UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFOMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

DOUTORADO EM ENGENHARIA DE TELEINFORMÁTICA

VINÍCIUS SILVA OSTERNE RIBEIRO

JOINT MODELING APPROACH WITH CHOLESKY DECOMPOSITION USING THE SCALE MIXTURES OF NORMAL DISTRIBUTIONS

FORTALEZA

2025

VINÍCIUS SILVA OSTERNE RIBEIRO

JOINT MODELING APPROACH WITH CHOLESKY DECOMPOSITION USING THE SCALE MIXTURES OF NORMAL DISTRIBUTIONS

Thesis submitted to the Graduate Program in Teleinformatics Engineering of the Technology Center at the Federal University of Ceará, as a partial requirement for obtaining the title of Doctor in Teleinformatics Engineering. Concentration Area: Signals and Systems.

Supervisor: Prof. Dr. Charles Casimiro Cavalcante

Co-supervisor: Prof. Dr. Lionel Bombrun

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

R372j Ribeiro, Vinicius Silva Osterne.
Joint Modeling Approach with Cholesky Decomposition Using the Scale Mixtures of normal Distributions / Vinicius Silva Osterne Ribeiro. – 2025.
136 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2025.
Orientação: Prof. Dr. Charles Casimiro Cavalcante.
Coorientação: Prof. Dr. Lionel Bombrun.

1. Longitudinal Data. 2. Scale Mixture of Normal distribution. 3. Modified Cholesky Decomposition. I. Título.

CDD 621.38

VINÍCIUS SILVA OSTERNE RIBEIRO

JOINT MODELING APPROACH WITH CHOLESKY DECOMPOSITION USING THE SCALE MIXTURES OF NORMAL DISTRIBUTIONS

Thesis submitted to the Graduate Program in Teleinformatics Engineering of the Technology Center at the Federal University of Ceará, as a partial requirement for obtaining the title of Doctor in Teleinformatics Engineering. Concentration Area: Signals and Systems.

Approved on: February 13th, 2025

EXAMINING COMMITTEE

Prof. Dr. Charles Casimiro Cavalcante (Supervisor)
Federal University of Ceará (PPGETI/UFC)

Prof. Dr. Lionel Bombrun
Bordeaux Sciences Agro, France

Prof. Dr. Juvêncio Santos Nobre
Federal University of Ceará (UFC)

Prof. Dr. Renato José de Sobral Cintra
Federal University of Pernambuco (UFPE)

Prof. Dr. Renato da Rocha Lopes
University of Campinas (UNICAMP)

Prof. Dr. Guilherme de Alencar Barreto
Federal University of Ceará (PPGETI/UFC)

Mãe e Pai, o suor de vocês valeu a pena.

ACKNOWLEDGEMENTS

First and foremost, I thank God for guiding me, giving me strength, and allowing me to walk this path with peace and purpose.

I am deeply grateful to my parents for their unconditional love, support, and encouragement throughout my journey.

Special thanks to Prof. Dr. Charles Casimiro Cavalcante for his continuous support and guidance.

I also sincerely thank Prof. Dr. Lionel Bombrun and Prof. Dr. Juvêncio Santos Nobre for their valuable insights, availability, and encouragement during this work.

“La vida es una milonga que hay que saber bailar.”

(Dicho popular rioplatense)

RESUMO

O processamento de dados longitudinais é de interesse significativo para vários campos, incluindo aplicações biomédicas e agronômicas. Para abordar isso, vários modelos mistos foram introduzidos na literatura (bem como modelos baseados em Equações de Estimativa Generalizadas, GEE), por exemplo. Esses modelos geralmente consideram a dependência dentro do sujeito ao parametrizar a matriz de dispersão usando a decomposição de Cholesky modificada (MCD) ou a decomposição de Cholesky alternativa (ACD). Os modelos tradicionais geralmente assumem que os resíduos seguem uma distribuição normal. No entanto, essa suposição pode não ser válida para muitos casos práticos. Extensões baseadas nas distribuições Student-t e Laplace foram propostas, mas podem ser muito restritivas devido à forma paramétrica fixa. Para abordar essas limitações, esta tese apresenta dois novos modelos de regressão que assumem que os resíduos são extraídos da família de misturas de escala de distribuições normais, considerando duas estruturas de decomposição diferentes para a matriz de dispersão. Desenvolvemos estimativas de máxima verossimilhança para esses modelos e as comparamos a modelos semelhantes que assumem distribuições diferentes, incluindo os modelos normal, Student-t e Laplace. Um estudo usando dados simulados demonstra que os modelos propostos são eficientes em comparação a modelos semelhantes e produzem resultados promissores para prever novas observações. Além disso, a metodologia é validada em dados reais, demonstrando bom desempenho em termos de precisão de estimativa e robustez a outliers.

Palavras-chave: Dados longitudinais. Mistura de escala de distribuições normais. Decomposição de Cholesky modificada.

ABSTRACT

Longitudinal data processing is of significant interest for several fields, including biomedical and agronomic applications. To address this approach, several mixed models have been introduced in the literature (as well as models based on Generalized Estimating Equations, GEE), for example. These models usually consider within-subject dependence by parameterizing the scatter matrix using the modified Cholesky decomposition (MCD) or the alternative Cholesky decomposition (ACD). Traditional models often assume that errors follow a normal distribution. However, this assumption may not be valid for many practical cases. Extensions based on the Student-*t* and Laplace distributions have been proposed, but they might be too restrictive due to the fixed parametric form. To address these limitations, this thesis presents a two new regression models that assum that the errors are drawn from the family of scale mixtures of normal distributions, considering two different decomposition structures for the scatter matrix. We develop maximum likelihood estimates for these models and compare it to similar models that assume different distributions, including the normal, Student-t, and Laplace models. A study using simulated data demonstrates that the proposed models are efficient compared to similar models and produces promising results for predicting new observations. Furthermore, the methodology is validated on real data, demonstrating good performance in terms of estimation accuracy and robustness to outliers.

Keywords: Longitudinal Data. Scale Mixture of Normal distribution. Modified Cholesky Decomposition.

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS AND ACRONYMS

ACD: Alternative Cholesky Decomposition

AIC: Akaike Information Criterion

ANOVA: Analysis of Variance

BIC: Bayesian Information Criterion

LJMM: Laplace Joint Modeling Model

MCD: Modified Cholesky Decomposition

MLE: Maximum Likelihood Estimation

MSE: Mean Squared Error

MSFE: Mean Square Forward Prediction Error

NJMM: Normal Joint Modeling Model

RIP: Relative Improvement Percentage

SE: Standard Error

SMN: Scale Mixture of Normal

SMNJMM: Scale Mixture of Normal Joint Modeling Model

TJMM: Student-t Joint Modeling Model

CONTENTS

1 INTRODUCTION

In this chapter, we introduce, in Section 1.1, the concept of what we call longitudinal data, a data framework that will be the basis for the development of our methodology and the importance of considering dependent data. Also in this chapter, in Section 1.2 we present the formulation of the chosen problem, in Section 1.3 we present the contribution of this thesis, and in Section 1.4 we present the thesis organization.

1.1 Longitudinal Data

1.1.1 Context about Longitudinal Data

Independence between observations is a common assumption for the application of numerous statistical techniques and is generally adequate when only one value is observed for each sample subject, as presented by (GELMAN; HILL, 2003; SEARLE *et al.*, 2009; MILLER; BEASLEY, 2015). In this structure, called a cross-sectional study, the sample is collected at a single time.

However, it is common for experiments to have more than one observation for each subject (a structure called repeated measures), and it is reasonable to consider the existence of some degree of dependence between these observations (intra-subject dependence) and, consequently, it is necessary to use more sophisticated methodologies to ensure more reliable results.

A special case of studies with repeated measures are the so-called longitudinal studies (called cohort or panel studies in health sciences), whose inherent characteristic is to consider the collection of the sample obtained at successive times ordered according to some dimension, be it time, speed, dosage, height, among others (see (SINGER; ANDRADE, 2000) for more examples), in which the measurement is made considering the same physical quantity.

There are, therefore, other cases that involve repeated measures, but which are not considered longitudinal, that is, they involve another approach. For example, we have time series (BOX *et al.*, 1976; HAMILTON, 1994; GRANGER; ENGLE, 1987), survival analysis (COX, 1972; KLEINBAUM; KLEIN, 2012; KALBFLEISCH; PRENTICE, 2002) and multivariate analysis (MARDIA *et al.*, 1979; ANDERSON, 2003; TABACHNICK; FIDELL, 2013), which consider, respectively, the monitoring of a single variable at many moments (historical series); the

monitoring, over time, until the outcome of the observed subject (lifetime); and the observation of a vector of measurements of the same subject, whose values can represent different physical quantities.

In the Figure 1, we graphically present the difference, based on simulated examples, between a study with time series data (on the left) and a study with longitudinal data (on the right). For the first, we considered the quantity of rice sold between the years 2000 and 2015, and for the second, we considered a study on the exposure index of a chemical product at different stages of a process.

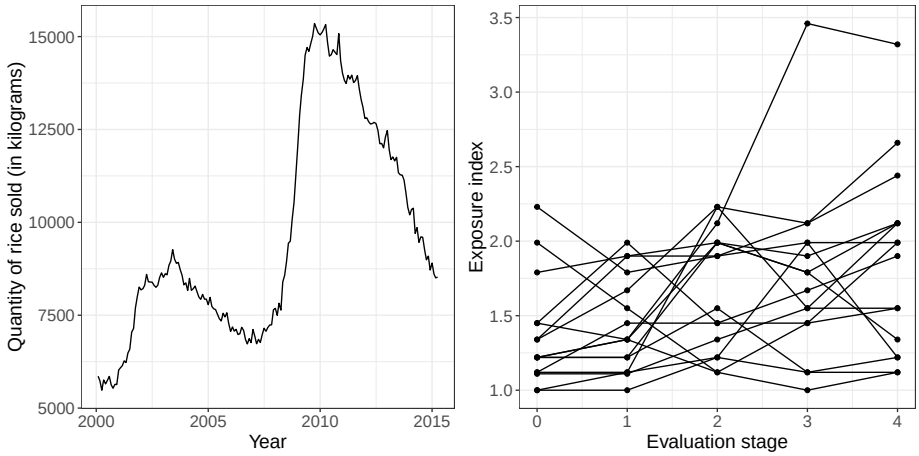


Figure 1 – Comparison between a study with time series data (left) and a study with longitudinal data (right).

Note that while time series evaluates single observations in each time period, longitudinal studies observe several subjects in each time period. Another important detail is that time series are usually evaluated over a long period of time, while longitudinal studies are generally shorter. In the next section, we highlight the importance of observing different subjects over time.

1.1.2 The application and importance of longitudinal data

The application of longitudinal studies is frequently used in several scientific fields, like psychology, economics, health and education, for example (see more about these application in (SINGER; ANDRADE, 2000; HEDEKER; GIBBONS, 2006; NEWSOM *et al.*, 2013; FITZMAURICE *et al.*, 2011)) and, not unlike other statistical methodologies, has advantages and disadvantages regarding its use. Monitoring sample subjects during the study, for example, may not be a simple task, as in some cases it generates high costs for the research.

However, according to (VENEZUELA, 2008), this approach requires fewer sample subjects than completely randomized designs; it provides more appropriate conditions for the study of covariates that may influence the response variable; in addition to assessing the behavior of the response over time and allowing the study of changes in the behavior of the average response of the sample subject in the different treatments (incorporating information on intra-individual variation in the analysis).

The difference, therefore, in longitudinal data analysis is to consider the dependence that exists within these individual response profiles of the observed subjects. Failure to consider this inherent characteristic can lead us to mistaken conclusions/interpretations of the results obtained. As an illustration, we simulated a hypothetical example, similar to that known in the literature as the *Sales paradox* (DEMIDENKO, 2013), in which we are interested in investigating whether the discount of a specific product (here we consider 5 product categories, totaling 10 products) affects its quantity sold. Note in Figure 2 the values collected.

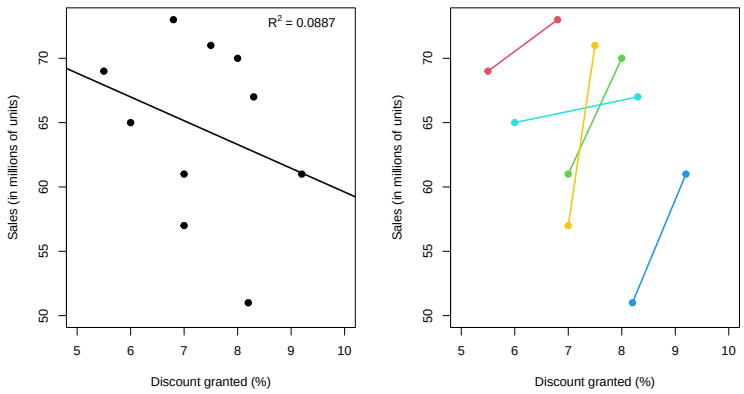


Figure 2 – Comparison between a cross-sectional approach study with the fitted regression line (left) and a longitudinal study approach with the identification of repeated observations per subject (right).

The graph on the left shows the plot of all observations collected considering that they come from the same reference population. If we were to fit a regression model in this case, we could conclude that product sales fall as their discount increases. However, if we consider the existence of dependence between observations, that is, considering that the products belong to different categories, as shown by the different colors in the graph on the right, the interpretation is completely different, that is, the greater the discount offered, the more the product sells, evidencing, in this case, a paradox and hence the importance of interpreting this data as longitudinal data.

1.2 Formulation of the thesis problem

As we initially presented, a longitudinal study involves evaluating sample subjects observed over time (unlike traditional studies called cross-sectional studies, which evaluate the sample subject at a single point in time). For this reason, longitudinal studies are more complex, because, in addition to needing to define a structure to model the mean, we also need to define a structure to model the dispersion matrix, which carries the correlation information of the components of the measurements collected within the same subject, which are not independent, as in classical models.

Thus, choosing a good approach to model the dispersion matrix will positively influence the inference about the mean and the covariance structure itself. It is worth noting that longitudinal models generalize to the case with dependence, but can be used when we do not have independence between observations.

It is precisely about "defining a good modeling approach for the scatter matrix" that the approach to this problem becomes challenging, as this matrix has peculiarities that make estimation processes difficult, such as high dimensionality (we will have many parameters to be estimated; if we do not impose any structure on the matrix, we will have more parameters to estimate than observations collected) and its inherent property: it is a positive definite matrix.

Over time, the literature has proposed several approaches that consider all these peculiarities, but in this thesis, we will focus on those works that proposed the use of joint modeling between mean and covariance to overcome these estimation obstacles. This type of joint modeling approach was initially presented by (POURAHMADI, 1999a), who advocated a data-driven method based on a modified Cholesky decomposition (MCD) of the within-subject marginal scatter matrix. The decomposition leads to a reparameterization where the inputs can be interpreted in terms of innovation variances and autoregressive coefficients.

Next, (CHEN; DUNSON, 2003a) present another Cholesky-type decomposition (called Alternative Cholesky decomposition, ACD), which can be understood as modeling certain innovation variances and moving average parameters. More recently, (ZHANG *et al.*, 2015) considered a regression approach based on the standard Cholesky decomposition of the correlation matrix and the hyperspherical parameterization (HPC) of its Cholesky factor, whose parameters are directly interpretable with respect to variance and correlation.

All these approaches to estimate the scatter matrix were developed in the context of regression for specific distributions. First, (POURAHMADI, 1999a) assumed the normal

assumption for the errors. Even though this assumption is commonly used in linear regression models, the model obtained is sensitive to outliers. To overcome this limitation, (LIN; WANG, 2009) and (GUNEY *et al.*, 2022) proposed, respectively, to use the Student-t and Laplace distributions, further encompassing the modeling options available in the literature.

However, even though these distributions have heavier tails compared to the normal distribution, these models suffer from having a fixed parametric form that can be very restrictive in practice. It is, therefore, from this limitation that we will develop this thesis. The idea is to introduce a more general and robust model by considering that the errors belong to the family of mixture of scale normal distributions (SMN) that encompasses many distributions, such as the normal, Student-t and Laplace.

1.3 Contributions of the thesis

As previously mentioned, the aim of this thesis is to present a more comprehensive and robust model, assuming that the errors follow a family of scale mixture of normal distributions (SMN), which includes several distributions, such as the normal, Student-t and Laplace. To better illustrate our contributions, we will highlight each of them below:

- By considering that the errors follow a scale mixture of normal distribution scale with unknown but deterministic mixing parameters, we introduce a new robust joint modeling framework for modeling longitudinal data;
- Based on the MCD (modified Cholesky decomposition) and ACD (alternative Cholesky decomposition), we obtain the maximum likelihood estimators and provide an iterative algorithm to estimate their parameters;
- We provide an analytical expression for the mean squared error predictor in the context of future forecasting;
- We present experiments on simulated and real data to illustrate the robustness of the proposed approach by comparing different probability distributions and scatter matrix structures..

Below, we present an illustration for a better understanding of all the topics we used to present this new proposal, going through the different distributions of the SMN family (normal, Student-t, Laplace and others), the different joint models already proposed (normal Joint Modeling Model, called as NJMM; Student-t Joint Modeling Model, called as TJMM, Laplace Joint Modeling Model, called as LJMM; and others), the different Cholesky decompositions

(MCD and ACD) and the joint robust estimation approach of the mean and the scatter matrix:

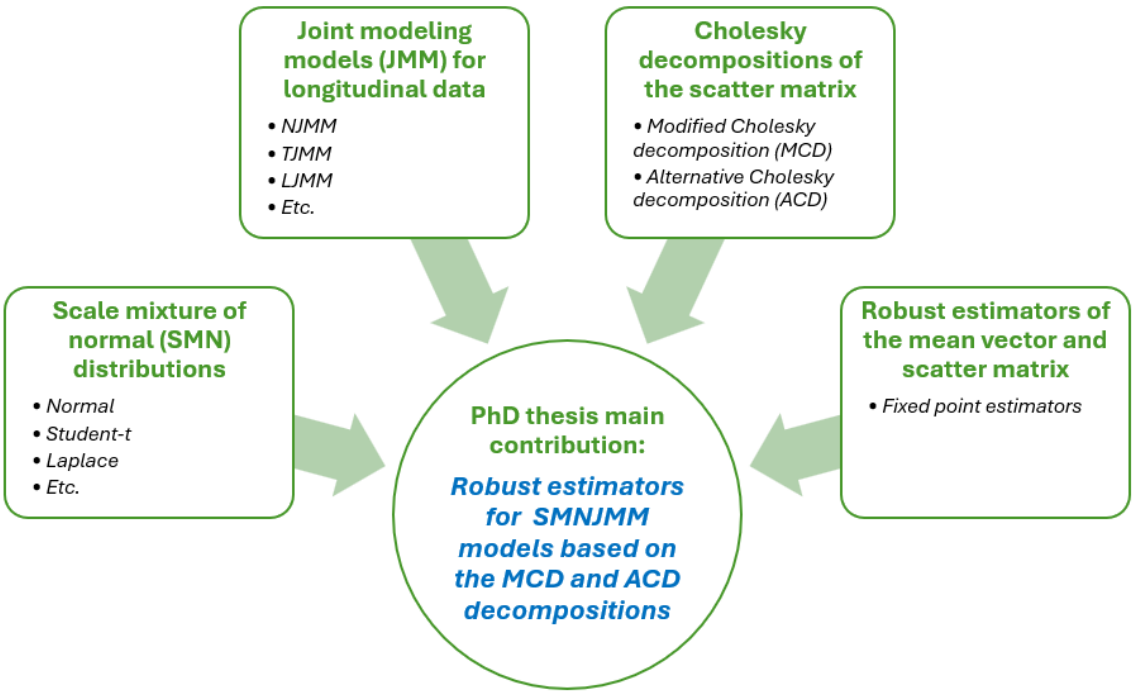


Figure 3 – Main contributions of the thesis.

It is important to highlight that, within this context, certain studies deserve special attention: linear mixed models based on scale mixtures of asymmetric normal distributions have been widely explored in recent literature to address longitudinal data and robust modeling. (FERREIRA *et al.*, 2022a) investigate linear mixed models based on scale mixtures of asymmetric normals, allowing for greater flexibility in capturing variability structures in the data. (ZELLER *et al.*, 2023) propose a robust mixture regression model grounded in these distributions, aiming to improve estimation in the presence of outliers.

(SCHUMACHER *et al.*, 2021) extend this approach by considering serial dependence within subjects, enabling more accurate modeling of the correlation structure in longitudinal data. (NUMMI *et al.*, 2025) explore enhanced methods for estimating mixture components in longitudinal data, emphasizing improvements in statistical inference. Furthermore, (SCHUMACHER *et al.*, 2023) develop approximate inference procedures for nonlinear mixed-effects models using scale mixtures of asymmetric normal distributions, expanding the applicability of these models.

In the next section, we present details of what will be presented in each chapter of this thesis aiming at a complete explanation of the proposed model.

1.4 Organization of the Thesis

In Chapter 2, we present the classical approaches considered for modeling longitudinal data and the structure that will serve as a basis for the development of the study of this thesis. We will also present the structures of the two Cholesky decompositions considered for the scatter matrix and conclude by presenting the family of elliptical distributions.

Next, in Chapter 3, we present the proposed bases for the development of our proposal, that is, the models that use the Cholesky decomposition from the normal, Student-t and Laplace distributions. After this introduction, we present the main content of this thesis, which is the development of the joint modeling approach using the family of scale mixture of normal distributions, including the estimation method, estimation algorithm and forward prediction. We conclude Chapter 3 with a discussion of the different models.

In Chapters 4 and 5, we present the model performance results based on simulated (synthetic) and real data, respectively, for the models using the normal, Student-t and Laplace distributions and for the proposed model considering the modified and alternative Cholesky decompositions. The results are presented in terms of the standard errors of the estimates and the BIC values adapted to the problem. Forward prediction estimates are also evaluated. All conclusions about the work performed and future works are presented in Chapter 6.

The thsis most results was divided into two separate manuscripts (one published and one under review), which we call "A general robust approach for joint modeling of the family of scale mixture of normal distribution", where we develop the joint modeling approach considering the MCD decomposition; and "Joint modeling approach with Alternative Cholesky Decomposition using the family of scale mixture of normal distribution" where we develop the joint modeling approach considering the ACD decomposition, as described below:

- Ribeiro, V. S. O., Bombrun, L., Nobre, J. S., Cavalcante, C. C., & Berthoumieu, Y. (2024). A general robust approach for joint modeling of the family of scale mixture of normal distribution. Signal Processing, 220, 109426.
- Ribeiro, V. S. O., Bombrun, L., Nobre, J. S., Cavalcante, C. C., & Berthoumieu, Y. (2025). Joint modeling approach with Alternative Cholesky Decomposition using the family of scale mixture of normal distribution. Under review.

2 STATE OF ART

In this chapter, we introduce, in Section 2.1 the existing classical approaches for longitudinal modeling data, in Section 2.2 we present the one we used to develop the proposed model. Also in this chapter, in Section 2.3 we present the two Cholesky decompositions considered for the scatter matrix and in Section 2.4 we conclude by presenting the family of elliptical distributions.

2.1 Classical approaches to Longitudinal Data

Additional work on the methodology of studies involving longitudinal data refers to the way in which the data dependence structure will be modeled. In some cases, we can reduce a multivariate study to a univariate study, without needing to use sophisticated techniques for the analysis. The use of the paired t-test or ANOVA (parametric or non-parametric), for example, may be sufficient. The first approach can be used when only two observations are made under the same subject and, therefore, a test comparing means will be efficient for the situation. In the case of ANOVA (which can be used when we have more than two observations from the same subject), however, it will be necessary to use summary measures, such as the area under the curve and outcome analysis. Details on this can be found in (SINGER; ANDRADE, 2000), for example.

Despite being widely applied in various situations, analyses using techniques such as those listed above can be limited in some cases (as the presence of dependence between observations). Based on this, many authors have proposed, over time, alternative approaches to longitudinal study approaches. Much of the initial research was aimed at cases in which the errors of the model follows a normal distribution. Among these, we can mention the analysis of variance with repeated measures (NETER *et al.*, 1996), univariate/multivariate analysis of profiles and the analysis of growth curves (SINGER; ANDRADE, 2000).

Another important approach for the analysis of longitudinal data is based on Generalized Estimating Equations (GEE), introduced by (LIANG; ZEGER, 1986) as an extension of generalized linear models to account for within-subject correlation. Unlike linear mixed models, which explicitly model the random effects structure, GEE provides a population-averaged approach by estimating parameters using quasi-likelihood methods. This framework has been widely extended to accommodate different correlation structures and link functions (DIGGLE

et al., 2002; FITZMAURICE *et al.*, 2011). Recent advancements have extended the Generalized Estimating Equations (GEE) framework to better handle bounded continuous responses. (VENEZUELA *et al.*, 2007) proposed a GEE approach based on the Beta distribution, providing a flexible alternative for modeling proportions in longitudinal data. More recently, (RIBEIRO *et al.*, 2021) introduced a Beta-Rectangular GEE model, further expanding the applicability of this methodology by accommodating responses with a wider range of distributional shapes.

Longitudinal data analysis has been extended beyond the classical assumption of normality, leading to the development of several flexible modeling approaches. General marginal models have been proposed to accommodate non-normal responses, including elliptical models (FANG *et al.*, 1990), asymmetric models (FERREIRA *et al.*, 2022b), and elliptical-asymmetric models (MATOS *et al.*, 2019). Multivariate models (MARDIA *et al.*, 2000) have also been explored to jointly model multiple correlated responses over time. Mixed-effects models (FITZMAURICE *et al.*, 2011) remain widely used due to their ability to incorporate subject-specific variability, while Generalized Estimating Equations (GEE) (LIANG; ZEGER, 1986) provide a robust alternative for marginal inference. Furthermore, Generalized Additive Models for Location, Scale, and Shape (GAMLSS) (RIGBY; STASINOPOULOS, 2005) has emerged as a powerful framework for modeling complex longitudinal data, allowing flexibility in distributional assumptions and incorporating covariance effects in multiple parameters of the response distribution.

The approaches mentioned above offer different ways of dealing with non-normality and other complex features of longitudinal data. However, depending on the structure of the data and the objectives of the analysis, some of these methodologies may present specific limitations. Given this, in the next section, we explore a particular alternative that best fits the context of our study.

2.2 Joint modeling approach of Longitudinal Data

Suppose that repeated measurements of a continuous random variable are observed over time in each of the m subjects under study. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^\top$ be the result vector of the ith subject (experimental subject), where n_i is the number of results per subject i, $i = 1, 2, ..., m$, which are allowed unevenly spaced.

One of the most common ways of modeling the mean of $\mathbf{y}_i$ is the linear regression

model given by

$$\mathbf{y}_i = \boldsymbol{\mu}_i + \mathbf{r}_i, \tag{2.1}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{i_{n_i}})^\top = \mathbf{X}_i\boldsymbol{\beta}$, with $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i})^\top$ being the $n_i \times (p+1)$ matrix of explanatory variables, i.e.

$$\mathbf{X}_i = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_i1} & x_{n_i2} & \cdots & x_{n_ip} \end{bmatrix},$$

where x_{ij} represents the value of variable j at observation i [1], n_i is the number of observations, and p is the number of variables, which is known and assumed to be of full row rank, $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^\top$ being the vector of regression coefficients and $\mathbf{r}_i = (r_{i1}, ..., r_{in_i})^\top = \mathbf{y}_i - \boldsymbol{\mu}_i$ is the error vector, which can assume different decompositions, among the most common is to consider that $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$.

In addition to mean modeling, the focus of regression models for longitudinal data is also on modeling the scatter matrix, given that we have repeated measures for different individuals and modeling this structure must be considered. However, as we have already mentioned previously, this matrix has peculiarities that make estimation processes difficult, such as high dimensionality, positive definition and, if we do not consider any restrictions, we will have more parameters to estimate than collected observations.

Based on these limitations, the joint modeling methodology considers using Cholesky decomposition as a methodology for this context, as we will see in the following section.

2.3 Cholesky decomposition

2.3.1 MCD approach

In the context of joint mean-covariance modeling, the idea proposed by Pourahmadi in (POURAHMADI, 1999b) and (POURAHMADI, 2000) to ensure that the properties (high dimensionality and positive definite) is to rewrite the scatter matrix using the Modified Cholesky

[1] in this notation, the index i refers both to the i-th subject, since it is an element of the matrix X_i, and to the i-th repeated measurement over time

Decomposition (MCD) approach, where such matrix is reparametrized as follows

$$\mathbf{L}_i \boldsymbol{\Sigma}_i \mathbf{L}_i^\top = \mathbf{D}_i, \tag{2.2}$$

where $\mathbf{L}_i = [l_{jk}]$ are lower triangular matrices with 1's as diagonal entries and (j,k)th entry being $-\phi_{jk}$, called autoregressive parameters, for $k \in \{1,...,j-1\}$, as denoted below

$$\mathbf{L}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_{21} & 1 & 0 & \cdots & 0 \\ -\phi_{31} & -\phi_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\phi_{n_i 1} & -\phi_{n_i 2} & -\phi_{n_i 3} & \cdots & 1 \end{bmatrix}, \tag{2.3}$$

and $\mathbf{D}_i = \text{diag}\{\sigma_1^2,...,\sigma_{n_i}^2\}$, where σ_j^2 are called scale innovation variances, as denoted below

$$\mathbf{D}_i = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{n_i}^2 \end{bmatrix}, \tag{2.4}$$

whose interpretations include: (i.) the below-diagonal entries of $\mathbf{L}_i$ are the negatives of the autoregressive parameters, namely $-\phi_{jk}$, in

$$y_{ij} = \mathbf{x}_{ij}^\top \beta + \sum_{k=1}^{j-1} \phi_{jk}(y_{ik} - \mu_{ik}), \tag{2.5}$$

and (ii.) the diagonal entries of $\mathbf{D}_i$ are the scale innovation variances $\sigma_j^2 = \text{var}(y_{ij} - \hat{y}_{ij})$. It is worth noting that we can also rewrite the inverse of the scatter matrix as $\boldsymbol{\Sigma}_i^{-1} = \mathbf{L}_i^\top \mathbf{D}_i^{-1} \mathbf{L}_i$.

Following (POURAHMADI, 1999b) and (POURAHMADI, 2000), we are able to model the scatter matrix from a joint regression model for variances and correlations considering

$$\phi_{jk} = \mathbf{z}_{jk}^\top \boldsymbol{\gamma} \quad \text{and} \quad \ln \sigma_j^2 = \mathbf{w}_j^\top \boldsymbol{\lambda}, \tag{2.6}$$

where $\boldsymbol{\gamma} = (\gamma_1,...,\gamma_d)^\top$ and $\boldsymbol{\lambda} = (\lambda_1,...,\lambda_q)^\top$ are d and q-dimensional vectors of parameters, $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ are d and q-dimensional covariate vectors.

The parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are assumed to be common for all $\boldsymbol{\Sigma}_i$'s exhibiting the same covariance structure. The covariates $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ may contain the baseline covariates, the time and the associated interactions. These covariates are presented as time polynomials for the

autoregressive components and time powers for innovation variances, respectively, as proposed in (YE; PAN, 2006), i.e.

$$\mathbf{z}_{jk} = \left[\begin{array}{ccccc} 1 & (t_{ik}-t_{ij}) & (t_{ik}-t_{ij})^2 & \cdots & (t_{ik}-t_{ij})^{d-1} \end{array} \right]^\top \tag{2.7}$$

and

$$\mathbf{w}_j = \left[\begin{array}{ccccc} 1 & t_{ij} & t_{ij}^2 & \cdots & t_{ij}^{q-1} \end{array} \right]^\top, \tag{2.8}$$

where t_{ij} (or t_{ik}) is the measure of time referring to the subject i in the repetition j (or repetition k), for $i \in \{1,\cdots,m\}$, $j \in \{1,\cdots,n_i\}$, and $k \in \{1,\cdots,j-1\}$.

2.3.2 ACD approach

Another structure also used in this context is the so-called ACD decomposition (Alternative Cholesky Decomposition), proposed by Chen et al. (CHEN; DUNSON, 2003b), it is presented as modeling innovation variances and moving average parameters. With this decomposition, the scatter matrix is rewritten as

$$\boldsymbol{\Sigma}_i = \mathbf{D}_i \mathbf{L}_i \mathbf{L}_i^\top \mathbf{D}_i, \tag{2.9}$$

where $\mathbf{L}_i$, in this case, is given by

$$\mathbf{L}_i = \left[\begin{array}{ccccc} 1 & 0 & 0 & \cdots & 0 \\ \phi_{21} & 1 & 0 & \cdots & 0 \\ \phi_{31} & \phi_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{n_i 1} & \phi_{n_i 2} & \phi_{n_i 3} & \cdots & 1 \end{array} \right]. \tag{2.10}$$

ϕ_{jk} are the moving average parameters such that:

$$(y_{ij}-\mu_{ik})/\sigma_i = \varepsilon_{ij} + \sum_{k=1}^{j-1} \phi_{jk}\varepsilon_{ij}, \tag{2.11}$$

and $\mathbf{D}_i = \text{diag}\{\sigma_1^2, ..., \sigma_{n_i}^2\}$ is a diagonal matrix containing the scale innovation variances $\sigma_j^2 = \text{var}(y_{ij} - \hat{y}_{ij})$. Following (POURAHMADI, 1999b) and (POURAHMADI, 2000), it is possible to describe the modeling structure of the scatter matrix assuming these two linear models

$$\phi_{jk} = \mathbf{z}_{jk}^\top \boldsymbol{\gamma} \quad \text{and} \quad \ln \sigma_j^2 = \mathbf{w}_j^\top \boldsymbol{\lambda}, \tag{2.12}$$

where $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_d)^\top$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_q)^\top$. $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ are known as the d and q dimensional covariate vectors. Note that vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are assumed to be the same for all subjects i.

The covariates $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ are presented as time polynomials and powers of time for innovation variances, respectively, as proposed in (YE; PAN, 2006), i.e,

$$\mathbf{z}_{jk} = \begin{bmatrix} 1 & (t_{ik} - t_{ij}) & (t_{ik} - t_{ij})^2 & \cdots & (t_{ik} - t_{ij})^{d-1} \end{bmatrix}^\top \tag{2.13}$$

and

$$\mathbf{w}_j = \begin{bmatrix} 1 & t_{ij} & t_{ij}^2 & \cdots & t_{ij}^{q-1} \end{bmatrix}^\top, \tag{2.14}$$

where t_{ij} (respectively t_{ik}) represents the time measurement for subject i at repetition j (respectively repetition k), with $i \in \{1, \cdots, m\}$, $j \in \{1, \cdots, n_i\}$, and $k \in \{1, \cdots, j-1\}$.

Based on these decompositions to model the structure of the scatter matrix, some authors have proposed several multivariate zero-mean models for the errors, considering the normal, Student-t and Laplace models. But before presenting these models in detail, let us return to the definition of the family of elliptical distributions that encompasses exactly the distributions of the proposed models, that is, the normal, Student-t and Laplace distributions. We will also recall the expressions of the maximum likelihood estimators when no restrictions are made on the shape of the scatter matrix.

2.4 Elliptical Symmetric Distributions

The family of elliptical distributions is a class of multivariate probability distributions that generalize other distributions (MUIRHEAD, 1982). They are so named because of the shape of their probability density distributions, which are elliptical in multidimensional space, that is, they have the shape of an ellipse around the median (JOHNSON *et al.*, 1994; KOTZ; NADARAJAH, 2000; JOHNSON *et al.*, 2005).

Elliptical distributions provide a flexible model for multivariate data and have applications in several areas, such as finance, statistics and engineering (FERNANDEZ; STEEL, 1996). The main characteristic of this family is that it encompasses both heavy-tailed distributions and more regular distributions, such as the normal distribution. Some works already use this distribution in the longitudinal context, such as (SAVALLI *et al.*, 2006; OSORIO *et al.*, 2007).

We say that an observed vector $\mathbf{y} = (y_1, ..., y_n)^\top$ with vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_n)^\top$ and scatter matrix $\boldsymbol{\Sigma}$ follows the elliptical symmetric distribution, if its density is of the form:

$$f(\mathbf{y}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} h\left\{(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}, \tag{2.15}$$

being $\mathbf{y}, \boldsymbol{\mu} \in \mathbb{R}^n$; $h : [0,\infty) \longrightarrow [0,\infty)$ is such that $\int_0^\infty u^{n/2-1} h(u)du < \infty$ known as density-generating function. In general, this density shows that elliptical distributions are defined by the function h, which is symmetric around 0, and by the scatter matrix ■ that controls the dispersion of the data. The notation for this case is given by $\mathbf{y} \sim \mathrm{EL}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, h)$.

In particular, for expression (2.15), where $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_n$, being $\mathbf{I}_n$ the identity matrix of order n, we have the called spherical distribution (PAULA *et al.*, 2009). Below we highlight some more interesting properties about this family of distributions:

- Symmetry: The distribution is symmetric about its center μ.
- Invariance property under linear transformations: If $\mathbf{X}$ follows an elliptical distribution, then for any matrix A and vector $\mathbf{b}$, the random variable $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ also follows an elliptical distribution.
- Heavy tails: Some elliptical distributions (such as Student-t) have heavier tails than the normal, which allows them to model data with more frequently observed extreme values (there are distributions with lighter tails too, such as the power exponential distribution).

There are several distributions that belong to the family of elliptical distributions, each with specific characteristics that make them suitable for different contexts and types of data. Below we highlight some examples of these distributions, which illustrate the flexibility and applicability of this family.

2.4.1 Normal distribution

If we consider that in expression (2.15), the function $h(\cdot)$ is given by $h(u) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp(-u/2)$, then we will have as a particular case the normal distribution, that is, its log-likelihood is of the form:

$$l_1(\mathbf{y}) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}),$$

where the Maximum Likelihood estimators of the mean and the scatter matrix, for a sample $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_m)$, are given by the follow expressions

$$\hat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{y}_i \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{m}\sum_{i=1}^{m} (\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})^\top.$$

2.4.2 Student-t distribution

If we consider that in expression (2.15), the function $h(\cdot)$ is given by $h(u) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{n/2}\pi^{n/2}}\left(1+\frac{u}{\nu}\right)^{-(\nu+n)/2}$, then we will have as a particular case the Student-t distribution, that is, its log-likelihood is of the form:

$$l_2(\mathbf{y}) = \log\Gamma\left(\frac{\nu+n}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2}\log\left(\nu^n\pi^n|\Sigma|\right) - \frac{(\nu+n)}{2}\log\left(1+\frac{u}{\nu}\right),$$

where $\Gamma(\cdot)$ é a chamada função Gamma ((ABRAMOWITZ; STEGUN, 1964; ARTIN, 2015)), ν is the degree of freedom and the Maximum Likelihood estimators of the mean and the scatter matrix, for a sample $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_m)$, can be obtained by iterative algorithms, considering that

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^m \hat{w}_i\mathbf{y}_i}{\sum_{i=1}^m \hat{w}_i} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{m}\sum_{i=1}^m \hat{w}_i(\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})^\top,$$

where $\hat{w}_i = (\nu+n)/(\nu+\hat{u}_i)$, with $\hat{u}_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})$.

To obtain the estimates, the iteratively reweighted algorithm which can be identified as an EM algorithm (Expectation–Maximization algorithm) can be used, considering that the score function to ν is given by

$$U_\nu = \frac{1}{2}\sum_{i=1}^m\left(\Psi\left(\frac{\nu+n}{2}\right) - \Psi\left(\frac{\nu}{2}\right) - \frac{n}{\nu} - \log(\nu+u_i) + \log(\nu) + \frac{(\nu+n)u_i}{(\nu+u_i)\nu}\right),$$

where $\Psi(\cdot)$ is the digamma function ((ABRAMOWITZ; STEGUN, 1964; TRICOMI, 1951)).

2.4.3 SMN distributions

We say that an n-dimensional random variable $\mathbf{y}$ belongs to a family of scale mixture of normal, presented by (ANDREWS; MALLOWS, 1974), with mean vector $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$, if its stochastic representation can be given by

$$\mathbf{y} = \boldsymbol{\mu} + \tau^{1/2}\mathbf{z}, \tag{2.16}$$

where τ and $\mathbf{z}$ are independent, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\tau \in \mathbb{R}^+$ is a scalar random variable. An important property of this structure is that the random variable $\mathbf{y}$ given τ has a normal distribution, where the mean vector and scatter matrix are given, respectively, by

$$\mathbb{E}(\mathbf{y}|\tau) = \sqrt{\tau}\mathbb{E}(\mathbf{z}) + \mu = \mu \tag{2.17}$$

and

$$\text{Var}(\mathbf{y}|\tau) = \mathbb{E}[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^\top|\tau] = \sqrt{\tau}\mathbb{E}(\mathbf{z}\mathbf{z}^\top)\sqrt{\tau} = \tau\boldsymbol{\Sigma}, \tag{2.18}$$

and consequently we can write its conditional distribution function as follows

$$\mathbf{y}|\tau \sim \mathcal{N}(\boldsymbol{\mu}, \tau\boldsymbol{\Sigma}) = \frac{1}{(2\pi\tau)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2\tau}\right\}. \tag{2.19}$$

Hence, each component of $\mathbf{y}$ is correlated with the other components from the variance dependence generated by τ. We can obtain the probability density function of the random variable $\mathbf{y}$ by integrating the joint distribution of $\mathbf{y}$ and τ over τ, i.e.

$$f(\mathbf{y}) = \int_0^\infty f(\mathbf{y}|\tau) f(\tau) d\tau. \tag{2.20}$$

Depending on the choice of the distribution of the parameter τ, we may have different structure for the multivariate vector $\mathbf{y}$, i.e.:

- if τ follows an Inverse Gamma distribution with shape parameter a and scale parameter b, denoted as $\tau \sim IG(a,b)$, whose probability density function is given by
  $$f(\tau;a,b) = \frac{b^a}{\Gamma(a)} \tau^{-a-1} e^{-b/\tau}, \tag{2.21}$$
  where $\tau \in (0,\infty)$ and $\Gamma(\cdot)$ is the complete Gamma function, then $\mathbf{y}$ follows a multivariate Student-t distribution (ROSCA *et al.*, 2006), as shown in section 2.4.2;
- if τ follows a Gamma distribution with shape parameter a and scale parameter b, denoted $\tau \sim G(a,b)$, whose probability density function is given by
  $$f(\tau;a,b) = \frac{1}{b^a\Gamma(a)} \tau^{a-1} e^{-\tau/b}, \tag{2.22}$$
  where $\tau \in (0,\infty)$, then $\mathbf{y}$ follows a K-distribution, a family of three parameters that arise from the composition of two Gamma distributions. In addition, if τ follows a Gamma distribution with shape parameter $a = n + 1/2$ and scale parameter $b = 1/8$, then $\mathbf{y}$ follows a Laplace distribution.

As mentioned before, by choosing a specific structure for the τ parameter, in the SMN family of distributions, a closed-form expression for $\mathbf{y}$ can be obtained by solving (2.20) such as the Student-t and Laplace distributions. Here, we will consider only the absolutely continuous cases, since depending on the choice of τ, we can also obtain discrete distributions.

2.4.4 SMN distribution to robust estimators

The main disadvantage of the previously mentioned approaches is that a well-defined parametric form is imposed for the structure of τ and, consequently, for the variable $\mathbf{y}$, which can be easily violated in practice. To overcome this problem, (PASCAL *et al.*, 2006), (CHITOUR; PASCAL, 2008) and (FORMONT *et al.*, 2010) propose to relax this restriction by not imposing

any explicit structure for the probability density function of τ. Furthermore, they consider that the value of τ varies with each subject, so that for a set of m individuals, we have the parameters $\tau_1, \tau_2, ..., \tau_m$ to be estimated, that are deterministic but unknown, and the ML estimator for this parameters are given by

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{m} \frac{\mathbf{y}_i}{[(\mathbf{y}_i-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\mu}})]^{1/2}}}{\sum_{i=1}^{m} \frac{1}{[(\mathbf{y}_i-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\mu}})]^{1/2}}} \tag{2.23}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{n}{m}\sum_{i=1}^{m} \frac{(\mathbf{y}_i-\hat{\boldsymbol{\mu}})(\mathbf{y}_i-\hat{\boldsymbol{\mu}})^\top}{(\mathbf{y}_i-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\mu}})} \tag{2.24}$$

$$\hat{\tau}_i = \frac{1}{n}(\mathbf{y}_i-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\mu}}). \tag{2.25}$$

An important aspect about the scatter matrix $\boldsymbol{\Sigma}$ is its identification problem. We can rewrite the stochastic representation presented in (2.16), given by $\mathbf{y} = \boldsymbol{\mu} + \tau^{1/2}\mathbf{z}$, considering $\mathbf{y} = \boldsymbol{\mu} + (a\tau)^{1/2}\frac{\mathbf{z}}{a^{1/2}}$ or simply $\mathbf{y} = \boldsymbol{\mu} + (\kappa)^{1/2}\mathbf{c}$, where $\kappa = a\tau$ is a positive scalar parameter and $\mathbf{c} \sim \mathcal{N}(0, \Sigma_c)$, with $\Sigma_c = \Sigma/a$.

This means that vectors (μ, τ, Σ) and $(\mu, a\tau, \Sigma/a)$ identify the same SMN model. In this way, we need to impose a restriction to identify

$$\boldsymbol{\Sigma} = \frac{n\boldsymbol{\Sigma}}{\text{tr}(\boldsymbol{\Sigma})}, \tag{2.26}$$

where $tr(\Sigma)$ is the trace of the scatter matrix, i.e., the sum of the main diagonal of this matrix. In this case, we impose the restriction that $\text{tr}(\boldsymbol{\Sigma}) = n$.

2.5 Joint Modeling Models

Within the longitudinal data approach, we have seen that the scatter matrix plays a very important role in the modeling process, since the subjects are not independent, it is likely that a good modeling approach for this matrix will lead to an improvement in the statistical inference of the covariance structure (and the mean of interest).

However, modeling the covariance structure is challenging because the estimated covariance matrices must generally be positive definite and there are many parameters in covariance matrices. To overcome these two obstacles, (POURAHMADI, 1999b) advocated a data-driven joint mean-covariance modeling method based on a modified Cholesky decomposition (MCD) of the within-subject marginal scatter matrix, and developed what the literature calls the NJMM. This decomposition leads to a reparameterization where the inputs can be interpreted in terms of innovation variances and autoregressive coefficients.

Taking into account that the normality assumption is easily violated in the presence of outliers, (LIN; WANG, 2009) proposed the TJMM model, which takes into account modeling using the multivariate Student-t distribution considering the MCD decomposition. More recently, (MAADOOLIAT *et al.*, 2013) proposed modeling using the ACD decomposition. A computationally efficient Fisher scoring algorithm was developed to perform maximum likelihood estimation and a technique for predicting future responses was investigated.

Aiming at greater robustness, (GUNEY *et al.*, 2022) proposed a model using the Laplace distribution, which is a heavy-tailed multivariate distribution with the same number of parameters as the multivariate normal distribution. To simplify the estimation procedure, the authors proposed a modified Cholesky decomposition to factorize the dependence structure in terms of scale innovation and unrestricted autoregressive parameters.

It is worth noting that the assumption of normality in mixed-effects models has been widely debated in the statistical literature, particularly in the presence of outliers or heavy-tailed data. An alternative that has gained attention is the use of the multivariate t-distribution, which provides greater robustness to deviations from normality while maintaining a flexible correlation structure (LANGE *et al.*, 1989; PINHEIRO *et al.*, 2001). The adoption of this distribution in linear mixed-effects models has been explored in several contexts, including robust estimation procedures and alternative formulations to deal with complex data structures.

(PINHEIRO *et al.*, 2001) proposed an efficient algorithm to estimate parameters in linear mixed-effects models under the multivariate t-distribution. His work focuses on robust estimation techniques that leverage the heavier tails of the t-distribution to mitigate the influence of extreme observations. Unlike other approaches that directly modify the covariance structure or apply data transformations, their method retains the fundamental mixed-effects structure while improving the robustness of inference.

In addition to the contribution of Pinheiro et al., several other studies have extended the use of the multivariate t-distribution to different statistical models. For example, (LANGE *et al.*, 1989) developed robust statistical procedures using the t-distribution, highlighting its advantages over normal models in the presence of non-normal errors. (MORRIS; LOCK, 2006) applied the t-distribution to hierarchical models, demonstrating its effectiveness in modeling clustered data with heavy-tailed residuals. Furthermore, recent developments in Bayesian statistics have also incorporated multivariate t-models to improve posterior inference in mixed-effects models (GELMAN *et al.*, 2013).

These advances highlight the growing recognition of the multivariate t-distribution as a powerful alternative to normality assumptions, particularly in robust statistical modeling. Given its ability to effectively handle outliers and model heavy-tailed data, it serves as an essential tool in applied longitudinal and hierarchical modeling frameworks.

2.5.1 NJMM model

The proposal by (POURAHMADI, 1999a) was the one that initiated the modeling of longitudinal data considering the data-driven joint mean-covariance modeling method based on a Cholesky decomposition, specifically the Modified Cholesky decomposition (MCD). The decomposition leads to a reparameterization where the inputs can be interpreted in terms of innovation variances and autoregressive coefficients.

This proposal therefore considers the joint modeling of a general linear model with errors following a multivariate normal distribution, i.e., assuming that $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_i^{-1} = \mathbf{L}_i^\top \mathbf{D}_i^{-1} \mathbf{L}_i$. Considering this structure, the author presented the estimation of the parameters μ and the parameters of the scattering matrix from the maximum likelihood method, whose score functions are given by

$$\mathbf{U}_\beta = \sum_{i=1}^{m} \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}), \tag{2.27}$$

$$\mathbf{U}_\gamma = \sum_{i=1}^{m} \mathbf{Z}_i^\top \mathbf{D}_i^{-1} (\mathbf{r}_i - \mathbf{Z}_i \boldsymbol{\gamma}), \tag{2.28}$$

$$\mathbf{U}_\lambda = -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{w}_j + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sigma_j^{-2} (r_{ij} - \hat{r}_{ij})^2 \mathbf{w}_j, \tag{2.29}$$

where $\hat{\mathbf{r}}_i = \sum_{k=1}^{j-1} r_{ik} \mathbf{z}_{jk}^\top \boldsymbol{\gamma}$, with the $n_i \times d$ matrix $\mathbf{Z}_i$, given by

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}(i,1) & \cdots & \mathbf{z}(i,n_i) \end{bmatrix}^\top, \tag{2.30}$$

with $\mathbf{z}(i,j) = \sum_{k=1}^{j-1} r_{ik} \mathbf{z}_{jk}$, considering that $\hat{\mathbf{r}}_i = \mathbf{Z}_i \boldsymbol{\gamma}$.

Next, (LIN; WANG, 2009) proposes a doubly robust procedure to model the correlation matrix of a longitudinal data set. By using the ACD framework instead of the MCD, the authors claim that the model is robust with respect to model misspecification for the innovation variances in the matrix $\mathbf{D}$ and is also robust with respect to the presence of outliers. The Fisher scoring algorithm is presented to compute the maximum likelihood estimator of the parameters of interest.

2.5.2 TJMM model

NJMM models have advantages because they assume errors with a normal distribution, including convenience of statistical interpretation and computational ease in parameter estimation. However, this assumption of normality for the error terms can be questionable in many practical situations when there are outliers or the underlying data exhibit fat tails.

Several authors in the literature have used a heavy-tailed distribution, such as the multivariate t-distribution, instead of the normal distribution for robust estimation of general linear models (but robustness depends on fixing the degrees of freedom, since only in this case the influence function is limited; when ν is estimated, the influence may become limited, compromising the resistance to outliers (HUBER; RONCHETTI, 2009; LANGE *et al.*, 1989).). (LIN; WANG, 2009), for example, considers the joint modeling of mean and covariance in the general linear model with errors following a multivariate Student-t distribution. The score functions for estimating the parameters of this structure are presented below.

$$\mathbf{U}_{\beta} = \sum_{i=1}^{m} \alpha_i \mathbf{X}_i^{\top} \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}), \tag{2.31}$$

$$\mathbf{U}_{\gamma} = \sum_{i=1}^{m} \alpha_i \mathbf{Z}_i^{\top} \mathbf{D}_i^{-1} (\mathbf{r}_i - \mathbf{Z}_i \boldsymbol{\gamma}), \tag{2.32}$$

$$\mathbf{U}_{\lambda} = -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{w}_j + \frac{1}{2} \sum_{i=1}^{m} \left(\alpha_i \sum_{i=1}^{m} \sigma_j^{-2} (r_{ij} - \hat{r}_{ij})^2 \mathbf{w}_j \right), \tag{2.33}$$

$$\begin{aligned} \mathbf{U}_{\nu} &= \frac{1}{2} \sum_{i=1}^{m} \phi \left(\frac{\nu + n_i}{2} \right) - \phi \left(\frac{\nu}{2} \right) - \frac{n_i}{\nu} \\ &- \log \left(1 + \frac{\Delta_i}{\nu} \right) + \log \frac{\alpha_i}{\nu} \Delta_i, \end{aligned} \tag{2.34}$$

where $\alpha_i = \dfrac{\nu + n_i}{\nu + \Delta_i}$, $\Delta_i = (\mathbf{y}_i - \mathbf{X}_i)^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \mathbf{X}_i)$, $\mathbf{D}_i = \{\sigma_1^2, ..., \sigma_{n_1}^2\}$, $n = \sum_{i=1}^{m} n_i$ and $\hat{r}_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$ are the $m_i \times (d+1)$ matrix of covariates and the $m_i \times 1$ vector of squared fitted errors respectively, and 1_{n_i} is the $n_i \times 1$ vector of 1's.

2.5.3 LJMM model

Similar to the Student-t distribution, the multivariate Laplace distribution is an alternative heavy-tailed distribution to the multivariate normal distribution. However, the Laplace distribution has an advantage over the multivariate Student-t distribution regarding the number of parameters to be estimated. As a heavy-tailed distribution, the multivariate Laplace distribution

has the same number of parameters as the multivariate normal distribution. This feature provides great convenience to the Laplace distribution in computational aspects.

On the other hand, since the multivariate Laplace distribution is heavy-tailed unlike the multivariate normal distribution, it will provide robust estimators that are less sensitive to outliers and heavy-tailed features of datasets. In this context, (GUNEY *et al.*, 2022), considers the joint modeling of mean and covariance in the general linear model with errors following a multivariate Laplace distribution, i.e., considering that $\boldsymbol{\epsilon} \sim L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The score functions for estimating the parameters of this structure are presented below.

$$\mathbf{U}_{\boldsymbol{\beta}} = \frac{1}{2}\sum_{i=1}^{m} \Delta_i^{-1/2} \mathbf{X}_i^{\top} \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}), \tag{2.35}$$

$$\mathbf{U}_{\boldsymbol{\gamma}} = \frac{1}{2}\sum_{i=1}^{m} \Delta_i^{-1/2} \mathbf{Z}_i^{\top} \mathbf{D}_i^{-1} (\mathbf{r}_i - \mathbf{Z}_i \boldsymbol{\gamma}), \tag{2.36}$$

$$\mathbf{U}_{\boldsymbol{\lambda}} = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{w}_j + \frac{1}{4}\sum_{i=1}^{m} \Delta_i^{-1/2} \sum_{j=1}^{m} \sigma_j^{-2} (r_{ij} - \hat{r}_{ij})^2 \mathbf{w}_j, \tag{2.37}$$

where $\Delta_i = (\mathbf{y}_i - \mathbf{X}_i)^{\top} \Sigma^{-1} (\mathbf{y}_i - \mathbf{X}_i)$, $\mathbf{D}_i = \{\sigma_1^2, ..., \sigma_{n_1}^2\}$, $n = \sum_{i=1}^{m} n_i$ and $\hat{r}_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$ are the $m_i \times (d+1)$ matrix of covariates and the $m_i \times 1$ vector of squared fitted errors respectivaly, and 1_{n_i} is the $n_i \times 1$ vector of 1's.

2.6 Conclusion

In this chapter, we present the context and motivation for using longitudinal data, showing the differences between the approaches to this type of study. Next, we present the methodology that will be used to model the mean and the scatter matrix for the proposed model, focusing on the use of two main decompositions for the scatter matrix (MCD and ACD). Finally, we present the family of distributions that will be considered for the assumption of the model's errors. In the following chapter, we will present the two proposed model structures for longitudinal data with the assumption that the errors follow a SMN distribution.

It is important to mention that although in this context, the scatter matrix Σ is unique for all individuals, in our approach, it will be unique for each individual. This becomes, therefore, a reason why a restriction must be imposed on the shape of the scatter matrix. Otherwise, we will not be able to estimate the parameters. More details about the estimation process and the structure of the scatter matrix will be presented during the next chapters.

3 JOINT MODELING MODELS USING THE SCALE MIXTURE OF NORMAL FAMILY

In this chapter, we will present the the main content of this thesis, which is the development of the joint modeling approach using the family of scale mixture of normal distributions, including the estimation method, estimation algorithm and forward prediction. We end this chapter with a discussion of the different models.

3.1 Our proposal

In summary, what differentiates all these works is the assumption that is made about the distribution of errors $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ (where $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, in the context of regression models). Thus, if we consider that $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \blacksquare_i)$, we have the NJMM model, as presented in (POURAHMADI, 1999b). If we consider that $\mathbf{r}_i \sim$ Student-t$(\mathbf{0}, \blacksquare_i, v)$, where v are the degrees of freedom, we have the TJMM model, as presented in (LIN; WANG, 2009). And if we consider that $\mathbf{r}_i \sim$ Laplace$(\mathbf{0}, \blacksquare_i)$, we have structures the LJMM model, as presented in (GUNEY *et al.*, 2022).

Considering that all these structures belong to the SMN family of distributions, the objective of this thesis is to generalize these works that consider the normal, Student-t and Laplace distributions for the errors assuming that $\mathbf{r}_i \sim$ SMN$(\mathbf{0}, \blacksquare_i, \tau_i)$. For this purpose, we will assume that the parameters τ_i are deterministic but unknown. By not imposing any explicit structure on their probability density function , the proposed regression model will be robust, since it has the ability to represent a wide range of distributions belonging to the SMN family, including normal, Student-t and Laplace structures as special cases.

However, note that the SMN model presented in equations (2.23) to (2.25) without any restrictions on the scatter matrix (except the trace) cannot be directly employed for the regression of longitudinal data. With this model, the number of parameters to be estimated is much larger than the total number of observations. The distribution parameters for this regression model cannot be estimated. A restriction on the scatter matrix needs to be imposed on the shape of the scatter matrix.

3.2 Modified Cholesky Decomposition

In the following section, we present the proposed model, linking the SMN distribution to the regression model for longitudinal data, considering the reparameterization for the scatter matrix using the MCD decomposition. The main contributions of this section are detailed below.

- Parameter estimates obtained from the Maximum Likelihood method, considering the **SMN distribution with robust estimators** for the errors and the **MCD structure** for the scatter matrix
- Section originated from **paper**: Ribeiro, V. S. O., Bombrun, L., Nobre, J. S., Cavalcante, C. C., & Berthoumieu, Y. (2024). A general robust approach for joint modeling of the family of scale mixture of normal distribution. Signal Processing, 220, 109426

3.2.1 Statistical model

Considering the stochastic representation defined in (2.16) and assuming that $\mathbf{r}_i|\tau_i \sim \mathcal{N}(\mathbf{0}, \tau_i\boldsymbol{\Sigma}_i)$, then the log-likelihood function for a sample of m individuals $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^\top$, for $i = 1, 2, ..., m$, with $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\tau}^\top, \boldsymbol{\Sigma})^\top$ is given by

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{1}{2}\log(2\pi)\sum_{i=1}^{m} n_i - \frac{1}{2}\sum_{i=1}^{m} n_i \log\tau_i - \frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| \\ &- \sum_{i=1}^{m}\frac{1}{2\tau_i}(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i). \end{aligned} \tag{3.1}$$

Assuming a regression structure with repeated measures, consider $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^\top$ the result vector of i-th sample subject, where n_i is the number of results of the subject i, $i = 1, 2, ..., m$. Here, we allow dependence (correlation) within the subject, but impose independence between subjects. Note that we defined $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\tau}^\top, \boldsymbol{\Sigma})$, but since the scatter matrix depends on the parameters λ and γ we can also define $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\tau}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top)^\top$ throughout the writing of this text.

3.2.2 Maximum likelihood estimation

In this section, we will present the maximum likelihood estimates (MLE) of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\tau}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top)^\top$ of the proposed model, assuming that we have independent observations of m subjects $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m$, considering the log-likelihood function given in (3.1) and a two-step estimation process, given that we assume that $\boldsymbol{\Sigma}_i$ is known to obtain the

expressions of the other estimates (all details about how the expressions in this section were obtained can be seen in the Appendix).

Taking the first derivative of (3.1), the maximum likelihood estimating equations for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\tau}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top)^\top$ of the proposed structure are given by

$$\mathbf{U}_\tau = -\frac{n_i}{2\tau_i} + \frac{1}{2\tau_i^2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{3.2}$$

$$\mathbf{U}_\beta = \sum_{i=1}^{m} \frac{1}{\tau_i}\mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{3.3}$$

$$\mathbf{U}_\gamma = \sum_{i=1}^{m} \frac{1}{\tau_i}\mathbf{Z}_i^\top \mathbf{D}_i^{-1}(\mathbf{r}_i - \mathbf{Z}_i\boldsymbol{\gamma}), \tag{3.4}$$

$$\mathbf{U}_\lambda = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{w}_j + \frac{1}{2}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{j=1}^{n_i} \sigma_j^{-2}(r_{ij} - \hat{r}_{ij})^2 \mathbf{w}_j, \tag{3.5}$$

where $\hat{\mathbf{r}}_i = \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{jk}^\top \boldsymbol{\gamma}$, with the $n_i \times d$ matrix $\mathbf{Z}_i$, given by

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}(i,1) & \cdots & \mathbf{z}(i,n_i) \end{bmatrix}^\top, \tag{3.6}$$

with $\mathbf{z}(i,j) = \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{jk}$, considering that $\hat{\mathbf{r}}_i = \mathbf{Z}_i\boldsymbol{\gamma}$.

Equating the scoring functions $\mathbf{U}_\tau$, $\mathbf{U}_\beta$ and $\mathbf{U}_\gamma$ to zero and rearranging them, we can write the following update equations

$$\hat{\tau}_i^{(h+1)} = \frac{(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(h)})^\top {\boldsymbol{\Sigma}_i^{(h)}}^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(h)})}{n_i} \tag{3.7}$$

$$\hat{\boldsymbol{\beta}}^{(h+1)} = \left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i^{(h)}}\mathbf{X}_i^\top {\boldsymbol{\Sigma}_i^{(h)}}^{-1}\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i^{(h)}}\mathbf{X}_i^\top {\boldsymbol{\Sigma}_i^{(h)}}^{-1}\mathbf{y}_i\right] \tag{3.8}$$

$$\hat{\boldsymbol{\gamma}}^{(h+1)} = \left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i^{(h)}}\hat{\mathbf{Z}}_i^{(h)^\top}\hat{\mathbf{D}}_i^{(h)^{-1}}\hat{\mathbf{Z}}_i^{(h)}\right]^{-1}\left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i^{(h)}}\hat{\mathbf{Z}}_i^{(h)^\top}\hat{\mathbf{D}}_i^{(h)^{-1}}\mathbf{r}_i^{(h)}\right], \tag{3.9}$$

where superscript (h) is the notation to indicate the parameter estimation in step h of the iterative estimation process.

The score function of the parameter $\boldsymbol{\lambda}$ depends on the parameter of interest. Thus, we use the Newton-Raphson method as an iterative estimation method, whose expression for the case of the $\boldsymbol{\lambda}$ parameter is given by

$$\hat{\boldsymbol{\lambda}}^{(h+1)} = \hat{\boldsymbol{\lambda}}^{(h)} + \left(\mathbf{H}_{\lambda\lambda}^{(h)}\right)^{-1}\mathbf{U}_\lambda^{(h)}, \tag{3.10}$$

where $\mathbf{U}_{\lambda}^{(h)}$ is a $d \times 1$ vector given the expression presented in (3.5) and $\mathbf{H}_{\lambda\lambda}^{(h)}$ is a $d \times d$ matrix given by

$$\mathbf{H}_{\lambda\lambda} = \frac{\partial^2}{\partial \lambda^2} l(\boldsymbol{\theta}) \quad = \quad -\frac{1}{2}\sum_{i=1}^{m} \frac{1}{\tau_i} \left(\sum_{j=1}^{n_i} \sigma_j^{-2} (r_{ij} - \hat{r}_{ij})^2 \mathbf{w}_j \mathbf{w}_j^{\top} \right). \tag{3.11}$$

To summarize, the steps to estimate the parameters of the proposed SMNJMM model with deterministic but unknown τ_i are presented in Algorithm 1.

Algorithm 1: Algorithm for the maximum likelihood estimate of the proposed SMN-JMM model

[1] Given a starting value $\boldsymbol{\theta}^{(h)} = (\boldsymbol{\beta}^{(h)}, \boldsymbol{\tau}^{(h)}, \boldsymbol{\gamma}^{(h)}, \boldsymbol{\lambda}^{(h)})^{\top}$, which can be obtained by fitting the NJMM model, use the models given in (2.12) to form $\mathbf{Z}_i^{(h)}$, $\hat{\mathbf{r}}_i^{(h)}$, $\mathbf{L}_i^{(h)}$, $\mathbf{D}_i^{(h)}$, and $\boldsymbol{\Sigma}_i^{(h)}$;
Update $\hat{\tau}_i^{(h+1)}$ using (3.7);
Update $\hat{\boldsymbol{\beta}}^{(h+1)}$ using (3.8);
Update $\hat{\boldsymbol{\gamma}}^{(h+1)}$ using (3.9);
Update $\hat{\boldsymbol{\lambda}}^{(h+1)}$ using (3.10);
Update the following matrices: $\mathbf{Z}_i^{(h+1)}$, using (3.6), $\mathbf{L}_i^{(h+1)}$ using (2.10), $\mathbf{D}_i^{(h+1)}$, using (2.4), and $\boldsymbol{\Sigma}_i^{(h+1)}$, using (2.9);
normalization of the scatter matrix, considering $\boldsymbol{\Sigma}_i^{(h+1)} = \frac{n_i \boldsymbol{\Sigma}_i^{(h+1)}}{tr(\boldsymbol{\Sigma}_i^{(h+1)})}$;
Repeat steps 2 to 7 until convergence.

It is important to emphasize that the estimation process of τ can be naturally interpreted in terms of the Mahalanobis distance, as it accounts for the scaled variability and correlation structure of the residuals. Specifically, the update formula for $\hat{\boldsymbol{\gamma}}$ involves a weighting mechanism that inversely depends on $\hat{\tau}_i^{(h)}$, effectively adjusting the contribution of each observation based on its deviation from the model expectations. Since τ governs the dispersion structure, its estimation indirectly measures how far an observation deviates in a multivariate sense, analogous to the Mahalanobis distance (MAHALANOBIS, 1936; HAMPEL *et al.*, 2011).

This interpretation highlights the role of τ in controlling the influence of each observation, reinforcing its connection with robust statistical techniques that aim to downweight extreme deviations while preserving efficiency in parameter estimation (i.e. observations for which $\mathbf{y}_i$ is far from $\mathbf{X}_i\boldsymbol{\beta}$ will have a higher value for the estimate of τ_i and a lower weight for the estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$).

And about the iterative update process for estimating $\boldsymbol{\beta}$, we can interpreted as a reweighted Generalized Least Squares (GLS) estimator, where the weights are adjusted at each

iteration based on the current estimates. This iterative reweighting mechanism naturally aligns with the concept of robustness in statistical estimation, as it reduces the influence of observations that deviate significantly from the assumed model structure. In particular, when heavy-tailed distributions or adaptive weighting schemes are used, the estimation process becomes more resistant to outliers and model misspecifications, mitigating their impact on parameter estimation (HUBER; RONCHETTI, 2009; HAMPEL *et al.*, 2011). This property is crucial in practical applications where deviations from normality are expected, ensuring that the estimator remains stable even in the presence of extreme values.

The SMN model presented in equations (2.23) and (2.24) without any constraint on the scatter matrix (except the trace) and the proposed SMNJMM model share many similarities. Actually, the only difference relies on the structure of the scatter matrix. For this application of regression with longitudinal data, we need to impose this structure. As these models are quite closed each other, we retrieve some similarities during the estimation process. For example:

- Equation (2.23) for the SMN model is retrieve as step 3 in Algorithm 1 (equation (3.8), estimation of the regression coefficients to obtain the mean vector).
- Equation (2.24) for the SMN model (estimation of Σ) is replaced by steps 4 and 5 in Algorithm 1 (equations (3.9) and (3.10) to estimate γ and λ).
- Equation (2.25) for the SMN model (estimation of τ_i) is retrieved by step 2 in Algorithm 1 (equation (3.7)). Note that it is exactly the same equation with the quadratic term on the numerator.
- For identification reason, as explained in the paragraph before, the scatter matrix should be normalized. In the proposed SMNJMM model, we have also considered that its trace is equal to n (see step 7 in Algorithm 1).

3.2.3 Forward prediction

- Development of a technique capable of predicting future values of the i-th subject, which is one of the m individuals under study, assuming that the values already observed and the new predictions follow the multivariate distribution under consideration.

In this section, we present a technique to predict the future values of the individual i ($\mathbf{y}_i$), which is one of the m individuals under consideration, for the proposed model (SMNJMM), by following the same approach as the one developed for the TJMM (LIN; WANG, 2009) and LJMM (GUNEY *et al.*, 2022) models, i.e. formulate an estimation structure capable of predicting

future values of the i-th subject, which is one of the m individuals under study, assuming that the values already observed and the new predictions follow the multivariate distribution under consideration.

For this, consider that $\tilde{\mathbf{y}}_i$ denotes a future vector, $(g \times 1)$, of measurements for the individual $\mathbf{y}_i$, $\tilde{\mathbf{X}}_i$ denotes a matrix, $(g \times (p+1))$, of prediction for the corresponding $\tilde{\mathbf{y}}_i$ with $\mathbb{E}(\tilde{\mathbf{y}}_i) = \tilde{\mathbf{X}}_i\boldsymbol{\beta}$, and $\tilde{\mathbf{z}}_{jk}$ and $\tilde{\mathbf{w}}_j$ denote the vectors of the covariates associated with $\tilde{\mathbf{y}}_i$, of dimension $d \times 1$ and $q \times 1$, being $\ln\sigma_i^2 = \tilde{\mathbf{w}}_i^\top\boldsymbol{\lambda}$ and $\phi_{jk} = \tilde{\mathbf{z}}_{jk}^\top\boldsymbol{\gamma}$, for $j \in \{n_i+1, ..., n_i+g\}$ and $k \in \{1, ..., j-1\}$, respectively.

Furthermore, we assume that the joint distribution of $\mathbf{y}_i$ and $\tilde{\mathbf{y}}_i$ given τ_i follows a (n_i+g)-variate normal distribution, i.e.

$$\begin{bmatrix} \mathbf{y}_i \\ \tilde{\mathbf{y}}_i \end{bmatrix} | \tau_i \sim \mathscr{N}_{n_i+g}(\mathbf{X}_i^*\boldsymbol{\beta}, \tau_i\boldsymbol{\Sigma}_i^*), \tag{3.12}$$

where $\mathbf{X}_i^* = (\mathbf{X}_i^\top, \tilde{\mathbf{X}}_i^\top)^\top$, $\boldsymbol{\Sigma}_i^{*-1} = \mathbf{L}_i^{*-1}\mathbf{D}_i^{*-1}\mathbf{L}_i^*$, $\mathbf{D}_i^* = \text{diag}(\sigma_1^2, ..., \sigma_{n_i+g}^2)$ and $\mathbf{L}_i^* = [l_{jk}]$ are subjectary lower triangular matrices, $(n_i+g) \times (n_i+g)$, with 1's as diagonal entries and (j,k)-th entry being $-\phi_{jk}$.

Partitioning $\boldsymbol{\Sigma}_i^*$ according to the size of $\mathbf{y}_i^* = (\mathbf{y}_i^\top, \tilde{\mathbf{y}}_i^\top)^\top$, which is a vector that joins current observations and future observations, we have

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \tag{3.13}$$

where $\boldsymbol{\Sigma}_{11}$ is the upper left $(n_i \times n_i)$ submatrix of $\boldsymbol{\Sigma}_i^*$ and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$ (here we suppress the subscript i of the partitioned matrices for notational convenience).

Considering the characteristics of the marginal and conditional distributions (CAMBANIS *et al.*, 1981; GÓMEZ *et al.*, 2003), we have that if $\mathbf{y}_i|\tau_i \sim \mathscr{N}_{n_i}(\boldsymbol{\mu}_i, \tau_i\boldsymbol{\Sigma}_i)$, then

$$\tilde{\mathbf{y}}_i|\mathbf{y}_i, \tau_i \sim \mathscr{N}_g\left(\boldsymbol{\mu}_{2\cdot1}, \tau_i\boldsymbol{\Sigma}_{22\cdot1}\right), \tag{3.14}$$

where $\boldsymbol{\mu}_{2\cdot1} = \mathbb{E}(\tilde{\mathbf{y}}_i|\mathbf{y}_i, \tau_i)$ and $\boldsymbol{\Sigma}_{22\cdot1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, and the Mean Squared Error (MSE) error predictor of $\tilde{\mathbf{y}}_i$, which is the conditional expectation of $\tilde{\mathbf{y}}_i$ given $\mathbf{y}_i$ and τ_i, as shown by (GUNEY *et al.*, 2022), is given by

$$\hat{\tilde{\mathbf{y}}}_i = \boldsymbol{\mu}_{2\cdot1} = \tilde{\mathbf{X}}_i\boldsymbol{\beta} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{3.15}$$

and the MSE scatter matrix for the predictor (3.15) is determined by

$$\mathbb{E}[(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)^\top|\tau_i] = \mathbb{E}(\text{cov}(\tilde{\mathbf{y}}_i|\mathbf{y}_i)|\tau_i) = \tau_i\boldsymbol{\Sigma}_{22\cdot1}. \tag{3.16}$$

This result is well known in the literature: the best predictor in the sense of minimizing the Mean Squared Error (MSE) is given by the conditional expectation. In the case of mixed linear models, this conditional expectation coincides with the Best Linear Unbiased Predictor (BLUP), a concept originally developed by (HENDERSON, 1975). This result is directly related to the Gauss-Markov Theorem, which guarantees that the Generalized Least Squares (GLS) estimator is the best unbiased linear estimator under the usual assumptions (AITKEN, 1936).

Furthermore, the BLUP has an interpretation within the Bayesian approach, being equivalent to a Bayes linear estimator when we assume random effects with normal prior distributions and a hierarchical model (GOLDSTEIN, 1995; ROBINSON, 1991). This equivalence between frequentist and Bayesian methods reinforces the importance of BLUP and its wide applicability in statistical modeling, especially in areas such as quantitative genetics and hierarchical inference.

In the next section, we will follow the same sequence of explanation, but considering the Alternative Cholesky Decomposition framework for the modeling process.

3.3 Alternative Cholesky Decomposition

In the following section, we present the proposed model, linking the SMN distribution to the regression model for longitudinal data, considering the reparameterization for the scatter matrix using the ACD decomposition. The main contributions of this section are detailed below.

- Parameter estimates obtained from the Maximum Likelihood method, considering the **SMN distribution with robust estimators** for the errors and the **ACD structure** for the scatter matrix
- Section originated from **paper**: Ribeiro, V. S. O., Bombrun, L., Nobre, J. S., Cavalcante, C. C., & Berthoumieu, Y. (2025). Joint modeling approach with Alternative Cholesky Decomposition using the family of scale mixture of normal distribution.

3.3.1 Statistical model

Assuming that $\boldsymbol{\epsilon}_i|\tau_i \sim \mathcal{N}(\mathbf{0}, \tau_i\boldsymbol{\Sigma}_i)$ and the stochastic representation defined in (2.16), we obtain the log-likelihood function for a sample of m individuals $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^\top$, for

$i = 1,2,...,m$, with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\Sigma})$ given by

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{1}{2}\ln(2\pi)\sum_{i=1}^{m} n_i - \frac{1}{2}\sum_{i=1}^{m} n_i \ln \tau_i - \frac{1}{2}\sum_{i=1}^{m} \ln|\boldsymbol{\Sigma}_i| \\ &- \sum_{i=1}^{m} \frac{1}{2\tau_i}(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i). \end{aligned} \tag{3.17}$$

3.3.2 Maximum likelihood estimation

In this section, we introduce an iterative algorithm to estimate by maximum likelihood the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\Sigma})$ of the proposed joint model. For that, the score functions are first derived and then the Fisher scoring algorithm is introduced as detailed in the next two subsections.

By taking the first derivative of (3.17), the score functions for each parameter are given by

$$\mathbf{U}_{\boldsymbol{\beta}} = \sum_{i=1}^{m} \frac{1}{\tau_i}\mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) \tag{3.18}$$

$$\mathbf{U}_{\gamma_r} = \sum_{i=1}^{m} \frac{1}{\tau_i}\left(\mathrm{tr}\left((\mathbf{D}_i\mathbf{L}_i)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top (\mathbf{D}_i\mathbf{L}_i)^{-\top}\mathbf{L}_i^{-1}\frac{\partial \mathbf{L}_i}{\partial \gamma_r}\right)\right) \tag{3.19}$$

$$\mathbf{U}_{\lambda_s} = \sum_{i=1}^{m} \mathrm{tr}\left(\left(\frac{1}{\tau_i}(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} - \mathbf{I}_i\right)\mathbf{D}_i^{-1}\frac{\partial \mathbf{D}_i}{\partial \lambda_s}\right) \tag{3.20}$$

$$\mathbf{U}_{\tau_i} = -\frac{n_i}{2\tau_i} + \frac{1}{2\tau_i^2}(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i, \tag{3.21}$$

where $\dfrac{\partial \mathbf{L}_i}{\partial \gamma_r}$ is the derivative of matrix $\mathbf{L}_i$ with respect to r^{st} element of vector $\boldsymbol{\gamma}$ and $\dfrac{\partial \mathbf{D}_i}{\partial \lambda_s}$ is the derivative of matrix $\mathbf{D}_i$ with respect to s^{st} element of vector $\boldsymbol{\lambda}$. Equalling the score function (3.18) to zero, we obtain the maximum likelihood estimator (MLE) of the regression coefficients as:

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i}\mathbf{X}_i^\top \hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i}\mathbf{X}_i^\top \hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{y}_i\right], \tag{3.22}$$

where $\hat{\tau}_i$ is the MLE of τ_i, i.e the result when equating the equation (3.21) to zero

$$\hat{\tau}_i = \frac{(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^\top \hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)}{n_i}, \tag{3.23}$$

and $\hat{\boldsymbol{\Sigma}}_i$ is the MLE of $\boldsymbol{\Sigma}_i$, whose components will be estimated from the Fisher information matrix, given by

$$\mathbf{I}_{\gamma_r\gamma_s} = \sum_{i=1}^{m} \frac{1}{\tau_i}\text{tr}\left(\frac{\partial \mathbf{L}_i}{\partial \gamma_r}\frac{\partial \mathbf{L}_i}{\partial \gamma_s}\mathbf{L}_i^{-\top}\mathbf{L}_i^{-1}\right) \tag{3.24}$$

$$\mathbf{I}_{\lambda_r\lambda_s} = \sum_{i=1}^{m} \frac{1}{\tau_i}\text{tr}\left(\mathbf{L}_i\mathbf{L}_i^{\top}\frac{\partial \mathbf{D}_i}{\partial \lambda_r}\boldsymbol{\Sigma}_i^{-1}\frac{\partial \mathbf{D}_i}{\partial \lambda_s} + \mathbf{D}_i^{-2}\frac{\partial \mathbf{D}_i}{\partial \lambda_r}\frac{\partial \mathbf{D}_i}{\partial \lambda_s}\right) \tag{3.25}$$

$$\mathbf{I}_{\gamma_r\lambda_s} = \mathbf{I}_{\lambda_s\gamma_r} = \sum_{i=1}^{m} \frac{1}{\tau_i}\text{tr}\left(\mathbf{D}_i\frac{\partial \mathbf{L}_i}{\partial \gamma_r}\mathbf{L}_i^{\top}\frac{\partial \mathbf{D}_i}{\partial \lambda_s}\boldsymbol{\Sigma}_i^{-1}\right). \tag{3.26}$$

From the closed expressions of $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}_i$, we were able to obtain the MLE in a simple way for these two parameters as given in (3.18) and (3.23). Unfortunately, to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, there is no close form expression. We recommend using the Fisher scoring algorithm, as described by (FISHER, 1925), by applying:

$$\hat{\boldsymbol{\eta}}^{(h+1)} = \hat{\boldsymbol{\eta}}^{(h)} + \left(\mathbf{I}_{\hat{\boldsymbol{\eta}}^{(h)}}\right)^{-1}\mathbf{U}_{\hat{\boldsymbol{\eta}}^{(h)}}, \tag{3.27}$$

where $\hat{\boldsymbol{\eta}}^{(h)} = \left(\hat{\boldsymbol{\gamma}}^{(h)}, \hat{\boldsymbol{\lambda}}^{(h)}\right)^{\top} = \left(\hat{\gamma}_1^{(h)}, \ldots, \hat{\gamma}_d^{(h)}, \hat{\lambda}_1^{(h)}, \ldots, \hat{\lambda}_q^{(h)}\right)^{\top}$ are the estimates at step h and

$$\mathbf{U}_{\hat{\boldsymbol{\eta}}^{(h)}} = \begin{pmatrix} \mathbf{U}_{\hat{\boldsymbol{\gamma}}^{(h)}} \\ \mathbf{U}_{\hat{\boldsymbol{\lambda}}^{(h)}} \end{pmatrix} \quad \text{and} \quad \mathbf{I}_{\hat{\boldsymbol{\eta}}^{(h)}} = \begin{pmatrix} \mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\gamma}}^{(h)}} & \mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} \\ \mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} & \mathbf{I}_{\hat{\boldsymbol{\lambda}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} \end{pmatrix}. \tag{3.28}$$

In addition, the score vectors are respectively defined as:

$$\mathbf{U}_{\hat{\boldsymbol{\gamma}}^{(h)}} = \left(\mathbf{U}_{\hat{\gamma}_1^{(h)}}, \ldots, \mathbf{U}_{\hat{\gamma}_d^{(h)}}\right)^{\top} \quad \text{and} \quad \mathbf{U}_{\hat{\boldsymbol{\lambda}}^{(h)}} = \left(\mathbf{U}_{\hat{\lambda}_1^{(h)}}, \ldots, \mathbf{U}_{\hat{\lambda}_q^{(h)}}\right)^{\top}, \tag{3.29}$$

and the elements of the Fisher Information matrix are given by:

$$\mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\gamma}}^{(h)}} = \begin{bmatrix} \mathbf{I}_{\hat{\gamma}_1^{(h)}\hat{\gamma}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\gamma}_1^{(h)}\hat{\gamma}_d^{(h)}} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{\hat{\gamma}_d^{(h)}\hat{\gamma}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\gamma}_d^{(h)}\hat{\gamma}_d^{(h)}} \end{bmatrix}, \tag{3.30}$$

$$\mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} = \begin{bmatrix} \mathbf{I}_{\hat{\gamma}_1^{(h)}\hat{\lambda}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\gamma}_1^{(h)}\hat{\lambda}_q^{(h)}} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{\hat{\gamma}_d^{(h)}\hat{\lambda}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\gamma}_d^{(h)}\hat{\lambda}_q^{(h)}} \end{bmatrix}, \tag{3.31}$$

$$\mathbf{I}_{\hat{\boldsymbol{\lambda}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} = \begin{bmatrix} \mathbf{I}_{\hat{\lambda}_1^{(h)}\hat{\lambda}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\lambda}_1^{(h)}\hat{\lambda}_q^{(h)}} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{\hat{\lambda}_q^{(h)}\hat{\lambda}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\lambda}_q^{(h)}\hat{\lambda}_q^{(h)}} \end{bmatrix}. \tag{3.32}$$

Algorithm 2: Algorithm for the maximum likelihood estimate of the proposed SMN-JMM model

[1] Given a starting value $\boldsymbol{\theta}^{(h)} = (\boldsymbol{\beta}^{(h)}, \boldsymbol{\tau}^{(h)}, \boldsymbol{\gamma}^{(h)}, \boldsymbol{\lambda}^{(h)})^\top$, which can be obtained by fitting the NJMM model, to form $\mathbf{L}_i^{(h)}$, $\mathbf{D}_i^{(h)}$, and $\boldsymbol{\Sigma}_i^{(h)}$;
Update $\hat{\boldsymbol{\tau}}_i^{(h+1)}$ using (3.23)
Update $\hat{\boldsymbol{\beta}}^{(h+1)}$ using (3.22)
Update $\hat{\boldsymbol{\gamma}}^{(h+1)}$ and $\hat{\boldsymbol{\lambda}}^{(h+1)}$ using Fisher Scoring, presented in (3.27)
Update $\hat{\boldsymbol{\Sigma}}_i^{(h+1)}$ using (2.9)
normalization of the scatter matrix, considering $\boldsymbol{\Sigma}_i^{(h+1)} = \frac{n_i \boldsymbol{\Sigma}_i^{(h+1)}}{tr(\boldsymbol{\Sigma}_i^{(h+1)})}$;
Repeat steps 2 to 6 until convergence.

The steps for estimating the parameters of the proposed model with deterministic but unknown τ_i are summarized in Algorithm 2.

As expected and as observed in Algorithm 2, the proposed SMNJMM model behaves as the NJMM when τ_i are the same for each subject i. In that case, the stochastic representation given in (2.16) vanishes to the one for a normally distributed random variable, and equations (3.22) and (3.27) reduces to the one obtained for the NJMM model (CHEN; DUNSON, 2003b). In addition, an interesting property of the proposed SMNJMM model is that it is robust to outliers. Indeed, if the observation $\mathbf{y}_i$ for ith subject has a large deviance from the estimated mean $\hat{\boldsymbol{\mu}}_i$, $\hat{\tau}_i$ will be large as shown in (3.23). And since $\hat{\tau}_i$ is on the denominator of each score functions, this ith subject will have a lower weight in the estimation of the regression coefficients.

The SMNJMM model proposed in this paper shares the same advantages as the one introduced by Ribeiro et al. in (RIBEIRO *et al.*, 2024) and presented in the section 3.2. The only difference relies on the definition of the structure of the scatter matrix. Here, the ACD decomposition is employed while the MCD decomposed was used in (RIBEIRO *et al.*, 2024). To better understand the added-value of each of these models, Section 5 will present a detail comparison between both on real dataset. Before showing these results, the next subsection gives some elements about forward prediction with the proposed SNMJMM model.

3.3.3 Forward prediction

By following the same methodology as done by Ribeiro et al. in (RIBEIRO *et al.*, 2024) for the SMNJMM model with the MCD decomposition, we introduce a technique for predicting the new g observations of subject i with the proposed model but now using the ACD decomposition for the scatter matrix.

We denote $\tilde{\mathbf{y}}_i$ the future vector of measurements of the individual i, which has dimension $(g \times 1)$. $\tilde{\mathbf{X}}_i$ is the $g \times (p+1)$ design matrix for the prediction part. With these notations, we have $\mathbb{E}(\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i) = \tilde{\mathbf{X}}_i\boldsymbol{\beta}$. In addition, $\tilde{\mathbf{z}}_{jk}$ and $\tilde{\mathbf{w}}_j$ are the covariates vectors of dimension $d \times 1$ and $q \times 1$, where $\ln \sigma_i^2 = \tilde{\mathbf{w}}_i^\top \boldsymbol{\lambda}$ and $\phi_{jk} = \tilde{\mathbf{z}}_{jk}^\top \boldsymbol{\gamma}$, for $j \in \{n_i+1, ..., n_i+g\}$ and $k \in \{1, ..., j-1\}$, respectively. For the prediction, we assume that the joint distribution of $\mathbf{y}_i$ and $\tilde{\mathbf{y}}_i$ given τ_i follows a multivariate normal distribution. It yields:

$$\begin{bmatrix} \mathbf{y}_i \\ \tilde{\mathbf{y}}_i \end{bmatrix} |\tau_i \sim \mathscr{N}_{n_i+g}(\mathbf{X}_i^*\boldsymbol{\beta}, \tau_i\boldsymbol{\Sigma}_i^*), \tag{3.33}$$

where $\mathbf{X}_i^* = (\mathbf{X}_i^\top, \tilde{\mathbf{X}}_i^\top)^\top$, $\boldsymbol{\Sigma}_i^{*-1} = \mathbf{D}_i^{*-1}\mathbf{L}_i^{*-\top}\mathbf{L}_i^{*-1}\mathbf{D}_i^{*-1}$, $\mathbf{D}_i^* = \text{diag}(\sigma_1^2, ..., \sigma_{n_i+g}^2)$ and $\mathbf{L}_i^* = [l_{jk}]$ is an $(n_i+g) \times (n_i+g)$ lower triangular matrix as defined in (2.10). The scatter matrix $\boldsymbol{\Sigma}_i^*$ can be decomposed as:

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \tag{3.34}$$

where $\boldsymbol{\Sigma}_{11}$ is the upper left $(n_i \times n_i)$ submatrix of $\boldsymbol{\Sigma}_i^*$ corresponding to the scatter matrix of the n_i observed data. $\boldsymbol{\Sigma}_{22}$ is the $g \times g$ scatter matrix for the future observations, and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$. Note that to simplify notations, the subscript i on partitioned matrices is removed but all these matrix depend on subject i.

By using properties of marginal and conditional distributions shown in (CAMBANIS *et al.*, 1981; GÓMEZ *et al.*, 2003) for elliptical distributions which extends scale mixture of normal distributions, we know that if $\mathbf{y}_i|\tau_i \sim \mathscr{N}_{n_i}(\boldsymbol{\mu}_i, \tau_i\boldsymbol{\Sigma}_i)$, then

$$\tilde{\mathbf{y}}_i|\mathbf{y}_i, \tau_i \sim \mathscr{N}_g\left(\boldsymbol{\mu}_{2\cdot1}, \tau_i\boldsymbol{\Sigma}_{22\cdot1}\right), \tag{3.35}$$

where $\boldsymbol{\mu}_{2\cdot1} = \mathbb{E}(\tilde{\mathbf{y}}_i|\mathbf{y}_i, \tau_i)$ and $\boldsymbol{\Sigma}_{22\cdot1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. With (3.35), it yields that the Mean Squared Error (MSE) predictor of $\tilde{\mathbf{y}}_i$ is given by (GUNEY *et al.*, 2022)

$$\hat{\tilde{\mathbf{y}}}_i = \boldsymbol{\mu}_{2\cdot1} = \tilde{\mathbf{X}}_i\boldsymbol{\beta} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{3.36}$$

and the MSE scatter matrix for the predictor (3.36) is determined by

$$\mathbb{E}[(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)^\top|\tau_i] = \mathbb{E}(\text{cov}(\tilde{\mathbf{y}}_i|\mathbf{y}_i)|\tau_i) = \tau_i\boldsymbol{\Sigma}_{22\cdot1}. \tag{3.37}$$

Note that these expressions are the same as the one obtained in (RIBEIRO *et al.*, 2024) for the SMNJMM model with the MCD decomposition. The only difference relies on the way the scatter matrix $\boldsymbol{\Sigma}_i^*$ is parameterized. For the MCD decomposition, $\boldsymbol{\Sigma}_i^*$ is decomposed using (2.2), while here for the proposed ACD decomposition, (2.9) is used.

3.4 Discussion about the models

In both models (MCD and ACD frameworks), we consider the approach that the parameters given by τ_i (which are the mixture parameters of the model) are deterministic but unknown, and we do not assume any specific distribution for them. Both models therefore have the ability to represent any other model issued from the family of scale mixtures of normal distributions.

This framework can be considered as a robust extension of the state-of-the-art models based on the normal, Student-t and Laplace distributions, as we have discussed throughout the text. Furthermore, for a practical application, the interest is that it is no longer necessary to perform the goodness-of-fit to validate the proposed model, since our SMNJMM model generalizes a wide range of distributions.

In addition, we consider the approach of joint estimation of the intra-subject marginal scatter matrix, with the aim of overcoming common problems in these processes, such as the positive definiteness of the scatter matrix and the number of parameters to be estimated.

Evaluating the structure of the covariance matrices, we have that the decompositions were considered, since both present a closer relationship, given that they are constructed in a similar way through the standardization of the Cholesky factor and the resulting unrestricted parameters have a good statistical interpretation in terms of innovation variance, autoregressive parameters and moving average, respectively.

However, due to the decomposition, the resulting correlation function of MCD depends on both the innovation variance and the autoregressive parameters, indicating that MCD is not robust against misspecification of the innovation variance when correlation is the main interest. We also need to note that MCD is more computationally efficient between the two approaches due to the fact that its Fisher information matrix is block diagonal.

3.5 Conclusion

In this chapter, we present the maximum likelihood estimates, the estimation algorithms, and the methods for predicting new observations for the two proposed models, using the MCD and ACD decompositions. In addition, we introduce the NJMM, TJMM, and LJMM models, previously presented in the literature, which served as motivation and basis for this work. Finally, we discuss the comparisons between considering an SMN model with an MCD-type

matrix and an ACD-type matrix. In the next chapters, we will evaluate the behavior of these models in simulated and real data.

4 SYNTHETIC DATA ANALYSIS

In this chapter, we present the model performance results based on simulated (synthetic), for the models using the normal, Student-t and Laplace distributions and for the proposed model, considering the modified and alternative Cholesky decompositions. The results are presented in terms of the standard errors of the estimates and the BIC values adapted for the problem. Forward prediction estimates are also evaluated.

4.1 Estimation performance

- Experiments on synthetic data confirmed the robustness property of the proposed SMN-JMM model compared to the NJMM, TJMM, and LJMM models with both the MCD and ACD structure.
- Hypothesis tests performed to evaluate the results of the different models validated that there is a significant difference between the models that use the Student-t and Laplace distributions compared to the proposed model, validating the efficiency of our model.

4.1.1 Description

Considering the **Modified Cholesky Decomposition** structure we compared the performance of the maximum likelihood estimates of the proposed model with that of recent models. For this, we generate $m = 100$ balanced random samples (with $n_i = 12$ observations for each individual) from the normal, Student-t (with 4 degrees of freedom) and Laplace distributions, with $\beta = [5, 27, 0.16]^\top$, $\gamma = [1.12, -0.58, 0.09, -0.0043]^\top$ and $\lambda = [-0.4, -1.37, 0.18, -0.007]^\top$, that is, $d = 4$, $q = 4$, $p = 1$, considering the design matrix $\mathbf{X}_i$ for the mean, the covariates $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ for autoregressive parameters and chosen scale innovation variances as follows

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{1} & \mathbf{k} \end{bmatrix}^\top, \quad \mathbf{z}_{jk} = \begin{bmatrix} \mathbf{1} & (j-k) & (j-k)^2 & (j-k)^3 \end{bmatrix}^\top \tag{4.1}$$

and

$$\mathbf{w}_j = \begin{bmatrix} 1 & j & j^2 & j^3 \end{bmatrix}^\top, \tag{4.2}$$

where $\mathbf{k} = [0, 1, 2.5, 3.5, 4.5, 6, 7, 8, 10, 11.5, 13, 14]^\top$, following the same simulation structure presented in (LIN; WANG, 2009) and (GUNEY *et al.*, 2022). It is important to say that, as described in the literature for the TJMM and LJMM models, we have that $\mathbf{z}(i, 1) = 0$, so that the

first line of $\mathbf{Z}_i$ is zero. To compute the standard error, the simulations were run with a total of 500 replications.

To evaluate the models, we consider three scenarios: data generated from the normal, from the Student-t and from the Laplace distributions. The results are presented in Tables 1–3 (the best result is highlighted in blue and the second best result is highlighted in red), with the mean values of estimates and standard errors (SE) of each model parameter, where $SE = \sigma/\sqrt{r}$, with σ the standard deviation of the estimated parameter and r the number of replications.

Results for the MCD approach

The first evaluation scenario aims at illustrating that the proposed model is suitable for any distribution chosen for the parameter τ of the SMN model. In this context, we consider that τ has the following distributions: inverse gamma, beta and inverse beta. For this, 100 samples were generated for each scenario, considering that the structure that generate the mixing parameter τ have the same mean.

To compare these models, we consider the use of the BIC metric (Bayesian Information Criterion), proposed by (SCHWARZ, 1978), which is given by

$$\text{BIC} = -\frac{2}{m}\hat{l}_{max} + \frac{k}{m}\ln(m), \tag{4.3}$$

where $\hat{l}_{max}$ is the maximized log-likelihood with respect to the chosen model. For the proposed SMNJMM, the number of parameters k is equal to $p+d+q+m+1$ since $\beta \in \mathbb{R}^{p+1}$, $\gamma \in \mathbb{R}^d$, $\lambda \in \mathbb{R}^q$ and $\tau \in \mathbb{R}^m$. For the NJMM and LJMM models, $k = p+d+q+1$, since we do not have τ_i, and for the TJMM model, $k = p+d+q+2$ since the degree of freedom ν must be additionally estimated.

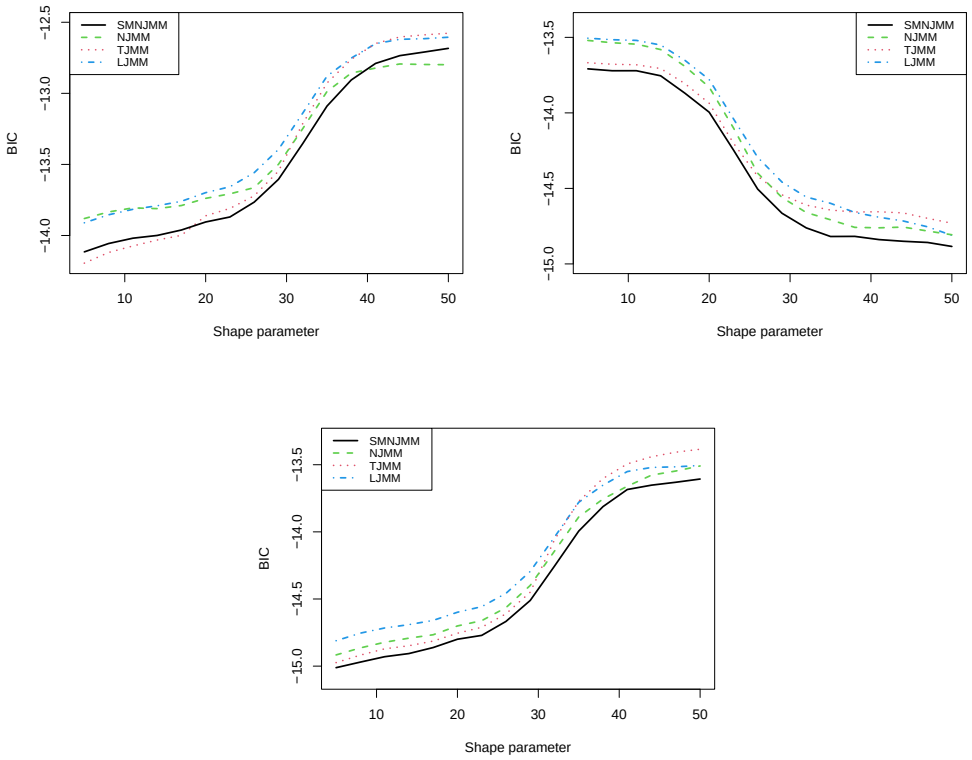


Figure 4 – BIC metric value for each model obtained from the shape parameter. First: inverse gamma distribution, second: beta distribution, third: inverse beta distribution.

The BIC metric is a tradeoff between fit performance and model complexity. It is based on the likelihood function, being related to the Akaike Information Criterion (AIC) and is used as a criterion for selecting models from a finite set of other models. The rule for using this measure is to choose models with a lower BIC (the values being positive or negative).

Figure 4 shows the evolution of the BIC criterion for the three scenario presented before when τ follows an inverse gamma, a beta and an inverse beta distribution.

As observed, the proposed SMNJMM model has generally the lowest BIC value for each of these scenario indicating its great potential. More precisely, for the first scenario when τ follows an inverse gamma distribution, the errors follow the Student-t distributions and unsurprisingly the best performance are obtained for the TJMM model displayed in red in the top left plot of Figure 4 especially when the shape parameter (degree of freedom) has a low value. Note also that in this context, the performances of the proposed SMNJMM in black are very closed to the one obtained with TJMM. For the second and third scenario of Figure 4, τ follows respectively a beta and an inverse beta distribution.

In this case, the errors follow much more complex distributions involving the Whit-

taker functions as presented in (BOMBRUN *et al.*, 2011). For these scenario, the NJMM, TJMM and LJMM models do not work well for low shape parameters. However, as expected since we have designed it for its robust behavior, the proposed SMNJMM model exhibits the best performances. It performs well whatever the considered distribution for τ. Note also that for a large shape parameter, all these models converge towards the normal distribution and in this case, the best performance is naturally obtained with the NJMM model.

According to Table 1, where data are generated from the normal distribution, the model that best fits the data is the NJMM model, followed by the proposed SMNJMM model. We also observed that the standard errors of the estimators are similar between the models, with emphasis on the values of the parameters λ_1, whose smallest standard error was observed in the estimates for the SMNJMM model.

To assess if there is a difference between the results obtained for each model, the pairwise Wilcoxon rank-sum test is employed (MANN; WHITNEY, 1947). We obtained a p-value greater than 5% for comparing BIC values between NJMM and SMNJMM, that is, there is no significant difference between the theoretical model and the proposed one. Furthermore, a p-value of less than 5% is obtained when comparing BIC values between SMNJMM versus TJMM, and SMNJMM versus LJMM models. That is, there is a significant difference between the models that use the Student-t and Laplace distributions, respectively, and the proposed one. These comparisons help us to validate the efficiency and robustness of the proposed model in relation to other models, when the data is generated from the normal distribution structure.

Table 1 – Parameter estimations and Standard Errors (SE) for NJMM, TJMM, LJMM and SMNJMM of the data generated data from normal Distribution - Scenario 1.

Parameter	True value	NJMM (POURAHMADI, 1999b; POURAHMADI, 2000)		TJMM (LIN; WANG, 2009)		LJMM (GUNEY *et al.*, 2022)		SMNJMM (proposed)	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	5.27	5.2791	0.0469	5.2742	0.0459	5.2779	0.0465	5.2813	0.0452
β_2	0.16	0.1893	0.0124	0.1998	0.0128	0.1907	0.0129	0.1905	0.0121
γ_1	1.12	1.1183	0.0581	1.1173	0.0578	1.1183	0.0638	1.1216	0.0605
γ_2	-0.58	-0.5691	0.0487	-0.5674	0.0481	-0.5688	0.0534	-0.5715	0.0489
γ_3	0.09	0.0934	0.0196	0.0931	0.0187	0.0932	0.0205	0.0931	0.0192
γ_4	-0.0043	-0.0046	0.0138	-0.0042	0.0133	-0.0046	0.0138	-0.0046	0.0139
λ_1	-0.4	-0.3883	0.2449	-0.3973	0.2414	-2.2568	0.2369	-0.4107	0.2432
λ_2	-1.37	-1.3771	0.1593	-1.3735	0.1569	-1.3562	0.1529	-1.3584	0.1627
λ_3	0.18	0.1861	0.0364	0.1859	0.0364	0.1828	0.0351	0.1822	0.037
λ_4	-0.007	-0.0033	0.0124	-0.0032	0.0124	-0.0033	0.0119	-0.0029	0.0121
BIC		-13.49		-12.67		-11.97		-13.01	

When data are generated from the Student-t distribution, the model that best fits the data is the TJMM model, followed by the proposed SMNJMM, as shown in the results presented in Table 2. As in the models with data from the normal distribution, in the case of the Student-t distribution, we also observed that the standard errors of the estimators are similar between the

models, with emphasis on the values of the parameters λ_1, λ_2, λ_3 and λ_4.

Again, using the pairwise Wilcoxon rank-sum test for comparison, we obtained a p-value greater than 5% for comparing BIC values between TJMM and SMNJMM, that is, there is no significant difference between the theoretical model and the proposed one. Furthermore, a p-value of less than 5% is obtained for comparing BIC values between the SMNJMM versus NJMM, and SMNJMM versus LJMM models, that is, there is a significant difference between the models that use the normal and Laplace distributions, respectively, and the proposed one. These comparisons help us to validate the efficiency and robustness of the proposed model in relation to other models, when the data is generated from the Student-t distribution structure.

Table 2 – Parameter estimations and Standard Errors (SE) for NJMM, TJMM, LJMM and SMNJMM of the data generated data from Student-t Distribution - Scenario 2.

Parameter	True value	NJMM (POURAHMADI, 1999b; POURAHMADI, 2000)		TJMM (LIN; WANG, 2009)		LJMM (GUNEY *et al.*, 2022)		SMNJMM (proposed)	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	5.27	5.2798	0.0454	5.2796	0.0443	5.2809	0.0395	5.2781	0.0453
β_2	0.16	0.1895	0.0118	0.1904	0.0126	0.1891	0.0119	0.1907	0.0129
γ_1	1.12	1.1095	0.0913	1.1135	0.1017	1.1158	0.0668	1.1145	0.1027
γ_2	-0.58	-0.5601	0.0729	-0.5652	0.0837	-0.5657	0.0554	-0.5665	0.0842
γ_3	0.09	0.0914	0.0245	0.0928	0.0261	0.0923	0.0206	0.0934	0.0275
γ_4	-0.0043	0.0049	0.0139	0.0041	0.0134	0.0043	0.0132	0.0046	0.0141
λ_1	-0.4	-0.4302	0.4378	-0.4141	0.4755	-2.4824	0.3031	-0.4764	0.4367
λ_2	-1.37	-1.3745	0.2629	-1.3792	0.2742	-1.3616	0.1808	-1.3437	0.2814
λ_3	0.18	0.1861	0.0545	0.1861	0.0566	0.1828	0.0398	0.1812	0.0574
λ_4	-0.007	-0.0033	0.0134	-0.0031	0.0135	-0.0031	0.0124	-0.0029	0.0135
BIC		-11.05		-14.51		-12.36		-13.97	

When the data are generated from the Laplace distribution, the model that best fits the data is the LJMM model, followed by the proposed SMNJMM model, according to the results presented in Table 3. For the last comparison using the pairwise Wilcoxon rank-sum test, we obtained a p-value greater than 5% for comparing BIC values between LJMM and SMNJMM, that is, there is no significant difference between the theoretical model and the proposed one. Furthermore, a p-value of less than 5% is obtained for comparing BIC values between the SMNJMM versus NJMM, and SMNJMM versus TJMM models, that is, there is a significant difference between the models that use the normal and Student-t distributions, respectively, and the proposed one. These comparisons help us to validate the efficiency and robustness of the proposed model in relation to other models, when the data is generated from the Laplace distribution structure.

Thus, as expected, the proposed SMNJMM model performs well whatever the distribution used for data generation, as long as this generating distribution belongs to the SMN family, which allows it to be used for a wide range of cases, making the model robust.

Table 3 – Parameter estimations and Standard Errors (SE) for NJMM, TJMM, LJMM and SMNJMM of the data generated data from Laplace Distribution - Scenario 3.

Parameter	True value	NJMM (POURAHMADI, 1999b; POURAHMADI, 2000)		TJMM (LIN; WANG, 2009)		LJMM (GUNEY *et al.*, 2022)		SMNJMM (proposed)	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	5.27	5.2829	0.1046	5.2771	0.1039	5.2851	0.1021	5.2778	0.0463
β_2	0.16	0.1891	0.0175	0.1903	0.0169	0.1906	0.0179	0.1911	0.0121
γ_1	1.12	1.1178	0.0616	1.1161	0.0601	1.1192	0.0597	1.1175	0.0611
γ_2	-0.58	-0.5690	0.0525	-0.5661	0.0512	-0.5671	0.0492	-0.5669	0.0509
γ_3	0.09	0.0935	0.0209	0.0935	0.0192	0.0931	0.0199	0.0934	0.0191
γ_4	-0.0043	0.0041	0.0141	0.0045	0.0135	0.0051	0.0132	0.0045	0.0133
λ_1	-0.4	1.4061	0.2508	1.4275	0.2253	-0.4135	0.2457	-0.3905	0.2391
λ_2	-1.37	-1.3539	0.1601	-1.3683	0.1475	-1.3594	0.1591	-1.3771	0.1603
λ_3	0.18	0.1831	0.0355	0.1858	0.0341	0.1838	0.0365	0.1861	0.0374
λ_4	-0.007	-0.0031	0.0125	-0.0032	0.0125	-0.0031	0.0125	-0.0032	0.0126
BIC		-12.32		-12.27		-14.87		-13.23	

4.1.2 Results for the ACD approach

Moving from the Modified Cholesky Decomposition structure to the **Alternative Cholesky Decomposition** structure, we conducted a similar experiment to compare the models, with the exception of the inclusion of the LJMM model (not yet presented in the literature), then we considered two scenarios: in scenario 1, a synthetic dataset is generated by using the normal distribution, while in scenario 2 the Student-t distribution with 4 degrees of freedom is employed. For both scenarios, $m = 100$ balanced random samples (with $n_i = 12$ observations for each sample subject) are generated.

Tables 4 and 5 present the estimation performance results in terms of mean estimates and standard errors (SE) computed on the r replications. To ease the understanding, the best and second best results are respectively displayed in blue and red. Table 4 shows the results when data are generated from a normal distribution while Table 5 shows the results for a Student-t generation model.

As expected and as observed in Table 4 for the first scenario, when data are generated from a normal distribution, the best performance is obtained with the NJMM model. The proposed SMNJMM is the second best fit. To identify whether there is a significant difference between these models in term of BIC values, the paired Wilcoxon hypothesis test is used. At the 5% significance level, no significant difference between the “theoretical" NJMM model and the proposed SMNJMM model are observed indicating that these two models perform similarly in the normal case. In addition, when compared with the TJMM model, a significant difference is obtained indicating that the TJMM is a worst fit compared to the NJMM and SMNJMM model for this first scenario.

Following the same approach, but now generating synthetic data from the Student-t distribution, Table 5 shows that the model presenting the best results is the TJMM, followed by

Table 4 – Estimation performance for the NJMM, TJMM and SMNJMM models - Scenario 1: data generated from the normal distribution.

Parameter	True value	NJMM (CHEN; DUNSON, 2003b)		TJMM (MAADOOLIAT *et al.*, 2013)		SMNJMM	
		Estimate	SE	Estimate	SE	Estimate	SE
β_1	5,27	5,2703	0,0394	5,2652	0,0489	5,2689	0,039
β_2	0,16	0,1605	0,0038	0,1607	0,0047	0,1606	0,0038
γ_1	1,12	1,1157	0,0489	1,114	0,0514	1,1152	0,0547
γ_2	-0,58	-0,5761	0,0385	-0,5749	0,0426	-0,5758	0,0427
γ_3	0,09	0,0899	0,0084	0,0898	0,0091	0,0899	0,0092
γ_4	-0,0043	-0,0036	0,0012	-0,0036	0,0021	-0,0036	0,0012
λ_1	-0,4	-0,3921	0,2362	-0,4014	0,2383	-2,2606	0,2287
λ_2	-1,37	-1,3807	0,1509	-1,3773	0,1525	-1,3598	0,1446
λ_3	0,18	0,1828	0,027	0,1827	0,0297	0,1797	0,0261
λ_4	-0,007	-0,0066	0,0018	-0,0066	0,0021	-0,0065	0,0018
BIC		-15,51		-11,94		-14,35	

the proposed SMNJMM model. The pairwise Wilcoxon test shows no difference between the "theoretical" TJMM model and the proposed SMNJMM model, while significant difference are observed when compared with the NJMM model.

Table 5 – Estimation performance for the NJMM, TJMM and SMNJMM models - Scenario 2: data generated from the Student-t Distribution.

Parameter	True value	NJMM (CHEN; DUNSON, 2003b)		TJMM (MAADOOLIAT *et al.*, 2013)		SMNJMM (proposed)	
		Estimate	SE	Estimate	SE	Estimate	SE
β_1	5,27	5,2702	0,0393	5,2651	0,0388	5,2688	0,0389
β_2	0,16	0,1603	0,0036	0,1605	0,0035	0,1604	0,0036
γ_1	1,12	1,1154	0,0486	1,1137	0,0481	1,1149	0,0544
γ_2	-0,58	-0,5762	0,0384	-0,575	0,0375	-0,5759	0,0426
γ_3	0,09	0,0897	0,0082	0,0896	0,0079	0,0897	0,009
γ_4	-0,0043	-0,0039	0,0009	-0,0039	0,0008	-0,0039	0,0009
λ_1	-0,4	-0,3922	0,2361	-0,4015	0,2342	-2,2607	0,2286
λ_2	-1,37	-1,3809	0,1507	-1,3775	0,1483	-1,36	0,1444
λ_3	0,18	0,1825	0,0267	0,1824	0,0264	0,1794	0,0258
λ_4	-0,007	-0,0076	0,0008	-0,0076	0,0008	-0,0075	0,0008
BIC		-11,13		-12,85		-12,18	

These two experiments on synthetic data clearly illustrate the robust behaviour of the proposed SMNJMM model with ACD structure. By do not imposing a fixed parametric form, the proposed model is able to fit the data as long as the hypothesis of belonging to SMN family of distribution is valid. In practice, this hypothesis is not so restrictive since the SMN family of distributions includes a very broad kind of distributions. Note also that the proposed SMNJMM is quite efficient since it performs as well as the theoretical model (NJMM in scenario 1 and TJMM in scenario 2).

4.2 Forward prediction

- The proposed model (SMNJMM with MCD structure) presents, in general, better more recent prediction results (up to step g = 4).

- The other proposed model (SMNJMM with ACD structure) presents, in general, better more recent prediction results (up to step g = 3).

4.2.1 Description

First, considering the **Modified Cholesky Decomposition**, we performed one more simulation study to evaluate the prediction performance of NJMM, TJMM, LJMM and SMNJMM models. To measure predictive accuracies, we report the empirical mean square forward prediction error (MSFE), defined as

$$\text{MSFE} = \frac{1}{mg}\sum_{i=1}^{m}(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)^\top(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i), \tag{4.4}$$

where m is the number of individuals, g is the number of steps forward for prediction, $\tilde{\mathbf{y}}_i$ is the predicted estimated vector of the response variables and $\mathbf{y}_i$ is the observed vector of the response variable.

The technique we used to obtain $\hat{\tilde{\mathbf{y}}}_i$ and assess the relative predictive performance of different models considers the pseudo-cross-validation approach, which combines the holdout cross-validation ((STONE, 1974)) and a predictive sample reuse ((GEISSER, 1975)), as follows: for each individual i, for $i = 1,...,m$, we do (i) holdout the last g measurements on the i th participant; (ii) compute ML estimates using the remaining data as the sample; (iii) prediction of the g true measurements $\mathbf{y}_i = (y_{i,n_i-g+1},...,y_{i,n_i})^\top$, denoted by $\hat{\tilde{\mathbf{y}}}_i = (\hat{\tilde{y}}_{i,n_i-g+1},...,\hat{\tilde{y}}_{i,n_i})^\top$, is made by using (3.36).

To evaluate the predictive power of each model, we generated $m = 100$ balanced random samples observed at $n_i = 12$ different times (the number of observations is the same for each individual) from the normal, Student-t (with 4 degrees of freedom) and Laplace distributions, and we estimated the parameters considering the previously mentioned pseudo-cross-validation approach. To compare the state of the art models with the proposed SMNJMM, we used the Relative Improvement Percentage (RIP) as classically employed in the literature (LIN; WANG, 2009; GUNEY *et al.*, 2022). It is defined as:

$$\text{RIP} = \frac{MSFE_{model} - MSFE_{SMNJMM}}{MSFE_{model}} \times 100\%,$$

where $MSFE_{model}$ can be the MSFE of NJMM, TJMM, or LJMM. A positive value for this coefficient indicates that the SMNJMM model has better future forecasting performance, while a negative value indicates that there has been no relative improvement. The obtained results are shown in Table 6.

Results for the MCD approach

Table 6 – MSFE and RIP (%) values for NJMM, TJMM, LJMM and SMNJMM models of g step-ahead forecasts (MCD structure).

Generating data from	Model	Metric	g step-ahead forecasts					
			g = 1	g = 2	g = 3	g = 4	g = 5	g = 6
normal	NJMM	**MSFE**	**0,0510**	**0,0731**	**0,1034**	**0,1137**	**0,1309**	**0,1740**
		RIP (%)	0,1961	0,1368	0,0967	0,0000	-1,7571	-2,2989
	TJMM	**MSFE**	**0,0516**	**0,0734**	**0,1075**	**0,1139**	**0,1339**	**0,1792**
		RIP (%)	1,3566	0,5450	3,9070	0,1756	0,5228	0,6696
	LJMM	**MSFE**	**0,0512**	**0,0736**	**0,1063**	**0,1145**	**0,1333**	**0,1830**
		RIP (%)	0,5859	0,8152	2,8222	0,6987	0,0750	2,7322
	SMNJMM	**MSFE**	**0,0509**	**0,0730**	**0,1033**	**0,1137**	**0,1332**	**0,1780**
Student-t	NJMM	**MSFE**	**0,0480**	**0,0567**	**0,0953**	**0,0954**	**0,1670**	**0,1895**
		RIP (%)	2,0833	1,0582	0,4197	0,3145	0,1198	1,1609
	TJMM	**MSFE**	**0,0472**	**0,0549**	**0,0951**	**0,0953**	**0,1664**	**0,1865**
		RIP (%)	0,4237	-2,1858	0,2103	0,2099	-0,2404	-0,4290
	LJMM	**MSFE**	**0,0481**	**0,0580**	**0,1004**	**0,0963**	**0,1675**	**0,1875**
		RIP (%)	2,2869	3,2759	5,4781	1,2461	0,4179	0,1067
	SMNJMM	**MSFE**	**0,0470**	**0,0561**	**0,0949**	**0,0951**	**0,1668**	**0,1873**
Laplace	NJMM	**MSFE**	**0,0481**	**0,1104**	**0,1379**	**0,1444**	**0,1635**	**0,1898**
		RIP (%)	0,4158	9,1486	2,7556	1,3850	0,6116	5,0053
	TJMM	**MSFE**	**0,0480**	**0,1171**	**0,1379**	**0,1444**	**0,1639**	**0,1819**
		RIP (%)	0,2083	14,3467	2,7556	1,3850	0,8542	0,8796
	LJMM	**MSFE**	**0,0477**	**0,1083**	**0,1345**	**0,1431**	**0,1615**	**0,1791**
		RIP (%)	-0,4193	7,3869	0,2974	0,4892	-0,6192	-0,6700
	SMNJMM	**MSFE**	**0,0479**	**0,1003**	**0,1341**	**0,1424**	**0,1625**	**0,1803**

We can observe that the MSFE metric increases as the number of steps for the prediction increases. This behavior is common for all steps presented, that is, when $g = 1$ (when we remove the last observation of each individual and use the prediction to estimate this value), when $g = 2$ (when we remove the last two observations of each individual and use prediction to estimate these values), and so on.

The proposed model (SMNJMM) generally presents better prediction results up to the step $g = 4$ for the three presented scenarios. After this prediction step ($g = 5$ and $g = 6$), the best model for prediction is the one generated with the same structure. Table 6 highlights these results.

4.2.2 Results for the ACD approach

Moving from the Modified Cholesky Decomposition structure to the **Alternative Cholesky Decomposition** structure, we conducted a similar experiment to compare the models. To compare the prediction performance of the models, we consider $m = 100$ balanced random

samples observed at $n_i = 12$ different times (we consider that the number of observations for each individual is the same), for the normal and Student-t distributions (with 4 distributions of degrees of freedom), and we estimated the parameters of the respective models through the pseudo-cross-validation mentioned above. The results are shown in Table 7.

Table 7 – MSFE and RIP (%) values for NJMM, TJMM, LJMM and SMNJMM models of g step-ahead forecasts (ACD structure).

Generating data from	Model	Metric	g step-ahead forecasts					
			g = 1	g = 2	g = 3	g = 4	g = 5	g = 6
Scenario 1: normal	NJMM	**MSFE**	**0,0823**	**0,1042**	**0,1343**	**0,1442**	**0,1620**	**0,2049**
		RIP (%)	0,2162	0,1478	0,1185	0,0000	-2,3146	-3,4123
	TJMM	**MSFE**	**0,0830**	**0,1048**	**0,1389**	**0,1455**	**0,1655**	**0,2108**
		RIP (%)	1,4981	0,5287	3,8291	0,1831	0,6210	0,6565
	SMNJMM (proposed)	**MSFE**	**0,0814**	**0,1036**	**0,1338**	**0,1443**	**0,1637**	**0,2086**
Scenario 2: Student-t	NJMM	**MSFE**	**0,0796**	**0,0887**	**0,1269**	**0,1278**	**0,1988**	**0,2213**
		RIP (%)	2,0112	1,7654	0,2322	0,6765	0,1234	1,1392
	TJMM	**MSFE**	**0,0789**	**0,0867**	**0,1268**	**0,1270**	**0,1971**	**0,2175**
		RIP (%)	0,4543	0,3336	0,2654	-0,2737	-0,2876	-0,4312
	SMNJMM (proposed)	**MSFE**	**0,0775**	**0,0866**	**0,1255**	**0,1275**	**0,1974**	**0,2178**

Unsurprisingly, as displayed in Table 7, the MSFE score increases with the number of steps g in the prediction. This behavior is expected since the model is less and less confident to predict long term observations. Table 7 also shows that the model proposed SMNJMM model exhibits better prediction performance up to the third step, that is, up to $g = 3$ for the two scenarios. The results after this prediction step ($g = 4, 5, 6$) are better for the theoretical models (NJMM for scenario 1 and TJMM for scenario 2).

In the next section, we will evaluate the proposed model on a real dataset and compare its performance with other models with similar structure (NJMM and TJMM). In addition, we perform a comparison between MCD and ACD decompositions.

5 REAL DATA ANALYSIS

In this chapter, we present the model performance results based on real data, for the models using the normal, Student-t and Laplace distributions and for the proposed model, considering the modified and alternative Cholesky decompositions. The results are presented in terms of the standard errors of the estimates and the BIC values adapted for the problem. Forward prediction estimates are also evaluated.

- For both MCD and ACD decompositions, the proposed SMNJMM model provides a lower BIC value than the other NJMM and TJMM models.
- For the same model (NJMM, TJMM or SMNJMM), ACD decomposition always leads to better performance than MCD.
- SMNJMM model is less influenced by outliers.

5.1 About the dataset

Now that the proposed SMNJMM model has been successfully validated on synthetic data, some experiments on real data are carried out in this section. For that, a standard dataset is used. It consists of dental growth from the pituitary gland up to the left pterygomaxillary distance (mm) measured repeatedly at the ages of 8, 10, 12 and 14 years, with 11 women and 16 men, presented in (POTTHOFF; ROY, 1964), as shown in the Figure 5. Data can be downloaded from the `nlme` (PINHEIRO *et al.*, 2007) package in R (R Core Team, 2020).

5.2 Estimation performance

The objective of the first experiment is to confirm the potential of the proposed SMNJMM model compared to the NJMM and TJMM in terms of fitting accuracy. In addition, a comparison between the MCD and ACD decomposition structures is done.

The regression model is defined as:

$$\mu_{ij} = \beta_0 + \delta_0 I_i(F) + (\beta_1 + \delta_1 I_i(F)) t_j, \tag{5.1}$$

with $i \in \{1,...,m = 27\}$ (27 subjects, 11 women and 16 men) and $j \in \{1,...,n_i = 4\}$ (4 times steps at 8, 10, 12 and 14 years). μ_{ij} is the average orthodontic distance for the i-th subject at age t_j. β_0 and β_1 are the intercept and slope for men, respectively. δ_0 and δ_1 are the difference in intercept and slope between women and men, respectively. $I_i(F)$ is the indicator variable for

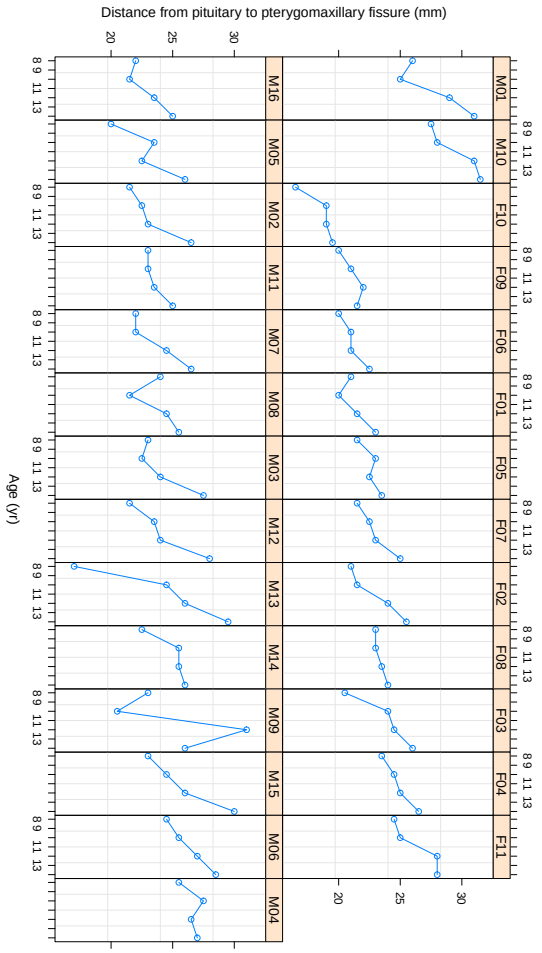


Figure 5 – Distance between the pituitary gland and the pterygomaxillary fissure of 27 children (16 males and 11 females) aged 8 to 14 years, measured every two years.

women. In the following, the Poly(1,1) model is considered (i.e. $d = q = 1$ for the covariate vectors).

The maximum likelihood estimates corresponding to the six compared models (NJMM, TJMM and SMNJMM for both MCD and ACD decompositions) are presented in Table 5.2. As observed, both for the MCD and ACD decompositions, the proposed SMNJMM model provides a lower BIC value than the other NJMM and TJMM counterparts. This highlights the effectiveness of the proposed SMNJMM model for this practical application. In addition, for a same model (NJMM, TJMM or SMNJMM), the ACD decomposition always leads to better performance than the MCD one. These results clearly illustrate the interest of jointly using the SMNJMM model with the ACD decomposition.

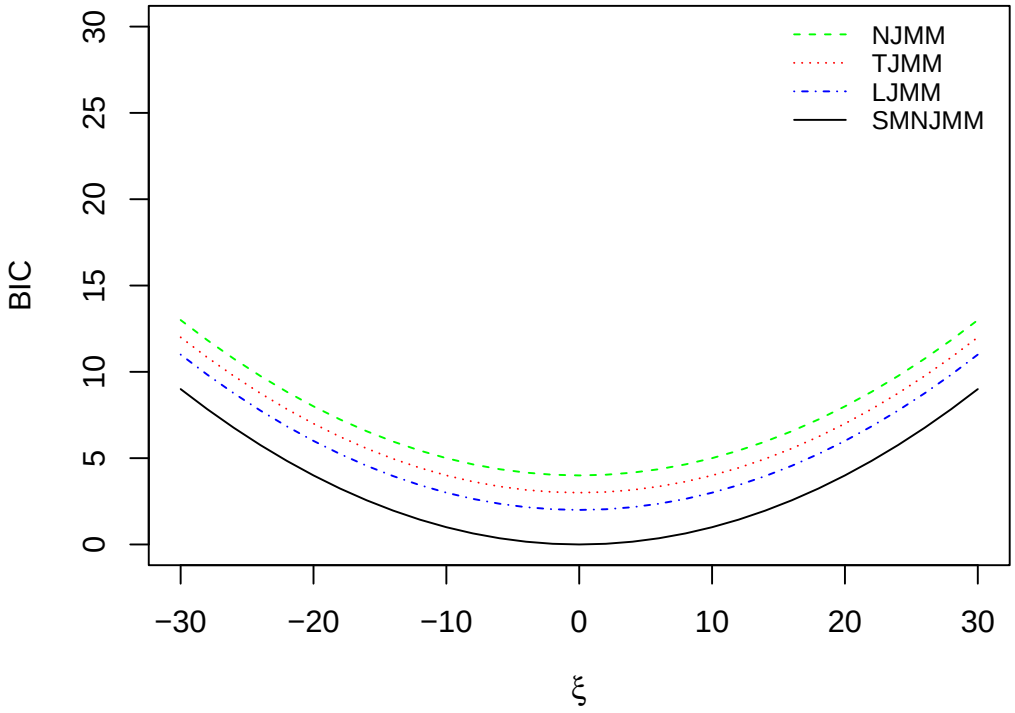
Parameter	MCD decomposition			ACD decomposition		
	NJMM (POURAHMADI, 1999a)	TJMM (LIN; WANG, 2009)	SMNJMM (proposed) (RIBEIRO *et al.*, 2024)	NJMM (CHEN; DUNSON, 2003b)	TJMM (MAADOOLIAT *et al.*, 2013)	SMNJMM (proposed) (RIBEIRO *et al.*, 2024)
β_1	20,9738	15,2569	21,5423	18,4050	12,7204	18,8457
β_2	1,6117	0,7955	1,4254	1,3313	0,5151	0,9674
δ_1	-0,6947	0,9785	2,3120	-1,3303	0,3429	1,8540
δ_2	-0,6572	-0,3127	0,5214	-0,93f76	-0,5931	0,0634
γ_1	0,8498	0,9894	1,0685	0,2142	0,3538	0,4329
γ_2	-0,2537	-0,3781	-0,3781	-0,5341	-0,6585	-0,6585
λ_1	1,9268	1,5298	1,9951	1,2912	0,8942	1,3595
λ_2	-0,2846	-0,3714	-0,3124	-0,5650	-0,6518	-0,5928
BIC	9,59	12,52	8,52	8,44	12,04	8,03

5.3 Robustness performance

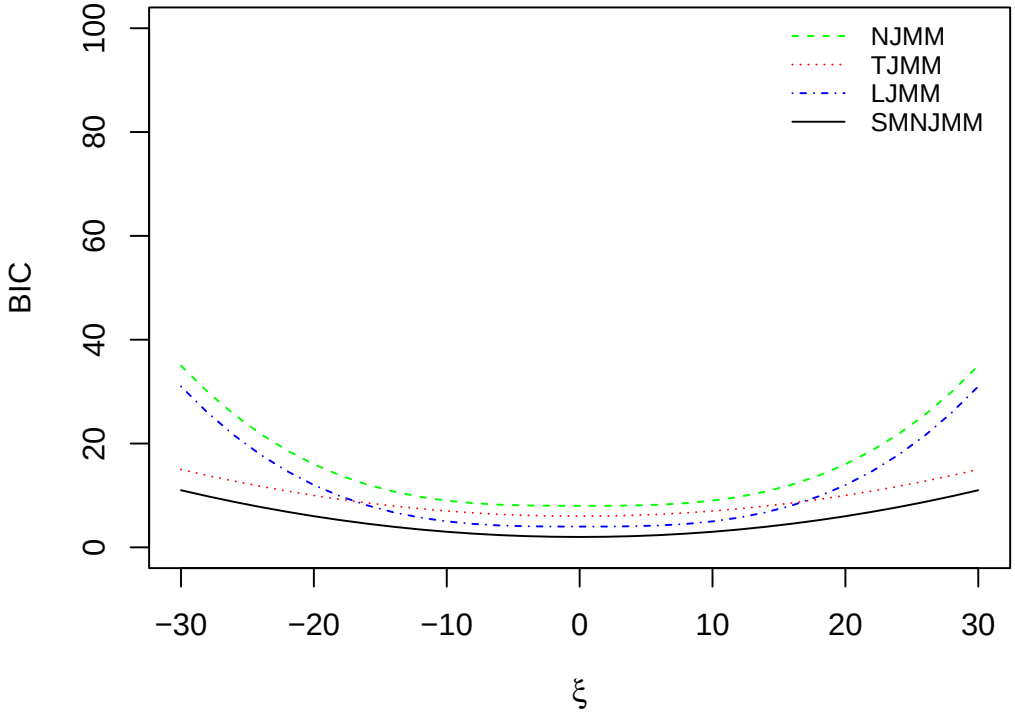
A second experiment is performed on this dataset to evaluate the robustness behavior of the proposed SMNJMM model. The same evaluation protocol is employed as the one presented in (LIN; WANG, 2009) and (GUNEY *et al.*, 2022). First a contamination denoted by ξ is added for a single observation as:

$$y_{ij}(\xi) = y_{ij} + \xi. \tag{5.2}$$

Then the NJMM (in green), TJMM (in red), LJMM (in blue) and and SMNJMM (in black) models are estimated and their performance are evaluated in term of BIC values. Here, only the ACD decomposition for the scatter matrix is considered. Figure 6 draws the evolution of the BIC criterion as a function of the disturbance intensity (ξ) for two subjects: a woman in Figure 6(a) and a man in Figure 6(b).



(a)



(b)

Figure 6 – Evolution of the BIC metric as a function of perturbation ξ for the NJMM, TJMM,

From Figure 6, we observe that for both subjects, the BIC values obtained with the proposed SMNJMM model are always smaller than the BIC values for the NJMM and TJMMM models. These results confirm that the SMNJMM model is less influenced by outliers as explained in Section 2. This robustness property of the proposed SMNJMM model allows it to satisfactorily accommodate extreme observations and makes it a relevant model to use in practice.

5.4 Conclusion

In this chapter, we present the performance results for the proposed models, considering both synthetic and real data. The experiments on synthetic data confirmed the robustness property of the proposed SMNJMM model compared to the NJMM, TJMM, and LJMM models with the MCD and ACD structures.

We used hypothesis testing to evaluate the results of the different models and validated that there is a significant difference between the models using the Student-t and Laplace distributions compared to the proposed model, validating the efficiency of our model.

We also observed that the proposed SMNJMM model provides a lower BIC value than the other NJMM and TJMM models. Furthermore, for the same model (NJMM, TJMM, or SMNJMM), the ACD decomposition always leads to better performance than MCD. We were also able to validate that the SMNJMM model is less influenced by outliers.

In the next chapter, we will discuss a summary of what was presented throughout this thesis, as well as present possible development topics for future work on the subject.

6 CONCLUSION AND FUTURE WORKS

6.1 Conclusion

The study of longitudinal data is of interest in several areas, including biomedical and agronomic applications. And like any other statistical methodology, it has advantages and disadvantages regarding its use. Monitoring sample subjects during the study, for example, may not be a simple task, due to the complexity of data collection and management. However, this approach provides more suitable conditions for the study of covariates that can influence the response variable; in addition to evaluating the behavior of the response over time and allowing the study of changes in the behavior of the mean response of the sample subject in the different treatments (incorporating information on intra-individual variation in the analysis).

Therefore, failing to consider the inherent characteristic of dependence of longitudinal data can lead us to erroneous conclusions/interpretations of the results obtained. The use of the paired t-test or ANOVA (parametric or non-parametric), for example, was widely used as simple alternatives for modeling this type of data. However, more robust proposals for including correlation in the modeling structure have been proposed over time. Much of the initial research focused on cases in which the model errors follow a normal distribution. Among them, we can mention the analysis of variance with repeated measures, univariate/multivariate profile analysis, growth curve analysis and models with random effects.

The proposals mentioned above, although still widely used, may not be suitable for certain problems, such as non-normality, for example, an intrinsic fact in a large part of the data set. Based on this central motivation of the limitation of the normality assumption and the computational advances inherent in the 1980s and 1990s, other approaches (more robust and more generalist) have been developed in the literature over time.

One of these approaches was considered in this thesis and is based on jointly modeling the mean and scale covariance, and our contribution is focused on developing this approach considering an SMN error framework. This approach generalizes recent models based on the normal, Student-t and Laplace distributions, using the distribution-free (unknown but deterministic) mixing parameter τ, i.e., the goal is to generalize particular cases, assuming that $\mathbf{r}_i \sim \text{SMN}(\mathbf{0}, \blacksquare_i, \tau_i)$. To achieve this, we assume that the parameters τ_i are deterministic but unknown. By not imposing any explicit structure on its probability density function, the proposed regression model will be robust, since it has the ability to represent a wide range of distributions

belonging to the SMN family, including normal, Student-t and Laplace structures as special cases.

The idea was based on working with the assumption made about the distribution of errors $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ (where $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, in the context of regression models). Thus, if we consider that $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \blacksquare_i)$, we have the NJMM model, as presented in (POURAHMADI, 1999b). If we consider that $\mathbf{r}_i \sim$ Student-t$(\mathbf{0}, \blacksquare_i, v)$, where v are the degrees of freedom, we have the TJMM model, as presented in (LIN; WANG, 2009). And if we consider that $\mathbf{r}_i \sim$ Laplace$(\mathbf{0}, \blacksquare_i)$, we have the structures of the LJMM model, as presented in (GUNEY *et al.*, 2022).

From this structure, we derive the MCD (Modified Cholesky Decomposition) and ACD (Alternative Cholesky Decomposition) decompositions, derive maximum likelihood estimators, and provide an algorithm for estimating the model parameters. We present estimators for future prediction, providing an explicit expression of the mean squared error predictor. Based on simulated and real data, we present extensive experiments to illustrate the robustness of the proposed model.

From the results obtained, we observed that the SMNJMM model with MCD structure presents a lower BIC value than the other models, that is, the results of the data analysis support the effectiveness of the proposed model. To evaluate the prediction of new observations for orthodontic growth data, we saw that, from the values presented, the SMNJMM model provides the lowest MSFE values. Compared to the modified Cholesky decomposition (MCD), the ACD structure proved to be more appropriate in some situations, validating the interest of applying the proposed model. [1]

6.2 Future works

Here, we present some directions that we consider as possible areas of future research for the commsubjecty and that we intend to explore. Below, we have compiled these suggestions and included others that may also be of interest. The main points are presented below:

- **Potential applications:** The models proposed in this thesis are very generic and can be used for many applications. Here, we show their interest in synthetic data and in a real dataset commonly used in the scientific commsubjecty. But, several other application perspectives can be considered in different domains: clinical and health studies (monitoring the progression of diseases, effects of treatments or changes in health param-

[1] The R codes will be available upon request to the authors.

eters over time), as discussed in (SCHAFER; KANG, 2008) and (PFOHL; YAMADA, 2016); psychopedagogical studies (monitoring the cognitive development of children or the impact of educational interventions over time), as discussed in (GORE; GIL, 2015) and (HUSSAIN; GOH, 2017); economic studies (studying the effects of public policies, investments, or macroeconomic changes over time), as discussed in (PISCHKE, 2005) and (CAMERON; TRIVEDI, 2009); sociological studies (tracking changes in attitude or behavior over time, such as in social interventions or public policy changes), as discussed in (BREEN; JONSSON, 2020) and (MANSKI; PEPPER, 2021); among others.

- **Sparse regression models:** In the context of developing new methodologies, it is possible to evaluate the use of the modified Cholesky decomposition (MCD) and the Alternative Cholesky Decomposition in the context of sparse regression models. These models use regularization techniques to select a subset of relevant predictor variables, forcing most of the coefficients to become zero, which results in simpler and more interpretable models. (ZHOU *et al.*, 2021) and (XU *et al.*, 2015) developed the approach following these models, but under the assumption of a Gaussian distribution and without considering any Cholesky decomposition for the scattering matrix Σ, for example.
- **Other distributions:** Regarding the possibility of considering other distributions that do not belong to the scale mixture family of normal distributions, we suggest applying it to distributions that generalize the normal distribution, such as the multivariate generalized Gaussian distribution (MGGD, also known as the exponential power distribution) and/or the Kotz type distribution. The MGGD generalizes the normal distribution by including one more parameter (a power parameter), while the Kotz type extends the MGGD by considering one more parameter. (PLUNGPONGPUN; NAIK, 2008), for example, uses the Kotz distribution for an ANOVA and this methodology can possibly be applied to a JMM model.
- **Other decompositions:** In addition to considering the use of other distributions, it is also worth considering the use of other Cholesky decompositions. In this work, we consider the MCD and the ACD, but structures such as the parameterization of the Hyperspherical Cholesky Factor (HPC), for example, can be considered. This decomposition shows that the dependencies and correlations between observations of the same individual over time can be better modeled through hyperspherical coordinates, which offer a way to capture nonlinear, nonstationary correlations with complex variance-covariance structures. (PAN;

PAN, 2017) presents this structure for the context of normality, but the idea can be extended to other distributions of the SMN family.

- **Machine learning based methods versus statistical models:** In the context of Machine Leaning, there are some models (such as decision trees, random forests or deep neural networks) that were specifically designed to process longitudinal data. In (SIMCHONI; ROSSET, 2023) and (SIMCHONI; ROSSET, 2021) some approaches of neural network structures were presented that aim to imitate the most well-known structure for longitudinal data: mixed linear models. It is possible that this proposal can be adapted when other hypotheses are made about the errors (Student, Laplace, SMN, for example) and also for different specific structures for the scatter matrix (MCD and ACD).

A PAPER 1: A GENERAL ROBUST APPROACH FOR JOINT MODELING ON THE FAMILY OF SCALE MIXTURE OF NORMAL DISTRIBUTION

Published paper:

Ribeiro, V. S. O., Bombrun, L., Nobre, J. S., Cavalcante, C. C., & Berthoumieu, Y. (2024).

A general robust approach for joint modeling of the family of scale mixture of normal distribution.

Signal Processing, 220, 109426.

A general robust approach for joint modeling of the family of scale mixture of Normal distribution

Vinícius Silva Osterne Ribeiro[a], Lionel Bombrun[b,c], Juvêncio Santos Nobre[d], Charles Casimiro Cavalcante[a], Yannick Berthoumieu[b]

[a] *Department of Teleinformatics Engineering, Federal University of Ceara, Campus do Pici, Block 725, Fortaleza, 60455-970, Ceará, Brazil*
[b] *Univ. Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218, Talence, F-33400, France*
[c] *Bordeaux Sciences Agro, Gradignan, F-33175, France*
[d] *Department of Statistics and Applied Mathematics, Federal University of Ceara, Campus do Pici, Block 910, Fortaleza, 60455-970, Ceará, Brazil*

Abstract

In this paper, we present the class of linear models with errors belonging to the family of scale mixture of normal distributions, considering the reparameterization of the scatter matrix based on the Modified Cholesky Decomposition approach and that the mixing parameters of the model are deterministic but unknown and vary from each observation. Scoring functions and Hessian matrices are derived for the parameters of interest and an iterative process is proposed for parameter estimation. Some aspects of predicting new observations and robustness of maximum likelihood estimates are discussed, as well as the performance of different models with different structures for one of the parameters of the family of distributions used. The methodology is applied in an extensive comparative approach with synthetic and real data. Performance results and prediction of new observations show that the proposed model performs better compared to recent models based on Normal, Student-t and Laplace distributions.

Keywords: Longitudinal Data, Scale Mixture of Normal distribution, Modified Cholesky Decomposition
PACS: 0000, 1111
2000 MSC: 0000, 1111

1. Introduction

An important category of repeated measures consideration is longitudinal studies (also known as cohort or panel studies), whose characteristic is to consider the collection of the samples obtained in successive moments ordered under some dimension (e.g. time, velocity, height), in which the measurement is done considering the same physical quantity (e.g. seconds, miles per hour, meters). The reason of the analysis of this type of data is to consider the existing dependence in the individual observations of responses of the observed units. Examples of this type of study are presented in [1], [2], [3] and [4] in different contexts, including econometrics, biomedical or agricultural applications.

In recent years, an approach that has gained relevance for analyzing longitudinal data is joint mean-covariance models (or multivariate heteroscedastic regression model), initially proposed in [5] and discussed in more details in [6], under the assumption of error normality, called the NJMM (Normal Joint Modeling Model). The NJMM considers that the mean vector is parameterized as a linear model and the covariance matrix is formed to represent the heteroscedastic and autoregressive structure of measurements within the same subject. This representation of the covariance matrix is constructed from the Modified Cholesky Decomposition (MCD) [7], which is ideal when high-dimensional (HD) and positive definition (PD) constraints are obstacles in modeling covariance and correlation matrices. Other decompositions of the covariance matrix were also developed [8], such as the Alternative Cholesky Decomposition (ACD) [9], and the Hyperspheric Parameterization of the Cholesky Factor (HPC) [10].

Considering that in many cases the normality assumption for the errors can be easily violated, some generalizations of the NJMM model were proposed in the literature by considering the Student-t distribution for the residuals, yielding to the TJMM structure (Student-t Joint Modeling Model). This idea was followed in [11] and [12] where the MCD and HPC decompositions are respectively used for the covariance matrix. This modification in the model is useful when residuals have heavier tails than the normal distribution. However, although the inclusion of one more degree of freedom makes the model more flexible, the estimation process becomes a little more complex. An alternative to the TJMM was recently developed in [13], which considered the joint modeling of the mean and covariance of the multivariate Laplace distribution, which is called LJMM (Laplace Joint Modeling Model),

an alternative structure of heavy tail to the multivariate normal distribution. Furthermore, the model is simpler than the TJMM, as the number of parameters to be estimated is the same as in the NJMM model.

Even if these models have successfully been used in the literature, they have one major drawback. They all assume a fixed parametric form for the distribution of the residuals which may be problematic especially in presence of outliers. To overcome this limitation, this paper introduces a novel joint modeling model adapted to a wider class of distributions. Considering that the normal, Student-t and Laplace distributions are particular cases of the class called Scale Mixtures of Normal (SMN), presented by [14], in this paper we propose the joint modeling of the mean and covariance for analysis of data with repetitions using this class, whose model we will call SMNJMM (Scale Mixtures of Normal Joint Modeling Model), in which the NJMM, TJMM and LJMM models will be particular cases.

Based on this context, the main contributions of this work can be summarized as follows:

- By considering that residuals follow a scale mixture of normal distribution with unknown but deterministic mixing parameters, we introduce a novel robust joint modeling model for the modeling of longitudinal data;

- Based on the MCD decomposition, we derive the maximum likelihood estimators and provide an iterative algorithm to estimate its parameters;

- In the context of forward prediction, we give an explicit expression of the mean square error predictor;

- Based on simulated and real data, we introduce extensive experiments to illustrate the robustness of the proposed approach for the modeling of longitudinal data.

The paper is organized as follows. Details on the structure of the distribution, approach for joint modeling of mean and covariance, and motivation for developing this paper are presented in Section 2. In Section 3, we present the proposed model, including the structure, parameter estimation and algorithm for this process, and estimation/prediction for new observations of the response variable. In Section 4, simulation studies and application to a

real dataset are presented to evaluate the sensitivity and robustness of the proposed model. Finally, the Section 5 is dedicated to conclusions and final considerations about the paper.

2. State of the art

In this section, we will present the structure of the distribution that will be considered for the modeling of the residuals of the regression model for longitudinal data, as well as the approach of reparameterization of the scatter matrix for the joint modeling of the mean and covariance.

The first subsection gives a brief recall of the SMN model with a specific focus on the case of deterministic but unknown mixing parameters τ_i. Then, subsection 2.2 introduces the main notations for the joint modeling of longitudinal data based on the MCD decomposition. And finally subsection 2.3 gives the main ingredients of our proposal which consists in modeling the residuals of the joint model with an SMN distribution with deterministic but unknown mixing parameters τ_i.

2.1. SMN distribution and its extension

We say that an n-dimensional random variable $\mathbf{y}$ follows a Scale Mixture of Normal distribution [14] with mean vector μ and scatter matrix $\mathbf{\Sigma}$, if its stochastic representation can be given by

$$\mathbf{y} = \mu + \tau^{1/2}\mathbf{z}, \tag{1}$$

where τ and $\mathbf{z}$ are independent, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and $\tau \in \mathbb{R}^+$ is a scalar random variable. An important property of this structure is that the random variable $\mathbf{y}$ given τ has a normal distribution, where the mean vector and covariance matrix are given, respectively, by

$$\mathbb{E}(\mathbf{y}|\tau) = \sqrt{\tau}\mathbb{E}(\mathbf{z}) + \mu = \mu \tag{2}$$

and

$$\mathrm{Var}(\mathbf{y}|\tau) = \mathbb{E}[(\mathbf{y}-\mu)(\mathbf{y}-\mu)^\top|\tau] = \sqrt{\tau}\mathbb{E}(\mathbf{z}\mathbf{z}^\top)\sqrt{\tau} = \tau\mathbf{\Sigma}, \tag{3}$$

and consequently we can write its conditional distribution function as follows

$$\mathbf{y}|\tau \sim \mathcal{N}(\mu, \tau\mathbf{\Sigma}) = \frac{1}{(2\pi\tau)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{(\mathbf{y}-\mu)^\top\mathbf{\Sigma}^{-1}(\mathbf{y}-\mu)}{2\tau}\right\}. \tag{4}$$

Hence, each component of $\mathbf{y}$ is correlated with the other components from the variance dependence generated by τ. We can obtain the probability density function of the random variable $\mathbf{y}$ by integrating the joint distribution of $\mathbf{y}$ and τ over τ, i.e.

$$f(\mathbf{y}) = \int_0^\infty f(\mathbf{y}|\tau) f(\tau) d\tau. \tag{5}$$

Depending on the choice of the distribution of the parameter τ, we may have different structure for the multivariate vector $\mathbf{y}$, i.e.:

- if τ follows an Inverse Gamma distribution with shape parameter a and scale parameter b, denoted as $\tau \sim IG(a, b)$, whose probability density function is given by

$$f(\tau; a, b) = \frac{b^a}{\Gamma(a)} \tau^{-a-1} e^{-b/\tau}, \tag{6}$$

  where $\tau \in (0, \infty)$ and $\Gamma(\cdot)$ is the complete Gamma function, then $\mathbf{y}$ follows a multivariate Student-t distribution [15];

- if τ follows a Gamma distribution with shape parameter a and scale parameter b, denoted $\tau \sim G(a, b)$, whose probability density function is given by

$$f(\tau; a, b) = \frac{1}{b^a \Gamma(a)} \tau^{a-1} e^{-\tau/b}, \tag{7}$$

  where $\tau \in (0, \infty)$, then $\mathbf{y}$ follows a K-distribution, a family of three parameters that arise from the composition of two Gamma distributions. In addition, if τ follows a Gamma distribution with shape parameter $a = n + 1/2$ and scale parameter $b = 1/8$, then $\mathbf{y}$ follows a Laplace distribution.

As mentioned before, by choosing a specific structure for the τ parameter, in the SMN family of distributions, a closed-form expression for $\mathbf{y}$ can be obtained by solving (5) such as the Student-t and Laplace distributions.

The main disadvantage of the previously mentioned approaches is that a well-defined parametric form is imposed for the structure of τ and, consequently, for the variable $\mathbf{y}$, which can be easily violated in practice. To overcome this problem, [16], [17] and [18] propose to relax this restriction

by not imposing any explicit structure for the probability density function of τ. Furthermore, they consider that the value of τ varies with each sampling unit, so that for a set of m individuals, we have the parameters $\tau_1, \tau_2, ..., \tau_m$ to be estimated, that are deterministic but unknown, and the ML estimator for this parameters are given by

$$\hat{\mu} = \frac{\sum_{i=1}^{m} \frac{\mathbf{y}_i}{[(\mathbf{y}_i-\hat{\mu})^\top \hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_i-\hat{\mu})]^{1/2}}}{\sum_{i=1}^{m} \frac{1}{[(\mathbf{y}_i-\hat{\mu})^\top \hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_i-\hat{\mu})]^{1/2}}} \tag{8}$$

$$\hat{\mathbf{\Sigma}} = \frac{n}{m}\sum_{i=1}^{m} \frac{(\mathbf{y}_i - \hat{\mu})(\mathbf{y}_i - \hat{\mu})^\top}{(\mathbf{y}_i - \hat{\mu})^\top \hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\mu})} \tag{9}$$

$$\hat{\tau}_i = \frac{1}{n}(\mathbf{y}_i - \hat{\mu})^\top \hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\mu}). \tag{10}$$

An important aspect about the scatter matrix $\mathbf{\Sigma}$ is its identification problem. We can rewrite the stochastic representation presented in (1), given by $\mathbf{y} = \mu + \tau^{1/2}\mathbf{z}$, considering $\mathbf{y} = \mu + (a\tau)^{1/2}\frac{\mathbf{z}}{a^{1/2}}$ or simply $\mathbf{y} = \mu + (\kappa)^{1/2}\mathbf{c}$, where $\kappa = a\tau$ is a positive scalar parameter and $\mathbf{c} \sim \mathcal{N}(0, \Sigma_c)$, with $\Sigma_c = \Sigma/a$. This means that vectors (μ, τ, Σ) and $(\mu, a\tau, \Sigma/a)$ identify the same SMN model. In this way, we need to impose a restriction to identify the parameters and this restriction will be made in the scatter matrix, as follows

$$\mathbf{\Sigma} = \frac{n\mathbf{\Sigma}}{tr(\mathbf{\Sigma})}, \tag{11}$$

where $tr(\mathbf{\Sigma})$ is the trace of the scatter matrix, i.e., the sum of the main diagonal of this matrix. In this case, we impose the restriction that $tr(\mathbf{\Sigma}) = n$.

In the following section, we will present the structure to be considered when modeling regression models for longitudinal data and understand how this distribution will be involved in the process.

2.2. Joint modeling of longitudinal data

Suppose that repeated measurements of a continuous random variable are observed over time in each of the m subjects under study. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^\top$ be the result vector of the ith subject (experimental unit), where n_i is the number of results per subject i, $i = 1, 2, ..., m$, which are allowed unevenly spaced.

One of the most common ways of modeling the mean of $\mathbf{y}_i$ is the linear regression model given by

$$\mathbf{y}_i = \mu_i + \mathbf{r}_i, \tag{12}$$

where $\mu_i = (\mu_{i1}, ..., \mu_{i_{n_i}})^\top = \mathbf{X}_i\beta$, with $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i})^\top$ being the $n_i \times (p+1)$ matrix of explanatory variables, which is known and assumed to be of full row rank, $\beta = (\beta_0, ..., \beta_p)^\top$ being the vector of regression coefficients and $\mathbf{r}_i = (r_{i1}, ..., r_{i_{n_i}})^\top = \mathbf{y}_i - \mu_i$ is the residual vector, which can assume different structures, among the most common is to consider that $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i)$.

In addition to mean modeling, the focus of regression models for longitudinal data is also on modeling the scatter matrix, given that we have repeated measures for different individuals and modeling this structure must be considered. However, this matrix has peculiarities that make estimation processes difficult, such as high dimensionality (we will have many parameters to be estimated if we do not impose any structure to the matrix), positive definition (property that these matrices must obey), and, if we do not consider any restrictions, we will have more parameters to estimate than collected observations.

In the context of joint mean-covariance modeling, the idea proposed by Pourhamadi in [5] and [6] to ensure these two properties is to rewrite the scatter matrix using the Modified Cholesky Decomposition (MCD) approach, where such matrix is reparametrized as follows

$$\mathbf{L}_i\mathbf{\Sigma}_i\mathbf{L}_i^\top = \mathbf{D}_i, \tag{13}$$

where $\mathbf{L}_i = [l_{jk}]$ are lower triangular matrices with 1's as diagonal entries and (j,k)th entry being $-\phi_{jk}$, called autoregressive parameters, for $k \in \{1, ..., j-1\}$, as denoted below

$$\mathbf{L}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_{21} & 1 & 0 & \cdots & 0 \\ -\phi_{31} & -\phi_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\phi_{n_i1} & -\phi_{n_i2} & -\phi_{n_i3} & \cdots & 1 \end{bmatrix}, \tag{14}$$

and $\mathbf{D}_i = \text{diag}\{\sigma_1^2, ..., \sigma_{n_i}^2\}$, where σ_j^2 are called scale innovation variances, as denoted below

$$\mathbf{D}_i = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{n_i}^2 \end{bmatrix}, \tag{15}$$

whose interpretations include: (i.) the below-diagonal entries of $\mathbf{L}_i$ are the negatives of the autoregressive parameters, namely $-\phi_{jk}$, in

$$y_{ij} = \mathbf{x}_{ij}^{\top}\beta + \sum_{k=1}^{j-1} \phi_{jk}(y_{ik} - \mu_{ik}), \tag{16}$$

and (ii.) the diagonal entries of $\mathbf{D}_i$ are the scale innovation variances $\sigma_j^2 = \text{var}(y_{ij} - \hat{y}_{ij})$. It is worth noting that we can also rewrite the inverse of the scatter matrix as $\mathbf{\Sigma}_i^{-1} = \mathbf{L}_i^{\top}\mathbf{D}_i^{-1}\mathbf{L}_i$.

Following [5] and [6], we are able to model the scatter matrix from a joint regression model for variances and correlations considering

$$\phi_{jk} = \mathbf{z}_{jk}^{\top}\gamma \quad \text{and} \quad \ln \sigma_j^2 = \mathbf{w}_j^{\top}\lambda, \tag{17}$$

where $\gamma = (\gamma_1, ..., \gamma_d)^{\top}$ and $\lambda = (\lambda_1, ..., \lambda_q)^{\top}$ are d and q-dimensional vectors of parameters, $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ are d and q-dimensional covariate vectors.

The parameters γ and λ are assumed to be common for all $\mathbf{\Sigma}_i$'s exhibiting the same covariance structure. The covariates $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ may contain the baseline covariates, the time and the associated interactions. These covariates are presented as time polynomials for the autoregressive components and time powers for innovation variances, respectively, as proposed in [19], i.e.

$$\mathbf{z}_{jk} = \left[\begin{array}{ccccc} 1 & (t_{ik} - t_{ij}) & (t_{ik} - t_{ij})^2 & \cdots & (t_{ik} - t_{ij})^{d-1} \end{array}\right]^{\top} \tag{18}$$

and

$$\mathbf{w}_j = \left[\begin{array}{ccccc} 1 & t_{ij} & t_{ij}^2 & \cdots & t_{ij}^{q-1} \end{array}\right]^{\top}, \tag{19}$$

where t_{ij} (or t_{ik}) is the measure of time referring to the sampling unit i in the repetition j (or repetition k), for $i \in \{1, \cdots, m\}$, $j \in \{1, \cdots, n_i\}$, and $k \in \{1, \cdots, j-1\}$.

From the presentation of this modeling structure for the average and the reparametrization for estimation of the scatter matrix, we will present the motivation for the development of this paper in the next section.

2.3. Our proposal

As mentioned in the introduction, in the literature there are several approaches for joint modeling of mean and covariance using the Modified

Cholesky Decomposition. What differentiates these approaches is the assumption on the distribution of the residuals $\mathbf{r}_i = \mathbf{y}_i - \mu_i$ (where $\mu_i = \mathbf{X}_i\beta$, in the context of regression models). We mentioned that if we consider that $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i)$, we have the Normal Joint Modeling Model structured, as presented in [5], if we consider that $\mathbf{r}_i \sim$ Student-t$(\mathbf{0}, \mathbf{\Sigma}_i, v)$, where v are the degrees of freedom, we have the Student-t Joint Modeling Model as presented in [11], and if we consider that $\mathbf{r}_i \sim$ Laplace$(\mathbf{0}, \mathbf{\Sigma}_i)$, we have the Laplace Joint Modeling Model structure as presented in [13]. Considering that all these structures belong to the family of SMN distributions, the purpose of this paper is to generalize these works that considers the Normal, Student-t and Laplace distributions for the residuals by assuming that $\mathbf{r}_i \sim$ SMN$(\mathbf{0}, \mathbf{\Sigma}_i, \tau_i)$. For that, we will assume that parameters τ_i are deterministic, but unknown. By not imposing any explicit structure for their probability density function, the proposed regression model will be robust since it has the ability to represent a wide kind of distributions belonging to the SMN family including the Normal, Student-t and Laplace structures as special cases. However, note that the SMN model presented in equations (8) to (10) without any constraint on the scatter matrix (except the trace) cannot be directly employed for the regression of longitudinal data. With such model, the number of parameters to estimate is much more larger than the total number of observations. The distribution parameters for this regression model cannot be estimated. A constraint on the scatter matrix need to be imposed on the shape of scatter matrix and in this paper, the MCD decomposition is considered.

In the following section, we present the proposed model, linking the SMN distribution to the regression model for longitudinal data, considering the reparameterization for the scatter matrix using the MCD decomposition.

3. SMN Joint Modeling Model

3.1. Structure

Considering the stochastic representation defined in (1) and assuming that $\mathbf{r}_i|\tau_i \sim \mathcal{N}(\mathbf{0}, \tau_i\mathbf{\Sigma}_i)$, then the log-likelihood function for a sample of m individuals $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^\top$, for $i = 1, 2, ..., m$, with $\theta = (\mu, \tau, \mathbf{\Sigma})$ is given

by

$$
\begin{aligned}
l(\theta) &= -\frac{1}{2}\ln(2\pi)\sum_{i=1}^{m} n_i - \frac{1}{2}\sum_{i=1}^{m} n_i \ln \tau_i - \frac{1}{2}\sum_{i=1}^{m} \ln|\mathbf{\Sigma}_i| \\
&- \sum_{i=1}^{m} \frac{1}{2\tau_i}(\mathbf{y}_i - \mu_i)^\top \mathbf{\Sigma_i}^{-1}(\mathbf{y}_i - \mu_i). \qquad (20)
\end{aligned}
$$

Assuming a regression structure with repeated measures, consider $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^\top$ the result vector of i-th sample unit, where n_i is the number of results of the subject i, $i = 1, 2, ..., m$. Here, we allow dependence (correlation) within the subject, but impose independence between subjects.

3.2. Parameter estimation

In this section, we will present the maximum likelihood estimates (MLE) of the parameter vector $\theta = (\beta, \tau, \gamma, \lambda)^\top$ of the proposed model, assuming that we have independent observations of m subjects $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m$ from model (12), considering the log-likelihood function given in (20) and a two-step estimation process, given that we assume that $\mathbf{\Sigma}_i$ is known to obtain the expressions of the other estimates.

3.3. Score functions

Taking the first derivative of (20), the maximum likelihood estimating equations for the parameter vector $\theta = (\beta, \tau, \gamma, \lambda)^\top$ of the proposed structure are given by

$$
\begin{aligned}
\mathbf{U}_\tau &= -\frac{n_i}{2\tau_i} + \frac{1}{2\tau_i^2}(\mathbf{y}_i - \mathbf{X}_i\beta)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta), \qquad (21) \\
\mathbf{U}_\beta &= \sum_{i=1}^{m} \frac{1}{\tau_i}\mathbf{X}_i^\top \mathbf{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta), \qquad (22) \\
\mathbf{U}_\gamma &= \sum_{i=1}^{m} \frac{1}{\tau_i}\mathbf{Z}_i^\top \mathbf{D}_i^{-1}(\mathbf{r}_i - \mathbf{Z}_i\gamma), \qquad (23) \\
\mathbf{U}_\lambda &= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{w}_j + \frac{1}{2}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{j=1}^{n_i} \sigma_j^{-2}(r_{ij} - \hat{r}_{ij})^2 \mathbf{w}_j, \qquad (24)
\end{aligned}
$$

where $\hat{\mathbf{r}}_i = \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{jk}^\top\gamma$, with the $n_i \times d$ matrix $\mathbf{Z}_i$, given by

$$
\mathbf{Z}_i = \left[\begin{array}{ccc} \mathbf{z}(i,1) & \cdots & \mathbf{z}(i,n_i) \end{array} \right]^\top,
$$

with $\mathbf{z}(i,j) = \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{jk}$, considering that $\hat{\mathbf{r}}_i = \mathbf{Z}_i\gamma$ (see appendix Appendix A for more details).

3.4. Estimation algorithm

Equating the scoring functions $\mathbf{U}_\tau$, $\mathbf{U}_\beta$ and $\mathbf{U}_\gamma$ to zero and rearranging them, we can write the following update equations

$$\hat{\tau}_i^{(h+1)} = \frac{(\mathbf{y}_i - \mathbf{X}_i\hat{\beta}^{(h)})^\top {\boldsymbol{\Sigma}_i^{(h)}}^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\beta}^{(h)})}{n_i} \tag{25}$$

$$\hat{\beta}^{(h+1)} = \left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i^{(h)}}\mathbf{X}_i^\top {\boldsymbol{\Sigma}_i^{(h)}}^{-1}\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i^{(h)}}\mathbf{X}_i^\top {\boldsymbol{\Sigma}_i^{(h)}}^{-1}\mathbf{y}_i\right] \tag{26}$$

$$\hat{\gamma}^{(h+1)} = \left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i^{(h)}}\hat{\mathbf{Z}}_i^{(h)^\top}\hat{\mathbf{D}}_i^{(h)^{-1}}\hat{\mathbf{Z}}_i^{(h)}\right]^{-1}\left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i^{(h)}}\hat{\mathbf{Z}}_i^{(h)^\top}\hat{\mathbf{D}}_i^{(h)^{-1}}\mathbf{r}_i^{(h)}\right], \tag{27}$$

where superscript (h) is the notation to indicate the parameter estimation in step h of the iterative estimation process.

The score function of the parameter λ depends on the parameter of interest. Thus, we use the Newton-Raphson method as an iterative estimation method, whose expression for the case of the λ parameter is given by

$$\hat{\lambda}^{(h+1)} = \hat{\lambda}^{(h)} + \left(\mathbf{H}_{\lambda\lambda}^{(h)}\right)^{-1}\mathbf{U}_\lambda^{(h)}, \tag{28}$$

where $\mathbf{U}_\lambda^{(h)}$ is a $d \times 1$ vector given the expression presented in (A.10) and $\mathbf{H}_{\lambda\lambda}^{(h)}$ is a $d \times d$ matrix given by

$$\mathbf{H}_{\lambda\lambda} = \frac{\partial}{\partial\lambda^2}l(\theta) = -\frac{1}{2}\sum_{i=1}^{m}\frac{1}{\tau_i}\left(\sum_{j=1}^{n_i}\sigma_j^{-2}(r_{ij} - \hat{r}_{ij})^2\mathbf{w}_j\mathbf{w}_j^\top\right). \tag{29}$$

To summarize, the steps to estimate the parameters of the proposed SMN-JMM model with deterministic but unknown τ_i are presented in Algorithm 1.

We can observe by means of these update equations, that observations that are outliers (i.e. observations for which $\mathbf{y}_i$ is far from $\mathbf{X}_i\beta$) will have a higher value for the estimate of τ_i and a lower weight for the estimates of β, γ and λ, as these estimators are divided by τ_i, making the approach robust. It is also important to note that if τ_i is constant, then the equations

Algorithm 1 Algorithm for the maximum likelihood estimate of the proposed SMNJMM model

1: Given a starting value $\theta^{(h)} = (\beta^{(h)}, \tau^{(h)}, \gamma^{(h)}, \lambda^{(h)})^\top$, which can be obtained by fitting the NJMM model, use the models given in (17) to form $\mathbf{Z}_i^{(h)}$, $\hat{\mathbf{r}}_i^{(h)}$, $\mathbf{L}_i^{(h)}$, $\mathbf{D}_i^{(h)}$, and $\mathbf{\Sigma}_i^{(h)}$;
2: Update $\hat{\tau}_i^{(h+1)}$ using (25);
3: Update $\hat{\beta}^{(h+1)}$ using (26);
4: Update $\hat{\gamma}^{(h+1)}$ using (27);
5: Update $\hat{\lambda}^{(h+1)}$ using (28);
6: Update the following matrices: $\mathbf{Z}_i^{(h+1)}$, using (25), $\hat{\mathbf{r}}_i^{(h+1)}$, using (A.3), $\mathbf{L}_i^{(h+1)}$ using (14), $\mathbf{D}_i^{(h+1)}$, using (15), and $\mathbf{\Sigma}_i^{(h+1)}$, using (13);
7: Normalization of the scatter matrix, considering $\mathbf{\Sigma}_i^{(h+1)} = \frac{n_i \mathbf{\Sigma}_i^{(h+1)}}{tr(\mathbf{\Sigma}_i^{(h+1)})}$;
8: Repeat steps 2 to 7 until convergence.

reduce for the case of the model with normal structure (NJMM), following the representation presented in (1), i.e. the structure reduces to the stochastic representation of a Gaussian distribution.

The SMN model presented in equations (8) to (10) without any constraint on the scatter matrix (except the trace) and the proposed SMNJMM model share many similarities. Actually, the only difference relies on the structure of the scatter matrix. For this application of regression with longitudinal data, we need to impose this structure. As these models are quite closed each other, we retrieve some similarities during the estimation process. For example:

- Equation (8) for the SMN model is retrieve as step 3 in Algorithm 1 (equation (26), estimation of the regression coefficients to obtain the mean vector).
- Equation (9) for the SMN model (estimation of Σ) is replaced by steps 4 and 5 in Algorithm 1 (equations (27) and (28) to estimate γ and λ).
- Equation (10) for the SMN model (estimation of τ_i) is retrieved by step 2 in Algorithm 1 (equation (25)). Note that it is exactly the same equation with the quadratic term on the numerator.
- For identification reason, as explained in the paragraph before (11), the scatter matrix should be normalized. In the proposed SMNJMM

model, we have also considered that its trace is equal to n (see step 7 in Algorithm 1).

3.5. Forward prediction

In this section, we present a technique to predict the future values of the individual i ($\mathbf{y}_i$), which is one of the m individuals under consideration, for the proposed model (SMNJMM), by following the same approach as the one developed for the TJMM [11] and LJMM [13] models, i.e. formulate an estimation structure capable of predicting future values of the i-th sampling unit, which is one of the m individuals under study, assuming that the values already observed and the new predictions follow the multivariate distribution under consideration.

For this, consider that $\tilde{\mathbf{y}}_i$ denotes a future vector, $(g\times 1)$, of measurements for the individual $\mathbf{y}_i$, $\tilde{\mathbf{X}}_i$ denotes a matrix, $(g \times (p+1))$, of prediction for the corresponding $\tilde{\mathbf{y}}_i$ with $\mathbb{E}(\tilde{\mathbf{y}}_i) = \tilde{\mathbf{X}}_i\beta$, and $\tilde{\mathbf{z}}_{jk}$ and $\tilde{\mathbf{w}}_j$ denote the vectors of the covariates associated with $\tilde{\mathbf{y}}_i$, of dimension $d \times 1$ and $q \times 1$, being $\ln \sigma_i^2 = \tilde{\mathbf{w}}_i^\top \lambda$ and $\phi_{jk} = \tilde{\mathbf{z}}_{jk}^\top \gamma$, for $j \in \{n_i+1, ..., n_i+g\}$ and $k \in \{1, ..., j-1\}$, respectively.

Furthermore, we assume that the joint distribution of $\mathbf{y}_i$ and $\tilde{\mathbf{y}}_i$ given τ_i follows a (n_i+g)-variate normal distribution, i.e.

$$\left[\begin{array}{c} \mathbf{y}_i \\ \tilde{\mathbf{y}}_i \end{array} \right] |\tau_i \sim \mathcal{N}_{n_i+g}(\mathbf{X}_i^*\beta, \tau_i\boldsymbol{\Sigma}_i^*), \tag{30}$$

where $\mathbf{X}_i^* = (\mathbf{X}_i^\top, \tilde{\mathbf{X}}_i^\top)^\top$, $\boldsymbol{\Sigma}_i^{*-1} = \mathbf{L}_i^{*-1}\mathbf{D}_i^{*-1}\mathbf{L}_i^*$, $\mathbf{D}_i^* = \text{diag}(\sigma_1^2, ..., \sigma_{n_i+g}^2)$ and $\mathbf{L}_i^* = [l_{jk}]$ are unitary lower triangular matrices, $(n_i+g) \times (n_i+g)$, with 1's as diagonal entries and (j,k)-th entry being $-\phi_{jk}$.

Partitioning $\boldsymbol{\Sigma}_i^*$ according to the size of $\mathbf{y}_i^* = (\mathbf{y}_i^\top, \tilde{\mathbf{y}}_i^\top)^\top$, which is a vector that joins current observations and future observations, we have

$$\boldsymbol{\Sigma}_i^* = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right], \tag{31}$$

where $\boldsymbol{\Sigma}_{11}$ is the upper left $(n_i \times n_i)$ submatrix of $\boldsymbol{\Sigma}_i^*$ and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$ (here we suppress the subscript i of the partitioned matrices for notational convenience).

Considering the characteristics of the marginal and conditional distributions [20, 21], we have that if $\mathbf{y}_i|\tau_i \sim \mathcal{N}_{n_i}(\mu_i, \tau_i\boldsymbol{\Sigma}_i)$, then

$$\tilde{\mathbf{y}}_i|\mathbf{y}_i, \tau_i \sim \mathcal{N}_g\left(\mu_{2\cdot 1}, \tau_i\boldsymbol{\Sigma}_{22\cdot 1}\right), \tag{32}$$

where $\mu_{2\cdot1} = \mathbb{E}(\tilde{\mathbf{y}}_i|\mathbf{y}_i, \tau_i)$ and $\boldsymbol{\Sigma}_{22\cdot1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, and the Mean Squared Error (MSE) error predictor of $\tilde{\mathbf{y}}_i$, which is the conditional expectation of $\tilde{\mathbf{y}}_i$ given $\mathbf{y}_i$ and τ_i, as shown by [13], is given by

$$\hat{\tilde{\mathbf{y}}}_i = \mu_{2\cdot1} = \tilde{\mathbf{X}}_i\beta + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta), \tag{33}$$

and the MSE covariance matrix for the predictor (33) is determined by

$$\mathbb{E}[(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)^\top|\tau_i] = \mathbb{E}(\mathrm{cov}(\tilde{\mathbf{y}}_i|\mathbf{y}_i)|\tau_i) = \tau_i\boldsymbol{\Sigma}_{22\cdot1}. \tag{34}$$

4. Experiments

In order to validate the proposed model, in this section we will present the main results of some applications on synthetic (estimation and prediction performance) and real data.

4.1. Synthetic data

4.1.1. ML estimation performance for mixing parameter

The first evaluation scenario consists of illustrating that the proposed model is suitable for any distribution chosen for the parameter τ of the SMN model (see (1)). In this context, we consider that τ has the following distributions: inverse gamma, beta and inverse beta. For this, 100 samples were generated for each scenario, considering that the structure that generate the mixing parameter τ have the same mean.

To compare these models, we consider the use of the BIC metric (Bayesian Information Criterion), proposed by [22], which is given by

$$\mathrm{BIC} = -\frac{2}{m}\hat{l}_{max} + \frac{k}{m}\ln(m), \tag{35}$$

where $\hat{l}_{max}$ is the maximized log-likelihood with respect to the chosen model. For the proposed SMNJMM, the number of parameters k is equal to $p+d+q+m+1$ since $\beta \in \mathbb{R}^{p+1}$, $\gamma \in \mathbb{R}^d$, $\lambda \in \mathbb{R}^q$ and $\tau \in \mathbb{R}^m$. For the NJMM and LJMM models, $k = p+d+q+1$, since we do not have τ_i, and for the TJMM model, $k = p+d+q+2$ since the degree of freedom ν must be additionally estimated.

The BIC metric is a tradeoff between fit performance and model complexity. It is based on the likelihood function, being related to the Akaike Information Criterion (AIC), proposed by [23], and is used as a criterion for

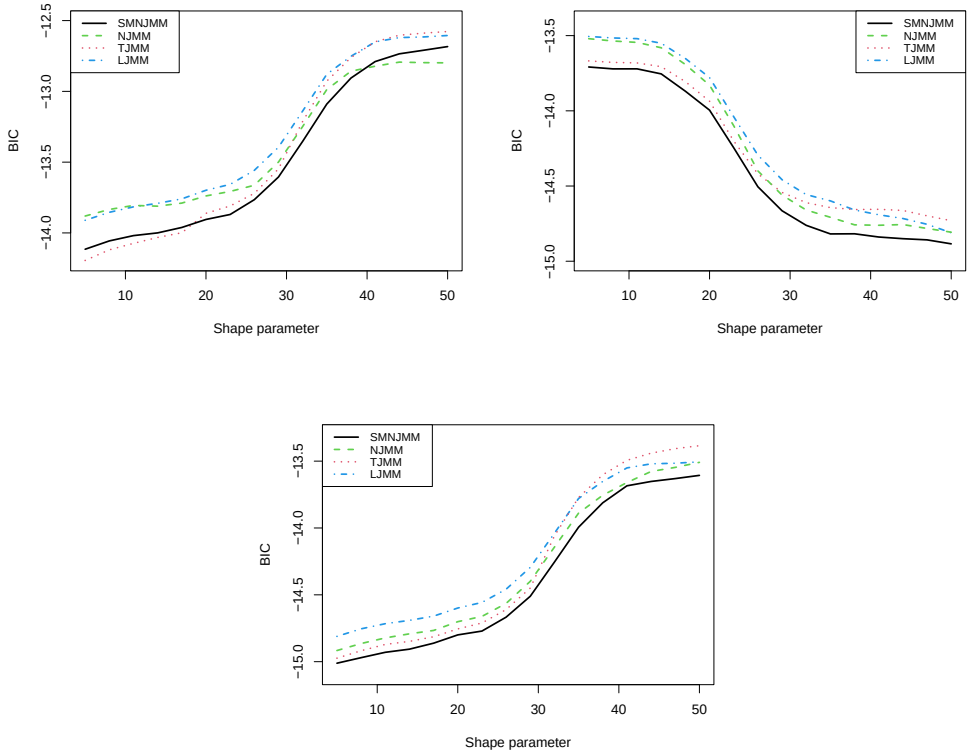


Figure 1: BIC metric value for each model obtained from the shape parameter. Top left: inverse gamma distribution, top right: beta distribution, bottom: inverse beta distribution.

selecting models from a finite set of other models. The rule for using this measure is to choose models with a lower BIC (the values being positive or negative).

Figure 1 shows the evolution of the BIC criterion for the three scenario presented before when τ follows an inverse gamma, a beta and an inverse beta distribution.

As observed, the proposed SMNJMM model has generally the lowest BIC value for each of these scenario indicating its great potential. More precisely, for the first scenario when τ follows an inverse gamma distribution, the residuals follow the Student-t distributions and unsurprisingly the best performance are obtained for the TJMM model displayed in red in the top

left plot of Figure 1 especially when the shape parameter (degree of freedom) has a low value. Note also that in this context, the performances of the proposed SMNJMM in black are very closed to the one obtained with TJMM. For the second and third scenario displayed in the top right and bottom plots of Figure 1, τ follows respectively a beta and an inverse beta distribution. In this case, the residuals follow much more complex distributions involving the Whittaker functions as presented in [24]. For these scenario, the NJMM, TJMM and LJMM models do not work well for low shape parameters. However, as expected since we have designed it for its robust behavior, the proposed SMNJMM model exhibits the best performances. It performs well whatever the considered distribution for τ. Note also that for a large shape parameter, all these models converge towards the Normal distribution and in this case, the best performance are naturally obtained with the NJMM model.

4.1.2. ML estimation performance result

The objective of this subsection is to compare the performance of the maximum likelihood estimates of the proposed model with that of recent models. For this, we generate $m = 100$ balanced random samples (with $n_i = 12$ observations for each individual) from the Normal, Student-t (with 4 degrees of freedom) and Laplace distributions, with $\beta = [5, 27, 0.16]^\top$, $\gamma = [1.12, -0.58, 0.09, -0.0043]^\top$ and $\lambda = [-0.4, -1.37, 0.18, -0.007]^\top$, that is, $d = 4$, $q = 4$, $p = 1$, considering the design matrix $\mathbf{X}_i$ for the mean, the covariates $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ for autoregressive parameters and chosen scale innovation variances as follows

$$\mathbf{X}_i = \left[\begin{array}{cc} \mathbf{1} & \mathbf{k} \end{array} \right]^\top, \quad \mathbf{z}_{jk} = \left[\begin{array}{cccc} \mathbf{1} & (j-k) & (j-k)^2 & (j-k)^3 \end{array} \right]^\top \tag{36}$$

and

$$\mathbf{w}_j = \left[\begin{array}{cccc} 1 & j & j^2 & j^3 \end{array} \right]^\top, \tag{37}$$

where $\mathbf{k} = [0, 1, 2.5, 3.5, 4.5, 6, 7, 8, 10, 11.5, 13, 14]^\top$, following the same simulation structure presented in [11] and [13]. It is important to say that, as described in the literature for the TJMM and LJMM models, we have that $\mathbf{z}(i, 1) = 0$, so that the first line of $\mathbf{Z}_i$ is zero. To compute the standard error, the simulations were run with a total of 500 replications.

To evaluate the models, we consider three scenarios: data generated from the Normal, from the Student-t and from the Laplace distributions. The

results are presented in Tables 1–3 (the best result is highlighted in blue and the second best result is highlighted in red), with the mean values of estimates and standard errors (SE) of each model parameter, where $SE = \sigma/\sqrt{r}$, with σ the standard deviation of the estimated parameter and r the number of replications.

According to Table 1, where data are generated from the Normal distribution, the model that best fits the data is the NJMM model, followed by the proposed SMNJMM model. We also observed that the standard errors of the estimators are similar between the models, with emphasis on the values of the parameters λ_1, whose smallest standard error was observed in the estimates for the SMNJMM model.

To assess if there is a difference between the results obtained for each model, the pairwise Wilcoxon rank-sum test is employed [25]. We obtained a p-value greater than 5% for comparing BIC values between NJMM and SMNJMM, that is, there is no significant difference between the theoretical model and the proposed one. Furthermore, a p-value of less than 5% is obtained when comparing BIC values between SMNJMM versus TJMM, and SMNJMM versus LJMM models. That is, there is a significant difference between the models that use the Student-t and Laplace distributions, respectively, and the proposed one. These comparisons help us to validate the efficiency and robustness of the proposed model in relation to other models, when the data is generated from the Normal distribution structure.

Table 1: Parameter estimations and Standard Errors (SE) for NJMM, TJMM, LJMM and SMNJMM of the data generated data from Normal Distribution - Scenario 1.

Parameter	True value	NJMM [5, 6]		TJMM [11]		LJMM [13]		SMNJMM (proposed)	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	5.27	5.2791	0.0469	5.2742	0.0459	5.2779	0.0465	5.2813	0.0452
β_2	0.16	0.1893	0.0124	0.1998	0.0128	0.1907	0.0129	0.1905	0.0121
γ_1	1.12	1.1183	0.0581	1.1173	0.0578	1.1183	0.0638	1.1216	0.0605
γ_2	-0.58	-0.5691	0.0487	-0.5674	0.0481	-0.5688	0.0534	-0.5715	0.0489
γ_3	0.09	0.0934	0.0196	0.0931	0.0187	0.0932	0.0205	0.0931	0.0192
γ_4	-0.0043	-0.0046	0.0138	-0.0042	0.0133	-0.0046	0.0138	-0.0046	0.0139
λ_1	-0.4	-0.3883	0.2449	-0.3973	0.2414	-2.2568	0.2369	-0.4107	0.2432
λ_2	-1.37	-1.3771	0.1593	-1.3735	0.1569	-1.3562	0.1529	-1.3584	0.1627
λ_3	0.18	0.1861	0.0364	0.1859	0.0364	0.1828	0.0351	0.1822	0.037
λ_4	-0.007	-0.0033	0.0124	-0.0032	0.0124	-0.0033	0.0119	-0.0029	0.0121
BIC		-13.49		-12.67		-11.97		-13.01	

When data are generated from the Student-t distribution, the model that

best fits the data is the TJMM model, followed by the proposed SMNJMM, as shown in the results presented in Table 2. As in the models with data from the normal distribution, in the case of the Student-t distribution, we also observed that the standard errors of the estimators are similar between the models, with emphasis on the values of the parameters λ_1, λ_2, λ_3 and λ_4.

Again, using the pairwise Wilcoxon rank-sum test for comparison, we obtained a p-value greater than 5% for comparing BIC values between TJMM and SMNJMM, that is, there is no significant difference between the theoretical model and the proposed one. Furthermore, a p-value of less than 5% is obtained for comparing BIC values between the SMNJMM versus NJMM, and SMNJMM versus LJMM models, that is, there is a significant difference between the models that use the Normal and Laplace distributions, respectively, and the proposed one. These comparisons help us to validate the efficiency and robustness of the proposed model in relation to other models, when the data is generated from the Student-t distribution structure.

Table 2: Parameter estimations and Standard Errors (SE) for NJMM, TJMM, LJMM and SMNJMM of the data generated data from Student-t Distribution - Scenario 2.

Parameter	True value	NJMM [5, 6]		TJMM [11]		LJMM [13]		SMNJMM (proposed)	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	5.27	5.2798	0.0454	5.2796	0.0443	5.2809	0.0395	5.2781	0.0453
β_2	0.16	0.1895	0.0118	0.1904	0.0126	0.1891	0.0119	0.1907	0.0129
γ_1	1.12	1.1095	0.0913	1.1135	0.1017	1.1158	0.0668	1.1145	0.1027
γ_2	-0.58	-0.5601	0.0729	-0.5652	0.0837	-0.5657	0.0554	-0.5665	0.0842
γ_3	0.09	0.0914	0.0245	0.0928	0.0261	0.0923	0.0206	0.0934	0.0275
γ_4	-0.0043	0.0049	0.0139	0.0041	0.0134	0.0043	0.0132	0.0046	0.0141
λ_1	-0.4	-0.4302	0.4378	-0.4141	0.4755	-2.4824	0.3031	-0.4764	0.4367
λ_2	-1.37	-1.3745	0.2629	-1.3792	0.2742	-1.3616	0.1808	-1.3437	0.2814
λ_3	0.18	0.1861	0.0545	0.1861	0.0566	0.1828	0.0398	0.1812	0.0574
λ_4	-0.007	-0.0033	0.0134	-0.0031	0.0135	-0.0031	0.0124	-0.0029	0.0135
BIC		-11.05		-14.51		-12.36		-13.97	

When the data are generated from the Laplace distribution, the model that best fits the data is the LJMM model, followed by the proposed SMNJMM model, according to the results presented in Table 3. For the last comparison using the pairwise Wilcoxon rank-sum test, we obtained a p-value greater than 5% for comparing BIC values between LJMM and SMNJMM, that is, there is no significant difference between the theoretical model and the proposed one. Furthermore, a p-value of less than 5% is obtained for comparing BIC values between the SMNJMM versus NJMM, and SMNJMM

versus TJMM models, that is, there is a significant difference between the models that use the Normal and Student-t distributions, respectively, and the proposed one. These comparisons help us to validate the efficiency and robustness of the proposed model in relation to other models, when the data is generated from the Laplace distribution structure.

Table 3: Parameter estimations and Standard Errors (SE) for NJMM, TJMM, LJMM and SMNJMM of the data generated data from Laplace Distribution - Scenario 3.

Parameter	True value	NJMM [5, 6]		TJMM [11]		LJMM [13]		SMNJMM (proposed)	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	5.27	5.2829	0.1046	5.2771	0.1039	5.2851	0.1021	5.2778	0.0463
β_2	0.16	0.1891	0.0175	0.1903	0.0169	0.1906	0.0179	0.1911	0.0121
γ_1	1.12	1.1178	0.0616	1.1161	0.0601	1.1192	0.0597	1.1175	0.0611
γ_2	-0.58	-0.5690	0.0525	-0.5661	0.0512	-0.5671	0.0492	-0.5669	0.0509
γ_3	0.09	0.0935	0.0209	0.0935	0.0192	0.0931	0.0199	0.0934	0.0191
γ_4	-0.0043	0.0041	0.0141	0.0045	0.0135	0.0051	0.0132	0.0045	0.0133
λ_1	-0.4	1.4061	0.2508	1.4275	0.2253	-0.4135	0.2457	-0.3905	0.2391
λ_2	-1.37	-1.3539	0.1601	-1.3683	0.1475	-1.3594	0.1591	-1.3771	0.1603
λ_3	0.18	0.1831	0.0355	0.1858	0.0341	0.1838	0.0365	0.1861	0.0374
λ_4	-0.007	-0.0031	0.0125	-0.0032	0.0125	-0.0031	0.0125	-0.0032	0.0126
BIC		-12.32		-12.27		-14.87		-13.23	

Thus, as expected, the proposed SMNJMM model performs well whatever the distribution used for data generation, as long as this generating distribution belongs to the SMN family, which allows it to be used for a wide range of cases, making the model robust.

4.1.3. Prediction performance

We performed one more simulation study to evaluate the prediction performance of NJMM, TJMM, LJMM and SMNJMM models. To measure predictive accuracies, we report the empirical mean square forward prediction error (MSFE), defined as

$$\text{MSFE} = \frac{1}{mg}\sum_{i=1}^{m}(\hat{\tilde{\mathbf{y}}}_i - \mathbf{y}_i)^\top(\hat{\tilde{\mathbf{y}}}_i - \mathbf{y}_i), \tag{38}$$

where m is the number of individuals, g is the number of steps forward for prediction, $\tilde{\mathbf{y}}_i$ is the predicted estimated vector of the response variables and $\mathbf{y}_i$ is the observed vector of the response variable.

The technique we used to obtain $\tilde{\mathbf{y}}_i$ and assess the relative predictive performance of different models considers the pseudo-cross-validation approach, which combines the holdout cross-validation [26] and a predictive sample reuse [27], as follows: for each individual i, for $i = 1, ..., m$, we do (i) holdout the last g measurements on the i th participant; (ii) compute ML estimates using the remaining data as the sample; (iii) prediction of the g true measurements $\mathbf{y}_i = (y_{i,n_i-g+1}, ..., y_{i,n_i})^\top$, denoted by $\hat{\tilde{\mathbf{y}}}_i = (\hat{\tilde{y}}_{i,n_i-g+1}, ..., \hat{\tilde{y}}_{i,n_i})^\top$, is made by using (33).

To evaluate the predictive power of each model, we generated $m = 100$ balanced random samples observed at $n_i = 12$ different times (the number of observations is the same for each individual) from the Normal, Student-t (with 4 degrees of freedom) and Laplace distributions, and we estimated the parameters considering the previously mentioned pseudo-cross-validation approach. To compare the state of the art models with the proposed SMNJMM, we used the Relative Improvement Percentage (RIP) as classically employed in the literature [11, 13]. It is defined as:

$$RIP = \frac{MSFE_{model} - MSFE_{SMNJMM}}{MSFE_{model}} * 100\%,$$

where $MSFE_{model}$ can be the MSFE of NJMM, TJMM, or LJMM. The obtained results are shown in Table 4.

Table 4: MSFE and RIP (%)* values for NJMM, TJMM, LJMM and SMNJMM models of g step-ahead forecasts.

Generating data from	Model	Metric	g step-ahead forecasts					
			g = 1	g = 2	g = 3	g = 4	g = 5	g = 6
Normal	NJMM	**MSFE**	**0,0510**	**0,0731**	**0,1034**	**0,1137**	**0,1309**	**0,1740**
		RIP (%)	0,1961	0,1368	0,0967	0,0000	-1,7571	-2,2989
	TJMM	**MSFE**	**0,0516**	**0,0734**	**0,1075**	**0,1139**	**0,1339**	**0,1792**
		RIP (%)	1,3566	0,5450	3,9070	0,1756	0,5228	0,6696
	LJMM	**MSFE**	**0,0512**	**0,0736**	**0,1063**	**0,1145**	**0,1333**	**0,1830**
		RIP (%)	0,5859	0,8152	2,8222	0,6987	0,0750	2,7322
	SMNJMM	**MSFE**	**0,0509**	**0,0730**	**0,1033**	**0,1137**	**0,1332**	**0,1780**
Student-t	NJMM	**MSFE**	**0,0480**	**0,0567**	**0,0953**	**0,0954**	**0,1670**	**0,1895**
		RIP (%)	2,0833	1,0582	0,4197	0,3145	0,1198	1,1609
	TJMM	**MSFE**	**0,0472**	**0,0549**	**0,0951**	**0,0953**	**0,1664**	**0,1865**
		RIP (%)	0,4237	-2,1858	0,2103	0,2099	-0,2404	-0,4290
	LJMM	**MSFE**	**0,0481**	**0,0580**	**0,1004**	**0,0963**	**0,1675**	**0,1875**
		RIP (%)	2,2869	3,2759	5,4781	1,2461	0,4179	0,1067
	SMNJMM	**MSFE**	**0,0470**	**0,0561**	**0,0949**	**0,0951**	**0,1668**	**0,1873**
Laplace	NJMM	**MSFE**	**0,0481**	**0,1104**	**0,1379**	**0,1444**	**0,1635**	**0,1898**
		RIP (%)	0,4158	9,1486	2,7556	1,3850	0,6116	5,0053
	TJMM	**MSFE**	**0,0480**	**0,1171**	**0,1379**	**0,1444**	**0,1639**	**0,1819**
		RIP (%)	0,2083	14,3467	2,7556	1,3850	0,8542	0,8796
	LJMM	**MSFE**	**0,0477**	**0,1083**	**0,1345**	**0,1431**	**0,1615**	**0,1791**
		RIP (%)	-0,4193	7,3869	0,2974	0,4892	-0,6192	-0,6700
	SMNJMM	**MSFE**	**0,0479**	**0,1003**	**0,1341**	**0,1424**	**0,1625**	**0,1803**

We can observe that the MSFE metric increases as the number of steps for the prediction increases. This behavior is common for all steps presented, that is, when $g = 1$ (when we remove the last observation of each individual and use the prediction to estimate this value), when $g = 2$ (when we remove the last two observations of each individual and use prediction to estimate these values), and so on.

The proposed model (SMNJMM) generally presents better prediction results up to the step $g = 4$ for the three presented scenarios. After this prediction step ($g = 5$ and $g = 6$), the best model for prediction is the one generated with the same structure. Table 4 highlights these results.

In the next section, we will evaluate the proposed model on a real dataset and compare its performance with other models with similar structure (NJMM, TJMM and LJMM).

4.2. Real data

4.2.1. Dataset description

To illustrate the application of the model on real data, we will consider the same dataset used in [11], for the TJMM model, and in [13], for the

LJMM model. It is worth noting that these data were originally analyzed in [28].

The dataset is open source, available in the package `nlme` [29] in the R software [30] and shows tooth growth from pituitary to pterygomaxillary cleft distances (mm) measured repeatedly at ages 8, 10, 12 and 14 years of 11 females and 16 males.

4.2.2. Fitting SMNJMM model

Our objective is to compare the performance of the NJMM, TJMM and LJMM models, with the proposed SMNJMM model for this dataset, considering the average response for a linear growth function that can be described with the Poly(1,1) model. The dependency structure is as follows

$$\mu_{ij} = \beta_0 + \delta_0 I_i(F) + (\beta_1 + \delta_1 I_i(F))t_j, \tag{39}$$

with $i \in \{1, ..., m = 27\}$ and $j \in \{1, ..., n_i = 4\}$, where μ_{ij} is the average orthodontic distance for the i-th subject at age t_j; β_0 and β_1 the intercept and slope for boys, respectively; δ_0 and δ_1 the difference in intercept and slope between girls and boys, respectively; $I_i(F)$ the indicator variable for the girl.

Corresponding ML estimates in Normal, Student-t, Laplace and SMN configurations can be found in Table 5. We observed that the SMNJMM model provides a lower BIC value than the other models, that is, the results of the data analysis support the effectiveness of the proposed model.

Table 5: Parameter estimations of Poly(1,1) model for NJMM, TJMM, LJMM and SMNJMM of the orthodontic growth data.

Parameter	NJMM	TJMM	LJMM	SMNJMM
β_0	20,9738	15,2569	15,5197	21,5423
β_1	1,6117	0,7955	0,7782	1,4254
δ_0	-0,6947	0,9785	1,0752	2,3120
δ_1	-0,6572	-0,3127	-0,3161	0,5214
γ_1	0,8498	0,9894	0,9710	1,0685
γ_2	-0,2537	-0,3781	-0,3512	-0,3781
λ_1	1,9268	1,5298	-0,5125	1,9951
λ_2	-0,2846	-0,3714	-0,3325	-0,3124
BIC	9,5935	12,5241	11,8510	8,5211

In order to evaluate the prediction results obtained from the four models under study (NJMM, TJMM, LJMM and SMNJMM) for orthodontic growth

data, we present Table 6 which shows the prediction, using the MSFE metric, for $g = 1$, $g = 2$ and $g = 3$, that is, removing the last, the last two and the last three observations of each individual. From the presented values, we can say that the SMNJMM model provides lower MSFE values for the orthodontic growth dataset.

These favorable results for the proposed SMNJMM model show that we do not need to impose any fixed shape on the residue distribution, as is done in the attributes for the NJMM, TJMM and LJMM models. This allows the SMNJMM model to be more flexible and therefore achieve better estimation and prediction performances, as illustrated in this real dataset.

Table 6: MSFE values for NJMM, TJMM, LJMM and SMNJMM models to compare the predictive accuracies of g step-ahead forecasts.

Model	g step-ahead forecasts		
	g = 1	g = 2	g = 3
NJMM [5, 6]	2,3136	4,4404	8,4312
TJMM [11]	2,3135	4,4121	8,4333
LJMM [13]	2,3108	4,4357	8,2111
SMNJMM (proposed)	2,3104	4,4064	8,2089

4.2.3. Diagnostic analysis

To evaluate the robustness of the estimates of the proposed SMNJMM model compared to NJMM, TJMM and LJMM models, we adopt the experiments introduced in [11] and [13]. It consists in a two step procedure detailed as below:

(i.) Add contaminated values, denoted by ξ, to a single observation, doing $y_{ij}(\xi) = y_{ij} + \xi$;

(ii.) Reestimate the models and collect the corresponding BIC values.

These two steps are applied to three different observations: the last observations of the first female, the last observation of the last male and a random observation for the different values of ξ between -20 and 20 by the increment of 2 mm. Figure 2 shows the BIC versus perturbation values (ξ) for the NJMM, TJMM, LJMM and SMNJMM models for the three aforementioned cases: a woman, a man and a random subject.

We can notice that the BIC values of the NJMM, TJMMM and LJMMM models are always greater than the BIC values of the SMNJMM model.

This difference tends to increase steadily with the increase in the perturbation value for each of these observations, mainly for the TJMM and LJMM models. For the SMNJMM model, perturbation values ranging from -20 to 20 have little effect on the BIC value of each fitted model, indicating that our model is the most robust one.

These results show the robustness of the model, which, in addition to not needing to impose any fixed shape on the distribution of residuals, is also capable of satisfactorily accommodating extreme observations.

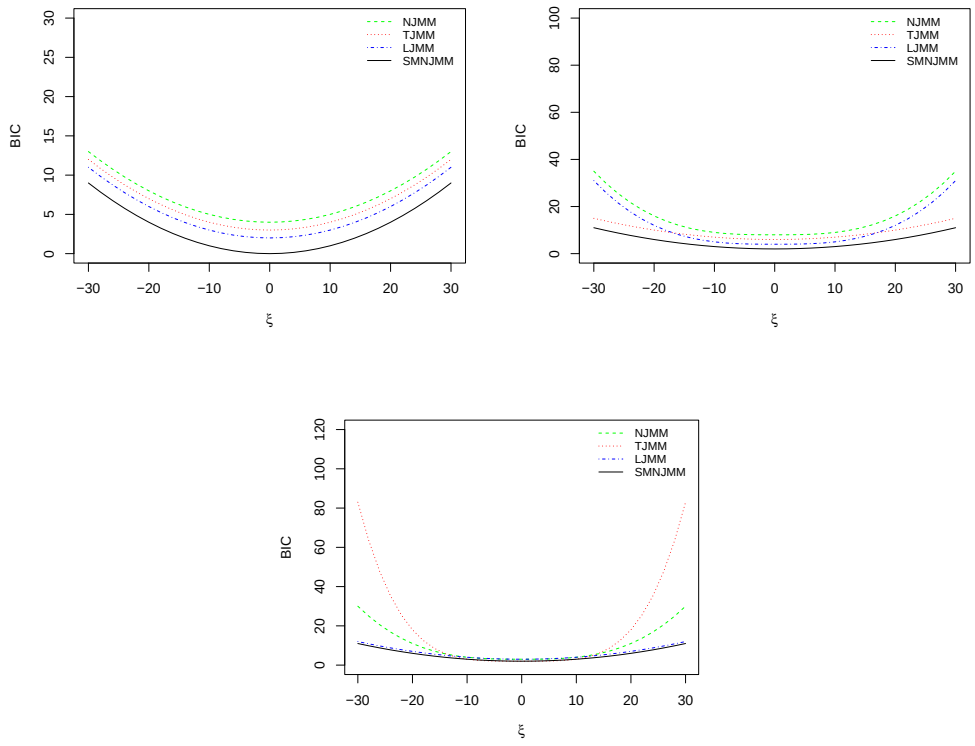


Figure 2: BIC versus perturbations. Top left: female observation, top right: male observation, bottom: a random observation.

4.3. Advantages and disadvantages of the proposed model

By considering that the mixing parameters τ_i are deterministic but unknown parameters, we do not assume any specific distribution. Our model

has hence the ability to represent any model issued from the family of scale mixture of Normal distributions. The proposed SMNJMM model can hence be viewed as a robust extension of the state-of-the-art models based on the Normal, Student-t and Laplace distributions. In addition, from a practical application, the interest is that it is no more necessary to perform goodness of fit to validate the fitting quality of the proposed model since our SMNJMM model generalizes a wide kind of distributions. Furthermore, we consider the joint estimation approach of the within-subject marginal covariance matrix, with the aim of overcoming common problems in these processes, such as the positive definition of the scatter matrix and number of parameters to be estimated.

The limitation of our proposal is considering that one of the matrices that form the MCD decomposition is formed by generalized autoregressive parameters (GARPs), that is, an autoregressive structure is assumed for the measurements of each sampling unit. A proposal that considers a moving average structure or a decomposition in the correlation matrix [31, 10] are alternatives already developed for Gaussian models and that can be considered for the approach with the SMN distribution.

5. Conclusion

We present a robust approach to the analysis of longitudinal data with the mean and the scale covariance modeled jointly in a framework of SMN errors. This approach generalizes recent models that are based on the Normal, Student-t and Laplace distributions, by using the distribution-free τ mixing parameter (unknown, but deterministic).

From the MCD decomposition, we derive maximum likelihood estimators and provide an algorithm to estimate model parameters. We present estimators for forward prediction, providing an explicit expression of the root mean square error predictor. Based on simulated and real data, we introduce extensive experiments to illustrate the robustness of the proposed model[1].

From the results obtained, we observed that the SMNJMM model provides a lower BIC value than the other models, that is, the results of the data analysis support the effectiveness of the proposed model. To evaluate the prediction of new observations for orthodontic growth data, we saw that,

[1]R codes for using the proposed model can be obtained from the authors upon request.

from the values presented, the SMNJMM model provides the lowest MSFE values.

Future works will include the development of diagnostic techniques for this type of scatter matrix reparameterization, such as global and local influence analysis and disturbance in the scatter matrix, are tools that can be developed in future works for the diagnostic evaluation of the proposed model.

Appendix A. Estimates of each parameter

MLE of τ and β

Taking the derivative of the SMN likelhood function with respect to the parameters τ_i and β, we have that the respective score functions are given by

$$\mathbf{U}_\tau = -\frac{n_i}{2\tau_i} + \frac{1}{2\tau_i^2}(\mathbf{y}_i - \mathbf{X}_i\beta)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta), \tag{A.1}$$

$$\mathbf{U}_\beta = \sum_{i=1}^{m} \frac{1}{\tau_i}\mathbf{X}_i^\top \mathbf{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta). \tag{A.2}$$

MLE of γ

For the estimation of the parameters γ and λ, we need to consider some changes in the structure of the log-likelihood function, in order to expose the presence of the parameters of interest in the estimation, as we will present below.

First, we consider that the residuals of the i-th individual are obtained from the difference $\mathbf{r}_i = \mathbf{y}_i - \mu_i$, where $\mu_i = \mathbf{X}_i\beta$. Thus, the estimate for this parameter can be obtained from

$$\hat{\mathbf{r}}_i = \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{jk}^\top \gamma, \tag{A.3}$$

where $\hat{\mathbf{r}}_i$ is the predictor of r_{ij} based on its predecessors $r_{i,j-1}, ..., r_{i,1}$, i.e., the AR model for the actual measurements on subject i, as presented in (16). Also, using the fact that $\mathbf{\Sigma}_i^{-1} = \mathbf{L}_i^\top \mathbf{D}_i^{-1}\mathbf{L}_i$ and $\mathbf{L}_i\mathbf{r}_i = (\mathbf{r}_i - \hat{\mathbf{r}}_i)$, as presented in [5], we have that the log-likelihood function (20) can be rewritten (without the constant term) as follows

$$\begin{aligned} l(\theta) &\propto -\frac{1}{2}\sum_{i=1}^{m} n_i \ln \tau_i - \frac{1}{2}\sum_{i=1}^{m} \ln|\mathbf{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m}\left\{\frac{1}{\tau_i}\mathbf{r}_i^\top(\mathbf{L}_i^\top \mathbf{D}_i^{-1}\mathbf{L}_i)\mathbf{r}_i\right\} \\ &= -\frac{1}{2}\sum_{i=1}^{m} n_i \ln \tau_i - \frac{1}{2}\sum_{i=1}^{m} \ln|\mathbf{\Sigma}_i| \\ &- \frac{1}{2}\sum_{i=1}^{m}\left\{\frac{1}{\tau_i}(\mathbf{r}_i - \hat{\mathbf{r}}_i)^\top \mathbf{D}_i^{-1}(\mathbf{r}_i - \hat{\mathbf{r}}_i)\right\}. \end{aligned} \tag{A.4}$$

Making some changes to the expression (A.4) in order to include the parameter γ, we consider the matrix $\mathbf{Z}_i$, of dimension $n_i \times d$, given by

$$\mathbf{Z}_i = \left[\begin{array}{ccc} \mathbf{z}(i,1) & \cdots & \mathbf{z}(i,n_i) \end{array} \right]^\top, \tag{A.5}$$

with $\mathbf{z}(i,j) = \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{jk}$. Furthermore, we can consider that $\hat{\mathbf{r}}_i = \mathbf{Z}_i\gamma$, such that the function (A.4) can be rewritten as follows

$$\begin{aligned} l(\theta) &\propto -\frac{1}{2}\sum_{i=1}^{m} n_i \ln \tau_i - \frac{1}{2}\sum_{i=1}^{m} \ln |\mathbf{\Sigma}_i| \\ &- \frac{1}{2}\sum_{i=1}^{m}\left\{ \frac{1}{\tau_i}(\mathbf{r}_i - \mathbf{Z}_i\gamma)^\top \mathbf{D}_i^{-1}(\mathbf{r}_i - \mathbf{Z}_i\gamma) \right\}. \end{aligned} \tag{A.6}$$

With the expression presented in (A.6), we will be able to estimate the vector of parameters γ. Taking the derivative of this log-likelihood function with respect to parameter γ, we have

$$\mathbf{U}_\gamma = \sum_{i=1}^{m} \frac{1}{\tau_i}\mathbf{Z}_i^\top \mathbf{D}_i^{-1}(\mathbf{r}_i - \mathbf{Z}_i\gamma). \tag{A.7}$$

MLE of λ

Now, we need to rewrite the expression (A.4) in order to estimate the vector of parameters λ. We consider the submodel given by $\ln \sigma_j^2 = \mathbf{w}_j^\top \lambda$ and the fact that $(\mathbf{r}_i - \hat{\mathbf{r}}_i)^\top \mathbf{D}_i^{-1}(\mathbf{r}_i - \hat{\mathbf{r}}_i)$ can be rewritten as $\sum_{j=1}^{n_i}\left\{(r_{ij} - \hat{r}_{ij})\sigma_j^{-2}(r_{ij} - \hat{r}_{ij})\right\}$. Thus, we can rewrite the log-likelihood function as follows

$$l(\theta) \propto -\frac{1}{2}\sum_{i=1}^{m} n_i \ln \tau_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{w}_j^\top \lambda \tag{A.8}$$

$$- \frac{1}{2}\sum_{i=1}^{m}\frac{1}{\tau_i}\left\{ \sum_{j=1}^{n_i}\left\{\sigma_j^{-2}(r_{ij} - \hat{r}_{ij})^2\right\} \right\}. \tag{A.9}$$

With the expression presented in (A.9), we will be able to estimate the vector of parameters λ. Taking the derivative of this log-likelihood function with respect to parameter λ, we have

$$\mathbf{U}_\lambda = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{w}_j + \frac{1}{2}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{j=1}^{n_i} \sigma_j^{-2}(r_{ij} - \hat{r}_{ij})^2 \mathbf{w}_j. \tag{A.10}$$

References

[1] J. J. Heckman, B. Singer, Econometric analysis of longitudinal data, Handbook of econometrics 3 (1986) 1689–1763.

[2] N. R. Chaganty, An alternative approach to the analysis of longitudinal data via generalized estimating equations, Journal of Statistical Planning and Inference 63 (1) (1997) 39–54.

[3] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger, et al., Analysis of longitudinal data, Oxford university press, 2002.

[4] E. Demidenko, Mixed models: theory and applications with R, John Wiley & Sons, 2013.

[5] M. Pourahmadi, Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, Biometrika 86 (3) (1999) 677–690.

[6] M. Pourahmadi, Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix, Biometrika 87 (2) (2000) 425–435.

[7] W. Zhang, C. Leng, A moving average cholesky factor model in covariance modelling for longitudinal data, Biometrika 99 (1) (2012) 141–150.

[8] J. Pan, Y. Pan, jmcm: An R package for joint mean-covariance modeling of longitudinal data, Journal of Statistical Software 82 (2017) 1–29.

[9] Z. Chen, D. B. Dunson, Random effects selection in linear mixed models, Biometrics 59 (4) (2003) 762–769.

[10] W. Zhang, C. Leng, C. Y. Tang, A joint modelling approach for longitudinal studies, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77 (1) (2015) 219–238.

[11] T.-I. Lin, Y.-J. Wang, A robust approach to joint modeling of mean and scale covariance for longitudinal data, Journal of Statistical Planning and Inference 139 (9) (2009) 3013–3026.

[12] A. Rhee, M.-S. Kwak, K. Lee, Robust modeling of multivariate longitudinal data using modified cholesky and hypersphere decompositions, Computational Statistics & Data Analysis 170 (2022) 107439.

[13] Y. Guney, O. Arslan, F. G. Yavuz, Robust estimation in multivariate heteroscedastic regression models with autoregressive covariance structures using em algorithm, Journal of Multivariate Analysis (2022) 105026.

[14] D. F. Andrews, C. L. Mallows, Scale mixtures of normal distributions, Journal of the Royal Statistical Society: Series B (Methodological) 36 (1) (1974) 99–102.

[15] J. Rosca, D. Erdogmus, J. C. Príncipe, S. Haykin, Independent component analysis and blind signal separation: 6th International Conference, ICA 2006, Charleston, SC, USA, March 5-8, 2006, Proceedings, Vol. 3889, Springer, 2006.

[16] F. Pascal, J.-P. Ovarlez, P. Forster, P. Larzabal, On a SIRV-CFAR detector with radar experimentations in impulsive clutter, in: 2006 14th European Signal Processing Conference, IEEE, 2006, pp. 1–5.

[17] Y. Chitour, F. Pascal, Exact maximum likelihood estimates for SIRV covariance matrix: Existence and algorithm analysis, IEEE Transactions on signal processing 56 (10) (2008) 4563–4573.

[18] P. Formont, F. Pascal, G. Vasile, J.-P. Ovarlez, L. Ferro-Famil, Statistical classification for heterogeneous polarimetric SAR images, IEEE Journal of selected topics in Signal Processing 5 (3) (2010) 567–576.

[19] H. Ye, J. Pan, Modelling of covariance structures in generalised estimating equations for longitudinal data, Biometrika 93 (4) (2006) 927–941. arXiv:https://academic.oup.com/biomet/article-pdf/93/4/927/593167/934927.pdf, doi:10.1093/biomet/93.4.927.
URL https://doi.org/10.1093/biomet/93.4.927

[20] S. Cambanis, S. Huang, G. Simons, On the theory of elliptically contoured distributions, Journal of Multivariate Analysis 11 (3) (1981) 368–385.

[21] E. Gómez, M. A. Gómez-Villegas, J. M. Marín, A survey on continuous elliptical vector distributions, Revista matemática complutense 16 (1) (2003) 345–361.

[22] G. Schwarz, Estimating the dimension of a model, The annals of statistics (1978) 461–464.

[23] H. Akaike, A new look at the statistical model identification, IEEE transactions on automatic control 19 (6) (1974) 716–723.

[24] L. Bombrun, S. N. Anfinsen, O. Harant, A complete coverage of log-cumulant space in terms of distributions for polarimetric SAR data, in: POLinSAR 2011-5th International Workshop on Science and Applications of SAR Polarimetry and Polarimetric Interferometry, 2011, pp. 1–8.

[25] H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, The annals of mathematical statistics (1947) 50–60.

[26] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the royal statistical society: Series B (Methodological) 36 (2) (1974) 111–133.

[27] S. Geisser, The predictive sample reuse method with applications, Journal of the American statistical Association 70 (350) (1975) 320–328.

[28] R. F. Potthoff, S. Roy, A generalized multivariate analysis of variance model useful especially for growth curve problems, Biometrika 51 (3-4) (1964) 313–326.

[29] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, R. C. Team, Linear and nonlinear mixed effects models, R package version 3 (57) (2007) 1–89.

[30] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2020).
URL `https://www.R-project.org/`

[31] M. Maadooliat, M. Pourahmadi, J. Z. Huang, Robust estimation of the correlation matrix of longitudinal data, Statistics and Computing 23 (2013) 17–28.

B PAPER 2: JOINT MODELING APPROACH WITH ALTERNATIVE CHOLESKY DECOMPOSITION USING THE FAMILY OF SCALE MIXTURE OF NORMAL DISTRIBUTION

Submitted paper:

Ribeiro, V. S. O., Bombrun, L., Nobre, J. S., Cavalcante, C. C., & Berthoumieu, Y. (2025).

Joint modeling approach with Alternative Cholesky Decomposition using the family of scale mixture of normal distribution.

Under review in Signal Processing.

Alternative Cholesky Decomposition and family of scale mixture of Normal distribution: a joint modeling approach

Vinícius Silva Osterne Ribeiro[a], Lionel Bombrun[b,c], Juvêncio Santos Nobre[d], Charles Casimiro Cavalcante[a], Yannick Berthoumieu[b]

[a]*Department of Teleinformatics Engineering, Federal University of Ceará, Campus do Pici, Block 725, Fortaleza, 60455-970, Ceará, Brazil*
[b]*Univ. Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218, Talence, F-33400, France*
[c]*Bordeaux Sciences Agro, Gradignan, F-33175, France*
[d]*Department of Statistics and Applied Mathematics, Federal University of Ceará, Campus do Pici, Block 910, Fortaleza, 60455-970, Ceará, Brazil*

Abstract

In Statistics, the analysis of longitudinal data is essential across various domains, including biomedical and agricultural research. Joint mean-covariance models have been widely used to capture within-subject dependence, often by parametrizing the scatter matrix via the Modified Cholesky Decomposition (MCD). However, the MCD has known drawbacks, such as sensitivity to the ordering of variables and challenges in parameter interpretation. As an alternative, the Alternative Cholesky Decomposition (ACD) offers improved numerical stability and interpretability, yet has been underexplored in robust modeling contexts. Traditional approaches also frequently assume normally distributed residuals, which may not hold in practice. While extensions based on the Student-t and Laplace distributions address heavier tails, they still rely on fixed parametric forms. To overcome both structural and distributional limitations, this paper proposes a novel joint regression model that combines the flexibility of ACD with the robustness of scale mixture of normal (SMN) distributions. We obtain maximum likelihood estimators and compare our model against classical and Student-t-based alternatives. Simulation studies show superior performance in estimation and prediction under outlier contamination. Real data applications further highlight the model's robustness and practical utility.

Keywords: Repeated measures, Scale Mixture of Normal distribution,

Cholesky Decomposition, Robust estimation.
PACS: 0000, 1111
2000 MSC: 0000, 1111

1. Introduction

Studies with repeated measures (or longitudinal data) consist of observing the same subjects at successive time points, and are widely used in econometrics, biomedicine, agriculture, and other scientific fields [1, 2, 3, 4, 5]. Unlike cross-sectional studies, repeated measures require accounting for within-subject dependence, as multiple observations are obtained for each individual.

To properly account for this dependence, joint mean-covariance models have been developed, extending classical linear regression by modeling both the mean vector and the covariance (or scatter) matrix of the residuals. One of the most commonly used approaches for parameterizing the covariance structure is the Modified Cholesky Decomposition (MCD), as introduced in [6]. While MCD enables a convenient regression-based parametrization of the covariance matrix, it suffers from critical drawbacks: it is highly sensitive to the ordering of the variables, and the regression coefficients it produces lack direct interpretation in terms of correlation or dependence strength. These limitations have been discussed in detail in [7, 8], where the ordering-dependence is shown to affect both estimation accuracy and interpretability, especially in high-dimensional or unbalanced designs.

To address these issues, Chen *et al.* introduced in [7] the Alternative Cholesky Decomposition (ACD), which reparametrizes the covariance structure in a way that is order-invariant and allows for a clearer interpretation of dependence patterns. ACD has been shown to produce more stable parameter estimates and enables modeling strategies that better reflect the underlying correlation structure, particularly in the presence of irregular time grids or missing data [9]. These structural advantages have motivated its use in several modern applications, yet robust joint models using ACD remain scarce in the literature.

Regarding robustness, several works have relaxed the Gaussian assumption by adopting alternative error distributions. For instance, the Student-t and Laplace-based joint models [10, 11] were proposed to improve robustness, but they still rely on fixed parametric forms, which may be too rigid in practice. A more flexible and general approach was introduced by Ribeiro et

al. [12], who modeled the residuals using the Scale Mixture of Normal (SMN) family, thus encompassing a broader class of distributions. However, their model was built entirely upon the MCD, inheriting its structural limitations.

To overcome both distributional and structural constraints, this paper proposes a new robust joint model that combines the flexibility of the SMN distribution with the structural advantages of the ACD decomposition. By integrating these two components, the proposed model provides a more interpretable, robust, and flexible framework for modeling longitudinal data.

- We obtain the maximum likelihood estimation framework and propose an iterative algorithm to estimate the parameters of the joint model using the ACD structure and residuals from the SMN family;
- We provide analytical expressions for the mean squared prediction error in the context of forecasting future responses;
- We perform simulation studies and real-data applications to illustrate the robustness and flexibility of the proposed approach, especially in comparison with existing models based on MCD and fixed-distribution assumptions.

The paper is organized as follows. Details about the structure of the regression model and different correlation structures are introduced in Section 2 and the presentation of our proposal is presented in Section 3. In Section 4, we present the structure, estimation and prediction method of the proposed model. In Sections 5 and 6, simulation and application studies to a set of real data are presented to evaluate the sensitivity and robustness of the proposed model. In Section 7, we present some final considerations and future work on the topic.

2. Joint modeling of repeated measures

Let's consider a sample of m subjects. In the context of repeated measurements, a continuous random variable $\mathbf{y}$ is observed over time for each subject. Denote by $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^\top$ the response profile of the ith subject ($i = 1, 2, \ldots, m$), where n_i is the number of observations for subject i, which may be irregularly spaced.

To model the mean of $\mathbf{y}_i$, we adopt the standard linear structure widely used in the literature:

$$\mathbf{y}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, with $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i})^\top$ being the $n_i \times (p+1)$ design matrix for the ith subject, containing the values of the explanatory variables at each time point. $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^\top$ is the vector of regression coefficients, and $\boldsymbol{\epsilon}_i$ is the residual vector.

While the classical assumption is that $\boldsymbol{\epsilon}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_i)$, in longitudinal modeling, the structure of the scatter matrix $\boldsymbol{\Sigma}_i$ plays a crucial role in accounting for within-subject dependence. To make inference feasible and the model identifiable, a parametrization of $\boldsymbol{\Sigma}_i$ is required.

Historically, the *Modified Cholesky Decomposition* (MCD), introduced by Pourahmadi [6, 13], has been the most widely used strategy. This approach reparametrizes the scatter matrix as:

$$\mathbf{L}_i\boldsymbol{\Sigma}_i\mathbf{L}_i^\top = \mathbf{D}_i, \tag{2}$$

where $\mathbf{L}_i$ is a unit lower triangular matrix with autoregressive parameters $-\phi_{jk}$ in the subdiagonal elements, and $\mathbf{D}_i = \text{diag}(\sigma_1^2, \ldots, \sigma_{n_i}^2)$ contains the innovation variances. However, the MCD suffers from well-documented limitations: it is highly sensitive to the ordering of variables, and the parameters lack direct interpretability in terms of marginal correlations [7, 8]. These drawbacks may lead to unstable estimates, particularly in irregular or unbalanced designs, which are common in longitudinal settings.

To address these issues, Chen et al. [7] introduced the *Alternative Cholesky Decomposition* (ACD), which reverses the reparametrization strategy by modeling the covariance matrix directly as:

$$\boldsymbol{\Sigma}_i = \mathbf{D}_i\mathbf{L}_i\mathbf{L}_i^\top\mathbf{D}_i, \tag{3}$$

where

$$\mathbf{L}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \phi_{21} & 1 & 0 & \cdots & 0 \\ \phi_{31} & \phi_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{n_i1} & \phi_{n_i2} & \phi_{n_i3} & \cdots & 1 \end{bmatrix}. \tag{4}$$

In this decomposition, the ϕ_{jk} parameters are interpreted as *moving average coefficients*, and the residuals can be expressed as:

$$(y_{ij} - \mu_{ij})/\sigma_i = \epsilon_{ij} + \sum_{k=1}^{j-1} \phi_{jk}\epsilon_{ij}, \tag{5}$$

with $\mathbf{D}_i = \text{diag}(\sigma_1^2, \ldots, \sigma_{n_i}^2)$ representing the innovation variances. Unlike MCD, the ACD provides an *order-invariant parametrization* and yields parameters that are more directly interpretable and numerically stable [9].

Following [6], both MCD and ACD decompositions allow for structured modeling of the scatter matrix through:

$$\phi_{jk} = \mathbf{z}_{jk}^\top \boldsymbol{\gamma}, \qquad \ln \sigma_j^2 = \mathbf{w}_j^\top \boldsymbol{\lambda}, \tag{6}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are global parameter vectors shared across subjects, and $\mathbf{z}_{jk}$, $\mathbf{w}_j$ are covariate vectors, often specified as time-based polynomial expansions [14]:

$$\mathbf{z}_{jk} = \left[1, (t_{ik} - t_{ij}), (t_{ik} - t_{ij})^2, \ldots, (t_{ik} - t_{ij})^{d-1}\right]^\top, \tag{7}$$

$$\mathbf{w}_j = \left[1, t_{ij}, t_{ij}^2, \ldots, t_{ij}^{q-1}\right]^\top. \tag{8}$$

These flexible parameterizations allow for time-varying and subject-specific variance-correlation patterns. However, despite the growing recognition of the theoretical and computational advantages of ACD, most robust joint models in the literature are still formulated using the MCD structure. In this work, we adopt the ACD as the foundation for our modeling approach, enabling us to incorporate scale mixture of normal (SMN) residuals within a robust and interpretable joint modeling framework.

3. SMN distribution and proposed method

When considering that $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, Pourahmadi introduces in [6] the NJMM model, which uses the normal distribution for the residual term and the MCD decomposition for the scatter matrix. Whereas this assumption for error terms may be questionable in situations where there are outliers or the underlying data exhibit thick tails, Lin et al. [10] and Guney et al. [11] present the TJMM and LJMM models, respectively, for modeling longitudinal data with Student-t and Laplace distribution for residuals. More recently, Ribeiro et al. [12] have proposed to consider models from the wider family of SMN distribution. This family of distributions admits the following stochastic representation:

$$\mathbf{y} = \boldsymbol{\mu} + \tau^{1/2}\mathbf{z}, \tag{9}$$

where $\boldsymbol{\mu}$ is the mean vector. τ is a positive scalar random variable and $\mathbf{z}$ is a zero-mean normal random vector, *i.e.* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. In addition, τ

and $\mathbf{z}$ are independent. By considering various distributions for the random variable τ, different distributions are obtained for $\mathbf{y}$ such as the normal, Student-t and Laplace distributions. In [12], inspired by results on robust estimation theory [15], Ribeiro et al. have proposed to consider, for the MCD decomposition, that τ are unknown but deterministic parameters. By do not imposing any fixed parametric form for τ, this robust model adapts to the observations and generalizes the state of the art NJMM [6], LJMM [11] and TJMM [10] models.

In the context of ACD decomposition, recent works also consider modeling from the normal distribution, as presented in [16] and from the Student-t distribution, as presented in [17]. Therefore, following the extension used by [12], our proposal in this paper is to consider that the residuals belong to the SMN class of distributions, but now considering the ACD decomposition, instead of the MCD.

4. ACD decompositon and joint model for scale mixtures of normal

4.1. Statistical model

In this paper, we follow a similar approach as the one developed in [18, 19]. We consider that τ_i are unknown but deterministic parameters. Assuming that $\boldsymbol{\epsilon}_i|\tau_i \sim \mathcal{N}(\mathbf{0}, \tau_i\boldsymbol{\Sigma}_i)$, the likelihood function to maximize with respect to $\boldsymbol{\mu}^\top, \boldsymbol{\tau}, \boldsymbol{\Sigma}$, for a sample of m individuals $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^\top$, for $i = 1, 2, ..., m$, with $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\tau}, \boldsymbol{\Sigma})^\top$ is given by:

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_m|\boldsymbol{\theta}) = \prod_{i=1}^{m} \frac{1}{(2\pi)^{\frac{n_i}{2}} \tau_i^{\frac{n_i}{2}} |\boldsymbol{\Sigma_i}|^{\frac{1}{2}}} \exp\left(-\frac{(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma_i}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)}{2\tau_i}\right) \tag{10}$$

It yields that the corresponding log-likelihood function can be expressed by:

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{1}{2}\ln(2\pi)\sum_{i=1}^{m} n_i - \frac{1}{2}\sum_{i=1}^{m} n_i \ln\tau_i - \frac{1}{2}\sum_{i=1}^{m} \ln|\boldsymbol{\Sigma}_i| \\ &- \sum_{i=1}^{m} \frac{1}{2\tau_i}(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma_i}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i). \end{aligned} \tag{11}$$

4.2. Maximum likelihood estimation

In this section, we introduce an iterative algorithm to estimate by maximum likelihood the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\tau}, \boldsymbol{\Sigma})^\top$ of the proposed joint model. For that, the score functions are first derived and then the Fisher scoring algorithm is introduced as detailed in the next two subsections.

4.2.1. Score functions and Fisher information matrix

By taking the first derivative of (11), the score functions for each parameter are given by (see Appendix A for more details)

$$\mathbf{U}_{\boldsymbol{\beta}} = \sum_{i=1}^{m} \frac{1}{\tau_i} \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \tag{12}$$

$$\mathbf{U}_{\gamma_r} = \sum_{i=1}^{m} \frac{1}{\tau_i} \left(\operatorname{tr} \left((\mathbf{D}_i \mathbf{L}_i)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top (\mathbf{D}_i \mathbf{L}_i)^{-\top} \mathbf{L}_i^{-1} \frac{\partial \mathbf{L}_i}{\partial \gamma_r} \right) \right) \tag{13}$$

$$\mathbf{U}_{\lambda_s} = \sum_{i=1}^{m} \operatorname{tr} \left(\left(\frac{1}{\tau_i} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} - \mathbf{I}_i \right) \mathbf{D}_i^{-1} \frac{\partial \mathbf{D}_i}{\partial \lambda_s} \right) \tag{14}$$

$$\mathbf{U}_{\tau_i} = -\frac{n_i}{2\tau_i} + \frac{1}{2\tau_i^2} (\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i, \tag{15}$$

where $\dfrac{\partial \mathbf{L}_i}{\partial \gamma_r}$ is the derivative of matrix $\mathbf{L}_i$ with respect to r^{st} element of vector $\boldsymbol{\gamma}$ and $\dfrac{\partial \mathbf{D}_i}{\partial \lambda_s}$ is the derivative of matrix $\mathbf{D}_i$ with respect to s^{st} element of vector $\boldsymbol{\lambda}$. Equalling the score function (12) to zero, we obtain the maximum likelihood estimator (MLE) of the regression coefficients as:

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i} \mathbf{X}_i^\top \hat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{X}_i \right]^{-1} \left[\sum_{i=1}^{m} \frac{1}{\hat{\tau}_i} \mathbf{X}_i^\top \hat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{y}_i \right], \tag{16}$$

where $\hat{\tau}_i$ is the MLE of τ_i, i.e the result when equating the equation (15) to zero

$$\hat{\tau}_i = \frac{(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^\top \hat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)}{n_i}, \tag{17}$$

and $\hat{\boldsymbol{\Sigma}}_i$ is the MLE of $\boldsymbol{\Sigma}_i$, whose components will be estimated from the Fisher information matrix, given by

$$\mathbf{I}_{\gamma_r\gamma_s} = \sum_{i=1}^{m} \frac{1}{\tau_i} \mathrm{tr}\left(\frac{\partial \mathbf{L}_i}{\partial \gamma_r}\frac{\partial \mathbf{L}_i}{\partial \gamma_s}\mathbf{L}_i^{-\top}\mathbf{L}_i^{-1}\right) \tag{18}$$

$$\mathbf{I}_{\lambda_r\lambda_s} = \sum_{i=1}^{m} \frac{1}{\tau_i} \mathrm{tr}\left(\mathbf{L}_i\mathbf{L}_i^{\top}\frac{\partial \mathbf{D}_i}{\partial \lambda_r}\boldsymbol{\Sigma}_i^{-1}\frac{\partial \mathbf{D}_i}{\partial \lambda_s} + \mathbf{D}_i^{-2}\frac{\partial \mathbf{D}_i}{\partial \lambda_r}\frac{\partial \mathbf{D}_i}{\partial \lambda_s}\right) \tag{19}$$

$$\mathbf{I}_{\gamma_r\lambda_s} = \mathbf{I}_{\lambda_s\gamma_r} = \sum_{i=1}^{m} \frac{1}{\tau_i} \mathrm{tr}\left(\mathbf{D}_i\frac{\partial \mathbf{L}_i}{\partial \gamma_r}\mathbf{L}_i^{\top}\frac{\partial \mathbf{D}_i}{\partial \lambda_s}\boldsymbol{\Sigma}_i^{-1}\right). \tag{20}$$

4.2.2. Estimation algorithm

From the closed expressions of $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}_i$, we were able to obtain the MLE in a simple way for these two parameters as given in (12) and (17). Unfortunately, to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, there is no close form expression. We hence suggest to use the Fisher scoring algorithm by using:

$$\hat{\boldsymbol{\eta}}^{(h+1)} = \hat{\boldsymbol{\eta}}^{(h)} + \left(\mathbf{I}_{\hat{\boldsymbol{\eta}}^{(h)}}\right)^{-1}\mathbf{U}_{\hat{\boldsymbol{\eta}}^{(h)}}, \tag{21}$$

where $\hat{\boldsymbol{\eta}}^{(h)} = \left(\hat{\boldsymbol{\gamma}}^{(h)}, \hat{\boldsymbol{\lambda}}^{(h)}\right)^{\top} = \left(\hat{\gamma}_1^{(h)}, \ldots, \hat{\gamma}_d^{(h)}, \hat{\lambda}_1^{(h)}, \ldots, \hat{\lambda}_q^{(h)}\right)^{\top}$ are the estimates at step h and

$$\mathbf{U}_{\hat{\boldsymbol{\eta}}^{(h)}} = \begin{pmatrix}\mathbf{U}_{\hat{\boldsymbol{\gamma}}^{(h)}} \\ \mathbf{U}_{\hat{\boldsymbol{\lambda}}^{(h)}}\end{pmatrix} \quad \text{and} \quad \mathbf{I}_{\hat{\boldsymbol{\eta}}^{(h)}} = \begin{pmatrix}\mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\gamma}}^{(h)}} & \mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} \\ \mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} & \mathbf{I}_{\hat{\boldsymbol{\lambda}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}}\end{pmatrix}. \tag{22}$$

In addition, the score vectors are respectively defined as:

$$\mathbf{U}_{\hat{\boldsymbol{\gamma}}^{(h)}} = \left(\mathbf{U}_{\hat{\gamma}_1^{(h)}}, \ldots, \mathbf{U}_{\hat{\gamma}_d^{(h)}}\right)^{\top} \quad \text{and} \quad \mathbf{U}_{\hat{\boldsymbol{\lambda}}^{(h)}} = \left(\mathbf{U}_{\hat{\lambda}_1^{(h)}}, \ldots, \mathbf{U}_{\hat{\lambda}_q^{(h)}}\right)^{\top}, \tag{23}$$

and the elements of the Fisher Information matrix are given by:

$$\mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\gamma}}^{(h)}} = \begin{bmatrix} \mathbf{I}_{\hat{\gamma}_1^{(h)}\hat{\gamma}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\gamma}_1^{(h)}\hat{\gamma}_d^{(h)}} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{\hat{\gamma}_d^{(h)}\hat{\gamma}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\gamma}_d^{(h)}\hat{\gamma}_d^{(h)}} \end{bmatrix}, \tag{24}$$

$$\mathbf{I}_{\hat{\boldsymbol{\gamma}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} = \begin{bmatrix} \mathbf{I}_{\hat{\gamma}_1^{(h)}\hat{\lambda}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\gamma}_1^{(h)}\hat{\lambda}_q^{(h)}} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{\hat{\gamma}_d^{(h)}\hat{\lambda}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\gamma}_d^{(h)}\hat{\lambda}_q^{(h)}} \end{bmatrix}, \tag{25}$$

$$\mathbf{I}_{\hat{\boldsymbol{\lambda}}^{(h)}\hat{\boldsymbol{\lambda}}^{(h)}} = \begin{bmatrix} \mathbf{I}_{\hat{\lambda}_1^{(h)}\hat{\lambda}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\lambda}_1^{(h)}\hat{\lambda}_q^{(h)}} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{\hat{\lambda}_q^{(h)}\hat{\lambda}_1^{(h)}} & \cdots & \mathbf{I}_{\hat{\lambda}_q^{(h)}\hat{\lambda}_q^{(h)}} \end{bmatrix}. \tag{26}$$

The steps for estimating the parameters of the proposed model with deterministic but unknown τ_i are summarized in Algorithm 1.

Algorithm 1 Algorithm for the maximum likelihood estimate of the proposed SMNJMM model

1: Given a starting value $\boldsymbol{\theta}^{(h)} = (\boldsymbol{\beta}^{(h)}, \boldsymbol{\tau}^{(h)}, \boldsymbol{\gamma}^{(h)}, \boldsymbol{\lambda}^{(h)})^\top$, which can be obtained by fitting the NJMM model, to form $\mathbf{L}_i^{(h)}$, $\mathbf{D}_i^{(h)}$, and $\boldsymbol{\Sigma}_i^{(h)}$;
2: Update $\hat{\boldsymbol{\tau}}_i^{(h+1)}$ using (17)
3: Update $\hat{\boldsymbol{\beta}}^{(h+1)}$ using (16)
4: Update $\hat{\boldsymbol{\gamma}}^{(h+1)}$ and $\hat{\boldsymbol{\lambda}}^{(h+1)}$ using Fisher Scoring, presented in (21)
5: Update $\hat{\boldsymbol{\Sigma}}_i^{(h+1)}$ using (3)
6: Normalization of the scatter matrix, considering $\boldsymbol{\Sigma}_i^{(h+1)} = \frac{n_i \boldsymbol{\Sigma}_i^{(h+1)}}{tr(\boldsymbol{\Sigma}_i^{(h+1)})}$;
7: Repeat steps 2 to 6 until convergence.

As expected and as observed in Algorithm 1, the proposed SMNJMM model behaves as the NJMM when τ_i are the same for each subject i. In that case, the stochastic representation given in (9) vanishes to the one for a normally distributed random variable, and equations (16) and (21) reduces to the one obtained for the NJMM model [16]. In addition, an interesting property of the proposed SMNJMM model is that it is robust to outliers. Indeed, let's consider that the i^{st} observation $\mathbf{y}_i$ is an outlier. It will have a large deviance from $\boldsymbol{\mu}_i$ according to the Mahalanobis distance $(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma_i}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)$.

Hence, the estimated $\hat{\tau}_i$ for this observation will be large according to (17). And since $\hat{\tau}_i$ is on the denominator of (16), this outlier will have a lower weight during the estimation of $\hat{\boldsymbol{\beta}}$. Now that we have shown that an outlier has a lower influence for the estimation of $\hat{\boldsymbol{\beta}}$, the same argument holds for the estimation of the scatter matrix $\boldsymbol{\Sigma}_i$ since τ_i is on the denominator of the two score functions used to estimate γ_r and λ_s in (13) and (14) and hence to estimate $\boldsymbol{\Sigma}_i$ using (3).

The SMNJMM model proposed in this paper shares the same advantages as the one introduced by Ribeiro et al. in [12]. The only difference relies on the definition of the structure of the scatter matrix. Here, the ACD decomposition is employed while the MCD decomposed was used in [12]. To better understand the added-value of the proposed model, Section 6 will present a detail comparison between both on real dataset. Before showing these results, the next subsection gives some elements about forward prediction with the proposed SNMJMM model.

4.3. Prediction values

By following the same methodology as done by Ribeiro et al. in [12] for the SMNJMM model with the MCD decomposition, we introduce a technique for predicting the new g observations of subject i with the proposed model but now using the ACD decomposition for the scatter matrix.

We denote $\tilde{\mathbf{y}}_i$ the future vector of measurements of the individual i, which has dimension $(g \times 1)$. $\tilde{\mathbf{X}}_i$ is the $g \times (p+1)$ design matrix for the prediction part. With these notations, we have $\mathbb{E}(\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i) = \tilde{\mathbf{X}}_i\boldsymbol{\beta}$. In addition, $\tilde{\mathbf{z}}_{jk}$ and $\tilde{\mathbf{w}}_j$ are the covariates vectors of dimension $d \times 1$ and $q \times 1$, where $\ln \sigma_i^2 = \tilde{\mathbf{w}}_i^\top \boldsymbol{\lambda}$ and $\phi_{jk} = \tilde{\mathbf{z}}_{jk}^\top \boldsymbol{\gamma}$, for $j \in \{n_i+1, ..., n_i+g\}$ and $k \in \{1, ..., j-1\}$, respectively. For the prediction, we assume that the joint distribution of $\mathbf{y}_i$ and $\tilde{\mathbf{y}}_i$ given τ_i follows a multivariate normal distribution. It yields:

$$\begin{bmatrix} \mathbf{y}_i \\ \tilde{\mathbf{y}}_i \end{bmatrix} |\tau_i \sim \mathcal{N}_{n_i+g}(\mathbf{X}_i^*\boldsymbol{\beta}, \tau_i\boldsymbol{\Sigma}_i^*), \tag{27}$$

where $\mathbf{X}_i^* = (\mathbf{X}_i^\top, \tilde{\mathbf{X}}_i^\top)^\top$, $\boldsymbol{\Sigma}_i^{*-1} = \mathbf{D}_i^{*-1}\mathbf{L}_i^{*-\top}\mathbf{L}_i^{*-1}\mathbf{D}_i^{*-1}$, $\mathbf{D}_i^* = \text{diag}(\sigma_1^2, ..., \sigma_{n_i+g}^2)$ and $\mathbf{L}_i^* = [l_{jk}]$ is an $(n_i+g) \times (n_i+g)$ lower triangular matrix as defined in (4). The scatter matrix $\boldsymbol{\Sigma}_i^*$ can be decomposed as:

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \tag{28}$$

where $\boldsymbol{\Sigma}_{11}$ is the upper left $(n_i \times n_i)$ submatrix of $\boldsymbol{\Sigma}_i^*$ corresponding to the scatter matrix of the n_i observed data. $\boldsymbol{\Sigma}_{22}$ is the $g \times g$ scatter matrix for the future observations, and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$. Note that to simplify notations, the subscript i on partitioned matrices is removed but all these matrix depend on subject i.

By using properties of marginal and conditional distributions shown in [20, 21] for elliptical distributions which extends scale mixture of normal distributions, we know that if $\mathbf{y}_i|\tau_i \sim \mathcal{N}_{n_i}(\boldsymbol{\mu}_i, \tau_i\boldsymbol{\Sigma}_i)$, then

$$\tilde{\mathbf{y}}_i|\mathbf{y}_i, \tau_i \sim \mathcal{N}_g\left(\boldsymbol{\mu}_{2\cdot1}, \tau_i\boldsymbol{\Sigma}_{22\cdot1}\right), \tag{29}$$

where $\boldsymbol{\mu}_{2\cdot1} = \mathbb{E}(\tilde{\mathbf{y}}_i|\mathbf{y}_i, \tau_i)$ and $\boldsymbol{\Sigma}_{22\cdot1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. With (29), it yields that the Mean Squared Error (MSE) predictor of $\tilde{\mathbf{y}}_i$ is given by [11]

$$\hat{\tilde{\mathbf{y}}}_i = \boldsymbol{\mu}_{2\cdot1} = \tilde{\mathbf{X}}_i\boldsymbol{\beta} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{30}$$

and the MSE scatter matrix for the predictor (30) is determined by

$$\mathbb{E}[(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)^\top|\tau_i] = \mathbb{E}(\mathrm{cov}(\tilde{\mathbf{y}}_i|\mathbf{y}_i)|\tau_i) = \tau_i\boldsymbol{\Sigma}_{22\cdot1}. \tag{31}$$

Note that these expressions are the same as the one obtained in [12] for the SMNJMM model with the MCD decomposition. The only difference relies on the way the scatter matrix $\boldsymbol{\Sigma}_i^*$ is parameterized. For the MCD decomposition, $\boldsymbol{\Sigma}_i^*$ is decomposed using (2), while for the proposed ACD decomposition, (3) is used.

5. Simulation studies

In this section, we present the estimation and prediction performance for synthetic data with the aim of evaluating and validating the proposed model. The presentation of these results aims at complementing the extensive study presented by Ribeiro et al. [12], which uses the same model, but considers the MCD structure for the scatter matrix.

5.1. ML estimation performance result

The first experiment aims at evaluating the efficiency of the proposed SMNJMM model compared to the state of the art NJMM [16] and TJMM [17] models. For that, two scenarios are considered. In scenario 1, a synthetic dataset is generated by using the normal distribution, while in scenario

2 the Student-t distribution with 4 degrees of freedom is employed. For both scenarios, $m = 100$ balanced random samples (with $n_i = 12$ observations for each sample unit) are generated. The following true parameters are considered. The regression coefficients are $\beta = [5, 27, 0.16]^\top$, and the weights of covariates vectors are $\gamma = [1.12, -0.58, 0.09, -0.0043]^\top$ and $\lambda = [-0.4, -1.37, 0.18, -0.007]^\top$. Hence $d = 4$, $q = 4$ and $p = 1$. In addition, the design matrix $\mathbf{X}_i$ for modeling the mean, and the covariates vectors $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ for modeling the moving average parameters and innovation variances are respectively given by:

$$\mathbf{X}_i = \left[\begin{array}{cc} \mathbf{1} & \mathbf{k} \end{array} \right]^\top, \quad \mathbf{z}_{jk} = \left[\begin{array}{cccc} \mathbf{1} & (j-k) & (j-k)^2 & (j-k)^3 \end{array} \right]^\top \tag{32}$$

and

$$\mathbf{w}_j = \left[\begin{array}{cccc} 1 & j & j^2 & j^3 \end{array} \right]^\top, \tag{33}$$

where $\mathbf{k} = [0, 1, 2.5, 3.5, 4.5, 6, 7, 8, 10, 11.5, 13, 14]^\top$.

To quantify the modeling performance of the proposed SMNJMM model compared to NJMM and TJMM models, the BIC criterion is used as proposed by [22]. This model comparison metric evaluates both tuning performance and model complexity. Its expression is given by

$$\text{BIC} = -\frac{2}{m}\hat{l}_{max} + \frac{k}{m}\ln(m), \tag{34}$$

where $\hat{l}_{max}$ is the obtained maximum log-likelihood value for the considered model and k is the number of estimated parameters. For example, for the proposed SMNJMM model, $k = p + d + q + m + 1$ since β has $(p + 1) \times 1$ dimension, γ has $d \times 1$ dimension, λ has $q \times 1$ dimension and τ has $m \times 1$ dimension. A lower BIC value corresponds to a better model. To calculate the standard error of the estimates, a total of r = 500 replications is considered.

Tables 1 and 2 present the estimation performance results in terms of mean estimates and standard errors (SE) computed on the r replications. To ease the understanding, the best and second best results are respectively displayed in blue and red. Table 1 shows the results when data are generated from a normal distribution while Table 2 shows the results for a Student-t generation model.

As expected and as observed in Table 1 for the first scenario, when data are generated from a normal distribution, the best performance is obtained

with the NJMM model. The proposed SMNJMM is the second best fit. To identify whether there is a significant difference between these models in term of BIC values, the paired Wilcoxon hypothesis test is used. At the 5% significance level, no significant difference between the “theoretical" NJMM model and the proposed SMNJMM model are observed indicating that these two models perform similarly in the normal case. In addition, when compared with the TJMM model, a significant difference is obtained indicating that the TJMM is a worst fit compared to the NJMM and SMNJMM model for this first scenario.

Table 1: Estimation performance for the NJMM, TJMM and SMNJMM models - Scenario 1: data generated from the normal distribution.

Parameter	True value	NJMM	[16]	TJMM	[17]	SMNJMM	
		Estimate	SE	Estimate	SE	Estimate	SE
β_1	5,27	5,2703	0,0394	5,2652	0,0489	5,2689	0,039
β_2	0,16	0,1605	0,0038	0,1607	0,0047	0,1606	0,0038
γ_1	1,12	1,1157	0,0489	1,114	0,0514	1,1152	0,0547
γ_2	-0,58	-0,5761	0,0385	-0,5749	0,0426	-0,5758	0,0427
γ_3	0,09	0,0899	0,0084	0,0898	0,0091	0,0899	0,0092
γ_4	-0,0043	-0,0036	0,0012	-0,0036	0,0021	-0,0036	0,0012
λ_1	-0,4	-0,3921	0,2362	-0,4014	0,2383	-2,2606	0,2287
λ_2	-1,37	-1,3807	0,1509	-1,3773	0,1525	-1,3598	0,1446
λ_3	0,18	0,1828	0,027	0,1827	0,0297	0,1797	0,0261
λ_4	-0,007	-0,0066	0,0018	-0,0066	0,0021	-0,0065	0,0018
BIC		-15,51		-11,94		-14,35	

Following the same approach, but now generating synthetic data from the Student-t distribution, Table 2 shows that the model presenting the best results is the TJMM, followed by the proposed SMNJMM model. The pairwise Wilcoxon test shows no difference between the “theoretical" TJMM model and the proposed SMNJMM model, while significant difference are observed when compared with the NJMM model.

Table 2: Estimation performance for the NJMM, TJMM and SMNJMM models - Scenario 2: data generated from the Student-t Distribution.

Parameter	True value	NJMM [16]		TJMM [17]		SMNJMM (proposed)	
		Estimate	SE	Estimate	SE	Estimate	SE
β_1	5,27	5,2702	0,0393	5,2651	0,0388	5,2688	0,0389
β_2	0,16	0,1603	0,0036	0,1605	0,0035	0,1604	0,0036
γ_1	1,12	1,1154	0,0486	1,1137	0,0481	1,1149	0,0544
γ_2	-0,58	-0,5762	0,0384	-0,575	0,0375	-0,5759	0,0426
γ_3	0,09	0,0897	0,0082	0,0896	0,0079	0,0897	0,009
γ_4	-0,0043	-0,0039	0,0009	-0,0039	0,0008	-0,0039	0,0009
λ_1	-0,4	-0,3922	0,2361	-0,4015	0,2342	-2,2607	0,2286
λ_2	-1,37	-1,3809	0,1507	-1,3775	0,1483	-1,36	0,1444
λ_3	0,18	0,1825	0,0267	0,1824	0,0264	0,1794	0,0258
λ_4	-0,007	-0,0076	0,0008	-0,0076	0,0008	-0,0075	0,0008
BIC		-11,13		-12,85		-12,18	

These two experiments on synthetic data clearly illustrate the robust behaviour of the proposed SMNJMM model with ACD structure. By do not imposing a fixed parametric form, the proposed model is able to fit the data as long as the hypothesis of belonging to SMN family of distribution is valid. In practice, this hypothesis is not so restrictive since the SMN family of distributions includes a very broad kind of distributions. Note also that the proposed SMNJMM is quite efficient since it performs as well as the theoretical model (NJMM in scenario 1 and TJMM in scenario 2).

5.2. Evaluation of prediction values

In order to measure the predictive accuracy of the models, we carried out another simulation study using synthetic data to compare the forward prediction performance of the NJMM, TJMM and SMNJMM models. A quantitative evalutation is performed in term of mean square forward prediction error (MSFE), defined as

$$\text{MSFE} = \frac{1}{mg}\sum_{i=1}^{m}(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i)^\top(\hat{\tilde{\mathbf{y}}}_i - \tilde{\mathbf{y}}_i). \tag{35}$$

m is the total number of subjects. g is the number of forward steps. $\tilde{\mathbf{y}}_i$ and $\hat{\tilde{\mathbf{y}}}_i$ are respectively the true and predicted vector of the response variable.

For this comparison, we considered what the literature calls pseudo-cross-validation, which is an approach of cross-validation [23] and predictive use of

sample [24]. This approach is based on holding out the last g measured in the i-th subject, obtaining maximum likelihood estimates using the remaining sampling units as a sample and predicting the true g measurements $\mathbf{y}_i = (y_{i,n_i-g+1}, ..., y_{i,n_i})^\top$, denoted by $\hat{\hat{\mathbf{y}}}_i = (\hat{\hat{y}}_{i,n_i-g+1}, ..., \hat{\hat{y}}_{i,n_i})^\top$, is done using (30).

To be able to compare the prediction performance of the models, we consider $m = 100$ balanced random samples observed at $n_i = 12$ different times (we consider that the number of observations for each individual is the same), for the normal and Student-t distributions (with 4 distributions of degrees of freedom), and we estimated the parameters of the respective models through the pseudo-cross-validation mentioned above. The results are shown in Table 3.

Table 3: MSFE values for NJMM, TJMM and SMNJMM models of g step-ahead forecasts.

Generating data from	Model	g step-ahead forecasts					
		g = 1	g = 2	g = 3	g = 4	g = 5	g = 6
Scenario 1: Normal	NJMM [16]	0,0823	0,1042	0,1343	0,1442	0,1620	0,2049
	TJMM [17]	0,0830	0,1048	0,1389	0,1455	0,1655	0,2108
	SMNJMM (proposed)	0,0814	0,1036	0,1338	0,1443	0,1637	0,2086
Scenario 2: Student-t	NJMM [16]	0,0796	0,0887	0,1269	0,1278	0,1988	0,2213
	TJMM [17]	0,0789	0,0867	0,1268	0,1270	0,1971	0,2175
	SMNJMM (proposed)	0,0775	0,0866	0,1255	0,1275	0,1974	0,2178

Unsurprisingly, as displayed in Table 3, the MSFE score increases with the number of steps g in the prediction. This behavior is expected since the model is less and less confident to predict long term observations. Table 3 also shows that the proposed SMNJMM model exhibits better prediction performance up to the third step, that is, up to $g = 3$ for the two scenarios. The results after this prediction step ($g = 4, 5, 6$) are better for the theoretical models (NJMM for scenario 1 and TJMM for scenario 2).

5.3. Robustness to data generated from models outside the SMN family

To assess the robustness of the proposed SMNJMM model to data generated from distributions outside the SMN family, we consider a scenario in which data are drawn from the Multivariate Generalized Gaussian Distribution (MGGD) [25], an extension of the multivariate normal that introduces a shape parameter $b > 0$. When $b = 1$, the MGGD reduces to the normal distribution; values of $b < 1$ produce heavy-tailed distributions, while

$b > 1$ lead to lighter tails. This experiment enables us to evaluate the performance of the NJMM, TJMM, and SMNJMM models under varying degrees of deviation from normality.

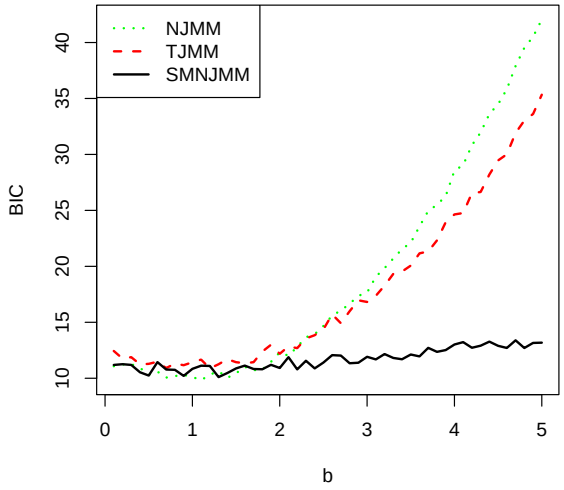


Figure 1: Evolution of the BIC criterion as a function of the MGGD shape parameter b for the NJMM, TJMM and SMNJMM models.

Figure 1 shows the evolution of the BIC criterion for the NJMM (in green), TJMM (in red), and SMNJMM (in black) models as a function of the shape parameter b of the MGGD. For this experiment, all other parameters: regression coefficients β, covariate vectors γ and λ, etc. are the same as those used in Section 5.1. As illustrated in Figure 1, when $b = 1$, the MGGD corresponds to the normal distribution, and in this case, the NJMM model performs best (i.e., yields the lowest BIC). The SMNJMM model, however, demonstrates greater robustness across the entire range of b, with notably superior performance in scenarios involving lighter tails ($b > 1$). This highlights its flexibility in handling various forms of non-normality, even when the data are generated from models outside the SMN family of distributions.

6. Real data analysis

Now that the proposed SMNJMM model has been successfully validated on synthetic data, some experiments on real data are carried out in this section. For that, a standard dataset is used. It consists of dental growth from the pituitary gland up to the left pterygomaxillary distance (mm) measured repeatedly at the ages of 8, 10, 12 and 14 years, with 11 women and 16 men, presented in [26]. Data can be downloaded from the `nlme` [27] package in R [28].

The objective of the first experiment is to confirm the potential of the proposed SMNJMM model compared to the NJMM and TJMM in terms of fitting accuracy. In addition, a comparison between the MCD and ACD decomposition structures is done. The regression model is defined as:

$$\mu_{ij} = \beta_0 + \delta_0 I_i(F) + (\beta_1 + \delta_1 I_i(F))t_j, \tag{36}$$

with $i \in \{1, ..., m = 27\}$ (27 subjects, 11 women and 16 men) and $j \in \{1, ..., n_i = 4\}$ (4 times steps at 8, 10, 12 and 14 years). μ_{ij} is the average orthodontic distance for the i-th subject at age t_j. β_0 and β_1 are the intercept and slope for men, respectively. δ_0 and δ_1 are the difference in intercept and slope between women and men, respectively. $I_i(F)$ is the indicator variable for women. In the following, the Poly(1,1) model is considered (i.e. $d = q = 1$ for the covariate vectors).

The maximum likelihood estimates corresponding to the six compared models (NJMM, TJMM and SMNJMM for both MCD and ACD decompositions) are presented in Table 4. As observed, both for the MCD and ACD decompositions, the proposed SMNJMM model provides a lower BIC value than the other NJMM and TJMM counterparts. This highlights the effectiveness of the proposed SMNJMM model for this practical application. In addition, for a same model (NJMM, TJMM or SMNJMM), the ACD decomposition always leads to better performance than the MCD one. These results clearly illustrate the interest of jointly using the SMNJMM model with the ACD decomposition.

A second experiment is performed on this dataset to evaluate the robustness behavior of the proposed SMNJMM model. The same evaluation protocol is employed as the one presented in [10] and [11]. First a contamination denoted by ζ is added for a single observation as:

$$y_{ij}(\zeta) = y_{ij} + \zeta. \tag{37}$$

Table 4: Maximum likelihood estimates for the Poly(1,1) structure model applied to orthodontic growth data.

Parameter	MCD decomposition			ACD decomposition		
	NJMM [29]	TJMM [10]	SMNJMM [12]	NJMM [16]	TJMM [17]	SMNJMM (proposed)
β_1	20,9738	15,2569	21,5423	18,4050	12,7204	18,8457
β_2	1,6117	0,7955	1,4254	1,3313	0,5151	0,9674
δ_1	-0,6947	0,9785	2,3120	-1,3303	0,3429	1,8540
δ_2	-0,6572	-0,3127	0,5214	-0,93f76	-0,5931	0,0634
γ_1	0,8498	0,9894	1,0685	0,2142	0,3538	0,4329
γ_2	-0,2537	-0,3781	-0,3781	-0,5341	-0,6585	-0,6585
λ_1	1,9268	1,5298	1,9951	1,2912	0,8942	1,3595
λ_2	-0,2846	-0,3714	-0,3124	-0,5650	-0,6518	-0,5928
BIC	9,59	12,52	8,52	8,44	12,04	8,03

Then the NJMM (in green), TJMM (in red) and SMNJMM (in black) models are estimated and their performance are evaluated in term of BIC values. Here, only the ACD decomposition for the scatter matrix is considered. Figure 2 draws the evolution of the BIC criterion as a function of the disturbance intensity (ζ) for two subjects: a woman on the left and a man on the right.

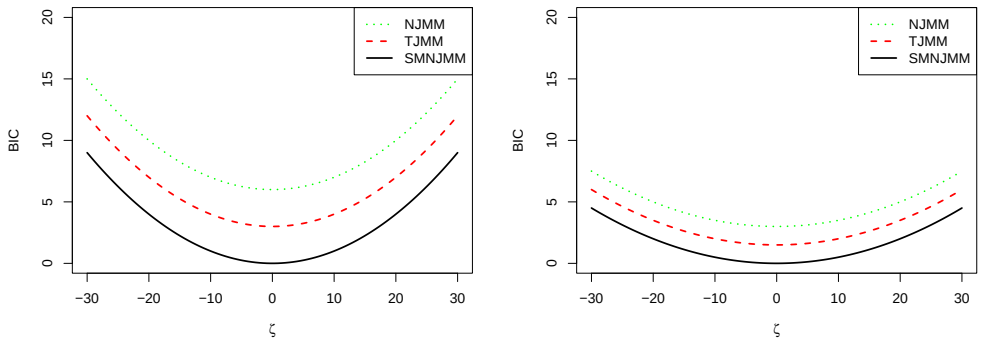


Figure 2: Evolution of the BIC metric as a function of perturbation ζ for the NJMM, TJMM and the proposed SMNJMM models (left: woman, right: man).

From Figure 2, we observe that for both subjects, the BIC values obtained with the proposed SMNJMM model are always smaller than the BIC values for the NJMM and TJMMM models. These results confirm that the SMNJMM model is less influenced by outliers as explained in Section 4.2.2. This robustness property of the proposed SMNJMM model allows it to satisfactorily accommodate extreme observations and makes it a relevant model

to use in practice.

7. Conclusions and future works

In this paper, we proposed a novel joint modeling approach, called SMNJMM, which assumes that residuals follow a scale mixture of normal (SMN) distribution and that the scatter matrix is decomposed using the Alternative Cholesky Decomposition (ACD). The SMNJMM framework allows the mixing parameters τ_i to be treated as deterministic but unknown, without imposing a fixed parametric distribution. This generalization enables the model to encompass a wide family of distributions—including the normal and Student-t as special cases—making it a robust extension of existing models such as NJMM and TJMM.

To perform inference, we derived a maximum likelihood estimation procedure based on the Fisher scoring algorithm. Simulation studies and applications to real data demonstrated the advantages of the SMNJMM model in terms of goodness of fit, forecasting accuracy, and robustness to outliers. In particular, we highlighted the benefits of using the ACD over the commonly adopted Modified Cholesky Decomposition (MCD): ACD provides order-invariant parametrization, improved numerical stability, and more interpretable parameter structures—especially relevant in irregular longitudinal designs. These results reinforce the practical value of adopting the ACD in robust modeling frameworks[1].

While both MCD and ACD offer parsimonious and unconstrained parameterizations for the scatter matrix, as discussed by [30], neither directly aligns the innovation variances with the marginal variances of repeated measures. This opens opportunities to explore alternative parametrizations, such as the Hyperspherical Cholesky Factor Parameterization [31], which may offer different interpretability or estimation advantages. Future work could therefore include extending the SMNJMM framework to incorporate such decompositions, exploring regularization techniques for high-dimensional data, or developing Bayesian implementations to incorporate prior knowledge and enhance interpretability. Additionally, diagnostic tools tailored for linear mixed models, as developed by [32], could be adapted and extended to this modeling framework, offering valuable insights into model fit and the identification of influential observations.

[1]R codes will be made available upon request from the authors.

Appendix A. Score functions and Fisher information matrix for the SMNJMM model with ACD decomposition

If $\boldsymbol{\epsilon}_i|\tau_i \sim \mathcal{N}(\mathbf{0}, \tau_i\boldsymbol{\Sigma}_i)$, then the log-likelihood function given by:

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{1}{2}\log(2\pi)\sum_{i=1}^{m} n_i - \frac{1}{2}\sum_{i=1}^{m} n_i \log\tau_i - \frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| \\ &- \frac{1}{2\tau_i}\sum_{i=1}^{m}(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i), \end{aligned}$$

where $\boldsymbol{\Sigma}_i = \mathbf{D}_i\mathbf{L}_i\mathbf{L}_i^\top\mathbf{D}_i$, then the score functions for each parameter is given by

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\beta} = \sum_{i=1}^{m}\frac{1}{\tau_i}\mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i).$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial\gamma_r} &= \frac{\partial}{\partial\gamma_r}\left(-\frac{1}{2\tau_i}\sum_{i=1}^{m}\mathrm{tr}\left((\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}\right)\right) \\ &= \frac{1}{2}\sum_{i=1}^{m}\left(\frac{1}{\tau_i}\mathrm{tr}\left((\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}\left(\frac{\partial\mathbf{D}_i\mathbf{L}_i\mathbf{L}_i^\top\mathbf{D}_i}{\partial\gamma_r}\right)\boldsymbol{\Sigma}_i^{-1}\right)\right) \\ &= \sum_{i=1}^{m}\left(\frac{1}{\tau_i}\mathrm{tr}\left((\mathbf{D}_i\mathbf{L}_i)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top(\mathbf{D}_i\mathbf{L}_i)^{-\top}\mathbf{L}_i^{-1}\frac{\partial\mathbf{L}_i}{\partial\gamma_r}\right)\right). \end{aligned}$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial\lambda_s} &= \frac{\partial}{\partial\lambda_s}\left(-\frac{1}{2\tau_i}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i|\right) \\ &+ \frac{\partial}{\partial\lambda_s}\left(-\frac{1}{2\tau_i}\sum_{i=1}^{m}\mathrm{tr}\left((\mathbf{y}_i - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}\right)\right) \\ &= -\sum_{i=1}^{m}\mathrm{tr}\left(\mathbf{D}_i^{-1}\frac{\partial\mathbf{D}_i}{\partial\lambda_s}\right) \\ &+ \sum_{i=1}^{m}\left(\mathrm{tr}\left(\frac{1}{2\tau_i}(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{D}_i^{-1}\frac{\partial\mathbf{D}_i}{\partial\lambda_s}\mathbf{D}_i^{-1}\right)\right)\right) \\ &= \sum_{i=1}^{m}\mathrm{tr}\left(\left(\frac{1}{\tau_i}(\mathbf{y}_i - \boldsymbol{\mu}_i)\boldsymbol{\epsilon}_i{}^\top\boldsymbol{\Sigma}_i^{-1} - \mathbf{I}\right)\mathbf{D}_i^{-1}\frac{\partial\mathbf{D}_i}{\partial\lambda_s}\right), \end{aligned} \tag{A.1}$$

where $|\boldsymbol{\Sigma}_i| = \mathbf{D}_i^2$.

To derive the Fisher information matrix in relation to γ and λ, we use the proposition presented in [33] and reused in [17], which states that let $\boldsymbol{\psi}$ be a generic parameterization of $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$ for the Student-t distribution with the scatter matrix $\boldsymbol{\Sigma}$, the contribution of a single observation to the Fisher information block for the scale parameter is given by:

$$\mathbf{I}_{ij}(\boldsymbol{\psi}) = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}_i}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}_j}\right) - \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}_i}\right)\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}_j}\right). \tag{A.2}$$

As the multivariate Student-t distribution can be reduced to a normal distribution when the degrees of freedom tend to infinity, we will use the same expressions to obtain the one for our case of interest. Then, following this proposition, we obtain the element of the Fisher information matrix in relation to the rth and sth components of vectors γ and/or λ, as shown below

$$\mathbf{I}_{\gamma_r\gamma_s} = \sum_{i=1}^{m}\frac{1}{\tau_i}\operatorname{tr}\left(\frac{\partial\mathbf{L}_i}{\partial\gamma_r}\frac{\partial\mathbf{L}_i^\top}{\partial\gamma_s}\mathbf{L}_i^{-\top}\mathbf{L}_i^{-1}\right), \tag{A.3}$$

$$\mathbf{I}_{\lambda_r\lambda_s} = \sum_{i=1}^{m}\frac{1}{\tau_i}\operatorname{tr}\left(\mathbf{L}_i\mathbf{L}_i^\top\frac{\partial\mathbf{D}_i}{\partial\lambda_r}\boldsymbol{\Sigma}_i^{-1}\frac{\partial\mathbf{D}_i}{\partial\lambda_s} + \mathbf{D}_i^{-2}\frac{\partial\mathbf{D}_i}{\partial\lambda_r}\frac{\partial\mathbf{D}_i}{\partial\lambda_s}\right), \tag{A.4}$$

$$\mathbf{I}_{\gamma_r\lambda_s} = \sum_{i=1}^{m}\frac{1}{\tau_i}\operatorname{tr}\left(\mathbf{D}_i\frac{\partial\mathbf{L}_i}{\partial\gamma_r}\mathbf{L}_i^\top\frac{\partial\mathbf{D}_i}{\partial\lambda_s}\boldsymbol{\Sigma}_i^{-1}\right). \tag{A.5}$$

References

[1] J. J. Heckman, B. Singer, Econometric analysis of longitudinal data, Handbook of econometrics 3 (1986) 1689–1763.

[2] N. R. Chaganty, An alternative approach to the analysis of longitudinal data via generalized estimating equations, Journal of Statistical Planning and Inference 63 (1) (1997) 39–54.

[3] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger, et al., Analysis of longitudinal data, Oxford university press, 2002.

[4] E. Demidenko, Mixed models: theory and applications with R, John Wiley & Sons, 2013.

[5] J. M. Singer, F. M. Rocha, J. S. Nobre, Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures, International Statistical Review 85 (2) (2017) 290–324.

[6] M. Pourahmadi, Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, Biometrika 86 (3) (1999) 677–690.

[7] Z. Chen, D. B. Dunson, Random effects selection in linear mixed models, Biometrics 59 (4) (2003) 762–769.

[8] A. J. Rothman, E. Levina, J. Zhu, Sparse multivariate regression with covariance estimation, Journal of Computational and Graphical Statistics 19 (4) (2010) 947–962.

[9] J. Fan, Y. Liao, H. Liu, An overview of the estimation of large covariance and precision matrices, The Econometrics Journal 19 (1) (2016) C1–C32.

[10] T.-I. Lin, Y.-J. Wang, A robust approach to joint modeling of mean and scale covariance for longitudinal data, Journal of Statistical Planning and Inference 139 (9) (2009) 3013–3026.

[11] Y. Guney, O. Arslan, F. G. Yavuz, Robust estimation in multivariate heteroscedastic regression models with autoregressive covariance structures using em algorithm, Journal of Multivariate Analysis (2022) 105026.

[12] V. S. O. Ribeiro, L. Bombrun, J. S. Nobre, C. C. Cavalcante, Y. Berthoumieu, A general robust approach for joint modeling of the family of scale mixture of Normal distribution, Signal Processing 220 (2024) 109426.

[13] M. Pourahmadi, Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix, Biometrika 87 (2) (2000) 425–435.

[14] H. Ye, J. Pan, Modelling of covariance structures in generalised estimating equations for longitudinal data, Biometrika 93 (4) (2006) 927–941.

[15] F. Pascal, J.-P. Ovarlez, P. Forster, P. Larzabal, On a SIRV-CFAR detector with radar experimentations in impulsive clutter, in: 2006 14th European Signal Processing Conference, IEEE, 2006, pp. 1–5.

[16] Z. Chen, D. B. Dunson, Random effects selection in linear mixed models, Biometrics 59 (4) (2003) 762–769.

[17] M. Maadooliat, M. Pourahmadi, J. Z. Huang, Robust estimation of the correlation matrix of longitudinal data, Statistics and Computing 23 (2013) 17–28.

[18] E. Conte, A. De Maio, G. Ricci, Recursive estimation of the covariance matrix of a compound-Gaussian process and its application to adaptive CFAR detection, IEEE Transactions on Signal Processing 50 (8) (2002) 1908–1915.

[19] F. Pascal, Y. Chitour, J.-P. Ovarlez, P. Forster, P. Larzabal, Covariance Structure Maximum-Likelihood Estimates in Compound Gaussian Noise: Existence and Algorithm Analysis, IEEE Transactions on Signal Processing 56 (1) (2008) 34–48.

[20] S. Cambanis, S. Huang, G. Simons, On the theory of elliptically contoured distributions, Journal of Multivariate Analysis 11 (3) (1981) 368–385.

[21] E. Gómez, M. A. Gómez-Villegas, J. M. Marín, A survey on continuous elliptical vector distributions, Revista matemática complutense 16 (1) (2003) 345–361.

[22] G. Schwarz, Estimating the dimension of a model, The annals of statistics (1978) 461–464.

[23] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the royal statistical society: Series B (Methodological) 36 (2) (1974) 111–133.

[24] S. Geisser, The predictive sample reuse method with applications, Journal of the American statistical Association 70 (350) (1975) 320–328.

[25] S. Kotz, Statistical Distributions in Scientific Work, I., Dordrecht: Reidel, 1968, ch. Multivariate distributions at a cross road, pp. 247–270.

[26] R. F. Potthoff, S. Roy, A generalized multivariate analysis of variance model useful especially for growth curve problems, Biometrika 51 (3-4) (1964) 313–326.

[27] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, R. C. Team, Linear and nonlinear mixed effects models, R package version 3 (57) (2007) 1–89.

[28] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2020). URL `https://www.R-project.org/`

[29] M. Pourahmadi, Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, Biometrika 86 (3) (1999) 677–690.

[30] J. Pan, Y. Pan, jmcm: An R Package for joint mean-covariance modeling of longitudinal data, Journal of Statistical Software 82 (2017) 1–29.

[31] W. Zhang, C. Leng, C. Y. Tang, A joint modelling approach for longitudinal studies, Journal of the Royal Statistical Society Series B: Statistical Methodology 77 (1) (2015) 219–238.

[32] J. S. Nobre, J. M. Singer, Residual analysis for linear mixed models, Biometrical Journal 49 (6) (2007) 863–875.

[33] K. L. Lange, R. J. Little, J. M. Taylor, Robust statistical modeling using the t distribution, Journal of the American Statistical Association 84 (408) (1989) 881–896.

REFERENCES

ABRAMOWITZ, M.; STEGUN, I. A. **Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables**. New York: Dover Publications, 1964.

AITKEN, A. C. Iv. on least squares and linear combination of observations. **Proceedings of the Royal Society of Edinburgh**, v. 55, p. 42–48, 1936.

ANDERSON, T. W. **Statistical Methods for Research Workers**. [S.l.]: Wiley, 2003.

ANDREWS, D. F.; MALLOWS, C. L. Scale mixtures of normal distributions. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 1, p. 99–102, 1974.

ARTIN, E. **The Gamma Function**. [S.l.]: Courier Dover Publications, 2015. ISBN 978-0-486-79555-2.

BOMBRUN, L.; ANFINSEN, S. N.; HARANT, O. A complete coverage of log-cumulant space in terms of distributions for polarimetric SAR data. In: **POLinSAR 2011-5th International Workshop on Science and Applications of SAR Polarimetry and Polarimetric Interferometry**. [S.l.: s.n.], 2011. p. 1–8.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time Series Analysis: Forecasting and Control**. [S.l.]: Wiley, 1976.

BREEN, R.; JONSSON, J. O. Educational expansion and its impact on social stratification: Longitudinal evidence from sweden. **European Sociological Review**, v. 36, n. 5, p. 744–758, 2020.

CAMBANIS, S.; HUANG, S.; SIMONS, G. On the theory of elliptically contoured distributions. **Journal of Multivariate Analysis**, Elsevier, v. 11, n. 3, p. 368–385, 1981.

CAMERON, A. C.; TRIVEDI, P. K. Microeconometrics: Methods and applications. **Journal of Econometrics**, v. 155, n. 2, p. 162–179, 2009.

CHEN, Z.; DUNSON, D. B. Random effects selection in linear mixed models. **Biometrics**, Wiley Online Library, v. 59, n. 4, p. 762–769, 2003.

CHEN, Z.; DUNSON, D. B. Random effects selection in linear mixed models. **Biometrics**, Oxford University Press, v. 59, n. 4, p. 762–769, 2003.

CHITOUR, Y.; PASCAL, F. Exact maximum likelihood estimates for SIRV covariance matrix: Existence and algorithm analysis. **IEEE Transactions on signal processing**, IEEE, v. 56, n. 10, p. 4563–4573, 2008.

COX, D. R. **Regression Models and Life-Tables**. [S.l.: s.n.], 1972. v. 34. 187-220 p.

DEMIDENKO, E. **Mixed models: theory and applications with R**. [S.l.]: John Wiley & Sons, 2013.

DIGGLE, P.; DIGGLE, P. J.; HEAGERTY, P.; LIANG, K.-Y.; ZEGER, S. *et al.* **Analysis of longitudinal data**. [S.l.]: Oxford university press, 2002.

FANG, K.-T.; KOTZ, S.; NG, K. W. **Symmetric multivariate and related distributions**. [S.l.]: Chapman and Hall, 1990.

FERNANDEZ, C.; STEEL, M. F. J. Elliptical distributions and their applications. **Statistics & Probability Letters**, v. 31, n. 2, p. 99–106, 1996.

FERREIRA, C. S.; BOLFARINE, H.; LACHOS, V. H. Linear mixed models based on skew scale mixtures of normal distributions. **Statistical Papers**, 2022.

FERREIRA, M. A.; LACHOS, V. H.; BOLFARINE, H. Skew elliptical distributions for longitudinal data analysis. **Statistical Papers**, 2022.

FISHER, R. A. Theory of statistical estimation. **Proceedings of the Cambridge Philosophical Society**, v. 22, p. 700–725, 1925.

FITZMAURICE, G. M.; LAIRD, N. M.; WARE, J. H. **Applied Longitudinal Analysis**. 2nd. ed. Hoboken, NJ: Wiley, 2011. (Wiley Series in Probability and Statistics).

FORMONT, P.; PASCAL, F.; VASILE, G.; OVARLEZ, J.-P.; FERRO-FAMIL, L. Statistical classification for heterogeneous polarimetric SAR images. **IEEE Journal of selected topics in Signal Processing**, IEEE, v. 5, n. 3, p. 567–576, 2010.

GEISSER, S. The predictive sample reuse method with applications. **Journal of the American statistical Association**, Taylor & Francis, v. 70, n. 350, p. 320–328, 1975.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian Data Analysis**. 3. ed. [S.l.]: Chapman and Hall/CRC, 2013.

GELMAN, A.; HILL, J. Applied regression analysis and generalized linear models. **Sage Publications**, 2003.

GOLDSTEIN, H. **Hierarchical Data Modeling**. [S.l.]: Wiley, 1995.

GÓMEZ, E.; GÓMEZ-VILLEGAS, M. A.; MARIN, J. M. A survey on continuous elliptical vector distributions. **Revista matemática complutense**, Citeseer, v. 16, n. 1, p. 345–361, 2003.

GORE, P. A.; GIL, E. L. Longitudinal data analysis of student learning behaviors: A psychopedagogical perspective. **Learning and Individual Differences**, v. 40, p. 31–42, 2015.

GRANGER, C. W. J.; ENGLE, R. F. Co-integration and error correction: Representation, estimation, and testing. **Econometrica**, v. 55, n. 2, p. 251–276, 1987.

GUNEY, Y.; ARSLAN, O.; YAVUZ, F. G. Robust estimation in multivariate heteroscedastic regression models with autoregressive covariance structures using em algorithm. **Journal of Multivariate Analysis**, Elsevier, p. 105026, 2022.

HAMILTON, J. D. **Time Series Analysis**. [S.l.]: Princeton University Press, 1994.

HAMPEL, F. R.; RONCHETTI, E. M.; ROUSSEEUW, P. J.; STAHEL, W. A. **Robust Statistics: The Approach Based on Influence Functions**. [S.l.]: Wiley, 2011.

HEDEKER, D.; GIBBONS, R. D. **Longitudinal data analysis**. [S.l.]: John Wiley & Sons, 2006.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, v. 31, n. 2, p. 423–447, 1975.

HUBER, P. J.; RONCHETTI, E. M. **Robust Statistics**. 2. ed. [S.l.]: Wiley, 2009.

HUSSAIN, M. S.; GOH, Y. M. Longitudinal analysis of developmental trajectories of children's cognitive abilities: A psychopedagogical study. **Journal of Educational Psychology**, v. 109, n. 5, p. 693–708, 2017.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. **Univariate Discrete Distributions**. 3. ed. Hoboken, NJ: Wiley-Interscience, 2005. (Wiley Series in Probability and Statistics). ISBN 978-0-471-71580-1.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous Univariate Distributions, Volume 1**. 2. ed. Hoboken, NJ:

Wiley-Interscience, 1994. (Wiley Series in Probability and Statistics).

KALBFLEISCH, J. D.; PRENTICE, R. L. **Statistical Analysis of Failure Time Data**. [S.l.]: Wiley-Interscience, 2002.

KLEINBAUM, D. G.; KLEIN, M. **Survival Analysis: A Self-Learning Text**. 3rd. ed. [S.l.]: Springer, 2012.

KOTZ, S.; NADARAJAH, S. **Multivariate Distributions: The Elliptical Case**. [S.l.]: World Scientific, 2000.

LANGE, K.; LITTLE, R. J. A.; TAYLOR, J. M. G. Robust statistical modeling using the t-distribution. **Journal of the American Statistical Association**, v. 84, n. 408, p. 881–896, 1989.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. **Biometrika**, Oxford University Press, v. 73, n. 1, p. 13–22, 1986.

LIN, T.-I.; WANG, Y.-J. A robust approach to joint modeling of mean and scale covariance for longitudinal data. **Journal of Statistical Planning and Inference**, Elsevier, v. 139, n. 9, p. 3013–3026, 2009.

MAADOOLIAT, M.; POURAHMADI, M.; HUANG, J. Z. Robust estimation of the correlation matrix of longitudinal data. **Statistics and Computing**, Springer, v. 23, p. 17–28, 2013.

MAHALANOBIS, P. C. On the generalized distance in statistics. **Proceedings of the National Institute of Sciences of India**, v. 2, p. 49–55, 1936.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **The annals of mathematical statistics**, JSTOR, p. 50–60, 1947.

MANSKI, C. F.; PEPPER, J. V. Longitudinal analysis of social mobility: Evidence from the national longitudinal survey of youth. **American Sociological Review**, v. 86, n. 3, p. 401–424, 2021.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate Analysis**. [S.l.]: Academic Press, 1979.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate Analysis**. [S.l.]: Academic Press, 2000.

MATOS, L. A.; LACHOS, V. H.; LIN, T.-I.; CASTRO, L. M. Heavy-tailed longitudinal regression models for censored data: a robust parametric approach. **TEST**, v. 28, p. 844—-878, 2019.

MILLER, J. L.; BEASLEY, D. H. Independence between observations in statistical analysis. **Journal of Statistical Theory and Practice**, v. 9, n. 3, p. 427–441, 2015.

MORRIS, C. N.; LOCK, K. F. Hierarchical models for combining information and smoothing. **Statistical Science**, v. 21, n. 1, p. 1–19, 2006.

MUIRHEAD, R. J. **Aspects of Multivariate Statistical Theory**. [S.l.]: Wiley, 1982.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. **Applied linear statistical models**. [S.l.]: Irwin Chicago, 1996. v. 4.

NEWSOM, J.; JONES, R. N.; HOFER, S. M. **Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences**. [S.l.]: Routledge, 2013.

NUMMI, T.; MöTTöNEN, J.; VäKEVäINEN, P.; SALONEN, J.; O'BRIEN, T. E. On the improved estimation of the normal mixture components for longitudinal data. **Journal of Statistical Theory and Practice**, 2025.

OSORIO, F.; PAULA, G. A.; GALEA, M. Assessment of local influence in elliptical linear models with longitudinal structure. **Computational Statistics & Data Analysis**, v. 51, n. 9, p. 4354–4368, 2007.

PAN, J.; PAN, Y. jmcm: An R package for joint mean-covariance modeling of longitudinal data. **Journal of Statistical Software**, v. 82, p. 1–29, 2017.

PASCAL, F.; OVARLEZ, J.-P.; FORSTER, P.; LARZABAL, P. On a SIRV-CFAR detector with radar experimentations in impulsive clutter. In: IEEE. **2006 14th European Signal Processing Conference**. [S.l.], 2006. p. 1–5.

PAULA, G. A.; MEDEIROS, M.; VILCA-LABRA, F. E. Influence diagnostics for linear models with first-order autoregressive elliptical errors. **Statistics & probability letters**, Elsevier, v. 79, n. 3, p. 339–346, 2009.

PFOHL, R. S.; YAMADA, M. Modeling longitudinal health data in the presence of informative dropout: Application to cardiovascular risk factors. **Statistics in Medicine**, v. 35, n. 4, p. 658–674, 2016.

PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D.; TEAM, R. C. Linear and nonlinear mixed effects models. **R package version**, v. 3, n. 57, p. 1–89, 2007.

PINHEIRO, J. C.; LIU, C.; WU, Y. N. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution. **Journal of Computational and Graphical Statistics**, v. 10, n. 2, p. 249–276, 2001.

PISCHKE, J.-S. Econometric analysis of panel data. **Handbook of Econometrics**, v. 6, p. 6071–6160, 2005.

PLUNGPONGPUN, K.; NAIK, D. Multivariate analysis of variance using a kotz type distribution. **Proc World Congr Eng**, Citeseer, v. 2, p. 2–4, 2008.

POTTHOFF, R. F.; ROY, S. A generalized multivariate analysis of variance model useful especially for growth curve problems. **Biometrika**, Oxford University Press, v. 51, n. 3-4, p. 313–326, 1964.

POURAHMADI, M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. **Biometrika**, Oxford University Press, v. 86, n. 3, p. 677–690, 1999.

POURAHMADI, M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. **Biometrika**, Oxford University Press, v. 86, n. 3, p. 677–690, 1999.

POURAHMADI, M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. **Biometrika**, Oxford University Press, v. 87, n. 2, p. 425–435, 2000.

R Core Team. **R: A language and environment for statistical computing**. Vienna, Austria, 2020.

RIBEIRO, V. S.; NOBRE, J. S.; SANTOS, J. R. S. dos; AZEVEDO, C. L. Beta rectangular regression models to longitudinal data. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 35, n. 4, p. 851–874, 2021.

RIBEIRO, V. S. O.; BOMBRUN, L.; NOBRE, J. S.; CAVALCANTE, C. C.; BERTHOUMIEU, Y. A general robust approach for joint modeling of the family of scale mixture of normal distribution. **Signal Processing**, Elsevier, p. 109426, 2024.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 54, n. 3, p. 507–554, 2005.

ROBINSON, G. K. That blup is a good thing: The estimation of random effects. **Statistical Science**, v. 6, n. 1, p. 15–32, 1991.

ROSCA, J.; ERDOGMUS, D.; PRINCIPE, J. C.; HAYKIN, S. **Independent component analysis and blind signal separation: 6th International Conference, ICA 2006, Charleston, SC, USA, March 5-8, 2006, Proceedings**. [S.l.]: Springer, 2006. v. 3889.

SAVALLI, C.; PAULA, G. A.; CYSNEIROS, F. J. A. Assessment of variance components in elliptical linear mixed models. **Statistical Modelling**, v. 6, n. 1, p. 59–76, 2006.

SCHAFER, J. L.; KANG, J. W. Average causal effects from nonrandomized studies: A practical guide and applications. **Psychological Methods**, v. 13, n. 4, p. 279–313, 2008.

SCHUMACHER, F. L.; DEY, D. K.; LACHOS, V. H. Approximate inferences for nonlinear mixed effects models with scale mixtures of skew-normal distributions. **Journal of Multivariate Analysis**, 2023.

SCHUMACHER, F. L.; LACHOS, V. H.; MATOS, L. A. Scale mixture of skew-normal linear mixed models with within-subject serial dependence. **Statistics in Medicine**, 2021.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978.

SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Generalized Linear Models**. 2nd. ed. [S.l.]: Wiley, 2009.

SIMCHONI, G.; ROSSET, S. Using random effects to account for high-cardinality categorical features and repeated measures in deep

neural networks. **Advances in Neural Information Processing Systems**, v. 34, p. 25111–25122, 2021.

SIMCHONI, G.; ROSSET, S. Integrating random effects in deep neural networks. **Journal of Machine Learning Research**, v. 24, n. 156, p. 1–57, 2023.

SINGER, J. M.; ANDRADE, D. F. Analysis of longitudinal data. Elsevier, Netherlands, p. 115–160, 2000.

STONE, M. Cross-validatory choice and assessment of statistical predictions. **Journal of the royal statistical society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 2, p. 111–133, 1974.

TABACHNICK, B. G.; FIDELL, L. S. **Using Multivariate Statistics**. 6th. ed. [S.l.]: Pearson, 2013.

TRICOMI, F. G. **The Special Functions**. [S.l.]: Interscience Publishers, 1951.

VENEZUELA, K. M.; BOTTER, D. A.; SANDOVAL, M. C. Diagnostic techniques in generalized estimating equations. **Journal of Statistical Computation and Simulation**, v. 77, n. 10, p. 879–888, 2007.

VENEZUELA, M. Generalized estimating equation and local influence to beta regression models with repeated measures. **Unpublished Ph. D. thesis, Departamento de Estatistica, Universidade de São Paulo, Brazil (in Portuguese)**, 2008.

XU, T.; SUN, J.; BI, J. Longitudinal lasso: Jointly learning features and temporal contingency for outcome prediction. In: **Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2015. p. 1345–1354.

YE, H.; PAN, J. Modelling of covariance structures in generalised estimating equations for longitudinal data. **Biometrika**, v. 93, n. 4, p. 927–941, 12 2006.

ZELLER, C.; CABRAL, C.; LACHOS, V. H. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. **Journal of Statistical Computation and Simulation**, 2023.

ZHANG, W.; LENG, C.; TANG, C. Y. A joint modelling approach for longitudinal studies. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press, v. 77, n. 1, p. 219–238, 2015.

ZHOU, J.; GUI, J.; VILES, W. D.; CHEN, H.; MADAN, J. C.; COKER, M. O.; HOEN, A. G. Identifying microbial interaction networks based on irregularly spaced longitudinal 16s rrna sequence data. **bioRxiv**, Cold Spring Harbor Laboratory, p. 2021–11, 2021.