Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Programa de Pós-Graduação em Matemática

Graciliano Márcio Santos Louredo

Estimação via EM e Diagnóstico em Modelos Misturas Assimétricas com Regressão

Juiz de Fora
2018

Graciliano Márcio Santos Louredo

Estimação via EM e Diagnóstico em Modelos Misturas Assimétricas com Regressão

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade Federal de Juiz de Fora, na área de concentração em Matemática Aplicada como requisito parcial para obtenção do título de Mestre em Matemática.

Orientadora: Camila Borelli Zeller

Coorientador: Clécio da Silva Ferreira

Juiz de Fora

2018

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Louredo, Graciliano Márcio Santos .
Estimação via EM e diagnóstico em modelos misturas assimétricas com regressão / Graciliano Márcio Santos Louredo. -- 2018.
154 f. : il.

Orientadora: Camila Borelli Zeller
Coorientador: Clécio da Silva Ferreira
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós Graduação em Matemática, 2018.

1. Modelos misturas assimétricas. 2. Regressão linear multivariada. 3. Estimação por máxima verossimilhança. 4. Algoritmo EM. 5. Influência global e local. I. Zeller, Camila Borelli , orient. II. Ferreira, Clécio da Silva, coorient. III. Título.

Graciliano Márcio Santos Louredo

Estimação via EM e Diagnóstico em Modelos Misturas Assimétricas com Regressão

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade Federal de Juiz de Fora, na área de concentração em Matemática Aplicada como requisito parcial para obtenção do título de Mestre em Matemática.

Aprovada em:

BANCA EXAMINADORA

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

Professora. Dra. Camila Borelli Zeller - Orientador
Universidade Federal de Juiz de Fora

Professora Dra. Lucy Tiemi Takahashi - Interno
Universidade Federal de Juiz de Fora

Professor Dr. Filidor Edilfonso Vilca Labra - Externo
Universidade Estadual de Campinas

AGRADECIMENTOS

Agradeço a Deus por me permitir estar em algum plano de existência.

Sou muito grato aos meus pais, Wanderlei e Maria de Fátima, familiares e amigos por muitas vezes acreditarem mais em mim do que eu mesmo.

A todos professores que fizeram parte desta longa trajetória por terem contribuído para a formação do meu modo de pensar. Aos do Departamento de Matemática por terem me fornecido grande parte do embasamento teórico que tenho atualmente e aos do Departamento de Estatística por me ajudarem na busca pela interdisciplinaridade através da compreensão do "charme" dos dados. Sem dúvida, uma boa mistura.

Em especial, agradeço à minha orientadora Camila pela enorme paciência com a minha compreensão sempre incompleta da estatística e por ter me incentivado durante todo o período de execução deste trabalho. Igualmente ao meu coorientador Clécio por me mostrar diversas vezes como fazer a ponte Matemática-Estatística pela qual ele mesmo passou.

Obrigado à professora Lucy e ao professor Filidor, não só por terem aceitado fazer parte da banca, mas também por terem dado contribuições importantes através dos comentários pertinentes e sugestões de correção valiosas.

Aos meus colegas que fizeram parte desses anos de Graduação e Mestrado, tornando-os um pouco menos cansativos. Não citarei nomes para não supervalorizar uns em detrimento de outros, porque todos foram igualmente importantes para minha formação pessoal.

À UFJF pela oportunidade de estudar numa instituição pública de qualidade e à FAPEMIG pelas bolsas, apesar dos frequentes atrasos...

Enfim, minha gratidão a todo o Cosmos que, embora parecesse continuamente estar conspirando contra mim, estava na verdade e sempre estará me desafiando! Acabou! É tetra!!!

“It’s human nature to make a symphony out of the cacophony of events going around us.”
(Morgan Freeman)

RESUMO

O objetivo deste trabalho é apresentar algumas contribuições para a melhoria do processo de estimação por máxima verossimilhança via algoritmo EM em modelos misturas assimétricas com regressão, além de realizar neles a análise de influência local e global. Essas contribuições, em geral de natureza computacional, visam à resolução de problemas comuns na modelagem estatística de maneira mais eficiente. Dentre elas está a substituição de métodos utilizados nas versões dos algoritmos GEM por outras que reduzem o problema aproximadamente a um algoritmo EM clássico nos principais exemplos das distribuições misturas de escala assimétricas de normais. Após a execução do processo de estimação, discutiremos ainda as principais técnicas existentes para o diagnóstico de pontos influentes com as adaptações necessárias aos modelos em foco. Desejamos com tal abordagem acrescentar ao tratamento dessa classe de modelos estatísticos a análise de regressão nas distribuições mais recentes na literatura. Também esperamos abrir caminho para o uso de técnicas similares em outras classes de modelos.

Palavras-chave: Modelos misturas assimétricas. Regressão linear multivariada. Estimação por máxima verossimilhança. Algoritmo EM. Influência global e local.

ABSTRACT

The objective of this work is to present some contributions to improvement the process of maximum likelihood estimation via the EM algorithm in skew mixtures models with regression, as well as to execute in them the global and local influence analysis. These contributions, usually with computational nature, aim to solving common problems in statistical modeling more efficiently. Among them is the replacement of used methods in the versions of the GEM algorithm by other techniques that reduce the problem approximately to a classic EM algorithm in the main examples of skew scale mixtures of normals distributions. After performing the estimation process, we will also discuss the main existing techniques for the diagnosis of influential points with the necessaries adaptations to the models in focus. We wish with this approach to add for the treatment of this statistical model class the regression analysis in the most recent distributions in the literature. We too hope to paving the way for use of similar techniques in other models classes.

Key-words: Skew mixtures models. Multivariate linear regression. Maximum Likelihood Estimation. EM algorithm. Global and local influence.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SÍMBOLOS

Θ	Espaço paramétrico
$\mathbb{R}^{m\times n}$	Conjunto das matrizes $m \times n$ com entradas reais
$\mathbf{1}$	Vetor de $\mathbb{R}^n$ com todas as coordenadas iguais a 1
$\mathbf{I}_p$	Matriz identidade de ordem p
$f(\,\cdot\,;\boldsymbol{\theta})$	Função densidade de probabilidade (f.d.p.) dependente de um vetor de parâmetros $\boldsymbol{\theta} \in \Theta$
$\phi(\,\cdot\,)$	Função densidade de probabilidade da normal padrão univariada
$\Phi(\,\cdot\,)$	Função de distribuição acumulada (f.d.a.) da normal padrão univariada
$S_{\mathbf{y}}$	Desvio padrão amostral do vetor de dados $\mathbf{y}$
$g_{\mathbf{y}}$	Coeficiente de assimetria amostral do vetor de dados $\mathbf{y}$
$\Gamma(\,\cdot\,)$	Função gama
$\Psi(\,\cdot\,)$	Função digama
$\Psi_1(\,\cdot\,)$	Função trigama
$TD_{(a,b)}(\boldsymbol{\theta})$	Distribuição da var. aleatória $D(\boldsymbol{\theta})$ truncada no intervalo (a, b)
$\mathbb{I}_X$	Função indicadora do conjunto X
$\lvert\mathbf{A}\rvert$	$\mathbf{A} \in \mathbb{R}^{m\times n} \Rightarrow \lvert\mathbf{A}\rvert = \det(\mathbf{A})$
$\mathbf{c}_j$	Vetor canônico com a j-ésima coordenada igual a 1

SUMÁRIO

1 INTRODUÇÃO

Em diversas áreas do conhecimento, ainda é bastante comum recorrer à distribuição normal (gaussiana) para modelar dados. Tal fato decorre em especial da facilidade de estimação dos seus parâmetros, inclusive pelo uso do tradicional método dos mínimos quadrados nos modelos de regressão. Igualmente há técnicas para análise de resíduos e diagnóstico de influência com resultados fechados ou consagrados na literatura que tornam os procedimentos envolvidos bem simples.

Porém, a distribuição normal ajusta mal dados com assimetria e caudas pesadas (excesso de probabilidade longe da média). Essas ocorrências são frequentes, tendo ficado famoso o conjunto de dados dos atletas australianos de Cook & Weisberg (1994) por ser uma das primeiras bases de dados usadas como aplicação para distribuições mais flexíveis com pelo menos uma das características mencionadas.

De modo independente, foram criadas em Andrews & Mallows (1974) as *distribuições misturas de escala de normais* com caudas mais pesadas que a normal e em Azzalini (1985) a *distribuição normal assimétrica*, interpretando a assimetria através de um parâmetro que reduzido a 0 equivaleria à normalidade. As duas distribuições surgiram na análise univariada, mas foram adaptadas à multivariada em Lange & Sinsheimer (1993) e Azzalini & Dalla-Valle (1996), respectivamente.

Com o tempo, concebeu-se em várias publicações a combinação dos dois tipos de distribuições em dois sentidos: a das *distribuições misturas de escala de normais assimétricas* e a das *distribuições misturas de escala assimétricas de normais*. Dentre essas publicações, orientaram este trabalho nas duas linhas Zeller, Lachos & Vilca-Labra (2009) e Ferreira, Lachos & Bolfarine (2016), respectivamente.

No Capítulo 2, estabelecemos conceitos e resultados sobre as distribuições envolvidas e a modelagem em si. No Capítulo 3, procedemos a estimação paramétrica dos modelos enfatizando novas técnicas e trabalhamos com um conjunto de dados dos municípios mineiros para ilustrar sua aplicação. No Capítulo 4, adaptamos os métodos clássicos do diagnóstico de influência de Zhu *et al.* (2001) e Zhu & Lee (2001) às distribuições em foco. Por fim, no Capítulo 5, apresentamos um balanço geral dos resultados e destacamos diretrizes para trabalhos futuros.

2 PRELIMINARES

Apresentaremos brevemente neste capítulo conceitos fundamentais recorrentes ao longo do texto. Para uma melhor compreensão do que será apresentado neste trabalho, recomenda-se ver Louredo (2016) para conhecer alguns resultados e as noções gerais acerca de modelos probabilísticos em geral e de regressão linear sob um ponto de vista mais básico.

2.1 ESTIMAÇÃO POR MÁXIMA VEROSSIMILHANÇA (EMV)

A ideia do chamado método de estimação por máxima verossimilhança já vinha sendo formulada desde meados do século XVIII, mas seus principais resultados foram apresentados na obra "The Mathematical Foundations of Theoretical Statistics" do estatístico inglês **Ronald Aylmer Fisher (1890-1962)** em 1922 conforme Stigler (2007). De maneira resumida, vamos descrever o método, que será a técnica básica considerada aqui para a estimação paramétrica.

Definição 2.1.1. Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ um vetor cujas coordenadas constituem uma amostra aleatória com função densidade de probabilidade (f.d.p.) f dependente do vetor de parâmetros $\boldsymbol{\theta} \in \Theta$. Supondo que a amostra $Y_1 = y_1, \ldots, Y_n = y_n$ foi coletada, definimos a *função log-verossimilhança* ℓ por

$$\ell(\boldsymbol{\theta}) = \ln f_{\mathbf{Y}}(y_1, \ldots, y_n; \boldsymbol{\theta}) = \ln \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\theta}). \tag{2.1}$$

Se $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ com $k < n$, então a(s) *estimativa(s) de máxima verossimilhança* é(são) o(s) valor(es) $\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1(y_1, \ldots, y_n), \ldots, \widehat{\theta}_k(y_1, \ldots, y_n)\right)$ que maximiza(m) a função ℓ, isto é, $\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta})$.

Os estimadores de máxima verossimilhança apresentam, sob certas condições, uma série de propriedades assintóticas como consistência e eficiência que são responsáveis por justificar seu uso frente a outros estimadores e podem ser encontradas em Ritter (2015) e Pawitan (2001). O que faremos na sequência é delinear consequências práticas de tais propriedades sem entrar em muitos pormenores.

Na prática, a consistência e a eficiência assintótica são verificadas computacionalmente por meio de medidas de viés e desvios para os conjuntos de dados simulados com diferentes tamanhos como faremos no Capítulo 3. Listamos adiante todas as medidas que usaremos para proceder tal verificação ao longo do texto, onde m é o número de simulações realizadas para a estimação de um vetor de parâmetros previamente estipulado $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ com coordenadas não nulas para uma geração pseudo-aleatória de dados para o modelo.

1. Medidas de viés puro
   - Viés médio: $\text{VME}_i = \dfrac{1}{m}\sum\limits_{j=1}^{m}(\widehat{\theta}_{ij} - \theta_i)$ para cada $i = 1, \ldots, k$;
   - Viés mediano: $\text{VMD}_i = \underset{1\leq j\leq m}{\text{med}}\,(\widehat{\theta}_{ij} - \theta_i)$ para cada $i = 1, \ldots, k$.
2. Desvios puros
   - Erro quadrático médio: $\text{EQM}_i = \sqrt{\dfrac{1}{m}\sum\limits_{j=1}^{m}(\widehat{\theta}_{ij} - \theta_i)^2}$ para cada $i = 1, \ldots, k$;
   - Desvio absoluto mediano: $\text{DAM}_i = \underset{1\leq j\leq m}{\text{med}}\left|\widehat{\theta}_{ij} - \underset{1\leq l\leq m}{\text{med}}\,\theta_{il}\right|$ para cada $i = 1, \ldots, k$.
3. Medidas de viés relativo
   - Viés médio relativo: $\text{VMER}_i = \dfrac{1}{m}\sum\limits_{j=1}^{m}\left(\dfrac{\widehat{\theta}_{ij} - \theta_i}{\theta_i}\right)$ para cada $i = 1, \ldots, k$;
   - Viés mediano relativo: $\text{VMDR}_i = \underset{1\leq j\leq m}{\text{med}}\left(\dfrac{\widehat{\theta}_{ij} - \theta_i}{\theta_i}\right)$ para cada $i = 1, \ldots, k$.
4. Desvios relativos
   - Erro quadrático médio relativo: $\text{EQMR}_i = \sqrt{\dfrac{1}{m}\sum\limits_{j=1}^{m}\left(\dfrac{\widehat{\theta}_{ij} - \theta_i}{\theta_i}\right)^2}$ para cada $i = 1, \ldots, k$;
   - Desvio absoluto mediano relativo: $\text{DAMR}_i = \underset{1\leq j\leq m}{\text{med}}\left|\dfrac{\widehat{\theta}_{ij}}{\theta_i} - \underset{1\leq l\leq m}{\text{med}}\,\dfrac{\theta_{il}}{\theta_i}\right|$ para cada $i = 1, \ldots, k$.

Mais detalhes acerca da definição, propriedades e aplicações dessas medidas podem ser encontrados em Bussab (2012), Cordeiro (1999), Karlsson & Laitila

(2014) e Jamalizadeh & Lin (2016). Outra medida bastante comum, que também utilizaremos aqui para avaliar consistência, é o erro padrão de cada parâmetro numérico. Ele é definido em termos da matriz de informação de Fisher observada do modelo e corresponde à raiz quadrada de cada elemento da diagonal da inversa da referida matriz. Ver Pawitan (2001).

2.2 ALGORITMO EM

O *Algoritmo EM*, onde a sigla EM significa Expectation-Maximization (Esperança-Maximização), foi desenvolvido na década de 1970 como uma alternativa aos métodos tradicionais de otimização normalmente utilizados na EMV, particularmente aqueles dos tipos Newton ou Quasi-Newton. Destacou-se nesse contexto a publicação do primeiro artigo completo sobre o assunto: Dempster, Leird & Rubin (1977).

De modo geral, o EM se baseia no tratamento dos dados observados como "incompletos". Notadamente casos em que realmente faltam observações podem ter a EMV feita a partir do Algoritmo EM conforme se vê em Pawitan (2001).

No entanto, esse algoritmo vem sendo amplamente usado em casos nos quais a função log-verossimilhança é considerada complicada demais para os algoritmos convencionais. Veremos posteriormente várias dessas situações, mas por agora nos limitaremos a estabelecer notações gerais utilizadas no EM.

Denotemos o chamado *vetor de dados completos* por $\mathbf{y}_C = (\mathbf{y}, \mathbf{w})$, onde $\mathbf{y} \in \mathscr{Y} \subset \mathbb{R}^{p\times n} \equiv \mathbb{R}^{pn}$ é um vetor de dados observados para p variáveis de n indivíduos e $\mathbf{w} \in \mathscr{W} \subset \mathbb{R}^{r\times s} \equiv \mathbb{R}^{rs}$ é um vetor de dados faltantes que pode ser visualmente identificado na prática ou hipoteticamente proposto.

Em Mclachlan & Krishnan (2001) é estabelecida a relação probabilística em termos de uma integral entre as f.d.p.s conjuntas $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ e $f_{\mathbf{Y}_C}(\mathbf{y}_C;\boldsymbol{\theta})$, respectivamente, dos dados incompletos e dos dados completos. Segundo o mesmo autor, tal relação pode ser reescrita em termos da f.d.p. condicional $h_{\mathbf{Y}_C|\mathbf{Y}}(\mathbf{y}_C|\mathbf{y};\boldsymbol{\theta})$ da maneira indicada a seguir:

$$f_{\mathbf{Y}_C}(\mathbf{y}_C;\boldsymbol{\theta}) = f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) h_{\mathbf{Y}_C|\mathbf{Y}}(\mathbf{y}_C|\mathbf{y};\boldsymbol{\theta}).$$

Um caso particular que será utilizado neste texto é aquele em que $n = s$ e há uma *relação hierárquica* entre as distribuições dos vetores aleatórios $\mathbf{Y}$ e $\mathbf{W}$ cuja f.d.p. é $g_{\mathbf{W}}(\mathbf{w};\boldsymbol{\theta})$, a qual resulta na seguinte expressão para $f_{\mathbf{Y}_C}(\mathbf{y}_C;\boldsymbol{\theta})$:

$$f_{\mathbf{Y}_C}(\mathbf{y}_C;\boldsymbol{\theta}) = h_{\mathbf{Y}|\mathbf{W}}(\mathbf{y}|\mathbf{w};\boldsymbol{\theta})g_{\mathbf{W}}(\mathbf{w};\boldsymbol{\theta}).$$

Sob a hipótese de independência entre as n coordenadas r-dimensionais de $\mathbf{W}$ e também das n coordenadas p-dimensionais de $\mathbf{Y}|\mathbf{W} = \mathbf{w}$ cujas f.d.p.s serão denotadas por g e h, respectivamente, temos que as coordenadas de $f_{\mathbf{Y}_C}$ são independentes e iguais a, digamos, f_C. Segue daí que podemos escrever a *função log-verossimilhança dos dados completos* ℓ_C na forma

$$\ell_C(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f_C(\mathbf{y}_{Ci};\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln g(\mathbf{w}_i;\boldsymbol{\theta}) + \sum_{i=1}^{n} \ln h(\mathbf{y}_i|\mathbf{w}_i;\boldsymbol{\theta}).$$

Evidentemente, espera-se que uma estimativa de máxima verossimilhança de ℓ_C seja igual a uma de ℓ e que maximizar ℓ_C seja mais simples do que maximizar ℓ. Dito isso, podemos estabelecer os passos básicos do Algoritmo EM e dos GEM na etapa $k+1$ supondo conhecido $\widehat{\boldsymbol{\theta}}^{(k)} \in \Theta$. Ver Mclachlan & Krishnan (2001).

* Passo E: Defina $Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) = E_{\widehat{\boldsymbol{\theta}}^{(k)}}(\ell_C(\boldsymbol{\theta})|\mathbf{Y} = \mathbf{y})$, com o intuito de determinar $\widehat{\varphi\mathbf{w}}_i^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}(\varphi(\mathbf{W}_i)|\mathbf{Y} = \mathbf{y})$ para qualquer função $\varphi(\,\cdot\,)$ do vetor aleatório r-dimensional $\mathbf{W}_i$ que eventualmente apareça na expressão de $\ell_C(\boldsymbol{\theta})$.

* Passo M:

$$\begin{cases} \text{Escolha } \widehat{\boldsymbol{\theta}}^{(k+1)} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) - \text{ algoritmo clássico} \\ \text{Escolha } \widehat{\boldsymbol{\theta}}^{(k+1)} \in \Theta \text{ tal que } Q(\widehat{\boldsymbol{\theta}}^{(k+1)};\widehat{\boldsymbol{\theta}}^{(k)}) \geq Q(\widehat{\boldsymbol{\theta}}^{(k)};\widehat{\boldsymbol{\theta}}^{(k)}) - \text{ GEM} \end{cases}$$

Em Meng & Rubin (1993) e Liu & Rubin (1994) foram desenvolvidas variantes do GEM conhecidas, respectivamente, por ECM e ECME. Resumidamente, ao particionarmos $\boldsymbol{\theta}$ em dois subvetores, substitui-se o Passo M por

1. Algoritmo ECM
$$\begin{cases} \text{Passo CM-1: Escolha } \widehat{\boldsymbol{\theta}_1}^{(k+1)} \in \underset{\theta_1\in\Theta_1}{\operatorname{argmax}}\, Q^{(k)}(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}_2}^{(k)}); \\ \text{Passo CM-2: Escolha } \widehat{\boldsymbol{\theta}_2}^{(k+1)} \in \underset{\theta_2\in\Theta_2}{\operatorname{argmax}}\, Q^{(k)}(\widehat{\boldsymbol{\theta}_1}^{(k+1)}, \boldsymbol{\theta}_2). \end{cases}$$

2. Algoritmo ECME
$$\begin{cases} \text{Passo CM: Escolha } \widehat{\boldsymbol{\theta}_1}^{(k+1)} \in \underset{\theta_1\in\Theta_1}{\operatorname{argmax}}\, Q^{(k)}(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}_2}^{(k)}); \\ \text{Passo CME: Escolha } \widehat{\boldsymbol{\theta}_2}^{(k+1)} \in \underset{\theta_2\in\Theta_2}{\operatorname{argmax}}\, \ell(\widehat{\boldsymbol{\theta}_1}^{(k+1)}, \boldsymbol{\theta}_2). \end{cases}$$

2.3 DISTRIBUIÇÕES MISTURAS NORMAIS ASSIMÉTRICAS

As distribuições de probabilidade presentes nos modelos a serem tratados neste trabalho serão classificadas em dois grupos básicos: misturas de escala e misturas finitas, ambas associadas à distribuição normal assimétrica – tema da primeira subseção. Na segunda e na terceira subseções abordaremos duas formas de ver as misturas de escala relacionadas à normal assimétrica. Por fim, mencionaremos brevemente o conceito de misturas finitas, que será útil em um exemplo específico.

2.3.1 Distribuição Normal Assimétrica

Para começar, vejamos algumas definições e resultados envolvendo a chamada *distribuição normal assimétrica multivariada* – SN na sigla em inglês – proposta originalmente em Azzalini & Dalla-Valle (1996).

Definição 2.3.1. Diz-se que um vetor aleatório p-dimensional $\mathbf{Y}$ segue uma *distribuição normal assimétrica p-variada* e escrevemos $\mathbf{Y} \sim \text{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ quando sua f.d.p. é dada por

$$f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})),\ \mathbf{y} \in \mathbb{R}^p. \tag{2.2}$$

Acima, denotamos por $\phi_p(\ .\ ; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ a f.d.p. da distribuição normal p-variada com vetor de médias $\boldsymbol{\mu}$ (parâmetro de locação) e matriz de covariâncias $\boldsymbol{\Sigma}$ positiva definida (parâmetro de escala) e por $\Phi_1(\ .\ )$ a f.d.a. da distribuição normal padrão univariada, na qual o vetor $\boldsymbol{\lambda}$ no argumento é o parâmetro de assimetria.

Observação: Fazendo $\boldsymbol{\lambda} = \mathbf{0}$ em (2.2), obtemos a f.d.p. da normal p-variada.

Em Lachos (2004), demonstra-se que um vetor aleatório $\mathbf{Y} \sim \text{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ possui a seguinte representação estocástica:

$$\mathbf{Y} \overset{d}{=} \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\left[\boldsymbol{\delta}|T_0| + (\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T)\mathbf{T}_1\right],\ \boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{\sqrt{1 + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}}. \tag{2.3}$$

Na expressão acima, o símbolo $\overset{d}{=}$ quer dizer igualdade em distribuição e $T_0 \sim N_1(0, 1)$ e $\mathbf{T}_1 \sim N_p(\mathbf{0}, \mathbf{I}_p)$ são independentes.

Da representação anterior, obtemos importantes resultados, dentre os quais uma representação hierárquica para o vetor $\mathbf{Y}$. Segundo Johnson, Kotz & Balakrishnan (1994), a v.a. $T = |T_0|$ segue uma distribuição normal truncada no intervalo $[0, +\infty)$, a qual denotaremos por $T \sim \ TN_{[0,+\infty)}(0,1)$. Dessa forma, ao fazermos $T = t$ em (2.3), obtemos $\mathbf{Y}|(T=t) \stackrel{d}{=} (\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}t) + \boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T)^{1/2}\mathbf{T}_1$. Utilizando $\mathbf{T}_1 \sim \ N_p(\mathbf{0}, \mathbf{I}_p)$ e as propriedades da distribuição normal assimétrica, vê-se em Lachos (2004) que

$$\begin{aligned} \mathbf{Y}|T=t &\sim N_p(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}t, \boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T)\boldsymbol{\Sigma}^{1/2}) \\ T &\sim TN_{[0,+\infty)}(0,1) \end{aligned} \tag{2.4}$$

Com a representação hierárquica de $\mathbf{Y}$ e resultados da Teoria da Probabilidade, provaremos a seguinte Proposição:

Proposição 2.3.1. Dado um vetor aleatório $\mathbf{Y} \sim \ \mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, seu valor esperado e sua variância são, respectivamente, dados por

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta} \ \text{ e } \ Var(\mathbf{Y}) = \boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_p - \frac{2}{\pi}\boldsymbol{\delta}\boldsymbol{\delta}^T\right)\boldsymbol{\Sigma}^{1/2}. \tag{2.5}$$

Demonstração:

Para provar o resultado, usaremos as identidades da esperança e da variância condicional que são encontradas em Casella (2002) numa versão para variáveis aleatórias, mas podem ser facilmente estendidas para o caso de vetores aleatórios. Com isso e as propriedades conhecidas da esperança e da variância, a representação dada em (2.5) nos dá que

$$\begin{aligned} E(\mathbf{Y}) &= E\big(E(\mathbf{Y}|T)\big) = E(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}T) = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}E(T); \\ Var(\mathbf{Y}) &= E\big(Var(\mathbf{Y}|T)\big) + Var\big(E(\mathbf{Y}|T)\big) = E\big(\boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T)\boldsymbol{\Sigma}^{1/2}\big) + Var\big(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}T\big) \\ &= \boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T)\boldsymbol{\Sigma}^{1/2} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}Var(T)\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{1/2}. \end{aligned}$$

Resta, pois, obter $E(T)$ e $Var(T)$. Como $T \sim \ TN_{[0,+\infty)}(0,1)$, então a f.d.p. de T é dada por $f_T(t) = \sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}t^2}\mathbb{I}_{[0,+\infty)}$ e daí

$$\begin{aligned} E(T) &= \sqrt{\frac{2}{\pi}}\int_0^{+\infty} te^{-\frac{1}{2}t^2}dt = \sqrt{\frac{2}{\pi}}\lim_{s\to+\infty}\big(-e^{-\frac{1}{2}t^2}\big)\Big|_0^s = \sqrt{\frac{2}{\pi}}\cdot 1 = \sqrt{\frac{2}{\pi}}; \\ E(T^2) &= \sqrt{\frac{2}{\pi}}\int_0^{+\infty} t^2e^{-\frac{1}{2}t^2}dt = \sqrt{\frac{2}{\pi}}\left[\lim_{s\to+\infty}\big(-te^{-\frac{1}{2}t^2}\big)\Big|_0^s + \int_0^{+\infty} e^{-\frac{1}{2}t^2}dt\right] = \sqrt{\frac{2}{\pi}}\left[0 + \sqrt{\frac{\pi}{2}}\right] = 1; \\ Var(T) &= E(T^2) - E(T)^2 = 1 - \frac{2}{\pi}. \end{aligned}$$

Substituindo os últimos resultados nas expressões iniciais, concluímos que

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}E(T) = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta};$$

$$Var(\mathbf{Y}) = \boldsymbol{\Sigma}^{1/2}\left[\mathbf{I}_p - \Big(Var(T) - 1\Big)\boldsymbol{\delta}\boldsymbol{\delta}^T\right]\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_p - \frac{2}{\pi}\boldsymbol{\delta}\boldsymbol{\delta}^T\right)\boldsymbol{\Sigma}^{1/2}.$$

■

2.3.2 Distribuições Misturas de Escala Assimétricas de Normais (SSMN)

Vejamos a definição de mistura de escala retirada de Souza Filho (2012).

Definição 2.3.2. Sejam $\mathbf{Y}$ um vetor aleatório e U uma variável aleatória positiva com função de probabilidade h tal que $P(U \in \varsigma) = 1$, ambos definidos em um mesmo espaço de probabilidade $(\Omega, \mathscr{A}, P)$. Considerando para cada $u \in \varsigma$, a função de probabilidade condicional $g(\mathbf{y}|u)$ de $\mathbf{Y}|U = u$, dizemos que $f(\mathbf{y}) = \int_\varsigma g(\mathbf{y}|u)h(u)du$ é uma *mistura de escala* da família $\{g(\mathbf{y}|u) : u \in \varsigma\}$ com *fator de escala* U e *densidade da mistura* h.

Em Lange & Sinsheimer (1993), foi proposta a definição das *distribuições misturas de escala de normais multivariadas* com vetor de locação $\boldsymbol{\mu}$ e matriz de escala $\boldsymbol{\Sigma}$ como aquelas cuja f.d.p. é dada por

$$f_0(\mathbf{y}) = \int_0^{+\infty} \phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right) h(u; \boldsymbol{\tau})du. \tag{2.6}$$

Remetendo à Definição 2.3.2, vemos na f.d.p. acima que o fator de escala U com função de probabilidade h dependendo de um hiper-parâmetro $\boldsymbol{\tau}$ determina de fato uma mistura de escala da família normal multivariada. Há de se observar que a própria distribuição normal multivariada com esperança $\boldsymbol{\mu}$ e variância $\boldsymbol{\Sigma}$ pode ser vista como caso particular das distribuições em questão tomando h como o delta de Dirac centrado no ponto 1.

O foco deste trabalho serão distribuições misturas de escala assimétricas envolvendo normais multivariadas, dentre as quais as chamadas *misturas de escala assimétricas de normais multivariadas* – SSMN na sigla em inglês. A característica

central desse tipo de mistura é combinar as caudas pesadas típicas das distribuições misturas de escala normais com a assimetria. A forma de definir essas distribuições que veremos foi proposta recentemente em Ferreira, Lachos & Bolfarine (2016).

Definição 2.3.3. Um vetor aleatório $\mathbf{Y}$ p-dimensional segue uma *distribuição mistura de escala assimétrica da família normal* e escreve-se $\mathbf{Y} \sim \text{SSMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ quando sua f.d.p. é

$$f(\mathbf{y}) = 2f_0(\mathbf{y})\Phi_1(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})),\ \mathbf{y} \in \mathbb{R}^p, \tag{2.7}$$

onde f_0 é a f.d.p. dada em (2.6).

Adotando uma técnica similar à utilizada em Lachos (2004), vamos deduzir a representação estocástica de $\mathbf{Y} \sim \text{SSMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$. Para tanto, necessitaremos do próximo lema.

Lema 2.3.1. Se $\mathbf{Y} \sim N_r(\boldsymbol{\zeta}, \boldsymbol{\Omega})$ e $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então

$$\begin{aligned}\phi_r(\mathbf{y}; \boldsymbol{\zeta} + \boldsymbol{\Psi}\mathbf{x}, \boldsymbol{\Omega})\phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \phi_r(\mathbf{y}; \boldsymbol{\zeta} + \boldsymbol{\Psi}\boldsymbol{\mu}, \boldsymbol{\Omega} + \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^T) \\ &\times \phi_p(\mathbf{x}; \boldsymbol{\mu} + \boldsymbol{\Pi}\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\zeta} - \boldsymbol{\Psi}\boldsymbol{\mu}), \boldsymbol{\Pi}),\end{aligned}$$

onde $\boldsymbol{\Pi} = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi})^{-1}$.

Demonstração:

Para provar a igualdade, usaremos os primeiros resultados do Anexo A. Fazendo $\mathbf{z} = \mathbf{y} - \boldsymbol{\zeta} - \boldsymbol{\Psi}\boldsymbol{\mu}$ e $\mathbf{w} = \mathbf{x} - \boldsymbol{\mu}$ na expressão direita da igualdade, temos por definição que

$$\phi_r(\mathbf{y}; \boldsymbol{\zeta} + \boldsymbol{\Psi}\boldsymbol{\mu}, \boldsymbol{\Omega} + \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^T)\phi_p(\mathbf{x}; \boldsymbol{\mu} + \boldsymbol{\Pi}\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\zeta} - \boldsymbol{\Psi}\boldsymbol{\mu}), \boldsymbol{\Pi}) =$$

$$= \frac{1}{(2\pi)^{\frac{r}{2}}\sqrt{|\boldsymbol{\Omega} + \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^T|}} e^{-\frac{1}{2}\mathbf{z}^T(\boldsymbol{\Omega}+\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^T)^{-1}\mathbf{z}} . \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\boldsymbol{\Pi}|}} e^{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\Pi}\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\mathbf{z})^T\boldsymbol{\Pi}^{-1}(\mathbf{w}-\boldsymbol{\Pi}\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\mathbf{z})}.$$

Pelo Teorema A.0.2, temos $|\boldsymbol{\Pi}| = |\boldsymbol{\Omega} + \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^T|^{-1}|\boldsymbol{\Omega}||\boldsymbol{\Sigma}|$ e daí $|\boldsymbol{\Omega} + \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^T||\boldsymbol{\Pi}| = |\boldsymbol{\Sigma}||\boldsymbol{\Omega}|$. Utilizando agora o Teorema A.0.1, obtemos a relação $(\boldsymbol{\Omega} + \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^T)^{-1} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}^T(\boldsymbol{\Sigma} + \boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi})\boldsymbol{\Psi}\boldsymbol{\Omega}^{-1}$ e assim podemos escrever a seguinte relação:

$$\mathbf{z}^T(\boldsymbol{\Omega}+\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^T)^{-1}\mathbf{z}+(\mathbf{w}-\boldsymbol{\Pi}\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Omega}^{-\mathbf{1}}\mathbf{z})^T\boldsymbol{\Pi}^{-1}(\mathbf{w}-\boldsymbol{\Pi}\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Omega}^{-\mathbf{1}}\mathbf{z})=$$

$$=\mathbf{z}^T[\boldsymbol{\Omega}^{-1}-\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}^T(\boldsymbol{\Sigma}+\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi})\boldsymbol{\Psi}\boldsymbol{\Omega}^{-1}]\mathbf{z}+(\mathbf{w}-\boldsymbol{\Pi}\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Omega}^{-\mathbf{1}}\mathbf{z})^T(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi})(\mathbf{w}-\boldsymbol{\Pi}\boldsymbol{\Psi}^{\mathbf{T}}\boldsymbol{\Omega}^{-\mathbf{1}}\mathbf{z})=$$

$$=\mathbf{z}^T\boldsymbol{\Omega}^{-1}\mathbf{z}-2\mathbf{z}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\boldsymbol{\Pi}\boldsymbol{\Sigma}^{-1}\mathbf{w}-2\mathbf{z}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\boldsymbol{\Pi}\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}^T\mathbf{w}+\mathbf{w}^T\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\mathbf{w}+\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}=$$

$$=\mathbf{z}^T\boldsymbol{\Omega}^{-1}\mathbf{z}-2\mathbf{z}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\mathbf{w}+2\mathbf{z}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\boldsymbol{\Pi}\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}^T\mathbf{w}-2\mathbf{z}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\boldsymbol{\Pi}\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}^T\mathbf{w}+\mathbf{w}^T\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\mathbf{w}+\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}=$$

$$=\mathbf{z}^T\boldsymbol{\Omega}^{-1}\mathbf{z}-2\mathbf{z}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\mathbf{w}+\mathbf{w}^T\boldsymbol{\Psi}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\mathbf{w}+\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}=(\mathbf{z}-\boldsymbol{\Psi}\mathbf{w})^T\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\Psi}\mathbf{w})+\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}.$$

Com os resultados acima, o produto inicial de normais é re-expresso por

$$\frac{1}{(2\pi)^{\frac{r}{2}}\sqrt{|\boldsymbol{\Omega}|}}e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\Psi}\mathbf{w})^T\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\Psi}\mathbf{w})}\cdot\frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\boldsymbol{\Sigma}|}}e^{-\frac{1}{2}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}}=$$

$$=\frac{1}{(2\pi)^{\frac{r}{2}}\sqrt{|\boldsymbol{\Omega}|}}e^{-\frac{1}{2}(\mathbf{y}-\zeta-\boldsymbol{\Psi}\mathbf{x})^T\boldsymbol{\Omega}^{-1}(\mathbf{y}-\zeta-\boldsymbol{\Psi}\mathbf{x})}\cdot\frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\boldsymbol{\Sigma}|}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}=\phi_r(\mathbf{y};\zeta+\boldsymbol{\Psi}\mathbf{x},\boldsymbol{\Omega})\phi_p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}).$$

■

Teorema 2.3.1. Dado $\mathbf{Y}\sim$ SSMN$_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\boldsymbol{\tau})$, temos a seguinte *representação estocástica*

$$\mathbf{Y}\overset{d}{=}\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\Big(U\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\left[\boldsymbol{\lambda}\left(\frac{U+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{U}\right)^{1/2}|T_0|+\Big(U\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{1/2}\mathbf{T}_1\right],\ (2.8)$$

onde $T_0\sim\ N_1(0,1),\mathbf{T}_1\sim\ N_p(\mathbf{0},\mathbf{I}_p)$ e $U\sim\ H(\ \cdot\ ;\boldsymbol{\tau})$ são independentes. Além disso, temos a seguinte *representação hierárquica* para $\mathbf{Y}$:

$$\begin{aligned}\mathbf{Y}|T=t,U=u\ &\sim\ N_p\left(\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}(u\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T)^{-1}\boldsymbol{\lambda}t,\boldsymbol{\Sigma}^{1/2}(u\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T)^{-1}\boldsymbol{\Sigma}^{1/2}\right);\\ T|U=u\ &\sim\ TN_{[0,+\infty)}\left(0,\frac{u+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}\right);\\ U\ &\sim\ H(\ .\ ;\boldsymbol{\tau}).\end{aligned}\qquad(2.9)$$

Demonstração: Defina a variável aleatória $T=\left(\frac{U+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{U}\right)^{1/2}|T_0|$ e o vetor aleatório $\mathbf{W}=\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}(U\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T)^{-1}\left[\boldsymbol{\lambda}T+\Big(U\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{1/2}\mathbf{T}_1\right]$. Dessa

maneira, procedendo os condicionamentos $T = t$ e $U = u$ em $\mathbf{W}$, vemos que a distribuição do vetor aleatório condicionado $\mathbf{W}\Big|T = t, U = u$ é

$$\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\Big(u\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\boldsymbol{\lambda}t + \boldsymbol{\Sigma}^{1/2}\Big(u\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1/2}\mathbf{T}_1.$$

Sendo $\mathbf{T}_1 \sim\ N_p(\mathbf{0}, \mathbf{I}_p)$, por resultado presente em Louredo (2016) o vetor condicionado possui distribuição normal p-variada com esperança $\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\Big(u\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\boldsymbol{\lambda}t$ e variância $\boldsymbol{\Sigma}^{1/2}\Big(u\mathbf{I}_p-\boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\boldsymbol{\Sigma}^{1/2}$, ou seja, concluímos que a distribuição condicional é $\mathbf{W}|T = t, U = u \sim\ N_p\left(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\Big(u\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\lambda t, \boldsymbol{\Sigma}^{1/2}\Big(uI_p - \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\right)$.

Além disso, ao condicionarmos a variável T em $U = u$, obtemos a distribuição de $\left(\dfrac{u + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}\right)^{1/2} |T_0|$. Como $|T_0| \sim\ TN_{[0,+\infty)}(0, 1)$, temos segundo Johnson, Kotz & Balakrishnan (1994) que $\left(\dfrac{u + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}\right)^{1/2} |T_0| \sim\ TN_{[0,+\infty)}\left(0, \dfrac{u + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}\right)$. Por fim, a suposição de que o fator de escala U segue uma distribuição dependente unicamente do hiper-parâmetro $\boldsymbol{\tau}$ determina a representação hierárquica de $\mathbf{W}$.

Resta, pois, mostrar que a distribuição de $\mathbf{W}$ é a mesma de $\mathbf{Y}$. Com efeito, por resultados da Teoria da Probabilidade, temos que a f.d.p. conjunta de $\mathbf{W}, T$ e U é o produto das f.d.p.s da representação hierárquica anterior e daí a f.d.p. de $\mathbf{W}$ pode ser obtida como a f.d.p. marginal integrando nas outras variáveis como segue:

$$\begin{aligned}
f_{\mathbf{W}}(\mathbf{w}) &= \int_0^{+\infty}\int_0^{+\infty} \phi_p\left(\mathbf{y}; \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\Big(u\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\boldsymbol{\lambda}t, \boldsymbol{\Sigma}^{1/2}\Big(uI_p - \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\right) \cdot 2\phi_1\left(t; 0, \frac{u + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}\right) h(u; \boldsymbol{\tau})dtdu \\
&= 2\int_0^{+\infty}\int_0^{+\infty} \phi_1\left(t; 0, 1 + \frac{\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}\right) \phi_p\left(\mathbf{y}; \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\Big(u\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\boldsymbol{\lambda}t, \boldsymbol{\Sigma}^{1/2}\Big(u\mathbf{I}_p - \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\boldsymbol{\Sigma}^{1/2}\right) h(u; \boldsymbol{\tau})dtdu \\
&= 2\int_0^{+\infty}\int_0^{+\infty} \phi_1\left(t; -\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu} + \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{y}, 1\right) \phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right) h(u; \boldsymbol{\tau})dtdu \\
&= 2\int_0^{+\infty}\int_0^{+\infty} \phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right) \phi_1\left(t; \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}), 1\right) h(u; \boldsymbol{\tau})dtdu \\
&= 2\int_0^{+\infty}\int_{-\infty}^{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})} \phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right) \phi_1(s)h(u; \boldsymbol{\tau})dsdu = 2f_0(\mathbf{y})\Phi_1\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})\right) = f(\mathbf{y}).
\end{aligned}$$

A igualdade da segunda para a terceira linha decorre do Lema 2.3.1. ∎

Assim como no caso da normal assimétrica, o condicionamento adequado de $\mathbf{Y} \sim\ \text{SSMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ permite deduzir seu valor esperado e sua variância, o que faremos na Proposição seguinte.

Proposição 2.3.2. Dado um vetor aleatório $\mathbf{Y} \sim$ SSMN$_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$, seu valor esperado e sua variância são, respectivamente,

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\varpi\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda} \text{ e } Var(\mathbf{Y}) = \boldsymbol{\Sigma}^{1/2}\left(\varrho\mathbf{I}_p - \frac{2}{\pi}\varpi^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right)\boldsymbol{\Sigma}^{1/2}, \quad (2.10)$$

onde $\varrho = E(U^{-1})$ e $\varpi = E\{[U(U + \boldsymbol{\lambda}^T\boldsymbol{\lambda})]^{-1/2}\}$.

Demonstração:

Aplicando aqui a mesma ideia utilizada na Proposição 2.3.1, procederemos o condicionamento do vetor aleatório $\mathbf{Y}$ na variável U a fim de que, pelo mesmo resultado mencionado naquela Proposição, possamos obter

$$E(\mathbf{Y}) = E\Big(E(\mathbf{Y}|U)\Big) \text{ e } Var(\mathbf{Y}) = Var\Big(E(\mathbf{Y}|U)\Big) + E\Big(Var(\mathbf{Y}|U)\Big).$$

Vamos deduzir agora a distribuição de $\mathbf{Y}|U = u$. Ora, das relações entre densidades conjuntas, marginais e condicionais devemos ter

$$f(\mathbf{y}) = \int_0^{+\infty} \bar{f}(\mathbf{y}, u)du = \int_0^{+\infty} \widetilde{f}(\mathbf{y}|u)h(u;\boldsymbol{\tau})du.$$

Ao compararmos a expressão anterior com (2.7), vemos que a f.d.p. condicional de $\mathbf{Y}$ dado $U = u$ é

$$\widetilde{f}(\mathbf{y}|u) = 2\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{u}\right)\Phi_1\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right) = 2\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{u}\right)\Phi_1\left(\left(\frac{\boldsymbol{\lambda}}{\sqrt{u}}\right)^T\left(\frac{\boldsymbol{\Sigma}}{u}\right)^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right).$$

Pela expressão acima, segue da Definição 2.2 que $\mathbf{Y}|U = u \sim$ SN$_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}, \frac{\boldsymbol{\lambda}}{\sqrt{u}}\right)$. Assim, temos pela Proposição 2.3.1 os seguintes resultados:

$$E(\mathbf{Y}|U=u) = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\cdot\frac{\left(\frac{\boldsymbol{\Sigma}}{u}\right)^{\frac{1}{2}}\frac{\boldsymbol{\lambda}}{\sqrt{u}}}{\sqrt{1+\frac{\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}}} = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\left[u(u+\boldsymbol{\lambda}^T\boldsymbol{\lambda})\right]^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda};$$

$$Var(\mathbf{Y}|U=u) = \left(\frac{\boldsymbol{\Sigma}}{u}\right)^{\frac{1}{2}}\left(\mathbf{I}_p - \frac{2}{\pi}\frac{\frac{\boldsymbol{\lambda}\boldsymbol{\lambda}^T}{u}}{1+\frac{\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}}\right)\left(\frac{\boldsymbol{\Sigma}}{u}\right)^{\frac{1}{2}} = \frac{1}{u}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\mathbf{I}_p - \frac{2}{\pi}\frac{\boldsymbol{\lambda}\boldsymbol{\lambda}^T}{u+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}\right)\boldsymbol{\Sigma}^{\frac{1}{2}}.$$

Dessa forma, usando o resultado citado no início da demonstração e definindo $\Upsilon = \left[U(U + \boldsymbol{\lambda}^T\boldsymbol{\lambda})\right]^{-\frac{1}{2}}$, concluímos de propriedades da esperança e da variância que

$$\begin{aligned} E(\mathbf{Y}) &= E\left(\boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\Upsilon\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda}\right) = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}E(\Upsilon)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda} = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\varpi\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda}; \\ Var(\mathbf{Y}) &= Var\left(\boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\Upsilon\boldsymbol{\Sigma}\boldsymbol{\lambda}\right) + E\left(U^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\mathbf{I}_p - \frac{2}{\pi}\frac{\boldsymbol{\lambda}\boldsymbol{\lambda}^T}{U + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}\right)\boldsymbol{\Sigma}^{\frac{1}{2}}\right) \\ &= \frac{2}{\pi}Var(\Upsilon)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{\frac{1}{2}} + E(U^{-1})\boldsymbol{\Sigma} - \frac{2}{\pi}E(\Upsilon^2)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &= \boldsymbol{\Sigma}^{\frac{1}{2}}\left[E(U^{-1})\mathbf{I}_p - \frac{2}{\pi}E(\Upsilon)^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right]\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{\Sigma}^{\frac{1}{2}}\left(\varrho\mathbf{I}_p - \frac{2}{\pi}\varpi^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right)\boldsymbol{\Sigma}^{\frac{1}{2}}. \end{aligned}$$

■

Vejamos mais um resultado útil para o processo de estimação dos parâmetros a ser discutido no próximo capítulo válido nesta forma de mistura de escala.

Proposição 2.3.3. Dado um vetor aleatório $\mathbf{Y} \sim \text{ SSMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$, temos a distribuição condicional $T|\mathbf{Y} = \mathbf{y} \sim \ TN_{[0,+\infty)}\Big(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}), 1\Big)$.

Demonstração:

Observe que na parte final da demonstração do Teorema 2.3.1, obtivemos após algumas manipulações algébricas que a f.d.p. conjunta de $\mathbf{Y}, T$ e U é $\check{f}(\mathbf{y}, t, u) = 2\phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right)\phi_1(t; \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}), 1)h(u; \boldsymbol{\tau})$, donde segue que podemos encontrar a f.d.p. conjunta de $\mathbf{Y}$ e T integrando a f.d.p. anterior na variável U. Isso nos dá

$$\bar{f}(\mathbf{y}, t) = 2\int_0^{+\infty}\check{f}(\mathbf{y}, t, u)du = 2f_0(\mathbf{y})\phi_1\Big(t; \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}), 1\Big).$$

Dessa forma, a distribuição condicional de T dado $\mathbf{Y} = \mathbf{y}$ possui a f.d.p.

$$\begin{aligned} \widetilde{f}(t|\mathbf{y}) &= \frac{\bar{f}(t, \mathbf{y})}{f(\mathbf{y})} = \frac{2f_0(\mathbf{y})\phi_1\Big(t; \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}), 1\Big)}{2f_0(\mathbf{y})\Phi_1\Big(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu})\Big)} \\ &= \frac{\phi_1\Big(t; \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}), 1\Big)}{\Phi_1\Big(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu})\Big)} = \frac{\phi_1\Big(t - \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu})\Big)}{1 - \Phi_1\Big(-\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu})\Big)}. \end{aligned}$$

De acordo com a definição de distribuição normal truncada presente em Johnson, Kotz & Balakrishnan (1994), $T|\mathbf{Y} = \mathbf{y} \sim \ TN_{[0,+\infty)}\Big(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}), 1\Big)$. ■

2.3.3 Distribuições Misturas de Escala de Normais Assimétricas (SMSN)

Consideraremos agora outro modo de definir misturas de escala assimétricas que envolve distribuições normais, proposto em Branco, Dey & Sahu (2003), a fim de discutir no início do Capítulo 3 uma pequena modificação no processo de estimação dos parâmetros que só pode ser feita (no contexto univariado) para essa forma de mistura. Tal família de distribuições constitui as denominadas *misturas de escala de normais assimétricas* – SMSN na sigla em inglês.

Definição 2.3.4. Dizemos que um vetor aleatório $\mathbf{Y}$ p-dimensional possui uma *distribuição mistura de escala da família normal assimétrica* e escrevemos a notação $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ quando este é da forma $\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}$, onde U é uma variável aleatória positiva com função de probabilidade $h(u; \boldsymbol{\tau})$ independente de $\mathbf{Z} \sim \text{SN}_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. Dessa forma, a f.d.p. de $\mathbf{Y}$ é dada por

$$f(\mathbf{y}) = 2\int_0^{+\infty} \phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right) \Phi_1\left(\sqrt{u}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})\right) h(u; \boldsymbol{\tau})du, \ \mathbf{y} \in \mathbb{R}^p. \quad (2.11)$$

Segue diretamente da Definição 2.3.4 e da representação dada em (2.3) para $\mathbf{Z} \sim \text{SN}_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ que o vetor $\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}$ anterior possui a representação estocástica expressa adiante:

$$\mathbf{Y} \overset{d}{=} \boldsymbol{\mu} + U^{-1/2}\boldsymbol{\Sigma}^{1/2}\left[\boldsymbol{\delta}|T_0| + (\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T)^{1/2}\mathbf{T}_1\right], \ \boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{\sqrt{1 + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}}. \quad (2.12)$$

Pela representação estocástica anterior, onde $T_0 \sim N_1(0, 1), \mathbf{T}_1 \sim N_p(\mathbf{0}, \mathbf{I}_p)$ e $U \sim H(\cdot\,; \boldsymbol{\tau})$ são independentes, podemos verificar a Proposição seguinte para esse tipo de mistura de escala.

Proposição 2.3.4. Dado um vetor aleatório $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$, seu valor esperado e sua variância são, respectivamente, dados por

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\psi\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta} \ \text{ e } \ Var(\mathbf{Y}) = \boldsymbol{\Sigma}^{1/2}\left(\xi\mathbf{I}_p - \frac{2}{\pi}\psi^2\boldsymbol{\delta}\boldsymbol{\delta}^T\right)\boldsymbol{\Sigma}^{1/2}, \quad (2.13)$$

onde $\psi = E(U^{-1/2})$ e $\xi = E(U^{-1})$.

Demonstração:

Lembrando que $\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}$ com $\mathbf{Z} \sim SN_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ e U é independente de $\mathbf{Z}$, a esperança de $\mathbf{Y}$ em termos da esperança de $\mathbf{Z}$ é expressa por

$$E(\mathbf{Y}) = \boldsymbol{\mu} + E(U^{-1/2})E(\mathbf{Z}).$$

Pela Proposição 2.3.1, temos $E(\mathbf{Z}) = \sqrt{\frac{2}{\pi}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ e daí, pondo $\psi = E(U^{-1/2})$, concluímos que $E(\mathbf{Y}) = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\psi\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$.

Para determinar $Var(\mathbf{Y})$, calculemos separadamente $E(\mathbf{Y}\mathbf{Y}^T)$ e $E(\mathbf{Y})E(\mathbf{Y})^T$. Utilizando a expressão de $\mathbf{Y}$ mencionada inicialmente, obtemos

$$E(\mathbf{Y}\mathbf{Y}^T) = \boldsymbol{\mu}\boldsymbol{\mu}^T + E(U^{-1/2})\left[\boldsymbol{\mu}E(\mathbf{Z})^T + E(\mathbf{Z})\boldsymbol{\mu}^T\right] + E(U^{-1})E(\mathbf{Z}\mathbf{Z}^T);$$

$$E(\mathbf{Y})E(\mathbf{Y})^T = \boldsymbol{\mu}\boldsymbol{\mu}^T + E(U^{-1/2})\left[\boldsymbol{\mu}E(\mathbf{Z})^T + E(\mathbf{Z})\boldsymbol{\mu}^T\right] + E(U^{-1/2})^2E(\mathbf{Z})E(\mathbf{Z})^T.$$

Com os resultados acima e considerando $\xi = E(U^{-1})$, vemos que

$$Var(\mathbf{Y}) = E(\mathbf{Y}\mathbf{Y}^T) - E(\mathbf{Y})E(\mathbf{Y})^T = \xi E(\mathbf{Z}\mathbf{Z}^T) - \psi^2 E(\mathbf{Z})E(\mathbf{Z})^T$$

Como $Var(\mathbf{Z}) = \boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_p - \frac{2}{\pi}\boldsymbol{\delta}\boldsymbol{\delta}^T\right)\boldsymbol{\Sigma}^{1/2}$ e $E(\mathbf{Z})E(\mathbf{Z})^T = \frac{2}{\pi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{1/2}$, então $E(\mathbf{Z}\mathbf{Z}^T) = Var(\mathbf{Z}) + E(\mathbf{Z})E(\mathbf{Z})^T = \boldsymbol{\Sigma}$ e assim $Var(\mathbf{Y}) = \xi\boldsymbol{\Sigma} - \frac{2}{\pi}\psi^2\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}\left(\xi\mathbf{I}_p - \frac{2}{\pi}\psi^2\boldsymbol{\delta}\boldsymbol{\delta}^T\right)\boldsymbol{\Sigma}^{1/2}$.

■

A Proposição seguinte mostra uma representação hierárquica de um vetor aleatório $\mathbf{Y}$ com a distribuição mistura de escala da família normal assimétrica.

Proposição 2.3.5. Se $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$, então

$$\begin{aligned} \mathbf{Y}|T = t, U = u &\sim N_p\left(\boldsymbol{\mu} + \boldsymbol{\Delta}t, \frac{\boldsymbol{\Gamma}}{u}\right); \\ T|U = u &\sim TN_{[0,+\infty)}\left(0, \frac{1}{u}\right); \\ U &\sim H(\,.\,;\boldsymbol{\tau}). \end{aligned} \tag{2.14}$$

Na representação acima, $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ e $\boldsymbol{\Gamma} = \boldsymbol{\Sigma} - \boldsymbol{\Delta}\boldsymbol{\Delta}^T$.

Demonstração:

Inicialmente defina a variável aleatória $T = U^{-1/2}|T_0|$, a qual é não negativa pois $|T_0|$ e U também o são. Dessa forma, fazendo os condicionamentos $T = t$ e $U = u$ no segundo membro da expressão (2.12), vemos que a distribuição de $\mathbf{Y}\Big|T = t, U = u$ é igual à de

$$\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}t + u^{-1/2}\boldsymbol{\Sigma}^{1/2}\left(I_p - \boldsymbol{\delta}\boldsymbol{\delta}^T\right)^{1/2}\mathbf{T}_1.$$

Como $\mathbf{T}_1 \sim\ N_p(\mathbf{0}, \mathbf{I}_p)$, temos pelo resultado presente em Louredo (2016) p. 25 que o vetor aleatório acima segue uma distribuição normal p-variada com esperança $\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}t$ e variância $\frac{1}{u}\boldsymbol{\Sigma}^{1/2}\Big(\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T\Big)\boldsymbol{\Sigma}^{1/2}$. Em outras palavras, escrevendo $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}$ e $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{1/2}\Big(\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^T\Big)\boldsymbol{\Sigma}^{1/2}$, vemos que $\mathbf{Y}|T = t, U = u \sim\ N_p\left(\boldsymbol{\mu} + \boldsymbol{\Delta}t, \frac{\boldsymbol{\Gamma}}{u}\right)$.

Por outro lado, condicionando a variável T em $U = u$, obtemos distribuição igual à de $u^{-1/2}|T_0|$. Sendo $|T_0| \sim\ TN_{[0,+\infty)}(0, 1)$, segue de uma propriedade da distribuição normal truncada que pode ser encontrada em Johnson, Kotz & Balakrishnan (1994) que $u^{-1/2}|T_0| \sim\ TN_{[0,+\infty)}\left(0, \frac{1}{u}\right)$. Por fim, a suposição de que o fator de escala U segue uma distribuição dependente unicamente do hiper-parâmetro $\boldsymbol{\tau}$ completa a descrição da representação hierárquica. ∎

Outros resultados úteis para este tipo de mistura são apresentados a seguir.

Proposição 2.3.6. Se $\mathbf{Y} \sim\ \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$, temos então a distribuição condicional $T|\mathbf{Y} = \mathbf{y}, U = u \sim\ TN_{[0,+\infty)}\left(\mu_T, \frac{\sigma_T}{\sqrt{u}}\right)$, onde temos

$$\mu_T = \frac{\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}(\mathbf{y} - \boldsymbol{\mu})}{1 + \boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}} \quad \text{e} \quad \sigma_T = \sqrt{\frac{1}{1 + \boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}}}.$$

Demonstração:

Da representação hierárquica do vetor $\mathbf{Y}$ dada em (2.14), considerando g a f.d.p. de $\mathbf{Y}|T = t, U = u$ e $\widetilde{f}$ a f.d.p. de $T|U = u$, segue que a f.d.p. conjunta de $\mathbf{Y}, T$ e U é dada por

$$\check{f}(\mathbf{y}, t, u) = g(\mathbf{y}|t, u)\widetilde{f}(t|u)h(u) = 2\phi_p\left(\mathbf{y}; \boldsymbol{\mu} + \boldsymbol{\Delta}t, \frac{\boldsymbol{\Gamma}}{u}\right)\phi_1\left(t; 0, \frac{1}{u}\right)h(u; \boldsymbol{\tau}).$$

Utilizando o Lema 2.3.1 e algumas manipulações algébricas, podemos reescrever a expressão acima da seguinte forma:

$$\check{f}(\mathbf{y},t,u) = 2\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{u}\right)\phi_1\left(t;\frac{\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}},\frac{1}{u(1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta})}\right)h(u;\boldsymbol{\tau}) =$$

$$= f(\mathbf{y})\cdot\frac{2\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{u}\right)\Phi_1\big(\sqrt{u}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\big)h(u;\boldsymbol{\tau})}{f(\mathbf{y})}\cdot\frac{\phi_1\left(t;\frac{\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}},\frac{1}{u(1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta})}\right)}{\Phi_1\big(\sqrt{u}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\big)}.$$

Lembrando que é possível escrever a f.d.p. conjunta de $\mathbf{Y},T$ e U como produto das densidades de $\mathbf{Y},U|\mathbf{Y}=\mathbf{y}$, $T|\mathbf{Y}=\mathbf{y},U=u$ e observando que o fator do meio no último produto acima é a f.d.p. de $U|\mathbf{Y}=\mathbf{y}$ (razão da f.d.p. conjunta de $\mathbf{Y}$ e U pela f.d.p. marginal de $\mathbf{Y}$), devemos ter obrigatoriamente o fator da direita igual à f.d.p. de $T|\mathbf{Y}=\mathbf{y},U=u$. Para caracterizar essa última densidade, segue das relações $\boldsymbol{\Delta}=\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}=\dfrac{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda}}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}$ e $\boldsymbol{\Gamma}^{-1}=\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{I}_p-\boldsymbol{\delta}\boldsymbol{\delta}^T)^{-1}\boldsymbol{\Sigma}^{-\frac{1}{2}}=\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T)\boldsymbol{\Sigma}^{-\frac{1}{2}}$ que

$$\frac{\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}}{\sqrt{1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}}}=\frac{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{\frac{1}{2}}}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}\frac{\boldsymbol{\Sigma}^{-\frac{1}{2}}\big(\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T\big)\boldsymbol{\Sigma}^{-\frac{1}{2}}}{\sqrt{1+\frac{\boldsymbol{\lambda}^T\big(\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T\big)\boldsymbol{\lambda}}{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}}=\frac{\boldsymbol{\lambda}^T(\mathbf{I}_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T)\boldsymbol{\Sigma}^{-\frac{1}{2}}}{\left(\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}\right)^2}=\frac{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(1+\boldsymbol{\lambda}^T\boldsymbol{\lambda})}{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}=\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}.$$

Com o resultado acima e denotando por $\widetilde{g}$ a f.d.p. desejada, concluímos enfim que

$$\begin{aligned}\widetilde{g}(t|\mathbf{y},u) &= \frac{\phi_1\left(t;\frac{\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}},\frac{1}{u(1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta})}\right)}{\Phi_1\big(\sqrt{u}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\big)}\\ &= \frac{\phi_1\left(t;\frac{\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}},\frac{1}{u(1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta})}\right)}{\Phi_1\left(\frac{\sqrt{u}\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\sqrt{1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}}}\right)}=\frac{\phi_1\left(t;\frac{\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}},\frac{1}{u(1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta})}\right)}{1-\Phi_1\left(-\frac{\sqrt{u}\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\sqrt{1+\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}}}\right)}.\end{aligned}$$

Logo, pela definição de Johnson, Kotz & Balakrishnan (1994), temos que $T|\mathbf{Y}=\mathbf{y},U=u\sim\ TN_{[0,+\infty)}\left(\mu_T,\frac{\sigma_T}{\sqrt{u}}\right)$.

■

Proposição 2.3.7. Seja $\mathbf{Y}\sim\ \mathrm{SMSN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\boldsymbol{\tau})$. Então,

$$\begin{aligned}u_1 &= E(U|\mathbf{Y}=\mathbf{y})=\frac{2f_0(\mathbf{y})}{f(\mathbf{y})}E\Big(U_{\mathbf{y}}\ \Phi(U_{\mathbf{y}}^{1/2}A)\Big);\\ z_1 &= E\Big(U^{1/2}\ W_\Phi(U^{1/2}A)\Big|\mathbf{Y}=\mathbf{y}\Big)=\frac{2f_0(\mathbf{y})}{f(\mathbf{y})}E\Big(U_{\mathbf{y}}^{1/2}\ \phi_1(U_{\mathbf{y}}^{1/2}A)\Big). \quad (2.15)\end{aligned}$$

Acima, $A = \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$, $W_\Phi(\ .\ ) = \frac{\phi_1(\ .\ )}{\Phi_1(\ .\ )}$, $f_0(\mathbf{y}) = \int_0^{+\infty} \phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right) h(u; \boldsymbol{\tau}) du$ é a f.d.p. do vetor aleatório $\mathbf{Y}_0$ e $U_{\mathbf{y}} \stackrel{d}{=} U | \mathbf{Y}_0 = \mathbf{y}$.

Demonstração:

Ver prova (mais geral) em Zeller, Lachos & Vilca-Labra (2009).

■

2.3.4 Misturas Finitas

Nesta subseção, faremos uma breve abordagem acerca das misturas misturas com o intuito de tratar um exemplo específico no capítulo seguinte. Para mais detalhes, ver Frühwirth-Schnatter (2006) e Mclachlan & Peel (2000).

Definição 2.3.5. Sejam $\mathbf{Y}$ um vetor aleatório contínuo e V uma variável aleatória discreta com função de probabilidade $h\colon \{1, \ldots, m\} \to [0, +\infty)$, ambos definidos no mesmo espaço de probabilidade. Considerando para cada $j \in \{1, \ldots, m\}$ que $f(\mathbf{y}|V = j) = f_j(\mathbf{y})$ e $h(j) = \nu_j$, a densidade $f(\mathbf{y}) = \sum_{j=1}^{m} \nu_j f_j(\mathbf{y})$ é dita uma *mistura finita*, onde f_j é o *j-ésimo componente da mistura* e ν_j é o *j-ésimo peso da mistura*.

Observação: Na definição acima, que pode ser vista em Souza Filho (2012), se cada $f_j(\mathbf{y})$ depende de um vetor de parâmetros $\boldsymbol{\theta}_j$, pondo $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ podemos escrever $f(\mathbf{y}; \boldsymbol{\theta}) = \sum_{j=1}^{m} \nu_j f_j(\mathbf{y}; \boldsymbol{\theta}_j)$.

Proposição 2.3.8. Se $\mathbf{Y}$ é um vetor aleatório p-dimensional que segue uma distribuição mistura finita com m componentes $f_1, \ldots, f_m$ de pesos respectivos $\nu_1, \ldots, \nu_m$, então

$$\begin{aligned} E(\mathbf{Y}) &= \sum_{j=1}^{m} \nu_j E_j(\mathbf{Y}); \\ Var(\mathbf{Y}) &= \sum_{j=1}^{m} \nu_j \left[Var_j(\mathbf{Y}) + E_j(\mathbf{Y}) E_j(\mathbf{Y})^T \right] - \left[\sum_{j=1}^{m} \nu_j E_j(\mathbf{Y}) \right] \left[\sum_{j=1}^{m} \nu_j E_j(\mathbf{Y}) \right]^T . \end{aligned} \quad (2.16)$$

Nas expressões anteriores, estamos considerando que $E_j(\mathbf{Y}) = \int_{\mathbb{R}^p} \mathbf{y} f_j(\mathbf{y}) d\mathbf{y}$ e $Var_j(\mathbf{Y}) = E_j\left(\Big(\mathbf{Y} - E_j(\mathbf{Y})\Big)\Big(\mathbf{Y} - E_j(\mathbf{Y})\Big)^T\right)$.

Demonstração:

Mantendo a notação da Definição 2.3.5, temos que a f.d.p. de $\mathbf{Y}$ é dada por $f(\mathbf{y}) = \sum_{i=1}^{m} \nu_j f_j(\mathbf{y})$, donde seu valor esperado pode ser calculado assim:

$$E(\mathbf{Y}) = \int_{\mathbb{R}^p} \mathbf{y} f(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^p} \mathbf{y} \left[\sum_{i=1}^{m} \nu_j f_j(\mathbf{y}) \right] d\mathbf{y} = \sum_{i=1}^{m} \nu_j \left[\int_{\mathbb{R}^p} \mathbf{y} f_j(\mathbf{y}) d\mathbf{y} \right] = \sum_{j=1}^{m} \nu_j E_j(\mathbf{Y}).$$

Para determinar $Var(\mathbf{Y})$, vamos computar separadamente $E(\mathbf{Y}\mathbf{Y}^T)$. Temos por definição de valor esperado e algumas propriedades que

$$\begin{aligned} E(\mathbf{Y}\mathbf{Y}^T) &= \int_{\mathbb{R}^p} \mathbf{y}\mathbf{y}^T f(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^p} \mathbf{y}\mathbf{y}^T \left[\sum_{i=1}^{m} \nu_j f_j(\mathbf{y}) \right] d\mathbf{y} = \sum_{i=1}^{m} \nu_j \left[\int_{\mathbb{R}^p} \mathbf{y}\mathbf{y}^T f_j(\mathbf{y}) d\mathbf{y} \right] \\ &= \sum_{j=1}^{m} \nu_j E_j(\mathbf{Y}\mathbf{Y}^T) = \sum_{j=1}^{m} \nu_j \left[Var_j(\mathbf{Y}) + E_j(\mathbf{Y}) E_j(\mathbf{Y})^T \right]. \end{aligned}$$

Dessa forma, concluímos do resultado acima que $Var(\mathbf{Y}) = E(\mathbf{Y}\mathbf{Y}^T) - E(\mathbf{Y})E(\mathbf{Y})^T = \sum_{j=1}^{m} \nu_j \left[Var_j(\mathbf{Y}) + E_j(\mathbf{Y}\mathbf{Y}^T) \right] - \left[\sum_{j=1}^{m} \nu_j E_j(\mathbf{Y}) \right] \left[\sum_{j=1}^{m} \nu_j E_j(\mathbf{Y}) \right]^T$. ∎

Vamos associar um vetor aleatório $\mathbf{V}$ de m coordenadas assumindo os valores 0 ou 1 à nova variável aleatória V como indicada na Definição 2.3.5 de modo que, para cada $j \in \{1, \ldots, m\}$, tenhamos $V_j = 1 \Leftrightarrow V = j$. Em Mclachlan & Peel (2000), vemos que $\mathbf{V}$ assim definido segue uma distribuição multinomial.

Remetendo novamente à definição inicial desta subseção e à observação subsequente, sendo $\mathbf{Y}$ um vetor aleatório com uma distribuição mistura finita, denotemos por $\mathbf{Y}_j$ o vetor $\mathbf{Y}|V_j = 1$. Dessa forma, se $F_j(\,.\,;\boldsymbol{\theta}_j)$ é a f.d.a. de $\mathbf{Y}_j$, temos a seguinte representação hierárquica para $\mathbf{Y}$:

$$\begin{aligned} \mathbf{Y}|V_j = 1 &\sim F_j(\,.\,;\boldsymbol{\theta}_j); \\ \mathbf{V} &\sim \text{Mult}(1, \nu_1, \ldots, \nu_m). \end{aligned} \tag{2.17}$$

Sendo h uma função de probabilidade na Definição 2.3.5, vale obrigatoriamente a relação $\sum_{j=1}^{m} \nu_j = 1$. Se $\mathbf{Y}$ possuir outra representação hierárquica, esta pode ser combinada à anterior para produzir uma nova que incorpore a estrutura de mistura finita. Veremos posteriormente que isso ocorre na *distribuição normal contaminada assimétrica* e na *distribuição normal assimétrica contaminada*, as quais serão vistas tanto como misturas de escala quanto como misturas finitas.

2.4 REGRESSÃO LINEAR MÚLTIPLA MULTIVARIADA

Os chamados *modelos de regressão linear múltipla multivariada* são generalizações dos modelos de regressão linear múltipla (univariada). Há pelo menos três maneiras de generalizar tais modelos: a forma clássica, que encontrada em Johnson & Wichern (2007) , além de duas formas alternativas: a primeira, vista em Branco, Dey & Sahu (2003) e a segunda é uma proposta que pode contribuir para uma seleção de variáveis mais flexível. A diferença entre as três formas se encontra na disposição das covariáveis, seja numa única matriz de planejamento (forma clássica), seja no uso de uma matriz de planejamento por indivíduo (formas alternativas) sendo a construção de cada uma dessas matrizes o que difere uma forma alternativa da outra. Mais precisamente, se $\mathbf{Y}$ é um vetor aleatório p-dimensional de respostas, as formas acima geram os seguintes modelos por indivíduo $i = 1, \ldots, n$ observado:

1. Forma clássica: $\mathbf{Y}_i^T = \mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\epsilon}_i^T$, onde $\boldsymbol{\beta}$ é uma matriz $q \times p$ de parâmetros desconhecidos, $\mathbf{x}_i^T$ é a i-ésima linha da matriz de planejamento $\mathbf{X}$ de ordem $n \times q$ e $\boldsymbol{\epsilon}_i$ é o vetor p-dimensional de erros aleatórios do i-ésimo indivíduo;

2. Forma alternativa I: $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, onde $\boldsymbol{\beta}$ é um vetor $q \times 1$ de parâmetros desconhecidos, $\mathbf{X}_i^T = \begin{bmatrix}\mathbf{x}_{i1} & \ldots & \mathbf{x}_{ip}\end{bmatrix}$ é a matriz $p \times q$ de covariáveis do i-ésimo indivíduo e $\boldsymbol{\epsilon}_i$ é o vetor p-dimensional de erros aleatórios do i-ésimo indivíduo;

3. Forma alternativa II: $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, onde $\boldsymbol{\beta}$ é um vetor $pq \times 1$ de parâmetros desconhecidos, $\mathbf{X}_i = \begin{bmatrix}\mathbf{x}_i^T & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_i^T & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{x}_i^T\end{bmatrix}$ é a matriz $p \times pq$ de covariáveis do i-ésimo indivíduo e $\boldsymbol{\epsilon}_i$ é o correspondente vetor p-dimensional de erros.

Em todas as formas, suporemos a distribuição dos erros teóricos por indivíduo como $\boldsymbol{\epsilon}_i \sim$ SSMN$_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ ou $\boldsymbol{\epsilon}_i \sim$ SMSN$_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$. Assim, obtém-se dos resultados das Subseções 2.3.2 e 2.3.3 os valores $E(\boldsymbol{\epsilon}_i), Var(\boldsymbol{\epsilon}_i), E(\mathbf{Y}_i)$ e $Var(\mathbf{Y}_i)$ em termos dos parâmetros de escala ($\boldsymbol{\Sigma}$), de forma ($\boldsymbol{\lambda}$ e $\boldsymbol{\tau}$) e da regressão ($\boldsymbol{\beta}$). A estimação desses parâmetros será objeto do Capítulo 3, em cuja Seção 3.2 adotaremos as formas alternativas optando por uma ou outra conforme a necessidade.

3 EMV VIA ALGORITMO EM NOS MODELOS MISTURAS NORMAIS ASSIMÉTRICAS COM REGRESSÃO

Neste capítulo, discutiremos o processo de estimação dos parâmetros nos modelos misturas apresentados no capítulo anterior. Separaremos os casos univariado e multivariado em virtude de diferenças de cunho algébrico entre eles, especialmente no que se refere ao parâmetro de escala. Em linhas gerais, seguiremos os processos de estimação apresentados em Zeller, Cabral & Lachos (2015) para o caso univariado e Ferreira, Lachos & Bolfarine (2016) para o caso multivariado.

A escolha dos referidos métodos de estimação em cada caso se deve ao fato de que procuraremos fornecer a eles contribuições específicas que não foram possíveis ou igualmente satisfatórias considerando a outra abordagem correspondente tanto no caso univariado quanto no multivariado. O caso univariado das misturas de escala assimétricas de normais (SSMN) foi explorado em Ferreira (2008) enquanto desenvolvimentos no caso multivariado das misturas de escala de normais assimétricas (SMSN) foram apresentados em Zeller, Lachos & Vilca-Labra (2009).

3.1 MODELOS MISTURAS UNIVARIADOS COM REGRESSÃO

Nesta seção, utilizaremos a abordagem de Zeller, Cabral & Lachos (2015) para tratar os modelos mistura de escala de normais assimétricas univariados. Em tais modelos, consideraremos Y uma variável aleatória que possui a distribuição discutida na Subseção 2.3.3 com $p = 1$ e parâmetros de locação μ, de escala $\sigma^2 > 0$, de assimetria λ e hiper-parâmetro $\boldsymbol{\tau}$.

A estrutura de tais modelos apresentada na Subseção 2.3.3 será agora adaptada ao caso univariado. Nessa situação particular, reformularemos a Definição 2.3.4 da seguinte maneira:

Definição 3.1.1. Uma variável aleatória Y segue uma *distribuição mistura de escala normal assimétrica* quando pode ser escrita na forma $Y = \mu + U^{-1/2}Z$, onde U é uma variável aleatória positiva com f.d.p. $h(u;\boldsymbol{\tau})$ e $Z \sim SN_1(0, \sigma^2, \lambda)$. Dessa forma, a f.d.p. de Y é dada por

$$f(y) = 2\int_0^{+\infty} \phi_1\left(y; \mu, \frac{\sigma^2}{u}\right) \Phi_1\left(\sqrt{u}\lambda\left(\frac{y-\mu}{\sigma}\right)\right) h(u;\boldsymbol{\tau})du,\ y \in \mathbb{R}. \tag{3.1}$$

Da Definição 3.1.1, obtemos a seguinte representação estocástica análoga à dada em (2.12) para a variável Y:

$$Y \stackrel{d}{=} \mu + U^{-1/2}\left(\Delta|T_0| + \omega T_1\right), \quad \text{onde} \quad \Delta = \frac{\sigma\lambda}{\sqrt{1+\lambda^2}} \text{ e } \omega = \frac{\sigma}{\sqrt{1+\lambda^2}}. \tag{3.2}$$

Na expressão (3.2), as variáveis T_0 e T_1 seguem distribuições normais padrão univariadas. Como casos particulares dos últimos resultados apresentados na Subseção 2.3.3, decorrem imediatamente as três proposições seguintes.

Proposição 3.1.1. Dada uma variável aleatória Y com distribuição mistura de escala normal assimétrica univariada e fator de escala U, seu valor esperado e sua variância são, respectivamente,

$$E(Y) = \mu + \sqrt{\frac{2}{\pi}}E(U^{-1/2})\Delta \quad \text{e} \quad Var(Y) = \sigma^2 E(U^{-1}) - \frac{2}{\pi}E(U^{-1/2})^2\Delta^2. \tag{3.3}$$

Proposição 3.1.2. Se Y segue uma distribuição mistura de escala normal assimétrica univariada e fator de escala U, então

$$\begin{aligned} Y|T=t, U=u &\sim N_1\left(\mu + \Delta t, \frac{\omega^2}{u}\right); \\ T|U=u &\sim TN_{[0,+\infty)}\left(0, \frac{1}{u}\right); \\ U &\sim H(\,.\,;\boldsymbol{\tau}). \end{aligned} \tag{3.4}$$

Proposição 3.1.3. Se Y é uma variável aleatória com f.d.p. f que segue uma distribuição mistura de escala normal assimétrica com fator de escala U, então $T|Y=y, U=u \sim TN_{[0,+\infty)}\left(\mu_T, \frac{\sigma_T}{\sqrt{u}}\right)$, onde $\mu_T = \frac{\Delta(y-\mu)}{\omega^2+\Delta^2}$ e $\sigma_T = \frac{\omega}{\sqrt{\omega^2+\Delta^2}}$.

Proposição 3.1.4. Seja Y uma variável aleatória com f.d.p. f que segue uma distribuição mistura de escala normal assimétrica univariada com fator de escala U. Temos então as seguintes esperanças condicionais:

$$u_1 = E(U|Y=y) = \frac{2f_0(y)}{f(y)}E\Big(U_y\Phi\Big(U_y^{1/2}A\Big)\Big|Y=y\Big);$$

$$z_1 = E\Big(U^{1/2}W_\Phi\big(U^{1/2}A\big)\Big|Y=y\Big) = \frac{2f_0(y)}{f(y)}E\Big(U_y^{1/2}\phi_1\big(U_y^{1/2}A\big)\Big|Y=y\Big).$$

Acima $A = \lambda\left(\frac{y-\mu}{\sigma}\right)$, $W_\Phi(\ .\ ) = \frac{\phi_1(\ .\ )}{\Phi_1(\ .\ )}$, $f_0(y) = \int_0^{+\infty}\phi_1\left(y;\mu,\frac{\sigma^2}{u}\right)h(u;\boldsymbol{\tau})du$ é a f.d.p. da variável aleatória Y_0 e $U_y \overset{d}{=} (U|Y_0 = y)$.

3.1.1 Exemplos Básicos

Nesta seção, explicitaremos os quatro exemplos básicos encontrados na literatura para as misturas de escala de normais assimétricas: são eles as distribuições normal assimétrica, T-Student assimétrica, Slash assimétrica e normal assimétrica contaminada.

Descreveremos brevemente cada um desses exemplos de distribuições misturas do tipo a ser considerado neste texto para, em seguida, discutir o processo de estimação de seus parâmetros.

Distribuição normal assimétrica

O caso univariado da distribuição normal assimétrica mostrada na Subseção 2.3.1 pode ser visto como uma particularização bastante peculiar de qualquer uma das formas de mistura de escala apresentadas. Para tanto, basta considerar $h(u;\boldsymbol{\tau})$ como o delta de Dirac centrado em $u = 1$. Dessa forma, obtemos a f.d.p. seguinte para $Y \sim\ \mathrm{SN}_1(\mu,\sigma^2,\lambda)$:

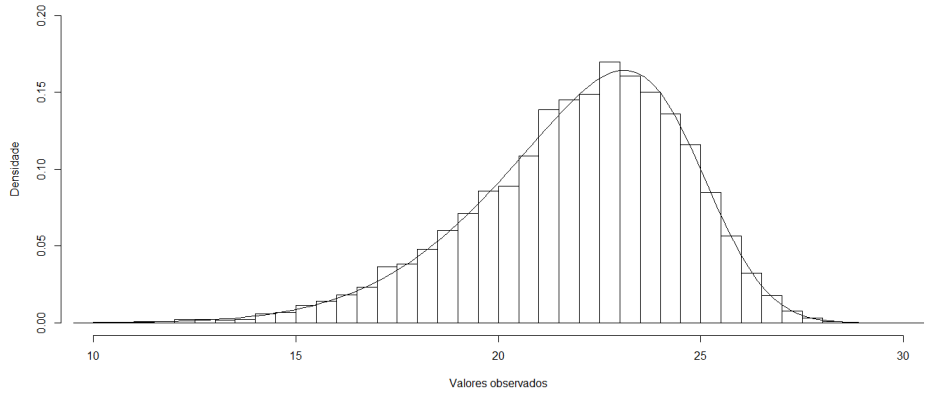
$$f(y;\mu,\sigma^2,\lambda) = 2\phi_1(y;\mu,\sigma^2)\Phi\left(\lambda\left(\frac{y-\mu}{\sigma}\right)\right),\ y\in\mathbb{R}. \tag{3.5}$$

Da definição do delta de Dirac, segue que $E(U^r) = \int_0^{+\infty}u^r h(u;\boldsymbol{\tau})du = 1$. Com isso, a Proposição 3.1.1 nos dá

$$E(Y) = \mu + \sqrt{\frac{2}{\pi}}\Delta \ \text{ e } \ Var(Y) = \sigma^2 - \frac{2}{\pi}\Delta^2. \tag{3.6}$$

Na Figura 1, vemos o gráfico da f.d.p. de uma distribuição normal assimétrica com $\mu = 25$, $\sigma = 4$ e $\lambda = -3$.

Figura 1 – Gráfico de uma normal assimétrica



Tal gráfico foi traçado sobre o histograma de uma amostra de 10000 observações geradas da referida distribuição no *software* `R` – R Core Team (2017). Note que o valor negativo da assimetria dada pelo parâmetro λ pode ser observado na cauda esquerda mais alongada da curva. Sendo os dados gerados dessa distribuição em quantidade suficientemente grande, vemos que seu comportamento registrado no histograma reflete o tipo de assimetria exemplificado: a assimetria negativa.

Distribuição T-Student assimétrica

Um dos exemplos mais conhecidos de mistura de escala da família normal assimétrica é a chamada distribuição T-Student assimétrica. Nessa distribuição, temos $\boldsymbol{\tau} = \nu$ numérico e $U \sim$ Gama $\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, donde segue que $h(u;\nu) = \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} u^{\nu/2-1} e^{-u\nu/2}$. Assim, escrevemos $Y \sim \ \mathrm{ST}_1(\mu, \sigma^2, \lambda, \nu)$ cuja f.d.p. é

$$f(y;\mu,\sigma^2,\lambda,\nu) = 2t_1(y;\mu,\sigma^2,\nu) T_1\left(A\sqrt{\frac{\nu+1}{\nu+d}}; \nu+1\right), \ y \in \mathbb{R}. \tag{3.7}$$

Acima, temos $A = \lambda\left(\frac{y-\mu}{\sigma}\right)$, $d = \left(\frac{y-\mu}{\sigma}\right)^2$ e $t_1(\,\cdot\,;\mu,\sigma^2,\nu)$, $T_1(\,\cdot\,;\nu)$, respectivamente, a f.d.p. e a f.d.a. de uma T-Student com ν graus de liberdade.

Da distribuição de U, obtemos $E(U^r) = \frac{\Gamma\left(\frac{\nu}{2}+r\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu}{2}\right)^{-r}$, o que nos dá em particular $E(U^{-1/2}) = \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\sqrt{\frac{\nu}{2}}$ e $E(U^{-1}) = \frac{\Gamma\left(\frac{\nu}{2}-1\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{\nu}{2} = \frac{\nu}{\nu-2}$.

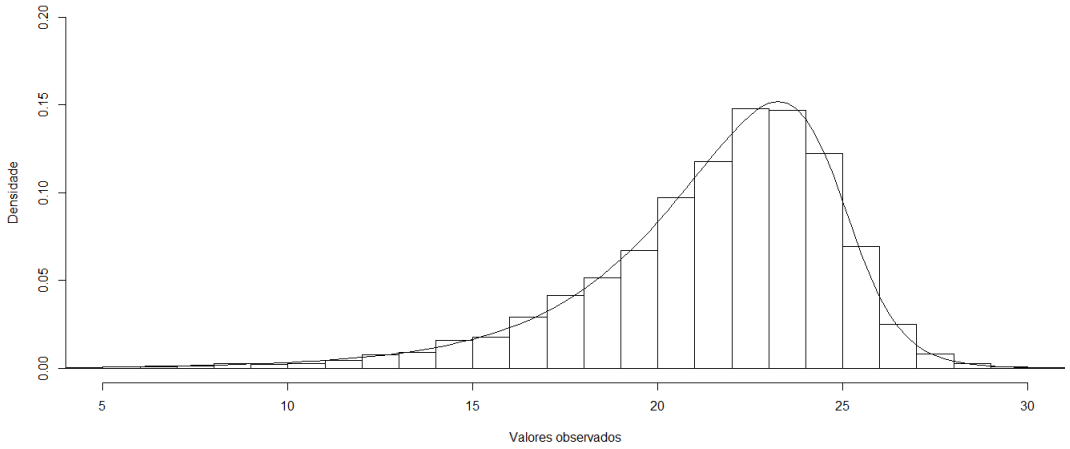
Pelos resultados anteriores e a Proposição 3.1.1, temos que

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

$$Var(Y) = \frac{\sigma^2\nu}{\nu-2} - \frac{\nu}{\pi}\frac{\Gamma\left(\frac{\nu-1}{2}\right)^2}{\Gamma\left(\frac{\nu}{2}\right)^2}\Delta^2,\ \nu > 2. \tag{3.8}$$

Na Figura 2, apresentamos o gráfico da f.d.p. de uma T-Student assimétrica com $\mu = 25$, $\sigma = 4$, $\lambda = -3$ e $\nu = 5$.

Figura 2 – Gráfico de uma T-Student assimétrica



Mais uma vez o gráfico foi traçado sobre o histograma de uma amostra de 10.000 observações geradas da referida distribuição no *software* `R`. Como se pode observar no gráfico da Figura 2, ambas as caudas são mais pesadas em relação à normal assimétrica apresentada acima devido ao hiper-parâmetro ν além do mesmo efeito de assimetria já comentado. Por isso, diz-se que a distribuição T-Student assimétrica combina duas características típicas de diversos conjunto de dados reais: assimetria e caudas pesadas conforme Azzalini & Capitanio (1999). Na mesma referência, podem ser encontradas as principais propriedades da distribuição T-Student assimétrica dentre as quais a escrita na forma indicada em (3.7).

Distribuição Slash assimétrica

Outro exemplo de mistura de escala da família normal assimétrica de que vamos tratar é a distribuição Slash assimétrica. Nesse caso, temos novamente $\boldsymbol{\tau} = \nu$ numérico e agora $U \sim$ Beta$(\nu, 1)$, donde se obtém $h(u;\nu) = \nu u^{\nu-1}$. Dessa forma, escrevemos $Y \sim$ SS$_1(\mu, \sigma^2, \lambda, \nu)$ com f.d.p. dada por

$$f(y;\mu,\sigma^2,\lambda,\nu) = 2\nu \int_0^1 u^{\nu-1}\phi_1\left(y;\mu,\frac{\sigma^2}{u}\right)\Phi_1(\sqrt{u}A)du,\ y \in \mathbb{R}. \tag{3.9}$$

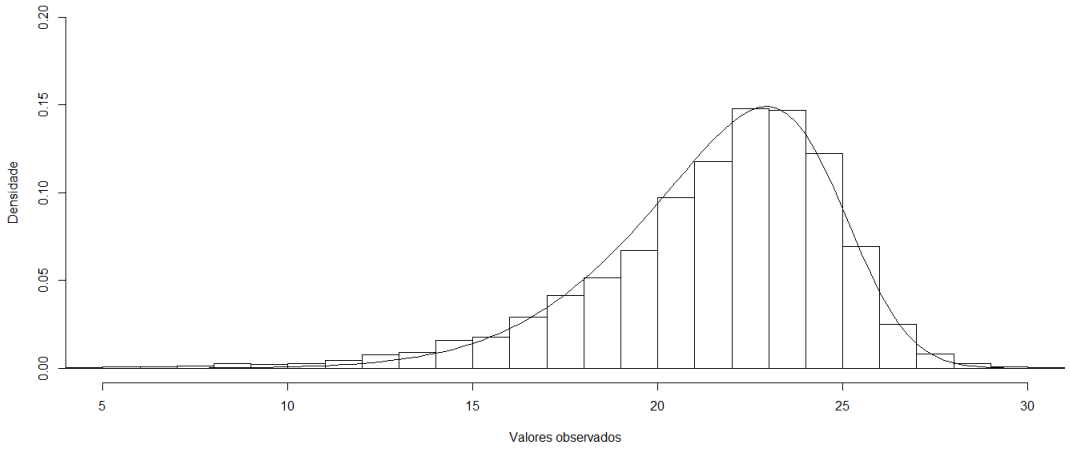
Novamente estamos considerando $A = \lambda\left(\frac{y-\mu}{\sigma}\right)$.

Da densidade de U, obtemos trivialmente que $E(U^r) = \frac{\nu}{\nu+r}$, donde vêm imediatamente as expressões $E(U^{-1/2}) = \frac{2\nu}{2\nu-1}$ e $E(U^{-1}) = \frac{\nu}{\nu-1}$. Dessa forma, pela Proposição 3.1.1, concluímos que

$$\begin{aligned} E(Y) &= \mu + \sqrt{\frac{2}{\pi}}\frac{2\nu}{2\nu-1}\Delta,\ \nu \neq \frac{1}{2}; \\ Var(Y) &= \frac{\sigma^2\nu}{\nu-1} - \frac{8\nu^2}{\pi(2\nu-1)^2}\Delta^2,\ \nu > 1. \end{aligned} \tag{3.10}$$

Na Figura 3, apresentamos o gráfico da f.d.p. de uma Slash assimétrica com $\mu = 25$, $\sigma = 4$, $\lambda = -3$ e $\nu = 5$.

Figura 3 – Gráfico de uma Slash assimétrica



Traçando outra vez o gráfico sobre o histograma de uma amostra de 10.000 observações da Slash assimétrica no *software* R, nota-se que a forma da Slash assimétrica é muito similar à da T-Student assimétrica. A primeira possui caudas ligeiramente mais pesadas que a segunda para hiper-parâmetros iguais, embora ambas apresentem essencialmente as mesmas características já mencionadas. Para mais resultados acerca da Slash assimétrica, ver Wang & Genton (2006).

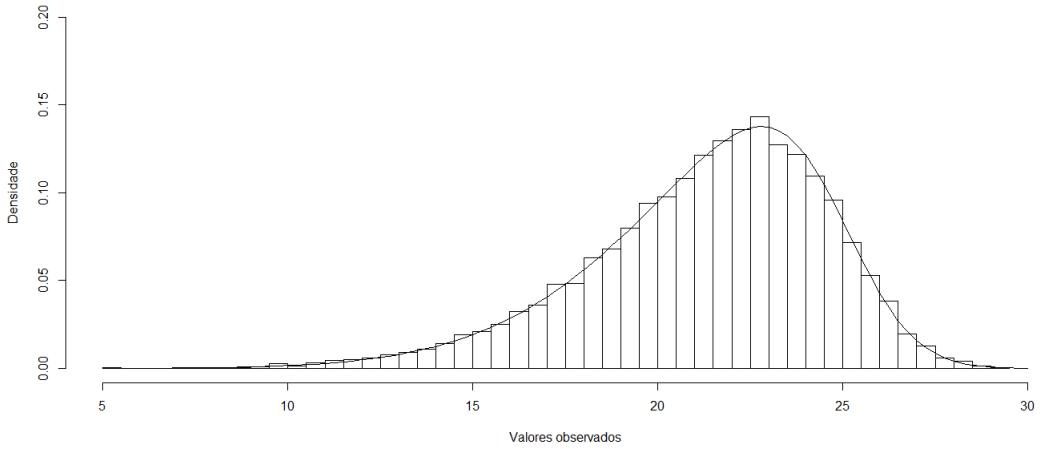
Distribuição normal assimétrica contaminada

Como dissemos anteriormente, a distribuição normal assimétrica contaminada será vista tanto como um exemplo de mistura de escala da família normal assimétrica quanto de mistura finita de duas normais assimétricas. A densidade de $Y \sim \text{SC}_1(\mu, \sigma^2, \lambda, \nu, \gamma)$ no caso univariado está definida para todo $y \in \mathbb{R}$ por

$$f(y;\mu,\sigma^2,\lambda,\nu,\gamma) = 2\left[\nu\phi_1\left(y;\mu,\frac{\sigma^2}{\gamma}\right)\Phi_1(\sqrt{\gamma}A) + (1-\nu)\phi_1\left(y;\mu,\sigma^2\right)\Phi_1(A)\right],\ A = \lambda\left(\frac{y-\mu}{\sigma}\right). \quad (3.11)$$

Os parâmetros μ, σ e λ possuem a mesma interpretação dos modelos anteriores (locação, escala e assimetria, respectivamente), ao passo que $\nu \in (0,1)$ e $\gamma \in (0,1)$ podem ser interpretados, respectivamente, como uma proporção ou peso da mistura e um fator de escala do componente da mistura (vide Definição 2.3.5). A Figura 4 mostra o gráfico da f.d.p. de uma normal assimétrica contaminada com $\mu = 25$, $\sigma = 4$, $\lambda = -3$, $\nu = 0,7$ e $\gamma = 0,6$.

Figura 4 – Gráfico de uma normal assimétrica contaminada



Esta distribuição também possui caudas mais pesadas do que a normal assimétrica e, portanto, também apresenta as características desejadas para a modelagem de vários conjuntos de dados reais. Comparada às outras distribuições misturas de escala assimétricas apresentadas, pode-se dizer que o parâmetro a mais deixa sua forma mais flexível. Na literatura mais recente, é comum tratar a distribuição normal assimétrica contaminada (e sua variante normal contaminada assimétrica a ser vista posteriormente) como mistura de escala conforme pode ser constatado em Lachos, Ghosh & Arellano-Valle (2010) e Ferreira, Lachos & Bolfarine (2016). Nas duas referências citadas, considera-se $\boldsymbol{\tau} = (\nu, \gamma)$ o hiper-parâmetro.

Na abordagem supracitada, considera-se na normal assimétrica contaminada $h(u; \nu, \gamma)$ uma função de probabilidade discreta que vale ν para $u = \gamma$, $1 - \nu$ para $u = 1$ e 0 para os demais valores de u (positivos). Assim, temos $E(U^r) = \nu\gamma^r + 1 - \nu$ e, portanto, a Proposição 3.1.1 fornece

$$\begin{aligned} E(Y) &= \mu + \sqrt{\frac{2}{\pi}}\left(\frac{\nu}{\sqrt{\gamma}} + 1 - \nu\right)\Delta; \\ Var(Y) &= \sigma^2\left(\frac{\nu}{\gamma} + 1 - \nu\right) - \frac{2}{\pi}\left(\frac{\nu}{\sqrt{\gamma}} + 1 - \nu\right)^2 \Delta^2. \end{aligned} \tag{3.12}$$

Optaremos pelo tratamento desta distribuição como uma mistura de escala no contexto univariado. Já na Seção 3.2 veremos outra maneira de enxergá-la.

Observação: As gerações dos dados das distribuições exemplificadas bem como os histogramas foram todos feitos no *software* `R`. No Anexo A, exibimos as funções programadas e utilizadas para esses procedimentos.

3.1.2 **Estimação dos Parâmetros**

Descrevemos nesta subseção o procedimento geral de estimação paramétrica para modelos com distribuição mistura de escala da família normal assimétrica, incorporando a regressão para assim obter casos particulares dos modelos propostos em Zeller, Cabral & Lachos (2015).

Apresentaremos aqui o método de estimação dos parâmetros do modelo da forma comum na literatura e com nossa proposta de modificação. Para tanto, vamos supor uma amostra aleatória $Y_1, \ldots, Y_n$ de uma variável Y que segue uma

distribuição mistura de escala da normal assimétrica com parâmetros de locação μ_i associado ao indivíduo i, de escala $\sigma > 0$, de assimetria λ e hiper-parâmetro $\boldsymbol{\tau}$.

Para cada um dos n indivíduos, suporemos ainda um vetor de covariáveis. De modo mais preciso, associaremos ao indivíduo i o vetor $\mathbf{x}_i$ de dimensão $q > 1$ com a primeira entrada 1 e as demais iguais aos valores das $q-1$ covariáveis conhecidas para esse indivíduo, de modo que $\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$ com $\boldsymbol{\beta}$ sendo o vetor de parâmetros desconhecidos da regressão.

Dessa forma, construímos o MRLM dado por $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + e_i;\ i = 1, \ldots, n$ com e_i o resíduo do modelo para o i-ésimo indivíduo. Da mesma forma que em Ferreira (2008), procederemos o ajuste do modelo não viesado (supondo erros com locação nula) através da estimação do vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \lambda, \boldsymbol{\tau})$ pelo método da máxima verossimilhança. A função log-verossimilhança das misturas de escala dadas na Definição 3.1.1 pode ser escrita na forma seguinte:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i;\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[\frac{1}{2}\ln\left(\frac{2}{\pi}\right) - \frac{1}{2}\Lambda + \ln K_i\right]. \tag{3.13}$$

Acima, temos $\Lambda = \ln\sigma^2$ e $K_i = \int_0^{+\infty} u^{1/2}e^{-ud_i/2}\Phi_1(\sqrt{u}A_i)h(u;\boldsymbol{\tau})du$. A partir de agora, fixaremos nas distribuições univariadas as notações $d_i = \frac{e_i^2}{\sigma^2}$ e $A_i = \frac{\lambda e_i}{\sigma}$.

De acordo com o que foi desenvolvido na Subseção 2.1, a maximização da função acima pode ser feita com o uso de qualquer algoritmo de otimização, particularmente do tipo Newton ou Quasi-Newton. Porém, a implementação desses algoritmos pode esbarrar em problemas de saída do conjunto viável ou ter complicada implementação devido ao grau de complexidade das funções envolvidas.

Em virtude disso, usaremos o algoritmo EM descrito na Seção 2.2. Nas distribuições SMSN, a representação hierárquica dada em (3.4) permite escrever a função log-verossimilhança dos dados completos $\mathbf{y}_{C_i} = (y_i, t_i, u_i)$ na forma

$$\begin{aligned}
\ell_C(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\ln f_C(y_i,t_i,u_i;\boldsymbol{\theta}) = \sum_{i=1}^{n}\ln\left[\phi_1\left(y_i;\mathbf{x}_i^T\beta+\Delta t_i,\frac{\omega^2}{u_i}\right)\phi_1\left(t_i;0,\frac{1}{u_i}\right)h(u_i;\boldsymbol{\tau})\right] \\
&= \sum_{i=1}^{n}\ln\left[2\frac{\sqrt{u_i}}{\omega\sqrt{2\pi}}e^{-\frac{u_i}{2\omega^2}(y_i-\mathbf{x}_i^T\beta-\Delta t_i)^2}\right] + \ln\left[\frac{2}{\sqrt{2\pi}}u_i e^{-ut_i^2/2}\right] + \sum_{i=1}^{n}\ln h(u_i;\boldsymbol{\tau}) \\
&= c - n\ln\omega - \frac{1}{2\omega^2}\sum_{i=1}^{n}u_i(y_i-\mathbf{x}_i^T\beta)^2 + \frac{\Delta}{\omega^2}\sum_{i=1}^{n}u_i t_i(y_i-\mathbf{x}_i^T\beta) - \frac{\Delta^2}{2\omega^2}\sum_{i=1}^{n}u_i t_i^2 + \sum_{i=1}^{n}\ln h(u_i;\boldsymbol{\tau}).
\end{aligned}$$

Na expressão anterior, o valor $c = n\ln 2 - n\ln 2\pi + \frac{1}{2}\sum_{i=1}^{n}(3\ln u_i - u_i t_i^2)$ é constante em relação aos parâmetros, sendo irrelevante na otimização da função ℓ_C. Note que estamos considerando a reparametrização $\Delta = \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}$ e $\omega = \frac{\sigma}{\sqrt{1+\lambda^2}}$.

Nesse caso, a função $Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) = E_{\widehat{\boldsymbol{\theta}}^{(k)}}(\ell_C(\boldsymbol{\theta})|\mathbf{y})$ na etapa k do algoritmo EM pode ser decomposta na forma $Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) = c + Q_1(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) + Q_2(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)})$, onde

$$\begin{aligned} Q_1(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) &= -n\ln\omega - \frac{1}{2\omega^2}\sum_{i=1}^{n}\widehat{u_i}^{(k)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \frac{\Delta}{\omega^2}\sum_{i=1}^{n}\widehat{u_i t_i}^{(k)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \frac{\Delta^2}{2\omega^2}\sum_{i=1}^{n}\widehat{u_i t_i^2}^{(k)}; \\ Q_2(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) &= \sum_{i=1}^{n} E_{\widehat{\boldsymbol{\theta}}^{(k)}}\big(\ln h(u_i;\boldsymbol{\tau})\big). \end{aligned} \tag{3.14}$$

Na notação introduzida na Subseção 2.2, o vetor de dados faltantes é $\mathbf{W} = (T, U)$. Dessa forma, a fim de encontrar $\widehat{u_i}^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}(U_i|Y_i = y_i)$, $\widehat{u_i t_i}^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}(U_i T_i|Y_i = y_i)$ e $\widehat{u_i t_i^2}^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}(U_i T_i^2|Y_i = y_i)$ para proceder o passo E do algoritmo, utilizaremos as Proposições 3.1.3 e 3.1.4 além de resultados da Teoria da Probabilidade. Para isso, fazendo $\widehat{e_i}^{(k)} = y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)}$, $\widehat{A_i}^{(k)} = \frac{\widehat{\lambda}^{(k)}\widehat{e_i}^{(k)}}{\widehat{\sigma}^{(k)}}$, $\widehat{\mu}_{T_i}^{(k)} = \frac{\widehat{\Delta}^{(k)}\widehat{e_i}^{(k)}}{\widehat{\Delta}^{(k)^2}+\widehat{\omega}^{(k)^2}}$, $\widehat{\sigma}_{T_i}^{(k)} = \frac{\widehat{\omega}^{(k)}}{\sqrt{\widehat{\Delta}^{(k)^2}+\widehat{\omega}^{(k)^2}}}$ e $\widehat{z}_i^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_i^{1/2}W_\Phi\Big(U_i^{1/2}\widehat{A_i}^{(k)}\Big)\middle|Y_i = y_i\right)$, mostraremos como obter as expressões a seguir:

$$\widehat{u_i t_i}^{(k)} = \widehat{u_i}^{(k)}\widehat{\mu}_{T_i}^{(k)} + \widehat{z}_i^{(k)}\widehat{\sigma}_{T_i}^{(k)} \quad \text{e} \quad \widehat{u_i t_i^2}^{(k)} = \widehat{u_i}^{(k)}\widehat{\mu}_{T_i}^{(k)^2} + \widehat{\sigma}_{T_i}^{(k)^2} + \widehat{z}_i^{(k)}\widehat{\mu}_{T_i}^{(k)}\widehat{\sigma}_{T_i}^{(k)}.$$

Inicialmente, fazendo $\mathbf{P}_i = (Y_i, U_i)$, a forma mais geral da identidade da esperança condicional nos permite concluir o seguinte para $s \in \{1, 2\}$:

$$E\big(U_i T_i^s|Y_i = y_i\big) = E\left(E\big(U_i T_i^s|\mathbf{P}_i\big)\big|Y_i = y_i\right) = E\left(E\big(U_i T_i^s|Y_i, U_i\big)\big|Y_i = y_i\right) = E\left(U_i E\big(T_i^s|Y_i, U_i\big)\big|Y_i = y_i\right).$$

Vamos agora adaptar ao nosso caso o resultado expresso na Proposição 3.1.3 com o propósito de determinar $E_{\widehat{\boldsymbol{\theta}}^{(k)}}\big(T_i^s|Y_i, U_i\big)$. Fazendo a adaptação mencionada, vemos que $T_i|Y_i = y_i, U_i = u_i;\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)} \sim TN_{[0,+\infty)}\left(\widehat{\mu}_{T_i}^{(k)}, \frac{\widehat{\sigma}_{T_i}^{(k)}}{\sqrt{u_i}}\right)$, onde temos $\widehat{\mu}_{T_i}^{(k)} = \dfrac{\widehat{\Delta}^{(k)}\big(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)}\big)}{\widehat{\omega}^{(k)^2} + \widehat{\Delta}^{(k)^2}}$ e $\widehat{\sigma}_{T_i}^{(k)} = \dfrac{\widehat{\omega}^{(k)}}{\sqrt{\widehat{\omega}^{(k)^2} + \widehat{\Delta}^{(k)^2}}}$.

Usando os resultados de Johnson, Kotz & Balakrishnan (1994) sobre a distribuição normal truncada e adotando a notação $W_\Phi\big(x\big) = \dfrac{\phi_1\big(x\big)}{\Phi_1\big(x\big)}$, obtemos após algumas manipulações as seguinte expressões:

$$E_{\widehat{\boldsymbol{\theta}}^{(k)}}(T_i|Y_i=y_i,U_i=u_i) = \widehat{\mu}_{T_i}^{(k)} + W_\Phi\left(\sqrt{u_i}\frac{\widehat{\mu}_{T_i}^{(k)}}{\widehat{\sigma}_{T_i}^{(k)}}\right)\frac{\widehat{\sigma}_{T_i}^{(k)}}{\sqrt{u_i}};$$

$$Var_{\widehat{\boldsymbol{\theta}}^{(k)}}(T_i|Y_i=y_i,U_i=u_i) = \left[1 - W_\Phi\left(\sqrt{u_i}\frac{\widehat{\mu}_{T_i}^{(k)}}{\widehat{\sigma}_{T_i}^{(k)}}\right)\frac{\widehat{\mu}_{T_i}^{(k)}}{\widehat{\sigma}_{T_i}^{(k)}}\sqrt{u_i} - W_\Phi\left(\sqrt{u_i}\frac{\widehat{\mu}_{T_i}^{(k)}}{\widehat{\sigma}_{T_i}^{(k)}}\right)^2\right]\frac{\widehat{\sigma}_{T_i}^{(k)^2}}{u_i}.$$

Sendo $\frac{\widehat{\mu}_{T_i}^{(k)}}{\widehat{\sigma}_{T_i}^{(k)}} = \dfrac{\widehat{\Delta}^{(k)}}{\widehat{\omega}^{(k)}\sqrt{\widehat{\omega}^{(k)^2}+\widehat{\Delta}^{(k)^2}}}\big(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)}\big) = \dfrac{\widehat{\lambda}^{(k)}}{\widehat{\sigma}^{(k)}}\big(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)}\big) = \widehat{A_i}^{(k)}$,
podemos escrever com base nos resultados e convenções anteriores o seguinte:

$$\begin{aligned}
E_{\widehat{\boldsymbol{\theta}}^{(k)}}(T_i|Y_i=y_i,U_i=u_i) &= \widehat{\mu}_{T_i}^{(k)} + W_\Phi(\sqrt{u_i}\widehat{A_i}^{(k)})\frac{\widehat{\sigma}_{T_i}^{(k)}}{\sqrt{u_i}};\\
E_{\widehat{\boldsymbol{\theta}}^{(k)}}(T_i^2|Y_i=y_i,U_i=u_i) &= Var_{\widehat{\boldsymbol{\theta}}^{(k)}}(T_i|Y_i=y_i,U_i=u_i) + E_{\widehat{\boldsymbol{\theta}}^{(k)}}(T_i|Y_i=y_i,U_i=u_i)^2\\
&= \widehat{\mu}_{T_i}^{(k)^2} + \frac{\widehat{\mu}_{T_i}^{(k)}\widehat{\sigma}_{T_i}^{(k)}}{\sqrt{u_i}}W_\Phi(\sqrt{u_i}\widehat{A_i}^{(k)}) + \frac{\widehat{\sigma}_{T_i}^{(k)}}{u_i}.
\end{aligned}$$

Dessa forma, concluímos das expressões acima e das propriedades do valor esperado que

$$\begin{aligned}
\widehat{u_it_i}^{(k)} &= E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_iT_i|Y_i=y_i\right) = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_iE_{\widehat{\boldsymbol{\theta}}^{(k)}}(T_i|Y_i,U_i)|Y_i=y_i\right)\\
&= \widehat{\mu}_{T_i}^{(k)}E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_i|Y_i=y_i\right) + \widehat{\sigma}_{T_i}^{(k)}E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_i^{1/2}W_\Phi\big(U_i^{1/2}\widehat{A_i}^{(k)}\big)\Big|Y_i=y_i\right)\\
&= \widehat{u}_i^{(k)}\widehat{\mu}_{T_i}^{(k)} + \widehat{z}_i^{(k)}\widehat{\sigma}_{T_i}^{(k)};\\
\widehat{u_it_i^2}^{(k)} &= E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_iT_i^2|Y_i=y_i\right) = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_iE_{\widehat{\boldsymbol{\theta}}^{(k)}}(T_i^2|Y_i,U_i)|Y_i=y_i\right)\\
&= \widehat{\mu}_{T_i}^{(k)^2}E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_i|Y_i=y_i\right) + \widehat{\mu}_{T_i}^{(k)}\widehat{\sigma}_{T_i}^{(k)}E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_i^{1/2}W_\Phi\big(U_i^{1/2}\widehat{A_i}^{(k)}\big)\Big|Y_i=y_i\right) + \widehat{\sigma}_{T_i}^{(k)^2}\\
&= \widehat{u}_i^{(k)}\widehat{\mu}_{T_i}^{(k)^2} + \widehat{z}_i^{(k)}\widehat{\mu}_{T_i}^{(k)}\widehat{\sigma}_{T_i}^{(k)} + \widehat{\sigma}_{T_i}^{(k)^2}.
\end{aligned}$$

Quanto aos valores de $\widehat{u_i}^{(k)}$ bem como de $\widehat{z}_i^{(k)}$, os quais são descritos na Proposição 3.1.4, veremos na Subseção 3.1.3 sua expressão para cada um dos exemplos a serem trabalhados.

No nosso caso, em vez do tradicional passo M que envolveria a maximização direta da função $Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)})$ determinada acima, utilizaremos aqui a versão ECME do algoritmo proposta por Liu & Rubin (1994). Esta consistirá, no nosso caso, em proceder a maximização condicional de $Q_1(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)})$ supondo o hiper-parâmetro $\boldsymbol{\tau} = \widehat{\boldsymbol{\tau}}^{(k)}$ fixo e, logo em seguida, maximizando a log-verossimilhança original ℓ em $\boldsymbol{\tau}$ após obtidos os valores $\widehat{\boldsymbol{\beta}}^{(k+1)}$, $\widehat{\sigma}^{(k+1)}$ e $\widehat{\lambda}^{(k+1)}$.

Esses últimos três valores são provenientes das condições estacionárias de primeira ordem $\dfrac{\partial Q_1}{\partial \boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(k)};\widehat{\boldsymbol{\theta}}^{(k)}\right) = \mathbf{0}$. Após o cálculo das derivadas de Q_1 correspondentes a cada parâmetro, obtém-se as seguintes estimativas explícitas conforme se pode ver como caso particular em Zeller, Cabral & Lachos (2015):

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(k+1)} &= \left[\sum_{i=1}^{n} \widehat{u_i}^{(k)}\mathbf{x}_i\mathbf{x}_i^T\right]^{-1}\left[\sum_{i=1}^{n} \widehat{u_i}^{(k)}y_i - \widehat{u_i t_i}^{(k)}\widehat{\Delta}^{(k+1)}\right];\\
\widehat{\Delta}^{(k+1)} &= \sum_{i=1}^{n} \widehat{u_i t_i}^{(k)}\left(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k+1)}\right) / \sum_{i=1}^{n} \widehat{u_i t_i^2}^{(k)};\\
\widehat{\omega}^{(k+1)} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{u_i}^{(k)}\left(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k+1)}\right)^2 - 2\widehat{u_i t_i}^{(k)}\left(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k+1)}\right)\widehat{\Delta}^{(k+1)} + \widehat{u_i t_i^2}^{(k)}\widehat{\Delta}^{(k+1)^2}\right]}.
\end{aligned}
$$

Observe que as expressões de $\widehat{\boldsymbol{\beta}}^{(k+1)}$ e $\widehat{\Delta}^{(k+1)}$ revelam uma dependência mútua entre os dois parâmetros. Por isso, na prática, ao implementar o algoritmo em um *software* como o `R` se considera $\widehat{\Delta}^{(k)}$ na expressão de $\widehat{\boldsymbol{\beta}}^{(k+1)}$ e não $\widehat{\Delta}^{(k+1)}$.

No entanto, uma modificação simples pode resolver o problema da dependência mútua mencionada. Para tanto, basta notar que ao derivarmos $Q_1(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)})$ em relação a $\boldsymbol{\beta}$, Δ e ω, respectivamente, obtemos

$$
\begin{cases}
\sum\limits_{i=1}^{n}\left[\frac{1}{\omega^2}\widehat{u_i}^{(k)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i - \frac{\Delta}{\omega^2}\widehat{u_i t_i}^{(k)}\mathbf{x}_i\right] = \mathbf{0}\\
\sum\limits_{i=1}^{n}\left[-\frac{1}{\omega} + \frac{1}{\omega^3}\widehat{u_i}^{(k)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 - \frac{2\Delta}{\omega^3}\widehat{u_i t_i}^{(k)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \frac{\Delta^2}{\omega^3}\widehat{u_i t_i^2}^{(k)}\right] = 0\\
\sum\limits_{i=1}^{n}\left[\frac{1}{\omega^2}\widehat{u_i t_i}^{(k)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \frac{\Delta}{\omega^2}\widehat{u_i t_i^2}^{(k)}\right] = 0
\end{cases}
$$

Manipulando as expressões acima, chegamos ao seguinte sistema de equações:

$$\begin{cases} \sum\limits_{i=1}^{n} \widehat{u_i}^{(k)} \mathbf{x}_i \mathbf{x}_i^T \boldsymbol{\beta} + \left(\sum_{i=1}^{n} \widehat{u_i t_i}^{(k)} \mathbf{x}_i \right) \Delta = \sum\limits_{i=1}^{n} \widehat{u_i}^{(k)} y_i \mathbf{x}_i \\ \sum\limits_{i=1}^{n} \widehat{u_i t_i}^{(k)} \mathbf{x}_i^T \boldsymbol{\beta} + \left(\sum\limits_{i=1}^{n} \widehat{u_i t_i^2}^{(k)} \right) \Delta = \sum\limits_{i=1}^{n} \widehat{u_i t_i}^{(k)} y_i \\ \sum\limits_{i=1}^{n} \left[\widehat{u_i}^{(k)} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 - 2 \widehat{u_i t_i}^{(k)} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \Delta + \widehat{u_i t_i^2}^{(k)} \Delta^2 \right] = n\omega^2 \end{cases}$$

Dessa forma, percebe-se que o subsistema de equações obtido pela retirada da última equação acima é linear em $\boldsymbol{\beta}$ e Δ. Dessa forma, a menos de problemas de posto da matriz $\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$, este possui solução única. Mais ainda, pelo fato de ω ser positivo por construção, $\widehat{\omega}^{(k+1)}$ fica unicamente determinado na última equação mediante o conhecimento de $\widehat{\boldsymbol{\beta}}^{(k+1)}$ e $\widehat{\Delta}^{(k+1)}$ das equações anteriores.

Resolvendo o sistema obtido a começar pelo subsistema linear, conseguimos as seguintes soluções explícitas exatas:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{(k+1)} &= \left[\left(\sum_{i=1}^{n} \widehat{u_i t_i^2}^{(k)} \right) \sum_{i=1}^{n} \widehat{u_i}^{(k)} \mathbf{x}_i \mathbf{x}_i^T - \left(\sum_{i=1}^{n} \widehat{u_i t_i}^{(k)} \mathbf{x}_i \right) \left(\sum_{i=1}^{n} \widehat{u_i t_i}^{(k)} \mathbf{x}_i \right)^T \right]^{-1} \\ &\times \left[\left(\sum_{i=1}^{n} \widehat{u_i t_i^2}^{(k)} \right) \left(\sum_{i=1}^{n} \widehat{u_i}^{(k)} y_i \mathbf{x}_i \right) - \left(\sum_{i=1}^{n} \widehat{u_i t_i}^{(k)} y_i \right) \left(\sum_{i=1}^{n} \widehat{u_i t_i}^{(k)} \mathbf{x}_i \right) \right]; \\ \widehat{\Delta}^{(k+1)} &= \sum_{i=1}^{n} \widehat{u_i t_i}^{(k)} \widehat{e_i}^{(k+1)} / \sum_{i=1}^{n} \widehat{u_i t_i^2}^{(k)}; \\ \widehat{\omega}^{(k+1)} &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[\widehat{u_i}^{(k)} \widehat{e_i}^{(k+1)^2} - 2 \widehat{u_i t_i}^{(k)} \widehat{e_i}^{(k+1)} \widehat{\Delta}^{(k+1)} + \widehat{u_i t_i^2}^{(k)} \widehat{\Delta}^{(k+1)^2} \right]}. \end{aligned}$$

Nos resultados acima, utilizamos a notação $\widehat{e_i}^{(k+1)} = y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}^{(k+1)}$ por simplicidade. Após obter os três parâmetros acima de qualquer uma das formas indicadas, concluímos a etapa k fazendo $\widehat{\sigma}^{(k+1)} = \sqrt{\widehat{\omega}^{(k+1)^2} + \widehat{\Delta}^{(k+1)^2}}$, $\widehat{\lambda}^{(k+1)} = \frac{\widehat{\Delta}^{(k+1)}}{\widehat{\omega}^{(k+1)}}$ e escolhendo $\widehat{\boldsymbol{\tau}}^{(k+1)} \in \underset{\boldsymbol{\tau}}{\operatorname{argmax}}\, \ell\left(\widehat{\boldsymbol{\beta}}^{(k+1)}, \widehat{\sigma}^{(k+1)}, \widehat{\lambda}^{(k+1)}, \boldsymbol{\tau} \right)$. Encontra-se um único valor de $\widehat{\boldsymbol{\tau}}^{(k+1)}$ na maioria das vezes, sendo relativamente simples verificar que a última função de $\boldsymbol{\tau}$ é côncava nos nossos exemplos. Porém, dois deles costumam apresentar um problema neste ponto, o qual mencionaremos na Subseção 3.1.3.

Na análise dos exemplos, faremos uma comparação do desempenho computacional de ambos métodos para constatar que o segundo é, até certo ponto, mais rápido que primeiro, fornecendo praticamente as mesmas estimativas.

Os valores iniciais para os três primeiros parâmetros são estimativas provenientes de uma suposição "ingênua" de normalidade dos modelos, ou seja, $\boldsymbol{\beta}^{(0)} = \left[\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\right]^{-1}\left[\sum_{i=1}^{n} \mathbf{y}_i\mathbf{x}_i\right]$, $\sigma^{(0)} = S_{\mathbf{e}^{(0)}}$ e $\lambda^{(0)} = g_{\mathbf{e}^{(0)}}$, onde $\mathbf{e}^{(0)} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(0)}$ é o vetor de resíduos inicial do modelo cuja utilização é preferível à do vetor $\mathbf{y}$ para evitar um inflacionamento dos parâmetros. Já os valores iniciais do hiperparâmetro podem variar consideravelmente de acordo com a distribuição e, por isso, apenas serão apresentados na Subseção 3.1.3. Como critério de parada, utilizaremos $\left\|\widehat{\boldsymbol{\theta}}^{(k)} - \widehat{\boldsymbol{\theta}}^{(k)}\right\| < \varepsilon$, o qual se mostrou mais eficaz na consistência das estimativas em relação a outro critério geralmente adotado: $\left|\ell\left(\widehat{\boldsymbol{\theta}}^{(k)}\right) - \ell\left(\widehat{\boldsymbol{\theta}}^{(k)}\right)\right| < \varepsilon$. Ver uso em Ferreira, Lachos & Bolfarine (2016) e Zeller, Cabral & Lachos (2015).

Para a obtenção das medidas que endossam a consistência da EMV de acordo com o exposto na Subseção 2.1 e também de critérios que controlem as simulações, precisaremos determinar o vetor escore e a matriz de informação de Fisher observada nos modelos misturas de escala. Estes correspondem, respectivamente, ao gradiente e à oposta da hessiana da log-verossimilhança (dos dados incompletos ou observados) indicada na expressão (3.13).

Dessa forma, o escore e a matriz de informação observada podem ser expressos de modo genérico na notação do cálculo matricial, respectivamente, por

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \boldsymbol{\theta}};\ \frac{\partial \ell_i}{\partial \boldsymbol{\theta}} = -\frac{1}{2}\frac{\partial \Lambda}{\partial \boldsymbol{\theta}} + \frac{1}{K_i}\frac{\partial K_i}{\partial \boldsymbol{\theta}};$$

$$-\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = -\sum_{i=1}^{n} \frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T};\ -\frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \frac{1}{2}\frac{\partial^2 \Lambda}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} + \frac{1}{K_i^2}\frac{\partial K_i}{\partial \boldsymbol{\theta}}\frac{\partial K_i}{\partial \boldsymbol{\theta}^T} - \frac{1}{K_i}\frac{\partial^2 K_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}.$$

Para explicitar as derivadas em relação a cada sub-parâmetro de $\boldsymbol{\theta}$ separadamente e possibilitar sua implementação computacional, utilizaremos as notações a seguir:

1. $I_i^G(r,s) = \sqrt{2\pi}\int_0^{+\infty} u^r(\ln u)^s \phi_1(\sqrt{ud_i})G(\sqrt{u}A_i)h(u;\boldsymbol{\tau})du;$
2. $J_i^G(r) = \sqrt{2\pi}\int_0^{+\infty} u^r \phi_1(\sqrt{ud_i})G(\sqrt{u}A_i)\frac{\partial h}{\partial \boldsymbol{\tau}}(u;\boldsymbol{\tau})du;$
3. $L_i^G(r) = \sqrt{2\pi}\int_0^{+\infty} u^r \phi_1(\sqrt{ud_i})G(\sqrt{u}A_i)\frac{\partial^2 h}{\partial \boldsymbol{\tau}\partial \boldsymbol{\tau}^T}(u;\boldsymbol{\tau})du.$

A função G que aparece como índice sobrescrito nas integrais acima pode ser $\phi := \phi_1$ ou $\Phi := \Phi_1$. Em particular, verifica-se de maneira imediata que $K_i = \int_0^{+\infty} u^{1/2} e^{-ud_i/2} \Phi_1(\sqrt{u}A_i) h(u;\boldsymbol{\tau}) du = I_i^{\Phi}\left(\frac{1}{2}, 0\right)$. As integrais $J_i^G(r)$ e $L_i^G(r)$ que aparecem nas expressões acima evidentemente possuem formulações distintas em cada exemplo e serão apresentadas na subseção seguinte para as distribuições T-Student e Slash assimétricas.

Especificamente no exemplo da normal assimétrica contaminada, em que a integral da mistura de escala se degenera numa soma pelo fato de $h(u;\nu,\gamma)$ ser uma função de probabilidade discreta, não faz sentido definir as expressões $J_i^G(r)$ e $L_i^G(r)$. Nesse caso, faremos $I_i^G(r,0) = \sqrt{2\pi}\left[\nu\gamma^r\phi(\sqrt{\gamma d_i})G(\sqrt{\gamma}A_i) + (1-\nu)\phi(\sqrt{d_i})G(A_i)\right]$, donde ainda obtemos a igualdade $K_i = I_i^{\Phi}\left(\frac{1}{2}, 0\right)$. Definimos adiante, porém, novas expressões $J_{1i}^G(r)$ e $J_{2i}^G(r)$ para permitir a expressão das derivadas parciais de ℓ numa notação relativamente simples.

1. $J_{1i}^G(r) = \sqrt{2\pi}\nu\gamma^r\phi(\sqrt{\gamma d_i})G(\sqrt{\gamma}A_i)$;
2. $J_{2i}^G(r) = \sqrt{2\pi}\left[\gamma^r\phi(\sqrt{\gamma d_i})G(\sqrt{\gamma}A_i) - \phi(\sqrt{d_i})G(\sqrt{\gamma}A_i)\right]$.

Com todas essas convenções, as derivadas de interesse podem ser efetivamente calculadas e se encontram listadas no Anexo B.

3.1.3 **Estudo de Casos**

O objetivo desta subseção é mostrar que a pequena modificação proposta na Subseção 3.1.2 acelera o processo de estimação até certo ponto, na medida em que zera o escore mais rapidamente e reduz o número de iterações necessárias para atingir o critério de parada, embora exija mais operações matriciais. Faremos essa constatação nos exemplos da Subseção 3.1.1, exceto na distribuição normal assimétrica cujos resultados da modificação não se mostraram satisfatórios.

Neste estudo de simulação com dados artificiais gerados no `R`, estipulamos cinco tamanhos de amostras: 50, 100, 200, 500 e 1.000 para testar o quão boa é a estimação e sua consistência, além de comparar os algoritmos padrão e modificado. Em todos os exemplos, geramos duas variáveis explicativas com distribuições uniformes $U(1,10)$ e $U(15,20)$ e realizamos 500 simulações por tamanho amostral.

Os parâmetros usados na geração foram: $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2]^T = [25\ \ -10\ 5]^T$, $\sigma = 2,8$, $\lambda = -1,4$ e hiper-parâmetro $\nu = 4,5$ na T-Student e na Slash assimétricas, ou hiper-parâmetro composto por $\nu = 0,65$ e $\gamma = 0,2$ na distribuição normal assimétrica contaminada.

Apresentaremos os resultados em duas tabelas: uma contendo a média das estimativas por parâmetro e medidas de variabilidade (desvio padrão e média dos erros padrão) a fim de verificar a recuperação dos parâmetros e da informação de Fisher; e outra com módulo dos vieses médio (VME) e mediano (VMD), erro quadrático médio (EQM) e desvio absoluto mediano (DAM) – ver Subseção 2.1 – para avaliar a consistência em ambos os algoritmos, além de apresentar gráficos de linha mostrando as medidas (relativas) baseadas na mediana com o intuito de evidenciar o comportamento mediano da estimação por parâmetro.

Além disso, para comparar o desempenho computacional dos métodos, mostraremos boxplots de tempo e iterações em três tamanhos de amostra n classificados da seguinte forma: pequeno ($n = 50$), médio ($n = 200$) e grande ($n = 1.000$).

O valor de ε no critério de parada foi fixado em 10^{-6} por heurística com o propósito principal de atender às condições necessárias para a convergência dos algoritmos a um máximo local não degenerado, ou seja, escore suficientemente pequeno e matriz de informação negativa definida (e bem condicionada).

Com esse intuito, eliminamos os conjuntos de dados simulados cujos parâmetros estimados não satisfaziam a maior coordenada do escore menor que 10^{-2} e o menor autovalor da matriz de informação maior que 10^{-4}, além de limitar superiormente as iterações de cada algoritmo por 4.000.

A partir de agora começaremos uma análise caso a caso de acordo com os aspectos discorridos acima, apresentando as expressões restantes para a implementação em cada um dos exemplos discutidos do algoritmo EM, além de um resumo dos resultados das simulações computacionais feitas no *software* `R` através das tabelas e gráficos já mencionados anteriormente.

T-Student assimétrica

Já vimos que na distribuição t assimétrica $h(u;\nu) = \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} u^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}u}$, donde

$$\frac{\partial h}{\partial \nu}(u;\nu) = \frac{1}{2}\left[\ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}+1\right) + \ln u - u\right] h(u;\nu);$$

$$\frac{\partial^2 h}{\partial \nu^2}(u;\nu) = \frac{1}{4}\left\{\left[\frac{2}{\nu} - \Psi_1\left(\frac{\nu}{2}\right)\right] + \left[\ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right) + 1 + \ln u - u\right]^2\right\} h(u;\nu).$$

Com as derivadas anteriores, obtemos as seguintes expressões:

$$J_i^G(r) = \frac{1}{2}\left\{\left[\ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right) + 1\right] I_i^G(r,0) - I_i^G(r+1,0) + I_i^G(r,1)\right\};$$

$$L_i(r) = \frac{1}{4}\left\{\left[\frac{2}{\nu} - \Psi_1\left(\frac{\nu}{2}\right) + \left(\ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right) + 1\right)^2\right] I_i^G(r,0) + \right.$$
$$\left. +2\left(\ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right) + 1\right)\left(I_i^G(r,1) - I_i^G(r+1,0)\right) + I_i^G(r,2) - 2I_i^G(r+1,1) + I_i^G(r+2,0)\right\}.$$

Resolveremos $I_i^G(r,s) = \sqrt{2\pi}\frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)}\int_0^{+\infty} u^{\frac{\nu}{2}-1+r}(\ln u)^s \phi\left(\sqrt{u(\nu+d_i)}\right) G(\sqrt{u}A_i)du$, utilizando a função `integrate` do `R` para seu cálculo. Mas, em particular, temos segundo Zeller (2009) os seguintes resultados fechados:

$$I_i^{\Phi}(r,0) = \left(\frac{2}{\nu+d_i}\right)^r \left(\frac{\nu}{\nu+d_i}\right)^{\frac{\nu}{2}} \frac{\Gamma\left(\frac{\nu}{2}+r\right)}{\Gamma\left(\frac{\nu}{2}\right)} T_1\left(A_i\sqrt{\frac{\nu+2r}{\nu+d_i}}; \nu+2r\right);$$

$$I_i^{\phi}(r,0) = \frac{2^r \nu^{\nu/2}}{\sqrt{2\pi}} \frac{\Gamma\left(\frac{\nu}{2}+r\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{1}{\nu+d_i+A_i^2}\right)^{\frac{\nu}{2}+r}.$$

Como consequência da Proposição 3.1.4, vemos em Zeller (2009) que os valores esperados condicionais do passo E na etapa k do algoritmo EM são

$$\widehat{u_i}^{(k)} = \frac{T_1\left(\widehat{A_i}^{(k)}\sqrt{\frac{\widehat{\nu}^{(k)}+3}{\widehat{\nu}^{(k)}+\widehat{d_i}^{(k)}}};\ \widehat{\nu}^{(k)}+3\right)}{T_1\left(\widehat{A_i}^{(k)}\sqrt{\frac{\widehat{\nu}^{(k)}+1}{\widehat{\nu}^{(k)}+\widehat{d_i}^{(k)}}};\ \widehat{\nu}^{(k)}+1\right)} \frac{\widehat{\nu}^{(k)}+1}{\widehat{\nu}^{(k)}+d_i^{(k)}};$$

$$\widehat{z_i}^{(k)} = \frac{\Gamma\left(\frac{\widehat{\nu}^{(k)}+2}{2}\right)/\Gamma\left(\frac{\widehat{\nu}^{(k)}+1}{2}\right)}{\sqrt{\pi}T_1\left(\widehat{A_i}^{(k)}\sqrt{\frac{\widehat{\nu}^{(k)}+1}{\widehat{\nu}^{(k)}+\widehat{d_i}^{(k)}}};\ \widehat{\nu}^{(k)}+1\right)} \frac{\left(\widehat{\nu}^{(k)}+\widehat{d_i}^{(k)}\right)^{\frac{\widehat{\nu}^{(k)}+1}{2}}}{\left(\widehat{\nu}^{(k)}+\widehat{d_i}^{(k)}+\widehat{A_i}^{(k)^2}\right)^{\frac{\widehat{\nu}^{(k)}+2}{2}}}.$$

Para esta distribuição, controlamos a limitação superior do parâmetro ν, descartando estimativas nas quais esse parâmetro supera 30 (caso já considerado próximo da normalidade). Esse último procedimento se justifica em virtude do fenômeno chamado *platô da log-verossimilhança* presente nas distribuições da família T-Student como se vê, por exemplo, em Faria (2011) para o caso simétrico. Por fim, usamos como valor inicial para o hiper-parâmetro $\nu^{(0)} = 2,01$ lembrando que um vetor $\mathbf{Y}$ com distribuição T-Student assimétrica possui variância quando $\nu > 2$.

- Recuperação dos parâmetros

Tabela 1 – Recuperação dos parâmetros no modelo T-Student assimétrico

(a) Algoritmo padrão

	n=50			n=200			n=1000		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
β_0(25)	24,38421	4,47325	4,05555	25,11404	2,27766	2,24251	24,96119	0,96562	0,95755
β_1(-10)	-9,98889	0,14133	0,12491	-10,00339	0,06610	0,06713	-9,99924	0,02843	0,02944
β_2(5)	5,01319	0,23802	0,21807	4,99466	0,12442	0,12136	5,00310	0,05454	0,05311
σ(2,8)	2,56906	0,66494	0,83797	2,85490	0,41958	0,44127	2,82572	0,19158	0,19498
λ(-1,4)	-1,62441	1,55571	1,43331	-1,54692	0,63343	0,58145	-1,44792	0,24028	0,23797
ν(4,5)	4,63705	2,38688	4,62185	5,36797	2,23262	2,52439	4,70656	0,81217	0,75283

(b) Algoritmo modificado

	n=50			n=200			n=1000		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
β_0(25)	24,38421	4,47325	4,05555	25,11404	2,27766	2,24251	24,96119	0,96562	0,95755
β_1(-10)	-9,98889	0,14133	0,12491	-10,00339	0,06610	0,06713	-9,99924	0,02843	0,02944
β_2(5)	5,01319	0,23802	0,21807	4,99466	0,12442	0,12136	5,00310	0,05454	0,05311
σ(2,8)	2,56906	0,66494	0,83797	2,85491	0,41958	0,44127	2,82572	0,19158	0,19498
λ(-1,4)	-1,62441	1,55572	1,43331	-1,54692	0,63343	0,58145	-1,44792	0,24028	0,23797
ν(4,5)	4,63706	2,38689	4,62186	5,36797	2,23263	2,52440	4,70656	0,81218	0,75283

De acordo com a Tabela 1, os resultados das estimativas médias e sua variabilidade são praticamente os mesmos utilizando ambos os métodos com algumas diferenças na quinta casa decimal. Essa quase igualdade também pode ser percebida na Tabela 2, o que garante nesse caso a consistência de ambos os algoritmos.

- Consistência da estimação

Tabela 2 – Consistência no modelo T-Student assimétrico

(a) Algoritmo padrão

n=50				n=200				n=1000			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,61579	4,51544	0,60680	3,22008	0,11404	2,28051	0,11433	1,62447	0,03881	0,96640	0,02011	0,61127
0,01111	0,14177	0,00973	0,09478	0,00339	0,06619	0,00306	0,04648	0,00076	0,02844	0,00096	0,01916
0,01319	0,23838	0,02239	0,16694	0,00534	0,12453	0,00433	0,08938	0,00310	0,05463	0,00027	0,03575
0,23094	0,70390	0,30104	0,45655	0,05490	0,42316	0,04081	0,29380	0,02572	0,19330	0,01764	0,12037
0,22441	1,57181	0,04868	0,69379	0,14692	0,65024	0,07057	0,34985	0,04792	0,24501	0,03997	0,16970
0,13705	2,39081	0,59468	1,35162	0,86797	2,39540	0,35474	1,15144	0,20656	0,83803	0,08883	0,44564

(b) Algoritmo modificado

n=50				n=200				n=1000			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,61579	4,51544	0,60681	3,22007	0,11404	2,28051	0,11433	1,62447	0,03881	0,96640	0,02010	0,61127
0,01111	0,14177	0,00973	0,09478	0,00339	0,06619	0,00306	0,04648	0,00076	0,02844	0,00096	0,01916
0,01319	0,23838	0,02239	0,16694	0,00534	0,12453	0,00433	0,08938	0,00310	0,05463	0,00027	0,03575
0,23094	0,70390	0,30104	0,45655	0,05491	0,42316	0,04082	0,29380	0,02572	0,19330	0,01763	0,12037
0,22441	1,57182	0,04869	0,69379	0,14692	0,65025	0,07058	0,34985	0,04792	0,24501	0,03998	0,16970
0,13706	2,39082	0,59466	1,35163	0,86797	2,39541	0,35475	1,15144	0,20656	0,83804	0,08884	0,44564

Nos gráficos das Figuras 5 e 6, vemos que a eficiência mediana de ambos os métodos na estimação de cada parâmetro é também praticamente a mesma. Sendo a mediana uma medida mais resistente do que a média, observamos uma tendência mais comportada nos resultados em relação às medidas correspondentes baseadas na média.

A vantagem da utilização do segundo método está no ganho de tempo até certo ponto decorrente da realização de menos iterações como evidenciado nos boxplots das Figuras 7 e 8. Eles foram traçados sem os pontos discrepantes segundo o critério de Tukey para facilitar a visualização.

Figura 5 – Viés mediano por parâmetro para a T-Student assimétrica univariada

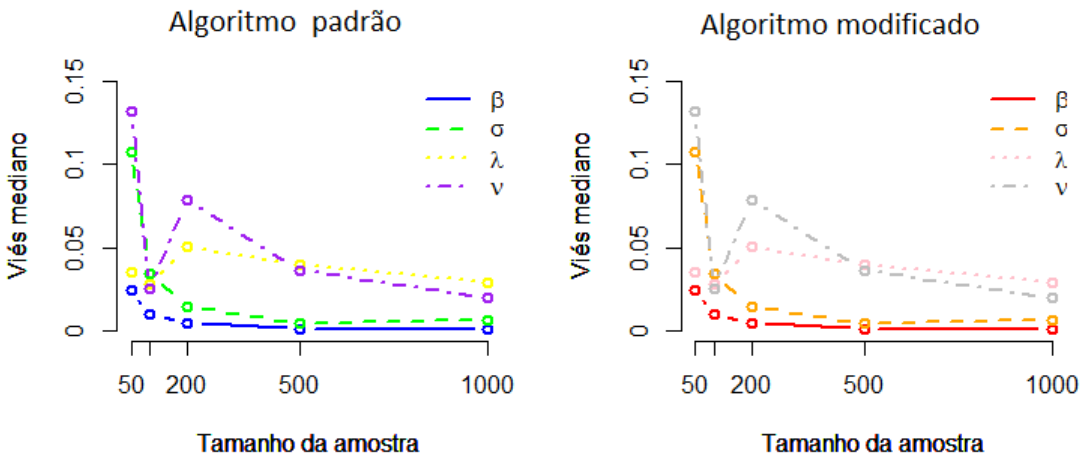
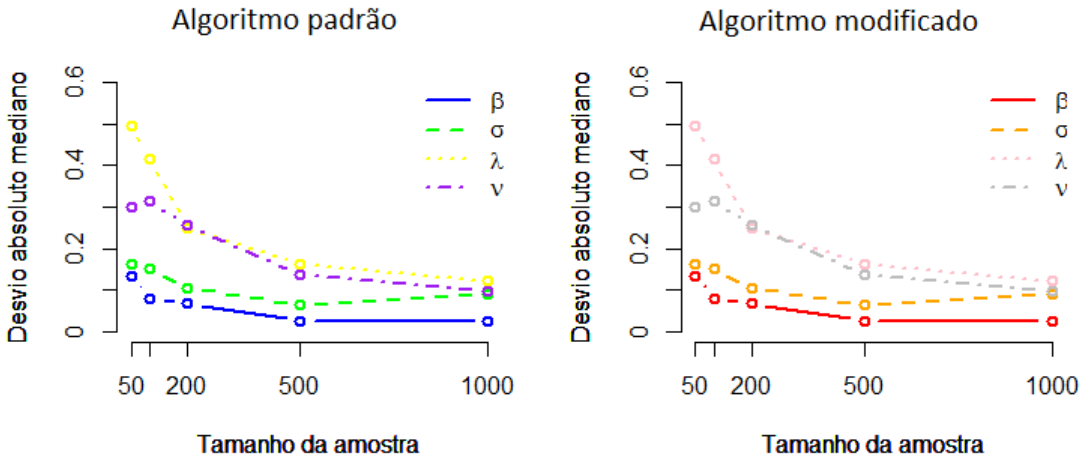


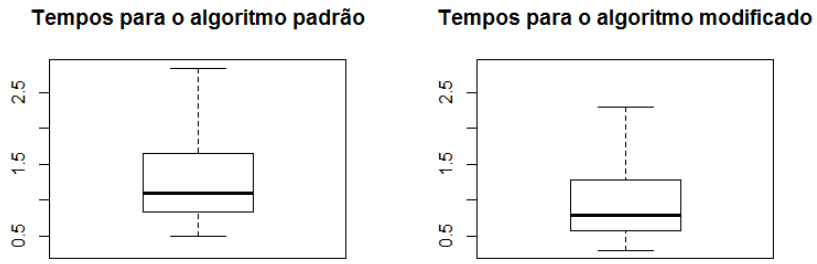
Figura 6 – Desvio absoluto mediano (DAM) por parâmetro para a T-Student assimétrica univariada



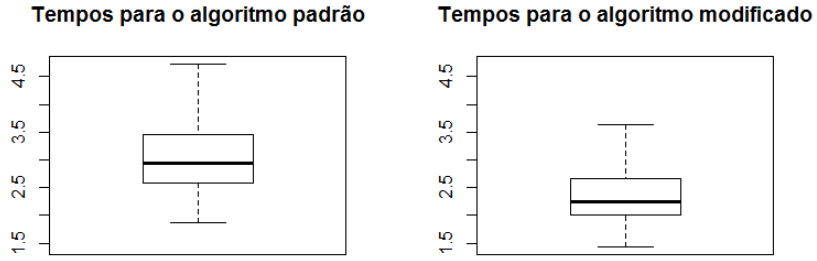
- Desempenho dos algoritmos

Figura 7 – Boxplots de tempos para a T-Student assimétrica univariada

(a) Para n=50



(b) Para n=200



(c) Para n=1000

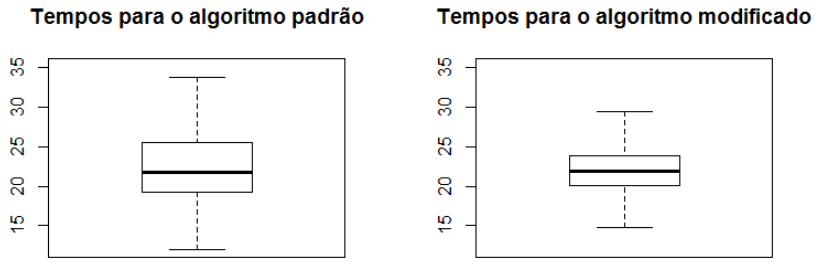
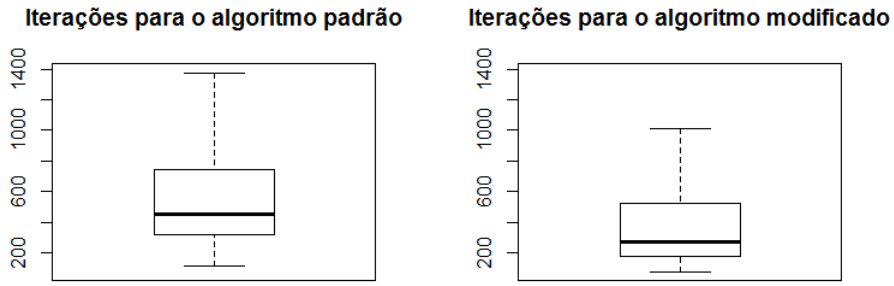
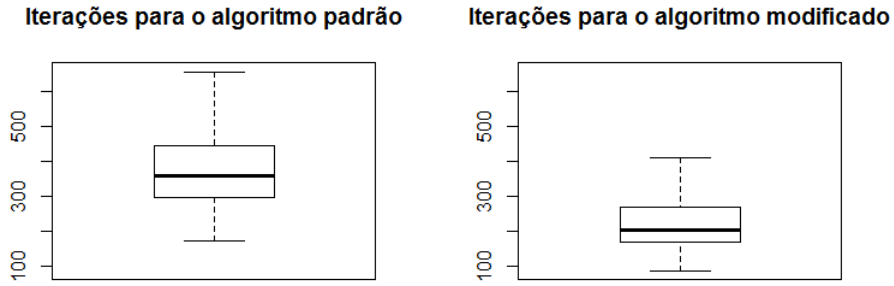


Figura 8 – Boxplots de iterações para a T-Student assimétrica univariada

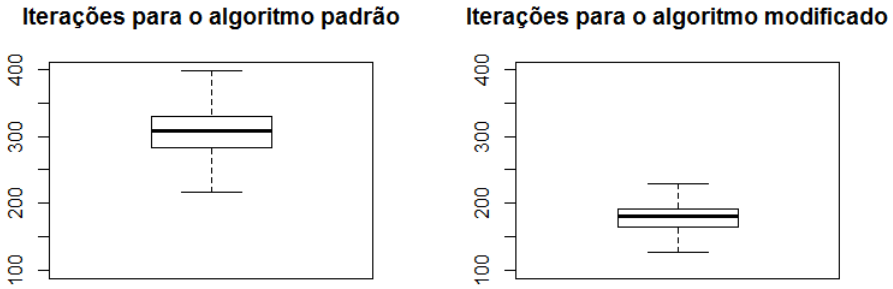
(a) Para n=50



(b) Para n=200



(c) Para n=1000



Slash assimétrica

Na distribuição Slash assimétrica $h(u;\nu) = \nu u^{\nu-1}$, donde temos

$$\frac{\partial h}{\partial \nu}(u;\nu) = \left(\frac{1}{\nu} + \ln u\right) h(u;\nu);$$

$$\frac{\partial^2 h}{\partial \nu^2}(u;\nu) = \left[\frac{2}{\nu}\ln u + (\ln u)^2\right] h(u;\nu).$$

Com o uso das derivadas anteriores, podemos escrever o seguinte:

$$J_i^G(r) = \frac{1}{\nu} I_i^G(r,0) + I_i^G(r,1);$$

$$L_i^G(r) = \frac{2}{\nu} I_i^G(r,1) + I_i^G(r,2).$$

Escrevemos em geral $I_i^G(r,s) = \sqrt{2\pi}\nu \int_0^{+\infty} u^{\nu-1+r}(\ln u)^s \phi(\sqrt{u d_i}) G(\sqrt{u} A_i) du$, cujo cálculo realizaremos através da função `integrate` do `R`. Em particular, de acordo com Zeller (2009), temos as seguintes expressões:

$$I_i^{\Phi}(r,0) = \left(\frac{2}{d_i}\right)^{\nu+r} \nu\Gamma(\nu+r) P\left(1;\nu+r,\frac{d_i}{2}\right) E\left(\Phi\left(S_i^{\frac{1}{2}} A_i\right)\right);$$

$$I_i^{\phi}(r,0) = \frac{2^{\nu+r}\nu}{\sqrt{2\pi}} \frac{\Gamma(\nu+r)^{\nu+r}}{d_i + A_i^2} P\left(1;\nu+r,\frac{d_i + A_i^2}{2}\right).$$

Nas identidades acima, $P(1;\nu_1,\nu_2)$ é a f.d.a. no ponto 1 de uma distribuição gama com parâmetros de forma ν_1, ν_2 e $S_i \sim$ TGama$_{(0,1)}\left(\nu + \frac{3}{2}, \frac{d_i}{2}\right)$. Da Proposição 3.1.4, vê-se em Zeller (2009) os seguintes resultados para os valores esperados condicionais do passo E na etapa k do algoritmo EM:

$$\widehat{u_i}^{(k)} = \frac{2}{\widehat{d_i}^{(k)}}(2\widehat{\nu}^{(k)}+1)\frac{f_0(y_i)}{f(y_i)} \frac{P\left(1;\widehat{\nu}^{(k)}+\frac{3}{2},\frac{\widehat{d_i}^{(k)}}{2}\right)}{P\left(1;\widehat{\nu}^{(k)}+\frac{1}{2},\frac{\widehat{d_i}^{(k)}}{2}\right)} E\left(\Phi\left(S_i^{\frac{1}{2}}\widehat{A_i}^{(k)}\right)\right);$$

$$\widehat{z_i}^{(k)} = \frac{2}{\sqrt{\pi}} \frac{\Gamma(\widehat{\nu}^{(k)}+1)}{\Gamma\left(\widehat{\nu}^{(k)}+\frac{1}{2}\right)} \frac{f_0(y_i)}{f(y_i)} \frac{\widehat{d_i}^{(k)^{\widehat{\nu}^{(k)}+\frac{1}{2}}}}{\left(\widehat{d_i}^{(k)} + \widehat{A_i}^{(k)^2}\right)^{\widehat{\nu}^{(k)}+1}} \frac{P\left(1;\widehat{\nu}^{(k)}+1,\frac{\widehat{d_i}^{(k)}+\widehat{A_i}^{(k)^2}}{2}\right)}{P\left(1;\widehat{\nu}^{(k)}+\frac{1}{2},\frac{\widehat{d_i}^{(k)}}{2}\right)}.$$

Na expressão, f é a f.d.p. da Slash assimétrica (3.9) trocando μ_i por $\mathbf{x}_i^T\boldsymbol{\beta}$ e $f_0(y_i) = \nu\int_0^1 u^{\nu-1}\phi_1\left(y_i; \mathbf{x}_i^T\boldsymbol{\beta}, \frac{\sigma^2}{u}\right) du = \frac{\nu\Gamma(\nu+\frac{1}{2})}{\sigma\sqrt{2\pi}}\left(\frac{2}{d_i}\right)^{\nu+\frac{1}{2}} P\left(1; \nu+\frac{1}{2}, \frac{d_i}{2}\right)$ com $d_i = \frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}$ corresponde à f.d.p. da Slash simétrica.

Para esta distribuição, vale a mesma observação feita no exemplo da T-Student assimétrica com relação à limitação superior do hiper-parâmetro ν por razão inteiramente similar. Neste caso, tomamos $\nu^{(0)} = 1,01$ como valor inicial para o hiper-parâmetro uma vez que um vetor $\mathbf{Y}$ com distribuição Slash assimétrica possui variância para $\nu > 1$.

- Recuperação dos parâmetros

Tabela 3 – Recuperação dos parâmetros no modelo Slash assimétrico

(a) Algoritmo padrão

	n=50			n=200			n=1000		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
β_0(25)	23,88240	4,74688	4,40311	24,57737	2,19972	2,24830	24,98809	1,01248	0,97356
β_1(-10)	-10,00595	0,14490	0,12978	-9,99732	0,06568	0,06168	-9,99856	0,02759	0,02874
β_2(5)	5,01658	0,25006	0,22277	5,00561	0,11652	0,11830	4,99683	0,05434	0,05224
σ(2,8)	2,13539	0,60757	1,00083	2,51849	0,38890	0,57524	2,74106	0,21975	0,28122
λ(-1,4)	-1,08243	1,45274	1,44523	-1,23932	0,64106	0,67320	-1,36589	0,25871	0,27096
ν(4,5)	2,36740	1,12127	3,66539	3,50819	1,40552	3,64630	4,57277	1,54822	2,88697

(b) Algoritmo modificado

	n=50			n=200			n=1000		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
β_0(25)	23,88241	4,74688	4,40310	24,57738	2,19973	2,24829	24,98809	1,01248	0,97356
β_1(-10)	-10,00595	0,14490	0,12978	-9,99732	0,06568	0,06168	-9,99856	0,02759	0,02874
β_2(5)	5,01658	0,25006	0,22277	5,00561	0,11652	0,11830	4,99683	0,05434	0,05224
σ(2,8)	2,13540	0,60757	1,00085	2,51850	0,38890	0,57524	2,74106	0,21974	0,28122
λ(-1,4)	-1,08243	1,45275	1,44522	-1,23933	0,64107	0,67319	-1,36589	0,25871	0,27096
ν(4,5)	2,36742	1,12128	3,66548	3,50821	1,40553	3,64636	4,57278	1,54822	2,88700

Vemos na Tabela 3 que as estimativas por ambos os métodos são praticamente as mesmas. Cabe destacar ainda o curioso e não desejável aumento do desvio padrão das estimativas do parâmetro ν à medida que n cresce. Esse fato mostra que o modelo Slash assimétrico exige muita

informação para recuperar o hiper-parâmetro e tal convergência é bastante lenta, mas vale ressaltar que o coeficiente de variação (razão entre o desvio padrão e a média) das respectivas estimativas diminui quando n cresce, indicando que ao menos a variabilidade relativa das estimativas tem um comportamento dentro do esperado. Quanto à consistência das estimativas, expressa em termos de medidas de tendência central e variabilidade baseadas na média e na mediana, também podemos dizer que não há diferenças significativas entre os dois algoritmos como se vê na Tabela 4.

- Consistência da estimação

Tabela 4 – Consistência no modelo Slash assimétrico

(a) Algoritmo padrão

n=50				n=200				n=1000			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
1,11760	4,87667	1,03615	3,38013	0,42263	2,23996	0,50058	1,38766	0,01192	1,01255	0,05203	0,67412
0,00595	0,14502	0,00292	0,09761	0,00268	0,06573	0,00144	0,04320	0,00144	0,02762	0,00083	0,01800
0,01658	0,25061	0,00833	0,17680	0,00560	0,11665	0,00746	0,07417	0,00317	0,05443	0,00453	0,03588
0,66461	0,90047	0,77964	0,41477	0,28151	0,48009	0,28343	0,31219	0,05894	0,22751	0,05021	0,14444
0,31757	1,48705	0,49409	0,83388	0,16068	0,66089	0,09891	0,39069	0,03411	0,26095	0,02574	0,16960
2,13260	2,40940	2,40619	0,63893	0,99181	1,72023	1,40555	0,75006	0,07276	1,54992	0,33427	0,75424

(b) Algoritmo modificado

n=50				n=200				n=1000			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
1,11759	4,87666	1,03613	3,38012	0,42261	2,23996	0,50057	1,38764	0,01191	1,01255	0,05204	0,67413
0,00595	0,14502	0,00029	0,09761	0,00268	0,06573	0,00144	0,04320	0,00144	0,02762	0,00083	0,01800
0,01658	0,25061	0,00833	0,17680	0,00560	0,11665	0,00746	0,07417	0,00317	0,05443	0,00453	0,03588
0,66460	0,90046	0,77963	0,41478	0,28150	0,48009	0,28342	0,31218	0,05894	0,22751	0,05021	0,14444
0,31757	1,48706	0,49406	0,83390	0,16068	0,66090	0,09890	0,39069	0,03411	0,26094	0,02574	0,16960
2,13258	2,40939	2,40619	0,63892	0,99179	1,72022	1,40554	0,75007	0,07278	1,54993	0,33425	0,75424

Os gráficos das Figuras 9 e 10 mostram que o comportamento mediano das estimativas em termos de tendência central e variabilidade também está dentro do esperado para cada grupo de parâmetros, exceto possivelmente pelo aumento do desvio absoluto mediano relativo para o parâmetro ν. Já as Figuras 11 e 12 revelam que os ganhos do algoritmo modificado são bastante consideráveis nesta distribuição.

Figura 9 – Viés mediano por parâmetro para a Slash assimétrica univariada

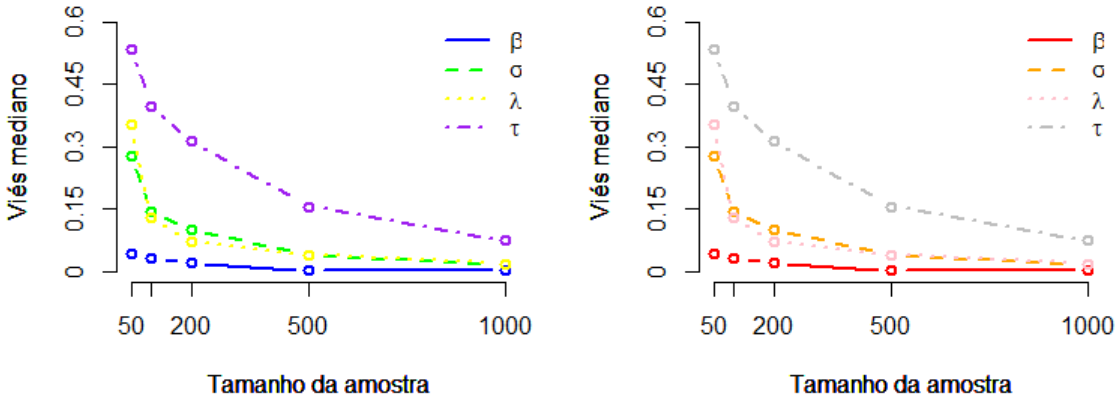
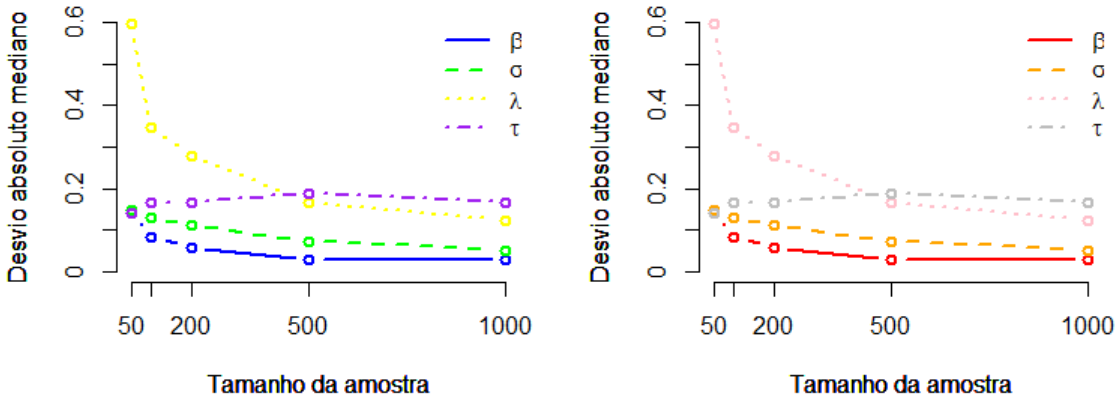


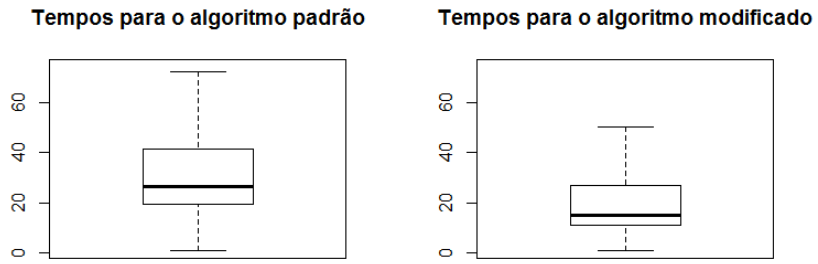
Figura 10 – Desvio absoluto mediano (MAD) por parâmetro para a Slash assimétrica univariada



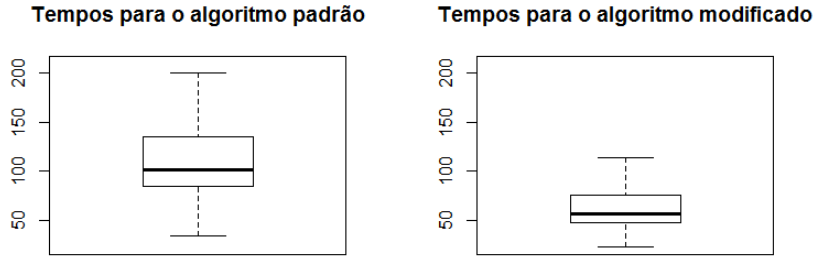
- Desempenho dos algoritmos

Figura 11 – Boxplots de tempos para a Slash assimétrica univariada

(a) Para n=50



(b) Para n=200



(c) Para n=1000

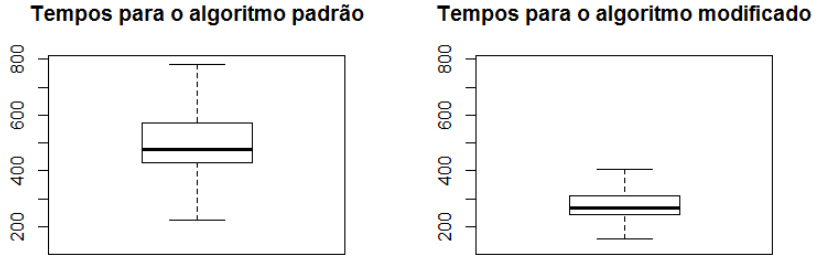
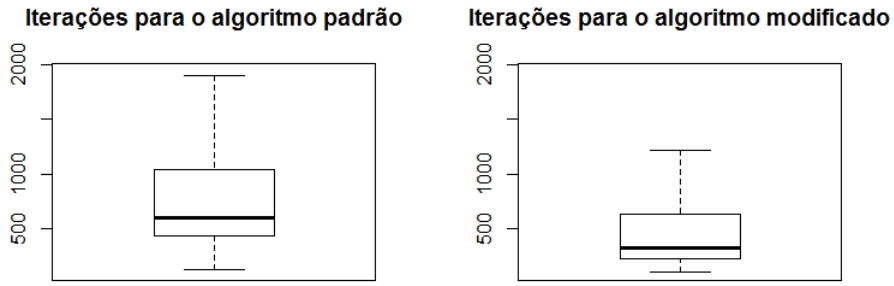
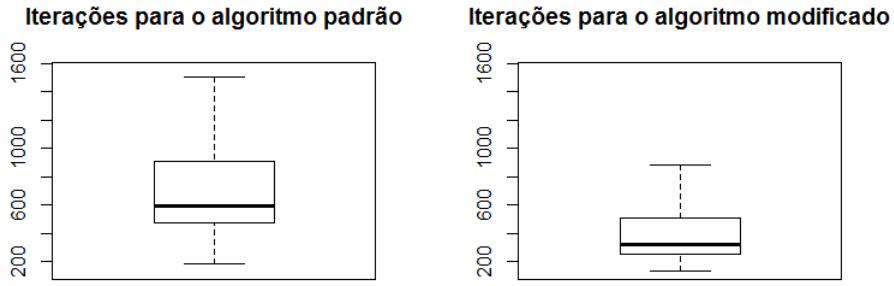


Figura 12 – Boxplots de iterações para a Slash assimétrica univariada

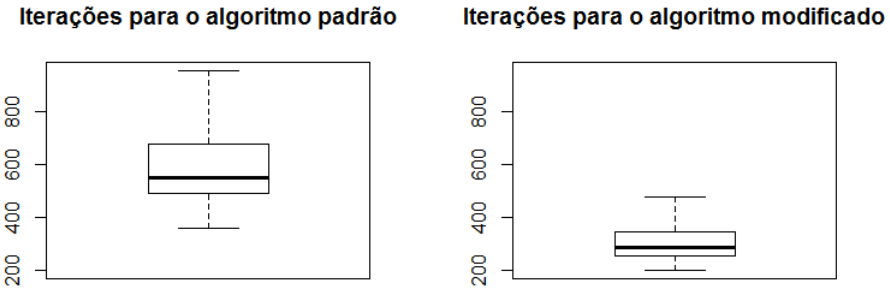
(a) Para n=50



(b) Para n=200



(c) Para n=1000



Normal assimétrica contaminada

Já mencionamos que na distribuição normal assimétrica contaminada vista como mistura de escala, a densidade $h(u;\gamma,\nu)$ é dada por

$$h(u;\gamma,\nu)=\begin{cases}\nu & ;\text{ se } u=\gamma\\ 1-\nu & ;\text{ se } u=1\\ 0 & ;\text{ caso contrário.}\end{cases}$$

Mediante a adoção de uma notação similar à dos dois exemplos anteriores, temos também que

$$I_i^{\Phi}(r,0)=\sqrt{2\pi}\left[\nu\gamma^r\phi(\sqrt{\gamma d_i})\Phi(\sqrt{\gamma}A_i)+(1-\nu)\phi(\sqrt{d_i})\Phi(A_i)\right];$$

$$I_i^{\phi}(r,0)=\sqrt{2\pi}\left[\nu\gamma^r\phi(\sqrt{\gamma d_i})\phi(\sqrt{\gamma}A_i)+(1-\nu)\phi(\sqrt{d_i})\phi(A_i)\right].$$

Note que não faz sentido considerar aqui, como nos exemplos anteriores, os casos em que $s\neq 0$, de modo que podemos escrever $I_i^G(r):=I_i^G(r,0)$. Além disso, observamos anteriormente que não existem análogos diretos para $J_i^G(r)$ e $L_i^G(r)$ neste exemplo devido ao fato de a função de probabilidade h ser discreta. Dessa forma, tivemos de definir novas expressões $J_{1i}^G(r)$ e $J_{2i}^G(r)$, as quais foram já apresentadas e utilizadas no cômputo das derivadas de primeira e segunda ordens de ℓ.

Novamente em Zeller (2009), foram obtidos com o uso da Proposição 3.1.4 os resultados adiante sobre os valores esperados condicionais do passo E na etapa k do algoritmo EM:

$$\widehat{u_i}^{(k)}=\frac{\widehat{\nu}^{(k)}\widehat{\gamma}^{(k)}\phi\left(y_i;\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)},\frac{\widehat{\sigma}^{(k)^2}}{\widehat{\gamma}^{(k)}}\right)\Phi\left(\sqrt{\widehat{\gamma}^{(k)}}\widehat{A_i}^{(k)}\right)+(1-\widehat{\nu}^{(k)})\phi\left(y_i;\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)},\widehat{\sigma}^{(k)^2}\right)\Phi\left(\widehat{A_i}^{(k)}\right)}{\widehat{\nu}^{(k)}\phi\left(y_i;\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)},\frac{\widehat{\sigma}^{(k)^2}}{\widehat{\gamma}^{(k)}}\right)\Phi\left(\sqrt{\widehat{\gamma}^{(k)}}\widehat{A_i}^{(k)}\right)+(1-\widehat{\nu}^{(k)})\phi\left(y_i;\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)},\widehat{\sigma}^{(k)^2}\right)\Phi\left(\widehat{A_i}^{(k)}\right)};$$

$$\widehat{z_i}^{(k)}=\frac{\widehat{\nu}^{(k)}\sqrt{\widehat{\gamma}^{(k)}}\phi\left(y_i;\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)},\frac{\widehat{\sigma}^{(k)^2}}{\widehat{\gamma}^{(k)}}\right)\phi\left(\sqrt{\widehat{\gamma}^{(k)}}\widehat{A_i}^{(k)}\right)+(1-\widehat{\nu}^{(k)})\phi\left(y_i;\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)},\widehat{\sigma}^{(k)^2}\right)\phi\left(\widehat{A_i}^{(k)}\right)}{\widehat{\nu}^{(k)}\phi\left(y_i;\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)},\frac{\widehat{\sigma}^{(k)^2}}{\widehat{\gamma}^{(k)}}\right)\Phi\left(\sqrt{\widehat{\gamma}^{(k)}}\widehat{A_i}^{(k)}\right)+(1-\widehat{\nu}^{(k)})\phi\left(y_i;\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{(k)},\widehat{\sigma}^{(k)^2}\right)\Phi\left(\widehat{A_i}^{(k)}\right)}.$$

Como não há restrições para a existência dos momentos de um vetor aleatório $\mathbf{Y}$ com distribuição normal assimétrica contaminada, tomaremos como valores iniciais para o hiper-parâmetro arbitrariamente $(\nu^{(0)},\gamma^{(0)})=(0,5,\ 0,5)$.

- Recuperação dos parâmetros

Tabela 5 – Recuperação dos parâmetros no modelo normal contaminado assimétrico

(a) Algoritmo padrão

	n=50			n=200			n=1000		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
β_0(25)	24,12641	7,92910	6,51659	25,02522	3,12706	3,08755	24,87325	1,51121	1,47801
β_1(-10)	-9,99614	0,14133	0,18880	-10,00243	0,10303	0,10285	-9,99816	0,04723	0,04472
β_2(5)	5,00358	0,42188	0,34348	4,99904	0,17135	0,16803	5,00494	0,08229	0,08107
σ(2,8)	2,97064	1,03469	1,26852	3,17514	0,87471	0,94348	2,84773	0,42982	0,44451
λ(-1,4)	-1,25825	1,10173	1,24364	-1,49526	0,54481	0,56617	-1,40819	0,21218	0,22100
γ(0,2)	0,18719	0,10410	0,19722	0,23040	0,09315	0,12434	0,204245	0,04618	0,04886
ν(0,65)	0,40291	0,19666	0,32130	0,53067	0,17481	0,23606	0,62744	0,08747	0,08997

(b) Algoritmo modificado

	n=50			n=200			n=1000		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
β_0(25)	24,12512	7,93616	6,50963	25,02522	3,12706	3,08755	24,87323	1,51121	1,47801
β_1(-10)	-9,99654	0,21771	0,18859	-10,00243	0,10303	0,10285	-9,99816	0,04723	0,04472
β_2(5)	5,00440	0,42119	0,34288	4,99904	0,17135	0,16803	5,00494	0,08229	0,08107
σ(2,8)	2,97999	1,03663	1,261023	3,17516	0,87473	0,94348	2,84773	0,42982	0,44451
λ(-1,4)	-1,26345	1,10118	1,23888	-1,49526	0,54482	0,56617	-1,40818	0,21218	0,22100
γ(0,2)	0,18821	0,10482	0,19547	0,23040	0,09316	0,12434	0,204244	0,04618	0,04886
ν(0,65)	0,40225	0,19666	0,31865	0,53066	0,17482	0,23607	0,62744	0,08747	0,08997

De acordo com a Tabela 5, mais uma vez obtivemos resultados muito próximos nas estimativas por ambos os métodos, o que fica cada vez mais evidente à medida que o tamanho da amostra aumenta.

Quanto às evidências de consistência, também já esperadas em ambos os algoritmos, podemos constatar pela observação dos dados da Tabela 6 que praticamente todos os resultados estão dentro do esperado.

- Consistência da estimação

Tabela 6 – Consistência no modelo normal contaminado assimétrico

(a) Algoritmo padrão

n=50				n=200				n=1000			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,87359	7,97708	0,77200	5,32022	0,02522	3,12716	0,11661	1,91301	0,12675	1,51651	0,09672	0,98813
0,00386	0,21840	0,00687	0,14640	0,00243	0,10305	0,00061	0,06401	0,00184	0,04726	0,00060	0,03046
0,00358	0,42190	0,02599	0,28291	0,00096	0,17135	0,00163	0,10765	0,0049	0,08244	0,00191	0,05683
0,17064	1,04866	0,00832	0,65826	0,37514	0,95176	0,16897	0,50284	0,04773	0,43247	0,00206	0,29504
0,14176	1,11082	0,33390	0,57082	0,09526	0,55308	0,01971	0,34414	0,00819	0,21234	0,00217	0,13363
0,01281	0,10489	0,04160	0,05274	0,03040	0,09799	0,00570	0,04558	0,00425	0,04638	0,00279	0,03133
0,24709	0,31580	0,24043	0,14692	0,11933	0,21166	0,08481	0,10810	0,02256	0,09033	0,01847	0,05138

(b) Algoritmo modificado

n=50				n=200				n=1000			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,87488	7,98423	0,77201	5,29144	0,02522	3,12716	0,11660	1,91303	0,12677	1,51652	0,09674	0,98814
0,00346	0,21774	0,66173	0,14640	0,00243	0,10305	0,00061	0,06401	0,00184	0,04726	0,00060	0,03046
0,00440	0,42122	0,02599	0,28236	0,00096	0,17135	0,00163	0,10765	0,0049	0,08244	0,00191	0,05683
0,17999	1,05214	0,00131	0,66057	0,37516	0,95179	0,16895	0,50282	0,04773	0,43247	0,00204	0,29505
0,13655	1,10862	0,33125	0,57159	0,09526	0,55308	0,01970	0,34413	0,00818	0,21234	0,00218	0,13362
0,01179	0,10548	0,04112	0,05322	0,03040	0,09799	0,00570	0,04558	0,00424	0,04638	0,00279	0,03133
0,24775	0,31637	0,24315	0,14951	0,11934	0,21166	0,08482	0,10810	0,02256	0,09033	0,01847	0,05138

As Figuras 13 e 14 complementam as informações da Tabela 6. Elas acrescentam a informação por parâmetro para as medidas baseadas na mediana e revelam que, embora haja algumas pequenas incoerências (especialmente no viés mediano do parâmetro de escala), ambos os algoritmos levam igualmente a estimativas que cumprem os requisitos da teoria assintótica.

Por outro lado, neste exemplo específico os "ganhos" em número de iterações e tempo com a modificação proposta são bem mais modestos e basicamente ocorrem para amostras pequenas. Isso pode ser constatado pela observação das Figuras 15 e 16.

Figura 13 – Viés mediano por parâmetro para a normal assimétrica contaminada univariada

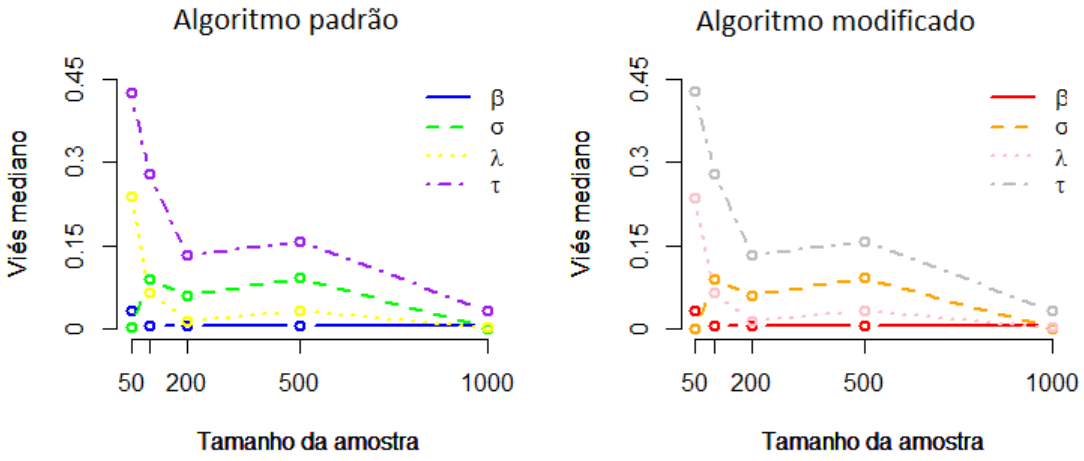
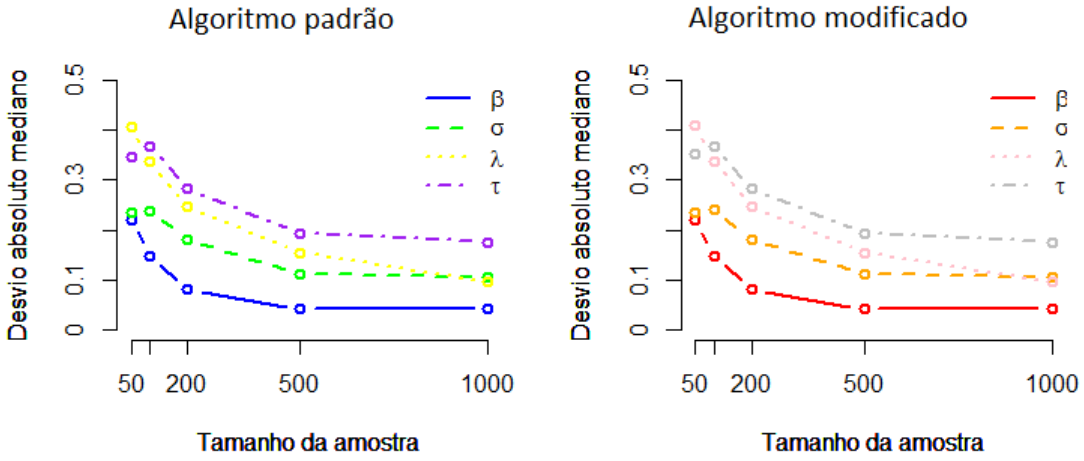


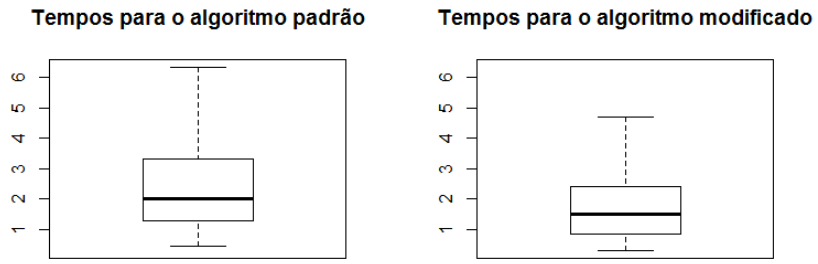
Figura 14 – Desvio absoluto mediano (DAM) por parâmetro para a normal contaminada assimétrica univariada



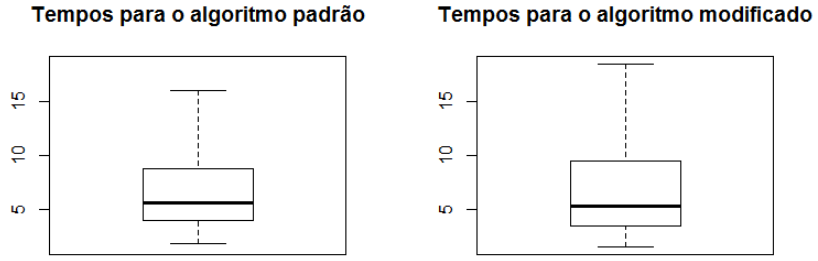
- Desempenho dos algoritmos

Figura 15 – Boxplots de tempos para a normal contaminada assimetrica univariada

(a) Para n=50



(b) Para n=200



(c) Para n=1000

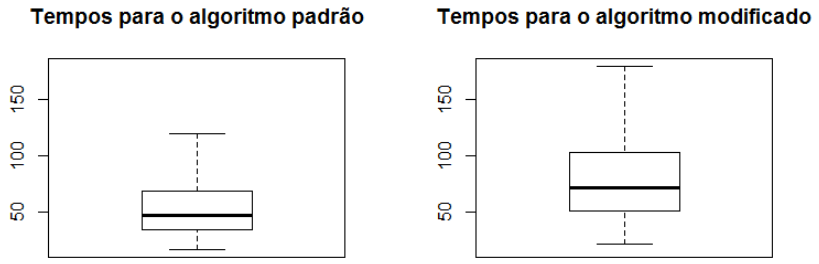
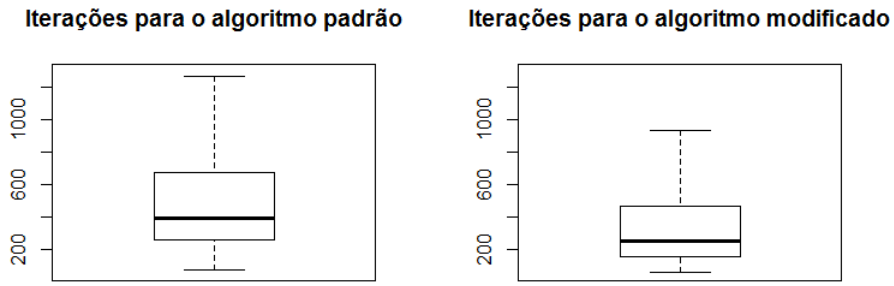
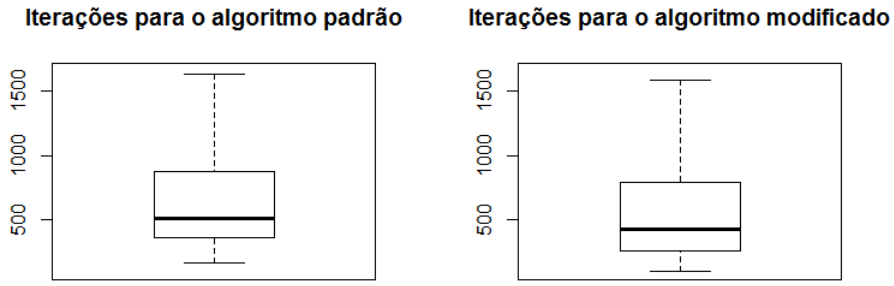


Figura 16 – Boxplots de iterações para a normal assimétrica contaminada univariada

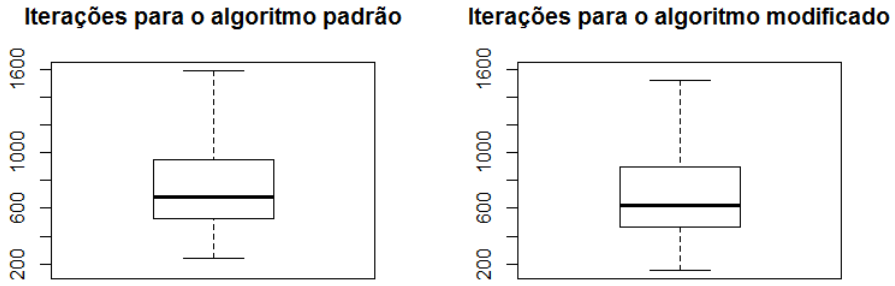
(a) Para n=50



(b) Para n=200



(c) Para n=1000



3.2 MODELOS MISTURAS MULTIVARIADOS COM REGRESSÃO

Como já mencionado, discutiremos nesta seção os modelos multivariados com a abordagem de Ferreira, Lachos & Bolfarine (2016), incluindo a regressão linear múltipla multivariada nas formas alternativas que descrevemos na Seção 2.4.

Inicialmente, apresentaremos os principais exemplos desse tipo de mistura, cujos resultados estão expressos na Subseção 2.3.2, considerando também um caso excepcional que pode ser enquadrado tanto como o tipo de mistura apresentado na referida subseção quanto como uma *mistura finita de normais assimétricas*.

Ademais, trataremos a estimação dos parâmetros em modelos de regressão linear múltipla multivariada com erros que seguem distribuições misturas de escala assimétricas de normais utilizando técnicas diferentes das que se encontra normalmente na literatura para esse tipo de modelo.

Basicamente, vamos estender duas técnicas utilizadas por Lange & Sinsheimer (1993) nas misturas de escala simétricas de normais univariadas para o caso assimétrico multivariado, além de apresentar uma nova técnica para um dos exemplos indicados. Ressaltamos que técnicas similares aparentemente não funcionam bem nas distribuições misturas de escala de normais assimétricas, fato sobre o qual discorremos no capítulo final.

Por fim, faremos um estudo de simulação para comparar os métodos usuais e os propostos com a finalidade de mostrar os ganhos provenientes da adoção desse último nos exemplos básicos citados, além de realizar um estudo com dados reais para ilustrar como funciona a nova sugestão na prática.

3.2.1 **Exemplos Básicos**

A distribuição normal assimétrica multivariada, definida na Subseção 2.3.1, pode ser vista como uma mistura de escala assimétrica degenerada conforme observado no início da Subseção 2.3.2. Além desse exemplo direto que não reproduziremos novamente, descreveremos nesta subseção sucintamente mais três exemplos das misturas multivariadas: T-Student normal assimétrica, Slash normal assimétrica e normal contaminada assimétrica.

Distribuição T-Student normal assimétrica multivariada

O principal exemplo de mistura de escala assimétrica da família normal é a chamada distribuição T-Student normal assimétrica multivariada. Nessa distribuição, temos assim como na T-Student assimétrica univariada da Subseção 3.1.1 que $\boldsymbol{\tau} = \nu$ numérico e $U \sim$ Gama $\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, donde $h(u;\nu) = \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} u^{\nu/2-1} e^{-u\nu/2}$. Dessa forma, um vetor aleatório p-dimensional com tal distribuição é escrito na forma $\mathbf{Y} \sim$ STN$_p\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\Big)$ e sua f.d.p. é dada por

$$f\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\Big) = 2t_p\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\Big) \Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu})\right), \ \mathbf{y} \in \mathbb{R}^p. \tag{3.15}$$

Em (3.15), $t_p\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\Big) = \dfrac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\pi\nu)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left[1 + \dfrac{(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+p}{2}}$ é a expressão da f.d.p. de uma T-Student p-variada com ν graus de liberdade, a qual é uma mistura de escala normal de acordo com Lange & Sinsheimer (1993).

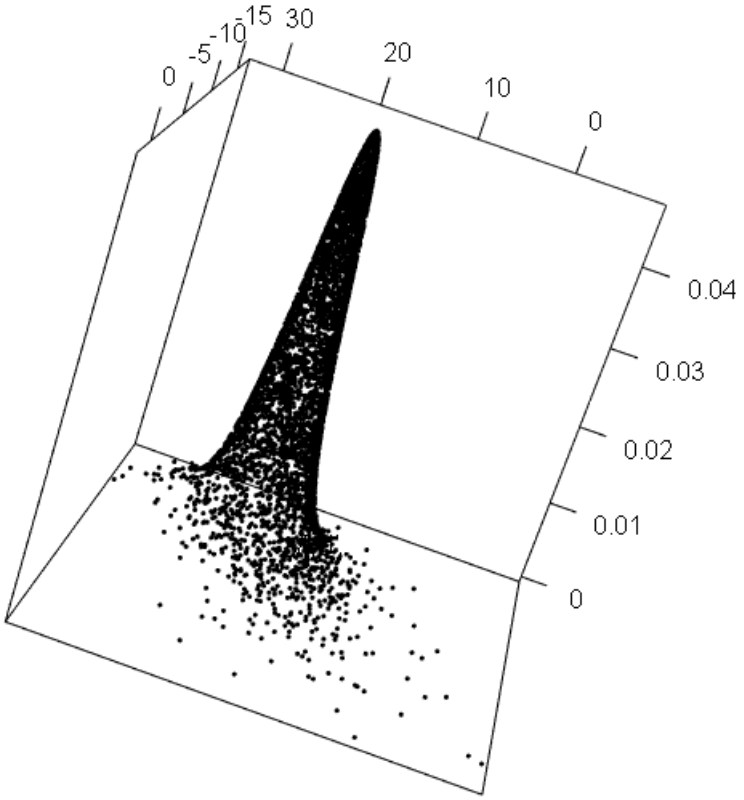
Lembrando os momentos de U obtidos no exemplo univariado, já vimos que $E(U^{-1}) = \frac{\Gamma\left(\frac{\nu}{2}-1\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{\nu}{2} = \frac{\nu}{\nu-2}$, o que combinado à Proposição 2.3.2 nos dá

$$\begin{aligned} E(\mathbf{Y}) &= \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}} \varpi \boldsymbol{\Sigma}^{1/2} \boldsymbol{\lambda}, \ \nu > 1; \\ Var(\mathbf{Y}) &= \boldsymbol{\Sigma}^{1/2} \left(\frac{\nu}{\nu-2} \mathbf{I}_p - \frac{2}{\pi} \varpi^2 \boldsymbol{\lambda}\boldsymbol{\lambda}^T\right) \boldsymbol{\Sigma}^{1/2}, \ \nu > 2. \end{aligned} \tag{3.16}$$

A expressão de $\varpi = E\left([U(U + \boldsymbol{\lambda}^T\boldsymbol{\lambda})]^{-\frac{1}{2}}\right)$ acima não possui forma fechada e é, nesse caso, dada por $\varpi = \dfrac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} \displaystyle\int_0^{+\infty} \frac{u^{\frac{\nu-3}{2}} e^{-\frac{\nu}{2}u}}{\sqrt{u + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}} du$, onde se pode verificar que a integral imprópria converge para $\nu > 1$.

Na Figura 17, vemos o gráfico da f.d.p. de uma distribuição T-Student normal assimétrica com $p = 2$ e parâmetros $\boldsymbol{\mu} = (20, -15)$, $\mathbf{B} = \boldsymbol{\Sigma}^{1/2} = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}$, $\boldsymbol{\lambda} = (-1, 2)$ e $\nu = 5$.

Figura 17 – Gráfico da f.d.p. de uma T-Student normal assimétrica multivariada



Tal gráfico foi esboçado com base numa amostra de 5.000 observações da distribuição em questão no `R`. A própria escala do gráfico permite notar o efeito da assimetria negativa na primeira coordenada e da assimetria positiva na segunda. Já a presença de pontos distantes da massa central revelam o efeito do hiper-parâmetro de conferir mais peso caudal. Portanto, valem em geral os mesmos comentários feitos para o caso univariado da distribuição T-Student assimétrica, ressaltando que no caso da T-Student normal assimétrica o efeito da assimetria é independente do efeito de caudas pesadas. Essa afirmação, válida para quaisquer misturas de escala assimétricas de normais, decorre da própria expressão dada em (2.7), na qual os parâmetros $\boldsymbol{\lambda}$ e ν não se "misturam", e ficará mais evidente na próxima subseção quando tratarmos da estimação dos parâmetros.

Tendo em vista a estimação paramétrica, explicitaremos ainda a distribuição de $U|\mathbf{Y}=\mathbf{y}$ cuja f.d.p. denotaremos por $h_0(u|\mathbf{y})$. Com as notações usadas nas proposições da Subseção 2.3.3, segue de (3.15) e da relação entre distribuição conjunta e condicional que

$$\begin{aligned} h_0(u|\mathbf{y}) &= \frac{\bar{f}(\mathbf{y},u)}{f(\mathbf{y})} = \frac{2\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{u}\right)\left(\frac{\nu}{2}\right)^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)^{-1}u^{\frac{\nu}{2}-1}e^{-\frac{\nu}{2}}\Phi\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right)}{2t_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right)\Phi\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right)} \\ &= \frac{\left(\frac{\nu+d}{2}\right)^{\frac{\nu+p}{2}}}{\Gamma\left(\frac{\nu+p}{2}\right)}u^{\frac{\nu+p}{2}-1}e^{-\frac{\nu+d}{2}u}. \end{aligned}$$

Logo, $U|\mathbf{Y}=\mathbf{y}\sim$ Gama$\left(\frac{\nu+p}{2},\frac{\nu+d}{2}\right)$, onde $d=(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$.

Distribuição Slash normal assimétrica multivariada

Analogamente ao caso univariado das misturas de escala de normais assimétricas, temos dentre as misturas de escala assimétricas de normais a distribuição Slash normal assimétrica multivariada. Nesta distribuição, temos também $\boldsymbol{\tau}=\nu$ numérico e $U\sim$ Beta$(\nu,1)$, donde segue que $h(u;\nu)=\nu u^{\nu-1}$. Dessa forma, um vetor aleatório p-variado $\mathbf{Y}$ possui distribuição Slash normal assimétrica e escrevemos $\mathbf{Y}\sim$ SSN$_p\left(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\nu\right)$ quando sua f.d.p. é

$$f\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\nu\right)=2s_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right)\Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right),\ \mathbf{y}\in\mathbb{R}^p. \tag{3.17}$$

Sendo $P\left(1;\nu+\frac{p}{2},\frac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}\right)$ a f.d.a. aplicada no ponto 1 de uma variável aleatória Gama$\left(\nu+\frac{p}{2},\frac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}\right)$, temos em (3.17) que $s_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right)=\frac{2^\nu\nu\Gamma\left(\nu+\frac{p}{2}\right)}{\pi^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left[(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]^{-\left(\nu+\frac{p}{2}\right)}P\left(1;\nu+\frac{p}{2},\frac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}\right)$ é a f.d.p. da Slash simétrica. Pode-se mostrar mediante algumas manipulações algébricas que a escrita (3.17) coincide com a indicada em Ferreira, Lachos & Bolfarine (2016).

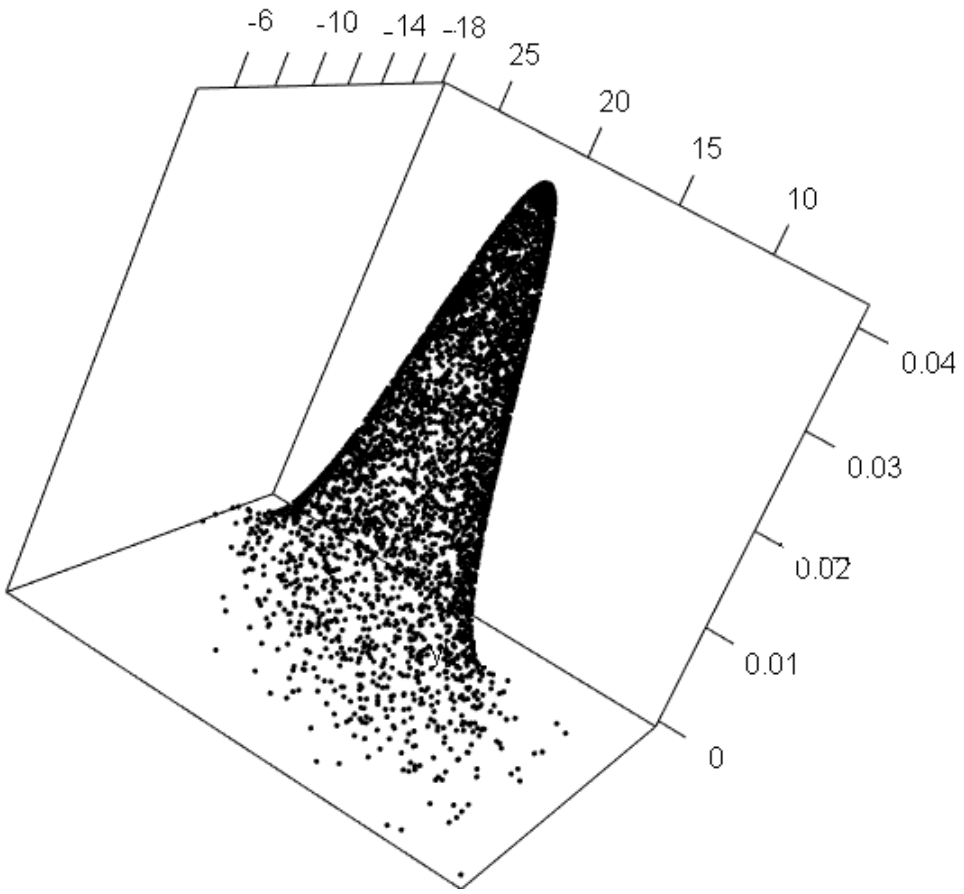
Segue do resultado obtido no caso univariado para a distribuição de U que $E(U^{-1})=\frac{\nu}{\nu-1}$ e, portanto, usando a Proposição 2.3.2, vemos que

$$\begin{aligned} E(\mathbf{Y}) &= \boldsymbol{\mu}+\sqrt{\frac{2}{\pi}}\varpi\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda},\ \nu>\frac{1}{2}; \\ Var(\mathbf{Y}) &= \boldsymbol{\Sigma}^{1/2}\left(\frac{\nu}{\nu-1}\mathbf{I}_p-\frac{2}{\pi}\varpi^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right)\boldsymbol{\Sigma}^{1/2},\ \nu>1. \end{aligned} \tag{3.18}$$

Assim como no exemplo anterior, $\varpi = E\Big([U(U + \boldsymbol{\lambda}^T\boldsymbol{\lambda})]^{-\frac{1}{2}}\Big)$ não possui forma fechada e, nesse caso, é dado por $\varpi = \nu\int_0^1 \frac{u^{\nu-\frac{3}{2}}}{\sqrt{u+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}du$, onde a integral imprópria converge se, e só se, $\nu > \frac{1}{2}$.

Na Figura 18, apresentamos o gráfico da f.d.p. de uma Slash normal assimétrica com $p = 2$ e parâmetros $\boldsymbol{\mu} = (20, -15)$, $\mathbf{B} = \boldsymbol{\Sigma}^{1/2} = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}$, $\boldsymbol{\lambda} = (-1, 2)$ e $\nu = 5$.

Figura 18 – Gráfico da f.d.p. de uma Slash normal assimétrica multivariada



Novamente o gráfico foi esboçado com base numa amostra de 5.000 observações da referida distribuição no *software* R. Neste exemplo, observa-se o mesmo efeito da assimetria constatado no anterior. Por outro lado, o peso caudal é menor no caso da Slash normal assimétrica para o mesmo valor do hiper-parâmetro. No mais, valem os mesmos comentários feitos no exemplo da T-Student normal assimétrica. Quanto à distribuição de $U|\mathbf{Y}=\mathbf{y}$ para fins de estimação dos parâmetros, ao adotarmos as mesmas notações e procedimentos do exemplo anterior, obtemos

$$\begin{aligned} h_0(u|\mathbf{y}) &= \frac{\bar{f}(\mathbf{y},u)}{f(\mathbf{y})} = \frac{2\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{u}\right)\nu u^{\nu-1}\Phi\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right)}{2s_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu\right)\Phi\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right)} \\ &= \frac{\left(\frac{d}{2}\right)^{\nu+\frac{p}{2}}u^{\nu+\frac{p}{2}-1}e^{-\frac{u}{2}d}}{\Gamma\left(\frac{\nu+p}{2}\right)P\left(1;\nu+\frac{p}{2},\frac{d}{2}\right)} = \frac{u^{\nu+\frac{p}{2}-1}e^{-\frac{d}{2}u}}{\displaystyle\int_0^1 u^{\nu+\frac{p}{2}-1}e^{-\frac{d}{2}u}du}. \end{aligned}$$

De acordo com a definição dada em Johnson, Kotz & Balakrishnan (1994), $U|\mathbf{Y}=\mathbf{y}\sim\ \text{TGama}_{(0,1)}\left(\nu+\frac{p}{2},\frac{d}{2}\right)$, onde $d=(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$.

Distribuição normal contaminada assimétrica multivariada

Dentre os mais conhecidos exemplos de misturas de escalas assimétricas de normais, temos ainda a distribuição normal contaminada assimétrica multivariada. Denotamos um vetor aleatório p-dimensional com tal distribuição por $\mathbf{Y}\sim\ \text{SCN}_p\left(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\nu,\gamma\right)$ quando sua f.d.p. é definida para todo $\mathbf{y}\in\mathbb{R}^p$ da forma expressa a seguir:

$$f(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\nu,\gamma)=2\left[\nu\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{\gamma}\right)+(1-\nu)\phi_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\right]\Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right). \quad (3.19)$$

Neste exemplo, assim como no caso univariado, o hiper-parâmetro $\boldsymbol{\tau}=(\nu,\gamma)$ caracteriza a mistura de escala através da função de probabilidade discreta definida da maneira abaixo:

$$h(u;\nu,\gamma)=\begin{cases}\nu & ;\text{ se } u=\gamma\\ 1-\nu & ;\text{ se } u=1\\ 0 & ;\text{ caso contrário}\end{cases}$$

Utilizando a função h acima, já vimos que $\varrho = E(U^{-1}) = \frac{\nu}{\gamma} + 1 - \nu$, donde podemos escrever as expressões

$$\begin{aligned} E(\mathbf{Y}) &= \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\varpi\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda}; \\ Var(\mathbf{Y}) &= \boldsymbol{\Sigma}^{1/2}\left[\left(\frac{\nu}{\gamma} + 1 - \nu\right)\mathbf{I}_p - \frac{2}{\pi}\varpi^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right]\boldsymbol{\Sigma}^{1/2}. \end{aligned} \tag{3.20}$$

Neste caso, temos ainda $\varpi = E\Big([U(U + \boldsymbol{\lambda}^T\boldsymbol{\lambda})]^{-\frac{1}{2}}\Big) = \dfrac{\nu}{\sqrt{\gamma(\gamma + \boldsymbol{\lambda}^T\boldsymbol{\lambda})}} + \dfrac{1-\nu}{\sqrt{1 + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}}$, gerando expressões fechadas para a esperança e a variância.

Observe agora que este exemplo também pode ser visto como uma mistura finita de duas normais assimétricas quando reescrevemos a f.d.p. dada em (3.19) do seguinte modo:

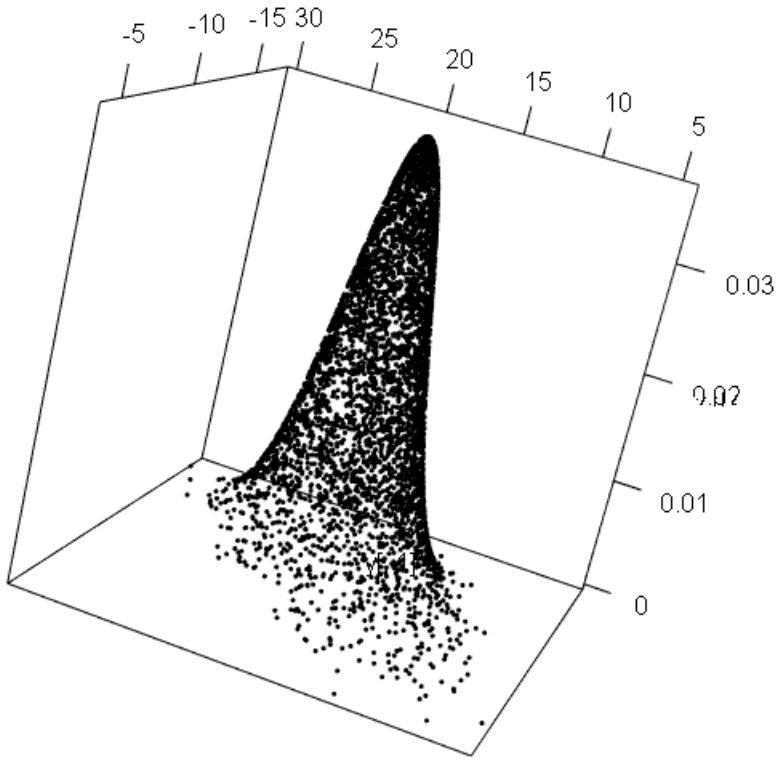
$$2\left[\nu\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{\gamma}\right)\Phi\left(\left(\frac{\boldsymbol{\lambda}}{\sqrt{\gamma}}\right)\left(\frac{\boldsymbol{\Sigma}}{\gamma}\right)^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right) + (1-\nu)\phi_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right)\right]. \tag{3.21}$$

Dessa forma, a distribuição acima pode ser considerada como mistura finita das normais assimétricas $\mathbf{Y}_1 \sim \text{SN}_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{\gamma}, \frac{\boldsymbol{\lambda}}{\sqrt{\gamma}}\right)$ e $\mathbf{Y}_2 \sim \text{SN}_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}\right)$. Como consequência das Proposições 2.3.1 e 2.3.8, podemos reobter a esperança e a variância de $\mathbf{Y}$ fazendo

$$\begin{aligned} E(\mathbf{Y}) &= \nu E(\mathbf{Y}_1) + (1-\nu)E(\mathbf{Y}_2) \\ &= \nu\left[\boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\frac{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda}/\gamma}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}/\gamma}}\right] + (1-\nu)\left[\boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\frac{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda}}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}\right] \\ &= \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\left(\frac{\nu}{\sqrt{\gamma(\gamma+\boldsymbol{\lambda}^T\boldsymbol{\lambda})}} + \frac{1-\nu}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}\right)\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda}; \\ Var(\mathbf{Y}) &= \nu[Var(\mathbf{Y}_1) + E(\mathbf{Y}_1)E(\mathbf{Y}_1)^T] + (1-\nu)[Var(\mathbf{Y}_2) + E(\mathbf{Y}_2)E(\mathbf{Y}_2)^T] - E(\mathbf{Y})E(\mathbf{Y})^T \\ &= \nu Var(\mathbf{Y}_1) + (1-\nu)Var(\mathbf{Y}_2) + \nu(1-\nu)[E(\mathbf{Y}_1) - E(\mathbf{Y}_2)][E(\mathbf{Y}_1) - E(\mathbf{Y}_2)]^T \\ &= \frac{\nu}{\gamma}\boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_p - \frac{2}{\pi}\frac{\boldsymbol{\lambda}\boldsymbol{\lambda}^T/\gamma}{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}/\gamma}\right)\boldsymbol{\Sigma}^{1/2} + (1-\nu)\boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_p - \frac{2}{\pi}\frac{\boldsymbol{\lambda}\boldsymbol{\lambda}^T}{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}\right)\boldsymbol{\Sigma}^{1/2} \\ &+ \frac{2}{\pi}\nu(1-\nu)\left(\frac{1}{\sqrt{\gamma(\gamma+\boldsymbol{\lambda}^T\boldsymbol{\lambda})}} - \frac{1}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}\right)^2\frac{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{\frac{1}{2}}}{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}} \\ &= \boldsymbol{\Sigma}^{1/2}\left[\left(\frac{\nu}{\gamma}+1-\nu\right)\mathbf{I}_p - \frac{2}{\pi}\left(\frac{\nu}{\sqrt{\gamma(\gamma+\boldsymbol{\lambda}^T\boldsymbol{\lambda})}} + \frac{1-\nu}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}\right)^2\frac{\boldsymbol{\lambda}\boldsymbol{\lambda}^T}{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}\right]\boldsymbol{\Sigma}^{1/2}. \end{aligned}$$

A Figura 19 mostra a f.d.p. de uma normal contaminada assimétrica com $p = 2$, $\boldsymbol{\mu} = (20, -15)$, $\mathbf{B} = \boldsymbol{\Sigma}^{1/2} = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}$, $\boldsymbol{\lambda} = (-1, 2)$, $\nu = 0,7$ e $\gamma = 0,6$.

Figura 19 – Gráfico de uma normal contaminada assimétrica multivariada



Tal distribuição também possui caudas mais pesadas do que a normal assimétrica. Ainda enxergando a normal contaminada assimétrica como mistura de escala, o mesmo raciocínio utilizado nos dois exemplos anteriores nos permite concluir que a distribuição de $U|\mathbf{Y} = \mathbf{y}$ é discreta e fica expressa por

$$\begin{aligned} h_0(u|\mathbf{y}) &= \frac{\bar{f}(\mathbf{y},u)}{f(\mathbf{y})} = \frac{2\left[\nu\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{\gamma}\right)\mathbb{I}_{\{\gamma\}}(u) + (1-\nu)\phi_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\mathbb{I}_{\{1\}}(u)\right]\Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right)}{2\left[\nu\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{\gamma}\right) + (1-\nu)\phi_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\right]\Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right)} \\ &= \frac{\nu\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{\gamma}\right)}{\nu\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{\gamma}\right) + (1-\nu)\phi_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)}\mathbb{I}_{\{\gamma\}}(u) + \frac{(1-\nu)\phi_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)}{\nu\phi_p\left(\mathbf{y};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{\gamma}\right) + (1-\nu)\phi_p\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)}\mathbb{I}_{\{1\}}(u). \end{aligned}$$

Optaremos pelo tratamento dessa distribuição como mistura finita no contexto multivariado para a estimação, o que analogamente seria possível na versão multivariada da normal assimétrica contaminada expressa em (3.11). Ademais, gerações de dados pseudoaleatórios e gráficos de alguns dos exemplos multivariados se encontram no Anexo A.

3.2.2 Estimação dos Parâmetros

Na mesma linha da Subseção 3.1.2, procuraremos mostrar primeiro o procedimento geral de estimação parâmetros para modelos multivariados com distribuição mistura de escala assimétrica da família normal, incorporando a regressão multivariada na forma alternativa em Ferreira, Lachos & Bolfarine (2016).

Como descrito na Subseção 2.4, será adotado o modelo $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ para cada indivíduo $i = 1, \ldots, n$ com uma das duas formas alternativas lá apresentadas. Mantendo a mesma notação, consideraremos p variáveis respostas e lembramos que a configuração da matriz de planejamento $\mathbf{X}_i$ de acordo com a forma alternativa I ou II é que definirá a dimensão de $\boldsymbol{\beta}$.

Além disso, faremos as seguintes suposições sobre os erros:

1. $\boldsymbol{\epsilon}_i \sim$ SSMN$_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$;
2. $\text{Cov}(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j) = \mathbf{0}\ \forall i \neq j$.

Com a primeira suposição, obtemos da Proposição 2.3.2 o seguinte:

$$E(\boldsymbol{\epsilon}_i) = \sqrt{\frac{2}{\pi}} E(\Upsilon)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\lambda} \ \text{ e } \ Var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}^{\frac{1}{2}}\left[E(U^{-1})\mathbf{I}_p - \frac{2}{\pi}E(\Upsilon)^2\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right]\boldsymbol{\Sigma}^{\frac{1}{2}}. \quad (3.22)$$

Acima, temos $\Upsilon = \left[U(U + \boldsymbol{\lambda}^T\boldsymbol{\lambda})\right]^{-\frac{1}{2}}$. Dessa forma, o estimador de regressão para a resposta será dado pela expressão:

$$\widehat{\mathbf{Y}_i} = \widehat{E(\mathbf{Y}_i|\mathbf{X}_i)} = \widehat{\mathbf{X}_i\boldsymbol{\beta}} + \widehat{E(\boldsymbol{\epsilon}_i)} = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \sqrt{\frac{2}{\pi}} E(\widehat{\Upsilon})\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\widehat{\boldsymbol{\lambda}}. \quad (3.23)$$

Na expressão (3.23), estamos considerando $E(\widehat{\Upsilon}) = E\left(\left[U(U + \widehat{\boldsymbol{\lambda}}^T\widehat{\boldsymbol{\lambda}})\right]^{-\frac{1}{2}}\right)$.

Já a segunda suposição indica que os erros do modelo associados a indivíduos distintos são não correlacionados. Por isso, é comum dizer de modelos com tal pressuposto que são não estruturados ou que não possuem estrutura de correlação. Note que pode haver correlação entre variáveis distintas associadas a um mesmo indivíduo em virtude de a matriz $Var(\boldsymbol{\epsilon}_i)$ não ser diagonal de modo geral.

Após detalharmos a estrutura do modelo, ressaltamos que o vetor de parâmetros a ser estimado será dado na forma $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\tau})$, onde $\boldsymbol{\alpha} = \text{vech}(\boldsymbol{\Sigma}^{\frac{1}{2}})$ – ver Harville (1997) – pois $\boldsymbol{\Sigma}$ é simétrica. A estimação será feita por máxima verossimilhança via algoritmo EM pelos mesmos motivos já comentadas no caso do modelo univariado.

Para tanto, escreveremos a função log-verossimilhança das misturas de escala da Definição 2.3.3 da seguinte maneira:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(\mathbf{y}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\ln 2 - \frac{p}{2} \ln (2\pi) - \frac{1}{2}\Lambda + \ln K_i + \ln \Phi(A_i) \right]. \quad (3.24)$$

Na expressão acima, $\Lambda = \ln |\boldsymbol{\Sigma}|$ e $K_i = \displaystyle\int_0^{+\infty} u^{p/2} e^{-ud_i/2} h(u; \boldsymbol{\tau}) du$. A partir de agora, fixaremos nas distribuições multivariadas as notações $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$, $d_i = \mathbf{e}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{e}_i$ e $A_i = \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-\frac{1}{2}} \mathbf{e}_i$.

Na notação da Subseção 2.2, o vetor de dados completos para cada indivíduo é $\mathbf{y}_{C_i} = (\mathbf{y}_i, \mathbf{w}_i)$, onde o vetor de dados faltantes $\mathbf{w}_i = (u_i, t_i)$ inclui as variáveis U e T da representação estocástica (2.9). Assim, usando a forma mais simples da f.d.p. dos dados completos deduzida na demonstração do Teorema 2.3.1, obtemos no nosso caso a seguinte log-verossimilhança dos dados completos:

$$\begin{aligned}
\ell_C(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \ln f_C(\mathbf{y}_i, \mathbf{w}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left[2\phi_p \left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}, \frac{\boldsymbol{\Sigma}}{u_i} \right) \phi_1 \left(t_i; \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-\frac{1}{2}} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), 1 \right) h(u_i; \boldsymbol{\tau}) \right] \\
&= \sum_{i=1}^{n} \ln \left[\frac{u_i^{p/2}}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{u_i}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})} \right] + \sum_{i=1}^{n} \ln \left[\frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(t_i - \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-\frac{1}{2}} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right)} \right] + \sum_{i=1}^{n} \ln h(u_i; \boldsymbol{\tau}) \\
&= c - \frac{n}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} u_i (y_i - \mathbf{X}_i\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (y_i - \mathbf{X}_i\boldsymbol{\beta}) + \sum_{i=1}^{n} t_i \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-\frac{1}{2}} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\
&- \frac{1}{2} \sum_{i=1}^{n} \left[\boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-\frac{1}{2}} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \right]^2 + \sum_{i=1}^{n} \ln h(u_i; \boldsymbol{\tau}).
\end{aligned}$$

Na expressão acima, o valor $c = n\ln 2 - \frac{(n+1)p}{2}\ln 2\pi + \frac{p}{2}\sum_{i=1}^{n}\ln u_i - \frac{1}{2}\sum_{i=1}^{n}\ln t_i^2$ constante em relação aos parâmetros pode ser desprezado para a otimização da função ℓ_C. Com isso, podemos decompor a função Q da etapa k do algoritmo EM em $Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) = E_{\widehat{\boldsymbol{\theta}}^{(k)}}(\ell_C(\boldsymbol{\theta})|\mathbf{Y}=\mathbf{y}) = c + \sum_{i=1}^{n} Q_{1i}(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) + \sum_{i=1}^{n} Q_{2i}(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)})$, onde

$$\begin{aligned} Q_{1i}\left(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}\right) &= -\tfrac{1}{2}\ln|\boldsymbol{\Sigma}| - \tfrac{1}{2}\widehat{u_i}^{(k)}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}) + \widehat{t_i}^{(k)}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}) - \tfrac{1}{2}\left[\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right]^2; \\ Q_{2i}\left(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}\right) &= E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(\ln h(u_i;\boldsymbol{\tau})|\mathbf{Y}=\mathbf{y}\right). \end{aligned} \tag{3.25}$$

O passo E é mais simples nas misturas de escala assimétricas de normais comparado ao outro tipo de mistura analisado, porque só demanda o cálculo de duas esperanças condicionais, a saber: $\widehat{t_i}^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(T_i|\mathbf{Y}_i=\mathbf{y}_i\right)$ e $\widehat{u_i}^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(U_i|\mathbf{Y}_i=\mathbf{y}_i\right)$. O segundo será explicitado na Subseção 3.2.3 para cada exemplo específico. Quanto ao primeiro, adaptando o resultado da Proposição 2.3.3, temos $T_i|\mathbf{Y}_i=\mathbf{y}_i;\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(k)} \sim\ TN_{[0,+\infty)}\left(\widehat{A_i}^{(k)},1\right)$, onde $\widehat{A_i}^{(k)} = \widehat{\boldsymbol{\lambda}}^{(k)^T}\widehat{\boldsymbol{\Sigma}}^{(k)^{-\frac{1}{2}}}\left(\mathbf{y}-\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\right)$. Assim, com a notação $W_\Phi\left(\widehat{A_i}^{(k)}\right) = \dfrac{\phi\left(\widehat{A_i}^{(k)}\right)}{\Phi\left(\widehat{A_i}^{(k)}\right)}$, concluímos novamente dos resultados de Johnson, Kotz & Balakrishnan (1994) para normais truncadas que

$$\widehat{t_i}^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(T_i|\mathbf{Y}_i=\mathbf{y}_i\right) = \widehat{A_i}^{(k)} + W_\Phi\left(\widehat{A_i}^{(k)}\right).$$

Assim como no caso univariado, o passo M tradicional do algoritmo costuma ser substituído pela versão ECME de Liu & Rubin (1994) como pode ser visto em Ferreira, Lachos & Bolfarine (2016) para o modelo sem regressão. Nesta subseção, descreveremos a maximização condicional de $Q_1(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)})$ para obter $\widehat{\boldsymbol{\beta}}^{(k+1)}$, $\widehat{\boldsymbol{\Sigma}}^{(k+1)}$ e $\widehat{\boldsymbol{\lambda}}^{(k+1)}$ em cada etapa k. Quanto ao hiper-parâmetro, no ECME se faz a escolha de $\widehat{\boldsymbol{\tau}}^{(k+1)} \in \underset{\boldsymbol{\tau}}{\operatorname{argmax}}\ \ell\left(\widehat{\boldsymbol{\beta}}^{(k+1)},\widehat{\boldsymbol{\Sigma}}^{(k+1)},\widehat{\boldsymbol{\lambda}}^{(k+1)},\boldsymbol{\tau}\right)$, o qual normalmente é único.

A fim de contornar complicações computacionais na implementação do ECME, desenvolveremos em cada um dos nossos três exemplos na subseção seguinte mecanismos para tornar possível a implementação pelo menos próxima de um ECM

por meio do tratamento, aproximação ou modificação da função $Q_2\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\Big)$. Por hora, vamos estabelecer o método geral de estimação dos parâmetros da regressão, de escala e assimetria impondo as condições estacionárias de primeira ordem $\dfrac{\partial Q_1}{\partial \boldsymbol{\theta}}\Big(\widehat{\boldsymbol{\theta}}; \widehat{\boldsymbol{\theta}}^{(k)}\Big) = \mathbf{0}$.

Como as derivadas são matriciais, há alguns detalhes técnicos que mostramos um pouco melhor no Apêndice A. Resumidamente, da observação de que a aplicação $\Big(\boldsymbol{\Sigma}, \boldsymbol{\lambda}\Big) \mapsto \Big(\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Delta}\Big)$, onde $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\lambda}$, é um difeomorfismo local – Lima (1999) – em $\mathrm{GL}_p \times \mathbb{R}^p \subset \mathbb{R}^{p\times p} \times \mathbb{R}^p$, obtemos o seguinte sistema de equações matriciais:

$$
\begin{cases}
\dfrac{\partial Q_1}{\partial \boldsymbol{\beta}}\big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\big) = \mathbf{0} \Rightarrow \sum\limits_{i=1}^{n} \Big[\widehat{u_i}^{(k)} \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) - \widehat{t_i}^{(k)} \mathbf{X}_i^T \boldsymbol{\Delta} + \mathbf{X}_i^T \boldsymbol{\Delta}\boldsymbol{\Delta}^T(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\Big] = \mathbf{0} \\
\dfrac{\partial Q_1}{\partial \boldsymbol{\Sigma}^{-1}}\big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\big) = \mathbf{0} \Rightarrow \frac{1}{2} \sum\limits_{i=1}^{n} \Big[2\boldsymbol{\Sigma} - \mathrm{diag}(\boldsymbol{\Sigma}) - \widehat{u_i}^{(k)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\Big] = \mathbf{0} \\
\dfrac{\partial Q_1}{\partial \boldsymbol{\Delta}}\big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\big) = \mathbf{0} \Rightarrow \sum\limits_{i=1}^{n} \Big[\widehat{t_i}^{(k)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) - (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \boldsymbol{\Delta}\Big] = \mathbf{0}
\end{cases}
$$

Diferentemente do caso univariado das misturas de escala de normais assimétricas, a menos que se consiga alguma reparametrização adequada, não é possível resolver o sistema acima removendo a dependência mútua entre os parâmetros. Dessa forma, a solução que se obtém na etapa k do algoritmo é apenas aproximada e é dada a seguir na forma (e na ordem) mais concisa (e interessante) obtida após cálculos relativamente simples, lembrando que $\widehat{\boldsymbol{\lambda}}^{(k+1)} = \widehat{\boldsymbol{\Sigma}}^{(k+1)^{1/2}} \widehat{\boldsymbol{\Delta}}^{(k+1)}$:

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^{n} \widehat{u_i}^{(k)} \big(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\big)\big(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\big)^T; \\
\widehat{\boldsymbol{\lambda}}^{(k+1)} &= \widehat{\boldsymbol{\Sigma}}^{(k+1)^{\frac{1}{2}}} \left[\sum_{i=1}^{n} \big(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\big)\big(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\big)^T\right]^{-1} \left[\sum_{i=1}^{n} \widehat{t_i}^{(k)} \big(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\big)\right] \\
\widehat{\boldsymbol{\beta}}^{(k+1)} &= \left[\sum_{i=1}^{n} \mathbf{X}_i^T \widehat{\boldsymbol{\Sigma}}^{(k+1)^{-\frac{1}{2}}} \big(\widehat{u_i}^{(k)} \mathbf{I}_p + \widehat{\boldsymbol{\lambda}}^{(k+1)} \widehat{\boldsymbol{\lambda}}^{(k+1)^T}\big) \widehat{\boldsymbol{\Sigma}}^{(k+1)^{-\frac{1}{2}}} \mathbf{X}_i\right]^{-1}; \\
&\times \left[\sum_{i=1}^{n} \mathbf{X}_i^T \Big(\widehat{\boldsymbol{\Sigma}}^{(k+1)^{-\frac{1}{2}}} \big(\widehat{u_i}^{(k)} \mathbf{I}_p + \widehat{\boldsymbol{\lambda}}^{(k+1)} \widehat{\boldsymbol{\lambda}}^{(k+1)^T}\big) \widehat{\boldsymbol{\Sigma}}^{(k+1)^{-\frac{1}{2}}} \mathbf{y}_i - \widehat{t_i}^{(k)} \widehat{\boldsymbol{\Sigma}}^{(k+1)^{-\frac{1}{2}}} \widehat{\boldsymbol{\lambda}}^{(k+1)}\Big)\right].
\end{aligned}
$$

Note que, ao trocarmos $\mathbf{X}_i$ por 1 para todo $i = 1, \ldots, n$, recuperamos as estimativas dos parâmetros dadas em Ferreira, Lachos & Bolfarine (2016) substituindo o parâmetro da regressão $\boldsymbol{\beta}$ pelo parâmetro de locação $\boldsymbol{\mu}$. Essa situação em que não há variáveis explicativas para os indivíduo permite ver os modelos de Ferreira, Lachos & Bolfarine (2016) como casos particulares dos que estão sendo apresentados.

Os valores iniciais usados para os três primeiros parâmetros também vêm da suposição "ingênua" de normalidade e são dados por $\boldsymbol{\beta}^{(0)} = \left[\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{y}_i\right]$, $\boldsymbol{\Sigma}^{\frac{1}{2}^{(0)}} = S_{\mathbf{e}^{(0)}}$ e $\boldsymbol{\lambda}^{(0)} = g_{\mathbf{e}^{(0)}}$, onde $\mathbf{e}_i^{(0)} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}^{(0)}$ é a i-ésima entrada do vetor de resíduos inicial do modelo. Os valores iniciais do hiper-parâmetro serão apresentados na Subseção 3.2.3 e o critério de parada do algoritmo será $\left\|\widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)}\right\| < \varepsilon$ pelas mesmas razões já comentadas no caso univariado.

Antes de passarmos ao estudo de cada caso particular das misturas de escala assimétricas de normais, vamos determinar o escore e a matriz de informação de Fisher observada, isto é, o gradiente e a oposta da hessiana da log-verossimilhança dos dados incompletos dada em (3.24). Ao derivarmos duas vezes a última equação, podemos escrever de maneira genérica $\dfrac{\partial \ell}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \dfrac{\partial \ell_i}{\partial \boldsymbol{\theta}}$ e $-\dfrac{\partial^2 \ell}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \sum_{i=1}^{n} \dfrac{\partial^2 \ell_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}$, onde a i-ésima parcela é dada em cada caso após alguns cálculos, respectivamente, por

$$\frac{\partial \ell_i}{\partial \boldsymbol{\theta}} = -\frac{1}{2}\frac{\partial \Lambda}{\partial \boldsymbol{\theta}} + \frac{1}{K_i}\frac{\partial K_i}{\partial \boldsymbol{\theta}} + W_\Phi(A_i)\frac{\partial A_i}{\partial \boldsymbol{\theta}};$$

$$\frac{\partial^2 \ell_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \frac{1}{2}\frac{\partial^2 \Lambda}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} + \frac{1}{K_i^2}\frac{\partial K_i}{\partial \boldsymbol{\theta}}\frac{\partial K_i}{\partial \boldsymbol{\theta}^T} - \frac{1}{K_i}\frac{\partial^2 K_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} - W_\Phi(A_i)\frac{\partial^2 A_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} + W'_\Phi(A_i)\frac{\partial A_i}{\partial \boldsymbol{\theta}}\frac{\partial A_i}{\partial \boldsymbol{\theta}^T}.$$

Na segunda expressão acima, temos $W'_\Phi(A_i) = -W_\Phi(A_i)[A_i + W_\Phi(A_i)]$. As expressões das derivadas primeira e segunda de ℓ_i com respeito a cada um dos subparâmetros de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ visando à implementação computacional se encontram no Anexo B.

3.2.3 **Estudo de Casos**

Objetivamos com esta subseção apresentar técnicas ainda não empregadas na literatura para a estimação dos hiper-parâmetros nos modelos de regressão linear envolvendo misturas de escala assimétricas de normais descritos nos exemplos da Subseção 3.2.1. O princípio por detrás de todos os métodos é substituir o tratamento do problema de estimação usando o ECME pela adoção aproximada de um EM clássico em cada caso (na verdade uma versão do ECM).

Após desenvolvermos tais métodos, mostraremos através de simulações computacionais controladas que as propostas realizadas são mais eficientes em

tempo e número de iterações do que o método geralmente empregado na literatura para os três exemplos básicos desenvolvidos. Além disso, compararemos ainda o desempenho na estimação por máxima verossimilhança para as mesmas distribuições do EM modificado com uma versão do método Quasi-Newton BFGS já presente como função programada no *software* `R`.

Ao fazermos isso, procuraremos refutar a tese de que o algoritmo EM é lento quando comparado a outros métodos de otimização tradicionais. Dessa forma, além dos resultados provenientes do passo E que serão úteis no Capítulo 4, veremos que pelo menos nos três exemplos abordados o EM também se mostra competitivo quando se trata apenas da otimização.

A fim de descrever os métodos que permitirão o tratamento diferenciado nos exemplos da T-Student normal assimétrica e da Slash normal assimétrica, vamos usar a seguinte definição de *família exponencial* dada em Casella (2002):

Definição 3.2.1. Diz-se que uma variável aleatória X pertence à *família exponencial de distribuições* quando sua f.d.p. é da forma

$$g(x;\boldsymbol{\vartheta}) = \eta(x)\kappa(\boldsymbol{\vartheta})e^{\sum_{i=1}^{k}\omega_i(\boldsymbol{\vartheta})t_i(x)}, \tag{3.26}$$

onde todas as funções ω_i, t_i, η e κ devem ser continuamente diferenciáveis sendo ainda as duas últimas funções positivas.

O próximo resultado, que consta em Casella (2002) p. 112, será a chave para o desenvolvimento dos métodos de estimação para dois exemplos.

Proposição 3.2.1. Dada uma variável aleatória X pertencente à família exponencial dependendo do vetor parâmetros $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_l)$, temos que

$$E\left(\sum_{i=1}^{k}\frac{\partial\omega_i(\boldsymbol{\vartheta})}{\partial\vartheta_j}t_i(X)\right) = -\frac{\partial}{\partial\vartheta_j}(\ln\kappa(\boldsymbol{\vartheta})).$$

Demonstração:

Provaremos o resultado considerando as derivadas relativas ao vetor $\boldsymbol{\vartheta}$ na notação do cálculo matricial (ver Apêndice A). Denotando por χ o conjunto dos

valores assumidos pela variável aleatória X, a aplicação do logaritmo natural na expressão (3.26) nos dá o seguinte:

$$\ln g(x;\boldsymbol{\vartheta}) = \ln \eta(x) + \ln \kappa(\boldsymbol{\vartheta}) + \sum_{i=1}^{n} t_i(x)\omega_i(\boldsymbol{\vartheta}),\ \forall x \in \chi.$$

Agora, ao sucessivamente derivarmos a expressão acima em relação a $\boldsymbol{\vartheta}$, multiplicarmos ambos os membros da igualdade resultante por $g(x;\boldsymbol{\vartheta})$ e tomarmos a integral em χ, obtemos

$$\begin{aligned}\int_{\chi} \frac{\partial g}{\partial \boldsymbol{\vartheta}}(x;\boldsymbol{\vartheta})dx &= \int_{\chi} \frac{\partial}{\partial \boldsymbol{\vartheta}}(\ln \kappa(\boldsymbol{\vartheta}))g(x;\boldsymbol{\vartheta})dx + \int_{\chi} \left[\sum_{i=1}^{k} \frac{\partial \omega_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} t_i(x)\right] g(x;\boldsymbol{\vartheta})dx \\ &= \frac{\partial}{\partial \boldsymbol{\vartheta}}(\ln \kappa(\boldsymbol{\vartheta})) + E\left(\sum_{i=1}^{k} \frac{\partial \omega_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} t_i(X)\right).\end{aligned}$$

Dessa forma, concluiremos o resultado mostrando que $\int_{\chi} \frac{\partial g}{\partial \boldsymbol{\vartheta}}(x;\boldsymbol{\vartheta})dx = 0$. Com efeito, o fato de g ser uma função de probabilidade implica $\int_{\chi} g(x;\boldsymbol{\vartheta})dx = 1$, o que já foi utilizado no cálculo acima. Como g é continuamente diferenciável em relação a x e $\boldsymbol{\vartheta}$, podemos aplicar a Regra de Leibniz – ver Lima (1999) – para obter

$$\int_{\chi} \frac{\partial g}{\partial \boldsymbol{\vartheta}}(x;\boldsymbol{\vartheta})dx = \frac{\partial}{\partial \boldsymbol{\vartheta}}\left(\int_{\chi} g(x;\boldsymbol{\vartheta})dx\right) = 0.$$

Note que se X for uma variável aleatória discreta, a integral acima se degenera numa soma e a prova não exigiria a recorrência à Regra de Leibniz. ∎

Mais especificamente, para nossos propósitos, precisaremos da Proposição 3.2.1 aplicada aos dois exemplos a seguir:

(i) Seja $X \sim$ Gama(α, β). Então, $\boldsymbol{\vartheta} = (\alpha, \beta)$ e $g(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = 1 \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot e^{-\beta x + (\alpha-1)\ln x}$. Assim, $\eta(x) = 1$, $\kappa(\boldsymbol{\vartheta}) = \frac{\beta^\alpha}{\Gamma(\alpha)}$, $\omega_1(\boldsymbol{\vartheta}) = -\beta$, $\omega_2(\boldsymbol{\vartheta}) = \alpha - 1$, $t_1(x) = x$ e $t_2(x) = \ln x$, donde segue que X pertence à família exponencial e pela Proposição 3.2.1 temos que

$$\begin{aligned}E(X) &= \frac{\partial}{\partial \beta}\left(\ln\left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]\right) = \frac{\partial}{\partial \beta}[\alpha \ln \beta - \ln \Gamma(\alpha)] = \frac{\alpha}{\beta}; \\ E(\ln X) &= -\frac{\partial}{\partial \alpha}\left(\ln\left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]\right) = \frac{\partial}{\partial \alpha}[\ln \Gamma(\alpha) - \alpha \ln \beta] = \Psi(\alpha) - \ln \beta.\end{aligned}$$

(ii) Seja $X \sim$ TGama$_{(0,1)}(\alpha, \beta)$. Assim, $\boldsymbol{\vartheta} = (\alpha, \beta)$ e $g(x; \alpha, \beta) = \dfrac{x^{\alpha-1}e^{-\beta x}}{\displaystyle\int_0^1 x^{\alpha-1}e^{-\beta x}dx} =$
$1 \cdot \dfrac{1}{\displaystyle\int_0^1 x^{\alpha-1}e^{-\beta x}dx} \cdot e^{-\beta x+(\alpha-1)\ln x}$ e daí $\eta(x) = 1$, $\kappa(\boldsymbol{\vartheta}) = \dfrac{1}{\displaystyle\int_0^1 x^{\alpha-1}e^{-\beta x}dx}$,
$\omega_1(\boldsymbol{\vartheta}) = -\beta$, $\omega_2(\boldsymbol{\vartheta}) = \alpha - 1$, $t_1(x) = x$ e $t_2(x) = \ln x$.
Dessa forma, X pertence à família exponencial e pela Proposição 3.2.1 temos

$$E(X) = -\frac{\partial}{\partial\beta}\left(\ln\int_0^1 x^{\alpha-1}e^{-\beta x}dx\right) = \frac{\int_0^1 x^{\alpha}e^{-\beta x}dx}{\int_0^1 x^{\alpha-1}e^{-\beta x}dx} = \frac{\alpha P(1;\alpha+1,\beta)}{\beta P(1;\alpha,\beta)};$$

$$E(\ln X) = \frac{\partial}{\partial\alpha}\left(\ln\int_0^1 x^{\alpha-1}e^{-\beta x}dx\right) = \frac{\int_0^1 x^{\alpha-1}\ln(x)e^{-\beta x}dx}{\int_0^1 x^{\alpha-1}e^{-\beta x}dx} = \frac{\beta^{\alpha}\int_0^1 x^{\alpha-1}\ln(x)e^{-\beta x}dx}{\Gamma(\alpha)P(1;\alpha,\beta)}.$$

Quanto à distribuição normal contaminada assimétrica, faremos o já mencionado tratamento dessa distribuição como a mistura finita expressa na equação (3.21) e estabeleceremos as relações desse método com a já conhecida abordagem da referida distribuição como mistura de escala.

3.2.3.1 *Estudo de Simulação*

Neste estudo de simulação com dados artificiais gerados no `R`, procederemos de modo bastante similar ao caso univariado. No entanto, pelo fato de termos em geral mais parâmetros no caso multivariado, usaremos os seguintes tamanhos de amostra: 100, 200, 300, 500 e 750 para verificar a estimação e sua consistência, além de comparar os algoritmos padrão (ECME usado na literatura) e o modificado (com as novas propostas).

Em todos os exemplos, consideramos duas variáveis respostas ($p = 2$) e geramos duas variáveis explicativas com distribuições uniformes $U(1, 10)$ e $U(-5, 0)$ adotando a forma alternativa I da regressão multivariada (ver Anexo A) e realizamos 200 simulações por tamanho amostral. Os parâmetros usados na geração dos modelos foram: $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2]^T = [40\ -30\ 15]^T$, $\boldsymbol{\alpha} = [2,5\ -1\ 1,5]^T$, $\boldsymbol{\lambda} = [2\ -1]^T$ e hiper-parâmetro $\nu = 3,5$ na T-Student e na Slash normais assimétricas, ou hiper-parâmetro composto por $\nu = 0,25$ e $\gamma = 0,4$ na normal contaminada assimétrica.

No mais, os procedimentos são idênticos aos adotados nos modelos de regressão univariada com o mesmo esquema de montagem das tabelas: uma para

avaliar a recuperação dos parâmetros e outra para a consistência. No caso da comparação do desempenho computacional dos três métodos, exibiremos boxplots de tempo e iterações para três tamanhos de amostra n classificados em pequeno ($n = 100$), médio ($n = 300$) e grande ($n = 750$). Finalmente, o critério de parada e as condições de teste para verificação da convergência dos algoritmos a um máximo local não degenerado são exatamente as mesmas utilizadas nos modelos univariados.

Na sequência, descreveremos a técnica de estimação adotada em cada exemplo específico, além de exibir as esperanças condicionais exigidas no passo E e o valor inicial utilizado para o hiper-parâmetro.

T-Student normal assimétrica

Neste exemplo, a função Q_2 na etapa k do algoritmo EM é dada por

$$Q_2\big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\big) = n\left[\frac{\nu}{2}\ln\left(\frac{\nu}{2}\right) - \ln\Gamma\left(\frac{\nu}{2}\right)\right] - \frac{\nu}{2}\sum_{i=1}^{n}\big(\widehat{u_i}^{(k)} - \widehat{lu_i}^{(k)}\big). \tag{3.27}$$

Como consequência do que vimos na Subseção 3.2.1, para a T-Student normal assimétrica vale $U_i|\mathbf{Y}_i = \mathbf{y}_i; \widehat{\boldsymbol{\theta}}^{(k)} \sim$ Gama $\left(\dfrac{\widehat{\nu}^{(k)} + p}{2}, \dfrac{\widehat{\nu}^{(k)} + \widehat{d}_i^{\,(k)}}{2}\right)$ e segue dos resultados da família exponencial que

$$\begin{aligned}
\widehat{u_i}^{(k)} &= E_{\widehat{\boldsymbol{\theta}}^{(k)}}\big(U_i|\mathbf{Y}_i = \mathbf{y}_i\big) = \frac{\widehat{\nu}^{(k)} + p}{\widehat{\nu}^{(k)} + \widehat{d}_i^{\,(k)}};\\
\widehat{lu_i}^{(k)} &= E_{\widehat{\boldsymbol{\theta}}^{(k)}}\big(\ln U_i|\mathbf{Y}_i = \mathbf{y}_i\big) = \Psi\left(\frac{\widehat{\nu}^{(k)} + p}{2}\right) - \ln\left(\frac{\widehat{\nu}^{(k)} + \widehat{d}_i^{\,(k)}}{2}\right).
\end{aligned}$$

Observe que podemos realizar a maximização direta da função Q na etapa k do algoritmo EM para a T-Student normal assimétrica resolvendo ao conjunto das condições estacionárias $\dfrac{\partial Q}{\partial \boldsymbol{\theta}}\big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\big) = \mathbf{0}$. Estas correspondem a $\dfrac{\partial Q_1}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\big) = \mathbf{0}$, já resolvidas para qualquer mistura de escala na Subseção 3.2.2 e $\dfrac{\partial Q_2}{\partial \nu}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\big) = 0$, visto que a função Q_2 depende apenas do hiper-parâmetro ν. Após a derivação de (3.27), vemos que a última equação fica expressa por

$$\frac{\partial Q_2}{\partial \nu}\big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\big) = \frac{n}{2}\left[\ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right) + 1 - \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{u_i}^{(k)} - \widehat{lu_i}^{(k)}\big)\right] = 0. \tag{3.28}$$

É notável que não há meio direto de explicitar ν na equação acima. Entretanto, com o intuito de tornar tal explicitação possível pelo menos aproximadamente, fazendo $t = \dfrac{\nu}{2}$ propomos a seguinte aproximação:

$$\ln t - \Psi(t) \approx e^{c_0} t^{c_1 + c_2 \ln t};\ (c_0, c_1, c_2) \in \underset{(c_0,c_1,c_2)}{\operatorname{argmin}} \int_1^{15} \frac{e^{-t}}{t} \left[\ln\left(\ln t - \Psi(t)\right) - \left(c_0 + c_1 \ln t + c_2 (\ln t)^2\right)\right] dt.$$

A técnica de aproximação consiste num método conhecido como *aproximação por mínimos quadrados ponderados* que pode ser encontrado em Powell (1981). Alguns detalhes sobre esse método, como ele foi empregado no nosso contexto e qual sua margem de erro constam no Apêndice B.

Assumindo que tal aproximação é suficientemente boa para $t \in [1, 15]$, ou equivalentemente, $\nu \in [2, 30]$ (intervalo de maior interesse prático para esse parâmetro), a equação (3.28) toma a forma aproximada dada por $c_0 t^{c_1 + c_2 \ln t} = \dfrac{1}{n} \sum\limits_{i=1}^{n} \left(\widehat{u_i}^{(k)} - \widehat{lu_i}^{(k)}\right) - 1$. Assim, fazendo $\widehat{s}^{(k)} = \dfrac{1}{n} \sum\limits_{i=1}^{n} \left(\widehat{u_i}^{(k)} - \widehat{lu_i}^{(k)}\right) - 1$ e aplicando o logaritmo em ambos os membros da última equação, obtemos o resultado abaixo:

$$c_2 (\ln t)^2 + c_1 \ln t + \ln\left(\frac{c_0}{\widehat{s}^{(k)}}\right) = 0 \Rightarrow t = e^{\frac{1}{2c_2}\left[-c_1 \pm \sqrt{c_1^2 - 4c_2 \ln\left(\frac{c_0}{\widehat{s}^{(k)}}\right)}\right]}.$$

Portanto, a estimativa do hiper-parâmetro que consideraremos na etapa $k+1$ do EM é $\widehat{\nu}^{(k+1)} = 2e^{-\frac{1}{2c_2}\left[c_1 + \sqrt{c_1^2 - 4c_2 \ln\left(\frac{c_0}{\widehat{s}^{(k)}}\right)}\right]}$ por heurística, visto que a outra raiz tende a se afastar do valor verdadeiro. As constantes possuem os seguintes valores aproximados: $e^{c_0} \approx 0,5768854$, $c_1 \approx -1,112673$ e $c_2 \approx 0,02783412$.

Destacamos ainda que o valor inicial utilizado para o hiper-parâmetro será $\nu^{(0)} = 2,01$ e controlaremos as simulações eliminando amostras que gerem estimativas maiores do que 30 para o hiper-parâmetro pelos mesmos motivos elencados no caso univariado.

Apresentaremos agora os resultados das simulações realizadas no `R` dentro das especificações anteriores. Reiteramos que as tabelas construídas a seguir trazem medidas para verificar a recuperação dos parâmetros originais e a consistência da estimação. Além disso, boxplots ilustram a comparação dos tempos para os dois algoritmos de interesse: padrão e modificado.

- Recuperação dos parâmetros

Tabela 7 – Recuperação dos parâmetros no modelo T-Student normal assimétrico

(a) Algoritmo padrão

	n=100			n=300			n=750		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
β_0(40)	40,00172	0,31504	0,28613	39,99628	0,19202	0,17482	40,00052	0,09833	0,10091
β_1(-30)	-30,00303	0,03992	0,03615	-29,99968	0,02258	0,02230	-29,99849	0,01271	0,01284
β_2(15)	15,00160	0,06216	0,06055	15,00067	0,03894	0,03679	15,00095	0,02356	0,02321
α_1(2,5)	2,54533	0,32692	0,22859	2,51564	0,18357	0,13046	2,49173	0,11043	0,08177
α_2(-1)	-1,00922	0,15454	0,12073	-1,00854	0,08124	0,06907	-1,00269	0,05256	0,04353
α_3(1,5)	1,50506	0,18305	0,13160	1,50342	0,09161	0,07575	1,50558	0,06292	0,04790
λ_1(2)	2,35885	0,71507	0,64643	2,09011	0,37218	0,30818	2,02189	0,21496	0,18606
λ_2(-1)	-1,11615	0,42457	0,35124	-1,04538	0,18668	0,17647	-1,01663	0,11868	0,10720
ν(3,5)	4,04921	1,52368	1,27298	3,64258	0,64864	0,56503	3,55147	0,35465	0,33801

(b) Algoritmo modificado

	n=100			n=300			n=750		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
β_0(40)	40,00174	0,31494	0,28613	39,99628	0,19200	0,17482	40,00048	0,09833	0,10091
β_1(-30)	-30,00303	0,03992	0,03615	-29,99968	0,02258	0,02230	-29,99849	0,01271	0,01284
β_2(15)	15,00160	0,06217	0,06055	15,00067	0,03894	0,03679	15,00095	0,02356	0,02321
α_1(2,5)	2,54523	0,32676	0,22857	2,51584	0,18334	0,13047	2,49212	0,11034	0,08178
α_2(-1)	-1,00912	0,15450	0,12073	-1,00862	0,08121	0,06908	-1,00282	0,05253	0,04354
α_3(1,5)	1,50502	0,18294	0,13160	1,50354	0,09153	0,07576	1,50577	0,06286	0,04791
λ_1(2)	2,35870	0,71500	0,64636	2,09029	0,37211	0,30821	2,02226	0,21498	0,18610
λ_2(-1)	-1,11608	0,42436	0,35122	-1,04547	0,18670	0,17649	-1,01674	0,11868	0,10722
ν(3,5)	4,04803	1,53015	1,27395	3,64305	0,64292	0,56491	3,55390	0,35266	0,33842

Na Tabela 7, vemos que as estimativas dos parâmetros por ambos os algoritmos são muito similares. As diferenças estão no máximo na 4ª casa decimal para $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$, e no máximo na 3ª casa decimal para $\boldsymbol{\lambda}$ (destacados em azul) e $\boldsymbol{\tau} = \nu$ (destacado em amarelo). Há de se destacar que a função objetivo log-verossimilhança (com valores absolutos da ordem de centenas ou milhares) tem diferenças ainda menores: da ordem de 10^{-5}.

À medida que o tamanho n da amostra cresce, vemos claramente que o desvio e o erro padrão de todos os parâmetros estão diminuindo e ficando cada vez mais próximos entre si, o que aponta a consistência da estimação pelos dois métodos. Essa evidência é corroborada pelas medidas da Tabela 8, todas tendendo a 0 como se espera.

- Consistência da estimação

Tabela 8 – Consistência no modelo T-Student normal assimétrico

(a) Algoritmo padrão

n=100				n=300				n=750			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,00172	0,31505	0,00271	0,21584	0,00372	0,19206	0,02016	0,11105	0,00052	0,09833	0,00007	0,06240
0,00302	0,04004	0,00242	0,02641	0,00032	0,02259	0,00135	0,01266	0,00151	0,01280	0,00183	0,00885
0,00160	0,06218	0,00073	0,04912	0,00067	0,03895	0,00065	0,02550	0,00095	0,02358	0,00028	0,01687
0,04533	0,33005	0,00416	0,18984	0,01564	0,18423	0,01335	0,12403	0,00827	0,11074	0,01195	0,08017
0,00922	0,15481	0,00327	0,11489	0,00854	0,08169	0,01148	0,04734	0,00269	0,05262	0,00150	0,03281
0,00506	0,18312	0,00467	0,11877	0,00342	0,09167	0,00198	0,06147	0,00558	0,06317	0,00257	0,04665
0,35885	0,80006	0,23136	0,45895	0,09011	0,38293	0,04220	0,23872	0,02189	0,21608	0,00401	0,13644
0,11615	0,44017	0,06012	0,25054	0,04538	0,19211	0,01103	0,13175	0,01663	0,11984	0,02208	0,06866
0,54921	1,61964	0,07499	0,67706	0,14258	0,66412	0,08552	0,36200	0,05147	0,35836	0,00922	0,23567

(b) Algoritmo modificado

n=100				n=300				n=750			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,00172	0,31505	0,00271	0,21584	0,00373	0,19204	0,02016	0,11106	0,00048	0,09834	0,00005	0,06240
0,00302	0,04004	0,00242	0,02641	0,00032	0,02259	0,00134	0,01266	0,00151	0,01280	0,00183	0,00885
0,00160	0,06218	0,00073	0,04912	0,00067	0,03895	0,00065	0,02551	0,00095	0,02358	0,00028	0,01688
0,04533	0,33005	0,00416	0,18984	0,01584	0,18403	0,01317	0,12426	0,00788	0,11062	0,01181	0,08013
0,00922	0,15481	0,00327	0,11489	0,00862	0,08167	0,01166	0,04703	0,00282	0,05261	0,00168	0,03271
0,00506	0,18312	0,00467	0,11877	0,00354	0,09160	0,00187	0,06165	0,00577	0,06313	0,00279	0,04652
0,35885	0,80006	0,23136	0,45895	0,09029	0,38291	0,04239	0,23825	0,02226	0,21613	0,00443	0,13643
0,11615	0,44017	0,06012	0,25054	0,04546	0,19215	0,01117	0,13168	0,01674	0,11986	0,02222	0,06866
0,54921	1,61964	0,07499	0,67706	0,14305	0,65864	0,08828	0,36039	0,05390	0,35676	0,01266	0,23530

Todos os resultados anteriores mostram que a convergência para o parâmetro real ocorre aproximadamente da mesma forma com qualquer um dos dois algoritmos. Nas Figuras 20 e 21, vemos graficamente essa convergência em viés e variabilidade relativos para cada grupo de parâmetros no método modificado (no padrão é praticamente igual).

Por fim, na Figura 22, os boxplots revelam o real ganho que se tem com a adoção do algoritmo modificado: uma estimação com tempo médio de duas a três vezes menor e muito mais homogêneo do que com o algoritmo padrão. Observe que esse ganho é maior quanto maior o tamanho da amostra e, apesar de não apresentadas aqui, outras simulações indicaram que esse ganho também aumenta com a elevação do número de parâmetros estimados.

- Comportamento mediano por parâmetro

Figura 20 – Viés mediano por parâmetro da T-Student normal assimétrica multivariada

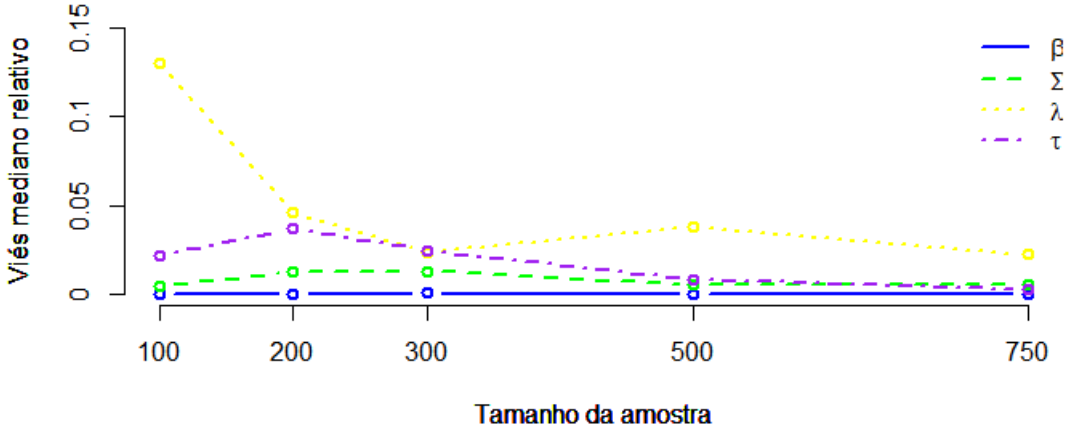
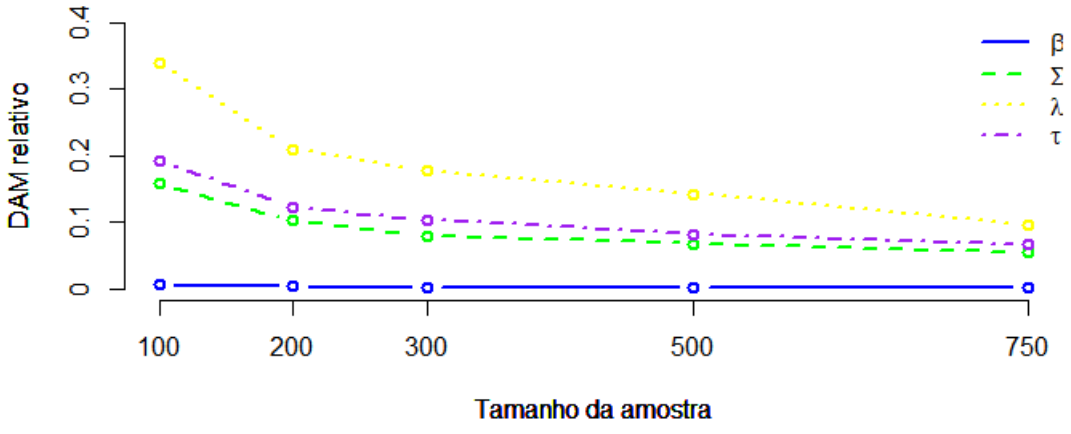


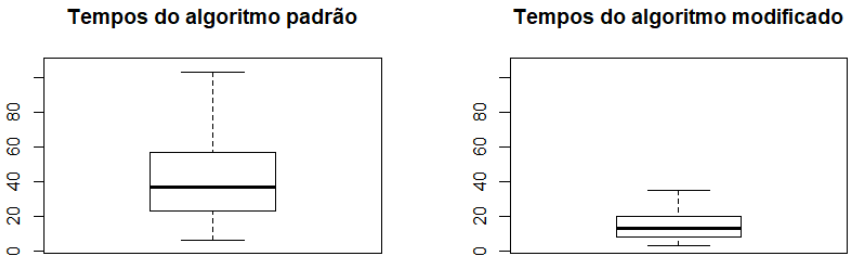
Figura 21 – Desvio absoluto mediano (DAM) por parâmetro da T-Student normal assimétrica multivariada



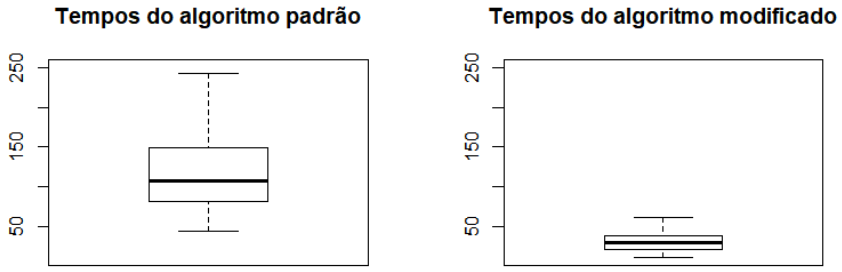
- Desempenho dos algoritmos

Figura 22 – Boxplots de tempos para a T-Student normal assimétrica multivariada

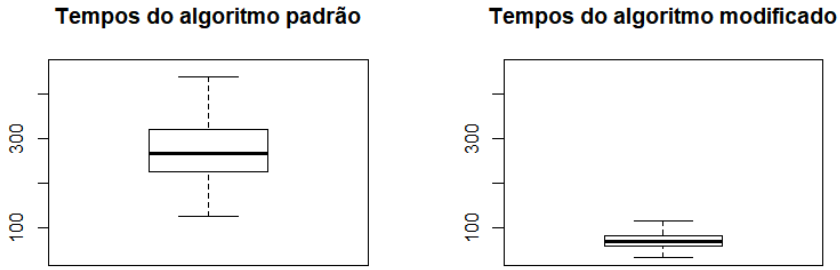
(a) Para n=100



(b) Para n=300



(c) Para n=750



Slash normal assimétrica

Nesta distribuição, a função Q_2 obtida na etapa k do algoritmo EM é

$$Q_2\Big(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}\Big) = n\ln\nu + (\nu-1)\sum_{i=1}^{n}\widehat{lu_i}^{(k)}. \tag{3.29}$$

Em decorrência do que vimos da Slash normal assimétrica na Subseção 3.2.1, temos $U_i|\mathbf{Y}_i=\mathbf{y}_i;\widehat{\boldsymbol{\theta}}^{(k)} \sim \text{ TGama}_{(0,1)}\left(\widehat{\nu}^{(k)}+\frac{p}{2},\frac{\widehat{d}_i^{(k)}}{2}\right)$ e, pelos resultados da família exponencial, concluímos que

$$\begin{aligned}
\widehat{u}_i^{(k)} &= E_{\widehat{\boldsymbol{\theta}}^{(k)}}(U_i|\mathbf{Y}_i=\mathbf{y}_i) = \frac{2\widehat{\nu}^{(k)}+p}{\widehat{d}_i^{(k)}/2}\frac{P(1;\widehat{\nu}^{(k)}+p/2+1,\widehat{d}_i^{(k)})}{P(1;\widehat{\nu}^{(k)}+p/2,\widehat{d}_i^{(k)}/2)};\\
\widehat{lu_i}^{(k)} &= E_{\widehat{\boldsymbol{\theta}}^{(k)}}(\ln U_i|\mathbf{Y}_i=\mathbf{y}_i) = \frac{(\widehat{d}_i^{(k)}/2)^{\widehat{\nu}^{(k)}+p/2}}{\Gamma(\widehat{\nu}^{(k)}+p/2)}\frac{\displaystyle\int_0^1 u^{\widehat{\nu}^{(k)}+p/2-1}\ln(u)e^{u\widehat{d}_i^{(k)}/2}du}{P(1;\widehat{\nu}^{(k)}+p/2,\widehat{d}_i^{(k)}/2)}.
\end{aligned}$$

Na mesma linha do exemplo anterior, podemos maximizar a função Q diretamente na etapa $k+1$ do EM, bastando para isso resolver a equação $\dfrac{\partial Q_2}{\partial\nu}\Big(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}\Big)=0$, uma vez que os demais parâmetros são determinados pelo caso geral da Subseção 3.2.2 com o conhecimento de $\widehat{u}_i^{(k)}$. Além disso, neste caso, a obtenção de ν é bem simples, pois derivando a expressão (3.29) e igualando-a a zero obtemos o seguinte:

$$\frac{\partial Q_2}{\partial\nu}\Big(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}\Big) = \frac{n}{\nu}+\sum_{i=1}^{n}\widehat{lu_i}^{(k)} = 0. \tag{3.30}$$

Dessa forma, ao explicitarmos o hiper-parâmetro na equação (3.30), conseguimos como estimativa na etapa $k+1$ do algoritmo $\widehat{\nu}^{(k+1)} = -\dfrac{n}{\sum_{i=1}^{n}\widehat{lu_i}^{(k)}}$. A escrita de $\widehat{lu_i}^{(k)}$ indicada acima gera resultados ligeiramente mais rápidos em simulações no `R` do que a proposta por Lange & Sinsheimer (1993).

Por fim, ressaltamos que o valor inicial usado para o hiper-parâmetro será $\nu^{(0)}=1,01$ e controlaremos as simulações eliminando amostras que gerem estimativas maiores do que 30 para o hiper-parâmetro assim como na distribuição T-Student normal assimétrica. Vamos agora aos resultados das simulações nos mesmos moldes do exemplo anterior.

- Recuperação dos parâmetros

Tabela 9 – Recuperação dos parâmetros na Slash normal assimétrica

(a) Algoritmo padrão

	n=100			n=300			n=750		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
$\beta_0(40)$	40,02062	0,34948	0,29631	40,00525	0,19107	0,17369	40,01382	0,11703	0,10786
β_1(-30)	-29,99979	0,03873	0,03605	-29,99962	0,02060	0,02102	-30,00141	0,01336	0,01284
β_2(15)	14,99984	0,06359	0,06055	14,99765	0,03544	0,03713	14,99947	0,02140	0,02344
α_1(2,5)	2,40690	0,29614	0,23186	2,47968	0,19426	0,13905	2,49878	0,12542	0,08801
α_2(-1)	-0,96632	0,14442	0,11779	-0,99396	0,07819	0,06955	-1,00278	0,05023	0,04409
α_3(1,5)	1,46880	0,18398	0,14357	1,50237	0,10268	0,08481	1,50658	0,06733	0,05359
λ_1(2)	2,25132	0,79367	0,62487	2,04979	0,38211	0,31117	2,01302	0,22337	0,19003
λ_2(-1)	-1,17080	0,50065	0,37524	-1,04719	0,21966	0,19055	-1,02183	0,13045	0,11616
ν(3,5)	3,32045	1,23072	1,76287	3,72836	1,19054	1,22722	3,74821	0,84228	0,72276

(b) Algoritmo modificado

	n=100			n=300			n=750		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
$\beta_0(40)$	40,02062	0,34948	0,29631	40,00524	0,19107	0,17369	40,01382	0,11703	0,10786
β_1(-30)	-29,99979	0,03873	0,03605	-29,99962	0,02060	0,02102	-30,00141	0,01336	0,01284
β_2(15)	14,99985	0,06359	0,06055	14,99765	0,03544	0,03713	14,99947	0,02140	0,02344
α_1(2,5)	2,40690	0,29614	0,23186	2,47968	0,19426	0,13905	2,49878	0,12542	0,08801
α_2(-1)	-0,96632	0,14442	0,11779	-0,99395	0,07819	0,06955	-1,00278	0,05023	0,04409
α_3(1,5)	1,46879	0,18398	0,14357	1,50236	0,10268	0,08481	1,50658	0,06733	0,05359
λ_1(2)	2,25132	0,79367	0,62487	2,04980	0,38212	0,31117	2,01303	0,22337	0,19003
λ_2(-1)	-1,17080	0,50065	0,37524	-1,04719	0,21966	0,19055	-1,02183	0,13045	0,11616
ν(3,5)	3,32041	1,23065	1,76278	3,72830	1,19047	1,22715	3,74817	0,84223	0,72273

Na Tabela 9, vemos novamente estimativas dos parâmetros por ambos os algoritmos bastante semelhantes entre si. As diferenças estão no máximo na 5ª casa decimal para $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$, e no máximo na 4ª casa decimal para $\boldsymbol{\lambda}$ (destacados em azul) e $\boldsymbol{\tau} = \nu$ (destacado em amarelo). Quanto à função objetivo (também com valores na ordem das centenas ou milhares), os dois algoritmos geram diferenças nos valores máximos da ordem de 10^{-9}.

Aqui vale a mesma observação feita na distribuição anterior sobre o comportamento do desvio e do erro padrão. Porém, um fenômeno anômalo acontece nesta distribuição: a estimativa média do hiper-parâmetro se distancia do parâmetro real à medida que o tamanho n da amostra aumenta.

- Consistência da estimação

Tabela 10 – Consistência no modelo Slash normal assimétrico

(a) Algoritmo padrão

n=100				n=300				n=750			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,02062	0,35008	0,02284	0,26913	0,00524	0,19114	0,01312	0,13335	0,01382	0,11784	0,00828	0,07432
0,00021	0,03873	0,00366	0,02393	0,00038	0,02061	0,00098	0,01327	0,00141	0,01344	0,00145	0,00860
0,00016	0,06359	0,00374	0,04606	0,00235	0,03552	0,00465	0,02174	0,00053	0,02141	0,00080	0,01417
0,09310	0,31043	0,10725	0,18212	0,02032	0,19532	0,02268	0,13349	0,00122	0,12542	0,00046	0,08857
0,03368	0,14830	0,03068	0,08840	0,00605	0,07842	0,00613	0,05723	0,00278	0,05031	0,00054	0,03583
0,03121	0,18661	0,04365	0,11998	0,00236	0,10270	0,00988	0,05868	0,00658	0,06765	0,00586	0,04732
0,25132	0,83251	0,11967	0,44467	0,04980	0,38535	0,02741	0,23478	0,01302	0,22374	0,00371	0,16139
0,17080	0,52898	0,10719	0,27844	0,04719	0,22467	0,03222	0,12509	0,02183	0,13226	0,02204	0,07985
0,17959	1,24368	0,49906	0,73462	0,22831	1,21216	0,09414	0,65814	0,24821	0,87809	0,04255	0,47942

(b) Algoritmo modificado

n=100				n=300				n=750			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,02062	0,35008	0,02284	0,26913	0,00525	0,19114	0,01312	0,13335	0,01382	0,11784	0,00828	0,07432
0,00021	0,03873	0,00366	0,02393	0,00038	0,02061	0,00098	0,01327	0,00141	0,01344	0,00145	0,00860
0,00016	0,06359	0,00374	0,04606	0,00235	0,03552	0,00465	0,02174	0,00053	0,02141	0,00080	0,01417
0,09310	0,31043	0,10725	0,18211	0,02032	0,19532	0,02268	0,13350	0,00122	0,12542	0,00046	0,08857
0,03368	0,14830	0,03068	0,08841	0,00604	0,07842	0,00613	0,05723	0,00278	0,05031	0,00054	0,03583
0,03120	0,18661	0,04365	0,11998	0,00237	0,10270	0,00988	0,05867	0,00658	0,06765	0,00586	0,04732
0,25131	0,83251	0,11967	0,44467	0,04979	0,38534	0,02740	0,23478	0,01303	0,22375	0,00371	0,16138
0,17080	0,52898	0,10718	0,27845	0,04719	0,22467	0,03222	0,12508	0,02183	0,13226	0,02204	0,07986
0,17955	1,24375	0,49905	0,73462	0,22836	1,21225	0,09412	0,65816	0,24817	0,87803	0,04255	0,47941

O último fato elencado anteriormente poderia contrariar a consistência, mas vemos na Tabela 10 que as medidas baseadas na mediana tendem a 0 como é esperado. Isso é o que podemos constatar também pela análise das Figuras 23 e 24 relativamente a cada grupo de parâmetros no método modificado, pois o padrão traz quase os mesmos resultados.

Mais uma vez, na Figura 25, temos boxplots que revelam ganho no tempo de estimação com o algoritmo modificado: cerca de duas a duas vezes e meia mais rápido em média e um pouco mais homogêneo do que o algoritmo padrão. Neste caso, o ganho diretamente proporcional ao tamanho da amostra não é tão evidente, mas também ocorre.

- Comportamento por parâmetro

Figura 23 – Viés mediano por parâmetro da Slash normal assimétrica multivariada

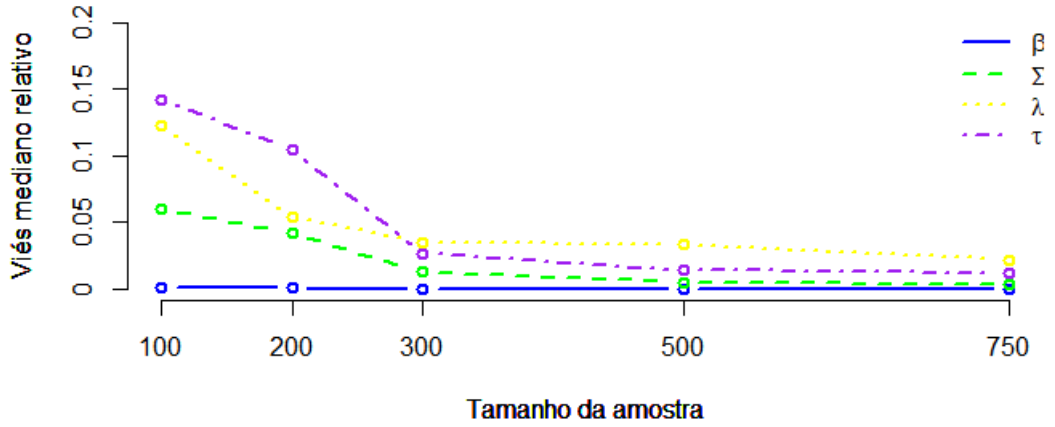
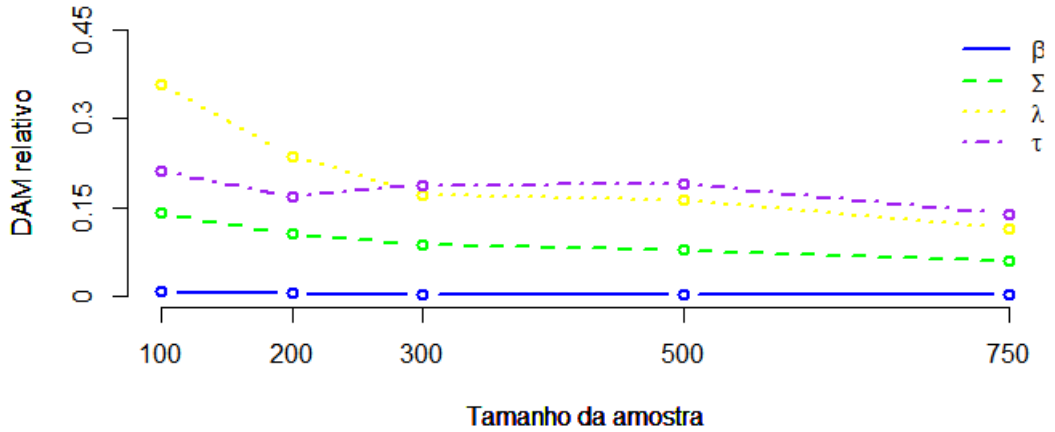


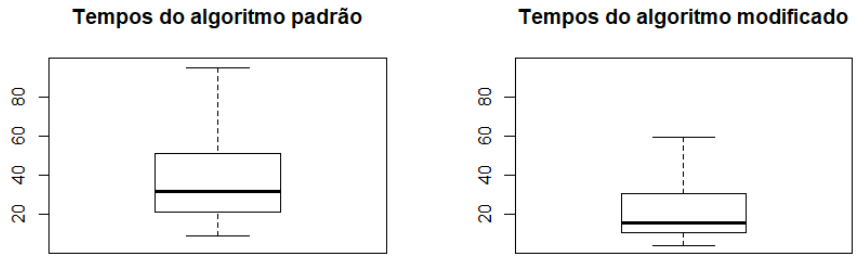
Figura 24 – Desvio absoluto mediano (DAM) por parâmetro da Slash normal assimétrica multivariada



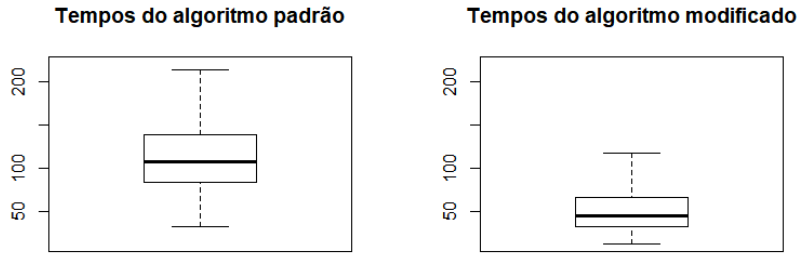
- Desempenho dos algoritmos

Figura 25 – Boxplots de tempos da Slash normal assimétrica multivariada

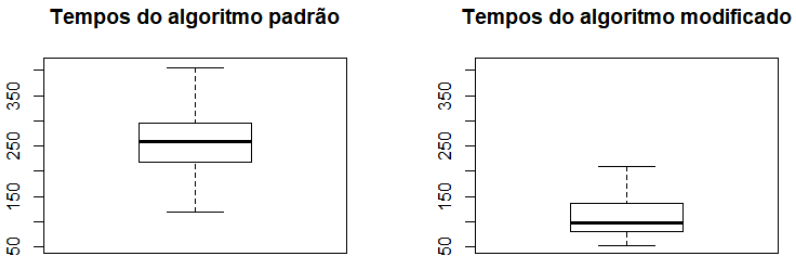
(a) Para n=100



(b) Para n=300



(c) Para n=750



Normal contaminada assimétrica

Antes de fornecer o novo tratamento proposto para este caso, observe que ao enxergarmos a normal contaminada assimétrica como mistura de escala a função Q decomposta na forma indicada em (3.25) não admite o mesmo tratamento dos casos anteriores. Isso ocorre, pois a função de probabilidade do fator de escala não pertence à família exponencial. Dessa forma, a saída adotada na literatura nesse exemplo para a estimação de todos os parâmetros é o uso do algoritmo ECME de Liu & Rubin (1994).

No entanto, estendendo a proposta de Lange & Sinsheimer (1993) da normal contaminada (simétrica) para esse caso, podemos expressar a última distribuição na forma indicada em (3.21). Com a notação introduzida na Subseção 2.3.4, podemos criar para cada indivíduo i a variável aleatória V_i assumindo um valor $j \in \{1, 2\}$ dependendo do componente da mistura (grupo) ao qual o indivíduo pertence.

Assim, ao considerarmos também o vetor aleatório $\mathbf{V}_i$ tal que $V_{ji} = 1 \Leftrightarrow V_i = j$, podemos escrever $\mathbf{Y}_{ij} \stackrel{d}{=} \mathbf{Y}_i \Big| V_{ji} = 1 \sim \text{SN}_p\Big(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j\Big)$, onde $\boldsymbol{\Sigma}_1 = \dfrac{\boldsymbol{\Sigma}}{\gamma}$, $\boldsymbol{\lambda}_1 = \dfrac{\boldsymbol{\lambda}}{\sqrt{\gamma}}$, $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ e $\boldsymbol{\lambda}_2 = \boldsymbol{\lambda}$.

Dessa forma, combinando a representação hierárquica de cada uma das duas normais assimétricas envolvidas com a que é dada em (2.17), obtemos a seguinte representação hierárquica para $\mathbf{Y}_i \sim \text{SCN}_p\Big(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma\Big)$:

$$\begin{aligned}
\mathbf{Y}_i | T_i = t_i, V_{ji} = 1 &\sim N_p\left(\mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Sigma}_j^{1/2}(\mathbf{I}_p + \boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^T)^{-1}\boldsymbol{\lambda}_j t_i, \boldsymbol{\Sigma}_j^{1/2}(\mathbf{I}_p + \boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^T)^{-1}\boldsymbol{\Sigma}_j^{1/2}\right); \\
T_i | V_{ji} = 1 &\sim TN_{[0,+\infty)}\left(0, 1 + \boldsymbol{\lambda}_j^T\boldsymbol{\lambda}_j\right); \\
\mathbf{V}_i &\sim \text{Mult}(1, \nu, 1-\nu).
\end{aligned} \tag{3.31}$$

Com isso, temos que o vetor de dados completos para cada indivíduo é dada por $\mathbf{y}_{Ci} = (\mathbf{y}_i, \mathbf{w}_i)$ com $\mathbf{w}_i = (v_{1i}, v_{2i}, t_i)$, o qual inclui as variáveis T, V_1 e V_2 sendo a soma dos valores das duas últimas sempre igual a 1 para cada indivíduo. Usando a representação (3.31) e fazendo algumas simplificações, chegamos de modo análogo ao exposto em Zeller, Cabral & Lachos (2015) à seguinte log-verossimilhança dos dados completos:

$$\begin{aligned}
\ell_C(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \ln\left\{2\left[\nu\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\frac{\boldsymbol{\Sigma}}{\gamma}\right)\right]^{v_{1i}}\left[(1-\nu)\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}\right)\right]^{v_{2i}}\phi_1\left(t_i;\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),1\right)\right\} \\
&= c-\frac{n}{2}\ln|\boldsymbol{\Sigma}|+\sum_{i=1}^{n}v_{1i}\left[\ln\nu+\frac{p}{2}\ln\gamma\right]+\sum_{i=1}^{n}v_{2i}\ln(1-\nu)-\frac{1}{2}\sum_{i=1}^{n}(\gamma v_{1i}+v_{2i})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}) \\
&+ \sum_{i=1}^{n}t_i\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})-\frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right]^2.
\end{aligned}$$

Mais uma vez, temos um valor $c = n\ln 2 - \frac{(n+1)p}{2}\ln 2\pi - \frac{1}{2}\sum\limits_{i=1}^{n}\ln t_i^2$ constante em relação aos parâmetros que não interfere na otimização da função ℓ_C. Para determinar $Q\left(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}\right) = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(\ell_C(\boldsymbol{\theta})|\mathbf{Y}_i=\mathbf{y}_i\right)$, devemos encontrar $\widehat{t}_i^{\,(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(T_i|\mathbf{Y}_i=\mathbf{y}_i\right)$ e $\widehat{v_{ji}}^{(k)} = E_{\widehat{\boldsymbol{\theta}}^{(k)}}\left(V_{ji}|\mathbf{Y}_i=\mathbf{y}_i\right)$ para $j\in\{1,2\}$.

Sendo a função de probabilidade conjunta de $(\mathbf{Y}_i,\mathbf{V}_i,T_i)$ dada por $\check{f}(\mathbf{y}_i,\mathbf{v}_i,t_i) = 2\left[\nu\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\frac{\boldsymbol{\Sigma}}{\gamma}\right)\right]^{v_{1i}}\left[(1-\nu)\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}\right)\right]^{v_{2i}}\phi_1\left(t_i;\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),1\right)$. Como $v_{1i},v_{2i}\in\{0,1\}$ e $v_{1i}+v_{2i}=1$, a função de probabilidade conjunta de $(\mathbf{Y}_i,T_i)$ é $\bar{f}(\mathbf{y}_i,t_i) = 2\left[\nu\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\frac{\boldsymbol{\Sigma}}{\gamma}\right)+(1-\nu)\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}\right)\right]\phi_1\left(t_i;\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),1\right)$ e daí a distribuição condicional de $T_i\Big|\mathbf{Y}_i=\mathbf{y}_i$ é caracterizada por

$$\begin{aligned}
\widetilde{f}(t_i|\mathbf{y}_i) &= \frac{\bar{f}(\mathbf{y}_i,t_i)}{\bar{f}(\mathbf{y}_i)} = \frac{2\left[\nu\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\frac{\boldsymbol{\Sigma}}{\gamma}\right)+(1-\nu)\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}\right)\right]\phi_1\left(t_i;\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),1\right)}{2\left[\nu\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\frac{\boldsymbol{\Sigma}}{\gamma}\right)+(1-\nu)\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}\right)\right]\Phi_1\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),1\right)} \\
&= \frac{\phi_1\left(t_i;\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),1\right)}{\Phi_1\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),1\right)}.
\end{aligned}$$

Como na Proposição 2.3.3 temos $T_i\Big|\mathbf{Y}_i=\mathbf{y}_i \sim TN_{[0,+\infty)}\Big(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}),1\Big)$ e, portanto, $\widehat{t}_i^{\,(k)} = \widehat{\boldsymbol{\lambda}}^{(k)^T}\widehat{\boldsymbol{\Sigma}}^{(k)^{-\frac{1}{2}}}(\mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}) + W_\Phi\Big(\widehat{\boldsymbol{\lambda}}^{(k)^T}\widehat{\boldsymbol{\Sigma}}^{(k)^{-\frac{1}{2}}}(\mathbf{y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)})\Big)$. Isso significa que chegamos ao mesmo resultado auferido anteriormente para qualquer mistura de escala assimétrica.

Por outro lado, fazendo $\nu_1=\nu$ e $\nu_2=1-\nu$, temos para cada $j\in\{1,2\}$ que a função de probabilidade conjunta de $(\mathbf{Y}_i,V_{ji},T_i)$ é expressa por $\breve{g}(\mathbf{y}_i,v_{ji},t_i) = 2\nu_j\phi_p\left(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}_j\right)\phi_1\left(t_i;\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),1\right)$. Assim, a distribuição conjunta de

$(\mathbf{Y}_i, V_{ji})$ é $\bar{g}(\mathbf{y}_i, v_{ji}) = 2\nu_j\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_j\right)\Phi_1\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), 1\right)$, donde concluímos que a distribuição condicional de $V_{ji}\Big|\mathbf{Y}_i = \mathbf{y}_i$ é caracterizada por

$$\begin{aligned}\widetilde{g}(v_{ji}|\mathbf{y}_i) &= \frac{\bar{g}(\mathbf{y}_i, v_{ji})}{f(\mathbf{y}_i)} = \frac{2\nu_j\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_j\right)\Phi_1\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), 1\right)\mathbb{I}_{\{1\}}(v_{ji})}{2\left[\nu\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}, \frac{\boldsymbol{\Sigma}}{\gamma}\right) + (1-\nu)\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)\right]\Phi_1\left(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), 1\right)} \\ &= \frac{\nu_j\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_j\right)}{\nu\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}, \frac{\boldsymbol{\Sigma}}{\gamma}\right) + (1-\nu)\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)}\mathbb{I}_{\{1\}}(v_{ji}).\end{aligned}$$

Como a distribuição deduzida acima é discreta, vemos de maneira imediata que $\widehat{v_{ji}}^{(k)} = \dfrac{\widehat{\nu_j}^{(k)}\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{\Sigma}_j}^{(k)}\right)}{\widehat{\nu}^{(k)}\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}, \frac{\widehat{\boldsymbol{\Sigma}}^{(k)}}{\widehat{\gamma}^{(k)}}\right) + (1-\widehat{\nu}^{(k)})\phi_p\left(\mathbf{y}_i; \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{\Sigma}}^{(k)}\right)};\ j \in \{1, 2\}.$

Pode-se observar que da distribuição condicional de $U_i\Big|\mathbf{Y}_i = \mathbf{y}_i$ para a distribuição normal contaminada assimétrica deduzida na Subseção 3.2.1 (e adaptada ao nosso caso com regressão) segue que $\widehat{u_i}^{(k)} = \widehat{\gamma}^{(k)}\widehat{v_{1i}}^{(k)} + \widehat{v_{2i}}^{(k)}$. Com esse resultado, ao fixarmos $\gamma = \widehat{\gamma}^{(k)}$ na função $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right)$ obtida agora, recuperamos exatamente a mesma função Q_1 oriunda da configuração geral das misturas de escala.

Dessa forma, mediante a imposição mencionada acima, obtemos na etapa $k+1$ as mesmas estimativas $\boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\Sigma}^{(k+1)}$ e $\boldsymbol{\lambda}^{(k+1)}$ provenientes do tratamento da distribuição normal contaminada assimétrica como mistura de escala. Olhando novamente para a função Q originalmente construída a partir do tratamento da distribuição em questão como mistura finita, chamaremos de Q_2 a parte da função Q que depende do hiper-parâmetro $\boldsymbol{\tau} = (\nu, \gamma)$, ou seja,

$$Q_2\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right) = \ln(\nu)\sum_{i=1}^{n}\widehat{v_{1i}}^{(k)} + \ln(1-\nu)\sum_{i=1}^{n}\widehat{v_{2i}}^{(k)} + \frac{p}{2}\ln(\gamma)\sum_{i=1}^{n}\widehat{v_{1i}}^{(k)} - \frac{\gamma}{2}\sum_{i=1}^{n}\widehat{v_{1i}}^{(k)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

Supondo na função acima obtidos $\boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\Sigma}^{(k+1)}$ e $\boldsymbol{\lambda}^{(k+1)}$, considere $d_i^{(k+1)} = \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}^{(k+1)}\right)^T \boldsymbol{\Sigma}^{(k+1)^{-1}} \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}^{(k+1)}\right)$. Assim, $\dfrac{\partial Q_2}{\partial \boldsymbol{\tau}}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\right) = \mathbf{0}$ nos fornece

$$\begin{cases} \dfrac{\partial Q_2}{\partial \nu}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\right) = 0 \Rightarrow \dfrac{1}{\nu}\sum\limits_{i=1}^{n} \widehat{v_{1i}}^{(k)} - \dfrac{1}{1-\nu}\sum\limits_{i=1}^{n} \widehat{v_{2i}}^{(k)} = 0 \\ \dfrac{\partial Q_2}{\partial \gamma}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}\right) = 0 \Rightarrow \dfrac{p}{2\gamma}\sum\limits_{i=1}^{n} \widehat{v_{1i}}^{(k)} - \dfrac{1}{2}\sum\limits_{i=1}^{n} \widehat{v_{1i}}^{(k)} \widehat{d}_i^{\,(k+1)} = 0 \end{cases}$$

Resolvendo as equações acima, obtemos estimativas explícitas para ν e γ na etapa $k+1$ do algoritmo EM como seguem:

$$\widehat{\nu}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{v_{1i}}^{(k)} \quad \text{e} \quad \widehat{\gamma}^{(k+1)} = \frac{p\sum\limits_{i=1}^{n} \widehat{v_{1i}}^{(k)}}{\sum\limits_{i=1}^{n} \widehat{v_{1i}}^{(k)} \widehat{d}_i^{\,(k+1)}}.$$

Note que embora tenhamos estimativas explícitas para todos os parâmetros, não recaímos num EM clássico pelo fato de termos que supor alguns parâmetros constantes para obter outros. O mesmo ocorre também nos dois outros exemplos anteriores (relativamente aos parâmetros das respectivas funções Q_1), revelando que todos esses algoritmos, mesmo modificados de modo a parecerem um EM clássico, são ainda enquadrados na categoria ECM proposta por Meng & Rubin (1993).

Finalmente, antes de começar propriamente o estudo de simulação neste exemplo, enfatizamos que os valores iniciais para o hiper-parâmetro serão os mesmos adotados no caso univariado, isto é, $\boldsymbol{\tau}^{(0)} = (0,5;\ 0,5)$. Porém, dada a complexidade maior do processo de estimação num modelo multivariado, tomaremos o cuidado de tratar o problema conhecido como "troca de rótulos", que é mencionado em Mclachlan & Peel (2000).

Resumidamente, tal problema inerente a misturas finitas em geral consiste na estimação de proporções de misturas intercambiadas entre os diferentes componentes que a constituem, pois o algoritmo de otimização não é capaz de identificar claramente a qual grupo cada indivíduo pertence.

Como no nosso caso, há apenas dois grupos e a proporção a ser recuperada é $\nu = 0,25$, controlaremos esse fenômeno eliminando no estudo de simulação as amostras cuja estimativa de ν no EM for maior do que 0,5. Dito isso, vamos aos resultados das simulações.

- Recuperação dos parâmetros

Tabela 11 – Recuperação dos parâmetros na normal contaminada assimétrica

(a) Algoritmo padrão

	n=100			n=300			n=750		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
$\beta_0(40)$	40,03815	0,32364	0,28701	40,02651	0,18997	0,16987	40,00555	0,10860	0,10495
$\beta_1(-30)$	-30,00228	0,03747	0,03593	-30,00107	0,02094	0,02049	-29,99998	0,01303	0,01297
$\beta_2(15)$	14,99681	0,07166	0,06250	15,00187	0,03698	0,03631	15,00045	0,02256	0,02286
$\alpha_1(2,5)$	2,40578	0,34691	0,23078	2,47528	0,20884	0,14392	2,49506	0,17367	0,09400
$\alpha_2(-1)$	-0,97874	0,15150	0,11944	-1,00088	0,09891	0,07240	-0,99999	0,07139	0,04672
$\alpha_3(1,5)$	1,47390	0,20357	0,14376	1,50805	0,13425	0,08889	1,49795	0,09231	0,05758
$\lambda_1(2)$	2,17610	0,77079	0,60098	2,05814	0,40819	0,31206	2,02435	0,24602	0,19225
$\lambda_2(-1)$	-1,13485	0,40928	0,36756	-1,05178	0,24228	0,19164	-1,01770	0,13307	0,11751
$\nu(0,25)$	0,23318	0,14237	0,17194	0,22991	0,12410	0,12153	0,24730	0,11108	0,08416
$\gamma(0,4)$	0,31108	0,09392	0,16055	0,35826	0,07388	0,10743	0,37789	0,05497	0,06200

(b) Algoritmo modificado

	n=100			n=300			n=750		
Paramêtro (real)	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão	Média das estimativas	Desvio padrão	Erro padrão
$\beta_0(40)$	40,03814	0,32364	0,28701	40,02651	0,18997	0,16987	40,00554	0,10860	0,10495
$\beta_1(-30)$	-30,00228	0,03747	0,03593	-30,00107	0,02094	0,02049	-29,99998	0,01303	0,01297
$\beta_2(15)$	14,99681	0,07166	0,06250	15,00187	0,03698	0,03631	15,00045	0,02256	0,02286
$\alpha_1(2,5)$	2,40577	0,34692	0,23078	2,47526	0,20884	0,14392	2,49502	0,17368	0,09400
$\alpha_2(-1)$	-0,97873	0,15150	0,11944	-1,00087	0,09891	0,07240	-0,99998	0,07140	0,04672
$\alpha_3(1,5)$	1,47390	0,20357	0,14376	1,50803	0,13425	0,08889	1,49792	0,09232	0,05758
$\lambda_1(2)$	2,17611	0,77081	0,60099	2,05815	0,40821	0,31207	2,02434	0,24602	0,19225
$\lambda_2(-1)$	-1,13485	0,40928	0,36756	-1,02373	0,24228	0,19164	-1,01768	0,13307	0,11750
$\nu(0,25)$	0,23319	0,14238	0,17194	0,22993	0,12411	0,12154	0,24734	0,11110	0,08417
$\gamma(0,4)$	0,31108	0,09393	0,16054	0,35827	0,07389	0,10742	0,37789	0,05497	0,06200

A Tabela 11 mostra que esta distribuição, mais do que as duas anteriores, possui estimativas extremamente parecidas pelos dois algoritmos com diferenças em todos os parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$ – destacados em azul – e $\boldsymbol{\tau} = (\nu, \gamma)$ – destacados em amarelo – que não passam da 5ª casa decimal. Já as diferenças nos valores máximos da função objetivo (assumindo valores na ordem das centenas ou milhares) são da ordem de 10^{-9}.

Outra vez o comportamento do desvio e do erro padrão está dentro do esperado à medida que o tamanho n da amostra aumenta. Para confirmar novamente a consistência dos dois algoritmos, vejamos a Tabela 12.

- Consistência da estimação

Tabela 12 – Consistência no modelo normal contaminado assimétrico

(a) Algoritmo padrão

n=100				n=300				n=750			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,03814	0,32588	0,02554	0,21901	0,02651	0,19181	0,02991	0,12090	0,00555	0,10874	0,00630	0,06913
0,00228	0,03754	0,00196	0,02593	0,00107	0,02097	0,00042	0,01172	0,00002	0,01303	0,00134	0,00946
0,00319	0,07174	0,00629	0,04939	0,00187	0,03702	0,00305	0,02403	0,00045	0,02256	0,00077	0,01629
0,09422	0,35948	0,10123	0,23951	0,02472	0,21029	0,05880	0,13096	0,00494	0,17374	0,00161	0,11614
0,02126	0,15298	0,03026	0,08367	0,00088	0,09891	0,00161	0,06392	0,00001	0,07139	0,00086	0,04796
0,02610	0,20523	0,04359	0,11357	0,00805	0,13449	0,00407	0,08545	0,00205	0,09234	0,00272	0,05397
0,17610	0,79065	0,05798	0,45731	0,05814	0,41231	0,00774	0,28062	0,02435	0,24722	0,00980	0,15523
0,13485	0,43092	0,08022	0,26864	0,05178	0,24775	0,03154	0,14745	0,01770	0,13423	0,00578	0,08211
0,01682	0,14336	0,03295	0,11425	0,02009	0,12571	0,02742	0,09314	0,00270	0,11111	0,02192	0,07985
0,08892	0,12933	0,09612	0,06791	0,04174	0,08486	0,05105	0,04893	0,02211	0,05925	0,03016	0,03838

(b) Algoritmo modificado

n=100				n=300				n=750			
VME	EQM	VMD	DAM	VME	EQM	VMD	DAM	VME	EQM	VMD	DAM
0,03814	0,32588	0,02554	0,21901	0,02651	0,19181	0,02991	0,12090	0,00554	0,10874	0,00630	0,06913
0,00228	0,03754	0,00196	0,02593	0,00107	0,02097	0,00042	0,01172	0,00002	0,01303	0,00134	0,00946
0,00319	0,07174	0,00629	0,04939	0,00187	0,03702	0,00305	0,02403	0,00045	0,02256	0,00077	0,01629
0,09423	0,35949	0,10125	0,23949	0,02474	0,21030	0,05900	0,13096	0,00498	0,17375	0,00164	0,11611
0,02127	0,15298	0,03024	0,08363	0,00087	0,09891	0,00161	0,06391	0,00001	0,07140	0,00085	0,04797
0,02610	0,20523	0,04360	0,11352	0,00803	0,13449	0,00407	0,08542	0,00208	0,09234	0,00273	0,05400
0,17611	0,79067	0,05798	0,45731	0,05815	0,41233	0,00776	0,28060	0,02434	0,24722	0,00981	0,15525
0,13485	0,43093	0,08022	0,26864	0,05177	0,24775	0,03135	0,14744	0,01768	0,13424	0,00583	0,08214
0,01681	0,14337	0,03296	0,11425	0,02007	0,12572	0,02739	0,09314	0,00266	0,11113	0,02191	0,07986
0,08892	0,12934	0,09613	0,06791	0,04173	0,08486	0,05105	0,04891	0,02211	0,05925	0,03015	0,03838

A Tabela 12 apresenta apenas uma medida fora do esperado: o viés mediano da estimação do parâmetro ν aumenta ligeiramente para $n = 750$. Essa situação também é perceptível no viés mediano relativo do hiper-parâmetro completo conforme a Figura 26. Quanto à variabilidade mediana, não há surpresas de acordo com a Figura 27. Mais uma vez, os gráficos são somente para o método modificado, pois o padrão gera resultados praticamente iguais.

Na Figura 28, os boxplots novamente revelam que o algoritmo modificado é cerca de 2 a 2,5 vezes mais rápido em média e um pouco mais homogêneo que o padrão. Neste caso, o ganho é inversamente proporcional ao tamanho da amostra, em virtude do número (e também do seu aumento com n) bem maior de iterações no algoritmo modificado em relação ao padrão.

- Comportamento por parâmetro

Figura 26 – Viés mediano por parâmetro da normal contaminada assimétrica multivariada

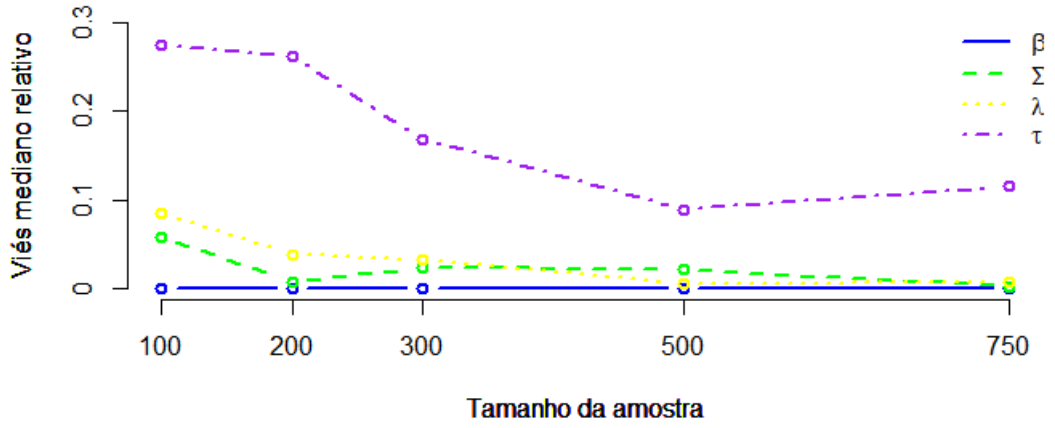
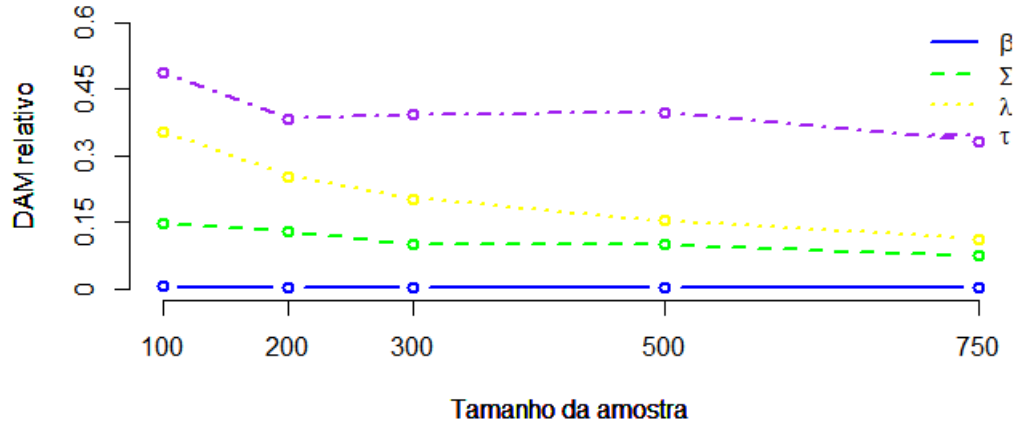


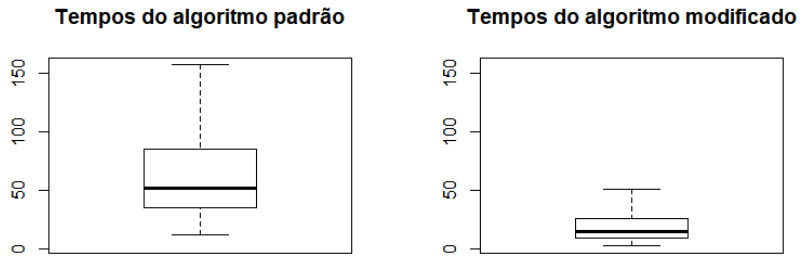
Figura 27 – Desvio absoluto mediano (DAM) por parâmetro da normal contaminada assimétrica multivariada



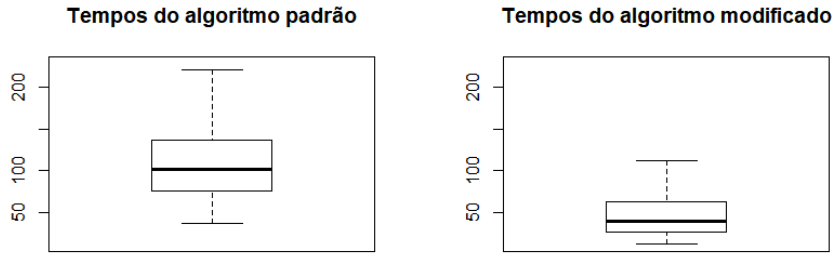
- Desempenho dos algoritmos

Figura 28 – Boxplots de tempos da normal contaminada assimétrica multivariada

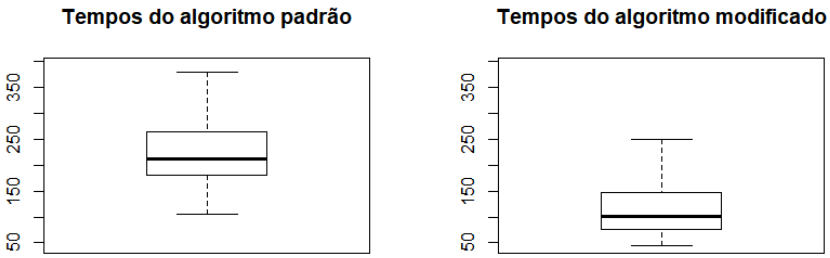
(a) Para n=100



(b) Para n=300



(c) Para n=750



3.2.3.2 *Estudo de Dados Reais*

Como aplicação para as distribuições SSMN, trabalharemos com um conjunto de dados provenientes do censo de 2010 do IBGE. A amostra que usaremos é constituída por dados dos 853 municípios de Minas Gerais, sendo a fonte dos mesmos o Atlas do Desenvolvimento Humano no Brasil – Atlas (2016), uma iniciativa do Programa das Nações Unidas para o Desenvolvimento (PNUD).

Ajustaremos um modelo de regressão multivariada para tentar explicar a esperança de vida e a renda *per capita* dos municípios mineiros com base nas taxas de atividade e analfabetismo e na razão de dependência do ano de 2010. Na sequência, faremos uma breve explanação acerca das variáveis relacionadas e também uma pequena análise exploratória.

- Respostas

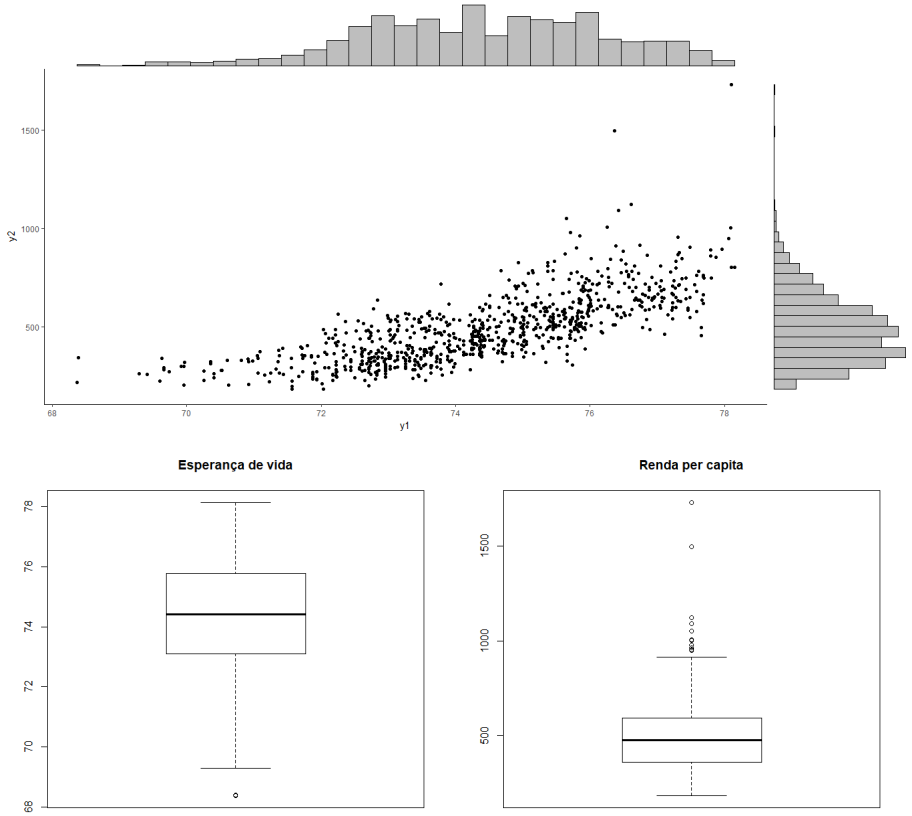
  1. Esperança de vida (Y_1): expectativa de anos de vida ao nascer.
  2. Renda *per capita* (Y_2): razão entre o Produto Interno Bruto (PIB) e a respectiva população total.

Tabela 13 – Medidas descritivas básicas das respostas

	Média	Mediana	Desvio padrão	Assimetria	Curtose
Y_1	74,42	74,41	1,7902	-0,2651	2,7735
Y_2	490,6	475,2	173,0778	1,1731	7,0184

Os valores de assimetria relativamente distantes de 0 (coeficiente de assimetria da normal) de ambas as variáveis justificam a adoção de um modelo assimétrico, enquanto os valores de curtose relativamente distantes de 3 (coeficiente de curtose da normal) apontam para a presença de *outliers* e, portanto, as distribuições SSMN são recomendáveis para sua modelagem. Há de se salientar também a relevância da relação entre as duas variáveis, pois $\text{cov}(Y_1, Y_2) = 230,977$ e $\text{cor}(Y_1, Y_2) = 0,74547$. Na Figura 29, vemos um panorama sobre o comportamento conjunto das duas variáveis.

Figura 29 – Análise exploratória das respostas



- Explicativas

    1. Taxa de atividade (X_1): porcentagem de pessoas ocupadas com mais de 18 relativa ao total nessa faixa etária.
    2. Razão de dependência (X_2): razão entre a população com menos de 15 anos ou mais de 64 anos e a população entre 15 e 64 anos.
    3. Taxa de analfabetismo (X_3): porcentagem de pessoas com mais de 18 anos que não sabem ler nem escrever relativa ao total nessa faixa etária.

A Tabela 14 e a Figura 30 mostram as relações existentes entre os três pares de variáveis explicativas.

Tabela 14 – Análise exploratória das explicativas

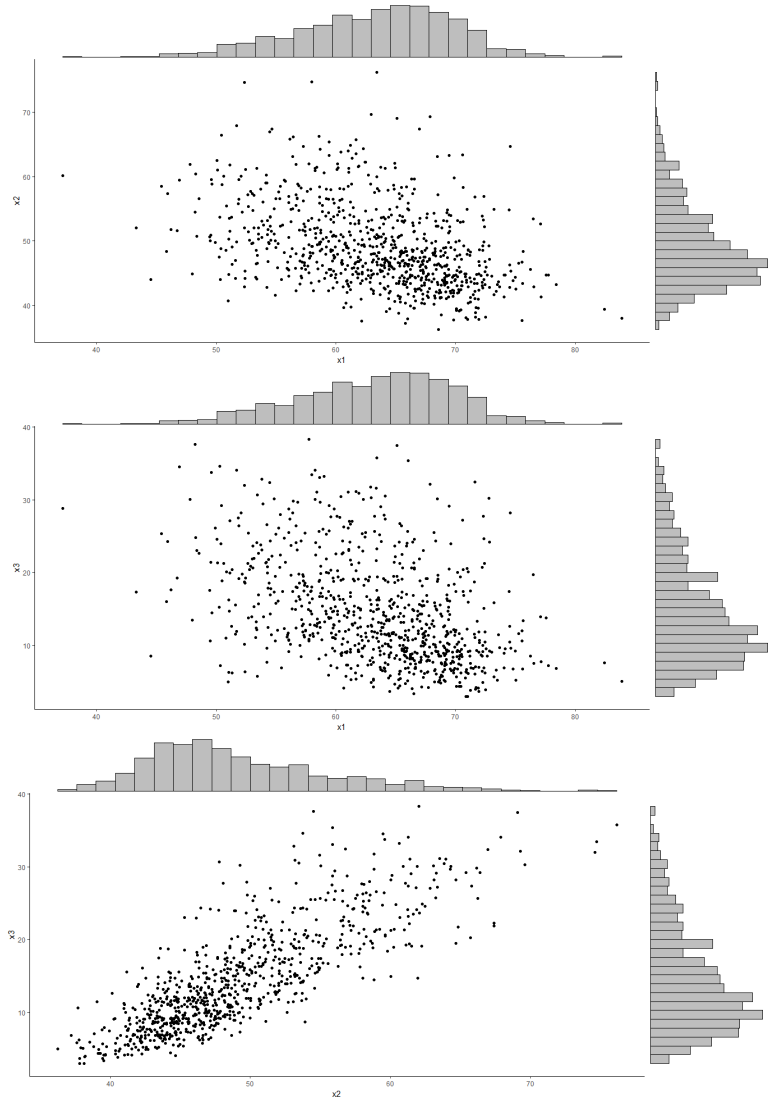
(a) Medidas descritivas básicas

	Média	Mediana	Desvio padrão
X_1	63,32	64,07	6,5078
X_2	49,06	47,73	6,4164
X_3	14,48	12,76	7,1125

(b) Tábua de correlações

	X_1	X_2	X_3
X_1	1	-0,410	-0,424
X_2	-0,410	1	0,819
X_3	-0,424	0,819	1

Figura 30 – Diagramas de dispersão entre os pares de variáveis explicativas



- Respostas × Explicativas

Tabela 15 – Correlações entre respostas e explicativas

	X_1	X_2	X_3
Y_1	0,358233	-0,650615	-0,697514
Y_2	0,492605	-0,725246	-0,757763

A fim de realizarmos o processo de estimação dos parâmetros de um modelo de regressão linear com duas variáveis respostas, consideraremos inicialmente a seguinte estrutura para a matriz de planejamento contendo as três variáveis explicativas de cada indivíduo (município) i :

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}_i^T & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_i^T \end{bmatrix} \tag{3.32}$$

Na composição da matriz (3.32), as três últimas entradas do vetor $\mathbf{x}_i^T = [1\ x_{1i}\ x_{2i}\ x_{3i}]$ correspondem, respectivamente, aos valores observados para o município i das variáveis taxa de atividade, razão de dependência e taxa de analfabetismo.

Como justificado anteriormente, a proposta é ajustar um modelo de regressão com cada distribuição da família SSMN. Contudo, também compararemos os ajustes feitos com as distribuições normal e normal assimétrica para ilustrar os ganhos no que diz respeito à qualidade do ajuste.

Em consonância com o que foi apresentado ao longo deste capítulo, adotaremos as notações $\boldsymbol{\beta} = [\beta_{01}\ \beta_{11}\ \beta_{21}\ \beta_{31}\ \beta_{02}\ \beta_{12}\ \beta_{22}\ \beta_{32}]^T$ para o parâmetro da regressão, $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2\ \alpha_3]^T$ para o parâmetro de escala, $\boldsymbol{\lambda} = [\lambda_1\ \lambda_2]^T$ para o parâmetro de assimetria (exceto no modelo normal), ν para o primeiro hiper-parâmetro (exceto nos modelos normal e normal assimétrico) e γ para o segundo hiper-parâmetro (somente no modelo normal contaminado assimétrico).

Além disso, indicaremos os modelos em questão pela sigla entre parênteses: normal (N), normal assimétrico (SN), T-Student normal assimétrico (STN), Slash normal assimétrico (SSN) e normal contaminado assimétrico (SCN).

Feitas essas convenções, apresentamos na Tabela 16 os resultados das estimativas de máxima verossimilhança dos parâmetros em cada um dos modelos via algoritmo EM seguidas do respectivo erro padrão.

Tabela 16 – Estimativas de máxima verossimilhança dos parâmetros do modelo

(a) Coeficientes da regressão

Modelo	$\widehat{\beta}_{01}$	$\widehat{\beta}_{11}$	$\widehat{\beta}_{21}$	$\widehat{\beta}_{31}$	$\widehat{\beta}_{02}$	$\widehat{\beta}_{12}$	$\widehat{\beta}_{22}$	$\widehat{\beta}_{32}$
N	78,322	0,016	-0,064	-0,122	706,695	4,978	-7,618	-10,883
	(0,724)	(0,007)	(0,012)	(0,011)	(60,245)	(0,611)	(0,977)	(0,888)
SN	78,249	0,015	-0,064	-0,122	602,744	4,049	-6,582	-10,521
	(0,729)	(0,007)	(0,012)	(0,011)	(50,637)	(0,502)	(0,830)	(0,750)
STN	78,042	0,018	-0,060	-0,122	616,686	4,140	-6,528	-10,267
	(0,711)	(0,007)	(0,011)	(0,010)	(48,682)	(0,488)	(0,792)	(0,711)
SSN	78,110	0,018	-0,061	-0,122	617,343	4,158	-6,632	-10,263
	(0,717)	(0,007)	(0,011)	(0,011)	(48,859)	(0,490)	(0,795)	(0,714)
SCN	78,151	0,017	-0,062	-0,122	619,350	4,110	-6,733	-10,276
	(0,727)	(0,007)	(0,012)	(0,011)	(49,299)	(0,493)	(0,802)	(0,719)

(b) Outros parâmetros

Modelo	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\nu}$	$\widehat{\gamma}$
N	1,134	0,536	104,371	-	-	-	-
	(0,024)	(0,040)	(2,260)				
SN	1,181	0,429	149,783	-0,909	3,325	-	-
	(0,029)	(0,079)	(4,315)	(0,178)	(0,350)		
STN	1,052	0,374	108,784	-0,713	1,741	7,561	-
	(0,033)	(0,069)	(4,217)	(0,146)	(0,222)	(1,132)	
SSN	0,923	0,321	96,605	-0,643	1,621	2,273	-
	(0,032)	(0,066)	(3,878)	(0,140)	(0,199)	(0,223)	
SCN	1,136	0,412	123,913	-0,792	2,228	0,026	0,132
	(0,035)	(0,091)	(4,550)	(0,186)	(0,260)	(0,013)	(0,046)

Observe que de modo geral a variabilidade da estimação dos coeficientes da regressão (expressa pelo erro padrão das estimativas) é menor nos modelos da família SSMN. Sendo esses os parâmetros de maior interesse, pode-se afirmar que tal fato é uma vantagem da utilização dos modelos propostos.

Após a estimação, perguntamo-nos se os parâmetros obtidos são significativos. Detalhes sobre a significância em um MRL normal podem ser encontrados em Louredo (2016). Nas demais distribuições não é possível avaliar a significância por meio de uma estatística com distribuição exata como no caso anterior. No entanto, o uso dos estimadores de máxima verossimilhança (EMV) nos permite recorrer a suas propriedades, dentre as quais a *normalidade assintótica*.

Uma discussão aprofundada sobre o assunto com algumas demonstrações se encontra em Ritter (2015), mas aqui nos limitaremos a mencionar um fato decorrente dessa discussão: denotando o vetor de parâmetros do modelo de interesse por $\boldsymbol{\theta}$ e a respectiva matriz de informação de Fisher observada por $\mathfrak{I} = \mathfrak{K}^{-1}$, temos que para cada coordenada θ_j de $\boldsymbol{\theta}$, a estatística $\widetilde{\theta}_j = \widehat{\theta}_j / \sqrt{\mathfrak{K}_{jj}}$ (razão entre cada estimativa e seu respectivo erro padrão) provém de uma distribuição que segue assintoticamente uma normal padrão sob a hipótese nula do teste de significância ($\theta_j = 0$).

Com isso, espera-se que o parâmetro seja significativo se tivermos o valor observado da referida estatística com módulo maior do que 2. Na Tabela 17, vemos os valores dessa estatística de teste para cada parâmetro.

Tabela 17 – Significância dos parâmetros do modelo

(a) Parâmetros de regressão

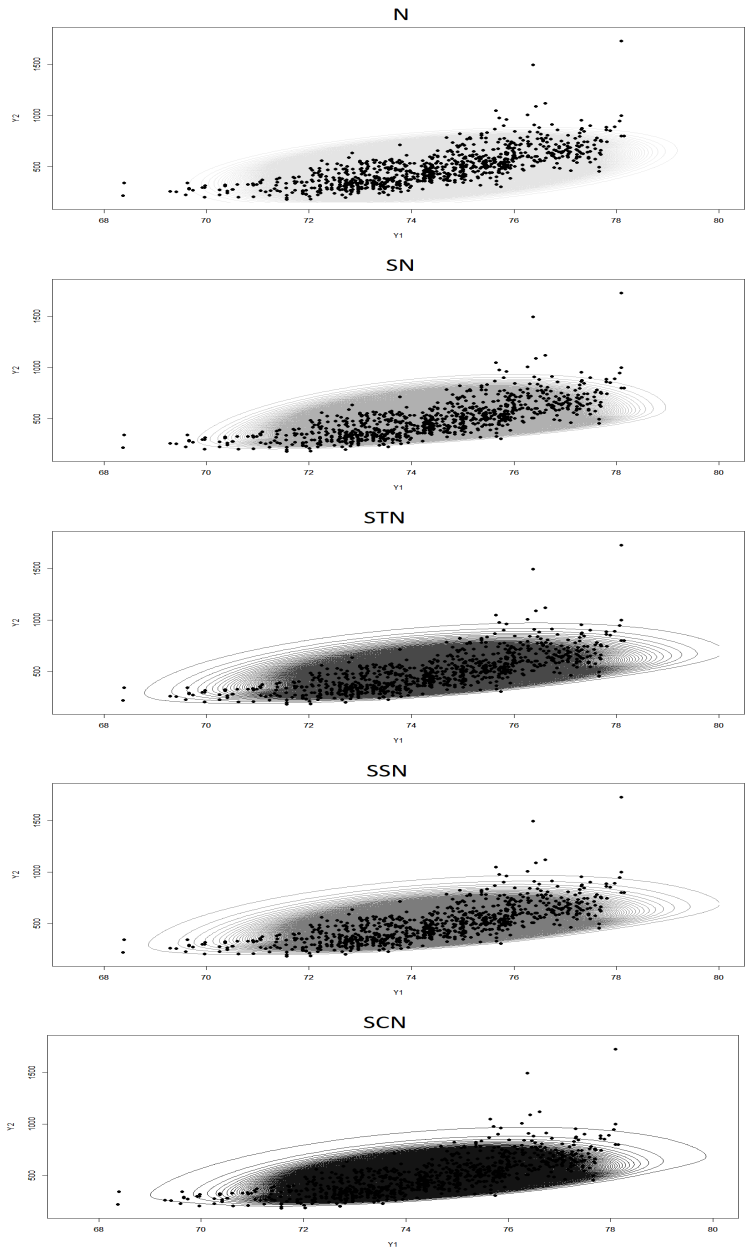
Modelo	$\widetilde{\beta}_{01}$	$\widetilde{\beta}_{11}$	$\widetilde{\beta}_{21}$	$\widetilde{\beta}_{31}$	$\widetilde{\beta}_{02}$	$\widetilde{\beta}_{12}$	$\widetilde{\beta}_{22}$	$\widetilde{\beta}_{32}$
N	108,183	2,191	-5,470	-11,422	11,730	8,145	-7,795	-12,256
SN	107,297	2,090	-5,393	-11,399	11,903	8,064	-7,928	-14,037
STN	109,822	2,526	-5,254	-11,631	12,668	8,489	-8,242	-14,444
SSN	109,009	2,438	-5,293	-11,581	12,635	8,485	-8,340	-14,383
SCN	107,449	2,309	-5,311	-11,551	12,563	8,336	-8,397	-14,295

(b) Outros parâmetros

Modelo	$\widetilde{\alpha}_1$	$\widetilde{\alpha}_2$	$\widetilde{\alpha}_3$	$\widetilde{\lambda}_1$	$\widetilde{\lambda}_2$	$\widetilde{\nu}$	$\widetilde{\gamma}$
N	46,288	13,441	46,179	-	-	-	-
SN	40,075	5,431	34,710	-5,110	9,511	-	-
STN	31,737	5,397	25,795	-4,886	7,829	6,682	-
SSN	29,067	4,826	24,910	-4,584	8,129	10,206	-
SCN	32,679	4,503	27,232	-4,250	8,577	1,955	2,895

Pelo critério anterior, pode-se considerar todos os parâmetros significativos. Porém, esse método não é eficaz para comparar a significância dos hiper-parâmetros em diferentes modelos. Esse tipo de comparação pode ser feito pela avaliação do quão bem cada modelo ajusta os dados nos gráficos de contorno da Figura 31.

Figura 31 – Gráficos de contorno das densidades médias dos modelos ajustados



Nas curvas de nível apresentadas (com níveis iguais para todos os modelos), maior densidade de linhas significa mais probabilidade na região. Como se pode observar facilmente, a distribuição normal (N) confere baixíssima probabilidade para os pontos correspondentes às respostas que se encontram nas partes inferior esquerda e superior direita. A distribuição normal assimétrica (SN) dá um pouco mais de probabilidade para os pontos extremos da parte inferior à esquerda (correspondentes a baixas expectativa de vida e renda), mas não produz diferença relevante quanto aos pontos extremos da parte superior à direita (correspondentes a altas expectativa de vida e renda). Por outro lado, as três distribuições da família SSMN promovem maior equilíbrio na atribuição da probabilidade inclusive entre os pontos extremos.

No caso das distribuições SSMN, é difícil dizer qual dentre elas está melhor adequada aos dados apenas com base na análise gráfica. Sendo assim, faz-se necessário recorrer aos critérios de informação AIC e BIC para selecionar o melhor modelo. Tais critérios de seleção são baseados nas seguintes estatísticas em termos da log-verossimilhança ℓ, da estimativa paramétrica $\widehat{\boldsymbol{\theta}}$, do comprimento r do vetor de parâmetros $\boldsymbol{\theta}$ e do tamanho n da amostra:

$$\begin{aligned} \text{AIC} &= -2\ell\Big(\widehat{\boldsymbol{\theta}}\Big) + 2r; \\ \text{BIC} &= -2\ell\Big(\widehat{\boldsymbol{\theta}}\Big) + r\ln n. \end{aligned} \tag{3.33}$$

A melhor qualidade de ajuste seria esperada para o modelo com a maior verossimilhança, mas devemos lembrar que estaremos comparando modelos distintos. Ambos os critérios de informação permitem uma comparação dos referidos modelos quanto à parcimônia ao penalizarem o excesso de parâmetros, sendo que o segundo penaliza proporcionalmente ao tamanho da amostra. Note que devido à convenção de sinais, o melhor modelo será aquele com menor AIC ou BIC. A Tabela 18 traz os valores da log-verossimilhança e das estatísticas para os três modelos em discussão.

Tabela 18 – Critérios de seleção do melhor modelo SSMN

Modelo	SSN	STN	SCN
Log-verossimilhança	-6367,640	-6368,894	-6370,446
AIC	12763,280	12765,788	12770,892
BIC	12829,763	12832,270	12842,124

Concluímos assim que a distribuição Slash normal assimétrica (SSN) é a que melhor se ajusta aos dados. Para finalizar este capítulo, vejamos algumas interpretações relevantes que podem ser extraídas do modelo selecionado. Inicialmente, temos que o viés aditivo da estimativa $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$ expresso em (3.23) é dado por:

$$\widehat{E(\boldsymbol{\epsilon}_i)} = \sqrt{\frac{2}{\pi}}E(\widehat{\Upsilon})\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\widehat{\boldsymbol{\lambda}} = \begin{bmatrix} -0,039124 \\ 83,533717 \end{bmatrix}.$$

Esse viés, que é nulo no modelo normal, justifica o deslocamento em direção aos pontos extremos no modelo SSN e também nas demais distribuições distintas da normal. De fato, vemos em particular para o nosso modelo escolhido que esse deslocamento se dá para o sentido dos menores valores da variável esperança de vida (Y_1) e no sentido dos maiores valores da variável renda *per capita* (Y_2).

Mais ainda, a melhor adequação do modelo SSN vista por meio da log-verossimilhança e dos critérios de informação pode sugerir que esse viés é aquele que permite a melhor distribuição da probabilidade entre os dados como vemos nos gráficos da Figura 31, embora seja difícil enxergar tal fato especialmente na comparação entre os modelos SSN e STN.

Apenas para se ter uma noção da variabilidade do processo de estimação da regressão, uma vez que o erro padrão do vetor de escala $\boldsymbol{\alpha}$ nos diferentes modelos não é comparável, vejamos a estimativa de variância do estimador de regressão. Esta é a mesma estimativa da variância do erro indicada na equação (3.22), a qual é dada pela seguinte matriz:

$$\widehat{\mathrm{Var}(\mathbf{Y}_i|\mathbf{X}_i)} = \widehat{\mathrm{Var}(\boldsymbol{\epsilon}_i)} = \begin{bmatrix} 1,7030 & 59,1519 \\ 59,1519 & 9685,8048 \end{bmatrix}.$$

Uma medida de variabilidade conjunta das variáveis Y_1 e Y_2 que poderia ser utilizada para visualizar o ganho com o ajuste do modelo SSN em relação ao modelo normal, por exemplo, é a soma dos desvios padrões das variáveis (expressos pela raiz quadrada de cada elemento da diagonal da mesma). No caso do modelo SSN, essa medida vale $99,7215$ enquanto no modelo normal vale $105,6269$.

Finalmente, discutiremos os coeficientes de regressão $\widehat{\boldsymbol{\beta}}$, interpretando seus valores. Porém, a fim de comparar as magnitudes desses coeficientes, devemos colocar as variáveis explicativas na mesma escala padronizada, retirando dos valores de cada uma delas sua média e dividindo o resultado pelo seu respectivo desvio padrão. Assim, todas as explicativas transformadas passam a ter média 0 e desvio padrão 1, de modo que podemos comparar os valores dos coeficientes da regressão para saber como cada aspecto demográfico explicativo impacta nas respostas.

Procedendo dessa forma, apresentamos na Tabela 19 as novas estimativas para os parâmetros da regressão no referido modelo seguidas dos seus respectivos erros padrão abaixo entre parênteses.

Tabela 19 – Coeficientes de regressão do modelo SSN com covariáveis padronizadas

$\widehat{\beta}_{01}$	$\widehat{\beta}_{11}$	$\widehat{\beta}_{21}$	$\widehat{\beta}_{31}$	$\widehat{\beta}_{02}$	$\widehat{\beta}_{12}$	$\widehat{\beta}_{22}$	$\widehat{\beta}_{32}$
74,481	0,115	-0,390	-0,867	406,675	27,06	-42,555	-72,998
(0,120)	(0,047)	(0,074)	(0,075)	(5,795)	(3,189)	(5,103)	(5,075)

Com base na análise dos resultados da Tabela 19, podemos depreender quatro conclusões básicas com relação ao modelo em questão:

1. O impacto negativo da razão de dependência (X_2) sobre a esperança de vida (Y_1) é cerca de 3 vezes maior do que o impacto positivo da taxa de atividade (X_1) sobre essa mesma resposta.

2. O impacto negativo da taxa de analfabetismo (X_3) sobre a esperança de vida (Y_1) é cerca de 7,5 vezes maior do que o impacto positivo da taxa de atividade (X_1) sobre essa mesma resposta.

3. O impacto negativo da razão de dependência (X_2) sobre a renda *per capita* (Y_2) é cerca de 1,5 vez maior do que o impacto positivo da taxa de atividade (X_1) sobre essa mesma resposta.

4. O impacto negativo da taxa de analfabetismo (X_3) sobre a renda *per capita* (Y_2) é cerca de 2,5 vezes maior do que o impacto positivo da taxa de atividade sobre essa mesma resposta.

4 DIAGNÓSTICO NOS MODELOS SSMN

Neste capítulo, discutiremos as principais técnicas de diagnóstico para detecção de *pontos influentes* presentes na literatura nos modelos de regressão que temos trabalhado com enfoque no conjunto de dados reais apresentado no final do Capítulo 3.

A adoção de modelos de regressão com erros que seguem distribuições misturas de escala assimétricas de normais tende a reduzir a quantidade de *outliers* observados em relação ao modelo normal. O objetivo deste capítulo é apresentar alguns métodos gráficos envolvendo o uso de medidas específicas para a detecção de *outliers* que podem impactar significativamente o ajuste do modelo: os chamados *pontos influentes*.

O primeiro trabalho que surgiu com esse tipo de abordagem foi o de Cook (1977), o qual introduziu a posteriormente denominada *distância de Cook*, uma medida do grau de impacto da deleção de cada observação ainda amplamente utilizada em modelo normais. Tal técnica levou à criação de outras medidas com o mesmo propósito para modelos mais gerais.

Em Zhu *et al.* (2001), a essência dessa técnica foi adaptada para modelos com dados incompletos cuja estimação paramétrica é baseada no algoritmo EM. Basicamente, apresentaremos essa técnica na Subseção 4.1, além de extensões e generalizações encontradas na literatura, aplicando-as no nosso conjunto de dados reais já apresentado.

Na subseção seguinte, apresentamos a denominada *análise de influência local* nos moldes definidos em Zhu & Lee (2001). Em linhas gerais, trata-se de mensurar como certas observações podem influenciar a estimação dos parâmetros e as variáveis envolvidas nos modelos através de perturbações.

Com isso, mostraremos de modo mais preciso que as referidas técnicas aplicadas às distribuições SSMN tendem a destacar menos pontos influentes do que no modelo normal, por exemplo. Como veremos, essa constatação se deve fortemente ao fato de aquelas distribuições atribuírem um peso menor aos *outliers*.

4.1 ANÁLISE DE INFLUÊNCIA GLOBAL

O objetivo desta seção é aplicar a metodologia da deleção de casos para análise de influência global nos modelos misturas de escala assimétricas de normais. Usaremos como base as definições e resultados desenvolvidos em Zhu *et al.* (2001), propondo uma generalização e mostrando os resultados efetivos da análise no conjunto de dados reais.

Para isso, necessitamos das derivadas de primeira e segunda ordens da última função Q oriunda do algoritmo EM de cada uma das distribuições trabalhadas. Supondo, em nossos modelos, obtida a estimativa de máxima verossimilhança $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\tau}})$ mediante os algoritmos propostos no Capítulo 3, usaremos a expressão $Q = Q_1 + Q_2$ para indicar a função dada por

$$Q\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big) = c + \sum_{i=1}^{n} Q_{1i}\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big) + \sum_{i=1}^{n} Q_{2i}\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big); \text{ onde } c \text{ não depende de } \boldsymbol{\theta}.$$

Para realizar a análise de influência, consideraremos medidas relacionadas à função $Q\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big) = E_{\widehat{\boldsymbol{\theta}}}\left(\ell_C(\boldsymbol{\theta}) \middle| \mathbf{Y} = \mathbf{y}\right)$ expressa na forma acima e suas derivadas primeira e segunda, que são dadas, respectivamente, por

$$\frac{\partial Q}{\partial \boldsymbol{\theta}}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{n} \frac{\partial Q_{1i}}{\partial \boldsymbol{\theta}}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\right) + \sum_{i=1}^{n} \frac{\partial Q_{2i}}{\partial \boldsymbol{\theta}}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\right) \text{ e } \frac{\partial^2 Q}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{n} \frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\right) + \sum_{i=1}^{n} \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\right).$$

Dessa forma, expressaremos as referidas derivadas de uma forma geral que englobe os três exemplos desenvolvidos neste texto no que se refere à função Q_1 e mostraremos os resultados específicos referentes à função Q_2 em cada caso.

Inicialmente, adotaremos a partir de agora ao longo de todo este capítulo as seguintes notações utilizadas anteriormente ou muito similares a outras já adotadas:

$$\Lambda = \ln|\boldsymbol{\Sigma}|,\ \mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta},\ d_i = \mathbf{e}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{e}_i,\ A_i = \boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i,\ \widehat{t_i} = E_{\widehat{\boldsymbol{\theta}}}\left(T_i \middle| \mathbf{Y}_i = \mathbf{y}_i\right);$$

$$\widehat{m_i} = \begin{cases} E_{\widehat{\boldsymbol{\theta}}}\left(U_i \middle| \mathbf{Y}_i = \mathbf{y}_i\right), \text{ para } \mathbf{Y}_i \sim \text{ STN}_p\left(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\right) \text{ ou SSN}_p\left(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\right) \\ E_{\widehat{\boldsymbol{\theta}}}\left(V_{2i} \middle| \mathbf{Y}_i = \mathbf{y}_i\right), \text{ para } \mathbf{Y}_i \sim \text{ SCN}_p\left(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma\right) \end{cases};$$

$$\widehat{v_{1i}} = E_{\widehat{\boldsymbol{\theta}}}\left(V_{1i} \middle| \mathbf{Y}_i = \mathbf{y}_i\right) \text{ para } \mathbf{Y}_i \sim \text{ SCN}_p\left(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma\right);$$

$$\widehat{lu_i} = E_{\widehat{\boldsymbol{\theta}}}\left(\ln U_i \middle| \mathbf{Y}_i = \mathbf{y}_i\right), \text{ para } \mathbf{Y}_i \sim \text{ STN}_p\left(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\right) \text{ ou SSN}_p\left(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\right).$$

Com as convenções anteriores e lembrando a expressão para Q_1 na etapa k do algoritmo EM dada na equação (3.25), obtemos os seguintes resultados válidos para todas as misturas de escala assimétricas de normais:

$$\begin{aligned} Q_{1i} &= \frac{1}{2}\Lambda - \frac{1}{2}\widehat{m_i}d_i + \widehat{t_i}A_i - \frac{1}{2}A_i^2;\ \frac{\partial Q_{1i}}{\partial \boldsymbol{\theta}} = -\frac{1}{2}\frac{\partial \Lambda}{\partial \boldsymbol{\theta}} - \frac{1}{2}\widehat{m_i}\frac{\partial d_i}{\partial \boldsymbol{\theta}} + (\widehat{t_i} - A_i)\frac{\partial A_i}{\partial \boldsymbol{\theta}}; \\ \frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} &= -\frac{1}{2}\frac{\partial^2 \Lambda}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} - \frac{1}{2}\widehat{m_i}\frac{\partial^2 d_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} + (\widehat{t_i} - A_i)\frac{\partial^2 A_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} - \frac{\partial A_i}{\partial \boldsymbol{\theta}}\frac{\partial A_i}{\partial \boldsymbol{\theta}^T}. \end{aligned} \tag{4.1}$$

Assim, separando os resultados por grupos de parâmetros e utilizando as derivadas auxiliares do Anexo B, deduzimos as expressões abaixo:

$$\frac{\partial Q_{1i}}{\partial \boldsymbol{\beta}} = -\frac{1}{2}\widehat{m_i}\frac{\partial d_i}{\partial \boldsymbol{\beta}} + (\widehat{t_i} - A_i)\frac{\partial A_i}{\partial \boldsymbol{\beta}};\ \frac{\partial Q_{1i}}{\partial \boldsymbol{\alpha}} = -\frac{1}{2}\frac{\partial \Lambda}{\partial \boldsymbol{\alpha}} - \frac{1}{2}\widehat{m_i}\frac{\partial d_i}{\partial \boldsymbol{\alpha}} + (\widehat{t_i} - A_i)\frac{\partial A_i}{\partial \boldsymbol{\alpha}};\ \frac{\partial Q_{1i}}{\partial \boldsymbol{\lambda}} = (\widehat{t_i} - A_i)\frac{\partial A_i}{\partial \boldsymbol{\lambda}};\ \frac{\partial Q_{1i}}{\partial \boldsymbol{\tau}} = \mathbf{0};$$

$$\frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = -\frac{1}{2}\widehat{m_i}\frac{\partial^2 d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} - \frac{\partial A_i}{\partial \boldsymbol{\beta}}\frac{\partial A_i}{\partial \boldsymbol{\beta}^T};\ \frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} = -\frac{1}{2}\widehat{m_i}\frac{\partial^2 d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} + (\widehat{t_i} - A_i)\frac{\partial^2 A_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} - \frac{\partial A_i}{\partial \boldsymbol{\beta}}\frac{\partial A_i}{\partial \boldsymbol{\alpha}^T};$$

$$\frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\lambda}^T} = (\widehat{t_i} - A_i)\frac{\partial^2 A_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\lambda}^T} - \frac{\partial A_i}{\partial \boldsymbol{\beta}}\frac{\partial A_i}{\partial \boldsymbol{\lambda}^T};\ \frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} = -\frac{1}{2}\frac{\partial^2 \Lambda}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} - \frac{1}{2}\widehat{m_i}\frac{\partial^2 d_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} + (\widehat{t_i} - A_i)\frac{\partial^2 A_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} - \frac{\partial A_i}{\partial \boldsymbol{\alpha}}\frac{\partial A_i}{\partial \boldsymbol{\alpha}^T};$$

$$\frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\lambda}^T} = (\widehat{t_i} - A_i)\frac{\partial^2 A_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\lambda}^T} - \frac{\partial A_i}{\partial \boldsymbol{\alpha}}\frac{\partial A_i}{\partial \boldsymbol{\lambda}^T};\ \frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}^T} = -\frac{\partial A_i}{\partial \boldsymbol{\lambda}}\frac{\partial A_i}{\partial \boldsymbol{\lambda}^T};\ \frac{\partial^2 Q_{1i}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\tau}^T} = \mathbf{0}.$$

Vejamos as derivadas de Q_2 em cada um dos exemplos da família SSMN.

1. T-Student normal assimétrica:

$$Q_{2i} = \frac{\nu}{2}\ln\left(\frac{\nu}{2}\right) - \ln\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu}{2}\left(\widehat{u_i} - \widehat{lu_i}\right) = \frac{\nu}{2}\ln\left(\frac{\nu}{2}\right) - \ln\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu}{2}\left[\frac{\widehat{\nu}+p}{\widehat{\nu}+\widehat{d_i}} + \ln\left(\frac{\widehat{\nu}+\widehat{d_i}}{2}\right) - \Psi\left(\frac{\widehat{\nu}+p}{2}\right)\right];$$

$$\frac{\partial Q_{2i}}{\partial \boldsymbol{\theta}} = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial Q_{2i}}{\partial \nu}\end{bmatrix}^T;\ \frac{\partial Q_{2i}}{\partial \nu} = \frac{1}{2}\left[\ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right) + 1 - \widehat{u_i} + \widehat{lu_i}\right];$$

$$\frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial^2 Q_{2i}}{\partial \nu^2}\end{bmatrix};\ \frac{\partial^2 Q_{2i}}{\partial \nu^2} = \frac{1}{4}\left[\frac{2}{\nu} - \Psi_1\left(\frac{\nu}{2}\right)\right].$$

2. Slash normal assimétrica:

$$Q_{2i} = \ln\nu + (\nu - 1)\widehat{lu_i} = \ln\nu + (\nu - 1)\frac{(\widehat{d_i}/2)^{\widehat{\nu}+p/2}}{\Gamma\left(\widehat{\nu}+p/2\right)}\frac{\int_0^1 u^{\widehat{\nu}+p/2-1}\ln(u)e^{u\widehat{d_i}/2}du}{P\left(1;\widehat{\nu}+p/2,\widehat{d_i}/2\right)};$$

$$\frac{\partial Q_{2i}}{\partial \boldsymbol{\theta}} = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial Q_{2i}}{\partial \nu}\end{bmatrix}^T;\ \frac{\partial Q_{2i}}{\partial \nu} = \frac{1}{\nu} + \widehat{lu_i};$$

$$\frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial^2 Q_{2i}}{\partial \nu^2}\end{bmatrix};\ \frac{\partial^2 Q_{2i}}{\partial \nu^2} = -\frac{1}{\nu^2}.$$

3. Normal contaminada assimétrica:

$$
\begin{aligned}
Q_{2i} &= \frac{1}{2}\widehat{v_{1i}}(p\ln\gamma - \gamma d_i + 2\ln\nu) + \widehat{v_{2i}}\ln(1-\nu) \\
&= \frac{1}{2}\frac{\widehat{\nu\gamma}^{p/2}e^{-\widehat{\gamma d_i}/2}}{\widehat{\nu\gamma}^{p/2}e^{-\widehat{\gamma d_i}/2} + (1-\widehat{\nu})e^{-\widehat{d_i}/2}}(p\ln\gamma - \gamma d_i + 2\ln\nu) + \frac{(1-\widehat{\nu})e^{-\widehat{\gamma d_i}/2}}{\widehat{\nu\gamma}^{p/2}e^{-\widehat{\gamma d_i}/2} + (1-\widehat{\nu})e^{-\widehat{d_i}/2}}\ln(1-\nu);
\end{aligned}
$$

$$
\frac{\partial Q_{2i}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial Q_{2i}}{\partial \boldsymbol{\beta}} & \frac{\partial Q_{2i}}{\partial \boldsymbol{\alpha}} & \mathbf{0} & \frac{\partial Q_{2i}}{\partial \nu} & \frac{\partial Q_{2i}}{\partial \gamma} \end{bmatrix}^T; \quad \begin{bmatrix} \frac{\partial Q_{2i}}{\partial \boldsymbol{\beta}} & \frac{\partial Q_{2i}}{\partial \boldsymbol{\alpha}} \end{bmatrix}^T = -\frac{1}{2}\widehat{v_{1i}}\gamma \begin{bmatrix} \frac{\partial d_i}{\partial \boldsymbol{\beta}} & \frac{\partial d_i}{\partial \boldsymbol{\alpha}} \end{bmatrix}^T;
$$

$$
\frac{\partial Q_{2i}}{\partial \nu} = \frac{\widehat{v_{1i}}}{\nu} - \frac{\widehat{v_{2i}}}{1-\nu}; \quad \frac{\partial Q_{2i}}{\partial \gamma} = \frac{1}{2\gamma}\widehat{v_{1i}}(p - \gamma d_i);
$$

$$
\frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \begin{bmatrix} \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\partial^2 Q_{2i}}{\partial \nu^2} & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 & \frac{\partial^2 Q_{2i}}{\partial \gamma^2} \end{bmatrix}; \quad \begin{bmatrix} \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} \\ \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 Q_{2i}}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} \end{bmatrix} = -\frac{1}{2}\widehat{v_{1i}}\gamma \begin{bmatrix} \frac{\partial^2 d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} \\ \frac{\partial^2 d_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 d_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} \end{bmatrix};
$$

$$
\frac{\partial^2 Q_{2i}}{\partial \nu^2} = -\left[\frac{\widehat{v_{1i}}}{\nu^2} + \frac{\widehat{v_{2i}}}{(1-\nu)^2}\right]; \quad \frac{\partial^2 Q_{2i}}{\partial \gamma^2} = -\frac{p\widehat{v_{1i}}}{2\gamma^2}.
$$

Algumas medidas de influência global alternativas usam a log-verossimilhança dos dados incompletos no lugar da função Q como pode ser visto em Pan, Fei & Foster (2013). Na mesma referência, também se considera a possibilidade de tomar o valor esperado $E\left(\frac{\partial^2 Q}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}})\right)$ no lugar da função $\frac{\partial^2 Q}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}})$, onde basicamente substitui-se o vetor de dados fixo $\mathbf{y}$ pelo vetor aleatório $\mathbf{Y}$ na referida função e calcula-se os valores esperados das funções de $\mathbf{Y}$ que surgem em decorrência de tal substituição.

As medidas de influência global consideradas aqui para os modelos misturas de escala assimétricas de normais serão embasadas na função Q conforme se vê em Zhu *et al.* (2001). Embora na mesma referência seja feita menção a medidas análogas envolvendo a própria função log-verossimilhança ℓ, vamos nos ater às primeiras por elas se adequarem aos nossos propósitos.

Essencialmente, a ideia por detrás de qualquer medida de influência global é mensurar o efeito da retirada ou deleção de observações. Essa análise geralmente é feita a partir da retirada das observações uma a uma, onde cada $\widehat{\boldsymbol{\theta}}_{[j]}$ é definida como a estimativa do novo modelo obtido após a exclusão da observação $j \in \{1,\ldots,n\}$.

Para evitar fazer uma nova estimação dos parâmetros para cada retirada de uma observação, em Zhu *et al.* (2001) considera-se a seguinte aproximação de $\widehat{\theta}_{[j]}$:

$$\widehat{\boldsymbol{\theta}}_{[j]} \approx \widehat{\boldsymbol{\theta}} - \left[\frac{\partial^2 Q}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}})\right]^{-1} \frac{\partial Q_{[j]}}{\partial\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}). \tag{4.2}$$

A equação (4.2) consiste basicamente na realização de um passo do método de Newton-Raphson com valor inicial $\widehat{\boldsymbol{\theta}}$ e onde $\dfrac{\partial Q_{[j]}}{\partial\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}) = \sum\limits_{\substack{i=1\\ i\neq j}}^{n} \dfrac{\partial Q_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}})$ e a matriz hessiana também correspondente à retirada da observação j é aproximada pela matriz $\dfrac{\partial^2 Q}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}})$. Embora mediante algumas hipóteses adicionais elencadas em Zhu *et al.* (2001) a aproximação seja considerada suficientemente boa, verificamos que isso não ocorre nos modelos em estudo neste texto de acordo com constatações em dados simulados.

Sem o uso da aproximação (4.2), definiremos como em Zhu *et al.* (2001) duas medidas de influência global relativas à remoção da j-ésima observação:

1. Distância Q:
$$QD_{[j]} = 2\Big[Q\big(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}\big) - Q\big(\widehat{\boldsymbol{\theta}}_{[j]};\widehat{\boldsymbol{\theta}}\big)\Big]; \tag{4.3}$$

2. Distância de Cook generalizada:
$$GD_{[j]} = \big(\widehat{\boldsymbol{\theta}}_{[j]} - \widehat{\boldsymbol{\theta}}\big)^T \left[\frac{\partial^2 Q}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\big(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}\big)\right] \big(\widehat{\boldsymbol{\theta}}_{[j]} - \widehat{\boldsymbol{\theta}}\big). \tag{4.4}$$

Cada valor $\widehat{\boldsymbol{\theta}}_{[j]}$ será, para todos os efeitos, obtido por nova estimação paramétrica via algoritmo EM considerando os $n-1$ indivíduos restantes, isto é, desconsiderando nas variáveis respostas e covariáveis o vetor $\mathbf{y}_j$ e a matriz $\mathbf{X}_j$, respectivamente, referentes ao indivíduo j. Com recursos de paralelização computacional, proceder a uma nova estimação em detrimento da aproximação citada anteriormente não é tão custoso como no passado. Além disso, tal procedimento se mostra necessário nos modelos misturas de escala assimétricas de normais a fim de tornar mais fidedigna a informação fornecida por ambas as medidas (4.3) e (4.4).

A propósito, podemos ainda constatar facilmente que ambas as medidas fornecem essencialmente a mesma informação. Com efeito, se considerarmos a

Fórmula de Taylor de 2ª ordem para a função Q no ponto $\widehat{\boldsymbol{\theta}}_{[j]}$ em torno de $\widehat{\boldsymbol{\theta}}$, obtemos a seguinte aproximação:

$$Q(\widehat{\boldsymbol{\theta}}_{[j]};\widehat{\boldsymbol{\theta}}) \approx Q(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}) + \frac{\partial Q}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}})^T(\widehat{\boldsymbol{\theta}}_{[j]} - \widehat{\boldsymbol{\theta}}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_{[j]} - \widehat{\boldsymbol{\theta}})^T \frac{\partial^2 Q}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{[j]} - \widehat{\boldsymbol{\theta}}). \quad (4.5)$$

Sendo nulo o gradiente da função Q em $\widehat{\boldsymbol{\theta}}$, a segunda parcela da aproximação (4.5) é igual a 0 e segue daí que $Q\big(\widehat{\boldsymbol{\theta}}_{[j]};\widehat{\boldsymbol{\theta}}\big) - Q\big(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}\big) \approx \frac{1}{2}\big(\widehat{\boldsymbol{\theta}}_{[j]} - \widehat{\boldsymbol{\theta}}\big)^T \frac{\partial^2 Q}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\big(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}\big)\big(\widehat{\boldsymbol{\theta}}_{[j]} - \widehat{\boldsymbol{\theta}}\big)$, donde resulta que $QD_{[j]} \approx GD_{[j]}$. A diferença entre essas duas medidas é, portanto, formada por termos de ordem 3 ou superior, de modo que será desprezada em nosso contexto.

Tendo em vista essa última aproximação, optaremos pelo uso da distância de Cook generalizada pelo fato de tal medida possibilitar o cálculo da influência global relativa a um dado grupo de parâmetros. Para evidenciar o que separa nossos modelos da normalidade, vamos decompor o vetor de parâmetros em $\boldsymbol{\theta} = \big(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\big)$ com $\boldsymbol{\theta}_1 = \big(\boldsymbol{\beta}, \boldsymbol{\alpha}\big)$ e $\boldsymbol{\theta}_2 = \big(\boldsymbol{\lambda}, \boldsymbol{\tau}\big)$. Dessa forma, consideraremos as seguintes distâncias de Cook generalizadas parciais:

- Distância de Cook generalizada parcial relativa a locação-escala:

$$GD_{1[j]} = \big(\widehat{\boldsymbol{\theta}}_{1[j]} - \widehat{\boldsymbol{\theta}}_1\big)^T \left[\frac{\partial^2 Q}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T}\big(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}\big)\right]\big(\widehat{\boldsymbol{\theta}}_{1[j]} - \widehat{\boldsymbol{\theta}}_1\big), \text{ onde } \frac{\partial^2 Q}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T} = \begin{bmatrix} \frac{\partial^2 Q}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 Q}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} \\ \frac{\partial^2 Q}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 Q}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} \end{bmatrix}. \quad (4.6)$$

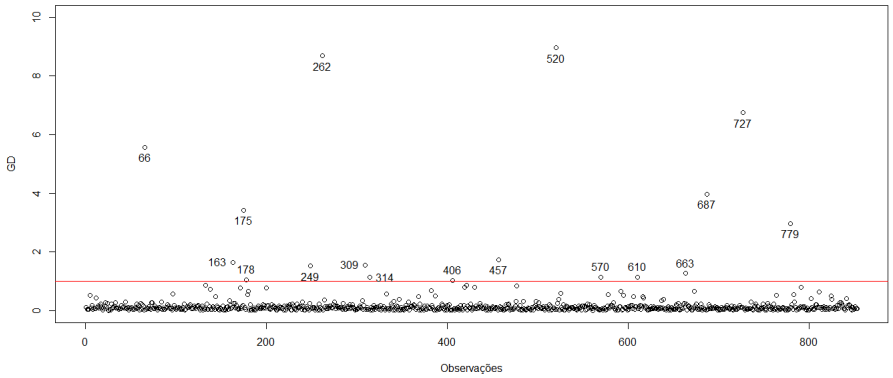
- Distância de Cook generalizada parcial relativa aos demais parâmetros:

$$GD_{2[j]} = \big(\widehat{\boldsymbol{\theta}}_{2[j]} - \widehat{\boldsymbol{\theta}}_2\big)^T \left[\frac{\partial^2 Q}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T}\big(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}\big)\right]\big(\widehat{\boldsymbol{\theta}}_{2[j]} - \widehat{\boldsymbol{\theta}}_2\big), \text{ onde } \frac{\partial^2 Q}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} = \begin{bmatrix} \frac{\partial^2 Q}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}^T} & \frac{\partial^2 Q}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\tau}^T} \\ \frac{\partial^2 Q}{\partial \boldsymbol{\tau}\partial \boldsymbol{\lambda}^T} & \frac{\partial^2 Q}{\partial \boldsymbol{\tau}\partial \boldsymbol{\tau}^T} \end{bmatrix}. \quad (4.7)$$

Realizaremos agora a aplicação das medidas (4.4), (4.6) e (4.7) para comparar a detecção de observações influentes nos modelos normal, normal assimétrico e Slash normal assimétrico ajustados ao conjunto de dados dos municípios mineiros no Capítulo 3. Analisaremos apenas esses modelos porque o último foi selecionado como o mais adequado aos dados e os outros dois são aqueles que permitem apontar uma gradação mais nítida na identificação de pontos influentes.

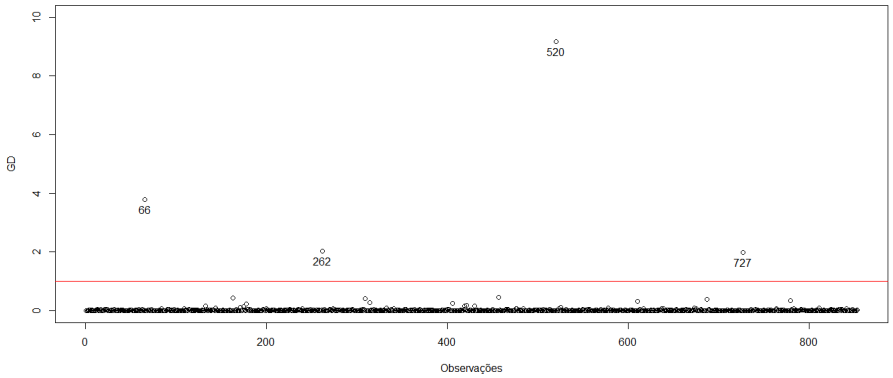
No modelo normal, temos $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, de modo que apresentamos apenas o gráfico de influência com a medida (4.4) na Figura 32.

Figura 32 – Distância de Cook generalizada no modelo normal



A linha horizontal, que corresponde à distância de Cook igual a 1, é o primeiro "ponto de corte" estabelecido na concepção original da medida em Cook & Weisberg (1982) para se considerar um ponto como influente. Na Figura 32, numeramos os 17 municípios (dispostos em ordem alfabética) que aparecem acima da linha de corte. Mantendo 1 como uma referência embora seja controverso na literatura, vemos na Figura 33 o gráfico da medida (4.4) para o modelo normal assimétrico, onde $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$.

Figura 33 – Distância de Cook generalizada no modelo normal assimétrico

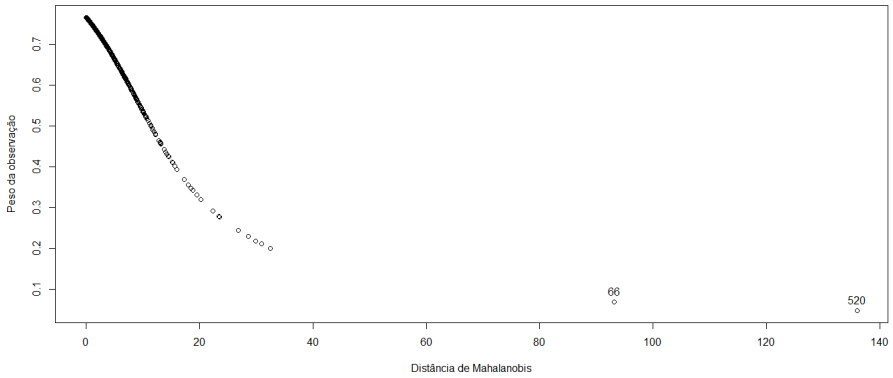


Vamos destacar os 4 municípios apontados como influentes nos dois modelos: 66 (Nova Lima), 262 (Doresópolis), 520 (Belo Horizonte) e 727 (São João da Lagoa).

Nova Lima e BH são pontos extremos por serem de longe os municípios com as duas maiores rendas *per capita* do estado e decerto possuem grande influência no ajuste dos modelos em questão. Quanto aos outros dois, embora Doresópolis esteja entre as 10% maiores esperanças de vida (57ª) e São João da Lagoa figura entre as 10% menores rendas *per capita* (776ª), são apenas medianos com relação à outra variável e não poderiam ser considerados influentes no sentido de alterar tão significativamente o ajuste. Tais fatos mostram que a identificação de pontos influentes nos modelos normal e normal assimétrico não é tão eficaz, embora no segundo seja um pouco melhor.

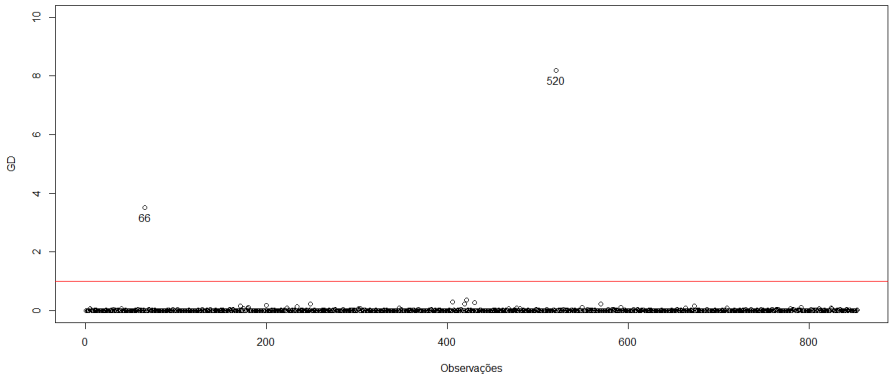
Antes de apresentarmos as distâncias de Cook generalizadas para o modelo Slash normal assimétrico, enfatizaremos uma das principais justificativas para a vantagem dos modelos SSMN, dentre os quais o SSN, na identificação de pontos influentes que efetivamente são determinantes no ajuste do modelo. Essa vantagem está na ponderação que tais distribuições fazem em relação aos *outliers*. O gráfico da Figura 34 mostra que as observações com maior distância de Mahalanobis $d_i(\widehat{\boldsymbol{\theta}})$ – indicador de *outliers*, que podem vir a ser influentes – apresentam menor peso na estimação dos parâmetros, indicado pelas estimativas do fator de escala $\widehat{u_i}$.

Figura 34 – Distâncias de Mahalanobis vs pesos no modelo SSN



Diante disso, vejamos como fica a identificação dos pontos influentes com a mesma referência 1 no modelo Slash normal assimétrico na Figura 35.

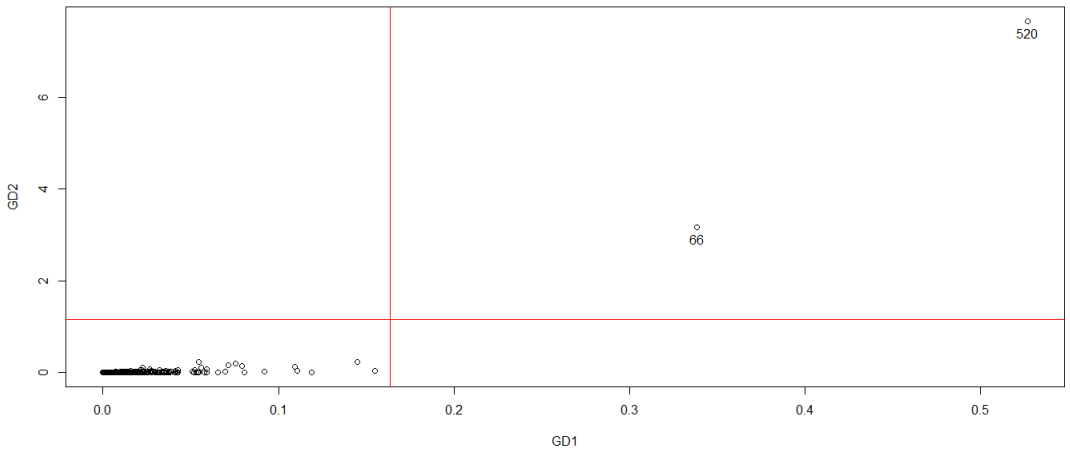
Figura 35 – Distância de Cook generalizada no modelo Slash normal assimétrico



A complementaridade dos gráficos das Figuras 34 e 35 fica nítida, quando vemos que se destacam apenas os municípios 66 (Belo Horizonte) e 520 (Nova Lima). Como já mencionado, ambos se destacam pela sua renda *per capita*, sendo que Belo Horizonte possui a segunda maior: R$1497,29 enquanto Nova Lima possui a maior: R$1731,84. O caso de BH é esperado por se tratar da capital do estado. Já o de Nova Lima se justifica pelo fato de este município fazer fronteira com a região mais rica de BH, o que leva muitos de seus moradores a migrarem para Nova Lima por ser um município com menos incidência de criminalidade, por exemplo. Essa migração em busca de melhor qualidade de vida fica evidente ao comparar a esperança de vida dos dois municípios: Nova Lima (78,10 anos – a 2ª maior de MG) e BH (76,37 anos – apenas a 127ª do estado).

Esses dois municípios são sem dúvida notadamente influentes no ajuste deste modelo. Para avaliar em qual grupo de parâmetros (locação-escala ou os demais – assimetria e hiper-parâmetro) a influência é mais representativa, vamos recorrer às distâncias de Cook generalizadas parciais definidas em (4.6) e (4.7). No modelo Slash normal assimétrico (SSN), temos $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ e $\boldsymbol{\theta}_2 = (\boldsymbol{\lambda}, \nu)$, cujas respectivas medidas de Cook associadas GD_1 e GD_2 estão dispostas na Figura 36.

Figura 36 – Distâncias de Cook generalizadas parciais no modelo SSN



Utilizando como referência para ambas as distâncias de Cook generalizadas parciais a média de cada medida individual somada a 4 desvios padrões respectivos, detectamos os mesmos pontos influentes em relação aos dois grupos de parâmetros: Nova Lima (520) e Belo Horizonte (66).

Apesar de a soma das distâncias parciais não ser necessariamente igual à distância total, é fácil ver que a interpretação extraída do gráfico da Figura 36 é essencialmente a mesma do que vemos na Figura 35.

Para finalizar essa discussão sobre a influência global, cabe enfatizar que o fato de o município de BH ser um pouco menos influente do que Nova Lima está de acordo com as constatações anteriores: enquanto este município possui valores extremos nas duas variáveis respostas, aquele só tem valor extremo de renda *per capita* apresentando um valor apenas relativamente alto na esperança de vida. Isso explica também o peso menor que é dado ao município de Nova Lima em relação a Belo Horizonte conforme vemos na Figura 34.

4.2 ANÁLISE DE INFLUÊNCIA LOCAL

A ideia da chamada *análise de influência local* consiste basicamente em perturbar o modelo proposto com o intuito de mensurar o efeito de cada observação na curvatura do gráfico de uma função dessa perturbação.

De modo mais preciso, Cook (1986) propôs considerar um vetor de perturbação $\boldsymbol{\omega} \in \Omega$, onde Ω é um conjunto aberto no espaço euclidiano $\mathbb{R}^n$ (sendo n a quantidade de indivíduos observados), assumindo ainda a existência de $\boldsymbol{\omega}_0 \in \Omega$ e uma função log-verossimilhança perturbada ℓ_0 de classe $\mathscr{C}^2$ tais que $\ell_0(\boldsymbol{\theta}, \boldsymbol{\omega}_0) = \ell(\boldsymbol{\theta})\ \forall \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r$. Segundo o referido autor, a influência local seria medida a partir da análise da curvatura normal do *gráfico de influência* da função $f_\ell \colon \Omega \to \mathbb{R}$ dada por

$$f_\ell(\boldsymbol{\omega}) = 2\left[\ell\Big(\widehat{\boldsymbol{\theta}}\Big) - \ell\Big(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})\Big)\right]. \tag{4.8}$$

Na expressão (4.8), $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}) \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ \ell_0\Big(\boldsymbol{\theta}, \boldsymbol{\omega}\Big)$, o qual suposto unitário define uma função $\boldsymbol{\omega} \mapsto \widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ de classe $\mathscr{C}^2$ com base em Boldrin & Montrucchio (1987). Note que a hipótese inicial sobre ℓ_0 implica $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}_0) = \widehat{\boldsymbol{\theta}}$ e a construção feita assegura que f_ℓ é não negativa em Ω.

Generalizando esse raciocínio, vemos em Zhu & Lee (2001) o emprego da mesma técnica com a hipótese adicional de que uma função log-verossimilhança completa perturbada ℓ_{C0} também de classe $\mathscr{C}^2$ satisfaz a igualdade $\ell_{C0}(\boldsymbol{\theta}, \boldsymbol{\omega}_0) = \ell_C(\boldsymbol{\theta})\ \forall \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r$. Com isso, define-se a função $f_Q \colon \Omega \to \mathbb{R}$ por

$$f_Q(\boldsymbol{\omega}) = 2\left[Q\Big(\widehat{\boldsymbol{\theta}}; \widehat{\boldsymbol{\theta}}\Big) - Q\Big(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}); \widehat{\boldsymbol{\theta}}\Big)\right]. \tag{4.9}$$

Na expressão (4.9), temos que $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}) \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ Q_0\Big(\boldsymbol{\theta}, \boldsymbol{\omega}; \widehat{\boldsymbol{\theta}}\Big)$ com $Q_0\Big(\boldsymbol{\theta}, \boldsymbol{\omega}; \widehat{\boldsymbol{\theta}}\Big) = E_{\widehat{\boldsymbol{\theta}}}\Big(\ell_{C0}(\boldsymbol{\theta}, \boldsymbol{\omega}) | \mathbf{Y} = \mathbf{y}\Big)$. O fato de a função $\boldsymbol{\omega} \mapsto \widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ definida agora ser duas vezes continuamente diferenciável é concluído de modo análogo ao caso anterior.

Como $\widehat{\boldsymbol{\theta}}$ é o maximizador de Q em Θ, então $f_Q(\boldsymbol{\omega}) \geq 0\ \forall \boldsymbol{\omega} \in \Omega$. Por outro lado, vem das suposições feitas que $\boldsymbol{\omega}_0$ é minimizador de f_Q em Ω, o que nos fornece os seguintes resultados úteis:

$$f_Q(\boldsymbol{\omega}_0) = 0, \quad \frac{\partial f_Q}{\partial \boldsymbol{\omega}}(\boldsymbol{\omega}_0) = \mathbf{0} \quad \text{e} \quad \frac{\partial^2 f_Q}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^T}(\boldsymbol{\omega}_0) \text{ positiva semidefinida.} \tag{4.10}$$

Com base nos conceitos discutidos, vejamos definições e resultados relevantes para a implementação da técnica proposta por Zhu & Lee (2001).

Definição 4.2.1. Seja f_Q a função definida em (4.9). Então, seu gráfico expresso pela função $\rho(\boldsymbol{\omega}) = (\boldsymbol{\omega}, f_Q(\boldsymbol{\omega}))$ é dito o *gráfico de influência do modelo*.

Antes de prosseguirmos, vamos convencionar três notações simplificadoras:

$$\mathscr{H}Q_{\boldsymbol{\theta}} = \frac{\partial^2 Q}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}), \quad \mathscr{H}Q_{\boldsymbol{\omega}} = \frac{\partial^2 Q_0}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}),\boldsymbol{\omega};\widehat{\boldsymbol{\theta}}) \text{ e } \Delta_{\boldsymbol{\omega}} = \frac{\partial^2 Q_0}{\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^T}(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}),\boldsymbol{\omega};\widehat{\boldsymbol{\theta}}). \quad (4.11)$$

A partir de uma generalização natural da noção de *curvatura normal* de uma superfície para dimensões quaisquer, provaremos o resultado a seguir citado em Zhu & Lee (2001):

Teorema 4.2.1. Com as mesmas notações anteriores, a curvatura normal da superfície $M = \rho(\Omega)$ de $\mathbb{R}^{n+1}$ no ponto $P = \rho(\boldsymbol{\omega}_0) = (\boldsymbol{\omega}_0, 0)$ na direção de um vetor unitário $\mathbf{v} \in \mathbb{R}^n$ é dada pela expressão seguinte:

$$C_{\rho,\mathbf{v}} = -2\mathbf{v}^T\mathscr{H}Q_{\boldsymbol{\omega}_0}\mathbf{v} = -2\mathbf{v}^T\Delta_{\boldsymbol{\omega}_0}^T\mathscr{H}Q_{\widehat{\boldsymbol{\theta}}}^{-1}\Delta_{\boldsymbol{\omega}_0}\mathbf{v}. \quad (4.12)$$

Demonstração:

Para provar este resultado, necessitaremos da linguagem da Geometria Diferencial e da Análise em Várias Variáveis, que adaptaremos de O'Neill (2006) e Lima (1999). No nosso caso, queremos calcular a curvatura normal, definida em O'Neill (2006) para superfícies bidimensionais em $\mathbb{R}^3$, para a superfície n-dimensional $M = \rho(\Omega)$ em $\mathbb{R}^{n+1}$, a qual seria tratada em Lima (1999) como uma hiperfície (imersa em $\mathbb{R}^{n+1}$).

Inicialmente, note que as duas primeiras condições dadas em (4.10) garantem que o *vetor normal unitário* a M no ponto P é dado por

$$U(P) = \frac{1}{\sqrt{1 + \frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0)^T\frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0)}}\begin{bmatrix} -\dfrac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0) \\ 1 \end{bmatrix} = \mathbf{c}_{n+1} \in \mathbb{R}^{n+1}. \quad (4.13)$$

Com isso, o *hiperplano tangente* $T_P(M)$ à superfície M em P é igual a $\mathbb{R}^n \times \{0\}$, de modo que o identificaremos com $\mathbb{R}^n$. Dessa forma, todo vetor $\mathbf{v}$ da hipótese pode ser considerado um *vetor tangente*.

Conforme a definição de O'Neill (2006) em termos do *operador forma* $S_P\colon T_P(M) \to T_P(M)$ para a curvatura normal da superfície M em P na direção de um vetor tangente unitário $\mathbf{v}$, teremos no nosso caso que

$$C_{\rho,\mathbf{v}} = \Big([S_P]\mathbf{v}\Big)^T\mathbf{v} = \mathbf{v}^T[S_P]\mathbf{v}.$$

Acima, representamos por $[S_P]$ a matriz na base canônica do referido operador forma, o qual é definido por $S_P(\mathbf{v}) = -\dfrac{d}{dt}U(P+t\mathbf{v})\Big|_{t=0}$.

Observando que $P + t\mathbf{v} = (\boldsymbol{\omega}_0 + t\mathbf{v}, 0)$, o uso de expressão análoga a (4.13) em tal ponto e a aplicação das regras de Cálculo Matricial nos dão o seguinte:

$$-\frac{d}{dt}U(P+t\mathbf{v}) = -\frac{d}{dt}\left(\frac{1}{\sqrt{1+\frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})^T\frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})}}\begin{bmatrix}-\dfrac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})\\ 1\end{bmatrix}\right) =$$

$$\frac{\mathbf{v}^T\frac{\partial^2 f_Q}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}(\boldsymbol{\omega}_0+t\mathbf{v})\frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v}) + \frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})^T\frac{\partial^2 f_Q}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}(\boldsymbol{\omega}_0+t\mathbf{v})\mathbf{v}}{2\left[1+\frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})^T\frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})\right]^{3/2}}\begin{bmatrix}-\dfrac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})\\ 1\end{bmatrix} +$$

$$\frac{1}{\sqrt{1+\frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})^T\frac{\partial f_Q}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0+t\mathbf{v})}}\begin{bmatrix}\dfrac{\partial^2 f_Q}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}(\boldsymbol{\omega}_0+t\mathbf{v}) & \mathbf{0}\\ \mathbf{0}^T & 0\end{bmatrix}\mathbf{v}.$$

Fazendo $t = 0$ na igualdade acima e identificando o espaço $\mathbb{R}^n \times \{0\}$ com $\mathbb{R}^n$, concluímos que $[S_P] = \dfrac{\partial^2 f_Q}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}(\boldsymbol{\omega}_0) = -2\mathscr{H}Q_{\boldsymbol{\omega}_0}$ por (4.9) e (4.11). Assim, a fórmula mencionada anteriormente para a curvatura normal coincide com a primeira expressão dada em (4.12). Por fim, usando sucessivamente a versão da regra da cadeia dada em (A.2), vemos que

(i) $\mathscr{H}Q_{\boldsymbol{\omega}_0} = \dfrac{\partial^2 Q_0}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}\big(\widehat{\boldsymbol{\theta}},\boldsymbol{\omega}_0;\widehat{\boldsymbol{\theta}}\big) = \dfrac{\partial^2(Q\circ\boldsymbol{\theta})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}(\boldsymbol{\omega}_0) = \dfrac{\partial\widehat{\boldsymbol{\theta}}}{\partial\boldsymbol{\omega}^T}(\boldsymbol{\omega}_0)\dfrac{\partial^2 Q}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\big(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}\big)\dfrac{\partial\widehat{\boldsymbol{\theta}}}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0);$

(ii) $Q_0|\boldsymbol{\omega}=\boldsymbol{\omega}_0 = Q \underset{\frac{\partial}{\partial\boldsymbol{\theta}}}{\Rightarrow} \dfrac{\partial Q_0}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \dfrac{\partial Q}{\partial\boldsymbol{\theta}} \underset{\frac{\partial}{\partial\boldsymbol{\omega}^T}}{\Rightarrow} \dfrac{\partial^2 Q}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\big(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\theta}}\big)\dfrac{\partial\widehat{\boldsymbol{\theta}}}{\partial\boldsymbol{\omega}}(\boldsymbol{\omega}_0) + \dfrac{\partial^2 Q_0}{\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^T}\big(\widehat{\boldsymbol{\theta}},\boldsymbol{\omega}_0;\widehat{\boldsymbol{\theta}}\big) = \mathbf{0}.$

Por (ii), $\dfrac{\partial\widehat{\boldsymbol{\theta}}}{\partial\boldsymbol{\omega}} = -\left[\dfrac{\partial^2 Q}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1}\dfrac{\partial^2 Q_0}{\partial\boldsymbol{\omega}\partial\boldsymbol{\theta}^T}$ em $\boldsymbol{\omega}=\boldsymbol{\omega}_0$, e substituindo em (i) vem $\dfrac{\partial^2 Q_0}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T} = \dfrac{\partial^2 Q_0}{\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^T}\left[\dfrac{\partial^2 Q}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1}\dfrac{\partial^2 Q_0}{\partial\boldsymbol{\omega}\partial\boldsymbol{\theta}^T}$, isto é, $\mathscr{H}Q_{\boldsymbol{\omega}_0} = \Delta_{\boldsymbol{\omega}_0}^T\mathscr{H}Q_{\widehat{\boldsymbol{\theta}}}^{-1}\Delta_{\boldsymbol{\omega}_0}$. Dessa forma, concluímos a prova da expressão (4.12).

■

Em virtude da dificuldade de se determinar valores de referência para detectar influência via curvatura normal, foi proposta com base nos resultados de Poon & Poon (1999) a definição indicada adiante:

Definição 4.2.2. Dada a superfície $M = \rho(\Omega)$ do gráfico de influência, sua *curvatura normal conformalizada* é dada pela expressão

$$B_{\rho,\mathbf{v}} = \frac{C_{\rho,\mathbf{v}}}{\mathrm{tr}\Big(-2\mathscr{H}Q_{\boldsymbol{\omega}_0}\Big)}. \tag{4.14}$$

A proposição seguinte apresenta algumas importantes propriedades da nova curvatura definida em (4.14).

Proposição 4.2.1. A curvatura normal conformalizada $B_{\rho,\mathbf{v}}$ é um-a-um com $C_{\rho,\mathbf{v}}$, pertence ao intervalo (0,1] e é invariante por reparametrizações quaisquer de $\boldsymbol{\theta}$ e reparametrizações conformes de $\boldsymbol{\omega}$.

Demonstração:

Verificar que as curvaturas $B_{\rho,\mathbf{v}}$ e $C_{\rho,\mathbf{v}}$ são um-a-um é muito simples. Como $-2\mathscr{H}Q_{\boldsymbol{\omega}_0} = -2\dfrac{\partial^2 Q_0}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}\Big(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}_0),\boldsymbol{\omega}_0;\widehat{\boldsymbol{\theta}}\Big) = \dfrac{\partial^2 f_Q}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}(\boldsymbol{\omega}_0)$ é positiva semidefinida, então $\mathrm{tr}\Big(-2\mathscr{H}Q_{\boldsymbol{\omega}_0}\Big)$ é uma constante $k>0$ visto que estamos considerando $\boldsymbol{\omega}_0\in\Omega$ fixo e descartando a irrelevante possibilidade de termos a matriz nula. Assim, $C_{\rho,\mathbf{v}} = kB_{\rho,\mathbf{v}}$ define uma homotetia bijetora entre as duas curvaturas.

Vejamos agora que a curvatura normal conformalizada satisfaz $0 < B_{\rho,\mathbf{v}} \leq 1$. Com efeito, a positividade é óbvia pois decorre do fato de $C_{\rho,\mathbf{v}}$ ser positiva de acordo com o Proposição 4.2.1. Para provar a limitação superior, consideraremos a decomposição espectral da matriz $-2\mathscr{H}Q_{\boldsymbol{\omega}_0}$ cujos autovalores denotaremos

ordenadamente por $\eta_1 \geq \ldots \geq \eta_r \geq \eta_{r+1} = \ldots = \eta_n = 0$ e seus respectivos autovetores (ortonormais) básicos associados por $\mathbf{w}_1, \ldots, \mathbf{w}_n$. Note que a relação $\mathscr{H}Q_{\boldsymbol{\omega}_0} = \Delta_{\boldsymbol{\omega}_0}^T \mathscr{H}Q_{\widehat{\boldsymbol{\theta}}}^{-1} \Delta_{\boldsymbol{\omega}_0}$ deduzida anteriormente implica a desigualdade $\text{posto}\Big(\mathscr{H}Q_{\boldsymbol{\omega}_0}\Big) \leq \text{posto}\Big(\mathscr{H}Q_{\boldsymbol{\theta}}\Big) = r$, donde vem que $-2\mathscr{H}Q_{\boldsymbol{\omega}_0}$ possui no máximo r autovalores positivos e pode ser escrita na forma a seguir:

$$-2\mathscr{H}Q_{\boldsymbol{\omega}_0} = \sum_{i=1}^{r} \eta_i \mathbf{w}_i \mathbf{w}_i^T. \tag{4.15}$$

Desse modo, vendo que $\text{tr}\Big(-2\mathscr{H}Q_{\boldsymbol{\omega}_0}\Big) = \sum\limits_{i=1}^{r} \eta_i$ e considerando $\widetilde{\eta}_i = \dfrac{\eta_i}{\sum_{i=1}^{r} \eta_i}$ para cada $i \in \{1, \ldots, r\}$, a expressão da curvatura normal conformalizada na direção de cada autovetor $\mathbf{w}_j$ ($j \in \{1, \ldots, n\}$) fica

$$B_{\rho, \mathbf{w}_j} = \sum_{i=1}^{n} \widetilde{\eta}_i (\mathbf{w}_j^T \mathbf{w}_i)^2 = \sum_{i=1}^{r} \widetilde{\eta}_i (\mathbf{w}_j^T \mathbf{w}_i)^2 = \widetilde{\eta}_j \leq 1.$$

Todo vetor unitário $\mathbf{v} \in \mathbb{R}^n$ cumpre $\mathbf{v} = \sum\limits_{j=1}^{n} \alpha_j \mathbf{w}_j$ com $\sum\limits_{j=1}^{n} \alpha_j^2 = 1$ e daí

$$B_{\rho, \mathbf{v}} = \sum_{j=1}^{n} \alpha_j^2 B_{\rho, \mathbf{w}_j} \leq \sum_{j=1}^{n} \alpha_j^2 = 1.$$

Finalmente, para mostrar $B_{\rho,\mathbf{v}}$ é invariante por reparametrizações adequadas de $\boldsymbol{\theta}$ e $\boldsymbol{\omega}$, é preciso recorrer novamente à Análise e à Geometria. De fato, se $\boldsymbol{\zeta} = \psi(\boldsymbol{\theta})$ é uma reparametrização qualquer de $\boldsymbol{\theta}$, o fato de ψ ser um difeomorfismo nos dá que $\widehat{\boldsymbol{\zeta}} = \psi(\widehat{\boldsymbol{\theta}})$ e permite ainda o uso sucessivo da regra da cadeia dada em (A.2) para obter a relação adiante com os termos sempre avaliados em $\boldsymbol{\omega} = \boldsymbol{\omega}_0$:

$$\frac{\partial^2 (Q \circ \boldsymbol{\zeta})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^T} = \frac{\partial \widehat{\boldsymbol{\zeta}}}{\partial \boldsymbol{\omega}^T} \frac{\partial^2 Q}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T}(\widehat{\boldsymbol{\zeta}}; \widehat{\boldsymbol{\zeta}}) \frac{\partial \widehat{\boldsymbol{\zeta}}}{\partial \boldsymbol{\omega}} = \frac{\partial \widehat{\boldsymbol{\theta}}}{\partial \boldsymbol{\omega}^T} \frac{\partial \psi}{\partial \boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}}) \frac{\partial \psi}{\partial \boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}})^{-1} \frac{\partial^2 Q}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\widehat{\boldsymbol{\theta}}; \widehat{\boldsymbol{\theta}}) \frac{\partial \psi}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})^{-1} \frac{\partial \psi}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) \frac{\partial \widehat{\boldsymbol{\theta}}}{\partial \boldsymbol{\omega}}.$$

Assim, tanto a curvatura normal quanto sua versão conformalizada são preservadas.

Já uma reparametrização conforme de $\boldsymbol{\varsigma} = \varphi(\boldsymbol{\omega})$ de $\boldsymbol{\omega}$ induz o *mapa conforme* $\Big(\boldsymbol{\omega}, f_Q(\boldsymbol{\omega})\Big) \overset{F}{\mapsto} \Big(\varphi(\boldsymbol{\omega}), f_Q(\varphi(\boldsymbol{\omega}))\Big)$ cuja aplicação tangente em $P = (\boldsymbol{\omega}_0, 0)$ é dada por $F_P^*(\mathbf{v}) = \dfrac{\partial \varphi}{\partial \boldsymbol{\omega}}(\boldsymbol{\omega}_0)\mathbf{v}$ satisfazendo $\dfrac{\partial \varphi}{\partial \boldsymbol{\omega}}(\boldsymbol{\omega}_0)^T \dfrac{\partial \varphi}{\partial \boldsymbol{\omega}}(\boldsymbol{\omega}_0) = a\mathbf{I}_n,\ a > 0$. Com o uso da última propriedade, concluímos que $B_{\rho\circ\varphi,\mathbf{v}} = B_{\rho,\mathbf{v}}$ e $C_{\rho\circ\varphi,\mathbf{v}} = aC_{\rho,\mathbf{v}}$.

■

Veremos enfim medidas de influência local propriamente ditas conforme a proposta de Zhu & Lee (2001). Antes disso, observe que usando novamente a decomposição (4.15), obtemos as seguintes expressões para as curvaturas normal e conformalizada na direção dos vetores canônicos de $\mathbb{R}^n$:

$$C_{\rho,\mathbf{c}_j} = \sum_{i=1}^{r} \eta_i w_{ij}^2 \quad \text{e} \quad B_{\rho,\mathbf{c}_j} = \sum_{i=1}^{r} \widetilde{\eta}_i w_{ij}^2. \tag{4.16}$$

Isso posto, definiremos o que vem a ser um autovetor influente ao nível $m_0 \in [0, r]$ (m_0-influente) e a soma ponderada dos autovetores m_0-influentes.

Definição 4.2.3. Dizemos que um autovetor $\mathbf{w}_i$ da matriz de influência $-2\mathscr{H}Q_{\boldsymbol{\omega}_0}$ é m_0-influente quando $B_{\rho,\mathbf{w}_i} \geq \dfrac{m_0}{r}$. Já a soma ponderada dos autovetores m_0-influentes, denotada por $M(m_0)$, possui a coordenada correspondente à observação $j \in \{1, \ldots, n\}$ dada por

$$M(m_0)_j = \sum_{i \in I_{m_0}} \widetilde{\eta}_i w_{ij}^2, \text{ onde } I_{m_0} = \left\{ i : \widetilde{\eta}_i \geq \frac{m_0}{r} \right\}.$$

Vejamos agora duas propriedades relativas à Definição 4.2.3.

Proposição 4.2.2. A soma ponderada $M(0)$ satisfaz $M(0)_j = B_{\rho,\mathbf{c}_j}$ $\forall j \in \{1, \ldots, n\}$ e possui média $\overline{M(0)} = \dfrac{1}{n}$.

Demonstração:

De acordo com a Definição 4.2.3, temos que $I_0 = \{i : \widetilde{\eta}_i \geq 0\} = \{1, \ldots, n\}$ e daí pela expressão dada em (4.16) vale a relação

$$M(0)_j = \sum_{i \in I_0} \widetilde{\eta}_i w_{ij}^2 = \sum_{i=1}^{n} \widetilde{\eta}_i w_{ij}^2 = \sum_{i=1}^{r} \widetilde{\eta}_i w_{ij}^2 = B_{\rho,\mathbf{c}_j} \ \forall j \in \{1, \ldots, n\}.$$

Além disso, como os autovetores são unitários, obtemos ainda

$$\overline{M(0)} = \frac{1}{n} \sum_{j=1}^{n} B_{\rho,\mathbf{c}_j} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{r} \widetilde{\eta}_i w_{ij}^2 = \frac{1}{n} \sum_{i=1}^{r} \widetilde{\eta}_i \sum_{j=1}^{n} w_{ij}^2 = \frac{1}{n} \sum_{i=1}^{r} \widetilde{\eta}_i = \frac{1}{n} \frac{\sum_{i=1}^{r} \eta_i}{\sum_{i=1}^{r} \eta_i} = \frac{1}{n}.$$

■

Tendo em mente todos os resultados anteriores, podemos apresentar o critério mais adotado na literatura para a detecção de observações localmente influentes. Trata-se da proposta de Lee & Xu (2004), que considera influente uma observação j satisfazendo a condição seguinte:

$$M(0)_j > \overline{M(0)} + c^* S_{M(0)}, \text{ onde } S_{M(0)} \text{ é o desvio padrão do vetor } M(0). \tag{4.17}$$

Na relação (4.17), a constante c^* é controversa; sendo em Zhu & Lee (2001), $c^* = 2$; em Russo, Paula & Aoki (2009), $c^* = 3$ e em Zeller *et al.* (2010), $c^* = 4$.

Vamos descrever agora em linhas gerais os resultados necessários à análise da influência local nas distribuições SSMN. A informação essencial para o cálculo das medidas de influência local associadas à curvatura normal conformalizada está contida na matriz $-2\mathscr{H}Q_{\boldsymbol{\omega}_0} = -2\Delta_{\boldsymbol{\omega}_0}^T \mathscr{H} Q_{\widehat{\boldsymbol{\theta}}}^{-1} \Delta_{\boldsymbol{\omega}_0}$. Como já descrevemos a hessiana da função $Q = Q\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big)$ na análise de influência global, resta-nos obter $\Delta_{\boldsymbol{\omega}_0}$, que consiste na derivada em relação a $\boldsymbol{\theta}$ e $\boldsymbol{\omega}$ da função aumentada $Q_0 = Q_0\Big(\boldsymbol{\theta}, \boldsymbol{\omega}; \widehat{\boldsymbol{\theta}}\Big)$ aplicada no ponto $\Big(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}_0), \boldsymbol{\omega}_0\Big) = \Big(\widehat{\boldsymbol{\theta}}, \boldsymbol{\omega}_0\Big)$.

De modo geral, tal função assume a forma $Q_0\Big(\boldsymbol{\theta}, \boldsymbol{\omega}; \widehat{\boldsymbol{\theta}}\Big) = \sum\limits_{i=1}^{n} Q_{0i}\Big(\boldsymbol{\theta}, \omega_i; \widehat{\boldsymbol{\theta}}\Big)$, onde cada $Q_{0i}\Big(\boldsymbol{\theta}, \omega_i; \widehat{\boldsymbol{\theta}}\Big)$ possui uma configuração associada a $Q_i\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big)$ que depende do esquema de perturbação considerado. Analisaremos quatro casos básicos presentes na literatura sobre modelos similares: **ponderação de casos**, **perturbação na resposta**, **perturbação nas explicativas** e **perturbação nos parâmetros**. Em todos os casos, denotaremos por $\mathbf{c}_i$ o i-ésimo vetor canônico de $\mathbb{R}^n$.

(a) **Ponderação de casos**: Consiste basicamente em tomar pesos ω_i para cada indivíduo que incidem sobre a função log-verossimilhança de maneira que $Q_{0i}\Big(\boldsymbol{\theta}, \omega_i; \widehat{\boldsymbol{\theta}}\Big) = \omega_i Q_i\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big)$. Nesse caso, as condições iniciais do método são atendidas para $\Omega = \{\boldsymbol{\omega} \in \mathbb{R}^n : \omega_i > 0 \ \forall i = 1, \ldots, n\}$ com $\boldsymbol{\omega}_0 = \mathbf{1}$ e temos

$$\Delta_{\boldsymbol{\omega}_0} = \sum_{i=1}^{n} \frac{\partial Q_i}{\partial \boldsymbol{\theta}}\Big(\widehat{\boldsymbol{\theta}}; \widehat{\boldsymbol{\theta}}\Big)\mathbf{c}_i^T. \tag{4.18}$$

(b) **Perturbação nas respostas**: Considera o efeito da alteração no vetor de respostas de cada indivíduo na forma $\mathbf{y}_i + \omega_i \mathbf{D}\mathbf{S}_\mathbf{y}$, onde $\mathbf{S}_\mathbf{y}$ é o vetor de $\mathbb{R}^p$

cujas coordenadas correspondem ao desvio padrão amostral de cada uma das p variáveis respostas e $\mathbf{D}$ é uma matriz diagonal $p \times p$ cujos elementos na diagonal são variáveis indicadoras que assumem os valores 1 ou 0, expressando a incidência ou não da perturbação na correspondente coordenada da resposta. Fazendo $\Omega = \mathbb{R}^n$ e $\boldsymbol{\omega}_0 = \mathbf{0}$, vemos que esse esquema permite considerar a perturbação de qualquer subconjunto das variáveis respostas em cada indivíduo i, de modo que a função $Q_{0i}\Big(\boldsymbol{\theta}, \omega_i; \widehat{\boldsymbol{\theta}}\Big)$ será obtida de $Q_i\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big)$ pela introdução dos efeitos da perturbação sobre as partes que dependem do vetor de respostas $\mathbf{y}_i$, a saber d_i e A_i. Dessa forma, Q_{0i} fica completamente determinada pelas quantidades

$$d_i^{\boldsymbol{\omega}} = d_i + \omega_i(2\mathbf{e}_i + \omega_i \mathbf{D}\mathbf{S}_\mathbf{y})^T \mathbf{B}^{-2} \mathbf{D}\mathbf{S}_\mathbf{y} \quad \text{e} \quad A_i^{\boldsymbol{\omega}} = A_i + \omega_i \boldsymbol{\lambda}^T \mathbf{B}^{-1} \mathbf{D}\mathbf{S}_\mathbf{y}.$$

Após algumas manipulações algébricas, é possível deduzir a seguinte expressão geral válida para qualquer distribuição da classe SSMN com as convenções adotadas nos capítulos anteriores:

$$\Delta_{\boldsymbol{\omega}_0} = -\frac{1}{2} \sum_{i=1}^{n} \widehat{u}_i \frac{\partial^2 d_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^T} + \sum_{i=1}^{n} \Big(\widehat{t}_i - A_i\Big) \frac{\partial^2 A_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^T} - \sum_{i=1}^{n} \frac{\partial A_i}{\partial \boldsymbol{\theta}} \frac{\partial A_i^{\boldsymbol{\omega_0}}}{\partial \boldsymbol{\omega}^T}. \tag{4.19}$$

Todas as derivadas envolvidas em (4.19) são dadas adiante:

$$\begin{aligned}
&\frac{\partial A_i^{\boldsymbol{\omega_0}}}{\partial \boldsymbol{\omega}^T} = \boldsymbol{\lambda}^T \mathbf{B}^{-1} \mathbf{D}\mathbf{S}_\mathbf{y} \mathbf{c}_i^T; \qquad \frac{\partial^2 A_i^{\boldsymbol{\omega_0}}}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\omega}^T} = \mathbf{B}^{-1} \mathbf{D}\mathbf{S}_\mathbf{y} \mathbf{c}_i^T; \\
&\frac{\partial^2 d_i^{\boldsymbol{\omega_0}}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^T} = -2\mathbf{X}_i^T \mathbf{B}^{-2} \mathbf{D}\mathbf{S}_\mathbf{y} \mathbf{c}_i^T; \qquad \frac{\partial^2 A_i^{\boldsymbol{\omega_0}}}{\partial \alpha_k \partial \boldsymbol{\omega}^T} = -\boldsymbol{\lambda}^T \mathbf{B}^{-1} \dot{\mathbf{B}}_k \mathbf{B}^{-1} \mathbf{D}\mathbf{S}_\mathbf{y} \mathbf{c}_i^T; \\
&\frac{\partial^2 d_i^{\boldsymbol{\omega_0}}}{\partial \alpha_k \partial \boldsymbol{\omega}^T} = -2\mathbf{e}_i^T \mathbf{B}^{-1} (\dot{\mathbf{B}}_k \mathbf{B}^{-1} + \mathbf{B}^{-1} \dot{\mathbf{B}}_k) \mathbf{B}^{-1} \mathbf{D}\mathbf{S}_\mathbf{y} \mathbf{c}_i^T.
\end{aligned} \tag{4.20}$$

(c) **Perturbação nas explicativas**: Tomando aqui matrizes de planejamento na forma alternativa II como no conjunto de dados reais, consideraremos o efeito de alterar as variáveis explicativas da matriz de planejamento de cada indivíduo na forma $\mathbf{X}_i + \omega_i \mathbf{E}\mathbf{M}_\mathbf{x}$, onde $\mathbf{M}_\mathbf{x}$ é uma matriz diagonal (com ordem igual ao número de colunas de $\mathbf{X}_i$) cujos elementos na diagonal correspondem às replicações do desvio padrão amostral de cada uma das variáveis explicativas associadas a cada resposta além de um zero para cada

termo independente, e $\mathbf{E}$ é a matriz com a mesma dimensão de $\mathbf{X}_i$ de entradas iguais a 1 nas posições correspondentes àquelas variáveis sobre as quais incide a perturbação e 0 nas demais. Assumindo novamente $\Omega = \mathbb{R}^n$ e $\boldsymbol{\omega}_0 = \mathbf{0}$, a função $Q_{0i}\Big(\boldsymbol{\theta}, \omega_i; \widehat{\boldsymbol{\theta}}\Big)$ será obtida de $Q_i\Big(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}\Big)$ pela introdução da perturbação sobre as partes que dependem da matriz $\mathbf{X}_i$, notadamente d_i e A_i. Assim, Q_{0i} fica completamente determinada por

$$d_i^{\boldsymbol{\omega}} = d_i - \omega_i(2\mathbf{e}_i - \omega_i \mathbf{E}\mathbf{M}_{\mathbf{x}}\boldsymbol{\beta})^T\mathbf{B}^{-2}\mathbf{E}\mathbf{M}_{\mathbf{x}}\boldsymbol{\beta} \quad \text{e} \quad A_i^{\boldsymbol{\omega}} = A_i - \omega_i \boldsymbol{\lambda}^T\mathbf{B}^{-1}\mathbf{E}\mathbf{M}_{\mathbf{x}}\boldsymbol{\beta}.$$

Sendo assim, o valor de $\Delta_{\boldsymbol{\omega}_0}$ possui a mesma forma genérica indicada em (4.19) e pode ser explicitado utilizando as seguintes expressões:

$$\begin{aligned}
\frac{\partial A_i^{\boldsymbol{\omega_0}}}{\partial \boldsymbol{\omega}^T} &= -\boldsymbol{\lambda}^T\mathbf{B}^{-1}\mathbf{E}\mathbf{M}_{\mathbf{x}}\boldsymbol{\beta}\mathbf{c}_i^T; \qquad \frac{\partial^2 A_i^{\boldsymbol{\omega_0}}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\omega}^T} = -\mathbf{E}\mathbf{M}_{\mathbf{x}}^T\mathbf{B}^{-1}\boldsymbol{\lambda}\mathbf{c}_i^T; \\
\frac{\partial^2 A_i^{\boldsymbol{\omega_0}}}{\partial \alpha_k \partial \boldsymbol{\omega}^T} &= \boldsymbol{\lambda}^T\mathbf{B}^{-1}\dot{\mathbf{B}}_k\mathbf{B}^{-1}\mathbf{E}\mathbf{M}_{\mathbf{x}}\boldsymbol{\beta}\mathbf{c}_i^T; \qquad \frac{\partial^2 A_i^{\boldsymbol{\omega_0}}}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\omega}^T} = -\mathbf{B}^{-1}\mathbf{E}\mathbf{M}_{\mathbf{x}}\boldsymbol{\beta}\mathbf{c}_i^T; \\
\frac{\partial^2 d_i^{\boldsymbol{\omega_0}}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\omega}^T} &= -2(\mathbf{E}\mathbf{M}_{\mathbf{x}}^T\mathbf{B}^{-2}\mathbf{e}_i - \mathbf{X}_i^T\mathbf{B}^{-2}\mathbf{E}\mathbf{M}_{\mathbf{x}}\boldsymbol{\beta})\mathbf{c}_i^T; \\
\frac{\partial^2 d_i^{\boldsymbol{\omega_0}}}{\partial \alpha_k \partial \boldsymbol{\omega}^T} &= 2\mathbf{e}_i^T\mathbf{B}^{-1}(\dot{\mathbf{B}}_k\mathbf{B}^{-1} + \mathbf{B}^{-1}\dot{\mathbf{B}}_k)\mathbf{B}^{-1}\mathbf{E}\mathbf{M}_{\mathbf{x}}\boldsymbol{\beta}\mathbf{c}_i^T.
\end{aligned} \tag{4.21}$$

(d) **Perturbação nos parâmetros**: Trata-se da medida do efeito direto da alteração sobre os parâmetros conforme um esquema específico de perturbação para cada um deles. Nesse caso, temos em geral $Q_{0i}\Big(\boldsymbol{\theta}, \omega_i; \widehat{\boldsymbol{\theta}}\Big) = Q_i\Big(\boldsymbol{\theta}_i; \widehat{\boldsymbol{\theta}}\Big)$, onde $\boldsymbol{\theta}_i = \Big(\boldsymbol{\beta}_i, \boldsymbol{\alpha}_i, \boldsymbol{\lambda}_i, \boldsymbol{\tau}_i\Big)$ é o vetor de parâmetros alterado pelo esquema de perturbação adotado para cada indivíduo i. Na prática, considera-se o efeito da perturbação sobre cada grupo de parâmetros (especialmente escala e assimetria) de maneira isolada. Apresentaremos os dois casos referidos anteriormente de acordo com o que é encontrado na literatura para modelos similares — ver Zeller, Cabral & Lachos (2015). Vale dizer que não existe referência sobre perturbação no hiper-parâmetro, o que também não faremos aqui dada a complexidade e extrema especificidade desse procedimento. Com relação à perturbação nos parâmetros de escala e assimetria, há diferentes formas presentes na literatura – ver Ferreira (2008), por exemplo – dentre as quais citaremos as duas mais simples. Em ambos os casos, verificamos

que $\Delta_{\boldsymbol{\omega}_0}$ respeita a forma padrão dada em (4.19), de maneira que apenas nos restará explicitar as derivadas não nulas envolvidas na expressão.

1. **Perturbação na escala**: Nesse esquema, fazemos $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}/\sqrt{\omega_i}$, o que resulta em $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}/\omega_i$. Naturalmente, é preciso considerar a perturbação no aberto $\Omega = \{\boldsymbol{\omega} \in \mathbb{R}^n : \omega_i > 0 \ \forall i = 1, \ldots, n\}$ e temos $\boldsymbol{\omega}_0 = \mathbf{1}$. Assim, as expressões dadas adiante determinam $Q_{0i}\Big(\boldsymbol{\theta}, \omega_i; \widehat{\boldsymbol{\theta}}\Big)$:

$$d_i^{\boldsymbol{\omega}} = \omega_i \mathbf{e}_i^T \mathbf{B}^{-2} \mathbf{e}_i \ \text{ e } \ A_i^{\boldsymbol{\omega}} = \sqrt{\omega_i} \boldsymbol{\lambda}^T \mathbf{B}^{-1} \mathbf{e}_i.$$

   Assim, as derivadas listadas abaixo determinam $\Delta_{\boldsymbol{\omega}_0}$.

$$\begin{aligned}
&\frac{\partial^2 d_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^T} = -2\mathbf{X}_i^T \mathbf{B}^{-2} \mathbf{e}_i \mathbf{c}_i^T; \qquad \frac{\partial^2 d_i^{\boldsymbol{\omega}_0}}{\partial \alpha_k \partial \boldsymbol{\omega}^T} = -\mathbf{e}_i^T \mathbf{B}^{-1} (\dot{\mathbf{B}}_k \mathbf{B}^{-1} + \mathbf{B}^{-1} \dot{\mathbf{B}}_k) \mathbf{B}^{-1} \mathbf{e}_i \mathbf{c}_i^T; \\
&\frac{\partial A_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\omega}^T} = \frac{1}{2} \boldsymbol{\lambda}^T \mathbf{B}^{-1} \mathbf{e}_i \mathbf{c}_i^T; \qquad \frac{\partial^2 A_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^T} = -\frac{1}{2} \mathbf{X}_i^T \mathbf{B}^{-1} \boldsymbol{\lambda} \mathbf{c}_i^T; \\
&\frac{\partial^2 A_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\omega}^T} = \frac{1}{2} \mathbf{B}^{-1} \mathbf{e}_i \mathbf{c}_i^T; \qquad \frac{\partial^2 A_i^{\boldsymbol{\omega}_0}}{\partial \alpha_k \partial \boldsymbol{\omega}^T} = -\frac{1}{2} \boldsymbol{\lambda}^T \mathbf{B}^{-1} \dot{\mathbf{B}}_k \mathbf{B}^{-1} \mathbf{e}_i \mathbf{c}_i^T.
\end{aligned} \tag{4.22}$$

2. **Perturbação na assimetria**: Nesse caso, $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}\omega_i$ com $\boldsymbol{\omega} \in \Omega = \mathbb{R}^n$ e temos $\boldsymbol{\omega}_0 = \mathbf{1}$. Aqui a única expressão sobre a qual tal perturbação tem efeito e que determina $Q_{0i}\Big(\boldsymbol{\theta}, \omega_i; \widehat{\boldsymbol{\theta}}\Big)$ é a seguinte:

$$A_i^{\boldsymbol{\omega}} = \omega_i \boldsymbol{\lambda}^T \mathbf{B}^{-1} \mathbf{e}_i.$$

   Assim, a caracterização de $\Delta_{\boldsymbol{\omega}_0}$ para essa perturbação se dá somente pelo uso das expressões indicadas abaixo:

$$\begin{aligned}
&\frac{\partial A_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\omega}^T} = -\boldsymbol{\lambda}^T \mathbf{B}^{-1} \mathbf{e}_i \mathbf{c}_i^T; \qquad \frac{\partial^2 A_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^T} = -\mathbf{X}_i^T \mathbf{B}^{-1} \boldsymbol{\lambda} \mathbf{c}_i^T; \\
&\frac{\partial^2 A_i^{\boldsymbol{\omega}_0}}{\partial \alpha_k \partial \boldsymbol{\omega}^T} = -\boldsymbol{\lambda}^T \mathbf{B}^{-1} \dot{\mathbf{B}}_k \mathbf{B}^{-1} \mathbf{e}_i \mathbf{c}_i^T; \qquad \frac{\partial^2 A_i^{\boldsymbol{\omega}_0}}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\omega}^T} = \mathbf{B}^{-1} \mathbf{e}_i \mathbf{c}_i^T.
\end{aligned} \tag{4.23}$$

Observe que o fato de o esquema de perturbação da assimetria não alterar a função d_i faz com que a expressão de $\Delta_{\boldsymbol{\omega}_0}$ dada em (4.19) seja a mesma para qualquer distribuição da família SSMN. Na verdade, essa expressão também é a mesma para a distribuição normal assimétrica, sendo portanto uma característica comum das famílias de distribuições assimétricas consideradas neste texto.

Realizaremos agora a aplicação dessa técnica de análise da influência local para detectar observações influentes com base na medida (4.17) nos modelos normal, normal assimétrico (SN) e Slash normal assimétrico (SSN) ajustados aos dados dos municípios de Minas Gerais no fim do Capítulo 3. Nas Figuras 37, 38, 39 e 40 apresentamos os casos dos quatro tipos de perturbação que mais destacam as diferenças entre os três modelos.

Figura 37 – Valores de M_0 para ponderação de casos nos modelos normal, SN e SSN

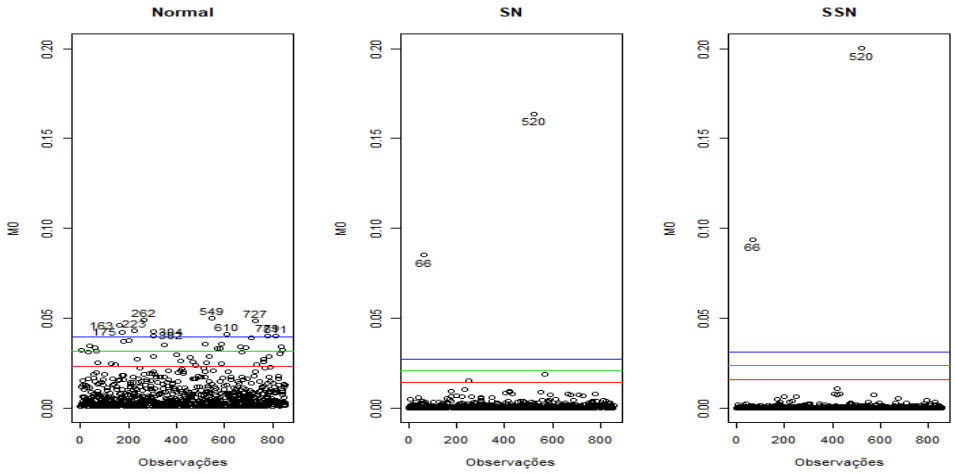


Figura 38 – Valores de M_0 para perturbação da renda *per capita* nos modelos normal, SN e SSN

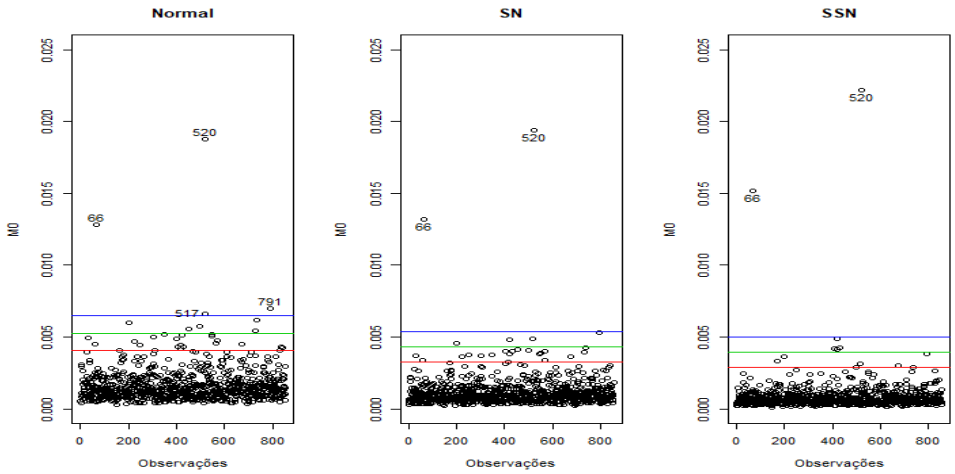


Figura 39 – Valores de M_0 para perturbação da taxa de analfabetismo relativamente à renda *per capita* nos modelos normal, SN e SSN

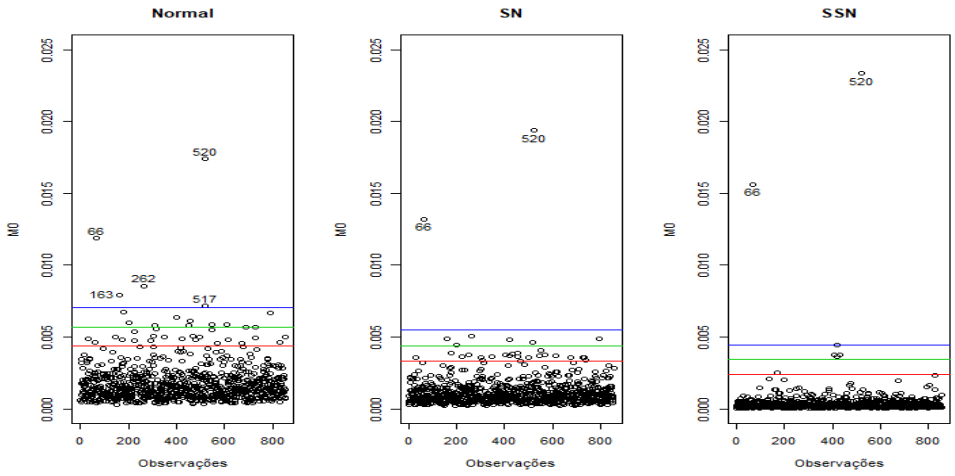
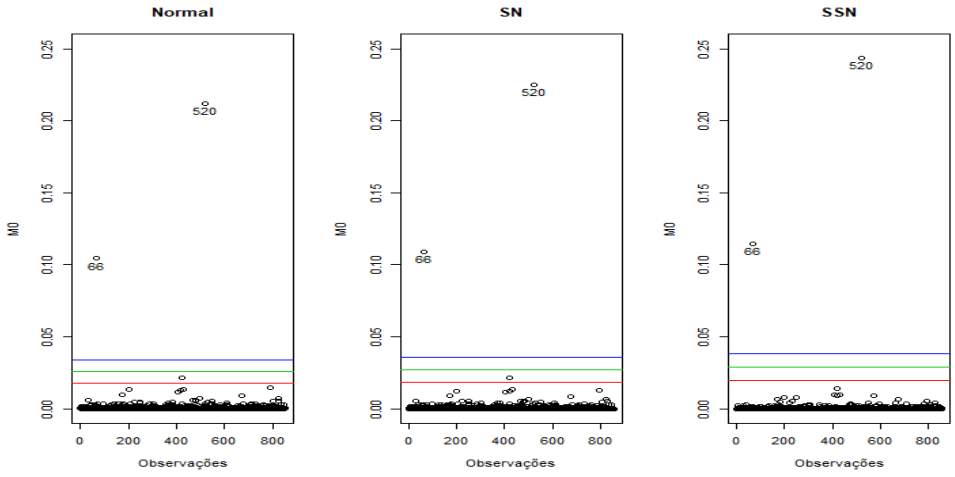


Figura 40 – Valores de M_0 para perturbação da escala nos modelos normal, SN e SSN



Em todos os 4 gráficos anteriores, as três linhas representam de baixo para cima os pontos de corte obtidos adotando para a constante c^* da expressão (4.17) os valores 2, 3 e 4. Diante do nosso propósito de apenas ilustrar a aplicação da técnica, adotaremos o valor $c^* = 4$ conforme Zeller *et al.* (2010) sem maiores discussões.

Procedendo dessa forma, vemos que o esquema de perturbação da escala é o único em que os três modelos identificam os mesmos pontos conforme a Figura 40, novamente BH e Nova Lima sobre os quais discorremos na Seção 4.1.

Como já destacamos na referida seção, ambos os municípios se notabilizam pela renda *per capita* muito elevada e, por isso, aparecem bem destacados também no esquema de perturbação dessa variável para os três modelos de acordo com a Figura 38. No entanto, o ponto de corte assumido também aponta outros dois municípios como influentes no modelo normal: 517 (Ninheira) e 791 (Setubinha). Esses municípios, por outro lado, destacam-se pela baixa renda. Enquanto Setubinha tem a 12ª pior renda do estado (R$ 224,63), Ninheira possui a 8ª pior (R$ 210,17). Além disso, apresentam outros indicadores negativos: Ninheira tem a 12ª pior taxa de atividade (48,22%) e a 2ª maior taxa de analfabetismo (37,61%), já Setubinha possui a 5ª pior esperança de vida (69,6 anos), a 4ª maior taxa de analfabetismo (35,76%) e a maior razão de dependência de Minas Gerais (76,19%).

Com relação aos outros esquemas de perturbação, os modelos SN e SSN só permitiram destacar os municípios de BH e Nova Lima, o que nos levou a investigar suas variáveis explicativas e concluir que duas delas são extremas: em Nova Lima, tem-se razão de dependência 38,12% (10ª menor) e taxa de analfabetismo 2,99% (a menor) e em BH, tem-se razão de dependência 37,79% (6ª menor) e taxa de analfabetismo 3,01% (2ª menor). Dos outros municípios que aparecem nas Figuras 37 e 39, destacamos os dois que aparecem em ambos os gráficos e também no gráfico de influência global da Figura 32: 163 (Casa Grande) e 262 (Doresópolis). O segundo, que foi mencionado na análise de influência global com ênfase na sua esperança de vida (57ª maior de MG), também tem algum destaque pela taxa de analfabetismo relativamente baixa: 6,97 (108ª menor). Já o primeiro se notabiliza por uma taxa de analfabetismo um pouco melhor: 6,22 (63ª menor do estado).

Sobre o fato de a análise de influência local aplicada aos modelos normal assimétrico e Slash normal assimétrico identificar os mesmos pontos influentes, cabe um breve comentário relativo aos valores obtidos. Nota-se que o modelo Slash normal assimétrico apresenta em geral valores ligeiramente mais homogêneos para M_0, donde poderíamos inferir uma adequação geral melhor desse modelo exceto nos dois pontos influentes.

Por fim, após as análises de influência global e local, podemos concluir que as informações extraídas das duas abordagens são praticamente equivalentes para a identificação de observações influentes, motivando a sua investigação particular. Na comparação do modelo normal com o modelo SN e os modelos da família SSMN, quando estes se ajustam melhor aos dados, fica evidente a clareza maior na detecção dos pontos influentes que os últimos modelos apresentam, ao passo que no modelo normal observações influentes acabam mascaradas por outras "pseudo-influentes" ou influentes somente em relação a uma variável ou parâmetro.

Neste texto, devido a limitações computacionais ficamos restritos às técnicas clássicas para o diagnóstico de influência sem muito aprofundamento. Contudo, encerraremos com uma menção a duas medidas que motivam a análise de influência conjunta: TRC e MRC, utilizadas em Lee, Lu & Song (2006). Como no nosso exemplo usaremos tais medidas apenas nas observações 66 e 520, vamos defini-las somente para os casos da retirada de uma ou duas observações. Denotando por $\widehat{\boldsymbol{\theta}}_{[i]}$ a estimativa do vetor de parâmetros (com comprimento r) do modelo sem a i-ésima observação e por $\widehat{\boldsymbol{\theta}}_{[i,j]}$ a estimativa sem as observações i e j, definimos as medidas associadas, respectivamente, por:

$$\begin{aligned} \mathrm{TRC}_{[i]} &= \sum_{k=1}^{r}\left|\frac{\widehat{\theta}_k-\widehat{\theta}_{[i]k}}{\widehat{\theta}_k}\right| \qquad \mathrm{TRC}_{[i,j]}=\sum_{k=1}^{r}\left|\frac{\widehat{\theta}_k-\widehat{\theta}_{[i,j]k}}{\widehat{\theta}_k}\right| \\ \mathrm{MRC}_{[i]} &= \max_{1\le k\le r}\left|\frac{\widehat{\theta}_k-\widehat{\theta}_{[i]k}}{\widehat{\theta}_k}\right| \qquad \mathrm{MRC}_{[i,j]}=\max_{1\le k\le r}\left|\frac{\widehat{\theta}_k-\widehat{\theta}_{[i,j]k}}{\widehat{\theta}_k}\right| \end{aligned} \tag{4.24}$$

Na Tabela 20, as medidas anteriores nos três modelos de maior interesse para $i, j \in \{66, 520\}$ mostram como o impacto da retirada das observações vai diminuindo em geral à medida que consideramos modelos com mais parâmetros.

Tabela 20 – Medidas MRC e TRC nos modelos normal, SN e SSN

Modelo/Medida	$\mathrm{TRC}_{[66]}$	$\mathrm{TRC}_{[520]}$	$\mathrm{TRC}_{[66,520]}$	$\mathrm{MRC}_{[66]}$	$\mathrm{MRC}_{[520]}$	$\mathrm{MRC}_{[66,520]}$
Normal	0,3148	0,4134	0,7413	0,0944	0,1355	0,2196
SN	0,3097	0,2838	0,5532	0,0878	0,1225	0,2083
SSN	0,1676	0,2100	0,4720	0,0730	0,0921	0,2114

Existem outros métodos mais robustos para realizar esse tipo de análise, para os quais apenas citaremos referência no Capítulo 5.

5 CONCLUSÃO

Para encerrar este trabalho, destacaremos alguns resultados importantes obtidos, pontos problemáticos e extensões em aberto para abordagens futuras.

Sem dúvida, os resultados mais relevantes que conseguimos obter foram as técnicas de estimação explícita dos hiper-parâmetros nos modelos SSMN não só pelo nítido ganho de tempo, mas também pela eliminação da única "caixa preta" presente no algoritmo. Sobre isso, talvez ainda seja possível melhorar a aproximação utilizada no modelo STN de modo a aumentar sua precisão sem que a estimativa do hiper-parâmetro deixe de ser explícita. Além disso, um aspecto a ser investigado é a possibilidade de se tratar as distribuições SMSN e SSMN como uma única família e generalizar o método de estimação apresentado.

Como principal problema deste trabalho no que tange à falta de investigações mais aprofundadas sobre aspectos numéricos do processo de estimação e outros procedimentos, apontamos a limitação computacional. Mesmo com os recursos de programação em paralelo implementados num computador de 8 núcleos, as simulações que realizamos se mostraram bastante demoradas e difíceis de controlar. Isso afetou principalmente o uso de métodos *bootstrap* na análise de influência como os utilizados em Zeller *et al.* (2010) para justificar os critérios adotados.

Além disso, outras possíveis formas mais amplas de se fazer análise de influência levam em conta a possibilidade de se medir o impacto da detecção de subconjuntos de observações livremente ou condicionadas à retirada prévia de outras. Isso remete às noções de *influência conjunta* e *influência condicional* apresentadas em Li, Xu & Zhu (2009) no contexto de modelos mistos, mas que poderiam em tese ser adaptadas para o caso das distribuições SSMN.

Dessa forma, entre as futuras extensões possíveis, pode-se mencionar a implementação de tais métodos, inclusive com vistas à determinação dos pontos de corte nas medidas de influência global e local. Mais ainda, é possível pensar até mesmo em usar métodos similares para realizar uma análise de resíduos dos modelos para a verificação de seus pressupostos de independência e adequação da distribuição correspondentes. Tudo isso são "cenas" de próximos capítulos.

REFERÊNCIAS

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* , 19 (6), 716–723.

ANDREWS, D. F.; MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* , 99–102.

Atlas do Desenvolvimento Humano no Brasil. Disponível em: `http://www.atlasbrasil.org.br/2013/pt/consulta/`. Acesso em: 29 de novembro de 2017.

AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* , 171–178.

AZZALINI, A.; CAPITANIO, A. (1999). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society* , 65 (3), 367–389.

AZZALINI, A.; DALLA-VALLE, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83 (4), 715–726.

BOLDRIN, M.; MONTRUCCHIO, L. (1987). Cyclic and Chaotic Behavior in Intertemporal Optimization Models. *Mathematical Modelling*, 8, 697–700.

BRANCO, M. D.; SAHU, S. K; DEY, D. K (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*, 31(2), 129–150.

BUSSAB, W. O.; MORETTIN, P. A. (2012). *Estatística Básica, Editora Saraiva, 5a. edição*. São Paulo, Saraiva.

CASELLA, G.; BERGER, R. L. (2002). *Statistical Inference*, volume 2. Duxbury, Pacific Grove, CA.

COOK, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics* , 19 (1), 15–18.

COOK, R. D.; WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York, Chapman & Hall.

COOK, R. D. (1986). Assessment of Local Influence. *Journal of the Royal Statistical Society, Series B*, 48 (2), 133–169.

COOK, R. D.; WEISBERG, S. (1994). *An Introduction to Regression Graphics*. New York, John Wiley & Sons.

CORDEIRO, G. M. (1999). *Introdução à Teoria Assintótica*. Rio de Janeiro, IMPA.

DEMPSTER, A.; LEIRD, N.; RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1), 1–38.

FARIA, V. B. (2011). Estimação de máxima verossimilhança via algoritmo EM. Trabalho de Conclusão de Curso (Graduação em Estatística).

FERREIRA, C. S. (2008). *Inferência e Diagnóstico em Modelos Assimétricos*. Tese de Doutorado, Universidade de São Paulo, São Paulo.

FERREIRA, C. S.; LACHOS, V. H.;BOLFARINE, H. (2016). Likelihood-based Inference for Multivariate Skew Scale Mixtures of Normal Distributions. *AStA Advances in Statistical Analysis*, 1–21.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. New York, Springer-Verlag.

HARVILLE, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York, Springer Verlag.

JAMALIZADEH, A.; LIN, T. (2016). A general class of scale-shape mixtures of skew-normal distributions: properties and estimation. *Computational Statistics*, 1–24.

JOHNSON, N. L.; KOTZ, S., BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions*. New York, Wiley.

JOHNSON, R. A.; WICHERN, D. W. (2007). *Applied Multivariate Statistical Analysis*. New Jersey, Prentice Hall.

KARLSSON, M.; LAITILA, T. (2014). Finite Mixture Modeling of Censored Regression Models. *Statistical papers*, 55(3), 627–642.

LACHOS, V. H. (2004). *Modelos Lineares Mistos Assimétricos*. Tese de Doutorado, Universidade de São Paulo, São Paulo.

LACHOS, V. H.; GHOSH, P.; ARELLANO-VALLE, R. B. (2010). Likelihood based inference for skew– normal/independent linear mixed models. *Statistica Sinica*, 303–322.

LACHOS, V. H.; VILCA, F. E. (2007). Skew-Normal/Independent Distributions, with Applications. RT-IMECC 02, IMECC-UNICAMP.

LANGE, K. L.; SINSHEIMER, J. S. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Data Analysis*, 2, 175—198.

LEE, S.; XU, L. (2004). Infuence analyses of nonlinear mixed-effects models. *Computational Statistics & Data Analysis*, 45, 321–341.

LEE, S. Y.; LU, B.; SONG, X. Y. (2006). Assessing local influence for nonlinear structural equation models with ignorable missing data. *Computational Statistics & Data Analysis* , 50, 1356–1377.

LIMA, E. L. (1999). *Curso de Análise vol. 2*. Rio de Janeiro, IMPA (Projeto Euclides).

LI, Z.; XU, W.; ZHU, W. (2009). Influence diagnostics and outlier tests for varying coefficient mixed models. *Journal of Multivariate Analysis*, 100(9), 2002–2017.

LIU, C.; RUBIN, D. B. (1994). The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence. *Biometrika*, 80, 267–278.

LOUREDO, G. M. S. (2016). Modelos de Regressão Linear com Erros Normais. Trabalho de Conclusão de Curso (Graduação em Matemática).

MASSUIA, M. B.; CABRAL, C. R. B.; Matos, L. A.; Lachos V. H. (2015). Influence diagnostics for T-Student censored linear regression models. *Statistics*, 49, 1074–1094.

MAGNUS, J. R.; NEUDECKER, H. (1988). *Matrix Differential Calculus with Spplications in Statistics and Econometrics*. Hoboken, John Wiley & Sons.

MCLACHLAN, G. J.; KRISHNAN, T. (2001). *The EM Algorithm and Extensions*. Hoboken, John Wiley & Sons.

MCLACHLAN, G. J.; PEEL, D. (2000). *Finite Mixture Models*. Hoboken, John Wiley & Sons.

MENG, X. L.; RUBIN, D. B.(1993) Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, 80 (2), 267–278.

O'NEILL, B. (2006). *Elementary Differential Geometry*. Philadelphia, Elsevier Inc.

PAN, J.; FEI, Y.; FOSTER, P. (2013). Case-deletion Diagnostics for Linear Mixed Models. *Technometrics*, 56 (3), 269–281.

PAWITAN, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford, Clarendon Press.

POON, W. Y.; POON, Y. S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society, Series B*, 51–61.

POWELL, M. J. D. (1981). *Approximation Theory and Methods*. Cambridge University Press.

RITTER, G. (2015). *Robust Cluster Analysis and Variable Selection*. Boca Raton, CRC Press.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

RUSSO, C. M.; PAULA, G. A.; AOKI, R. (2009). Influence diagnostics in nonlinear mixed-effects elliptical models. *Computational Statistics and Data Analysis*, 53 (12), 4143–4156.

SOUZA FILHO, N. L. (2012). *Modelagem Bayesiana Flexível em Regressão com Erros nas Variáveis*. Dissertacão de Mestrado, Universidade Federal do Amazonas, Manaus.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6 (2), 461–464.

STIGLER, S. M. (2007). The Epic Story of Maximum Likelihood. *Statistical Science*, 598–620.

WANG, J.; GENTON, M. G. (2006). The multivariate skew-slash distribution. *Journal of Statistical Planning and Inference*, 136 (1), 209–220.

WU, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11 (1), 95–103.

ZELLER, C. B.; CABRAL, C. R. B.; LACHOS, V. H (2015). Robust Mixture Regression Modeling Based on Scale Mixtures Skew-Normal Distributions. *TEST*, 25 (2), 375–396.

ZELLER, C. B. (2009). *Distribuições Misturas de Escala Skew-normal: Estimação e Diagnóstico em Modelos Lineares*. Tese de Doutorado, Unicamp, Campinas.

ZELLER, C. B.; LACHOS, V. H; VILCA-LABRA, F. E. (2009). Local Influence Analysis for Regression Models with Scale Mixtures of Skew-Normal Distributions. *Journal of Applied Statistics*, 1–21.

ZELLER, C. B.; LACHOS, V. H; VILCA-LABRA, F. E.; BALAKRISHNAN, N. (2010). Influence analyses of skew normal independent linear mixed models. *Computational Statistics and Data Analysis*, 54 (5), 1266–1280.

ZHU, H.; ZHOU, J.; WEI, B. C.; LEE, S. Y. (2001). Case-deletion Measures for Models with Incomplete-data. *Biometrika*, 88 (3), 727–737.

ZHU, H.; LEE, S. (2001). Local Influence for Incomplete-data Models. *Journal of the Royal Statistical Society, Series B*, 63, 111–126.

APÊNDICE A – Tópicos de Álgebra e Cálculo Matricial

Teorema A.0.1 (**Identidade de Woodbury**)**.** Dadas matrizes $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{B} \in \mathbb{R}^{n\times k}, \mathbf{C} \in \mathbb{R}^{k\times k}$ e $\mathbf{D} \in \mathbb{R}^{k\times n}$, vale a relação

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}.$$

Teorema A.0.2. Dadas matrizes $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{B} \in \mathbb{R}^{n\times k}, \mathbf{C} \in \mathbb{R}^{k\times k}$ e $\mathbf{D} \in \mathbb{R}^{k\times n}$, temos que

$$|\mathbf{A}+\mathbf{BCD}| = |\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B}||\mathbf{C}||\mathbf{A}|.$$

A partir de agora, apresentaremos definições e resultados do chamado Cálculo Matricial, cujas notações foram utilizadas nas condições de primeira ordem para estimação dos parâmetros nos modelos multivariados e nas expressões das derivadas de log-verossimilhanças em relação aos parâmetros. Não se trata de um novo tipo de Cálculo, mas tão somente de notações especiais, que estendem a notação de Leibniz para englobar casos de funções que tenham matrizes (ou em particular vetores) como variáveis com o objetivo de tornar o processo de derivação mais intuitivo. Referências que utilizam essa abordagem aplicada a problemas estatísticos são Harville (1997) e Magnus & Neudecker (1988).

Definição A.0.1. Seja $F\colon \mathbb{R}^{r\times s} \to \mathbb{R}^{k\times l}$ uma aplicação diferenciável. Denotaremos a derivada de cada função coordenada F_{ij} com respeito a $\mathbf{X} \in \mathbb{R}^{r\times s}$ por

$$\frac{\partial F_{ij}}{\partial \mathbf{X}} = \begin{bmatrix} \dfrac{\partial F_{ij}}{\partial x_{1\,1}} & \cdots & \dfrac{\partial F_{ij}}{\partial x_{1\,s}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_{ij}}{\partial x_{r\,1}} & \cdots & \dfrac{\partial F_{ij}}{\partial x_{r\,s}} \end{bmatrix}, \text{ sendo } \mathbf{X} = \begin{bmatrix} x_{1\,1} & \cdots & x_{1\,s} \\ \vdots & \ddots & \vdots \\ x_{r\,1} & \cdots & x_{r\,s} \end{bmatrix}. \tag{A.1}$$

Observação: Foram bastante utilizados no texto os casos particulares que mostraremos adiante.

- Para $s = l = k = 1$ e $\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_r \end{bmatrix}$ na Definição A.0.1, escreveremos $\dfrac{\partial F}{\partial \mathbf{X}} = \begin{bmatrix} \dfrac{\partial F}{\partial x_1} \\ \vdots \\ \dfrac{\partial F}{\partial x_r} \end{bmatrix}$ e $\dfrac{\partial F}{\partial \mathbf{X}^T} = \begin{bmatrix} \dfrac{\partial F}{\partial x_1} & \cdots & \dfrac{\partial F}{\partial x_r} \end{bmatrix}$, indicando o vetor gradiente de F e seu transposto, respectivamente. Além disso, se F for de classe $\mathscr{C}^2$, sua matriz hessiana é expressa por

$$\begin{aligned} \frac{\partial^2 F}{\partial \mathbf{X} \partial \mathbf{X}^T} &= \frac{\partial}{\partial \mathbf{X}^T}\left(\frac{\partial F}{\partial \mathbf{X}}\right) = \frac{\partial}{\partial \mathbf{X}}\left(\frac{\partial F}{\partial \mathbf{X}^T}\right) = \left[\frac{\partial}{\partial \mathbf{X}}\left(\frac{\partial F}{\partial x_1}\right) \quad \cdots \quad \frac{\partial}{\partial \mathbf{X}}\left(\frac{\partial F}{\partial x_r}\right)\right] \\ &= \begin{bmatrix} \dfrac{\partial^2 F}{\partial x_1^2} & \cdots & \dfrac{\partial^2 F}{\partial x_r \partial x_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 F}{\partial x_1 \partial x_r} & \cdots & \dfrac{\partial^2 F}{\partial x_r^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 F}{\partial x_1^2} & \cdots & \dfrac{\partial^2 F}{\partial x_1 \partial x_r} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 F}{\partial x_1 \partial x_r} & \cdots & \dfrac{\partial^2 F}{\partial x_r^2}. \end{bmatrix} \end{aligned}$$

- Para $s = l = 1$ e $\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_r \end{bmatrix}$ na Definição A.0.1, sabendo que $\dfrac{\partial F_j}{\partial \mathbf{X}} = \begin{bmatrix} \dfrac{\partial F_j}{\partial x_1} \\ \vdots \\ \dfrac{\partial F_j}{\partial x_r} \end{bmatrix}$ para todo $j \in \{1, \ldots, k\}$, indicamos a matriz jacobiana de F por

$$\frac{\partial F}{\partial \mathbf{X}} = \begin{bmatrix} \dfrac{\partial F_1}{\partial \mathbf{X}^T} \\ \vdots \\ \dfrac{\partial F_k}{\partial \mathbf{X}^T} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial F_1}{\partial x_1} & \cdots & \dfrac{\partial F_k}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_1}{\partial x_r} & \cdots & \dfrac{\partial F_k}{\partial x_r} \end{bmatrix}.$$

Com as convenções anteriores, podemos escrever a Regra da Cadeia do Cálculo de Várias Variáveis da seguinte forma:

Proposição A.0.1. Se $F\colon \mathbb{R}^p \to \mathbb{R}^n$ e $G\colon \mathbb{R}^q \to \mathbb{R}^p$ são funções diferenciáveis, então

$$\frac{\partial F(G(\mathbf{z}))}{\partial \mathbf{z}} = \left.\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=G(\mathbf{z})} \frac{\partial G(\mathbf{z})}{\partial \mathbf{z}}. \tag{A.2}$$

Observação: Se $q = p$ e a lei $\mathbf{x} = G(\mathbf{z})$ define um difeomorfismo, então $\dfrac{\partial F(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0} \Leftrightarrow \dfrac{\partial F(G(\mathbf{z}))}{\partial \mathbf{z}} = \mathbf{0}$. Esse fato foi utilizado para substituir as derivadas de Q_1 por outras relativas a uma reparametrização na estimação dos modelos misturas multivariados com o intuito de facilitar as contas.

Por fim, o cálculo de derivadas no processo de estimação e também do escore e da matriz de informação, especialmente no caso dos modelos multivariados, decorre da aplicação sucessiva dos resultados que reunimos adiante. Eles podem ser encontrados em Harville (1997) ou Magnus & Neudecker (1988).

Proposição A.0.2. Dada $\mathbf{X} \in \mathbb{R}^{p\times p}$, temos

(a) $\dfrac{\partial}{\partial \mathbf{X}}(\mathbf{v}^T\mathbf{X}\mathbf{v}) = \mathbf{v}\mathbf{v}^T \quad \forall \mathbf{v} \in \mathbb{R}^p$;

(b) $\dfrac{\partial}{\partial \mathbf{X}} \ln|\mathbf{X}| = 2\mathbf{X}^{-1} - \operatorname{diag}(\mathbf{X}^{-1})$ para $\mathbf{X}$ simétrica invertível.

Proposição A.0.3. Para qualquer $\mathbf{x} \in \mathbb{R}^p$, temos

(a) $\dfrac{\partial}{\partial \mathbf{x}}(\mathbf{A}\mathbf{x}) = \mathbf{A}^T \quad \forall \mathbf{A} \in \mathbb{R}^{m\times p}$;

(b) $\dfrac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x} \quad \forall \mathbf{A} \in \mathbb{R}^{p\times p}$.

Proposição A.0.4. Seja $F\colon \mathbb{R}^p \to \mathbb{R}^{k\times k}$ diferenciável. Se $\mathbf{x} \in \mathbb{R}^p$, então

(a) $\dfrac{\partial}{\partial x_j} \ln|F(\mathbf{x})| = \operatorname{tr}\left(F(\mathbf{x})^{-1}\dfrac{\partial F}{\partial x_j}\right)$;

(b) $\dfrac{\partial^2}{\partial x_i \partial x_j} \ln|F(\mathbf{x})| = \operatorname{tr}\left(F(\mathbf{x})^{-1}\dfrac{\partial^2 F}{\partial x_i \partial x_j}\right) - \operatorname{tr}\left(F(\mathbf{x})^{-1}\dfrac{\partial F}{\partial x_i}F(\mathbf{x})^{-1}\dfrac{\partial F}{\partial x_j}\right)$;

(c) $\dfrac{\partial}{\partial x_j}\left(\mathbf{A}F(\mathbf{x})^{-1}\mathbf{B}\right) = -\left(\mathbf{A}F(\mathbf{x})^{-1}\dfrac{\partial F}{\partial x_j}F(\mathbf{x})^{-1}\mathbf{B}\right) \quad \forall \mathbf{A} \in \mathbb{R}^{r\times k}, \mathbf{B} \in \mathbb{R}^{k\times s}$.

Observação: O último resultado é particularmente útil para computar as derivadas de $\Lambda = \ln|\boldsymbol{\Sigma}| = 2\ln|\mathbf{B}(\boldsymbol{\alpha})|$, bem como de formas lineares e quadráticas envolvendo a matriz $\mathbf{B}^{-1}$, em relação às coordenadas de $\boldsymbol{\alpha} = \operatorname{vech}(\mathbf{B})$ nos modelos multivariados.

APÊNDICE B – Elementos de Teoria da Aproximação

A **Teoria da Aproximação** surgiu no final do século XIX com os trabalhos do matemático russo Chebyshev e se desenvolveu com trabalhos de seus alunos Markov e Bernstein e também de Jackson no início do século XX. Todos esses trabalhos, além de generalizações posteriores como o Teorema de Stone-Weierstrass, eram de natureza muito teórica e pouco prática, porém de fundamental importância para embasar técnicas aplicadas recentemente na **Análise Numérica**.

Neste texto, empregamos uma dessas técnicas considerada variante da chamada *aproximação por mínimos quadrados ponderados* para propor um método de estimação do hiper-parâmetro no modelo T-Student normal assimétrico. Vamos desenvolver aqui os principais conceitos e resultados que fundamentaram essa noção de aproximação, que podem ser vistos com mais detalhes em Powell (1981).

Definição B.0.1. Sejam $(\mathscr{B}, \|\cdot\|)$ um espaço vetorial normado e $\mathscr{A} \subset \mathscr{B}$. Dado $f \in \mathscr{B}$, dizemos que $g \in \mathscr{A}$ é uma *melhor aproximação* para f em $\mathscr{A}$ quando $\|f - g\| \leq \|f - q\| \ \ \forall q \in \mathscr{A}$.

O resultado que enunciaremos a seguir é de grande relevância para a garantia de validade da técnica utilizada neste texto. Sua demonstração decorre dos teoremas 1.2 e 2.4 de Powell (1981), respectivamente. Antes, porém, de enunciá-lo vejamos a definição abaixo.

Definição B.0.2. Se $(\mathscr{B}, \|\cdot\|)$ um espaço vetorial normado, dizemos que a norma $\|\cdot\|$ é *estritamente convexa* quando toda bola $B[f; r] = \{h \in \mathscr{B} : \ \|h - f\| \leq r\}$ é um conjunto estritamente convexo, isto é, para quaisquer $p \neq q$ em $B[f; r]$ os pontos da forma $(1 - k)p + kq, \ k \in (0, 1)$ pertencem ao interior de $B[f; r]$.

Teorema B.0.1. Sejam $(\mathscr{B}, \|\cdot\|)$ um espaço vetorial normado e $\mathscr{A}$ um subespaço vetorial com dimensão finita de $\mathscr{B}$. Se a norma $\|\cdot\|$ é estritamente convexa, então existe uma única melhor aproximação em $\mathscr{A}$ para cada $f \in \mathscr{B}$.

Vejamos agora alguns pontos referentes especificamente à aproximação pelo método dos mínimos quadrados ponderados, os quais estão presentes em Powell (1981) a partir da página 123.

Definição B.0.3. Considere $\mathscr{B} = \mathscr{C}[a,b]$ o espaço vetorial de todas as funções $f\colon [a,b] \to \mathbb{R}$ contínuas. Se $\mathscr{A} \subset \mathscr{B}$ e $w \in \mathscr{B}$ é uma função positiva (função peso), então dada $f \in \mathscr{B}$, dizemos que $p \in \mathscr{A}$ é uma *melhor aproximação de mínimos quadrados ponderados* em $\mathscr{A}$ para f quando p minimiza a expressão $\int_a^b w(t)[f(t) - q(t)]^2 dt,\ q \in \mathscr{A}$.

Observação: A função $\langle\ \cdot,\ \cdot\ \rangle : \mathscr{C}[a,b] \times \mathscr{C}[a,b] \to \mathbb{R}$ dada por $\langle f, g\rangle = \int_a^b w(t)f(t)g(t)dt$ define um produto interno em $\mathscr{C}[a,b]$. Consequentemente, temos $\int_a^b w(t)[f(t) - q(t)]^2 dt = \|f - q\|^2$, onde $\|f\| = \sqrt{\langle f, f\rangle}$, na Definição B.0.3. Dessa forma, a noção de aproximação dessa última definição condiz com a forma geral dada na Definição B.0.1.

O teorema a seguir é uma versão adaptada do resultado cujo enunciado e demonstração se encontra em Powell (1981). Esse teorema caracteriza a aproximação por mínimos quadrados em um subespaço finito-dimensional como a projeção ortogonal sobre esse subespaço.

Teorema B.0.2. Sejam $\mathscr{B} = \mathscr{C}[a,b]$ com a norma dada por $\|f\| = \sqrt{\int_a^b w(t)[f(t)]^2}$ $\forall f \in \mathscr{B}$ e $\mathscr{A}$ um subespaço vetorial com dimensão finita de $\mathscr{B}$. Dado $f \in \mathscr{B}$, o ponto $p \in \mathscr{A}$ é a (única) melhor aproximação de mínimos quadrados ponderados em $\mathscr{A}$ para f se, e somente se, o erro de aproximação $e = f - p$ satisfaz $\langle e, q\rangle = 0\ \forall q \in \mathscr{A}$.

O cálculo da melhor aproximação de mínimos quadrados p de f nas condições (verdadeiras) do Teorema B.0.2, é feita da seguinte maneira: se $\mathscr{U} = \{f_0, f_1, \ldots, f_n\}$ base de $\mathscr{A}$, podemos escrever $p = \sum_{j=0}^{n} c_j f_j$ e segue do teorema que $0 = \langle f - p, f_i\rangle = \langle f, f_i\rangle - \sum_{j=0}^{n} c_j \langle f_j, f_i\rangle\ \ \forall i \in \{0, 1, \ldots, n\}$. Esse resultado indica que a determinação de p equivale a encontrar a solução $(c_0, c_1, \ldots, c_n)$ do sistema a seguir:

$$\begin{cases} \sum_{j=0}^{n} c_j \langle f_j, f_0\rangle = \langle f, f_0\rangle \\ \sum_{j=0}^{n} c_j \langle f_j, f_1\rangle = \langle f, f_1\rangle \\ \vdots \\ \sum_{j=0}^{n} c_j \langle f_j, f_n\rangle = \langle f, f_n\rangle \end{cases} \Leftrightarrow \begin{bmatrix} \langle f_0, f_0\rangle & \langle f_0, f_1\rangle & \ldots & \langle f_0, f_n\rangle \\ \langle f_1, f_0\rangle & \langle f_1, f_1\rangle & \ldots & \langle f_1, f_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle f_n, f_0\rangle & \langle f_n, f_1\rangle & \ldots & \langle f_n, f_n\rangle \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \langle f, f_0\rangle \\ \langle f, f_1\rangle \\ \vdots \\ \langle f, f_n\rangle \end{bmatrix}. \tag{B.1}$$

Observação: Se $\mathscr{U}$ for uma base ortogonal, então $p = \sum_{j=0}^{n} \frac{\langle f_j, f\rangle}{\|f_j\|^2} f_j$. Nesse caso, ao qual o caso geral pode ser reduzido via ortogonalização de Gram-Schmidt, descrevemos abaixo em três passos como obter o valor de n necessário a se atingir uma precisão desejada δ na aproximação, ou seja, encontrar n tal que $\|f - p_n\| < \delta$, onde p_n é a aproximação de mínimos quadrados ponderados de f utilizando os primeiros $n+1$ elementos da base ortogonal $\mathscr{U}_m = \{f_0, f_1, \ldots, f_m\}$ $(m > n)$ de algum subespaço de $\mathscr{B}$. Em suma, usamos o Teorema B.0.2 e propriedades do produto interno para verificar que

(i) $\|f - p_n\| < \delta \Leftrightarrow \|p_n\|^2 > \|f\|^2 - \delta^2$;

(ii) $p_n = \sum_{j=0}^{n} \frac{\langle f_j, f\rangle}{\|f_j\|^2} f_j \Leftrightarrow \|p_n\|^2 = \sum_{j=0}^{n} \frac{\langle f_j, f\rangle^2}{\|f_j\|^2}$;

(iii) $n = \min\left\{k \in \mathbb{N} : \sum_{j=0}^{k} \frac{\langle f_j, f\rangle^2}{\|f_j\|^2} > \|f\|^2 - \delta^2\right\}$.

No nosso caso concreto que consistia em aproximar a parte da função (3.28) dependente de $t = \frac{\nu}{2}$, consideramos $\mathscr{B} = \mathscr{C}[1, 15]$ com a norma dada por $\int_1^{15} w(t)[f(t)]^2 dt$, onde a função peso $w(t) = \frac{e^{-t}}{t}$ foi escolhida por heurística com o intuito de melhorar a aproximação para os menores valores do intervalo $[1, 15]$ (os de maior interesse prático).

A função aproximada pelo método dos mínimos quadrados ponderados foi $f\colon [1, 15] \to \mathbb{R}$ dada por $f(t) = \ln\Big(\ln t - \Psi(t)\Big)$, que é claramente contínua. Buscamos a melhor aproximação nos subespaços de $\mathscr{C}[1, 15]$ da forma $\mathscr{A}_n = \langle 1, \ln, \ldots, \ln^n\rangle$, onde $\ln^n(t) = (\ln t)^n$ $\forall t \in [1, 15]$.

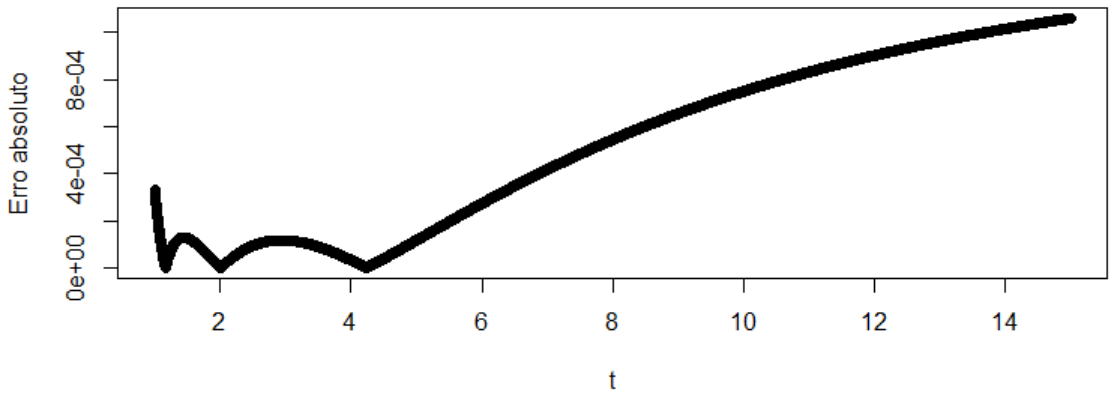
Para obter uma precisão de $\delta = 10^{-4}$ na aproximação, verificamos seguindo os três passos acima indicados que bastava tomar $n = 2$, ou seja, aproximar f por uma função do tipo $g(t) = c_0 + c_1 \ln t + c_2 (\ln t)^2$. Dessa forma, ao resolvermos um sistema da forma indicada em (B.1), obtemos os seguintes resultados: $e^{c_0} \approx 0,5766371$, $c_1 \approx -1,110662$ e $c_2 \approx 0,02637187$.

A justificativa para explicitar e^{c_0} vem do fato de que nosso interesse original era obter uma aproximação para $e^{f(t)} = \ln t - \Psi(t)$. Inicialmente observamos graficamente que a função $\ln t - \Psi(t)$ podia ser aproximada por uma função da forma $k_0 t^{c_1 + c_2 \ln t}$, a qual permitiria obter de maneira explícita uma solução para a equação de interesse.

Entretanto, diante da impossibilidade de obter diretamente pelo método dos mínimos quadrados ponderados uma aproximação para a função desejada, conseguimos aproximar seus logaritmos fazendo $k_0 = e^{c_0}$. Note que a precisão de aproximação obtida para a distância na norma da integral entre as funções $f(t)$ e $g(t)$ não se transmite diretamente para a distância entre $e^{f(t)}$ e $e^{g(t)}$.

Apesar de parecer que a aproximação poderia ser piorada pelo problema mencionado acima, verificamos que o erro absoluto $|e^{f(t)} - e^{g(t)}|$ para $t \in [1, 15]$ ainda fica menor que 10^{-3}, sendo menor para os menores valores de t no intervalo (os de maior interesse prático). Na Figura 41, mostramos um gráfico do erro absoluto da aproximação de $e^{f(t)}$ por $e^{g(t)}$ para 10000 pontos no intervalo $[1, 15]$.

Figura 41 – Erro da aproximação proposta



ANEXO A – Plataforma R

O *software* R, utilizado em todo o trabalho, pode ser baixado gratuitamente no link http://www.r-project.org. O repositório oficial dos pacotes usados no programa é https://cran.r-project.org/. A seguir mostramos alguns comandos e funções construídas no R que foram usados ao longo do texto.

1. Gerar as variáveis explicativas nos modelos de regressão

```
run<-function(n){
X0<-matrix(0,n,q-1)
for(i in 1:n){
X0[i,]=cbind(runif(1,1,10),runif(1,15,20))
}
X=abind(matrix(1,n,1),X0,along=2)
return(X)} #geracao das q-1 covariaveis de regressao univariada para tamanho de amostra n

Run<-function(n){
X0<-array(0,c(p,q-1,n))
for(i in 1:n){
X0[,,i]=cbind(runif(p,1,10),runif(p,-5,0))
}
X=abind(array(1,c(p,1,n)),X0,along=2)
return(X)} #geracao das covariaveis de regressao multivariada para tamanho de amostra n
```

2. Gerar dados de misturas univariadas (normal assimétrica contaminada)

```
mu=25
sigma=4
lambda=-3
nu=0.7
gamma=0.6
P0=c(mu,sigma,lambda,nu,gamma) #parametros normal assimetrica contaminada univariada


rscr<-function(n,P,X){
if(missing(X)) {X<-matrix(1,n,1)}
q=ncol(X)
b=as.matrix(P[1:q])
s=as.numeric(P[q+1])
h=as.numeric(P[q+2])
g=as.numeric(P[q+3])
z=as.numeric(P[q+4])
D=s*h/sqrt(1+h^2)
o=s/sqrt(1+h^2)
y=matrix(0,n,1)
a=rep(0,n)
```

```
for(i in 1:n){
a[i]=runif(1,0,1)
if(a[i]<1-z){
y[i]<-t(X[i,])%*%b+D*abs(rnorm(1))+o*rnorm(1)
}
else
y[i]<-t(X[i,])%*%b+1/sqrt(g)*(D*abs(rnorm(1))+o*rnorm(1))
}
return(y)}
y=rscr(10000,P0) #geracao normal assimetrica contaminada univariada
```

3. Histogramas e gráficos das densidades univariadas (Slash assimétrica)

```
dssr<-function(P,y,X){
n=length(y)
if(missing(X)) {X<-matrix(1,n,1)}
q=ncol(X)
b=as.matrix(P[1:q])
s=as.numeric(P[q+1])
h=as.numeric(P[q+2])
v=as.numeric(P[q+3])
e=y-X%*%b
d=e^2/s^2
aux=pmax(as.vector(h*e/s),-37)
f<-matrix(0,n,1)
for(i in 1:n){
fauxi<-function(u) 2*v/s*u^(v-1)*sqrt(u)*dnorm(sqrt(u*d[i]))*
pnorm(sqrt(u)*aux[i])
f[i]<-integrate(fauxi,0,1)$val}
return(f)} #densidade Slash assimetrica univariada

hist(y,prob=T,xlim=c(5,35),breaks=50,ylim=c(0,0.2),
xlab='Valores observados',ylab='Densidade',
main=paste('Histograma da Slash assimétrica'))
x=y
curve(dssr(P0,x),add=T,col='blue')
```

4. Gerar dados de misturas multivariadas (T-Student normal assimétrica)

```
p=2
mu=c(40,-30)
rSigma=matrix(c(2,-1,-1,1),p,p)
lambda=c(3,-2)
nu=5
P0=c(mu,vech(rSigma),lambda,nu) #parametros t-student normal assimetrica multivariada


rmstr<-function(n,Q,B,h,v,X){
```

```
p=length(h)
if(missing(X)) {X<-matrix(1,n,1)}
y<-matrix(0,n,p)
for(i in 1:n){
u=rgamma(1,shape=v/2,scale=2/v) #fator de escala da t
y[i,]=as.matrix(Q)%*%as.matrix(X[i,])+B%*%((u*(u+sum(h*h)))^(-1/2)
*(h*abs(rnorm(1)))+solve(sqrtm(u*diag(p)+h%*%t(h)))%*%mvrnorm(1,rep(0,p),diag(p)))}
return(y)}
y=rmstr(10000,mu,rSigma,lambda,nu) #geracao t-student normal assimetrica multivariada
```

5. Gráficos das densidades multivariadas (normal contaminada assimétrica):

```
dmscr<-function(P,y,X){
n=nrow(y)
p=ncol(y)
if(missing(X)) {X<-matrix(1,n,1)}
q=ncol(X)
Q=matrix(P[1:(p*q)],p,q)
B=xpnd(P[(p*q+1):(p*q+p*(p+1)/2)])
invB=solve(B)
h=as.matrix(P[(p*q+p*(p+1)/2+1):(length(P)-2)])
g=as.numeric(P[length(P)-1])
z=as.numeric(P[length(P)])
e=y-X%*%t(Q)
#d=diag(e%*%invB%*%invB%*%t(e))
aux=pmax(as.vector(e%*%invB%*%h),-37)
f=2*(z/det(1/sqrt(g)*B)*dmvnorm(sqrt(g)*e%*%invB)+(1-z)/det(B)*dmvnorm(e%*%invB))*pnorm(aux)
return(f)} #densidade normal contaminada assimetrica multivariada

plot3d(y[,1],y[,2],dmscr(P0,y),col='red')
```

ANEXO B – Escore e matriz de informação observada das misturas

Misturas de escala univariadas da família normal assimétrica

1. Escore

$$\underbrace{\frac{\partial \ell_i}{\partial \boldsymbol{\beta}} = \frac{1}{K_i}\left[\frac{e_i}{\sigma^2} I_i^{\Phi}\left(\frac{3}{2},0\right) - \frac{\lambda e_i}{\sigma^2} I_i^{\phi}(1,0)\right]\mathbf{x}_i; \quad \frac{\partial \ell_i}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{K_i}\left[\frac{e_i^2}{\sigma^3} I_i^{\Phi}\left(\frac{3}{2},0\right) - \frac{\lambda e_i}{\sigma^2} I_i^{\phi}(1,0)\right]; \quad \frac{\partial \ell_i}{\partial \lambda} = \frac{e_i}{\sigma K_i} I_i^{\phi}(1,0);}_{\text{Gerais}}$$

$$\underbrace{\frac{\partial \ell_i}{\partial \boldsymbol{\tau}} = \frac{1}{K_i}\mathbf{J}_i^{\Phi}\left(\frac{1}{2}\right)}_{\text{Fator de escala contínuo}} \quad ; \quad \underbrace{\frac{\partial \ell_i}{\partial \gamma} = \frac{1}{2\gamma}\frac{1}{K_i}\left[(1-\gamma d_i) J_{1i}^{\Phi}\left(\frac{1}{2}\right) + A_i J_{1i}^{\phi}(1)\right], \quad \frac{\partial \ell_i}{\partial \nu} = \frac{1}{K_i} J_{2i}^{\Phi}\left(\frac{1}{2}\right)}_{\text{Normal assimétrica contaminada}}.$$

2. Matriz de informação

- Gerais

$$\begin{aligned}
-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} &= \left\{\frac{1}{K_i^2}\left[\frac{e_i^2}{\sigma^4} I_i^{\Phi}\left(\frac{3}{2},0\right)^2 - \frac{2\lambda e_i}{\sigma^3} I_i^{\Phi}\left(\frac{3}{2},0\right) I_i^{\phi}(1,0) + \frac{\lambda^2}{\sigma^2} I_i^{\phi}(1,0)^2\right]\right. \\
&\quad \left. -\frac{1}{K_i}\left[\frac{e_i^2}{\sigma^4} I_i^{\Phi}\left(\frac{5}{2},0\right) - \frac{1}{\sigma^2} I_i^{\Phi}\left(\frac{3}{2},0\right) - \frac{\lambda(2+\lambda^2)e_i}{\sigma^3} I_i^{\phi}(2,0)\right]\right\}\mathbf{x}_i\mathbf{x}_i^T; \\
-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \sigma} &= \left\{\frac{1}{K_i^2}\left[\frac{e_i^3}{\sigma^5} I_i^{\Phi}\left(\frac{3}{2},0\right)^2 - \frac{2\lambda e_i^2}{\sigma^4} I_i^{\Phi}\left(\frac{3}{2},0\right) I_i^{\phi}(1,0) + \frac{\lambda^2 e_i}{\sigma^3} I_i^{\phi}(1,0)^2\right]\right. \\
&\quad \left. -\frac{1}{K_i}\left[\frac{e_i^3}{\sigma^5} I_i^{\Phi}\left(\frac{5}{2},0\right) - \frac{2e_i}{\sigma^3} I_i^{\Phi}\left(\frac{3}{2},0\right) + \frac{\lambda}{\sigma^2} I_i^{\phi}(1,0) - \frac{\lambda(2+\lambda^2)e_i^2}{\sigma^4} I_i^{\phi}(2,0)\right]\right\}\mathbf{x}_i; \\
-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \lambda} &= \left\{\frac{1}{K_i^2}\left[\frac{e_i^2}{\sigma^3} I_i^{\Phi}\left(\frac{3}{2},0\right) I_i^{\phi}(1,0) - \frac{\lambda e_i}{\sigma^2} I_i^{\phi}(1,0)^2\right] - \frac{1}{K_i}\left[\frac{(1+\lambda^2)e_i^2}{\sigma^3} I_i^{\phi}(2,0) - \frac{1}{\sigma} I_i^{\phi}(1,0)\right]\right\}\mathbf{x}_i; \\
-\frac{\partial^2 \ell_i}{\partial \sigma^2} &= -\frac{1}{\sigma^2} + \frac{1}{K_i^2}\left[\frac{e_i^4}{\sigma^6} I_i^{\Phi}\left(\frac{3}{2},0\right)^2 - \frac{2\lambda e_i^3}{\sigma^5} I_i^{\Phi}\left(\frac{3}{2},0\right) I_i^{\phi}(1,0) + \frac{\lambda^2 e_i^2}{\sigma^4} I_i^{\phi}(1,0)^2\right] \\
&\quad - \frac{1}{K_i}\left[\frac{e_i^4}{\sigma^6} I_i^{\Phi}\left(\frac{5}{2},0\right) - \frac{3e_i^2}{\sigma^4} I_i^{\Phi}\left(\frac{3}{2},0\right) + \frac{2\lambda e_i}{\sigma^3} I_i^{\phi}(1,0) - \frac{\lambda(2+\lambda^2)e_i^3}{\sigma^5} I_i^{\phi}(2,0)\right]; \\
-\frac{\partial^2 \ell_i}{\partial \sigma \partial \lambda} &= \frac{1}{K_i^2}\left[\frac{e_i^3}{\sigma^4} I_i^{\Phi}\left(\frac{3}{2},0\right) I_i^{\phi}(1,0) - \frac{\lambda e_i^2}{\sigma^3} I_i^{\phi}(1,0)^2\right] - \frac{1}{K_i}\left[\frac{(1+\lambda^2)e_i^3}{\sigma^4} I_i^{\phi}(2,0) - \frac{e_i}{\sigma^2} I_i^{\phi}(1,0)\right]; \\
-\frac{\partial^2 \ell_i}{\partial \lambda^2} &= \frac{1}{K_i^2}\frac{e_i^2}{\sigma^2} I_i^{\phi}(1,0)^2 + \frac{1}{K_i}\frac{\lambda e_i^3}{\sigma^3} I_i^{\phi}(2,0).
\end{aligned}$$

- Fator de escala contínuo

$$\begin{aligned}
-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\tau}^T} &= \mathbf{x}_i\left\{\frac{1}{K_i^2}\left[\frac{e_i}{\sigma^2} I_i^{\Phi}\left(\frac{3}{2},0\right) - \frac{\lambda}{\sigma} I_i^{\phi}(1,0)\right]\mathbf{J}_i^{\Phi}\left(\frac{1}{2}\right)^T - \frac{1}{K_i}\left[\frac{e_i}{\sigma^2}\mathbf{J}_i^{\Phi}\left(\frac{3}{2}\right) - \frac{\lambda}{\sigma}\mathbf{J}_i^{\phi}(1)\right]^T\right\}; \\
-\frac{\partial^2 \ell_i}{\partial \sigma \partial \boldsymbol{\tau}^T} &= \left\{\frac{1}{K_i^2}\left[\frac{e_i^2}{\sigma^3} I_i^{\Phi}\left(\frac{3}{2},0\right) - \frac{\lambda e_i}{\sigma^2} I_i^{\phi}(1,0)\right]\mathbf{J}_i^{\Phi}\left(\frac{1}{2}\right) - \frac{1}{K_i}\left[\frac{e_i^2}{\sigma^3}\mathbf{J}_i^{\Phi}\left(\frac{3}{2}\right) - \frac{\lambda e_i}{\sigma^2}\mathbf{J}_i^{\phi}(1)\right]\right\}^T; \\
-\frac{\partial^2 \ell_i}{\partial \lambda \partial \boldsymbol{\tau}^T} &= \left\{\frac{1}{K_i^2}\frac{e_i}{\sigma} I_i^{\phi}(1,0)\mathbf{J}_i^{\Phi}\left(\frac{1}{2}\right) - \frac{1}{K_i}\frac{e_i}{\sigma}\mathbf{J}_i^{\phi}(1)\right\}^T; \quad -\frac{\partial^2 \ell_i}{\partial \boldsymbol{\tau}\partial \boldsymbol{\tau}^T} = \left\{\frac{1}{K_i^2}\mathbf{J}_i^{\Phi}\left(\frac{1}{2}\right)\mathbf{J}_i^{\Phi}\left(\frac{1}{2}\right)^T - \frac{1}{K_i}\mathbf{L}_i^{\Phi}\left(\frac{1}{2}\right)\right\}.
\end{aligned}$$

- Normal assimétrica contaminada

$$\begin{aligned}
-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \gamma} &= \left\{\frac{1}{2\gamma K_i^2}\left[\frac{(1-\gamma d_i)e_i}{\sigma^2} I_i^{\Phi}\left(\frac{3}{2},0\right) J_{1i}^{\Phi}\left(\frac{1}{2}\right) + \frac{\lambda e_i^2}{\sigma^3} I_i^{\Phi}\left(\frac{3}{2},0\right) J_{1i}^{\phi}(1) - \frac{(1-\gamma d_i)\lambda}{\sigma} I_i^{\phi}(1,0) J_{1i}^{\Phi}\left(\frac{1}{2}\right) - \frac{\lambda^2 e_i}{\sigma^2} I_i^{\phi}(1,0) J_{1i}^{\phi}(1)\right]\right. \\
&\quad \left. -\frac{1}{K_i}\left[\frac{(3-\gamma d_i)e_i}{\sigma^2} J_{1i}^{\Phi}\left(\frac{1}{2}\right) + \frac{\gamma A_i e_i - \sigma\lambda[2-\gamma(d_i+A_i^2)]}{2\sigma^2\gamma} J_{1i}^{\phi}(1)\right]\right\}\mathbf{x}_i; \\
-\frac{\partial^2 \ell_i}{\partial \sigma \partial \gamma} &= \frac{1}{2\gamma K_i^2}\left[\frac{(1-\gamma d_i)e_i^2}{2\sigma^3} I_i^{\Phi}\left(\frac{3}{2},0\right) J_{1i}^{\Phi}\left(\frac{1}{2}\right) + \frac{\lambda e_i^3}{\sigma^4} I_i^{\Phi}\left(\frac{3}{2},0\right) J_{1i}^{\phi}(1) - \frac{(1-\gamma d_i)\lambda e_i}{\sigma^2} I_i^{\phi}(1,0) J_{1i}^{\Phi}\left(\frac{1}{2}\right) - \frac{\lambda^2 e_i^2}{\sigma^3} I_i^{\phi}(1,0) J_{1i}^{\phi}(1)\right] \\
&\quad - \frac{1}{K_i}\left[\frac{(3-\gamma d_i)e_i^2}{2\sigma^3} J_{1i}^{\Phi}\left(\frac{1}{2}\right) + \frac{\gamma A_i e_i^2 - \sigma\lambda e_i[2-\gamma(d_i+A_i^2)]}{2\sigma^3\gamma} J_{1i}^{\phi}(1)\right]; \\
-\frac{\partial^2 \ell_i}{\partial \lambda \partial \gamma} &= \frac{e_i}{2\sigma\gamma K_i^2}\left[(1-\gamma d_i) J_{1i}^{\Phi}\left(\frac{1}{2}\right) + A_i J_{1i}^{\phi}(1)\right] I_i^{\phi}(1,0) - \frac{1}{K_i}\frac{2-\gamma(d_i+A_i^2)}{2\sigma\gamma} e_i J_{1i}^{\phi}(1);
\end{aligned}$$

$$-\frac{\partial^2 \ell_i}{\partial \gamma^2} = \frac{1}{4\gamma^2 K_i^2}\left[(1-\gamma d_i^2)J_{1i}^{\Phi}\left(\tfrac{1}{2}\right)^2 + 2(1-\gamma d_i^2)A_i J_{1i}^{\Phi}\left(\tfrac{1}{2}\right) J_{1i}^{\phi}(1) + A_i^2 J_{1i}^{\phi}(1)^2\right]$$
$$- \frac{1}{4\gamma^2 K_i}\left[\Big((1-\gamma d_i)^2 - 2\Big)J_{1i}^{\Phi}\left(\tfrac{1}{2}\right) + A_i\Big(1-\gamma(2d_i + A_i^2)\Big)J_{1i}^{\phi}(1)\right];$$
$$-\frac{\partial^2 \ell_i}{\partial \gamma \partial \nu} = \frac{1}{2\gamma K_i^2}\left[(1-\gamma d_i)J_{1i}^{\Phi}\left(\tfrac{1}{2}\right) + A_i J_{1i}^{\phi}(1)\right] J_{2i}^{\Phi}\left(\tfrac{1}{2}\right) - \frac{1}{2\nu\gamma K_i}\left[(1-\gamma d_i)J_{1i}^{\Phi}\left(\tfrac{1}{2}\right) + A_i J_{1i}^{\phi}(1)\right];$$
$$-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta} \partial \nu} = \left\{\frac{1}{K_i^2}\left[\frac{e_i}{\sigma^2} I_i^{\Phi}\left(\tfrac{3}{2}, 0\right) - \frac{\lambda}{\sigma} I_i^{\phi}(1,0)\right] J_{2i}^{\Phi}\left(\tfrac{1}{2}\right) - \frac{1}{K_i}\left[\frac{e_i}{\sigma^2} J_{2i}^{\Phi}\left(\tfrac{3}{2}\right) - \frac{\lambda}{\sigma} J_{2i}^{\phi}(1)\right]\right\}\mathbf{x}_i;$$
$$-\frac{\partial^2 \ell_i}{\partial \sigma \partial \nu} = \frac{1}{K_i^2}\left[\frac{e_i^2}{\sigma^3} I_i^{\Phi}\left(\tfrac{3}{2}, 0\right) - \frac{\lambda e_i}{\sigma^2} I_i^{\phi}(1,0)\right] J_{2i}^{\Phi}\left(\tfrac{1}{2}\right) - \frac{1}{K_i}\left[\frac{e_i^2}{\sigma^3} J_{2i}^{\Phi}\left(\tfrac{3}{2}\right) - \frac{\lambda e_i}{\sigma^2} J_{2i}^{\phi}(1)\right];$$
$$-\frac{\partial^2 \ell_i}{\partial \lambda \partial \nu} = \frac{1}{K_i^2}\frac{e_i}{\sigma} I_i^{\phi}(1,0) J_{2i}^{\Phi}\left(\tfrac{1}{2}\right) - \frac{1}{K_i}\frac{e_i}{\sigma} J_{2i}^{\phi}(1); \quad -\frac{\partial^2 \ell_i}{\partial \nu^2} = \frac{1}{K_i^2} J_{2i}^{\Phi}\left(\tfrac{1}{2}\right)^2.$$

Misturas de escala assimétricas multivariadas da família normal

Inicialmente, vamos indicar algumas derivadas auxiliares adotando a notação $\dot{\mathbf{B}}_j = \dfrac{\partial \mathbf{B}(\boldsymbol{\alpha})}{\partial \alpha_j}$ para cada $j \in \left\{1, \ldots, p_0 = \frac{p(p+1)}{2}\right\}$, onde $\mathbf{B}(\boldsymbol{\alpha}) = \boldsymbol{\Sigma}^{\frac{1}{2}}$.

$\frac{\partial \Lambda}{\partial \boldsymbol{\beta}} = \mathbf{0}$; $\frac{\partial \Lambda}{\partial \alpha_j} = 2\mathrm{tr}(\mathbf{B}^{-1}\dot{\mathbf{B}}_j)$; $\frac{\partial \Lambda}{\partial \boldsymbol{\lambda}} = \mathbf{0}$; $\frac{\partial \Lambda}{\partial \boldsymbol{\tau}} = \mathbf{0}$;
$\frac{\partial d_i}{\partial \boldsymbol{\beta}} = -2\mathbf{X}_i^T\mathbf{B}^{-2}\mathbf{e}_i$; $\frac{\partial d_i}{\partial \alpha_j} = -\mathbf{e}_i^T\mathbf{B}^{-1}(\dot{\mathbf{B}}_j\mathbf{B}^{-1} + \mathbf{B}^{-1}\dot{\mathbf{B}}_j)\mathbf{B}^{-1}\mathbf{e}_i$; $\frac{\partial d_i}{\partial \boldsymbol{\lambda}} = \mathbf{0}$; $\frac{\partial d_i}{\partial \boldsymbol{\tau}} = \mathbf{0}$;
$\frac{\partial A_i}{\partial \boldsymbol{\beta}} = -\mathbf{X}_i^T\mathbf{B}^{-1}\boldsymbol{\lambda}$; $\frac{\partial A_i}{\partial \alpha_j} = -\boldsymbol{\lambda}^T\mathbf{B}^{-1}\dot{\mathbf{B}}_j\mathbf{B}^{-1}\mathbf{e}_i$; $\frac{\partial A_i}{\partial \boldsymbol{\lambda}} = \mathbf{B}^{-1}\mathbf{e}_i$; $\frac{\partial A_i}{\partial \boldsymbol{\tau}} = \mathbf{0}$;
$\frac{\partial^2 \Lambda}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\partial^2 \Lambda}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 \end{bmatrix}$; $\frac{\partial^2 \Lambda}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} = \left[\frac{\partial^2 \Lambda}{\partial \alpha_j \partial \alpha_k}\right]_{j,k=1,\ldots,p_0} = \left[-\mathrm{tr}(\mathbf{B}^{-1}\dot{\mathbf{B}}_k\mathbf{B}^{-1}\dot{\mathbf{B}}_j)\right]_{j,k=1,\ldots,p_0}$;
$\frac{\partial^2 d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = 2\mathbf{X}_i^T\mathbf{B}^{-2}\mathbf{X}_i$; $\frac{\partial^2 d_i}{\partial \boldsymbol{\beta}\partial \alpha_j} = 2\mathbf{X}_i^T\mathbf{B}^{-1}(\dot{\mathbf{B}}_j\mathbf{B}^{-1} + \mathbf{B}^{-1}\dot{\mathbf{B}}_j)\mathbf{B}^{-1}\mathbf{e}_i$, $\frac{\partial^2 d_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\lambda}^T} = \mathbf{0}$; $\frac{\partial^2 d_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\tau}^T} = \mathbf{0}$;
$\frac{\partial^2 d_i}{\partial \alpha_j \partial \alpha_k} = \mathbf{e}_i^T\mathbf{B}^{-1}(\dot{\mathbf{B}}_k\mathbf{B}^{-1}\dot{\mathbf{B}}_j\mathbf{B}^{-1} + \dot{\mathbf{B}}_j\mathbf{B}^{-1}\dot{\mathbf{B}}_k\mathbf{B}^{-1} + \dot{\mathbf{B}}_j\mathbf{B}^{-2}\dot{\mathbf{B}}_k + \dot{\mathbf{B}}_k\mathbf{B}^{-2}\dot{\mathbf{B}}_j + \mathbf{B}^{-1}\dot{\mathbf{B}}_k\mathbf{B}^{-1}\dot{\mathbf{B}}_j + \mathbf{B}^{-1}\dot{\mathbf{B}}_j\mathbf{B}^{-1}\dot{\mathbf{B}}_k)\mathbf{B}^{-1}\mathbf{e}_i$;
$\frac{\partial^2 A_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \mathbf{0}$; $\frac{\partial^2 A_i}{\partial \boldsymbol{\beta}\partial \alpha_j} = \mathbf{X}_i^T\mathbf{B}^{-1}\dot{\mathbf{B}}_j\mathbf{B}^{-1}\boldsymbol{\lambda}$; $\frac{\partial^2 A_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\lambda}^T} = -\mathbf{X}_i^T\mathbf{B}^{-1}$, $\frac{\partial^2 A_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\tau}^T} = \mathbf{0}$;
$\frac{\partial^2 A_i}{\partial \alpha_j \partial \alpha_k} = -\boldsymbol{\lambda}^T\mathbf{B}^{-1}(\dot{\mathbf{B}}_k\mathbf{B}^{-1}\dot{\mathbf{B}}_j + \dot{\mathbf{B}}_j\mathbf{B}^{-1}\dot{\mathbf{B}}_j)\mathbf{B}^{-1}\mathbf{e}_i$; $\frac{\partial^2 A_i}{\partial \alpha_j \partial \boldsymbol{\lambda}^T} = -\mathbf{e}_i^T\mathbf{B}^{-1}\dot{\mathbf{B}}_j\mathbf{B}^{-1}$; $\frac{\partial^2 A_i}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}^T} = \mathbf{0}$.

Com as notações $EU_{\mathbf{y}_i} = E(U|\mathbf{Y}_i = \mathbf{y}_i)$ e $VU_{\mathbf{y}_i} = Var(U|\mathbf{Y}_i = \mathbf{y}_i)$, obtemos

1. Escore

   - Gerais

     $\frac{\partial \ell_i}{\partial \boldsymbol{\beta}} = -\frac{1}{2}EU_{\mathbf{y}_i}\frac{\partial d_i}{\partial \boldsymbol{\beta}} + W_{\Phi}(A_i)\frac{\partial A_i}{\partial \boldsymbol{\beta}}$; $\frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} = -\frac{1}{2}\frac{\partial \Lambda}{\partial \boldsymbol{\alpha}} - \frac{1}{2}EU_{\mathbf{y}_i}\frac{\partial d_i}{\partial \boldsymbol{\alpha}} + W_{\Phi}(A_i)\frac{\partial A_i}{\partial \boldsymbol{\alpha}}$;
     $\frac{\partial \ell_i}{\partial \boldsymbol{\lambda}} = W_{\Phi}(A_i)\frac{\partial A_i}{\partial \boldsymbol{\lambda}}$; $\frac{\partial \ell_i}{\partial \boldsymbol{\tau}} = \frac{1}{K_i}\frac{\partial K_i}{\partial \boldsymbol{\tau}}$.

   - Específicos

     $\underbrace{\frac{\partial \ell_i}{\partial \nu} = \frac{1}{2}\left[\ln\left(\frac{\nu}{\nu + d_i}\right) + \Psi\left(\frac{\nu+p}{2}\right) - \Psi\left(\frac{\nu}{2}\right) + \frac{d_i - p}{\nu + d_i}\right]}_{\text{T-Student normal assimétrica}}$; $\underbrace{\frac{\partial \ell_i}{\partial \nu} = \frac{1}{\nu} + \frac{\int_0^1 u^{\nu+p/2-1}\ln(u)e^{ud_i/2}du}{\int_0^1 u^{\nu+p/2-1}e^{ud_i/2}du}}_{\text{Slash normal assimétrica}}$;

     $\underbrace{\frac{\partial \ell_i}{\partial \nu} = \frac{\gamma^{p/2}e^{-\gamma d_i/2} - e^{-d_i/2}}{\nu\gamma^{p/2}e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}}, \quad \frac{\partial \ell_i}{\partial \gamma} = \frac{\nu}{2}\left[\frac{(p-\gamma d_i)\left(\gamma^{p/2-1}e^{-\gamma d_i/2} - e^{-d_i/2}\right)}{\nu\gamma^{p/2}e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}}\right]}_{\text{Normal contaminada assimétrica}}$.

2. Matriz de informação

- Gerais

$$-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \frac{1}{2} EU_{\mathbf{y}_i} \frac{\partial d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} - \frac{1}{4} VU_{\mathbf{y}_i} \frac{\partial d_i}{\partial \boldsymbol{\beta}} \frac{\partial d_i}{\partial \boldsymbol{\beta}^T} - W'_\Phi(A_i) \frac{\partial A_i}{\partial \boldsymbol{\beta}} \frac{\partial A_i}{\partial \boldsymbol{\beta}^T};$$
$$-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} = \frac{1}{2} EU_{\mathbf{y}_i} \frac{\partial d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} - \frac{1}{4} VU_{\mathbf{y}_i} \frac{\partial d_i}{\partial \boldsymbol{\beta}} \frac{\partial d_i}{\partial \boldsymbol{\alpha}^T} - W_\Phi(A_i) \frac{\partial^2 A_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} - W'_\Phi(A_i) \frac{\partial A_i}{\partial \boldsymbol{\beta}} \frac{\partial A_i}{\partial \boldsymbol{\alpha}^T};$$
$$-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} = \frac{1}{2} \frac{\partial^2 \Lambda}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} + \frac{1}{2} EU_{\mathbf{y}_i} \frac{\partial d_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} - \frac{1}{4} VU_{\mathbf{y}_i} \frac{\partial d_i}{\partial \boldsymbol{\alpha}} \frac{\partial d_i}{\partial \boldsymbol{\alpha}^T} - W_\Phi(A_i) \frac{\partial^2 A_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} - W'_\Phi(A_i) \frac{\partial A_i}{\partial \boldsymbol{\alpha}} \frac{\partial A_i}{\partial \boldsymbol{\alpha}^T};$$
$$-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\lambda}^T} = -W_\Phi(A_i) \frac{\partial^2 A_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\lambda}^T} - W'_\Phi(A_i) \frac{\partial A_i}{\partial \boldsymbol{\beta}} \frac{\partial A_i}{\partial \boldsymbol{\lambda}^T}; \; -\frac{\partial^2 \ell_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\lambda}^T} = -W_\Phi(A_i) \frac{\partial^2 A_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\lambda}^T} - W'_\Phi(A_i) \frac{\partial A_i}{\partial \boldsymbol{\beta}} \frac{\partial A_i}{\partial \boldsymbol{\lambda}^T};$$
$$-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}^T} = -W'_\Phi(A_i) \frac{\partial A_i}{\partial \boldsymbol{\lambda}} \frac{\partial A_i}{\partial \boldsymbol{\lambda}^T}; \; -\frac{\partial^2 \ell_i}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\tau}^T} = \mathbf{0}; \; -\frac{\partial^2 \ell_i}{\partial \boldsymbol{\tau}\partial \boldsymbol{\tau}^T} = -\frac{1}{K_i} \frac{\partial^2 K_i}{\partial \boldsymbol{\tau}\partial \boldsymbol{\tau}^T} + \frac{1}{K_i^2} \frac{\partial K_i}{\partial \boldsymbol{\tau}} \frac{\partial K_i}{\partial \boldsymbol{\tau}^T};$$
$$-\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\tau}^T} = -\frac{1}{K_i} \frac{\partial^2 K_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\tau}^T} + \frac{1}{K_i^2} EU_{\mathbf{y}_i} \frac{\partial d_i}{\partial \boldsymbol{\beta}} \frac{\partial K_i}{\partial \boldsymbol{\tau}^T}; \; -\frac{\partial^2 \ell_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\tau}^T} = -\frac{1}{K_i} \frac{\partial^2 K_i}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\tau}^T} + \frac{1}{K_i^2} EU_{\mathbf{y}_i} \frac{\partial d_i}{\partial \boldsymbol{\alpha}} \frac{\partial K_i}{\partial \boldsymbol{\tau}^T}.$$

- Específicos

  * T-Student normal assimétrica:

$$\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \nu} = \frac{1}{2} \frac{p-d_i}{(\nu+d_i)^2} \frac{\partial d_i}{\partial \boldsymbol{\beta}}; \quad \frac{\partial^2 \ell_i}{\partial \boldsymbol{\alpha}\partial \nu} = \frac{1}{2} \frac{p-d_i}{(\nu+d_i)^2} \frac{\partial d_i}{\partial \boldsymbol{\alpha}}; \quad \frac{\partial^2 \ell_i}{\partial \nu^2} = \frac{1}{4} \left[\Psi_1\left(\frac{\nu+p}{2}\right) - \Psi_1\left(\frac{\nu}{2}\right) + 2\frac{d_i+p\nu}{\nu(\nu+d_i)^2} \right].$$

  * Slash normal assimétrica:

$$\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \nu} = -\frac{1}{2} \left[\frac{I(0,1)}{I(1,0)} - \frac{I(1,1)I(0,0)}{I(1,0)^2} \right] \frac{\partial d_i}{\partial \boldsymbol{\beta}}; \quad \frac{\partial^2 \ell_i}{\partial \boldsymbol{\alpha}\partial \nu} = \frac{1}{2} \left[\frac{I(0,1)}{I(1,0)} - \frac{I(1,1)I(0,0)}{I(1,0)^2} \right] \frac{\partial d_i}{\partial \boldsymbol{\alpha}};$$
$$\frac{\partial^2 \ell_i}{\partial \nu^2} = \frac{I(1,2)}{I(1,0)} - \left[\frac{I(1,1)}{I(1,0)}\right]^2 - \frac{1}{\nu^2}; \quad I(r,s) = \int_0^1 u^{\nu+p/2-r} (\ln u)^s e^{-ud_i/2} du.$$

  * Normal contaminada assimétrica

$$\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \nu} = \frac{1}{2} \left[\frac{(1-\gamma)\gamma^{p/2} e^{-(\gamma+1)d_i/2}}{\left(\nu\gamma^{p/2} e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}\right)^2} \right] \frac{\partial d_i}{\partial \boldsymbol{\beta}}; \quad \frac{\partial^2 \ell_i}{\partial \boldsymbol{\alpha}\partial \nu} = \frac{1}{2} \left[\frac{(1-\gamma)\gamma^{p/2} e^{-(\gamma+1)d_i/2}}{\left(\nu\gamma^{p/2} e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}\right)^2} \right] \frac{\partial d_i}{\partial \boldsymbol{\alpha}};$$
$$\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}\partial \gamma} = \frac{\nu}{4} \left\{ \frac{(1-\nu)[p-\gamma(2+p+d_i)+\gamma^2 d_i]\gamma^{p/2-1} e^{-(\gamma+1)d_i/2} - 2\nu\gamma^p e^{-\gamma d_i}}{\left(\nu\gamma^{p/2} e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}\right)^2} \right\} \frac{\partial d_i}{\partial \boldsymbol{\beta}};$$
$$\frac{\partial^2 \ell_i}{\partial \boldsymbol{\alpha}\partial \gamma} = \frac{\nu}{4} \left\{ \frac{(1-\nu)[p-\gamma(2+p+d_i)+\gamma^2 d_i]\gamma^{p/2-1} e^{-(\gamma+1)d_i/2} - 2\nu\gamma^p e^{-\gamma d_i}}{\left(\nu\gamma^{p/2} e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}\right)^2} \right\} \frac{\partial d_i}{\partial \boldsymbol{\alpha}};$$
$$\frac{\partial^2 \ell_i}{\partial \nu^2} = -\frac{\left(\gamma^{p/2} e^{-\gamma d_i/2} - e^{-d_i/2}\right)^2}{\left(\nu\gamma^{p/2} e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}\right)^2}; \quad \frac{\partial^2 \ell_i}{\partial \nu \partial \gamma} = \frac{1}{2} \frac{(p-\gamma d_i)\gamma^{p/2-1} e^{-(\gamma+1)d_i/2}}{\left(\nu\gamma^{p/2} e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}\right)^2};$$
$$\frac{\partial^2 \ell_i}{\partial \gamma^2} = \frac{\nu}{4} \left\{ \frac{(1-\nu)[(p-\gamma d_i)^2 - 2p]\gamma^{p/2-2} e^{-(\gamma+1)d_i/2} - 2p\nu\gamma^{p-2} e^{-\gamma d_i}}{\left(\nu\gamma^{p/2} e^{-\gamma d_i/2} + (1-\nu)e^{-d_i/2}\right)^2} \right\}.$$