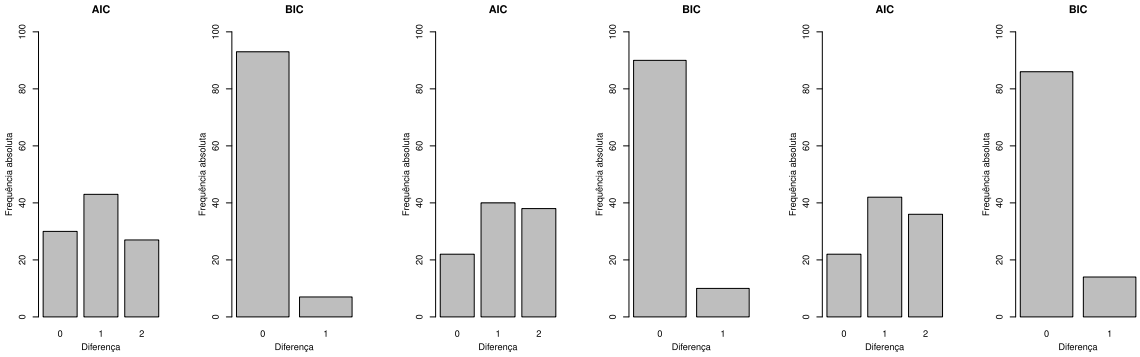
Em seguida, são apresentadas as Figuras 21 e 22, onde descrevem as frequências absolutas de subajustes, ajustes corretos e sobreajustes, em relação ao verdadeiro número de componentes indicados por cada critério, de acordo com cada valor adotado no parâmetro $\eta$. Note que, para selecionar o número de componentes do modelo FM-SNGH, o critério AIC apresentou um baixo desempenho.

Figura 21 – Amostras do modelo FM-SNGH univariado com duas componentes e as frequências absolutas dos critérios AIC e BIC ao considerar o modelo proposto com vários componentes, quando $\eta = -1/2$ (à esquerda), $\eta = 1/2$ (ao centro) e $\eta = 1$ (à direita).




Fonte: O autor (2020).

Figura 22 – Amostras do modelo FM-SNGH univariado com três componentes e as frequências absolutas dos critérios AIC e BIC ao considerar o modelo proposto com vários componentes, quando $\eta = -1/2$ (à esquerda), $\eta = 1/2$ (ao centro) e $\eta = 1$ (à direita).




Fonte: O autor (2020).

Pode-se concluir que o critério BIC apresentou um desempenho superior, comparado ao critério AIC, no decorrer das análises feitas neste estudo de simulação projetado para o modelo FM-SNGH univariado com duas e três componentes. Desta maneira, o critério BIC será utilizado, principalmente, para analisar os ajustes realizados pelo modelo FM-SNGH univariado nas aplicações à dados reais que serão apresentadas na Seção 4.2.

#### 4.1.4.2 *Critério de seleção - Caso multivariado*

Para o contexto multivariado do modelo FM-SNGH, também com duas e três componentes, é realizada a mesma comparação feita no caso anterior. Assim, a partir das porcentagens de preferência dos critérios AIC e BIC, descritas na Tabela 3, de acordo com os valores adotados no parâmetro $\eta$, pôde-se verificar que os critérios favoreceram, mais

uma vez, os modelos propostos com duas e três componentes (modelos verdadeiros).

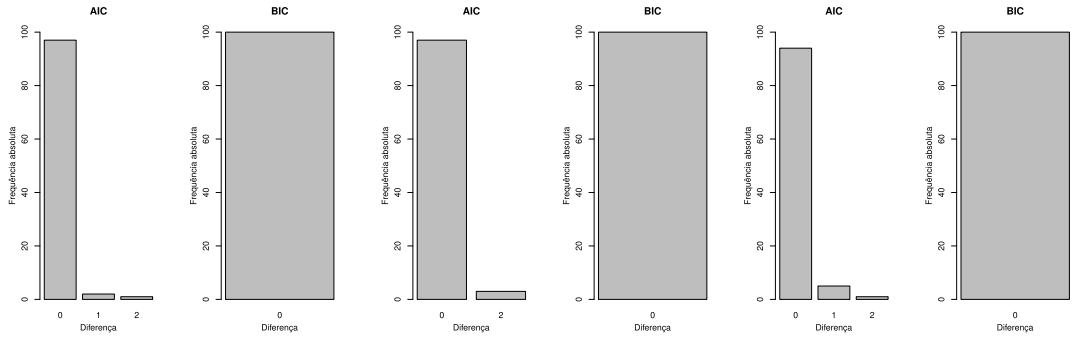
Tabela 3 – Porcentagens que o modelo FM-SNGH multivariado com duas e três componentes é escolhido sobre os modelos FM-SN e FM-ST ajustados.

| | **Valor de $\eta$** | **Critérios** | **Número de componentes** | | | | |
|---|---|---|---|---|---|---|---|
| | | | **1** | **2** | **3** | **4** | **5** |
| **Modelo FM-SNGH com duas componentes** | | | | | | | |
| **FM-SN** | $\eta = -1/2$ | AIC | 100 | 99 | 99 | 100 | 100 |
| | | BIC | 100 | 99 | 100 | 100 | 100 |
| | $\eta = 1/2$ | AIC | 100 | 99 | 100 | 100 | 100 |
| | | BIC | 100 | 99 | 100 | 100 | 100 |
| | $\eta = 1$ | AIC | 100 | 99 | 100 | 100 | 99 |
| | | BIC | 100 | 99 | 100 | 100 | 100 |
| **FM-ST** | $\eta = -1/2$ | AIC | 100 | 80 | 97 | 100 | 100 |
| | | BIC | 100 | 80 | 100 | 100 | 100 |
| | $\eta = 1/2$ | AIC | 100 | 94 | 98 | 100 | 100 |
| | | BIC | 100 | 94 | 100 | 100 | 100 |
| | $\eta = 1$ | AIC | 100 | 81 | 96 | 100 | 100 |
| | | BIC | 100 | 81 | 100 | 100 | 100 |
| **Modelo FM-SNGH com três componentes** | | | | | | | |
| **FM-SN** | $\eta = -1/2$ | AIC | 100 | 70 | 70 | 62 | 63 |
| | | BIC | 100 | 62 | 70 | 66 | 63 |
| | $\eta = 1/2$ | AIC | 100 | 72 | 69 | 70 | 67 |
| | | BIC | 100 | 66 | 69 | 73 | 67 |
| | $\eta = 1$ | AIC | 100 | 74 | 67 | 65 | 68 |
| | | BIC | 99 | 68 | 67 | 67 | 68 |
| **FM-ST** | $\eta = -1/2$ | AIC | 100 | 62 | 60 | 63 | 56 |
| | | BIC | 100 | 61 | 60 | 64 | 58 |
| | $\eta = 1/2$ | AIC | 100 | 66 | 67 | 70 | 67 |
| | | BIC | 100 | 66 | 67 | 74 | 67 |
| | $\eta = 1$ | AIC | 100 | 68 | 66 | 76 | 69 |
| | | BIC | 100 | 67 | 66 | 76 | 69 |

Fonte: Elaborada pelo autor (2020).

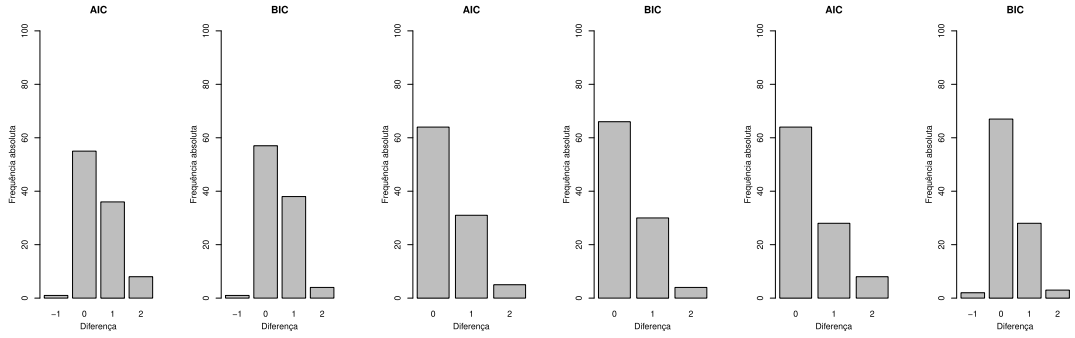
A seguir são apresentadas as Figuras 23 e 24, onde estão descritas as frequências absolutas de subajustes, ajustes corretos e sobreajustes, em relação ao verdadeiro número de componentes selecionado pelos critérios, conforme varia os valores adotados para o parâmetro $\eta$. Observe que, para escolher o número de componentes do modelo FM-SNGH multivariado, novamente o critério AIC apresentou um baixo desempenho.

Figura 23 – Amostras do modelo FM-SNGH multivariado com duas componentes: Frequências absolutas de subajuste, ajuste correto e sobreajuste dos critérios AIC e BIC, considerando o modelo proposto com vários componentes, quando $\eta = -1/2$ (à esquerda), $\eta = 1/2$ (ao centro) e $\eta = 1$ (à direita).




Fonte: O autor (2020).

Figura 24 – Amostras do modelo FM-SNGH multivariado com três componentes: Frequências absolutas de subajuste, ajuste correto e sobreajuste dos critérios AIC e BIC, considerando o modelo proposto com vários componentes, quando $\eta = -1/2$ (à esquerda), $\eta = 1/2$ (ao centro) e $\eta = 1$ (à direita).




Fonte: O autor (2020).

Assim como aconteceu no caso univariado, o critério de seleção BIC apresentou o melhor desempenho, evidenciando a escolha de utilizar o este critério nas aplicações à dados reais apresentadas a seguir, no contexto multivariado também.

## 4.2 APLICAÇÕES EM DADOS REAIS

Nesta seção, com intenção de exemplificar a aplicabilidade do modelo FM-SNGH, considerou-se os conjuntos de dados reais *Body Mass Index* [63], *Swiss Bank* [37] e *Old Faithful* [71] que se destacam por terem sido utilizados em ilustrações numéricas por um grande número de programas estatísticos e autores no contexto de misturas finitas. Consequentemente, tais conjuntos de dados foram revisitados com o intuito de aplicar a

metodologia desenvolvida, neste trabalho, no contexto de misturas finitas de SNGH.

Portanto, diferentes modelos, tais como, FM-SNGH, FM-SN e FM-ST, são ajustados a esses dados para $G = 1, \cdots, 5$. Verifica-se adequação dos modelos aos dados inspecionando os critérios de informação AIC e BIC. Sendo assim, constata-se que, para todos os dados reais aqui analisados, o modelo de misturas finitas com $G = 2$ foi o que se ajustou melhor.

As Estimativas de Máxima Verossimilhança (MLE) dos parâmetros do modelo FM-SNGH são obtidas via algoritmo EM, proposto na Seção 3.2.4, e seus erros padrão correspondentes (SD) calculados por meio do procedimento paramétrico de *bootstrap*. Na versão paramétrica do método *bootstrap*, considera-se $\widehat{\boldsymbol{\Theta}}$ como o valor verdadeiro do parâmetro a fim de gerar 500 amostras do modelo FM-SNGH. Esta estimativa também foi usada como chute inicial do algoritmo EM para obter cada amostra de *bootstrap*.

Sendo assim, calcula-se o desvio padrão amostral dessas replicações. Uma questão importante é se o problema da troca de rótulos (*labels*) ocorre na geração das amostras de *bootstrap*. Todavia, conforme observado por McLachlan & Peel (2004) [62] a escolha de $\widehat{\boldsymbol{\Theta}}$ como chute inicial do algoritmo EM para cada amostra de *bootstrap* evita, na prática, outras ocorrências de troca de rótulo, uma vez que foi considerado o valor verdadeiro do parâmetro do modelo na simulação das amostras de *bootstrap*. Além disso, esta estratégia foi combinada com a escolha dos rótulos minimizando a distância para os verdadeiros valores dos parâmetros, como foi feito na Seção 4.1.

Uma informação importante sobre o modelo FM-SNGH é sobre como obter o valor do parâmetro $\eta$ em uma aplicação à dados reais. Considera-se, um procedimento inspirado pelo trabalho de Snoussi & Idier (2006) [75], onde escolhe-se o valor do parâmetro $\eta$ de acordo com o desempenho da log-verossimilhança dos dados observados dentro do intervalo fechado de -5 à 5, com incremento de 0,5. Em outras palavras, é escolhido o valor de $\eta$ que maximiza a log-verossimilhança dos dados observados, dado que os demais parâmetros estão avaliados nas suas respectivas estimativas de máxima verossimilhança, obtidas via algoritmo EM. Cabe ressaltar que após a escolha do valor de tal parâmetro, ele permanece fixo durante todo processo de estimação dos parâmetros do modelo de interesse.

### 4.2.1 *Body Mass Index*

Para a primeira aplicação, no contexto univariado, considera-se o índice de massa corporal de homens com idades entre 18 e 80 anos. Esses dados foram obtidos no exame nacional de saúde e nutrição, realizado pelo *National Center for Health Statistics* (NCHS), do *Center for Disease Control* (CDC) nos EUA. Esse estudo foi estimulado pelo problema de obesidade que tem chamado bastante atenção nos últimos anos devido sua forte relação com muitas doenças crônicas recorrentes. O índice de massa corporal, mais conhecido por IMC, tornou-se uma média padrão quando o assunto é sobrepeso e obesidade. O índice é dado pela razão do peso corporal em quilos e o quadrado da altura em metros [17].

Na literatura, Lin et al. (2007) [55] utilizou o conjunto de dados *Body Mass Index*, no contexto dos modelos de misturas finitas normais (FM-NOR), *t-Student* (FM-T), normal assimétrica (FM-SN) e $t$ assimétrica (FM-ST), com $G = 2$. Em seu trabalho, ele considerou os relatórios feitos nos anos 1999-2000 e 2001-2002, de modo que para explorar o comportamento de misturas, somente os dados dos participantes que tinham seus pesos entre 39,50 kg à 70,00 kg e 95,01 kg à 196,80 kg foram usados. O conjunto de dados original possui 4579 registros de IMC, após essa seleção dos dados, o conjunto resultante consistiu um total de 2107, onde 1061 participantes pertenciam ao primeiro grupo e 1046 no segundo [66]. Estes dados também foram analisados por Basso et al. (2010) [18] sob os modelos de misturas finitas das distribuições misturas de escala normal assimétrica e notaram que as distribuições SMSN com caudas mais pesadas são mais adequadas aos dados do que a distribuição normal assimétrica. Sendo assim, este conjunto de dados foi revisitado agora no contexto do modelo FM-SNGH.

A Tabela 4 apresenta as estimativas de máxima verossimilhança (MLE) e seus respectivos desvios padrões (SD) sob os modelos ajustados aos dados. Para mais detalhes sobre a estimação dos parâmetros sob os modelos FM-SN e FM-ST, veja [18]. Além disso, a Tabela 4 mostra os valores dos AIC e BIC para os modelos ajustados. Observe que os modelos FM-SNGH e FM-SN não parecem ajustar bem aos dados. Também nota-se que os critérios AIC e BIC favorecem o modelo FM-ST.

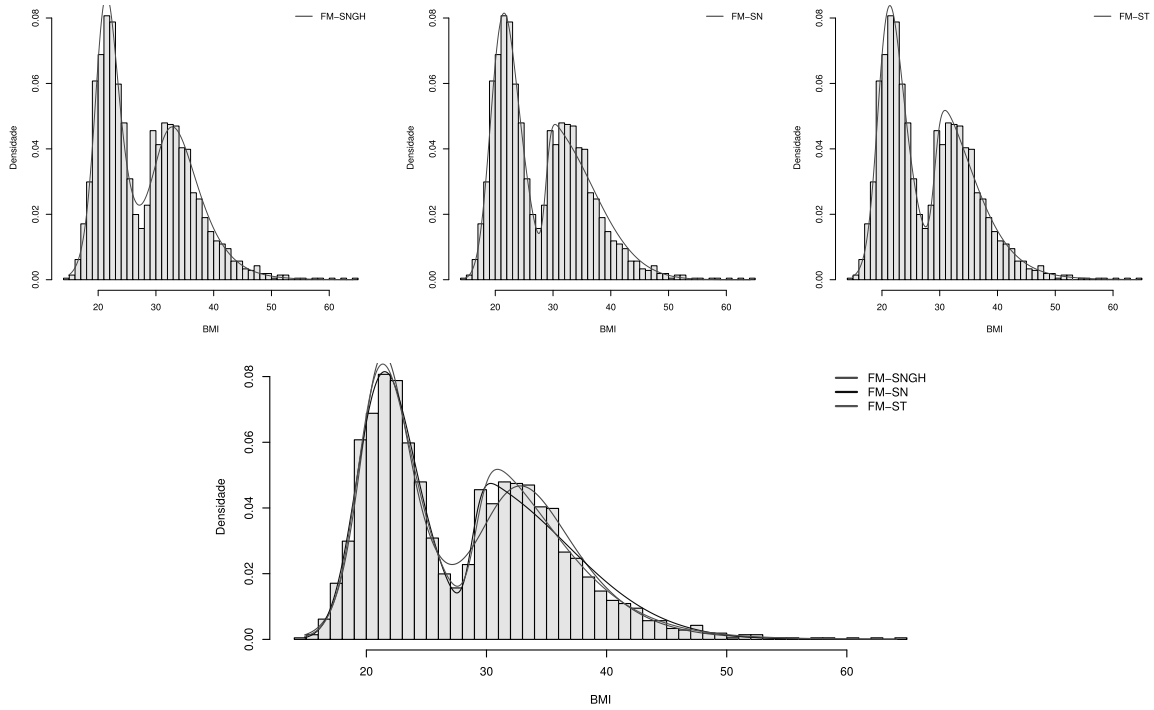
Tabela 4 – *Body Mass Index*: MLE para os parâmetros dos modelos ajustados com os correspondentes desvios padrões. Melhor ajuste indicado por (*1).

| **Parâmetros** | | **FM-SNGH** | | **FM-SN** | | **FM-ST**(*1) | |
|---|---|---|---|---|---|---|---|
| | | **MLE** | **SD** | **MLE** | **SD** | **MLE** | **SD** |
| **Grupo 1** | $\rho$ | 0,497 | 0,0062 | 0,472 | 0,0125 | 0,462 | 0,0142 |
| | $\mu$ | 20,132 | 4,9642 | 19,500 | 0,2429 | 15,572 | 0,2432 |
| | $\sigma^2$ | 15,146 | 0,4929 | 14,365 | 0,2841 | 12,916 | 0,3072 |
| | $\lambda$ | 1,120 | 0,1538 | 1,902 | 0,3446 | 1,900 | 0,3723 |
| **Grupo 2** | $\rho$ | 0,503 | 0,0062 | 0,528 | 0,0125 | 0,538 | 0,0142 |
| | $\mu$ | 30,206 | 4,9700 | 28,760 | 0,1456 | 29,100 | 0,1652 |
| | $\sigma^2$ | 61,172 | 0,4979 | 63,217 | 0,1580 | 45,841 | 0,3100 |
| | $\lambda$ | 1,436 | 0,1573 | 10,588 | 2,7408 | 7,131 | 1,8474 |
| | $\eta$ | -1,500 | - | - | | - | |
| | $\omega$ | 2,279 | 0,0194 | - | | - | |
| | $\nu$ | - | | - | | 8,759 | 2,1238 |
| **AIC** | | 13751,56 | | 13750,89 | | 13726,67 | |
| **BIC** | | 13791,13 | | 13790,46 | | 13771,89 | |

Fonte: Elaborada pelo autor (2020).

Em seguida, é interessante comparar os ajustes das densidades dos modelos FM-ST (o melhor ajustado), FM-SN e FM-SNGH. Assim, na Figura 25, encontra-se o histograma dos dados sobreposto pelas curvas de densidades ajustadas pelos modelos mencionados.

Figura 25 – Histograma dos dados *Body Mass Index* com as curvas ajustadas pelos modelos FM-SNGH, FM-SN e FM-ST.




Fonte: O autor (2020).

Portanto, pode-se concluir que o modelo proposto, neste trabalho, FM-SNGH é bastante flexível no sentido que pode acomodar simultaneamente multimodalidade, assimetria e caudas pesadas dependendo dos valores assumidos por seus parâmetros. Ou seja, o modelo FM-SNGH pode ser um concorrente dos modelos FM-SMSN em termos de ajuste e classificação. Veremos a seguir outras aplicações em dados reais.

### 4.2.2 *Swiss Bank*

Analisado por alguns autores como Flury & Riedwyl (1988) [37], Ma & Genton (2004) [59], Lin (2009) [53] e Basso (2009) [17], o conjunto de dados *Swiss Bank* será utilizado como primeira aplicação no contexto multivariado para ilustrar a flexibilidade do modelo FM-SNGH proposto neste trabalho.

Os dados pertencentes ao conjunto *Swiss Bank* consistem na análise de seis medidas de dimensões adquiridas de 100 notas verdadeiras e 100 notas falsas de mil francos suíços. Seguindo os autores citados acima, para explorar estes dados no contexto de misturas finitas de densidades, considera-se apenas duas das seis medidas tomadas nas notas, de

maneira que $X_1$ representa a largura da borda direita das notas e $X_2$ o comprimento diagonal da imagem central. Assim, pode-se utilizar os modelos de misturas com intuito de classificar as observações em dois grupos, tratando o problema como sendo de classificação supervisionada em grupos de notas verdadeiras e falsas, levando em conta que tal conhecimento é ignorado [17].

A Tabela 5 mostra as estimativas de máxima verossimilhança (MLE) e seus desvios padrões (SD) correspondentes sob os modelos ajustados aos dados. Para mais detalhes sobre a estimação dos parâmetros sob os modelos FM-SN e FM-ST, veja [17]. Adicionalmente, a Tabela 5 apresenta os resultados do ajuste em termos da log-verossimilhança dos dados observados e dos critérios AIC e BIC. Observe que o modelo FM-SN não parece ajustar bem os dados. Também observa-se que os critérios AIC e BIC favorecem o modelo FM-SNGH seguido do modelo FM-ST, ou seja, modelos que têm caudas pesadas e comportamentos assimétricos. Em seguida, compara-se estes modelos em termos de classificação.

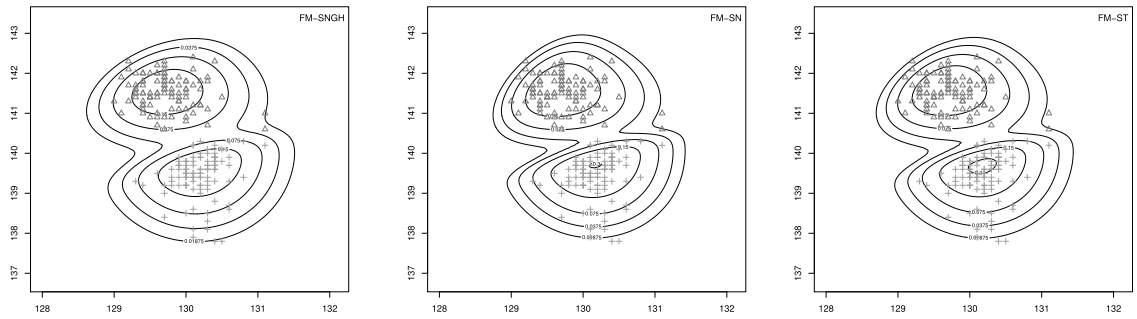
Tabela 5 – *Swiss Bank*: MLE para os parâmetros dos modelos ajustados com os correspondentes desvios padrões. Melhor ajuste indicado por (*1).

| **Parâmetros** | | **FM-SNGH**(*1) | | **FM-SN** | | **FM-ST** | |
|---|---|---|---|---|---|---|---|
| | | **MLE** | **SD** | **MLE** | **SD** | **MLE** | **SD** |
| **Grupo 1** | $\rho$ | 0,498 | 0,0208 | 0,501 | 0,0360 | 0,504 | 0,0365 |
| | $\mu_{11}$ | 130,121 | 5,8123 | 130,114 | 0,0656 | 130,116 | 0,0660 |
| | $\mu_{12}$ | 139,969 | 6,2520 | 140,011 | 0,0780 | 139,986 | 0,7758 |
| | $\sigma_{11}$ | 0,237 | 0,0198 | 0,085 | 0,0256 | 0,070 | 0,0302 |
| | $\sigma_{12}$ | -0,004 | 0,0050 | -0,0152 | 0,0894 | -0,009 | 0,0867 |
| | $\sigma_{22}$ | 0,608 | 0,0311 | 0,624 | 0,0726 | 0,496 | 0,0855 |
| | $\lambda_{11}$ | 1,091 | 0,0490 | 1,4528 | 0,9394 | 1,2679 | 0,7844 |
| | $\lambda_{12}$ | -3,724 | 0,1083 | -5,0930 | 2,0816 | -4,4342 | 1,7948 |
| **Grupo 2** | $\rho$ | 0,502 | 0,0208 | 0,498 | 0,0360 | 0,495 | 0,0365 |
| | $\mu_{11}$ | 129,424 | 5,7852 | 129,331 | 0,0678 | 129,376 | 0,0828 |
| | $\mu_{12}$ | 141,804 | 6,3409 | 141,772 | 0,1211 | 141,791 | 0,1273 |
| | $\sigma_{11}$ | 0,374 | 0,0292 | 0,299 | 0,0607 | 0,225 | 0,0833 |
| | $\sigma_{12}$ | -0,093 | 0,0066 | -0,129 | 0,1067 | -0,108 | 0,0108 |
| | $\sigma_{22}$ | 0,407 | 0,0321 | 0,239 | 0,0749 | 0,225 | 0,0855 |
| | $\lambda_{11}$ | 1,611 | 0,1132 | 2,713 | 1,1659 | 2,125 | 1,0739 |
| | $\lambda_{12}$ | -1,297 | 0,0842 | -1,426 | 1,0470 | -1,390 | 1,0820 |
| | $\eta$ | 2,000 | - | - | | - | |
| | $\omega$ | 5,497 | 0,1721 | - | | - | |
| | $\nu$ | - | | - | | 12,766 | 9,8835 |
| **log-likelihood** | | -306,52 | | -307,94 | | -306,21 | |
| **AIC** | | 643,058 | | 645,887 | | 644,433 | |
| **BIC** | | 692,532 | | 695,362 | | 697,207 | |

Fonte: Elaborada pelo autor (2020).

Assim, o objetivo principal dessa aplicação foi verificar qual dos modelos de misturas ajustados com duas componentes, especialmente o modelo FM-SNGH, produz resultados mais satisfatórios quando ajustados a este conjunto de dados. Na Figura 26, encontram-se os contornos ajustados aos dados pelos modelos FM-SNGH, FM-SN e FM-ST, com os pontos classificados entre notas verdadeiras, em vermelho, e notas falsas, na cor verde, de acordo com cada modelo aqui mencionado. Note que, em termos de classificação os ajustes feitos pelos modelos FM-SNGH, FM-ST e FM-SN tiveram desempenhos satisfatórios.

Figura 26 – Contornos e pontos classificados pelos modelos FM-SNGH, FM-SN e FM-ST para os dados *Swiss Bank*.




Fonte: O autor (2020).

Dessa forma, como pode ser visto nos resultados obtidos pelos critérios de seleção AIC e BIC, principalmente pelo critério BIC, o modelo FM-SNGH se ajustou melhor aos dados, evidenciando sua flexibilidade e utilidade em aplicações com dados reais. Além disso, pode-se notar que o modelo também conseguiu classificar de maneira satisfatória os dados pertencentes para cada grupo.

### 4.2.3 ***Old Faithful***

A segunda aplicação dos contextos univariado e multivariado, baseia-se no conjunto de dados chamado *Old Faithful*. Este conjunto é composto por 272 medições em minutos referentes aos intervalos de tempo entre as partidas de erupções sucessivas e a duração da erupção subsequente do gêiser conhecido por *The Old Faitful* que está localizado no *Yellowstone National Park*, Wyoming, EUA [71]. Esses dados tem sido analisados por diversos autores como Denby & Pregibon (1987) [34], Silverman (1985) [74], Azzalini & Bowman (1990) [12], Lin et al. (2007) [56] e Basso (2009) [17] que apresentaram diferentes metodologias e perceptivas, sob o ponto de vista frequentista e bayesiano.

Desta maneira, neste trabalho utiliza-se do conjunto de dados *Old Faithful* para ilustrar a aplicabilidade do modelo FM-SNGH, nos contextos univariado e multivariado. Sendo assim, no primeiro caso, considera-se apenas os dados referentes as 272 medições

de tempos das erupções em minutos. Já no segundo, para ilustração dos dados são considerados os 272 pares de medidas relacionados aos tempos das erupções e o tempo de espera registrado entre uma erupção e a próxima, ambos em minutos. Então, torna-se possível explorar o comportamento de misturas finitas nesses dados.

As estimativas de máxima verossimilhança (MLE) dos parâmetros dos modelos ajustados, seus desvios padrões (SD) correspondentes e os critérios AIC e BIC podem ser encontrados na Tabela 6. Para mais detalhes sobre a estimação dos parâmetros sob os modelos FM-SN e FM-ST, veja [17]. Note que os resultados em termos de ajuste, para o conjunto de dados Old Faithful, apontam para o modelo FM-SN.

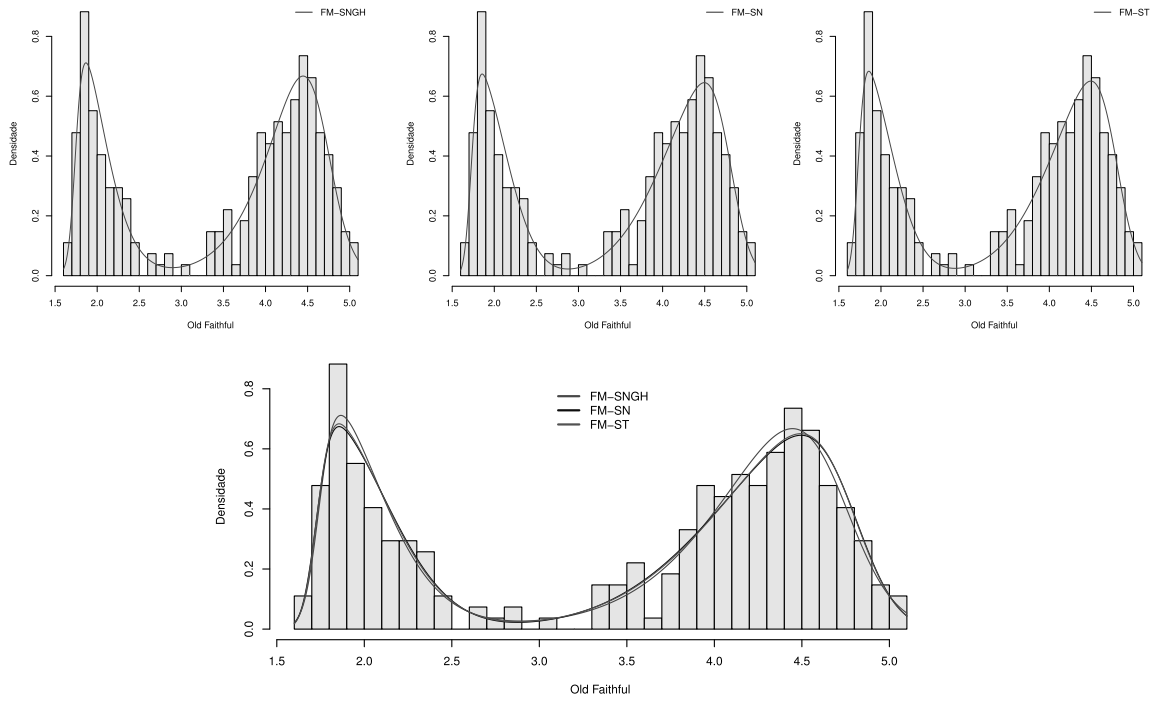
Tabela 6 – *Old Faithful*: MLE para os parâmetros dos modelos ajustados com os correspondentes desvios padrões. Melhor ajuste indicado por (*1).

| **Parâmetros** | | **FM-SNGH** | | **FM-SN**(*1) | | **FM-ST** | |
|---|---|---|---|---|---|---|---|
| | | **MLE** | **SD** | **MLE** | **SD** | **MLE** | **SD** |
| **Grupo 1** | $\rho$ | 0,348 | 0,0299 | 0,348 | 0,0293 | 0,348 | 0,0294 |
| | $\mu$ | 1,740 | 0,0113 | 1,726 | 0,0290 | 1,728 | 0,0287 |
| | $\sigma^2$ | 0,138 | 0,0072 | 0,145 | 0,0415 | 0,137 | 0,0430 |
| | $\lambda$ | 4,853 | 0,3669 | 5,811 | 2,1462 | 5,707 | 2,1269 |
| **Grupo 2** | $\rho$ | 0,652 | 0,0299 | 0,652 | 0,0293 | 0,652 | 0,0294 |
| | $\mu$ | 4,731 | 0,0262 | 4,797 | 0,0514 | 4,793 | 0,0524 |
| | $\sigma^2$ | 0,406 | 0,0313 | 0,465 | 0,0620 | 0,448 | 0,0667 |
| | $\lambda$ | -2,371 | 0,1394 | -3,438 | 1,1329 | -3,366 | 1,1329 |
| | $\eta$ | -0,500 | - | - | | - | |
| | $\omega$ | 5,327 | 0.1593 | - | | - | |
| | $\nu$ | - | | - | | 51,520 | 16,376 |
| **AIC** | | 530,216 | | 529,135 | | 531,067 | |
| **BIC** | | 555,457 | | 554,376 | | 554,913 | |

Fonte: Elaborada pelo autor (2020).

Na Figura 27, são apresentados os comportamentos das curvas de densidades ajustadas por cada um dos modelos aqui considerados sobre o histograma do conjunto de dados *Old Faithful*. Pôde-se observar que o modelo FM-SNGH se ajusta tão bem aos dados como os outros modelos, principalmente o modelo FM-SN que obteve o melhor resultado de critério como foi visto na tabela anterior.

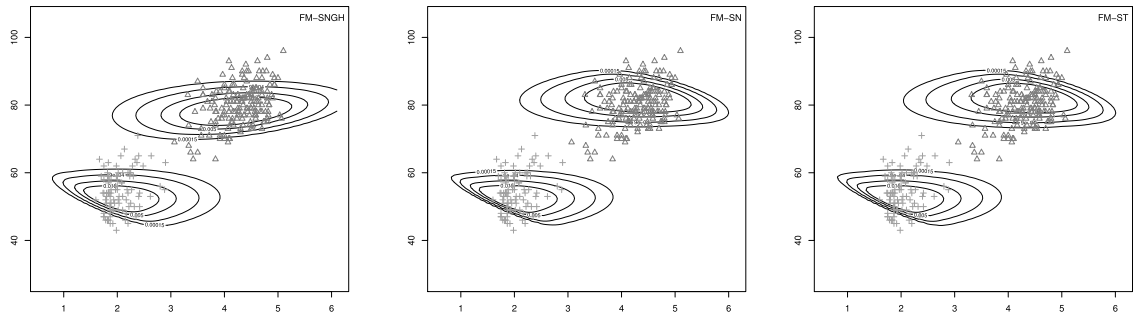
Figura 27 – Histograma dos dados *Old Faithful* com as curvas ajustadas pelos modelos FM-SNGH, FM-SN e FM-ST.




Fonte: O autor (2020).

Por fim, na Figura 28, são mostrados os contornos gerados, juntos dos pontos classificados de acordo com cada modelo. Observe que, a classificação realizada pelo modelo FM-SNGH comparada com as classificações dos demais modelos obtiveram os mesmos resultados, apesar de não haver um conhecimento prévio da realidade dos dados. Já na Tabela 7, encontram-se os resultados obtidos dos modelos FM-SNGH, FM-SN e FM-ST ajustados, para o contexto multivariado dos dados.

Figura 28 – Contornos e pontos classificados pelos modelos FM-SNGH, FM-SN e FM-ST para os dados *Old Faithful*.




Fonte: O autor (2020).

Tabela 7 – *Old Faithful*: MLE para os parâmetros dos modelos ajustados com os correspondentes desvios padrões. Melhor ajuste indicado por (*1).

| **Parâmetros** | | **FM-SNGH** | | **FM-SN**(*1) | | **FM-ST** | |
|---|---|---|---|---|---|---|---|
| | | **MLE** | **SD** | **MLE** | **SD** | **MLE** | **SD** |
| **Grupo 1** | $\rho$ | 0,642 | 0,0277 | 0,642 | 0,0293 | 0,642 | 0,0293 |
| | $\mu_{11}$ | 4,494 | 0,2128 | 4,464 | 0,1410 | 4,499 | 0,1642 |
| | $\mu_{12}$ | 77,142 | 3,4935 | 76,515 | 2,2350 | 77,498 | 2,9356 |
| | $\sigma_{11}$ | 0,451 | 0,0369 | 0,440 | 0,0633 | 0,438 | 0,0777 |
| | $\sigma_{12}$ | 0,041 | 0,0005 | 0,040 | 0,0577 | 0,050 | 0,0713 |
| | $\sigma_{22}$ | 6,650 | 0,1099 | 6,911 | 1,2533 | 6,283 | 1,2360 |
| | $\lambda_{11}$ | -1,057 | 0,0750 | -1,012 | 0,5718 | -1,041 | 0,6718 |
| | $\lambda_{12}$ | 0,970 | 0,0151 | 1,169 | 0,9118 | 0,828 | 1,0329 |
| **Grupo 2** | $\rho$ | 0,358 | 0,0277 | 0,358 | 0,0293 | 0,358 | 0,0293 |
| | $\mu_{11}$ | 1,742 | 0,0766 | 1,737 | 0,0301 | 1,742 | 0,0298 |
| | $\mu_{12}$ | 51,624 | 2,3056 | 51,664 | 1,3304 | 51,703 | 1,3299 |
| | $\sigma_{11}$ | 0,346 | 0,0269 | 0,352 | 0,0356 | 0,334 | 0,0360 |
| | $\sigma_{12}$ | 0,187 | 0,0042 | 0,192 | 0,0619 | 0,176 | 0,0619 |
| | $\sigma_{22}$ | 6,542 | 0,1251 | 6,481 | 0,7301 | 6,271 | 0,7332 |
| | $\lambda_{11}$ | 4,147 | 0,2634 | 4,430 | 1,6241 | 4,145 | 1,5177 |
| | $\lambda_{12}$ | 2,753 | 0,1704 | 2,959 | 1,7680 | 2,669 | 1,6335 |
| | $\eta$ | -0,500 | - | - | | - | |
| | $\omega$ | 11,451 | 0,2622 | - | | - | |
| | $\nu$ | - | | - | | 31,764 | 34,5275 |
| **log-likelihood** | | -1114,5 | | -1110,2 | | -1110,2 | |
| **AIC** | | 2259,04 | | 2250,45 | | 2252,54 | |
| **BIC** | | 2313,13 | | 2304,54 | | 2310,23 | |

Fonte: Elaborada pelo autor (2020).

Para este conjunto de dados, o modelo FM-SN parece se ajustar ligeiramente melhor aos dados. Isso aconteceu em ambos contextos dos dados aqui analisados. Estes resultados estão de acordo com os apresentados em [17]. Contudo, o modelo FM-SNGH produziu ajustes bem próximos ao modelo FM-SN, mostrando ser bastante competitivo tanto no ajuste quanto na classificação quando comparado com os modelos de misturas finitas sob a classe de distribuições SMSN.

# 5 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho teve a finalidade de propor um novo modelo abrangente/flexível baseado em misturas finitas de densidades quando segue uma distribuição SNGH. Esta distribuição pode ser vista como um membro particular da classe de distribuições SMSN, em que utiliza-se como distribuição mistura a distribuição GIG. Devido essas características, o modelo de misturas finitas de SNGH é capaz de acomodar simultaneamente multimodalidade, assimetria e caudas pesadas. Além de admitir uma representação estocástica e hierárquica interessantes, muitas de suas propriedades podem ser facilmente derivadas de resultados já conhecidos das distribuições SMSN. Portanto, a metodologia proposta nessa dissertação de mestrado mostra ter grande aplicabilidade em inúmeras situações encontradas na natureza.

O esforço principal deste trabalho foi dado no processo de obtenção dos estimadores de máxima verossimilhança, através do algoritmo EM, para os parâmetros do modelo de misturas proposto. É importante ressaltar que valores iniciais são necessários para implementação deste algoritmo. Neste trabalho, utiliza-se de uma metodologia similar à adotada por Biernacki (2000) [20] e Browne et al. (2015) [24] que pode ser considerada como uma pré-estimação dos parâmetros, onde são provenientes de um modelo de misturas finitas de normal assimétrica e da maximização da log-verossimilhança perfilada com relação ao parâmetro de forma, em um grid aleatório uniforme. Assim, os resultados aqui obtidos foram aplicados em conjunto de dados reais ou simulados.

Para a ilustração numérica de aplicações à dado reais, considerou-se os conjuntos de dados mencionados na Seção 4.2, na qual foram revisitados com o intuito de aplicar toda a metodologia aqui desenvolvida no contexto (univariado e multivariado) do modelo de misturas finitas de SNGH. Comparando os resultados obtidos, em termos de ajustes e classificação, com outros modelos conhecidos na literatura estatística de misturas finitas, constatou-se que o modelo de misturas finitas de SNGH mostrou-se competitivo em relação aos modelos de misturas finitas normal assimétrica e *skew*-t. Foi verificada a adequação dos modelos aos dados inspecionando alguns critérios de informação.

As aplicações numéricas foram realizadas no programa R [70]. A estimação por máxima verossimilhança dos parâmetros, via algoritmo EM, no modelo de misturas finitas sob a classe de distribuições SMSN foi feita usando o pacote *mixsmsn* do R. Os demais procedimentos deste trabalho foram programados e os códigos estão disponíveis a pedido.

Diante dos resultados obtidos neste trabalho, vários trabalhos futuros podem ainda ser considerados, tais como misturas finitas de SNGH no contexto de modelos de regressão e modelos lineares mistos, por exemplo. Por fim, outra sugestão de pesquisas futuras é utilizar do modelo proposto no desenvolvimento de estudos sob uma perspectiva bayesiana, devido sua representação hierárquica apresentada na Seção 3.2.3.

# REFERÊNCIAS


[1] ABRAMOWITZ, Milton; STEGUN, Irene A. (Ed.). **Handbook of mathematical functions with formulas, graphs, and mathematical tables**. US Government printing office, 1948.

[2] AHMAD, Khalaf E. Identifiability of finite mixtures using a new transform. **Annals of the Institute of Statistical Mathematics**, v. 40, n. 2, p. 261-265, 1988.

[3] AITKIN, Murray; AITKIN, Irit. A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. **Statistics and Computing**, v. 6, n. 2, p. 127-130, 1996.

[4] AKAIKE, Hirotugu. A new look at the statistical model identification. **IEEE transactions on automatic control**, v. 19, n. 6, p. 716-723, 1974.

[5] ANDREWS, David F.; MALLOWS, Colin L. Scale mixtures of normal distributions. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 36, n. 1, p. 99-102, 1974.

[6] ARELLANO-VALLE, Reinaldo B.; BOLFARINE, Heleno; LACHOS, Victor H. Skew-normal linear mixed models. **Journal of data science**, v. 3, n. 4, p. 415-438, 2005.

[7] ARELLANO-VALLE, Reinaldo B.; DEL PINO, Guido E. From Symmetric to Asymmetric Distributions: A Unified Approach. In: **Skew-Elliptical Distributions and Their Applications**. Chapman and Hall/CRC, p. 127-144, 2004.

[8] ARELLANO-VALLE, Reinaldo B.; GENTON, Marc G. On fundamental skew distributions. **Journal of Multivariate Analysis**, v. 96, n. 1, p. 93-116, 2005.

[9] ATIENZA, N.; GARCIA-HERAS, J.; MUNOZ-PICHARDO, J. M. A new condition for identifiability of finite mixture distributions. **Metrika**, v. 63, n. 2, p. 215-221, 2006.

[10] AZZALINI, Adelchi. **A class of distributions which includes the normal ones**. Scandinavian journal of statistics, p. 171-178, 1985.

[11] AZZALINI, Adelchi. The skew-normal distribution and related multivariate families. **Scandinavian Journal of Statistics**, v. 32, n. 2, p. 159-188, 2005.

[12] AZZALINI, Adelchi; BOWMAN, Adrian W. A look at some data on the Old Faithful geyser. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 39, n. 3, p. 357-365, 1990.

[13] AZZALINI, Adelchi; VALLE, A. Dalla. The multivariate skew-normal distribution. **Biometrika**, v. 83, n. 4, p. 715-726, 1996.

[14] BARNDORFF-NIELSEN, Ole Eiler. Hyperbolic distributions and distributions on hyperbolae. **Scandinavian Journal of statistics**, p. 151-157, 1978.

[15] BARNDORFF-NIELSEN, Ole Eiler. Normal inverse Gaussian processes and the modelling of stock returns. 1994.

[16] BARNDORFF-NIELSEN, Ole E. Normal inverse Gaussian distributions and stochastic volatility modelling. **Scandinavian Journal of statistics**, v. 24, n. 1, p. 1-13, 1997.

[17] BASSO, Rodrigo Marreiro. **Misturas finitas de misturas de escala skew-normal**. 2009. Dissertação (Mestrado em Estatística) - Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas, 2009.

[18] BASSO, Rodrigo M. et al. Robust mixture modeling based on scale mixtures of skew-normal distributions. Computational Statistics & Data Analysis, v. 54, n. 12, p. 2926-2941, 2010.

[19] BERKANE, Maia; KANO, Yutaka; BENTLER, Peter M. Pseudo maximum likelihood estimation in elliptical theory: effects of misspecification. **Computational statistics & data analysis**, v. 18, n. 2, p. 255-267, 1994.

[20] BIERNACKI, Christophe; CELEUX, Gilles; GOVAERT, Gérard. Assessing a mixture model for clustering with the integrated completed likelihood. **IEEE transactions on pattern analysis and machine intelligence**, v. 22, n. 7, p. 719-725, 2000.

[21] BÖHNING, Dankmar. Computer Assisted Analysis of Mixtures and Applications: Meta Analysis. **Disease Mapping and Others. Technometrics**, v. 42, p. 442-442, 2000.

[22] BÖHNING, Dankmar et al. Advances in mixture models. **Computational Statistics & Data Analysis**, v. 51, n. 11, p. 5205-5210, 2007.

[23] BRANCO, Márcia D.; DEY, Dipak K. A general class of multivariate skew-elliptical distributions. **Journal of Multivariate Analysis**, v. 79, n. 1, p. 99-113, 2001.

[24] BROWNE, Ryan P.; MCNICHOLAS, Paul D. A mixture of generalized hyperbolic distributions. **Canadian Journal of Statistics**, v. 43, n. 2, p. 176-198, 2015.

[25] CALEGARI, Hugo. **Estimação na distribuição hiperbólica skew-normal: algoritmo EM**. 2020. Dissertação (Mestrado em Estatística) - Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas, 2020.

[26] COELHO, Carina Figueiredo **Misturas finitas de normais assimétricas e de t assimétricas aplicadas em análise discriminante**. 2013. Dissertação (Mestrado em Matemática) - Universidade Federal do Amazonas, Manaus, 2013.

[27] COHEN, A. Clifford. Estimation in mixtures of two normal distributions. **Technometrics**, v. 9, n. 1, p. 15-28, 1967.

[28] CRUVINEL, Evelyn de Castro. **Discriminante para mistura de distribuições GEV**. 2017. Dissertação (Mestrado em Estatística) - Instituto de Ciências Exatas, Universidade de Brasília, Brasília, 2017.

[29] DA SILVA, Nívea Bispo. **Modelagem Bayesiana semi-paramétrica via misturas**. 2017. Tese (Doutorado em Estatística) - Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, 2017.

[30] DÁVILA, Victor Hugo Lachos. **Modelos lineares mistos assimétricos**. 2004. Tese (Doutorado em Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2004.

[31] DÁVILA, Víctor Hugo Lachos; CABRAL, Celso Rômulo Barbosa; ZELLER, Camila Borelli. **Finite Mixture of Skewed Distributions**. Springer, 2018.

[32] DAY, Neil E. Estimating the components of a mixture of normal distributions. **Biometrika**, v. 56, n. 3, p. 463-474, 1969.

[33] DEMPSTER, Arthur P.; LAIRD, Nan M.; RUBIN, Donald B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 39, n. 1, p. 1-22, 1977.

[34] DENBY, Lorraine; PREGIBON, Daryl. An example of the use of graphics in regression. **The American Statistician**, v. 41, n. 1, p. 33-38, 1987.

[35] EBERLEIN, Ernst; PRAUSE, Karsten. The generalized hyperbolic model: financial derivatives and risk measures. In: **Mathematical Finance—Bachelier Congress 2000**. Springer, Berlin, Heidelberg, p. 245-267, 2002.

[36] FENG, Ziding D.; MCCULLOCH, Charles E. Using bootstrap likelihood ratios in finite mixture models. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 3, p. 609-617, 1996.

[37] FLURY, Bernhard; RIEDWYL, Hans. The data. In: **Multivariate Statistics**. Springer Netherlands, p. 1-10, 1988.

[38] GRADSHTEYN, Izrail Solomonovich; RYZHIK, Iosif Moiseevich. **Table of integrals, series, and products**. Academic press, 2014.

[39] HALPHEN, Étienne. Étude statistique des débit du Rhin à Bâle. **Statistical Study of Rhine Flow in Basel**. Annuaire Hydrologique de la France, v. 3, p. 5–23, 1941.

[40] HASSELBLAD, Victor. Estimation of parameters for a mixture of normal distributions. **Technometrics**, v. 8, n. 3, p. 431-444, 1966.

[41] HO, Hsiu J.; LIN, Tsung-I. Robust linear mixed models using the skew t distribution with application to schizophrenia data. **Biometrical Journal**, v. 52, n. 4, p. 449-469, 2010.

[42] HOSMER JR, David W. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. **Biometrics**, p. 761-770, 1973.

[43] JØRGENSEN, Bent. **Statistical properties of the generalized inverse Gaussian distribution. Lecture Notes in Statistics. 9**. New York–Berlin: Springer-Verlag, 1982.

[44] KANFER, Frans; MILLARD, Sollie. A Scale Mixture Approach to t-Distributed Mixture Regression. In: **Computational and Methodological Statistics and Biostatistics**. Springer, Cham, p. 329-347, 2020.

[45] KONLACK SOCGNIA, Virginie; WILCOX, Diane. A Comparison of Generalized Hyperbolic Distribution Models for Equity Returns. **Journal of Applied Mathematics**, v. 2014, 2014.

[46] KOTZ, Samuel; KOZUBOWSKI, Tomasz J.; PODGORSKI, Krzysztof. The Laplace distribution and generalizations: A Revisit with Applications to Communications. **Economics, Engineering, and Finance**, v. 183, 2001.

[47] KOZUBOWSKI, Tomasz J.; PODGORSKI, Krzysztof. Asymmetric Laplace distributions. **Mathematical Scientist**, v. 25, n. 1, p. 37-46, 2000.

[48] LACHOS, Victor Hugo; LABRA, Filidor Vilca. Skew-normal/independent distributions, with applications. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica, 2007.

[49] LANGE, Kenneth L.; LITTLE, Roderick JA; TAYLOR, Jeremy MG. Robust statistical modeling using the t distribution. **Journal of the American Statistical Association**, v. 84, n. 408, p. 881-896, 1989.

[50] LANGE, Kenneth; SINSHEIMER, Janet S. Normal/independent distributions and their applications in robust regression. **Journal of Computational and Graphical Statistics**, v. 2, n. 2, p. 175-198, 1993.

[51] LEE, Sharon; MCLACHLAN, Geoffrey J. Finite mixtures of multivariate skew t-distributions: some recent and new results. **Statistics and Computing**, v. 24, n. 2, p. 181-202, 2014.

[52] LIMA, Elon Lages. Algebra linear, 2a. ediçao. **IMPA**, Rio de Janeiro, p. 22, 1996.

[53] LIN, Tsung I. Maximum likelihood estimation for multivariate skew normal mixture models. **Journal of Multivariate Analysis**, v. 100, n. 2, p. 257-265, 2009.

[54] LIN, Tsung-I. Robust mixture modeling using multivariate skew t distributions. **Statistics and Computing**, v. 20, n. 3, p. 343-356, 2010.

[55] LIN, Tsung-I.; LEE, Jack C.; HSIEH, Wan J. Robust mixture modeling using the skew t distribution. **Statistics and computing**, v. 17, n. 2, p. 81-92, 2007.

[56] LIN, Tsung-I.; LEE, Jack C.; YEN, Shu Y. Finite mixture modelling using the skew normal distribution. **Statistica Sinica**, p. 909-927, 2007.

[57] LINDSAY, Bruce G. Mixture models: theory, geometry and applications. In: **NSF-CBMS regional conference series in probability and statistics**. Institute of Mathematical Statistics and the American Statistical Association, p. i-163, 1995.

[58] LIU, Chuanhai. Efficient ML estimation of the multivariate normal distribution from incomplete data. **Journal of Multivariate Analysis**, v. 69, n. 2, p. 206-217, 1999.

[59] MA, Yanyuan; GENTON, Marc G. Flexible class of skew-symmetric distributions. **Scandinavian Journal of Statistics**, v. 31, n. 3, p. 459-468, 2004.

[60] MCLACHLAN, Geoffrey J.; BASFORD, Kaye E. **Mixture models: Inference and applications to clustering**. New York: M. Dekker, 1988.

[61] MCLACHLAN, Geoffrey J.; KRISHNAN, Thriyambakam. **The EM algorithm and extensions**. John Wiley & Sons, 2007.

[62] MCLACHLAN, Geoffrey J.; PEEL, David. **Finite mixture models**. John Wiley & Sons, 2004.

[63] MUST, Aviva; DALLAL, Gerard E.; DIETZ, William H. Reference data for obesity: 85th and 95th percentiles of body mass index (wt/ht2) and triceps skinfold thickness. **The American journal of clinical nutrition**, v. 53, n. 4, p. 839-846, 1991.

[64] OSORIO, Felipe; PAULA, Gilberto A.; GALEA, Manuel. Assessment of local influence in elliptical linear models with longitudinal structure. **Computational Statistics & Data Analysis**, v. 51, n. 9, p. 4354-4368, 2007.

[65] PEEL, David; MCLACHLAN, Geoffrey J. Robust mixture modelling using the t distribution. **Statistics and computing**, v. 10, n. 4, p. 339-348, 2000.

[66] PRATES, Marcos Oliveira; LACHOS, Victor Hugo; CABRAL, Celso Rômulo Barbosa. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. **Journal of Statistical Software**, v. 54, n. 12, p. 1-20, 2013.

[67] PRAUSE, Karsten. **The generalized hyperbolic model: Estimation, financial derivatives and risk measures**. 1999. Tese (Doutorado em Matemática) - Institut für Mathematische Stochastik, Albert-Ludwigs-Universität Freiburg, Eckerstraße 1, D–79104 Freiburg im Breisgau 1999.

[68] PROTASSOV, Rostislav S. EM-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed $\boldsymbol{\lambda}$. **Statistics and Computing**, v. 14, n. 1, p. 67-77, 2004.

[69] QUANDT, Richard E.; RAMSEY, James B. Estimating mixtures of normal distributions and switching regressions. **Journal of the American statistical Association**, v. 73, n. 364, p. 730-738, 1978.

[70] R Development Core Team. R: A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2020. Disponível em: <http://www.r-project.org>. Acesso em: 20 nov. 2020.

[71] RINEHART, John S. Thermal and seismic indications of Old Faithful Geyser's inner workings. **Journal of Geophysical Research**, v. 74, n. 2, p. 566-573, 1969.

[72] SCHWARZ, Gideon et al. Estimating the dimension of a model. **The annals of statistics**, v. 6, n. 2, p. 461-464, 1978.

[73] SHOHAM, Shy. Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. **Pattern Recognition**, v. 35, n. 5, p. 1127-1142, 2002.

[74] SILVERMAN, Bernhard W. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 47, n. 1, p. 1-21, 1985.

[75] SNOUSSI, Hichem; IDIER, Jérôme. Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures. **IEEE Transactions on Signal Processing**, v. 54, n. 9, p. 3257-3269, 2006.

[76] TEICHER, Henry. Identifiability of finite mixtures. **The annals of Mathematical statistics**, p. 1265-1269, 1963.

[77] TITTERINGTON, D. Michael; SMITH, Adrian FM; MAKOV, Udi E. **Statistical analysis of finite mixture distributions**.John Wiley & Sons, 1985.

[78] VILCA, Filidor; BALAKRISHNAN, Narayanaswamy; ZELLER, Camila Borelli. Multivariate skew-normal generalized hyperbolic distribution and its properties. **Journal of Multivariate Analysis**, v. 128, p. 73-85, 2014.

[79] WANG, Wan-Lun; JAMALIZADEH, Ahad; LIN, Tsung-I. Finite mixtures of multivariate scale-shape mixtures of skew-normal distributions. **Statistical Papers**, p. 1-28, 2018.

[80] WOLFE, John H. **NORMIX: Computational methods for estimating the parameters of multivariate normal mixtures of distributions**. Naval Personnel Research Activity San Diego Calif, 1967.

[81] WRAITH, Darren; FORBES, Florence. Location and scale mixtures of Gaussians with flexible tail behaviour: Properties, inference and application to multivariate clustering. **Computational Statistics & Data Analysis**, v. 90, p. 61-73, 2015.

[82] YAKOWITZ, Sidney J.; SPRAGINS, John D. On the identifiability of finite mixtures. **The Annals of Mathematical Statistics**, p. 209-214, 1968.

[83] ZELLER, Camila Borelli. **Distribuições misturas de escala skew-normal: estimação e diagnostico em modelos lineares**. 2009. Tese (Doutorado em Estatística) - Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas, 2009.

[84] ZELLER, Camila Borelli; CABRAL, Celso Rômulo Barbosa; DÁVILA, Víctor Hugo Lachos; BENITES, Luis. Finite mixture of regression models for censored data based on scale mixtures of normal distributions. **Advances in Data Analysis and Classification**, v. 13, n. 1, p. 89-116, 2018.

[85] ZELLER, Camila Borelli; CABRAL, Celso Rômulo Barbosa; LACHOS, Víctor Hugo. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. **Test**, v. 25, n. 2, p. 375-396, 2016.

# APÊNDICE A – Algumas propriedades da função de Bessel

Algumas propriedades da função modificada de Bessel do terceiro tipo $K_\eta(\mathrm{x})$ são apresentadas abaixo. Para mais detalhes, veja [38].

(i) $K_\eta(\mathrm{x}) = K_{-\eta}(\mathrm{x})$,

(ii) $K_{\eta+1}(\mathrm{x}) = \dfrac{2\eta}{\mathrm{x}} K_\eta(\mathrm{x}) + K_{\eta-1}(\mathrm{x})$,

(iii) Para $r \in \mathbb{N} \cup \{0\}$, $K_{r+1/2}(\mathrm{x}) = \sqrt{\dfrac{\pi}{2\mathrm{x}}}\, e^{(-\mathrm{x})} \displaystyle\sum_{k=0}^{r} \frac{(r+k)!(2u)^{-k}}{(r-k)!k!}$.

Em particular, tem-se:

- $K_{1/2}(\mathrm{x}) = \sqrt{\dfrac{\pi}{2\mathrm{x}}} e^{(-\mathrm{x})}$,
- $K_{3/2}(\mathrm{x}) = \sqrt{\dfrac{\pi}{2}} e^{-\mathrm{x}} \mathrm{x}^{-3/2}(\mathrm{x}+1)$,
- $K_{5/2}(\mathrm{x}) = \sqrt{\dfrac{\pi}{2}} e^{-\mathrm{x}} \mathrm{x}^{-7/2}(\mathrm{x}^2 + 3\mathrm{x} + 3)$,

(iv) $\dfrac{\partial K_\eta(\mathrm{x})}{\partial \mathrm{x}} = \dfrac{\eta}{\mathrm{x}} K_\eta(\mathrm{x}) - K_{\eta-1}(\mathrm{x})$,

(v) $\dfrac{\partial \log K_\eta(\mathrm{x})}{\partial \mathrm{x}} = -\dfrac{\eta}{\mathrm{x}} - \dfrac{K_{\eta-1}(\mathrm{x})}{K_\eta(\mathrm{x})}$,

(vi) $\lim\limits_{\mathrm{z}\to 0} \mathrm{z}^\eta K_\eta(\mathrm{z}) = 2^{\eta-1}\Gamma(\eta),\ \eta > 0.$

# APÊNDICE B – Resultados adicionais do Capítulo 3

Cabe ressaltar que os resultados aqui expressos podem ser encontrados com mais detalhes em Calegari (2020) [25]. Desta forma, a seguir são apresentados os seguintes valores esperados $\widehat{ut_{ij}}^{(k)}$, $\widehat{ut^2_{ij}}^{(k)}$, $\widehat{s_{ij}}^{(k)}$ e $\widehat{v_{ij}}^{(k)}$, onde para o contexto univariado, assuma $p = 1$.

$$\widehat{u_{ij}}^{(k)} = E\left[{U_i}^{-1}|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \frac{2\widehat{f_0\left(\mathbf{y}_i\right)}}{\widehat{f_{\mathbf{Y}}\left(\mathbf{y}_i\right)}} \frac{c(\eta - {}^{p}/_{2}, \widehat{q\left(\mathbf{y}_i\right)^2}, \omega)}{c\left(\eta - {}^{p}/_{2} - 1, \widehat{q\left(\mathbf{y}_i\right)^2}, \omega\right)} F_X\left(\widehat{C_{\mathbf{y}_i}}\right), \quad \text{(B.1)}$$

$$\widehat{ut_{ij}}^{(k)} = E\left[{U_i}^{-1}T_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \widehat{\boldsymbol{\mu}_{T_i}}\widehat{u_i} + \widehat{M_T}\widehat{\eta_{1i}}, \quad \text{(B.2)}$$

$$\widehat{ut^2_{ij}}^{(k)} = E\left[{U_i}^{-1}{T_i}^2|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \widehat{\boldsymbol{\mu}_{T_i}}^2\widehat{u_i} + \widehat{M_T}^2 + \widehat{M_T}\widehat{\boldsymbol{\mu}_{T_i}}\widehat{\eta_{1i}}, \quad \text{(B.3)}$$

$$\widehat{s_{ij}}^{(k)} = E\left[\log U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \frac{2\widehat{f_0\left(\mathbf{y}_i\right)}}{f_{\mathbf{Y}(\mathbf{y}_i)}} E\left[\log(U)\Phi\left(U^{-1/2}\widehat{C_{y_i}}\right)\right] \quad \text{(B.4)}$$

$$\widehat{v_{ij}}^{(k)} = E\left[U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \frac{2\widehat{f_0\left(\mathbf{y}_i\right)}}{\widehat{f_{\mathbf{Y}}\left(\mathbf{y_i}\right)}} \frac{c\left(\hat{\eta} - {}^{p}/_{2}, \widehat{\omega} + \widehat{d\left(\mathbf{y}_i\right)}, \widehat{\omega}\right)}{c\left(\hat{\eta} - {}^{p}/_{2} + 1, \widehat{\omega} + \widehat{d\left(\mathbf{y}_i\right)}, \widehat{\omega}\right)} F_W\left(\widehat{C_{\mathbf{y}_i}}\right), \quad \text{(B.5)}$$

onde $\widehat{\eta_{1i}} = \dfrac{2\widehat{f_0\left(\mathbf{y}_i\right)}}{\widehat{f_{\mathbf{Y}}\left(\mathbf{y}_i\right)}} \dfrac{c(\eta - {}^{p}/_{2}, \widehat{q\left(\mathbf{y}_i\right)^2}, \omega)}{c\left(\eta - {}^{p}/_{2} - 1, \widehat{C_{\mathbf{y}_i}}^2 + \widehat{q\left(\mathbf{y}_i\right)^2}, \omega\right)\sqrt{2\pi}}$, $\widehat{M_T}^2 = \dfrac{1}{1 + \widehat{\boldsymbol{\Delta}}^{\top}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}}$, $\widehat{\boldsymbol{\mu}_{T_i}} = \widehat{M_T}^2\widehat{\boldsymbol{\Delta}}^{\top}\widehat{\boldsymbol{\Gamma}}^{-1}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}\right)$, $\widehat{C_{\mathbf{y}_i}} = \hat{\boldsymbol{\lambda}}^{\top}\hat{\boldsymbol{\Sigma}}^{-1/2}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}\right)$, $\widehat{d(\mathbf{y}_i)} = \left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}\right)\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}\right)$, $\mathbf{X} \sim \text{GH}(0, 1; \boldsymbol{\nu}_{-1/2})$ com $\boldsymbol{\nu}_{-1/2} = (\eta - {}^{p}/_{2} - 1, \widehat{q\left(\mathbf{y}_i\right)^2}, \omega)^{\top}$ e $\mathbf{W} \sim \text{GH}(0, 1; \boldsymbol{\nu}_{\mathbf{y}_i})$ em que $\boldsymbol{\nu}_{\mathbf{y}_i} = (\eta - {}^{p}/_{2} + 1, \widehat{\omega} + \widehat{d\left(\mathbf{y}_i\right)}, \omega)^{\top}$.

# APÊNDICE C – Resumos das estimativas dos parâmetros no modelo de FM-SNGH

Nas Tabelas 8 à 20, encontram-se descritos os valores médios (AV) e os correspondentes desvio padrão (SD) das estimativas via algoritmo EM em todas as amostras, assim como o viés médio (BIAS) e erro quadrático médio (MSE), de cada parâmetro analisado de acordo com seus referidos cenários e contextos, definidos na Seção 4.1, conforme varia-se o tamanho amostral $n$.

Tabela 8 – Resumo das estimativas de $\rho$ em ambos cenários univariado, quando $\rho_1 = 0{,}4$.

| $n$ | Medidas | Cenário 1 | Cenário 2 |
|---|---|---|---|
| | | $\rho_1$ | $\rho_1$ |
| 100 | AV | 0,39512 | 0,39530 |
| | SD | 0,04795 | 0,06969 |
| | BIAS | 0,03854 | 0,04767 |
| | MSE | 2,3e-03 | 0,00485 |
| 300 | AV | 0,39571 | 0,39929 |
| | SD | 0,02965 | 0,03057 |
| | BIAS | 0,02427 | 0,02401 |
| | MSE | 0,00089 | 0,00092 |
| 500 | AV | 0,39861 | 0,40098 |
| | SD | 0,02190 | 0,02246 |
| | BIAS | 0,01795 | 0,01773 |
| | MSE | 0,00047 | 0,00050 |
| 1000 | AV | 0,40056 | 0,40368 |
| | SD | 0,01448 | 0,01608 |
| | BIAS | 0,01163 | 0,01317 |
| | MSE | 0,00020 | 0,00026 |

Fonte: Elaborada pelo autor (2020).

Tabela 9 – Resumo das estimativas de $\mu_1$ e $\mu_2$ em ambos cenários univariado, quando $\mu_1 = 15$ e $\mu_2 = 20$.

| $n$ | Medidas | Cenário 1 | | Cenário 2 | |
|---|---|---|---|---|---|
| | | $\mu_1$ | $\mu_2$ | $\mu_1$ | $\mu_2$ |
| 100 | AV | 15,10371 | 19,82773 | 15,0340 | 19,5695 |
| | SD | 0,82359 | 0,89479 | 0,82040 | 1,07455 |
| | BIAS | 0,24024 | 0,39163 | 0,30814 | 0,72895 |
| | MSE | 6,8e-01 | 8,2e-01 | 0,67073 | 1,33397 |
| 300 | AV | 15,00445 | 20,02827 | 14,9829 | 19,9094 |
| | SD | 0,06741 | 0,18476 | 0,10754 | 0,36534 |
| | BIAS | 0,05328 | 0,14787 | 0,08487 | 0,29648 |
| | MSE | 0,00454 | 0,03480 | 0,01179 | 0,14091 |
| 500 | AV | 15,00542 | 20,04605 | 15,0079 | 19,9016 |
| | SD | 0,06292 | 0,12116 | 0,08541 | 0,27310 |
| | BIAS | 0,03829 | 0,10608 | 0,06857 | 0,21340 |
| | MSE | 0,00397 | 0,01674 | 0,00730 | 0,08366 |
| 1000 | AV | 15,01099 | 20,02047 | 15,0255 | 19,8761 |
| | SD | 0,03292 | 0,08796 | 0,04804 | 0,19082 |
| | BIAS | 0,02816 | 0,07178 | 0,04259 | 0,18107 |
| | MSE | 0,00120 | 0,00812 | 0,00291 | 0,05106 |

Fonte: Elaborada pelo autor (2020).

Tabela 10 – Resumo das estimativas de $\sigma_1^2$ e $\sigma_2^2$ em ambos cenários univariado, quando $\sigma_1^2 = \sigma_2^2 = 1$.

| $n$ | Medidas | Cenário 1 | | Cenário 2 | |
|---|---|---|---|---|---|
| | | $\sigma_1^2$ | $\sigma_2^2$ | $\sigma_1^2$ | $\sigma_2^2$ |
| 100 | AV | 0,97268 | 1,26441 | 0,65917 | 0,71916 |
| | SD | 0,38173 | 0,73333 | 0,61355 | 0,73396 |
| | BIAS | 0,29006 | 0,47006 | 0,61878 | 0,65890 |
| | MSE | 1,4e-01 | 6,0e-01 | 0,49065 | 0,61477 |
| 300 | AV | 0,97295 | 1,06181 | 0,73890 | 0,75802 |
| | SD | 0,21017 | 0,27217 | 0,35058 | 0,39844 |
| | BIAS | 0,16815 | 0,21433 | 0,37277 | 0,39941 |
| | MSE | 0,04473 | 0,07761 | 0,19037 | 0,21640 |
| 500 | AV | 0,99612 | 1,04604 | 0,79052 | 0,76343 |
| | SD | 0,16724 | 0,20148 | 0,23642 | 0,26947 |
| | BIAS | 0,12816 | 0,16305 | 0,27487 | 0,31012 |
| | MSE | 0,02787 | 0,04255 | 0,09933 | 0,12800 |
| 1000 | AV | 0,99340 | 1,00865 | 0,81986 | 0,75753 |
| | SD | 0,11134 | 0,13289 | 0,13539 | 0,16547 |
| | BIAS | 0,09084 | 0,10415 | 0,19415 | 0,26410 |
| | MSE | 0,01239 | 0,01766 | 0,05042 | 0,08564 |

Fonte: Elaborada pelo autor (2020).

Tabela 11 – Resumo das estimativas de $\lambda_1$ e $\lambda_2$ em ambos cenários univariado, quando $\lambda_1$ = -4 e $\lambda_2$ = -1.

| $n$ | Medidas | Cenário 1 | | Cenário 2 | |
|---|---|---|---|---|---|
| | | $\lambda_1$ | $\lambda_2$ | $\lambda_1$ | $\lambda_2$ |
| 100 | AV | -8,23336 | -1,39590 | -5,8019 | -1,0464 |
| | SD | 11,74563 | 1,50092 | 7,01793 | 2,04064 |
| | BIAS | 5,71539 | 0,99628 | 4,15913 | 1,33002 |
| | MSE | 1,5e+02 | 2,4e+00 | 52,2433 | 4,14479 |
| 300 | AV | -4,92495 | -1,20529 | -5,2017 | -0,9991 |
| | SD | 2,24262 | 0,59434 | 3,47347 | 0,69035 |
| | BIAS | 1,64950 | 0,47782 | 2,01198 | 0,53172 |
| | MSE | 5,86557 | 0,39403 | 13,4407 | 0,47388 |
| 500 | AV | -4,58130 | -1,17437 | -5,0600 | -0,9300 |
| | SD | 1,39585 | 0,39243 | 1,80547 | 0,45221 |
| | BIAS | 1,09121 | 0,33186 | 1,50084 | 0,36640 |
| | MSE | 2,27864 | 0,18380 | 4,35759 | 0,20777 |
| 1000 | AV | -4,40788 | -1,08264 | -5,0586 | -0,8491 |
| | SD | 0,93496 | 0,27182 | 1,04234 | 0,34711 |
| | BIAS | 0,75990 | 0,21766 | 1,16597 | 0,32287 |
| | MSE | 1,03705 | 0,08042 | 2,18642 | 0,14092 |

Fonte: Elaborada pelo autor (2020).

Tabela 12 – Resumo das estimativas de $\omega$ em ambos cenários univariado, quando $\omega$ = 1.

| $n$ | Medidas | Cenário 1 | Cenário 2 |
|---|---|---|---|
| | | $\omega$ | $\omega$ |
| 100 | AV | 2,91244 | 0,69599 |
| | SD | 3,90344 | 0,83129 |
| | BIAS | 2,29137 | 0,73621 |
| | MSE | 1,8e+01 | 0,77988 |
| 300 | AV | 0,81434 | 0,69475 |
| | SD | 1,11441 | 0,41854 |
| | BIAS | 0,84658 | 0,42799 |
| | MSE | 1,27390 | 0,26736 |
| 500 | AV | 0,73276 | 0,70025 |
| | SD | 0,92856 | 0,26767 |
| | BIAS | 0,76116 | 0,35181 |
| | MSE | 0,93191 | 0,16092 |
| 1000 | AV | 1,26571 | 0,72629 |
| | SD | 0,34910 | 0,15518 |
| | BIAS | 0,32758 | 0,28856 |
| | MSE | 0,19199 | 0,09853 |

Fonte: Elaborada pelo autor (2020).

Tabela 13 – Resumo das estimativas de $\rho$ em ambos cenários multivariado, quando $\rho_1 = 0{,}7$.

| $n$ | Medidas | Cenário 1 | Cenário 2 |
|---|---|---|---|
| | | $\rho_1$ | $\rho_1$ |
| 100 | AV | 0,69810 | 0,69740 |
| | SD | 0,05660 | 0,05732 |
| | BIAS | 0,03833 | 0,03954 |
| | MSE | 0,00320 | 0,00328 |
| 300 | AV | 0,69770 | 0,69897 |
| | SD | 0,04141 | 0,04211 |
| | BIAS | 0,02289 | 0,02355 |
| | MSE | 0,00171 | 0,00177 |
| 500 | AV | 0,69869 | 0,69838 |
| | SD | 0,03789 | 0,03754 |
| | BIAS | 0,01867 | 0,01752 |
| | MSE | 0,00143 | 0,00140 |
| 1000 | AV | 0,69811 | 0,69903 |
| | SD | 0,03434 | 0,03453 |
| | BIAS | 0,01259 | 0,01304 |
| | MSE | 0,00118 | 0,00119 |

Fonte: Elaborada pelo autor (2020).

Tabela 14 – Resumo das estimativas de $\boldsymbol{\mu}_1$ em ambos cenários multivariado, quando $\boldsymbol{\mu}_1 = (0, 0)$.

| $n$ | Medidas | Cenário 1 | | Cenário 2 | |
|---|---|---|---|---|---|
| | | $\boldsymbol{\mu}_1$ | | $\boldsymbol{\mu}_1$ | |
| 100 | AV | 0,03476 | 0,03773 | 0,07140 | 0,04926 |
| | SD | 0,33045 | 0,13961 | 0,40906 | 0,18529 |
| | BIAS | 0,24783 | 0,09816 | 0,30538 | 0,12640 |
| | MSE | 0,11019 | 0,02087 | 0,17209 | 0,03669 |
| 300 | AV | 1,5e-02 | 2,1e-05 | 0,02644 | 0,00688 |
| | SD | 0,15753 | 0,05869 | 0,19814 | 0,08397 |
| | BIAS | 0,12077 | 0,04576 | 0,15435 | 0,06511 |
| | MSE | 0,02501 | 0,00343 | 0,03988 | 0,00708 |
| 500 | AV | 0,00812 | -0,0023 | 0,00765 | 0,00105 |
| | SD | 0,11233 | 0,04292 | 0,15607 | 0,06164 |
| | BIAS | 0,08959 | 0,03410 | 0,11910 | 0,04706 |
| | MSE | 0,01265 | 0,00184 | 0,02436 | 0,00379 |
| 1000 | AV | 0,00428 | -0,0011 | -0,0006 | -0,0079 |
| | SD | 0,07511 | 0,03150 | 0,10065 | 0,03718 |
| | BIAS | 0,05869 | 0,02484 | 0,08182 | 0,02980 |
| | MSE | 0,00564 | 0,00099 | 0,01011 | 0,00144 |

Fonte: Elaborada pelo autor (2020).

Tabela 15 – Resumo das estimativas de $\boldsymbol{\mu}_2$ em ambos cenários multivariado, quando $\boldsymbol{\mu}_2 = (5, 5)$.

| $n$ | Medidas | Cenário 1 | | Cenário 2 | |
|---|---|---|---|---|---|
| | | $\boldsymbol{\mu_2}$ | | $\boldsymbol{\mu_2}$ | |
| 100 | AV | 5,22068 | 5,28377 | 5,40612 | 5,34793 |
| | SD | 0,82569 | 0,60004 | 1,00408 | 0,68941 |
| | BIAS | 0,66041 | 0,45665 | 0,83272 | 0,53316 |
| | MSE | 0,72910 | 0,43986 | 1,17110 | 0,59539 |
| 300 | AV | 5,11289 | 5,08861 | 5,15857 | 5,13174 |
| | SD | 0,15753 | 0,05869 | 0,65031 | 0,45701 |
| | BIAS | 0,37878 | 0,20948 | 0,47875 | 0,28847 |
| | MSE | 0,31061 | 0,13464 | 0,44720 | 0,22579 |
| 500 | AV | 5,04784 | 5,04873 | 5,08169 | 5,08528 |
| | SD | 0,43629 | 0,31113 | 0,34676 | 0,21245 |
| | BIAS | 0,27676 | 0,16646 | 0,11910 | 0,04706 |
| | MSE | 0,19226 | 0,09898 | 0,28630 | 0,14340 |
| 1000 | AV | 5,01884 | 5,01569 | 5,03627 | 5,01733 |
| | SD | 0,33932 | 0,25389 | 0,39556 | 0,27654 |
| | BIAS | 0,19261 | 0,10242 | 0,23008 | 0,12426 |
| | MSE | 0,11526 | 0,06457 | 0,15747 | 0,07662 |

Fonte: Elaborada pelo autor (2020).

Tabela 16 – Resumo das estimativas de $\boldsymbol{\lambda}_1$ em ambos cenários multivariado, quando $\boldsymbol{\lambda}_1 = (1, 4)$.

| $n$ | Medidas | Cenário 1 | | Cenário 2 | |
|---|---|---|---|---|---|
| | | $\boldsymbol{\lambda_1}$ | | $\boldsymbol{\lambda_1}$ | |
| 100 | AV | 1,29633 | 5,34998 | 1,10110 | 5,10655 |
| | SD | 1,82886 | 3,42520 | 1,80964 | 3,84173 |
| | BIAS | 0,12801 | 0,10557 | 0,13563 | 0,10816 |
| | MSE | 0,03401 | 0,01761 | 0,03727 | 0,01950 |
| 300 | AV | 1,02273 | 4,38453 | 0,99078 | 4,36051 |
| | SD | 0,65201 | 1,13349 | 0,71758 | 1,51904 |
| | BIAS | 0,05500 | 0,05689 | 0,06137 | 0,06183 |
| | MSE | 0,00663 | 0,00517 | 0,00821 | 0,00610 |
| 500 | AV | 1,01982 | 4,23287 | 1,03409 | 4,22319 |
| | SD | 0,50348 | 0,81435 | 0,52921 | 0,94124 |
| | BIAS | 0,04345 | 0,04566 | 0,04942 | 0,04853 |
| | MSE | 0,00463 | 0,00333 | 0,00577 | 0,00385 |
| 1000 | AV | 1,01713 | 4,13783 | 1,04172 | 4,19959 |
| | SD | 0,31175 | 0,58827 | 0,35076 | 0,59490 |
| | BIAS | 0,03029 | 0,02902 | 0,03658 | 0,03504 |
| | MSE | 0,00322 | 0,00135 | 0,00391 | 0,00186 |

Fonte: Elaborada pelo autor (2020).

Tabela 17 – Resumo das estimativas de $\boldsymbol{\lambda}_2$ em ambos cenários multivariado, quando $\boldsymbol{\lambda}_2 = (1, 2)$.

| $n$ | Medidas | Cenário 1 | | Cenário 2 | |
|---|---|---|---|---|---|
| | | $\boldsymbol{\lambda_2}$ | | $\boldsymbol{\lambda_2}$ | |
| 100 | AV | 1,30687 | 3,54677 | 1,14239 | 3,20341 |
| | SD | 6,62357 | 7,10203 | 5,49259 | 6,29529 |
| | BIAS | 0,12254 | 0,55622 | 0,13083 | 0,52409 |
| | MSE | 0,02809 | 0,38406 | 0,03215 | 0,36710 |
| 300 | AV | 0,93666 | 2,06555 | 0,91161 | 2,05482 |
| | SD | 1,07352 | 1,09989 | 1,09228 | 1,65795 |
| | BIAS | 0,05887 | 0,57531 | 0,07390 | 0,54791 |
| | MSE | 0,00724 | 0,36822 | 0,01091 | 0,34305 |
| 500 | AV | 0,99624 | 2,00986 | 0,95311 | 1,96426 |
| | SD | 0,74699 | 0,81683 | 0,78035 | 0,86747 |
| | BIAS | 0,04500 | 0,57834 | 0,05628 | 0,54497 |
| | MSE | 0,00493 | 0,35975 | 0,00676 | 0,32194 |
| 1000 | AV | 0,98353 | 1,98401 | 0,97477 | 2,02041 |
| | SD | 0,49509 | 0,48557 | 0,52832 | 0,54227 |
| | BIAS | 0,03296 | 0,60054 | 0,04047 | 0,57396 |
| | MSE | 0,00353 | 0,37565 | 0,00424 | 0,34417 |

Fonte: Elaborada pelo autor (2020).

Tabela 18 – Resumo das estimativas de $\boldsymbol{\Sigma}_1$ em ambos cenários multivariado, quando $\sigma_{11} = 1$, $\sigma_{12} = 0$ e $\sigma_{22} = 1$.

| $n$ | Medidas | Cenário 1 | | | Cenário 2 | | |
|---|---|---|---|---|---|---|---|
| | | $\boldsymbol{\sigma_{11}}$ | $\boldsymbol{\sigma_{12}}$ | $\boldsymbol{\sigma_{22}}$ | $\boldsymbol{\sigma_{11}}$ | $\boldsymbol{\sigma_{12}}$ | $\boldsymbol{\sigma_{22}}$ |
| 100 | AV | 1,04458 | -0,0197 | 0,96125 | 1,02468 | -0,0278 | 0,95614 |
| | SD | 0,17915 | 0,13137 | 0,16324 | 0,19166 | 0,13700 | 0,17403 |
| | BIAS | 0,45753 | 0,74691 | 1,33120 | 0,44454 | 0,72025 | 1,24170 |
| | MSE | 0,25926 | 0,63547 | 3,42587 | 0,25555 | 0,60053 | 3,27849 |
| 300 | AV | 0,99904 | -0,0144 | 0,98229 | 1,01270 | -0,0128 | 1,01051 |
| | SD | 0,08155 | 0,07054 | 0,08334 | 0,08982 | 0,07715 | 0,10405 |
| | BIAS | 0,38711 | 0,67689 | 0,49735 | 0,38691 | 0,64836 | 0,54417 |
| | MSE | 0,17607 | 0,49769 | 0,42479 | 0,17878 | 0,46042 | 0,51398 |
| 500 | AV | 0,99108 | -0,0073 | 0,98735 | 1,02282 | -0,0043 | 1,01516 |
| | SD | 0,06754 | 0,05737 | 0,06916 | 0,07252 | 0,06195 | 0,08094 |
| | BIAS | 0,36036 | 0,64635 | 0,39188 | 0,36145 | 0,62490 | 0,41472 |
| | MSE | 0,14598 | 0,44628 | 0,25338 | 0,14813 | 0,42005 | 0,28067 |
| 1000 | AV | 0,99398 | -0,0032 | 0,98934 | 1,01604 | 0,00116 | 1,02178 |
| | SD | 0,05652 | 0,03669 | 0,05851 | 0,06054 | 0,04324 | 0,06143 |
| | BIAS | 0,34070 | 0,63424 | 0,24495 | 0,33597 | 0,59708 | 0,28893 |
| | MSE | 0,12505 | 0,41755 | 0,09728 | 0,12209 | 0,37348 | 0,12452 |

Fonte: Elaborada pelo autor (2020).

Tabela 19 – Resumo das estimativas de $\mathbf{\Sigma}_2$ em ambos cenários multivariado, quando $\sigma_{11} = 2$, $\sigma_{12} = 1/2$ e $\sigma_{22} = 2$.

| $n$ | Medidas | Cenário 1 | | | Cenário 2 | | |
|---|---|---|---|---|---|---|---|
| | | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ | $\sigma_{11}$ | $\sigma_{12}$ | $\sigma_{22}$ |
| 100 | AV | 1,47785 | 0,04564 | 1,25588 | 1,51389 | 0,06008 | 1,28012 |
| | SD | 0,33415 | 0,23006 | 0,28624 | 0,36203 | 0,24932 | 0,28718 |
| | BIAS | 2,39837 | 3,07256 | 3,41027 | 2,42346 | 2,85252 | 2,94065 |
| | MSE | 13,5310 | 43,8781 | 52,7305 | 15,9538 | 30,1285 | 40,9996 |
| 300 | AV | 1,42555 | 0,11291 | 1,32310 | 1,45544 | 0,11308 | 1,35163 |
| | SD | 0,19574 | 0,16213 | 0,19897 | 0,21588 | 0,17071 | 0,20032 |
| | BIAS | 0,87696 | 0,79835 | 0,83523 | 1,04943 | 0,84105 | 0,95612 |
| | MSE | 1,43010 | 1,15416 | 1,21164 | 2,43285 | 1,19850 | 2,74633 |
| 500 | AV | 1,42165 | 0,13963 | 1,35364 | 1,45502 | 0,13854 | 1,37509 |
| | SD | 0,15910 | 0,12710 | 0,16899 | 0,15809 | 0,13234 | 0,17207 |
| | BIAS | 0,64552 | 0,54451 | 0,61464 | 0,72134 | 0,57844 | 0,65765 |
| | MSE | 0,71608 | 0,55690 | 0,66598 | 0,93399 | 0,60993 | 0,75227 |
| 1000 | AV | 1,39945 | 0,15929 | 1,36575 | 1,42603 | 0,16402 | 1,40291 |
| | SD | 0,12261 | 0,09482 | 0,12378 | 0,12153 | 0,09606 | 0,13040 |
| | BIAS | 0,44546 | 0,36677 | 0,37531 | 0,47798 | 0,38287 | 0,41399 |
| | MSE | 0,36437 | 0,24489 | 0,23556 | 0,39303 | 0,27920 | 0,29389 |

Fonte: Elaborada pelo autor (2020).

Tabela 20 – Resumo das estimativas de $\omega$ em ambos cenários multivariado, quando $\omega = 2$.

| $n$ | Medidas | Cenário 1 | Cenário 2 |
|---|---|---|---|
| | | $\omega$ | $\omega$ |
| 100 | AV | 4,08117 | 2,96956 |
| | SD | 4,15339 | 2,12152 |
| | BIAS | 2,64118 | 1,45769 |
| | MSE | 21,5474 | 5,43192 |
| 300 | AV | 2,63312 | 2,41139 |
| | SD | 1,25589 | 0,69620 |
| | BIAS | 0,84080 | 0,59117 |
| | MSE | 1,97496 | 0,65298 |
| 500 | AV | 2,43798 | 2,38542 |
| | SD | 0,68087 | 0,55091 |
| | BIAS | 0,60264 | 0,49745 |
| | MSE | 0,65449 | 0,45144 |
| 1000 | AV | 2,28427 | 2,27236 |
| | SD | 0,43369 | 0,32330 |
| | BIAS | 0,38997 | 0,33344 |
| | MSE | 0,26853 | 0,17850 |

Fonte: Elaborada pelo autor (2020).