



## Convergência do Algoritmo EM para Estimação em Modelos Lineares Mistos: Enfrentando o Desafio das Distribuições Assimétricas e com Caudas Pesadas

**Palavras-chave:** Valores iniciais, modelos mistos, assimetria

**Autores:**

Sara dos Anjos de Souza Dias, IMECC - UNICAMP

Larissa Avila Matos, IMECC - UNICAMP

## Introdução

O uso de modelos lineares mistos para a análise de dados provenientes de medidas repetidas tem se tornado um tema cada vez mais presente em discussões acadêmicas, surgindo em diversas áreas do conhecimento, como Economia, Medicina, Psicologia e Educação, entre outras. Dados de medidas repetidas ocorrem quando cada unidade experimental é observada repetidamente sob diferentes condições experimentais. Esse tipo de dado geralmente apresenta certo grau de correlação entre as observações da mesma unidade experimental. Nosso interesse especial recai sobre estudos longitudinais, onde os dados são coletados ao longo do tempo ou em qualquer escala ordenada. Para uma discussão mais detalhada sobre dados longitudinais, veja, por exemplo, [11], [4], e [13].

Para o estudo de dados longitudinais, modelos lineares mistos gaussianos são os mais comumente utilizados para análise, não apenas por sua simplicidade, mas também porque servem como aproximações para modelos de outras classes e apresentam facilidade de implementação computacional ([12]). No entanto, assumir que os efeitos aleatórios e erros seguem distribuições normais pode levar a inferências estatísticas inválidas, especialmente quando os dados apresentam caudas pesadas e assimetria ([5, 9]).

Recentemente, [9] desenvolveram uma abordagem de verossimilhança, via algoritmo EM [3], para modelos lineares mistos de mistura de escalas de normais assimétricas (SMSN-LMM), considerando algumas estruturas de correlação, como correlação exponencial amortecida ([8]) e correlação autorregressiva de ordem  $p$  [1]. A classe de distribuições SMSN foi introduzida por [2] e cobre distribuições como normal assimétrica,  $t$ -assimétrica, slash-assimétrica (SSL) e normal contaminada assimétrica (SCN), além de ter a classe SNI como um caso especial.

Os modelos propostos por [9] levam em conta assimetria, caudas pesadas e dependência serial, o que pode ser necessário para modelar mais apropriadamente dados observados ao longo do tempo. Para possibilitar o uso desses modelos na análise de dados da vida real, os métodos para estimação e avaliação do modelo foram codificados e estão disponíveis no pacote R: `skewlmm` [10]. A escolha de valores iniciais apropriados para os parâmetros é crucial para facilitar a convergência do algoritmo EM nos modelos estudados neste projeto.

Neste relatório, temos como objetivo descrever os procedimentos realizados da pesquisa sobre simulações utilizando a distribuição  $t$ -student com assimetria positiva e normal com assimetria positiva e negativa, focando na investigação do impacto e importância dos valores iniciais.

## Metodologia

### Modelo Linear Misto (MLM)

Foi realizada uma série de simulações com base no Modelo Linear Misto, que foi a estrutura central de todas as análises, por se tratar do objeto principal de estudo. As simulações foram estruturadas por meio do pacote `skewlmm`, a partir de diferentes configurações. O modelo considerado assume que  $\mathbf{Y}_i$  é um vetor de respostas contínuas de dimensão  $n_i \times 1$ , observado na  $i$ -ésima unidade experimental, com  $i = 1, \dots, n$ . A formulação geral do modelo linear misto, conforme apresentada por Laird and Ware [7], é dada por:

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

onde  $\boldsymbol{\beta}$  é um vetor de  $l \times 1$  de coeficientes de regressão médios da população chamados efeitos fixos,  $\mathbf{X}_i$  de dimensão  $n_i \times l$  é a matriz de desenho correspondente aos efeitos fixos,  $\mathbf{Z}_i$  de dimensão  $q \times q$  é a matriz de

desenho correspondente ao vetor de  $q \times 1$  de variáveis latentes  $\mathbf{b}_i$  (comumente chamadas de efeitos aleatórios) e  $\epsilon_i$  é um vetor de dimensão  $n_i \times 1$  de erros aleatórios.

Geralmente, assume-se que os efeitos aleatórios  $\mathbf{b}_i$  e os erros aleatórios  $\epsilon_i$  são independentes, com  $\mathbf{b}_i \stackrel{\text{i.i.d.}}{\sim} N_q(\mathbf{0}, \mathbf{D})$  e  $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} N_{n_i}(\mathbf{0}, \Sigma_i)$ , onde  $\mathbf{D}$  é uma matriz de covariância de dimensão  $q \times q$  e pode ser estruturada ou não estruturada, e  $\Sigma_i$  é uma matriz de covariância de dimensão  $n_i \times n_i$ . Além disso,  $\Sigma_i$  é comumente escrita em termos de um parâmetro de dispersão como  $\sigma_\epsilon^2 \mathbf{R}_i$ , com  $\mathbf{R}_i$  sendo uma matriz  $(p \times 1)$  conhecida ou uma matriz estruturada dependendo de um vetor de parâmetros  $\phi$ .

Seguindo [9], o SMSN-LMM é definido assumindo

$$\begin{pmatrix} \mathbf{b}_i \\ \epsilon_i \end{pmatrix} \stackrel{\text{i.i.d.}}{\sim} \text{SMSN}_{q+n_i} + n_i \left( \begin{pmatrix} c\mathbf{\Delta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \Sigma_i \end{pmatrix}, \begin{pmatrix} \lambda \\ \mathbf{0} \end{pmatrix}; H \right), \quad i = 1, \dots, n;$$

e, pelo Lema 1 do Apêndice A em [6], temos marginalmente

$$\mathbf{b}_i \stackrel{\text{i.i.d.}}{\sim} \text{SMSN}_q(c\mathbf{\Delta}, \mathbf{D}, \lambda; H) \quad \text{e} \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \text{SMN}_{n_i}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{R}_i; H), \quad i = 1, \dots, n, \quad (2)$$

onde  $c = c(\nu) = -\sqrt{\frac{2}{\pi}} k_1$ , com  $k_1 = \mathbb{E}(U^{-1/2})$ ,  $\mathbf{\Delta} = \mathbf{D}^{1/2} \boldsymbol{\delta}$ ,  $\boldsymbol{\delta} = \frac{\lambda}{\sqrt{1+\lambda^\top \lambda}}$ ,  $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$  depende do vetor de parâmetros desconhecidos e reduzidos  $\boldsymbol{\alpha}$ , e  $\Sigma_i = \sigma_\epsilon^2 \mathbf{R}_i$ , com  $\mathbf{R}_i(\phi)$  sendo uma matriz estruturada dependendo de  $\phi = (\phi_1, \dots, \phi_p)^\top$ . Para mais detalhes, veja [9].

## Resultados e Discussão

### Simulação 1

Para investigar o impacto dos valores iniciais na estimação, começamos com o Modelo Linear Misto somente com o intercepto aleatório. Realizamos um estudo de simulação com dois cenários: no Modelo 1, os valores iniciais foram estimados automaticamente pelo pacote `skewlmm` e no Modelo 2, foram dados os valores reais do conjunto simulado. Foram geradas 500 amostras, de Monte Carlo, a partir de um Modelo Linear Misto com distribuição normal assimétrica (positiva e negativa) e t-student assimétrica positiva, permitindo avaliar o desempenho da estimação em diferentes contextos. Especificamente, os dados simulados para ambos os modelos foram gerados a partir do seguinte modelo:

$$\mathbf{Y}_i = (\beta_0 + b_i) \mathbf{1}_k + \beta_1 \mathbf{x}_i + \varepsilon_i, \quad i = 1, \dots, n,$$

onde  $\mathbf{1}_k$  é um vetor de uns de comprimento  $k$  e  $k = x_i = (1, 2, \dots, 7)^T$ . Para o caso normal, assumimos que  $b_i \sim \text{SN}_1\left(-\sqrt{\frac{2\alpha\lambda^2}{\pi(1+\lambda^2)}}, \alpha, \lambda\right)$  e  $\varepsilon_i \sim \mathcal{N}_7(0, \sigma^2 \mathbf{R}_i)$ . Consideramos também uma estrutura de dependência temporal modelada por um processo autorregressivo de ordem 1 (AR(1)), incorporado por meio da matriz  $\mathbf{R}_i$  através do parâmetro  $\phi$ . Por fim, repetimos as simulações para diferentes tamanhos amostrais, considerando  $n = 50, 100, 300, 500$  e  $1000$ .

Iniciamos o estudo de simulação pela distribuição normal assimétrica positiva, considerando como valores dos parâmetros aqueles apresentados na Tabela 1.

| Parâmetro  | $\beta$       | $\sigma^2$ | $\alpha$ | $\lambda$ | $\phi$ |
|------------|---------------|------------|----------|-----------|--------|
| Valor real | $(1, 2)^\top$ | 0,25       | 1        | 3         | 0,6    |

Tabela 1: Simulação 1 - Valores verdadeiros adotados para cada parâmetro do estudo.

A Figura 1 apresenta uma versão ampliada dos boxplots das estimativas do parâmetro  $\lambda$ , com a remoção dos outliers exclusivamente para facilitar a visualização. Observa-se que, à medida que o tamanho amostral aumenta, há uma redução na variabilidade das estimativas, indicando maior precisão. Esses resultados evidenciam que a estimação realizada pelo pacote `skewlmm` é adequada e se torna ainda mais precisa com o aumento do tamanho amostral.

Optamos por não apresentar os gráficos dos demais parâmetros devido à limitação de páginas. No entanto, os resultados obtidos foram consistentes, com estimativas adequadas para todos os parâmetros, independentemente do tamanho amostral. Observa-se, como observado em  $\lambda$ , que a variabilidade das estimativas diminui à medida que o tamanho da amostra aumenta, enquanto a mediana e a média permanecem próximas ao valor real. Uma análise completa de todos os parâmetros pode ser consultada no Relatório Final.

Replicamos o estudo considerando a distribuição normal com assimetria negativa, e os resultados obtidos foram, em geral, semelhantes aos anteriores. A principal diferença, contudo, refere-se ao comportamento da variabilidade de  $\lambda$ : enquanto na distribuição com assimetria positiva, observamos maior dispersão para valores positivos, na assimetria negativa, a variabilidade concentrou-se em valores negativos, como era esperado teoricamente.

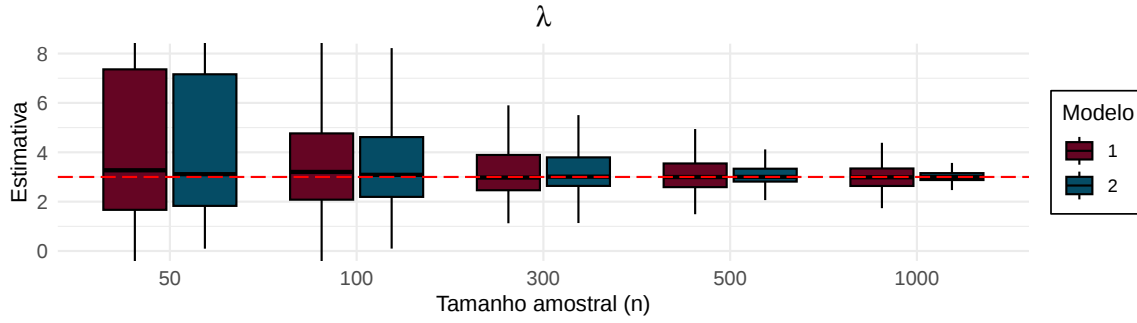


Figura 1: Simulação 1 - Boxplots das estimativas do parâmetro  $\lambda$  para cada um dos cenários considerados

De forma semelhante, conduzimos as simulações utilizando a distribuição t-student com assimetria positiva, fixando o parâmetro  $\nu = 6$ . Os resultados foram consistentes com os obtidos anteriormente, apresentando comportamento semelhante para todos os parâmetros, incluindo a estimação de  $\lambda$ . Especificamente para  $\nu$ , obtivemos boa estimação em todos os tamanhos amostrais.

## Simulação 2

Após verificarmos a boa performance do pacote na estimação dos parâmetros na Simulação 1, conduzimos novas simulações com tamanho amostral reduzido ( $n = 50$ ), avaliando diferentes estratégias de inicialização para o parâmetro de assimetria  $\lambda$ . O objetivo foi comparar o procedimento padrão do pacote com alternativas que pudessem melhorar a precisão das estimativas. Os valores reais foram mantidos os mesmos da simulação anterior para garantir comparabilidade.

| Ajuste   | Valor inicial de $\lambda$                      | Ajuste    | Valor inicial de $\lambda$                |
|----------|-------------------------------------------------|-----------|-------------------------------------------|
| Modelo 1 | Valor sugerido pelo pacote <code>skewlmm</code> | Modelo 7  | $3 \times$ Coef. de assimetria de Pearson |
| Modelo 2 | Valor real ( $\lambda = 3$ )                    | Modelo 8  | Coefficiente de assimetria de Fisher      |
| Modelo 3 | Valor fixado em 0,1                             | Modelo 9  | Coefficiente de assimetria de Groeneveld  |
| Modelo 4 | Valor fixado em 5                               | Modelo 10 | Coefficiente de assimetria MedCouple      |
| Modelo 5 | Valor fixado em $-3$                            | Modelo 11 | Coefficiente de assimetria de Yule        |
| Modelo 6 | Coefficiente de assimetria de Pearson           | Modelo 12 | Coefficiente de assimetria de Kelly       |

Tabela 2: Simulação 2 - Ajustes propostos com diferentes inicializações para o parâmetro  $\lambda$ .

Ao realizarmos a simulação para todos os modelos considerados, observamos que, em comparação com o Modelo 2 (que representa o melhor valor inicial possível - o valor real), o Modelo 1 apresentou o melhor desempenho entre as alternativas avaliadas, não sendo superado por nenhuma outra inicialização. Esses resultados indicam que, no contexto deste estudo, não foi identificado nenhum valor inicial para  $\lambda$  que superasse aquele sugerido automaticamente pelo pacote `skewlmm`.

## Simulação 3

Nesta etapa, avaliamos modelos mistos com intercepto e inclinação aleatórios, utilizando dados simulados a partir de uma distribuição t-student assimétrica multivariada para os efeitos aleatórios. O estudo foi dividido em duas partes: na primeira, considerou-se assimetria positiva para o intercepto e negativa para a inclinação (resultados apresentados na Simulação 3.1); na segunda, adotou-se assimetria positiva para ambos os efeitos (resultados apresentados na Simulação 3.2). Em ambos os cenários, os parâmetros utilizados para a geração dos dados estão apresentados na Tabela 3, com  $\lambda = (3, -2)$  para o Cenário 1 e  $\lambda = (2, 5)$  para o Cenário 2.

| Parâmetro  | $\beta$       | $\sigma^2$ | $\alpha$                | $\phi$ | $\nu$ |
|------------|---------------|------------|-------------------------|--------|-------|
| Valor real | $(1, 3)^\top$ | 0,25       | $(1, 0, 2, 0, 2, 0, 5)$ | 0,6    | 5     |

Tabela 3: Simulação 3 - Valores verdadeiros adotados para cada parâmetro no estudo.

### Simulação 3.1

De forma análoga às simulações anteriores, foram consideradas quatro estratégias distintas de inicialização para o vetor de parâmetros de assimetria  $\lambda$ , conforme apresentado na Tabela 4.

| Ajuste | Valor inicial de $\lambda$                                                                                           |
|--------|----------------------------------------------------------------------------------------------------------------------|
| 1      | Valores iniciais calculados pelo pacote                                                                              |
| 2      | Valores iniciais reais do conjunto de dados simulado                                                                 |
| 3      | Valores iniciais obtidos pelo pacote <i>lme</i> a partir do ajuste de um MLM com intercepto e inclinação aleatórios. |
| 4      | Valores iniciais de $\lambda$ fixados em $(1, -1)$                                                                   |

Tabela 4: Simulação 3.1 - Ajustes propostos com diferentes inicializações para o vetor de parâmetros  $\lambda$ .

Na Figura 2, apresentamos os boxplots das estimativas do vetor  $\lambda = (\lambda_1, \lambda_2)^\top$ , considerando dois tamanhos amostrais  $n = 50$  e  $n = 300$ . Em ambos os gráficos, observa-se uma redução clara na variabilidade das estimativas à medida que o tamanho amostral aumenta, indicando maior precisão para  $n = 300$  em todos os modelos avaliados.

Para ambos os parâmetros, o Modelo 2 apresentou desempenho superior, como era esperado, uma vez que utiliza os valores reais como ponto de partida. Os demais modelos apresentaram desempenho semelhante entre si e próximos aos do Modelo 2, o que evidencia, que os valores iniciais calculados automaticamente pelo pacote *skewlmm* fornecem boas estimativas.

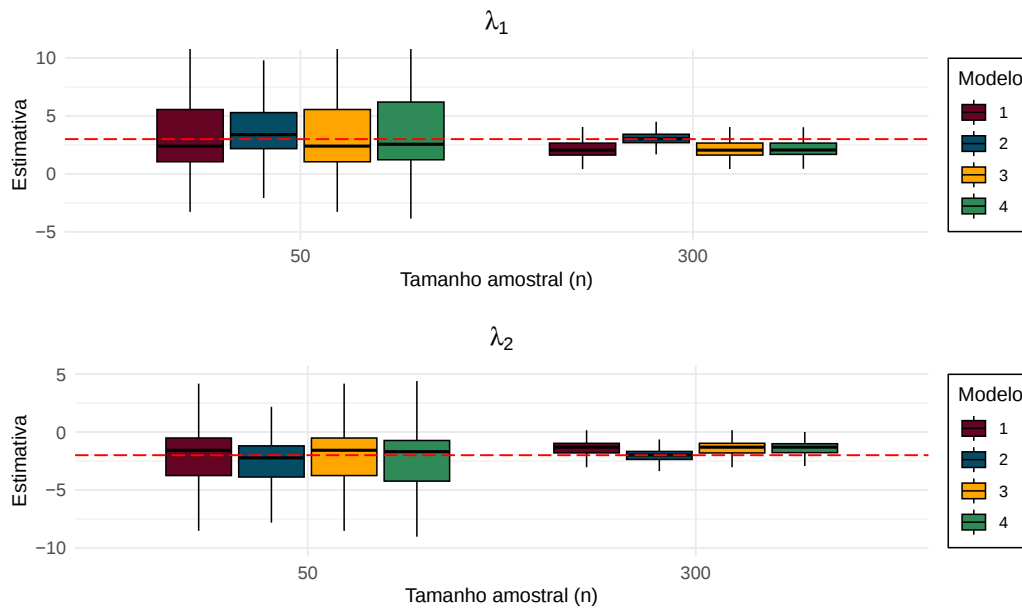


Figura 2: Simulação 3.1 — Boxplots das estimativas do vetor de parâmetros  $\lambda = (\lambda_1, \lambda_2)^\top$  para os cenários considerados.

### Simulação 3.2

Nesse cenário, como dito, adotamos  $\lambda = (2, 5)$ , mantendo os mesmos tamanhos amostrais utilizados anteriormente. Foram novamente considerados os Modelos 1, 2 e 3 da Simulação 3.1, alterando-se apenas a especificação do Modelo 4, onde o valor da inicialização para o vetor de parâmetros  $\lambda$  foi alterado para  $(1, 1)$ .

Na Figura 3, observa-se uma boa capacidade de estimação, especialmente do Modelo 2. Nos demais modelos, a estimação também se mostra satisfatória, embora, para alguns parâmetros, a mediana das estimativas não coincida exatamente com a linha que indica o valor real.

## Conclusões

A partir das diversas simulações realizadas, concluímos que os valores iniciais sugeridos pelo pacote *skewlmm* demonstraram ser os valores ótimos para alcançar estimativas precisas ao lidarmos com Modelos Lineares Mistos, quando os dados apresentam caudas pesadas e assimetria. Observamos que, mesmo em cenários com diferentes tamanhos amostrais e configurações de parâmetros, as rotinas de estimação implementadas no pacote foram capazes de convergir adequadamente para os valores verdadeiros dos parâmetros. Esses resultados reforçam a boa estimação, a partir de uma boa inicialização, o que acontece no pacote.

Descrições mais detalhadas dos resultados e estruturação das simulações encontram-se no Relatório Final.

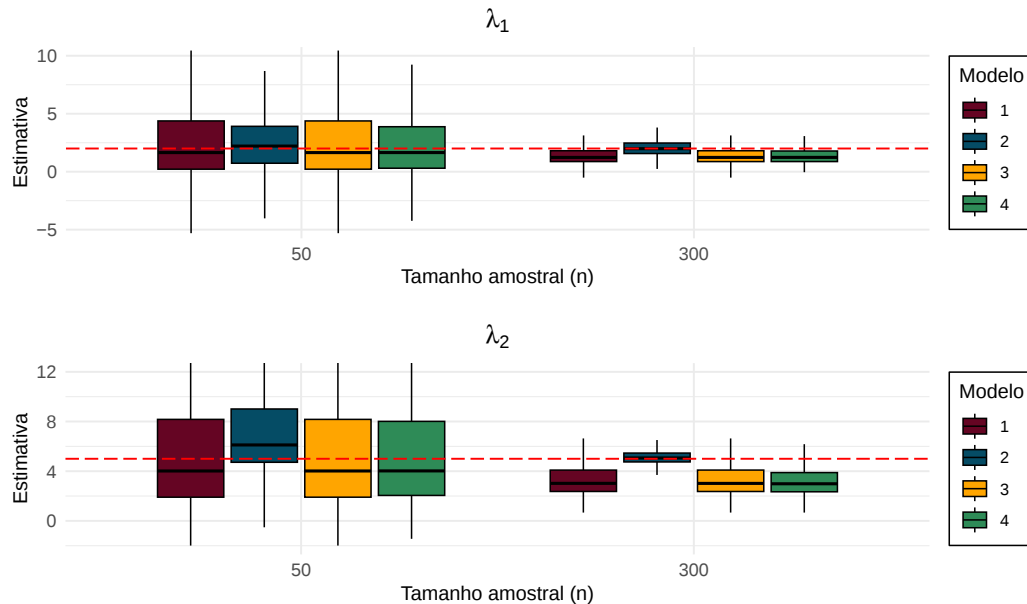


Figura 3: Simulação 3.2 — Boxplots das estimativas do vetor de parâmetros  $\lambda = (\lambda_1, \lambda_2)^\top$  para os cenários considerados.

## Bibliografia

- [1] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA, 1976.
- [2] M. D. Branco and D. K. Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79:99–113, 2001.
- [3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246.
- [4] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of longitudinal data*. Oxford University Press, 2nd edition, 2002.
- [5] R. Drikvandi, G. Verbeke, and G. Molenberghs. Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics*, 73(1):63–71, 2017.
- [6] V. H. Lachos, P. Ghosh, and R. B. Arellano-Valle. Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, 20:303–322, 2010.
- [7] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.
- [8] A. Muñoz, V. Carey, J. P. Schouten, M. Segal, and B. Rosner. A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics*, 48:733–742, 1992.
- [9] F. L. Schumacher, V. H. Lachos, and L. A. Matos. Scale mixture of skew-normal linear mixed models with within-subject serial dependence. *Statistics in Medicine*, 40(7):1790–1810, 2021. doi: 10.1002/sim.8870. URL <https://onlinelibrary.wiley.com/journal/sim>. Received: 23 June 2020; Revised: 9 December 2020; Accepted: 16 December 2020.
- [10] F. L. Schumacher, L. A. Matos, V. H. Lachos, and K. A. L. Valeriano. *skewlmm: Scale Mixture of Skew-Normal Linear Mixed Models*, 2023. URL <https://CRAN.R-project.org/package=skewlmm>. R package version 1.1.0.
- [11] J. M. Singer and D. F. Andrade. Analysis of longitudinal data. In *Bioenvironmental and Public Health Statistics*, volume 18 of *Handbook of Statistics*, pages 115–160. Elsevier, 2000.
- [12] J. M. Singer, F. M. Rocha, and J. S. Nobre. Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review*, 85(2):290–324, 2017.
- [13] R. E. Weiss. *Modeling longitudinal data*, volume 1. Springer, 2005.