Universidad de Valparaíso
CHILE
Instituto de Estadística
Facultad de Ciencias

Diagnostic Analysis of Autism Data Using Partially Varying-Coefficient Mixed Models under Elliptical Distributions

Esteban Adolfo Cárcamo Cerda
January 2026

Professor Guide

Germán Ibacache-Pulgar, Ph.D.
Instituto de Estadística, Universidad de Valparaíso, Chile.

Degree work to obtain the title of:

Magíster en Estadística.

Abstract

In this degree work we extend the partially varying-coefficients mixed model with Normal errors to symmetric errors in order to permit distributions with heavier and lighter tails than the Normal distribution. This class of elliptical models includes all symmetric continuous distributions, such as Student's t, Pearson VII, Power Exponential and Logistic, among others. Estimation is performed by maximum penalized likelihood method and by using smoothing splines, also, a diagnostic analysis is performed by considering different perturbation schemes, in addition to a residual analysis. A real data set is analyzed using partially varying-coefficients mixed models with Normal errors and then under heavy-tailed distributions.

Keywords: Local influence measure, maximum penalized likelihood estimators, semiparametric additive model, weighted back-fitting algorithm.

Chapter 1

Introduction and motivation

The partially varying-coefficients mixed model emerge as a powerful tool in statistical modeling because of its flexibility to model explanatory variables effects that can contribute, parametric way through constant coefficients that quantify the contribution in a linear way, random way through random coefficients that quantify the individual or particular effect of the different observations and nonparametric way, in which the effects of explanatory variables is quantified through coefficients that are allowed to vary as nonlinear smooth functions of other variables (for example, time variable). These models are often used in research related to longitudinal, clustered and spatial sampling schemes. In this class of models, usually it is assumed that random errors follow a Normal distribution. However, it is well known that in many cases the Normal distribution is not appropriate and that the least-squares estimates are sensitive to outlying observations. A possible alternative for dealing with this deficiency is to assume, for example, heavy-tailed distributions for the errors. A class of distributions containing distributions with such features is the class of elliptical distributions. The elliptical class includes all symmetric continuous distributions such as Normal, Student's t, Power Exponential, Logistics I and II, Contaminated Normal, among others. The variety of error distributions with different kurtosis coefficients gives more flexibility for analyzing data sets using light and heavy-tailed distributions.

In the last years various works on regression models under elliptical errors were developed, including on the class of semiparametric mixed models. For example, an application of a model, under Power Exponential distribution, was carried out, in which repeated measurements problems were presented (see, for instance, Lindsey, 1999). On the other hand, Lange et al. (1989) worked on an approach to modeling the multivariate Student's t distribution in which the degrees of freedom were assumed known and were also assumed unknown and therefore an estimation process was de-

rived. More recently, elliptical linear mixed models were developed, in wich the emphasis was the assessment of variance components (see, for instance, Savalli et al., 2006), whereas Staudenmayer et al. (2009) worked on mixed models under semiparametric framework, considering a Student's t distribution.

In addition to work related to multivariate elliptic distributions, several diagnostic techniques have also been developed for linear mixed models under errors with Normal distribution. For example, Kim et al. (2002) worked on diagnostic methods for semiparametric regression models, whereas Fung et al. (2002) discussed influence diagnostics methods for outliers under semiparametric mixed models. Beckman (1987) applied the local influence methodology (Cook, 1986) to detect influential observations through linear mixed models considering the Normal distribution, focusing on a single observation, whereas an approach of the local influence method was carried out in linear models under elliptical distributions considering various perturbation schemes in the model (see, for instance, Galea et al., 1977; Liu, 2002). It is important to mention that in nonparametric and semiparametric mixed models the local influence method has not had much theoretical development, however, Zhu et al. (2003) worked on measures of local influence in partial linear models and semiparametric mixed models considering a Normal distribution in the model errors.

Several authors have also contributed to the development of mixed models under semiparametric and varying coefficient frameworks. For instance, Hastie and Tibshirani (1993) introduced varying coefficient models as a flexible extension of classical regression models, whereas Fan and Zhang (1999) studied statistical properties and estimation procedures for models with varying coefficients. The incorporation of random effects into these models was later considered in order to accommodate correlated and longitudinal data structures, leading to mixed models with partially varying coefficients (see, for instance, Wu and Zhang, 2006; Zhang et al., 2007).

Regarding semiparametric mixed models, Ruppert et al. (2003) provided a comprehensive treatment of semiparametric regression using penalized splines, including extensions to mixed-effects representations. Lin and Zhang (1999) discussed inference in generalized additive mixed models, emphasizing the connection between smoothing techniques and random effects. These approaches established an important link between semiparametric modeling and mixed models, facilitating estimation and inference in complex data settings.

In parallel, further advances have been made in the study of elliptical distributions within regression and mixed modeling contexts. Fang et al. (1990) presented a general theory of elliptically contoured distributions, which later served as a foundation for robust modeling in multivariate analysis. Extensions of regression models under elliptical errors were explored by Kano et al. (1993),

while more recent works considered the integration of elliptical distributions into mixed-effects models to improve robustness against heavy-tailed behavior and model misspecification (see, for instance, Pinheiro et al., 2001).

With respect to diagnostic analysis, several authors have extended classical influence and residual-based methods to more complex regression structures. For example, Wei et al. (1998) discussed diagnostics for generalized linear and semiparametric models, whereas Lesaffre and Verbeke (1998) examined local influence techniques in mixed models with clustered data. Additionally, extensions of diagnostic methods to varying coefficient and semiparametric models have been proposed to assess the impact of influential observations and model perturbations beyond the linear framework (see, for instance, Carroll et al., 1998; Zhu and Wei, 2007).

In this degree work the class of multivariate elliptical distributions is introduced, with the objective of extending the distribution of errors beyond a multivariate Normal distribution, allowing the inclusion of distributions with heavier or lighter tails than the Normal one. In addition, an iterative process is introduced to find the estimates of the parameters involved, under the penalized maximum likelihood method. This degree work is presented in different chapters as follows. Chapter 1 presents an example of a data set that serves to motivate the use of the partially varying-coefficients mixed models and some of the most commonly used multivariate models. In Chapter 2 the partially varying-coefficients mixed model is presented under elliptical distributions, also, a penalty is proposed in the likelihood function to specify the parameter estimates. In Chapter 3 an estimation process is discussed, the observed and expected information matrices are presented and an iterative process is derived to maximize the penalized log-likelihood function of the proposed model, in addition, a way to predict the random effects is discussed. Also, some inferential aspects are discussed, the approximate standard errors are presented as well as the bands for these errors, in addition, a discussion on the degrees of freedom is carried out along with a way to select the smoothing parameters. In Chapter 4, some types of residual are presented and the local influence methodology is developed under different perturbation schemes, also the relative change is presented as a metric to confirm the detected local influence. In Chapter 5, the motivation example is reanalyzed using the partially varying-coefficients mixed model under elliptical distributions with different kurtosis coefficients. In the last chapter some conclusions are presented.

1.1 Motivating example

To motivate the use of the partially varying-coefficients mixed model, the data set collected by researchers at the University of Michigan (Oti et al., 2006) will be considered, whose structure is

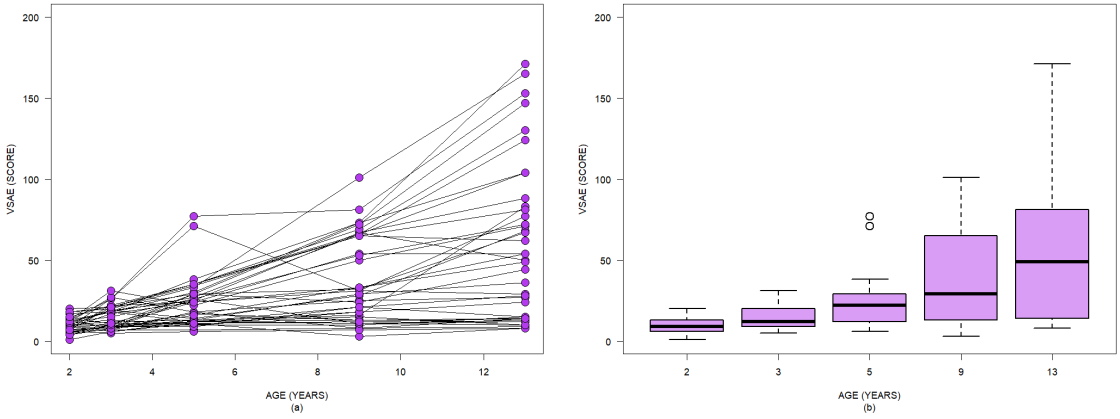


Figure 1.1: Individual profiles (a) and boxplot (b) for children's VSAE score by age.

Statistic	2 years	3 years	5 years	9 years	13 years
Mean	9.81	14.53	23	36.68	57.07
Standard d.	4.38	6.88	14.84	26.29	47.83
Minimum	1	5	6	3	8
Maximum	20	31	77	101	171
Skewness	0.47	0.51	1.92	0.58	0.91

Table 1.1: Main descriptive statistics of the VSAE score calculated by age

longitudinal. In the study carried out, 214 children were considered and they were divided into three groups when they were 2 years old, the first group corresponds to children with autism, the second group corresponds to children with pervasive developmental disorder (PDD) and the last group corresponds to nonspectrum children. The main idea of the study is to collect information about children considering ages 2, 3, 5, 9 and 13 years, however, there are children who do not have measurements at certain ages, and therefore, for simplicity, we will work with balanced profiles, that is, profiles with the same number of observations over time, in this case five measurements. In this motivation example, a subset of 41 children with autism spectrum disorder (ASD) were considered, including autistic and PDD children. Children's social development was assessed at each age using the Vineland Adaptive Behavior Interview survey form, for this purpose, we considered the variable defined as Vineland Socialization Age Equivalent (VSAE) and will be a combined score of differents factors that includes assessments of interpersonal relationships, coping skills, time activities, and play/leisure. The measurements of this variable correspond to the VSAE score obtained

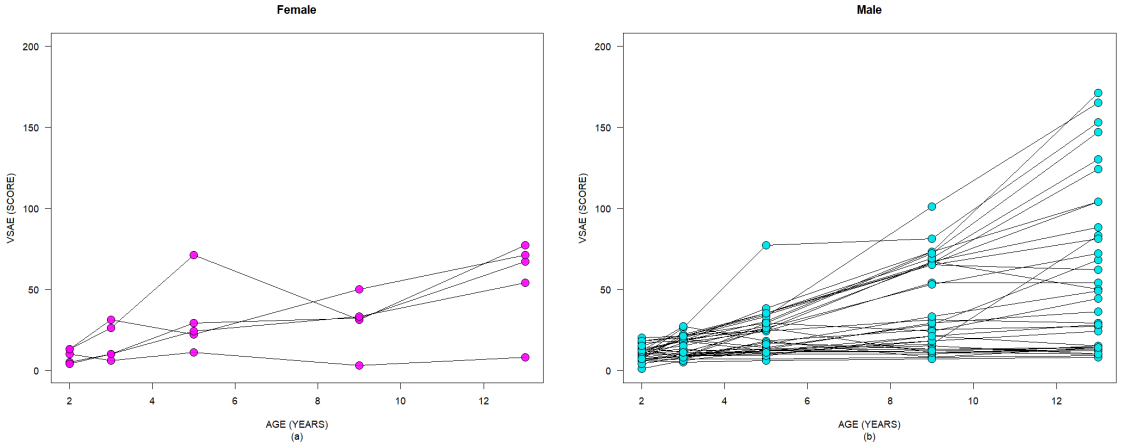


Figure 1.2: Individual VSAE scores for female (a) and male (b) children by age.

by children at different ages, particularly at 2, 3, 5, 9 and 13 years old. Another variable considered was gender, whose levels are female and male. In addition the variable defined as diagnosis at the age of 2 years will be incorporated, the levels of this variable are defined as autism and PDD. Finally, the time variable to use will be defined as age, whose levels are 2, 3, 5, 9 and 13 years. The main idea is to see if language proficiency at age 2, gender, the child's diagnosis (autism or PDD), or other covariates influence the developmental trajectory of the socialization of the children.

Figure 1.1 shows the individual behavior of the VSAE score obtained by the children over time. In Figure 1.1a it can be observed that the scores are increasing over time, in other words, for older ages the children's scores are larger in general, however, the observed distribution of the scores, at the different ages, can be better visualized in Figure 1.1b, in this one it can be observed that the distribution of the scores is quite symmetrical for 2 and 3 years, while for 5 years there are outliers causing a positive asymmetry, for 9 years there are no outliers, however, the dispersion is much larger compared to the previous ages, while the distribution of scores for 13 years is also quite dispersed reaching scores too high compared to the other ages, indicating that there are individual profiles that are atypical with respect to the rest, so that the VSAE score of children can be seen as a nonlinear function of time. Table 1.1 presents the most informative descriptive statistics for the VSAE score, calculated by age. In this table it can be observed that both the mean and the standard deviation increase as age increases, in addition the asymmetry coefficient is positive in all cases, however for the case of 5 years this value is much larger than the rest, confirming the asymmetry presented in Figure 1.1b.

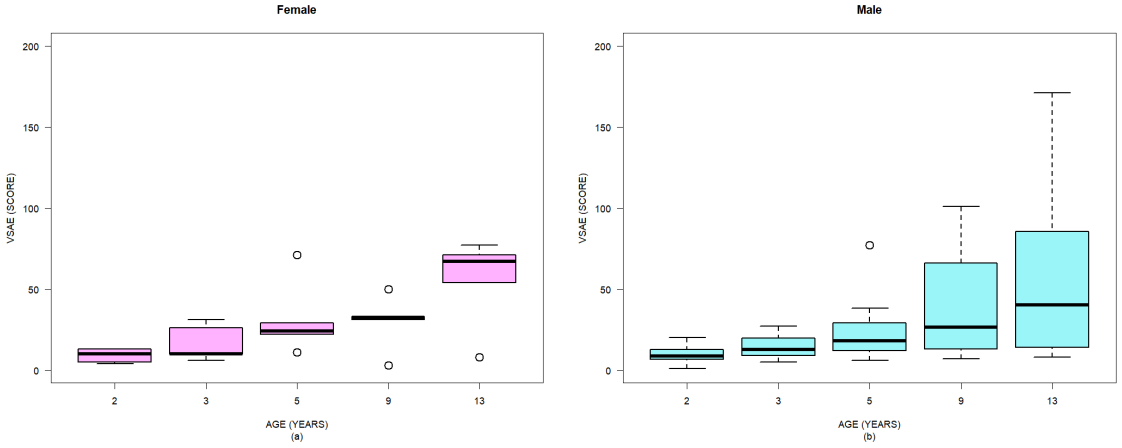


Figure 1.3: Boxplot of VSAE scores for female (a) and male (b) children by age.

Figure 1.2 shows the behavior of the VSAE scores separated by gender, in which it can be observed that for male profiles the scores tend to be higher as age increases, while for female profiles these scores are lower, even for 13 years, where it was previously shown that the scores were quite high and dispersed. In the case of the male gender, it can be observed that individual profiles tend to be more atypical compared to the female gender and despite the fact that there are not many female profiles, the trend in these can be seen.

Figure 1.3 shows a boxplot of the VSAE scores separated by gender. In Figure 1.3b it can be observed that the distribution of scores for the male gender is quite symmetrical for 2 and 3 years, while for 5 years an atypical score is presented. On the other hand, for 9 and 13 years there are no atypical values but the dispersion of the scores increases considerably. In the case of the female profiles, several atypical values can be observed at different ages, however, due to the small number of female profiles, it is not possible to clearly establish how the scores are distributed at different ages, but the trend of the profiles can be established in general terms.

Figure 1.4 shows the behavior of the VSAE scores separated by diagnosis, in which it can be observed that in the profiles of children diagnosed with PDD the scores are slightly higher compared to the scores of children diagnosed with autism, however, in both cases the profiles indicate that the scores are higher for older ages.

Figure 1.5 shows in a better way the distribution of the scores for the different ages through a box plot. In Figure 1.5a it can be observed that, in the case of children diagnosed with PDD, the

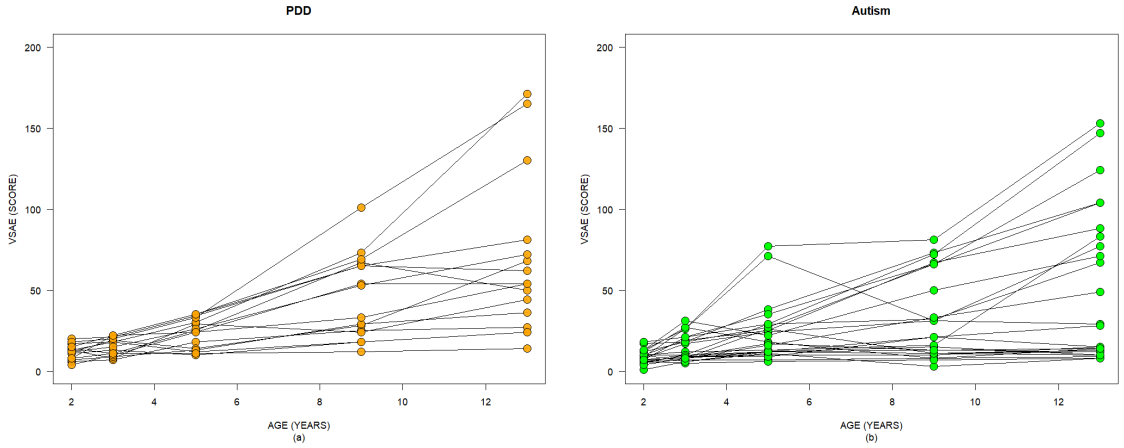


Figure 1.4: Individual VSAE scores for children diagnosed with PDD (a) and autism (b) by age.

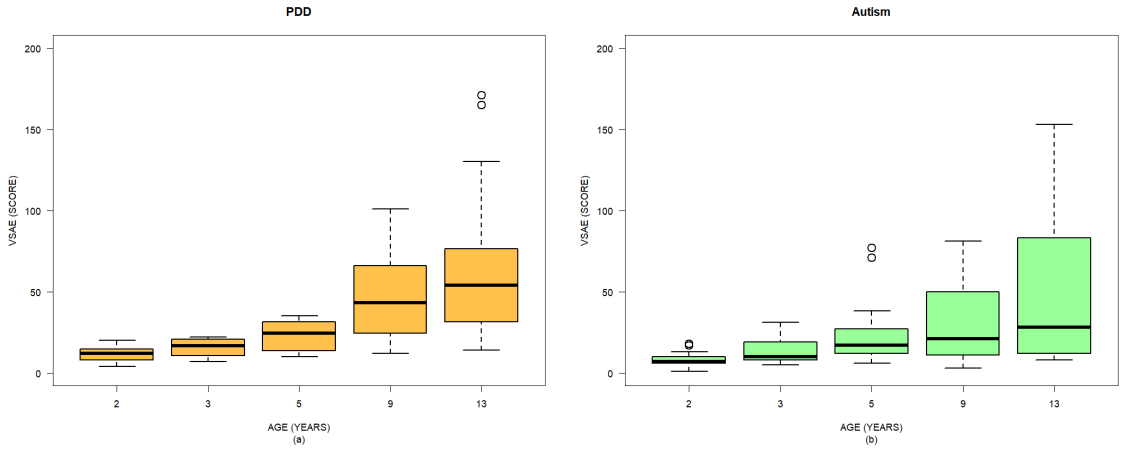


Figure 1.5: Boxplot of VSAE scores for children diagnosed with PDD (a) and autism (b) by age.

distribution of the scores is quite symmetrical at all ages, except for 13 years, where there is clearly a positive asymmetry due to the presence of outliers, while in Figure 1.5b it can be observed that, for children diagnosed with autism, the symmetry is lost a little due to the presence of outliers, particularly at 2 and 5 years. In both cases it can be observed how the dispersion becomes increasingly larger as the ages advance. The observed behavior of the VSAE scores of the children leads us to think that the scores vary over time but in a different way between female and male gender, or, in a different way between diagnosed condition and not only that, since it can also be observed in Figure 1.1a how the intercepts of the profiles seem to be different between children, in addition to their respective slopes, which seem to be different between certain profiles, especially those that are more atypical, so one could think that children could have a particular or individual effect due to this behavior in the intercepts and slopes.

Figure 1.6 shows how VSAE scores are distributed according to the level of expressive language development. This figure shows how VSAE scores tend to be larger as the level of expressive development increases.

1.2 Elliptical Distributions

An elliptic probability distribution is a member of a family of multivariate distributions that have elliptic contours equal to the multivariate Normal distribution and therefore generalize that distribution in that sense. In this family there are distributions whose tails are heavier or lighter than the Normal distribution, allowing this to model data sets with higher or lower kurtosis than the Normal distribution, or to model data sets with outliers. In the case of two-dimensional multivariate distributions the contours form an ellipse, while in the case of three-dimensional the contours form an ellipsoid. Therefore, a random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)^T$ is said to have a multivariate elliptic distribution if its density function can be written as:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1/2} \mathrm{g}((\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})) \quad ,$$

where:

- $\mathrm{g} : [0, \infty) \to [0, \infty)$ is a density generating function,
- $\boldsymbol{\mu}$ is the mean vector,
- $\boldsymbol{\Sigma}$ is the dispersion matrix,

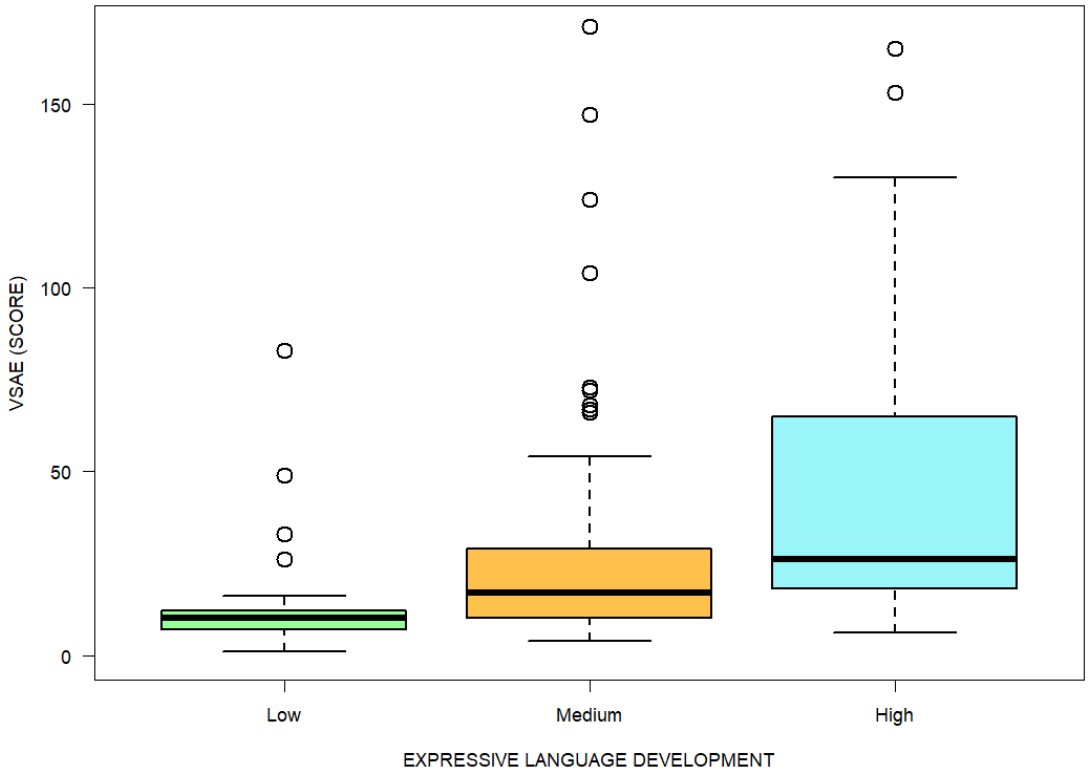


Figure 1.6: Boxplot of VSAE scores by expressive language development level.

- $(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$ is the squared Mahalanobis distance.

The characteristic function of a random vector Y has the form:

$$\varphi_{\boldsymbol{Y}}(\boldsymbol{t}) = \mathbb{E}\Big[e^{i\boldsymbol{t}^\top \boldsymbol{Y}}\Big] = e^{i\boldsymbol{t}^\top \boldsymbol{\mu}}\, \psi\big(\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}\big)\,, \qquad \boldsymbol{t} \in \mathbb{R}^k,$$

where:

- $\psi : [0, \infty) \to \mathbb{C}$ is a scalar function, called the characteristic generator function.

This type of distributions can be used to generalize some statistical procedures based on the Normal distribution, particularly when considering spherical distributions (with finite variance), that is, when considering multivariate distributions whose mean vector is the null vector and whose covariance matrix is the identity matrix multiplied by a constant. Some of the most commonly used elliptic distributions are detailed below.

Multivariate Normal distribution

The multivariate Normal distribution is an elliptic distribution and is a generalization of the classical univariate Normal distribution, therefore, a random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)^T$ is said to have multivariate Normal distribution of dimension k if its probability density function is given by:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{k/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})} \quad ,$$

where $\boldsymbol{\mu}$ is the localization parameter and $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix of dimension (k x k). The vector of means and covariance matrix are $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $V(\boldsymbol{Y}) = \boldsymbol{\Sigma}$ respectively.

Multivariate Power Exponential distribution

Within the class of elliptically contoured distributions, the multivariate Power Exponential distribution extends the univariate case to higher dimensions while allowing for different degrees of kurtosis. A random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)^T$ is defined to have a k-dimensional multivariate Power Exponential distribution when its density function takes the form:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \rho) = \frac{k\Gamma(k/2)}{\pi^{k/2} 2^{1+k/2\rho} \Gamma(1+k/2\rho) |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}((\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}))^\rho} \quad ,$$

where $\boldsymbol{\mu}$ is the localization parameter, $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix of dimension (k x k) and ρ is a shape parameter. The vector of means and covariance matrix are $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $V(\boldsymbol{Y}) = \frac{2^{1/\rho}\Gamma((k+2)/2\rho)}{k\Gamma(k/2\rho)}\boldsymbol{\Sigma}$ respectively. A special case is the multivariate Normal distribution and this results when $\rho = 1$.

Multivariate Student's t distribution

The multivariate Student's t distribution is an elliptically contoured distribution characterized by heavier tails than the multivariate Normal distribution. It generalizes the univariate Student's t distribution to the multivariate setting. A random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)^T$ is said to have a k-dimensional multivariate Student's t distribution if its density function is defined as:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma((\nu + k)/2)}{\Gamma(\nu/2)(\nu\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}}(1 + ((\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}))/\nu)^{-(\nu+k)/2} \quad ,$$

where $\boldsymbol{\mu}$ is the localization parameter, $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix of dimension (k x k) and ν are the degrees of freedom. The vector of means and covariance matrix are $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $V(\boldsymbol{Y}) = (\frac{\nu}{\nu-2})\boldsymbol{\Sigma}$ respectively. A special case is the multivariate Cauchy distribution and this results when $\nu = 1$.

Multivariate Laplace distribution

The multivariate Laplace distribution is an elliptically contoured distribution that generalizes the univariate Laplace distribution to the multivariate context. It is characterized by sharper peaks and heavier tails compared to the multivariate Normal distribution. A random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)^T$ is said to have a multivariate Laplace distribution of dimension k if its probability density function is given by:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \rho) = \frac{\Gamma(k/2)}{\pi^{k/2}2^{1+k}\Gamma(k)|\boldsymbol{\Sigma}|^{1/2}}e^{-\frac{1}{2}((\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}))^{1/2}} \quad ,$$

where $\boldsymbol{\mu}$ is the localization parameter and $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix of dimension (k x k). The vector of means and covariance matrix are $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $V(\boldsymbol{Y}) = 4(k + 1)\boldsymbol{\Sigma}$ respectively. This distribution is a particular case of the multivariate Power Exponential distribution and results when the shape parameter of this distribution is equal to $\frac{1}{2}$.

1.3 Objectives of the thesis

The objectives of this work are:

General objective:

- To study the parameter estimation and diagnostic analysis in the partially varying-coefficients mixed model under elliptical distributions.

Specific objectives:

- To study the theory of the partially varying-coefficients mixed model.
- To derive an iterative process for obtain the parameter estimates.
- To developed a diagnostic analysis for the partially varying-coefficients mixed model.
- To illustrate the methodology proposed with the autism data set.

1.4 Work Hypothesis

- We believe that it is possible to estimate the parameters and to study some aspects of statistical inference under partially varying-coefficients mixed model based on smoothing spline and by using the Backfitting and score of Fisher algorithms.
- We think that is possible to apply some diagnostic techniques for assess the influence of data on parameters estimates.

1.5 Methodology

The following methodology will be used to achieve the specific objectives raised and in the same order as indicated, the steps to do this are listed as follows:

- Review of material related to partially varying-coefficients mixed model.

- Review of material related to Backffiting and non-linear optimization algorithms. Perform an algorithm using R software to implement the results.
- Review of material related to diagnostic analysis for the partially varying-coefficients mixed model.
- To carry out a new application of the model to real data by performed computational implementation.

Chapter 2

The Partially Varying-Coefficients Mixed Models

The partially varying-coefficients model (PVCM) emerge as a powerful tool in statistical modeling because of its flexibility to model explanatory variables effects that can contribute parametric way and explanatory variables effects in which the coefficients are allowed to vary as smooth functions of other variables (for example, time variable), contributing in a non-parametric way. These models are often used in research related to longitudinal, clustered, spatial and hierarchical sampling schemes. However, there are occasions where the individual effect of each observation under study can be quantified and its contribution can be seen randomly, so the effect of the explanatory variable considered can be quantified through random coefficients. In this case the PVCM becomes a partially varying-coefficients mixed model (PVCMM) due to the inclusion of a component whose coefficients are random variables.

2.1 The proposed model

The PVCMM assume that the relationship between the response variable and the explanatory variables can be represented as shown below

$$\mathrm{y}_{ij} = \mathbf{w}_{ij}^T\boldsymbol{\alpha} + \sum_{k=1}^{s} \mathrm{x}_{ij}^{(k)}\, \beta_k(\mathrm{t}_{kij}) + \mathbf{z}_{ij}^T\mathbf{b}_i + \epsilon_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i)\,, \tag{2.1}$$

where y_{ij} denotes the *j*-th measurement associated with the *i*-th cluster at time point t_{kij}; $\mathbf{w}_{ij}$ and

$\mathbf{z}_{ij}$ are $(p \times 1)$ and $(q \times 1)$ vectors of explanatory variables, respectively; $\boldsymbol{\alpha}$ is a $(p \times 1)$ vector of fixed parameters; $\beta_k(\cdot)$ $(k = 1, \ldots, s)$ are unknown smooth functions of $\mathrm{t_k}$ associated with the covariates $\mathrm{x}_{ij}^{(k)}$; $\mathbf{b}_i$ denotes the $(q \times 1)$ vector of random effects; and ε_{ij} is a random error term.

The PVCMM is an extension of several models previously proposed in the literature. In particular,

(i) when the functions β_k are identically zero for $k = 1, \ldots, s$, the PVCMM reduces to the classical linear mixed model (LMM) proposed by Laird and Ware (1982);

(ii) when the random effects $\mathbf{b}_i$ are all zero and $m_i = 1$ for $i = 1, \ldots, n$, the PVCMM reduces to the partially varying-coefficients model (PVCM); see, for instance, Zhang et al. (2007);

(iii) when both the random effects $\mathbf{b}_i$ and the fixed effects vector $\boldsymbol{\alpha}$ are zero, the PVCMM reduces to the varying-coefficients model (VCM) proposed, for example, by Hastie and Tibshirani (1993) and Eubank et al. (2004); and

(iv) when $\boldsymbol{\alpha}$ is zero, the PVCMM reduces to the varying-coefficients mixed model (VCMM), as explored, for example, by Li et al. (2009).

In this class of models, the fixed effects are introduced to describe the expected value of y_{ij}. These effects are represented by the regression coefficient $\boldsymbol{\alpha}$, which lies in the Euclidean space $\mathcal{R}^p$, and by the functions β_k $(k = 1, ..., s)$. The functions β_k are treated as parameters associated with the mean structure of y_{ij}. Consequently, during the estimation process, it is necessary to specify the functional space to which they belong. This specification is typically guided by the domain of the function, prior knowledge about its form, identifiability constraints, or particular features of the application under study.

2.2 Representation matrix

To express model 2.1 in matrix form, we obtain

$$\mathbf{y}_i = \mathbf{W}_i\boldsymbol{\alpha} + \sum_{k=1}^{s} \widetilde{\mathbf{N}}_{ki}\boldsymbol{\beta}_k + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n,\ j = 1, \ldots, m_i. \tag{2.2}$$

where $\mathbf{y}_i$ is an $(m_i \times 1)$ random vector of observed responses from the *i*-th cluster; $\mathbf{W}_i$ is an $(m_i \times p)$ design matrix with rows $\mathbf{w}_{ij}^T$; $\widetilde{\mathbf{N}}_{ki} = \mathbf{X}_i^{(k)}\mathbf{N}_{ki}$, where $\mathbf{X}_i^{(k)} = \mathrm{diag}_{1 \le j \le m_i}\left(\mathrm{x}_{ij}^{(k)}\right)$; and

$\mathbf{N}_{ki}$ is an $(m_i \times r_k)$ incidence matrix whose *(j, l)*-th element equals the indicator $I(\mathrm{t}_{kij} = \mathrm{t}^0_{kl})$, $j = 1, \ldots, m_i$. Here, t^0_{kl} $(l = 1, \ldots, r_k)$ denotes the distinct and ordered values of the vector $\mathbf{t}_k = (\mathbf{t}_{k1}, \mathbf{t}_{k2}, \ldots, \mathbf{t}_{kn})^T$, with $\mathbf{t}_{ki} = (t_{ki1}, t_{ki2}, \ldots, t_{kim_i})^T$. Moreover, $\boldsymbol{\beta}_k = (\psi_{k1}, \psi_{k2}, \ldots, \psi_{kr_k})^T$ is an $(r_k \times 1)$ parameter vector, where $\psi_{kl} = \beta_k(\mathrm{t}^0_{kl})$ for $l = 1, \ldots, r_k$. Furthermore, $\mathbf{Z}_i$ is an $(m_i \times q)$ design matrix with rows $\mathbf{z}^T_{ij}$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^T$ is an $(m_i \times 1)$ vector of within-cluster errors.

A compact representation of model (2.2) is obtained by defining $\mathbf{y} = (\mathbf{y}^T_1, \ldots, \mathbf{y}^T_n)^T$ and constructing $\widetilde{\mathbf{W}}$, $\widetilde{\mathbf{N}}_k$, $\mathbf{b}$, and $\boldsymbol{\epsilon}$ analogously. The model can then be written as

$$\mathbf{y} = \widetilde{\mathbf{W}}\boldsymbol{\alpha} + \widetilde{\mathbf{N}}_1\boldsymbol{\beta}_1 + \cdots + \widetilde{\mathbf{N}}_s\boldsymbol{\beta}_s + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}.$$

where $\mathbf{Z} = \mathrm{blockdiag}_{1 \le i \le n}(\mathbf{Z}_i)$.

2.3 Distribution assumption

In this thesis, we assume that

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{b}_i \end{pmatrix} \sim \mathrm{El}_{(m_i+q)} \left[\begin{pmatrix} \boldsymbol{\mu}_i \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_i & \mathbf{Z}_i\mathbf{D} \\ \mathbf{D}\mathbf{Z}^T_i & \mathbf{D} \end{pmatrix}, \mathrm{g} \right]. \tag{2.3}$$

where $\boldsymbol{\mu}_i = \mathbf{W}_i\boldsymbol{\alpha} + \sum_{k=1}^{s} \widetilde{\mathbf{N}}_{ki}\boldsymbol{\beta}_k$. The random effects $\mathbf{b}_i$ and the error vector $\boldsymbol{\epsilon}_i$ are uncorrelated, but not necessarily independent, except in the Normal case. The matrices $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}^T_i + \mathbf{V}_i$, $\mathbf{D}$, $\mathbf{Z}_i\mathbf{D}$, and $\mathbf{V}_i$ are proportional to the variance–covariance matrices of $\mathbf{y}_i$, $\mathbf{b}_i$, $(\mathbf{y}_i, \mathbf{b}_i)^T$, and $\boldsymbol{\epsilon}_i$, respectively, through a scalar $\xi_i > 0$. Here, g denotes the density generating function defined in Subsection 1.2.

Using properties of the elliptical family in the distribution defined in (2.3), it can be shown that the marginal distribution of $\mathbf{y}_i$ is given by

$$\mathbf{y}_i \sim \mathrm{El}_{m_i}\left[\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\boldsymbol{i}}\right],$$

whose density function, assuming that the matrix $\boldsymbol{\Sigma}_i$ is positive definite ($\boldsymbol{\Sigma}_i > 0$), is given by

$$f(\mathbf{y}_i) = |\boldsymbol{\Sigma}_i|^{-1/2}\, \mathrm{g}(\delta_i), \qquad i = 1, \ldots, n. \tag{2.4}$$

Here, $\delta_i = \mathbf{r}^T_i\boldsymbol{\Sigma}^{-1}_i\mathbf{r}_i$ denotes the squared Mahalanobis distance, with $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$, and $\mathrm{g}(\cdot)$ is the function defined in Subsection 1.2, referred to as the density generating function, satisfying

$$\int_0^\infty \delta^{m_i/2-1}\mathrm{g}(\delta)\,\mathrm{d}\delta < \infty.$$

For simplicity, throughout this work we assume that the matrix $\mathbf{D}$ is parameterized as $\mathbf{D} = \mathbf{D}(\boldsymbol{\tau})$, where $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_d)^T$, and that $\mathbf{V}_i = \phi \mathbf{I}_{m_i}$, where $\mathbf{I}_{m_i}$ denotes the $(m_i \times m_i)$ identity matrix.

2.4 Penalized function

Since the functions β_k belong to an infinite-dimensional space and are treated as parameters in the specification of the expected value of $\mathbf{y}_i$, it is necessary to restrict them to an appropriate subspace in order to ensure identifiability of the model parameters. Therefore, we assume that each function β_k belongs to the Sobolev space

$$\mathcal{W}_2^{(l)} = \left\{ \beta_k : \beta_k, \beta_k^{(1)}, \ldots, \beta_k^{(l-1)} \text{ are absolutely continuous, } \beta_k^{(l)} \in \mathcal{L}^2[a_k, b_k] \right\},$$

where $\beta_k^{(l)}(\mathrm{t}_k) = \frac{\mathrm{d}^l}{\mathrm{dt}^l}\beta_k(\mathrm{t}_k)$, with $\mathrm{t}_k \in [a_k, b_k]$. The log-likelihood function associated with the model (2.4) is then given by

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} L_i(\boldsymbol{\theta}), \tag{2.5}$$

where $L_i(\boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log \mathrm{g}(\delta_i)$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\boldsymbol{T}}, \boldsymbol{\beta}^{\boldsymbol{T}}{}_1, \ldots, \boldsymbol{\beta}^{\boldsymbol{T}}{}_s, \boldsymbol{\tau}^{\boldsymbol{T}})^T \in \boldsymbol{\Theta} \subseteq \mathcal{R}^{p^*}$, and $p^* = p + r + d + 1$, with $r = \sum_{k=1}^{s} r_k$.Let $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_d)^T$, where $\gamma_0 = \phi$ and $\gamma_\ell = \tau_\ell$ for $\ell = 1, \ldots, d$. By incorporating a penalty function for each function β_k, the penalized log-likelihood function can be written as

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda_1, \ldots, \lambda_s) = L(\boldsymbol{\theta}) + \sum_{k=1}^{s} \lambda_k^* J(\beta_k),$$

where $J(\beta_k)$ denotes the penalty function associated with β_k, and $\lambda_k^* = \lambda^*(\lambda_k)$ is a constant depending on the smoothing parameter $\lambda_k \geq 0$. In this work, we will consider penalty functions of type

$$J(\beta_k) = \lambda^* \int_{a_k}^{b_k} [\beta_k^{(l)}(\mathrm{t}_k)]^2 \mathrm{dt}_k,$$

where, for $l = 2$, the estimation of β_k leads to a smooth cubic spline with knots at the points $\mathrm{t}_{k_q}^0$, for $q = 1, \ldots, r_k$. From a functional perspective, the penalization in PVCMMs can be interpreted as the introduction of a norm on the space to which the nonparametric components belong. Since these functions live in infinite-dimensional spaces, the norm induced by the penalty provides a

measure of the complexity of the function and serves to control its regularity. In this setting, the penalty term is typically defined as a norm or seminorm of the form

$$\|\boldsymbol{\beta_k}\|^2_{\mathcal{H}},$$

where $\mathcal{H}$ denotes an appropriate functional space, such as a Sobolev space or a reproducing kernel Hilbert space. This norm penalizes undesirable features of the function, for instance excessive roughness, and promotes smoother solutions. Maximization of the penalized likelihood function can therefore be viewed as a trade-off between goodness of fit, measured by the likelihood, and the magnitude of the function in terms of the chosen norm. The smoothing parameters regulates this balance by controlling the relative weight assigned to data fidelity and functional regularity. According to Green and Silverman (1994), we may express the penalty function as

$$J(\beta_k) = \lambda_k^* \, \boldsymbol{\beta}_k^T \mathbf{K}_k \boldsymbol{\beta}_k,$$

where $\mathbf{K}_k$ is a $(r_k \times r_k)$ nonnegative definite smoothing matrix associated with the *k*-th explanatory variable that depends only on the knots. Then, if we consider $\lambda_k^* = -\lambda_k/2$, the penalized log-likelihood function can be expressed as

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = L_{\mathrm{p}}(\boldsymbol{\theta}) - \sum_{k=1}^{s} \frac{\lambda_k}{2} \, \boldsymbol{\beta}_k^T \mathbf{K}_k \boldsymbol{\beta}_k \,, \tag{2.6}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s)^T$ denotes an $(s \times 1)$ vector of smoothing parameters that controls the tradeoff between goodness of fit and the smoothness of estimated functions. The choice of the smoothing parameters associated with the nonparametric components plays a crucial role in nonparametric modeling, as it directly determines the balance between model flexibility and regularity. An inadequate selection may lead either to oversmoothing, masking important features of the data, or to undersmoothing, resulting in overfitting and unstable estimates. Common selection methods, such as cross-validation, generalized cross-validation, or information criteria, can be interpreted as data-driven mechanisms for calibrating this trade-off. Rather than relying on fixed or subjective choices, these approaches adapt the degree of smoothness to the information contained in the data, thereby enhancing the interpretability and reliability of the estimated functions. From a practical perspective, the importance of these methods lies in their ability to translate abstract regularization concepts into objective and reproducible procedures. By selecting smoothing parameters that balance goodness of fit and model complexity, these criteria help ensure that the nonparametric components capture systematic structure rather than noise, leading to models with better predictive performance and greater stability. Consequently, smoothing parameter selection should be regarded not as a secondary technical detail, but as an integral part of the modeling process, with a direct impact on inference, prediction, and scientific interpretation.

Chapter 3

Estimation and Inference

In this chapter we discuss the estimation of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$'s and $\boldsymbol{\tau}$ based on penalized log-likelihood. First, we present a maximization procedure to obtain the MPLEs and some sufficient conditions on the existence of the MPLEs in PVCMM under elliptical errors. Then, from the application of the maximization procedure we derive a back-fitting algorithm for obtaining the MPLEs. A conditional explicit solution is derived for the parameter estimates, which will be useful for obtaining some diagnostic quantities. On the other hand, we also discuss some aspects of inference in elliptical PVCMMs. A procedure for calculating the approximate standard errors of the parameter estimates and approximate standard error bands is derived for the nonparametric functions. In the following subsections we present a discussion on degrees of freedom estimation and smoothing parameters selection.

3.1 Penalized Score Function

In statistical inference, the score function is defined as the gradient of the log-likelihood function with respect to the parameter vector. Let $L(\boldsymbol{\theta})$ denote the log-likelihood function associated with the observed data and parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. The score function is given by

$$\mathbf{U}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta}) = \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_p} \right)^T .$$

From a geometric perspective, the score vector represents the direction of steepest local increase of the log-likelihood in then parameter space. Consequently, the maximum likelihood estimators

(MLEs) are obtained as solutions of the score equations

$$\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0},$$

which correspond to stationary points of the log-likelihood function. Under standard regularity conditions, the score function has zero expectation at the true parameter value, reflecting the fact that, on average, the log-likelihood exhibits no systematic direction of increase in a neighborhood of the true parameter. Moreover, the covariance matrix of the score function is directly related to the Fisher information matrix, which characterizes the local curvature of the log-likelihood and plays a fundamental role in the asymptotic theory of maximum likelihood estimation. In more complex settings, such as models involving nonparametric components, smoothing penalties or random components, the score function remains the gradient of the penalized log-likelihood. In this context, the corresponding score equations provide the basic mechanism for deriving estimators, although their solutions typically require numerical optimization techniques.

Assuming that the penalized log-likelihood in (2.6) is regular with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_k$'s and $\boldsymbol{\tau}$, the penalized score function of $\boldsymbol{\theta}$ is given by

$$\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial L_{\mathrm{p}_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}}\,.$$

In particular, we obtain

$$\begin{aligned}
\frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\alpha}} &= \mathbf{W}^T \mathbf{W}_v \, \mathbf{r}\,, \\
\frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}_k} &= \widetilde{\mathbf{N}}_k^T \mathbf{W}_v \, \mathbf{r} - \lambda_k \mathbf{K}_s \boldsymbol{\beta}_k \qquad\qquad (k = 1, \ldots, s)
\end{aligned}$$

and

$$\frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \tau_\ell} = \sum_{i=1}^{n} -\frac{1}{2}\left[\mathrm{tr}\Big(\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \ell} \Big) - v_i \, \mathbf{r}_i^T \, \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \ell} \, \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i \right],$$

where $\mathbf{W}_v = \mathrm{blockdiag}_{1 \le i \le n}\big(v_i \, \boldsymbol{\Sigma}_i^{-1} \big)$, $v_i = -2\zeta_g(\delta_i)$, $\zeta_g(\delta_i) = \frac{\mathrm{d} \log g(\delta_i)}{\mathrm{d}\delta_i}$.

Since estimating equations in non-Normal semiparametric models are typically nonlinear and require iterative solution methods, several approaches have been proposed in the literature. Notable examples include the Fisher-scoring and Newton-Raphson algorithms (see, for instance, Green, 1987; Rigby and Stasinopoulus, 2005) as well as the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

3.2 Penalized Observed Information Matrix

Let $\mathbf{L}_{\rm p}$ $(p^* \times p^*)$ the Hessian matrix. The (j^*, ℓ^*)-element of $\mathbf{L}_{\rm p}$, with respect to the parameters $\theta^*_{j^*}$ and $\theta^*_{\ell^*}$, is given by

$$\mathbf{L}_{\rm p} = \frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \theta_{j^*} \theta_{\ell^*}} \, ,$$

for $j^*, \ell^* = 1, \ldots, p^*$. For simplicity, let $\boldsymbol{\Psi}_i = 2\,\boldsymbol{\Psi}_{1i} + \boldsymbol{\Psi}_{2i}$, $\boldsymbol{\Psi}^*_i = \boldsymbol{\Psi}_{1i} + \boldsymbol{\Psi}_{2i}$ and $\boldsymbol{\Psi}^{**}_i = \boldsymbol{\Psi}_{1i} + 2\,\boldsymbol{\Psi}_{2i}$, with $\boldsymbol{\Psi}_{1i} = v_i^{'} \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i \mathbf{r}_i^T \boldsymbol{\Sigma}_i^{-1}$ and $\boldsymbol{\Psi}_{2i} = v_i \boldsymbol{\Sigma}_i^{-1}$. After some algebraic manipulations we find

$$\begin{aligned}
\frac{\partial^2 L_{\rm p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} &= -\mathbf{W}_i^T \boldsymbol{\Psi}_i \mathbf{W}_i \, , \\
\frac{\partial^2 L_{\rm p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_{k'}^T} &= \begin{cases} -\widetilde{\mathbf{N}}_{ki}^T \boldsymbol{\Psi}_i \widetilde{\mathbf{N}}_{ki} - \frac{\lambda_k}{n} \, \mathbf{K}_k & k = k^{'} \\ -\widetilde{\mathbf{N}}_{ki}^T \boldsymbol{\Psi}_i \widetilde{\mathbf{N}}_{k^{'}i} & k \neq k^{'} , \end{cases} \\
\frac{\partial^2 L_{\rm p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}_k^T} &= -\mathbf{W}_i^T \boldsymbol{\Psi}_i \widetilde{\mathbf{N}}_{ki} \, , \\
\frac{\partial^2 L_{\rm p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\alpha} \partial \tau_j} &= -\mathbf{W}_i^T \boldsymbol{\Psi}_i^* \frac{\partial \boldsymbol{\Sigma}_i}{\partial j} \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i \, , \\
\frac{\partial^2 L_{\rm p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}_k \partial \tau_j} &= -\widetilde{\mathbf{N}}_{ki}^T \boldsymbol{\Psi}_i^* \frac{\partial \boldsymbol{\Sigma}_i}{\partial j} \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i \qquad \text{and}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 L_{\rm p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \tau_j \partial \tau_\ell} &= \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}_i^{-1} \Big[\frac{\partial \boldsymbol{\Sigma}_i}{\partial \ell} \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial j} - \frac{\partial^2 \boldsymbol{\Sigma}_i}{\partial j \partial \ell} \Big] \right) - \frac{1}{2} \, \mathbf{r}_i^T \boldsymbol{\Sigma}_i^{-1} \times \\
&\qquad \left[\frac{\partial \boldsymbol{\Sigma}_i}{\partial \ell} \, \boldsymbol{\Psi}_i^{**} \frac{\partial \boldsymbol{\Sigma}_i}{\partial j} - v_i \frac{\partial^2 \boldsymbol{\Sigma}_i}{\partial j \partial \ell} \right] \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i \, ,
\end{aligned}$$

where $v_i^{'} = \frac{\mathrm{d}\, v_i}{\mathrm{d}\delta_i}$.

The Hessian matrix plays a central role in statistics and model-based inference because it captures the local curvature of an objective function, typically the penalized log-likelihood. This curvature is essential for inference, diagnostic analysis, and model interpretation. First, in the computation of the Fisher information matrix, the Hessian arises naturally as the matrix of second-order derivatives of the log-likelihood with respect to the parameters. Under regularity conditions,

the negative expected Hessian coincides with the Fisher information, providing a measure of the amount of information the data carry about the parameters. In practice, especially for complex models, the observed Hessian is commonly used as an approximation to the Fisher information to obtain standard errors, confidence intervals, and hypothesis tests. Thus, the Hessian directly links the local geometry of the likelihood surface to statistical inference. Second, in the context of local influence analysis and Normal curvatures, the Hessian is fundamental for assessing model sensitivity to small perturbations in the data or model assumptions. Normal curvatures depend explicitly on the Hessian of the objective function, as it determines how rapidly the gradient changes in different directions of the parameter space. Large curvatures indicate directions of high sensitivity, allowing the identification of influential observations or perturbation schemes. In this sense, the Hessian provides a geometric characterization of local influence. Third, in the modeling process, particularly for a partially varying coefficient mixed model, the Hessian is crucial from both computational and conceptual perspectives. Such models combine fixed effects, random effects, and coefficients that vary smoothly with covariates, leading to high-dimensional and complex likelihood surfaces.

The Hessian is used to guide optimization algorithms, to assess the stability of the fitted model, and to understand interactions between parametric, nonparametric, and random components. Moreover, its structure reflects how information is distributed across constant and varying components, which is useful for diagnosing identifiability issues and potential overfitting. Overall, the Hessian matrix is not merely a technical object, but a central tool that connects inference, sensitivity analysis, and geometric interpretation of modern statistical models.

3.3 Penalized Expected Information Matrix

In general, by calculating the expectation of the matrix $-\mathbf{L}_{\mathrm{p}}$ we obtain the $(p^* \times p^*)$ penalized expected information matrix, denoted by

$$\boldsymbol{\mathcal{I}}_{\mathrm{p}} \;=\; -\mathrm{E}\left(\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right).$$

Following to Lange et al. (1989), we have that the (j^*,ℓ^*)-element of the matrix $\boldsymbol{\mathcal{I}}_{\mathrm{p}}$ for *i*-th cluster, with respect to the parameters $\theta^*_{j^*}$ and $\theta^*_{\ell^*}$, can be obtained as

$$\boldsymbol{\mathcal{I}}_{\mathrm{p}_i} = \mathrm{E}\left(\frac{\partial L_{\mathrm{p}_i}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\theta_{j^*}}\frac{\partial L_{\mathrm{p}_i}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\theta_{\ell^*}}\right).$$

Let $d_{g_i} = \mathrm{E}(\zeta_g^2(\delta_i)\delta_i)$ and $f_{g_i} = \mathrm{E}(\zeta_g^2(\delta_i)\delta_i^2)$, with $\delta_i = \mathbf{e}_i^T\mathbf{e}_i$, $\mathbf{e}_i \sim \mathrm{El}_{m_i}(\mathbf{0}, \mathbf{I}_{m_i})$, and $\mathbf{W}^* = \mathrm{blockdiag}_{1\leq i\leq n}\big(\frac{4d_{g_i}}{m_i}\,\boldsymbol{\Sigma}_i^{-1}\big)$. After some algebraic manipulations we find

$$\boldsymbol{\mathcal{I}}_{\mathrm{p}} \;=\; \mathrm{blockdiag}\Big(\boldsymbol{\mathcal{I}}_{\theta\theta}\,,\,\boldsymbol{\mathcal{I}}_{\tau\tau}\Big)\,,$$

where

$$\boldsymbol{\mathcal{I}}_{\theta\theta} \;=\; \begin{pmatrix} \mathbf{W}^T\mathbf{W}^*\mathbf{W} & \mathbf{W}^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 & \dots & \mathbf{W}^T\mathbf{W}^*\widetilde{\mathbf{N}}_s \\ \widetilde{\mathbf{N}}_1^T\mathbf{W}^*\mathbf{W} & \widetilde{\mathbf{N}}_1^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 + \lambda_1\mathbf{K}_1 & \dots & \widetilde{\mathbf{N}}_1^T\mathbf{W}^*\widetilde{\mathbf{N}}_s \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathbf{N}}_s^T\mathbf{W}^*\mathbf{W} & \widetilde{\mathbf{N}}_s^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 & \dots & \widetilde{\mathbf{N}}_s^T\mathbf{W}^*\widetilde{\mathbf{N}}_s + \lambda_s\mathbf{K}_s \end{pmatrix}$$

and

$$\boldsymbol{\mathcal{I}}_{\tau\tau} \;=\; \sum_{i=1}^{n} \boldsymbol{\mathcal{I}}_{\mathrm{p}_{i_{j\ell}}}^{\tau\tau_i}\,,$$

where the *(j , ℓ)*-th element of $\boldsymbol{\mathcal{I}}_{\mathrm{p}_i}^{\tau\tau}$ is given by

$$\mathcal{I}_{\mathrm{p}_{i_{j\ell}}}^{\tau\tau_i} \;=\; \left[\frac{b_{i_{j\ell}}}{4}\left(\frac{4f_{g_i}}{m_i(m_i+2)} - 1\right) + \frac{2f_{g_i}}{m_i(m_i+2)}\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial j}\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial \ell}\right)\right],$$

where $b_{i_{j\ell}} = \mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial j}\big)\mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial \ell}\big)$. For some multivariate elliptical distributions it is possible to obtain closed-form expressions for the quantities d_{g_i} and f_{g_i}; for example, for Student's t and Power Exponential distributions.

3.4 Maximizing the penalized log-likelihood function

Since the parameters β_k are infinite-dimensional, we adopt a maximum penalized likelihood approach for the estimation of $\boldsymbol{\theta}$, which naturally yields cubic spline estimators for the functions β_k. Specifically, the value of $\boldsymbol{\theta}$ that maximizes the penalized log-likelihood function $L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ over the parameter space $\boldsymbol{\Theta}$, denoted by $\widehat{\boldsymbol{\theta}}$, is referred to as the maximum penalized likelihood estimator (MPLE) and satisfies

$$L_{\mathrm{p}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\lambda}) \geq \sup_{\theta\in\Theta} L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}).$$

The computation of the MPLE $\widehat{\boldsymbol{\theta}}$ can be carried out through a sequence of successive maximizations, as described, for example, in Gourieroux and Monfort (1995, Chapter 7). In particular, for a fixed value of $\boldsymbol{\lambda}$ and for $s = 2$, the estimator $\widehat{\boldsymbol{\theta}}$ is obtained as the solution to the optimization problem

$$\max_{\theta \in \Theta} L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \max_{\alpha, \beta_1, \beta_2, \tau} L_{\mathrm{p}}(\boldsymbol{\alpha}, \boldsymbol{\beta}1, \boldsymbol{\beta}2, \boldsymbol{\tau}, \boldsymbol{\lambda}) \, ,$$

which can be solved using the following four-step procedure (see Ibacache-Pulgar et al., 2013):

(a) First, the function $L_{\mathrm{p}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda})$ is maximized with respect to $\boldsymbol{\alpha}$, while holding the parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\tau}$ fixed. The resulting maximizer, $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau})$, is attained for values of $\boldsymbol{\alpha}$ belonging to a set $\mathcal{B}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau})$, which depends on the parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\tau}$. Consequently, for $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau})$, the penalized log-likelihood reduces to

$$L_{\mathrm{p}}^{c}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda}) = \max_{\alpha} L_{\mathrm{p}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda}).$$

Here, L_{p}^{c} denotes the penalized log-likelihood concentrated with respect to $\boldsymbol{\alpha}$.

(b) Next, the concentrated penalized log-likelihood

$$L_{\mathrm{p}}^{c}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda}) = L_{\mathrm{p}}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda})$$

is maximized with respect to $\boldsymbol{\beta}_1$, while holding $\boldsymbol{\beta}_2$ and $\boldsymbol{\tau}$ fixed. The corresponding maximizer, $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_1(\boldsymbol{\beta}_2, \boldsymbol{\tau})$, belongs to a set $\mathcal{F}_1(\boldsymbol{\beta}_2, \boldsymbol{\tau})$ depending on $\boldsymbol{\beta}_2$ and $\boldsymbol{\tau}$. Hence, for $\boldsymbol{\beta}_1 \in \mathcal{F}_1(\boldsymbol{\beta}_2, \boldsymbol{\tau})$, the penalized log-likelihood further concentrates as

$$L_{\mathrm{p}}^{c}(\boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda}) = \max_{\beta_1} L_{\mathrm{p}}^{c}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda}) \, .$$

At this stage, L_{p}^{c} denotes the penalized log-likelihood concentrated with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_1$.

(c) Subsequently, the function

$$L_{\mathrm{p}}^{c}(\boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda}) = L_{\mathrm{p}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda})$$

is maximized with respect to $\boldsymbol{\beta}_2$, while keeping $\boldsymbol{\tau}$ fixed. The resulting maximizer, $\widehat{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_2(\boldsymbol{\tau})$, is attained for values of $\boldsymbol{\beta}_2$ in a set $\mathcal{F}_2(\boldsymbol{\tau})$ depending on $\boldsymbol{\tau}$. Therefore, for $\boldsymbol{\beta}_2 \in \mathcal{F}_2(\boldsymbol{\tau})$, the penalized log-likelihood becomes

$$L_{\mathrm{p}}^{c}(\boldsymbol{\tau}, \boldsymbol{\lambda}) = \max_{\beta_2} L_{\mathrm{p}}^{c}(\boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda}) \, .$$

Here, L_{p}^{c} represents the penalized log-likelihood concentrated with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$.

(d) Finally, the function

$$L_{\mathrm{p}}^{c}(\boldsymbol{\tau}, \boldsymbol{\lambda}) = L_{\mathrm{p}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \boldsymbol{\tau}, \boldsymbol{\lambda})$$

is maximized with respect to $\boldsymbol{\tau}$. The resulting estimator $\widehat{\boldsymbol{\tau}}$ is attained over a set $\mathcal{C}$ of admissible values of $\boldsymbol{\tau}$.

The four-step procedure can be generalized for $3 \leq k \leq s$.

3.5 Existence of the MPLE

From a general perspective, the existence of maximum likelihood estimators (MLEs) is not guaranteed without additional regularity conditions on the model and the parameter space. In finite-dimensional parametric models, the existence of the MLE typically follows from the continuity of the log-likelihood function and the compactness of the parameter space, which together ensure that the likelihood attains its supremum. However, in many practical settings the parameter space is not compact, and the likelihood function may be unbounded, leading to non-existence or degeneracy of the MLE.

In models involving infinite-dimensional parameters, such as nonparametric or semiparametric models, the issue of existence becomes more delicate. In these cases, the likelihood function may fail to achieve a maximum due to excessive flexibility of the parameter space, often resulting in ill-posed estimation problems. Penalization techniques are commonly introduced to address this issue by restricting the effective complexity of the model. The addition of a penalty term typically restores well-posedness by enforcing smoothness or regularity conditions on the parameter estimates.

The existence of the MPLE can then be established by analyzing the concavity properties of the penalized log-likelihood function. When the penalized objective function is concave (or strictly concave) over the relevant parameter space, any local maximizer is also a global maximizer, and existence follows under mild conditions. Therefore, concavity plays a central role in ensuring the existence and stability of MPLEs, particularly in models combining finite- and infinite-dimensional parameters.

To establish the existence of the MPLE, it is necessary to study the concavity of the penalized log-likelihood function $L_{\mathrm{p}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda})$ with respect to the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\tau}$. To this end, let

$$\boldsymbol{\Psi} = \mathrm{blockdiag}_{1 \leq i \leq n}\big(\boldsymbol{\Psi}_i\big),$$

where $\boldsymbol{\Psi}_i$ is defined in Subsection 3.2. The concavity conditions corresponding to each step of the estimation procedure are summarized below.

(a') In step (a), concavity with respect to $\boldsymbol{\alpha}$ of $L_{\mathrm{p}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda})$ is ensured if and only if the matrix

$$\mathbf{L}_{\mathrm{p}}^{\alpha\alpha} = -\mathbf{Z}^T \boldsymbol{\Psi} \mathbf{Z}$$

is negative semidefinite, or equivalently, if $-\mathbf{L}_{\mathrm{p}}^{\alpha\alpha}$ is positive semidefinite. This condition holds whenever $\boldsymbol{\Psi} \geq 0$, that is, when $\boldsymbol{\Psi}_i \geq 0$ for all $i = 1, \ldots, n$. The latter is satisfied if and only if $\boldsymbol{\Psi}_{1i} \geq 0$, or equivalently, if $v_i^{'} \mathbf{r}_i \mathbf{r}_i^T \geq 0$ for all $i = 1, \ldots, n$.

(b') In step (b), concavity with respect to $\boldsymbol{\beta}_1$ of $L_{\mathrm{p}}^{c}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda})$ holds if and only if the matrix

$$\mathbf{L}_{\mathrm{p}}^{\beta_1\beta_1} = -\big(\widetilde{\mathbf{N}}_1^T \boldsymbol{\Psi} \widetilde{\mathbf{N}}_1 + \lambda_1 \mathbf{K}_1\big)$$

is negative semidefinite. Equivalently, this requires $-\mathbf{L}_{\mathrm{p}}^{\beta_1\beta_1} \geq 0$, which is satisfied if $\widetilde{\mathbf{N}}_1^T \boldsymbol{\Psi} \widetilde{\mathbf{N}}_1 \geq 0$ and $\lambda_1 \mathbf{K}_1 \geq 0$. Since λ_1 is a positive scalar and $\mathbf{K}_1 \geq 0$, the latter condition always holds. Moreover, $\widetilde{\mathbf{N}}_1^T \boldsymbol{\Psi} \widetilde{\mathbf{N}}_1 \geq 0$ whenever $\boldsymbol{\Psi} \geq 0$, that is, if $\boldsymbol{\Psi}_i \geq 0$ for all $i = 1, \ldots, n$. As in step (a), this is equivalent to requiring $\boldsymbol{\Psi}_{1i} \geq 0$, or $v_i^{'} \mathbf{r}_i \mathbf{r}_i^T \geq 0$ for all $i = 1, \ldots, n$.

(c') Analogously, in step (c), concavity with respect to $\boldsymbol{\beta}_2$ of $L_{\mathrm{p}}^{c}(\boldsymbol{\beta}_2, \boldsymbol{\tau}, \boldsymbol{\lambda})$ is guaranteed if and only if

$$\mathbf{L}_{\mathrm{p}}^{\beta_2\beta_2} = -\big(\widetilde{\mathbf{N}}_2^T \boldsymbol{\Psi} \widetilde{\mathbf{N}}_2 + \lambda_2 \mathbf{K}_2\big)$$

is negative semidefinite. This condition is equivalent to $-\mathbf{L}_{\mathrm{p}}^{\beta_2\beta_2} \geq 0$, which holds whenever $\widetilde{\mathbf{N}}_2^T \boldsymbol{\Psi} \widetilde{\mathbf{N}}_2 \geq 0$ and $\lambda_2 \mathbf{K}_2 \geq 0$. Since $\lambda_2 > 0$ and $\mathbf{K}_2 \geq 0$, the latter inequality is automatically satisfied. Furthermore, $\widetilde{\mathbf{N}}_2^T \boldsymbol{\Psi} \widetilde{\mathbf{N}}_2 \geq 0$ whenever $\boldsymbol{\Psi} \geq 0$, which again is equivalent to $\boldsymbol{\Psi}_{1i} \geq 0$, or $v_i^{'} \mathbf{r}_i \mathbf{r}_i^T \geq 0$ for all $i = 1, \ldots, n$.

(d') Finally, in step (d), concavity with respect to $\boldsymbol{\tau}$ of $L_{\mathrm{p}}^{c}(\boldsymbol{\tau}, \boldsymbol{\lambda})$ is ensured if and only if

$$Ł_{\mathrm{p}}^{\tau\tau} = \frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T} < 0.$$

Therefore, if $\widehat{\boldsymbol{\theta}}$ is obtained as the solution of the four-step procedure (a)–(d), then $\widehat{\boldsymbol{\theta}}$ is the MPLE of $\boldsymbol{\theta}$ if and only if $\boldsymbol{\Psi}_{1i} \geq 0$, or equivalently, $v_i^{'} \mathbf{r}_i \mathbf{r}_i^T \geq 0$ for all $i = 1, \ldots, n$. In particular, under the Normal distribution one has $v_i^{'} = 0$, $v_i = 1$, and $\boldsymbol{\Sigma}_i > 0$, and hence the required concavity conditions are satisfied.

3.6 Fisher Score and weighted Backfitting algorithms

In statistical estimation, iterative methods play a central role in obtaining parameter estimates, particularly in complex or high-dimensional models. Two widely used iterative approaches are Fisher Scoring and Backfitting, which, while sharing the idea of successive updates, differ fundamentally in their context and application.

Fisher Scoring is an adaptation of the Newton-Raphson method for maximum likelihood estimation in parametric models. Instead of using the observed Hessian of the log-likelihood, it employs the expected Fisher information to update parameters iteratively. This approach often enhances numerical stability and ensures rapid convergence near the maximum likelihood estimate, making it particularly suitable for generalized linear models and other finite-dimensional parametric problems.

In contrast, Backfitting is primarily used in semiparametric or nonparametric additive models, where the response depends on the sum of unknown smooth functions of several predictors. The algorithm iteratively updates each function while keeping the others fixed, smoothing residuals at each step, until convergence. Backfitting efficiently decomposes the effects of multiple covariates and allows flexible estimation of each component function, which is crucial when dealing with infinite-dimensional parameter spaces.

Overall, while Fisher Scoring is focused on finite-dimensional parametric estimation with likelihood-based criteria, Backfitting provides a practical solution for high-dimensional or nonparametric settings, highlighting the complementary nature of these iterative methods in modern statistical modeling.

Consider for simplicity $\boldsymbol{\lambda}$ and $\mathbf{W}^*$ fixed. According to Berhane and Tibshirani (1998), the four-step procedure (a)-(d) described in Subsection 3.4 can be solved, for $1 \leq k \leq s$, by using the following Fisher scoring algorithm:

$$\begin{pmatrix} \mathbf{I} & \mathbf{S}_0^{(u)}\widetilde{\mathbf{N}}_1 & \dots & \mathbf{S}_0^{(u)}\widetilde{\mathbf{N}}_s \\ \mathbf{S}_1^{(u)}\mathbf{W} & \mathbf{I} & \dots & \mathbf{S}_1^{(u)}\widetilde{\mathbf{N}}_s \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_s^{(u)}\mathbf{W} & \mathbf{S}_s^{(u)}\widetilde{\mathbf{N}}_1 & \dots & \mathbf{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_0^{(u+1)} \\ \boldsymbol{\beta}_1^{(u+1)} \\ \vdots \\ \boldsymbol{\beta}_s^{(u+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_0^{(u)}\boldsymbol{\eta}^{(u)} \\ \mathbf{S}_1^{(u)}\boldsymbol{\eta}^{(u)} \\ \vdots \\ \mathbf{S}_s^{(u)}\boldsymbol{\eta}^{(u)} \end{pmatrix}, \tag{3.1}$$

where $\boldsymbol{\eta}^{(u)} = \boldsymbol{\mu} + \mathbf{W}^{*^{-1}}\mathbf{W}_v(\mathbf{y} - \boldsymbol{\mu})\big|_{\boldsymbol{\theta}^{(u)}}$ and $\mathbf{S}_k^{(u)} = \mathbf{S}_k\big|_{\boldsymbol{\theta}^{(u)}}$, with

$$\mathbf{S}_k^{(u)} = \begin{cases} (\widetilde{\mathbf{N}}_0^T\mathbf{W}^*\widetilde{\mathbf{N}}_0)^{-1}\widetilde{\mathbf{N}}_0^T\mathbf{W}^*\,\big|_{\boldsymbol{\theta}^{(u)}} & k = 0 \\ (\widetilde{\mathbf{N}}_k^T\mathbf{W}^*\widetilde{\mathbf{N}}_k + \lambda_k\,\mathbf{K}_k)^{-1}\widetilde{\mathbf{N}}_k^T\mathbf{W}^* & k = 1, \ldots, s \end{cases} \tag{3.2}$$

Then, the back-fitting (Gauss-Seidel) iterations that are used to solve the equations system in (3.1) take the form

$$\mathbf{f}_k^{(u+1)} = \mathbf{S}_k^{(u)}\left(\boldsymbol{\eta}^{(u)} - \sum_{l=0, l\neq k}^{s} \widetilde{\mathbf{N}}_l\boldsymbol{\beta}_l^{(u)}\right). \tag{3.3}$$

The system of equations in (3.1) is consistent, and the backfitting algorithm in (3.3) converges to a solution from any set of starting values, provided that the weight matrix is symmetric and positive definite (see, for instance, Berhane and Tibshirani, 1998). Moreover, this solution is unique in the absence of concurvity in the data. Specifically, for model (2.1) with smooth terms β_1 and β_2 but without the constant terms $\boldsymbol{\alpha}$, the following considerations apply:

1. If $\|\mathbf{S}_1\mathbf{S}_2\| < 1$, the estimating equations are consistent and have a unique solution, and the final iterates from the back-fitting algorithm are independent of the starting values and starting order.

2. If $\|\mathbf{S}_1\mathbf{S}_2\| = 1$, this give indication of concurvity in the data (strict collinearity), and therefore the back-fitting algorithm converges to one of the solutions of estimating equations system, and the starting functions determine the final solutions.

3. If the $\mathbf{S}_k$ smoothers are not centered ($\mathbf{S}_k^T\mathbf{1} = \mathbf{1}$), typically $\|\mathbf{S}_1\mathbf{S}_2\| = 1$. In this case, we can consider centered smoother such that $\mathbf{S}_k^T\mathbf{1} = \mathbf{0}$, whit $\mathbf{1}$ denoting a $(r_k \times 1)$ vetor of ones, and defined as

$$\mathbf{S}_k = \left(\mathbf{I}_{(r_k, r_k)} - \frac{\mathbf{1}\mathbf{1}^T}{r_k}\right)\left(\widetilde{\mathbf{N}}_k^T\mathbf{W}^*\widetilde{\mathbf{N}}_k + \lambda_k\mathbf{K}_k\right)^{-1}\widetilde{\mathbf{N}}_k^T\mathbf{W}^*.$$

Note that the back-fitting algorithm in (3.3) depend on the elliptical distribution through the weight matrices $\mathbf{W}^*$. In particular, under Normal distribution for which $v_i = 1$ and $d_{g_i} = \frac{m_i}{4}$, we have that $\mathbf{W}_v = \mathbf{W}^* = \text{blockdiag}_{1\leq i\leq n}\left(\boldsymbol{\Sigma}_i^{-1}\right)$ and, therefore, $\boldsymbol{\eta} = \mathbf{y}$.

3.7 Joint iterative process

The solution of the estimating equation system in (3.1) to obtain the MPLE of $\boldsymbol{\theta}$ may be attained by iterating between a weighted back-fitting algorithm with weight matrix $\mathbf{W}^*$ and a Fisher score algorithm to obtain MPLE of the parameter $\boldsymbol{\tau}$, which is equivalent to the following iterative process:

(i) Initialize:

(a) Fitting a PVCM under Normal errors to get $\boldsymbol{\beta}_j^{(0)}$ $(j = 0, 1, \ldots, s)$.

(b) Get starting value for $\boldsymbol{\tau}$ by using the fitted values from (a).

(c) From the current value $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}_0^{(0)}, \boldsymbol{\beta}_1^{(0)}, \ldots, \boldsymbol{\beta}_s^{(0)}, \boldsymbol{\tau}^{(0)})$ obtaining the weight matrix $\boldsymbol{\Sigma}_i^{(0)} = \boldsymbol{\Sigma}_i\big|_{\boldsymbol{\theta}^{(0)}}$, $\mathbf{W}^{*(0)}$, $v_i^{(0)} = v_i\big|_{\boldsymbol{\theta}^{(0)}}$ and $\mathbf{W}_v^{(0)} = \mathrm{blockdiag}_{1\leq i\leq n}\big(v_i^{(0)}\,{\boldsymbol{\Sigma}_i^{-1}}^{(0)}\big)$. Then, obtain

$$\begin{aligned}
\boldsymbol{\eta}^{(0)} &= \boldsymbol{\mu}^{(0)} + {\mathbf{W}^{*(0)}}^{-1}\mathbf{W}_v^{(0)}(\mathbf{y} - \boldsymbol{\mu}^{(0)}),\\
\mathbf{S}_0^{(0)} &= (\widetilde{\mathbf{N}}_0^T\mathbf{W}^{*(0)}\widetilde{\mathbf{N}}_0)^{-1}\widetilde{\mathbf{N}}_0^T\mathbf{W}^{*(0)} \qquad \text{and}\\
\mathbf{S}_k^{(0)} &= (\widetilde{\mathbf{N}}_k^T\mathbf{W}^{*(0)}\widetilde{\mathbf{N}}_k + \lambda_k\mathbf{K}_k)^{-1}\widetilde{\mathbf{N}}_k^T\mathbf{W}^{*(0)}, \qquad (k = 1, \ldots, s).
\end{aligned}$$

(ii) Step 1: Iterate repeatedly by cycling between the following equations:

$$\begin{aligned}
\boldsymbol{\beta}_0^{(u+1)} &= \mathbf{S}_0^{(u)}\left(\boldsymbol{\eta}^{(u)} - \sum_{l=1}^{s}\widetilde{\mathbf{N}}_l\boldsymbol{\beta}_l^{(u)}\right),\\
\boldsymbol{\beta}_1^{(u+1)} &= \mathbf{S}_1^{(u)}\left(\boldsymbol{\eta}^{(u)} - \widetilde{\mathbf{N}}_0\boldsymbol{\beta}_0^{(u+1)} - \sum_{l=2}^{s}\widetilde{\mathbf{N}}_l\boldsymbol{\beta}_l^{(u)}\right),\\
&\vdots\\
\boldsymbol{\beta}_s^{(u+1)} &= \mathbf{S}_s^{(u)}\left(\boldsymbol{\eta}^{(u)} - \sum_{l=0}^{s-1}\widetilde{\mathbf{N}}_l\boldsymbol{\beta}_l^{(u+1)}\right),
\end{aligned}$$

for $u = 0, 1, \ldots$. Repeat (ii) replacing $\boldsymbol{\beta}_j^{(u)}$ by $\boldsymbol{\beta}_j^{(u+1)}$ until convergence criterion $\Delta_u(\boldsymbol{\beta}_j^{(u+1)}, \boldsymbol{\beta}_j^{(u)}) = \sum_{j=0}^{s} \parallel \boldsymbol{\beta}_j^{(u+1)} - \boldsymbol{\beta}_j^{(u)} \parallel / \sum_{j=0}^{s} \parallel \boldsymbol{\beta}_j^{(u)} \parallel$ is below some small threshold (Hastie and Tibshirani, 1990).

(iii) Step 2: For current values $\boldsymbol{\beta}_j^{(u+1)}$ $(j = 0, 1, \ldots, s)$, obtaining $\boldsymbol{\tau}^{(u+1)}$ by using

$$\boldsymbol{\tau}^{(u+1)} = \boldsymbol{\tau}^{(u)} - \mathrm{E}\left\{ \frac{\partial^2 L_{\mathrm{p}}^c(\boldsymbol{\tau}, \boldsymbol{\lambda})}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T} \right\}^{-1} \frac{\partial L_{\mathrm{p}}^c(\boldsymbol{\tau}, \boldsymbol{\lambda})}{\partial \boldsymbol{\tau}} \Big|_{\boldsymbol{\theta}^{(u)}} .$$

(iv) Iterating between steps (ii) and (iii) by replacing $\boldsymbol{\beta}_j^{(0)}$ $(j = 0, 1, \ldots, s)$ and $\boldsymbol{\tau}^{(0)}$ by $\boldsymbol{\beta}_j^{(u+1)}$ and $\boldsymbol{\tau}^{(u+1)}$, respectively, until convergence.

Note that under Student's t distribution (df = ν) the current weight $v_i^{(u)} = (\nu+1)/(\nu+\delta_i^{(u)})$, with $\delta_i^{(u)} = \delta_i\big|_{\boldsymbol{\theta}^{(u)}}$, is inversely proportional to the distance between the observed value $\boldsymbol{y}_i$ and its current predicted value $\boldsymbol{\mu}_i^{(u)}$, so that outlying observations tend to have small weights in the estimation process.

3.8 Prediction of the random effects

Prediction of the subject-specific random effects $\mathbf{b}_i$ can be performed by computing their conditional expectation given $\mathbf{y}_i$, thereby obtaining, analogously to Normal semiparametric mixed-effects models, the empirical Bayes estimates for the random effects (see, e.g., Zhang et al., 1998). Following Fang et al. (1990), the conditional distribution of $\mathbf{b}_i$ given $\mathbf{y}_i$ is

$$\mathbf{b}_i \mid \mathbf{y}_i \sim \mathrm{El}_q\Big(\mathbf{D}\mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i, \mathbf{D} - \mathbf{D}\mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i\mathbf{D}\Big).$$

Assuming that $\boldsymbol{\Sigma}_i$ is known and using arguments analogous to those for Normal semiparametric mixed-effects models, the empirical Bayes estimate of $\mathbf{b}_i$ is given by

$$\widehat{\mathbf{b}}_i = \widehat{\mathrm{E}}\big(\mathbf{b}_i \mid \mathbf{y}_i\big) = \mathbf{D}\mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\widehat{\mathbf{r}}_i,$$

where $\widehat{\mathbf{r}}_i = \mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i$, with $\widehat{\boldsymbol{\mu}}_i = \mathbf{W}_i\widehat{\boldsymbol{\alpha}} + \sum_{k=1}^s \widetilde{\mathbf{N}}_{ki}\widehat{\boldsymbol{\beta}}_k$, and $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ denote the MPLEs. In practice, since $\boldsymbol{\tau}$ is unknown, it is typically replaced by its MPLE, $\widehat{\boldsymbol{\tau}}$.

Moreover, the shrinkage property observed in Normal mixed-effects models extends naturally to elliptical semiparametric mixed models. Specifically, the predicted profile for the *i*-th subject can be expressed as

$$\begin{aligned} \widehat{\mathbf{y}}_i &= \widehat{\boldsymbol{\mu}}_i + \mathbf{Z}_i\widehat{\mathbf{b}}_i \\ &= \widehat{\phi}\widehat{\boldsymbol{\Sigma}}_i^{-1}\widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{\Psi}}_i\mathbf{y}_i, \end{aligned}$$

which can be interpreted as a weighted average of the population profile $\widehat{\boldsymbol{\mu}}_i$ and the observed data $\mathbf{y}_i$. The corresponding weights are $\widehat{\phi}\widehat{\boldsymbol{\Sigma}}_i^{-1}$ and $\widehat{\boldsymbol{\Psi}}_i = \mathbf{I}_{m_i} - \widehat{\phi}\widehat{\boldsymbol{\Sigma}}_i^{-1}$, respectively. Note that more weight is assigned to the overall mean profile when the residual variability $\widehat{\phi}\mathbf{I}_{m_i}$ is large relative to the between-subject variability $\boldsymbol{\Sigma}_i$, whereas more weight is given to the observed data when the opposite holds.

3.9 Approximate standard errors

In the context of nonparametric regression Segal et al. (1994) noted that the variance estimate for the MPLE developed by Wahba (1983) and Silverman (1985), in the Bayesian context, corresponds to the inverse of the observed information matrix obtained by treating the penalized likelihood as a usual likelihood. Following this same procedure we derive the covariance matrix of $\widehat{\boldsymbol{\theta}}$ from the inverse of the expected information matrix, which is obtained by treating the penalized likelihood function $L_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$ as a usual likelihood. Therefore, the approximate covariance matrix of $\widehat{\boldsymbol{\theta}}$ is given as

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}}) \;\approx\; \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{-1}\big|_{\widehat{\boldsymbol{\theta}}}\,.$$

By using elemental propriedade to partitioned matrix we can show that the inverse matrix of $\boldsymbol{\mathcal{I}}_{\mathrm{p}}$ assumes the following block-diagonal form:

$$\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{-1} \;=\; \begin{pmatrix} \left(\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\alpha} - \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\beta}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\alpha}\right)^{-1} & \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\alpha}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\beta^{-1}} & \mathbf{0} \\ -\left(\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\alpha}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\beta^{-1}}\right)^{T} & \left(\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\beta} - \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta^{T}}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\alpha^{-1}}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\tau\tau^{-1}} \end{pmatrix},$$

where

$$\begin{aligned} \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\alpha} &= \mathbf{Z}^T\mathbf{W}^*\mathbf{Z}\,, \\ \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta} &= \begin{pmatrix} \mathbf{Z}^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 & \ldots & \mathbf{Z}^T\mathbf{W}^*\widetilde{\mathbf{N}}_s \end{pmatrix} \end{aligned}$$

and

$$\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\beta} \;=\; \begin{pmatrix} \widetilde{\mathbf{N}}_1^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 + \lambda_1\mathbf{K}_1 & \ldots & \widetilde{\mathbf{N}}_1^T\mathbf{W}^*\widetilde{\mathbf{N}}_s \\ \vdots & \ddots & \vdots \\ \widetilde{\mathbf{N}}_s^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 & \ldots & \widetilde{\mathbf{N}}_s^T\mathbf{W}^*\widetilde{\mathbf{N}}_s + \lambda_s\mathbf{K}_s \end{pmatrix}.$$

In particular, when the goal is to draw inferences for $\boldsymbol{\alpha}$ and $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s)$, their approximate covariance matrices can be estimated using the corresponding block-diagonal components of $\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{-1}$, that is,

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\alpha}}) \approx \left(\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\alpha} - \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\beta^{-1}}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta^T}\right)^{-1}\Big|_{\widehat{\boldsymbol{\theta}}}$$

and

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_s) \approx \left(\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\beta} - \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta^T}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\alpha^{-1}}\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\alpha\beta}\right)^{-1}\Big|_{\widehat{\boldsymbol{\theta}}}. \tag{3.4}$$

.

3.10 Approximate standard errors bands

By using variance-covariance matrix defined in (3.4) we can construct an approximate pointwise standard error band (SEB) for $\beta_k(\cdot)$ that allows us to assess how accurate the estimator $\widehat{\beta}_k(\cdot)$ at different locations within the range of interest. For example, we can consider the following approximate pointwise SEB (Hastie and Tibshirani, 1990):

$$\widehat{\beta}_k(\mathrm{t}_{kh}) \pm 2\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k(\mathrm{t}_{kh}))},$$

where $\mathrm{Var}(\widehat{\beta}_k(\mathrm{t}_{kh}))$ denotes the *h*-th diagonal element of the matrix (3.4), for $h = 1, \ldots, r_k$ and $k = 1, \ldots, s$. Since the SEB depends on the elliptical distribution through the estimates of the nonparametric functions and their standard errors, in this work we employ the SEB as an empirical tool to explore the robustness of the estimates of $\widehat{f}_k$, in the sense of Staudernmayer et al. (2009).

3.11 On degrees of freedom

In the literature on additive models, different definitions of the degrees of freedom (df) exist depending on the context in which they are applied. Analogously to the classical linear case, the degree of freedom associated with a specific term is defined in terms of the trace of the corresponding smoother matrix. In the elliptical PVCMM, the df associated with the *k*-th nonparametric component is given by (see, for instance, Hastie and Tibshirani, 1990)

$$\mathrm{df}(\lambda_k) = \mathrm{tr}\big(\widetilde{\mathbf{N}}_k\mathbf{S}_k\big) = \mathrm{tr}\big(\widetilde{\mathbf{S}}_k\big), \tag{3.5}$$

which quantifies the individual contribution of the *k*-th component, with $\mathbf{S}_k$ defined in (3.2).

Let $\mathbf{Q}_{\mathbf{N}_k} = \widetilde{\mathbf{N}}_k^T \mathbf{W}^* \widetilde{\mathbf{N}}_k$ and $\mathbf{Q}_{\lambda_k} = \lambda_k \mathbf{K}_k$. Since $\mathbf{W}^* > 0$ and $\mathrm{rank}(\widetilde{\mathbf{N}}_k^T) \leq r_k$, then $\mathbf{Q}_{\widetilde{\mathbf{N}}_k} \geq 0$. Therefore, there exists a matrix $\mathbf{Q}_{\widetilde{\mathbf{N}}_k}^{1/2} \geq 0$ such that $\mathbf{Q}_{\widetilde{\mathbf{N}}_k} = \mathbf{Q}_{\widetilde{\mathbf{N}}_k}^{1/2} \mathbf{Q}_{\widetilde{\mathbf{N}}_k}^{1/2}$. Thus, we can write $\mathrm{tr}\{\widetilde{\mathbf{S}}_k\}$ as (Eilers and Marx, 1996)

$$\begin{aligned} \mathrm{tr}\{\widetilde{\mathbf{S}}_k\} &= \sum_{j=1}^{r_k} \frac{1}{1+\lambda_k \ell_j} \\ &\approx 2 + \sum_{j=3}^{r_k} \frac{1}{1+\lambda_k \ell_j}, \end{aligned}$$

where ℓ_j, for $j = 1, \ldots, r_k$, are the eigenvalues of the matrix $\mathbf{Q}_{\widetilde{\mathbf{N}}_k}^{-1/2} \mathbf{Q}_{\lambda_k} \mathbf{Q}_{\widetilde{\mathbf{N}}_k}^{-1/2}$, for $k = 1, \ldots, s$. It is important to note that (i) $\mathrm{df}(\lambda_k) = \mathrm{tr}\{\widetilde{\mathbf{S}}_k\}$ is a monotonically decreasing function of λ_k; (ii) $\mathrm{df}(\lambda_k) \to 2 + r_k$ as $\lambda_k \to 0$; (iii) $\mathrm{df}(\lambda_k) \to 2$ as $\lambda_k \to \infty$; and (iv) $2 \leq \mathrm{df}(\lambda_k) \leq 2 + r_k$.

3.12 Choosing the smoothing parameters

In the previous chapters, the smoothing parameters λ_k were assumed to be fixed. However, in practical applications, these parameters should be selected from the data. When a smoothing spline is employed, it is common to use either the cross-validation method or the generalized cross-validation method (Craven and Wahba, 1979). Alternatively, the smoothing parameters can be chosen by minimizing information criteria, such as the Akaike Information Criterion (AIC) (Akaike, 1973) or the Bayesian Information Criterion (BIC) (Schwarz, 1978). More details on these approaches can be found in Hurvich et al. (1998), Buja et al. (1989), Simonoff and Tsai (1999), and Rigby and Stasinopoulos (2005). In this work, we adopt the following procedure to select the smoothing parameters based on the AIC (see, for instance, Ibacache-Pulgar et al., 2013):

(i) For simplicity, consider $s = 2$.

(i.1) Select m values $u_{k_\ell} \in (0, 1)$ and obtain the smoothing parameter values $\lambda_{k_\ell} = u_{k_\ell}/(1 - u_{k_\ell})$, for $\ell = 1, \ldots, m$.

(i.2) From the equation 3.5 obtain $\mathrm{df}_{k_\ell} = \mathrm{df}(\lambda_{k_\ell})$ and perform a dispersion graph between λ_{k_ℓ} and df_{k_ℓ}.

(ii) Select a range for the smoothing parameters.

(ii.1) Obtain an appropriate regression obtaining the fitted equation $\widehat{\lambda}_{k_\ell} = \psi(\mathrm{df}_{k_\ell})$, where $\psi(\cdot)$ denotes the regression function.

(ii.2) Since the relationship between λ_k and df_k is monotonically decreasing we may obtain from the fitted regression a range $[\lambda_k^{L_k}, \lambda_k^{U_k}]$ for λ_k given a range for the degrees of freedom. For example, if we consider the range $[2, 16]$ we have that $\lambda_k^{U_k} = \psi(16)$ and $\lambda_k^{L_k} = \psi(2)$.

(iii) Minimizing the AIC.

The suggestion is to select a grid of values from the range $[\lambda_k^{L_k}, \lambda_k^{U_k}]$ and choosing the smoothing parameters values λ_k that minimizes

$$\mathrm{AIC}(\boldsymbol{\lambda}) = -2L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\widehat{\boldsymbol{\theta}}} + 2\Big[1 + p + \mathrm{df}(\boldsymbol{\lambda})\Big],$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$, p denote the number of parameters in $\boldsymbol{\alpha}$, and $\mathrm{df}(\boldsymbol{\lambda}) = \sum_{k=1}^{2} \mathrm{df}(\lambda_k)$ denote approximately the number of effective parameters involved in modelling of the non-parametric effects.

Chapter 4

Diagnostic analysis

In this chapter we present two standardized residual based on the estimates $\widehat{\mathbf{f}}_j$ given in equation 4.1 and matrix representation of the random effects presented in Subsection 3.8, which can be used to detect, for instance, misspecifications of the error distribution as well as the presence of outlying observations and potentially influential data point. Also, we propose the local influence method and derive the perturbation matrix for different perturbation schemes under PVCMM.

4.1 Residual Analysis

In general, in the class of the elliptical semiparametric additive mixed models can be defined three types of residuals; specifically, the marginal residuals, the conditional residuals and scaled residuals; see, for instance, Schabenberger (2004) and Fung et al. (2002). This is discussed to follows.

A Conditional Explicit Solution for the Estimates

By assuming $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$ and $\mathbf{W}_v$ fixed, it is possible to write from this equation the estimates of $\mathbf{f}_j$ directly as (see Opsomer, 1999)

$$\widehat{\mathbf{f}}_j \;=\; \boldsymbol{\mathcal{H}}_j\,\mathbf{y}\,, \quad (j=0,1,\ldots,s)\,, \tag{4.1}$$

where $\boldsymbol{\mathcal{H}}_j = \mathbf{E}_j \mathbf{M}^{-1} \mathbf{S}$ is the smoother matrix obtained when fitting by smoothing spline the *j*-th explanatory variable only, with $\mathbf{E}_j$ being a partitioned matrix given by

$$\mathbf{E}_j = \begin{cases} \Big(\mathbf{I}_{(p\times p)}\ \mathbf{0}_{(p\times r_1)}\ \ldots\ \mathbf{0}_{(p\times r_s)}\Big), & j = 0\,, \\ \Big(\mathbf{0}_{(r_1\times p)}\ \mathbf{I}_{(r_1\times r_1)}\ \ldots\ \mathbf{0}_{(r_1\times r_s)}\Big), & j = 1\,, \\ \qquad\vdots \qquad\qquad\qquad\qquad & \qquad\vdots \\ \qquad\vdots \qquad\qquad\qquad\qquad & \qquad\vdots \\ \Big(\mathbf{0}_{(r_s\times p)}\ \mathbf{0}_{(r_s\times r_1)}\ \ldots\ \mathbf{I}_{(r_s\times r_s)}\Big), & j = s\,, \end{cases}$$

$$\mathbf{M} = \begin{pmatrix} \mathbf{I}_{(p\times p)} & \mathbf{S}_0\mathbf{N}_1 & \ldots & \mathbf{S}_0\mathbf{N}_s \\ \mathbf{S}_1\mathbf{N}_0 & \mathbf{I}_{(r_1\times r_1)} & \ldots & \mathbf{S}_1\mathbf{N}_s \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_s\mathbf{N}_0 & \mathbf{S}_s\mathbf{N}_1 & \ldots & \mathbf{I}_{(r_s\times r_s)} \end{pmatrix}$$

and $\mathbf{S} = (\mathbf{S}_0^T, \mathbf{S}_1^T, \ldots, \mathbf{S}_s^T)^T$.

The MPLEs obtained from equation (4.1) are rarely used in practice, since the back-fitting algorithm is generally more efficient for computing $\widehat{\mathbf{f}}_j$. This is because it avoids the need to manipulate high-dimensional matrices and their inverses. Nevertheless, the expressions above can be useful when studying theoretical properties of the MPLEs or when deriving diagnostic measures based on the residuals.

The Hat Matrix

Let $\widetilde{\boldsymbol{\mathcal{H}}}_j = \mathbf{N}_j \boldsymbol{\mathcal{H}}_j$ $(j = 0, 1, \ldots, s)$, $M = \sum_{i=1}^{n} m_i$. The marginal fitted values vector, denoted by $\widehat{\mathbf{y}}^{[\mathrm{marg}]}$, is given by

$$\begin{aligned} \widehat{\mathbf{y}}^{[\mathrm{marg}]} &= \sum_{j=0}^{s} \mathbf{N}_j \widehat{\mathbf{f}}_j \\ &= \widetilde{\mathbf{H}}(\boldsymbol{\alpha})\,\mathbf{y}\,, \end{aligned} \tag{4.2}$$

where the matrix

$$\widetilde{\mathbf{H}}(\boldsymbol{\alpha}) = \sum_{j=0}^{s} \widetilde{\boldsymbol{\mathcal{H}}}_j = \{\widetilde{h}_{ii'}(\boldsymbol{\alpha})\}_{i,i'=1,\ldots,M}$$

is equivalent to the hat matrix defined in the class of parametric regression models. The leverage points associated with the fixed effects, denoted by $\widetilde{h}_{ii}(\boldsymbol{\alpha})$, are related to the $\widetilde{\mathbf{H}}(\boldsymbol{\alpha})$ matrix and play an important role in the construction of diagnostic techniques in the class of nonparametric and semiparametric regression models. On the other hand, we can define the conditional fitted values vector, denoted by $\widehat{\mathbf{y}}^{[\mathrm{cond}]}$, in the form

$$\begin{aligned} \widehat{\mathbf{y}}^{[\mathrm{cond}]} &= \widehat{\mathbf{y}}^{[\mathrm{marg}]} + \mathbf{Z}\widehat{\mathbf{b}} \\ &= \widetilde{\mathbf{H}}(\boldsymbol{\alpha})\,\mathbf{y} + \boldsymbol{\Delta}\mathbf{Z}^T\boldsymbol{\Phi}\Big(\mathbf{y} - \widehat{\mathbf{y}}^{[\mathrm{marg}]}\Big) \\ &= \mathbf{H}(\boldsymbol{\alpha})\,\mathbf{y}\,, \end{aligned} \tag{4.3}$$

where the matrix

$$\mathbf{H}(\boldsymbol{\alpha}) = \widetilde{\mathbf{H}}(\boldsymbol{\alpha}) + \boldsymbol{\Delta}\mathbf{Z}^T\boldsymbol{\Phi}\left\{\mathbf{I}_{(M,M)} - \widetilde{\mathbf{H}}(\boldsymbol{\alpha})\right\}$$

is equivalent to the hat matrix defined in the class of parametric regression models. Similarly to such models, the principal diagonal elements $h_{ii}(\boldsymbol{\alpha})$ are called leverage points.

REMARK 1: The properties of the matrix $\widetilde{\mathbf{H}}(\boldsymbol{\alpha})$ in the context of nonparametric regression are discussed, for example, in Eubank (1984) and Silverman (1985), and its application to construct diagnostic techniques in the class of semiparametric mixed models is presented in Fung et al. (2002).

Marginal Residual

It follows from equation (4.2) that the marginal residuals vector is the difference between the observed data vector and estimated (marginal) mean vector, this is,

$$\widehat{\mathbf{r}}^{[\mathrm{marg}]} = \mathbf{y} - \widehat{\mathbf{y}}^{[\mathrm{marg}]} = \left[\mathbf{I} - \widetilde{\mathbf{H}}(\boldsymbol{\alpha})\right]\mathbf{y}\,.$$

Since that $\widetilde{\mathbf{H}}(\boldsymbol{\alpha})$ is not a projection operator, this is, $\widetilde{\mathbf{H}}^2(\boldsymbol{\alpha}) \neq \widetilde{\mathbf{H}}(\boldsymbol{\alpha})$, we have that the approximate variance of the marginal residual vector is given by

$$\mathrm{Var}_{\mathrm{approx}}(\widehat{\mathbf{r}}^{[\mathrm{marg}]}) = \left[\mathbf{I} - \widetilde{\mathbf{H}}(\boldsymbol{\alpha})\right]\mathrm{Var}(\mathbf{y})\left[\mathbf{I} - \widetilde{\mathbf{H}}(\boldsymbol{\alpha})\right]^T$$

where $\mathrm{Var}(\mathbf{y})$=$\mathrm{blockdiag}_{1\leq i\leq n}\big(\xi_i\boldsymbol{\Sigma}_i\big)$. Then, we have that the *l*-th studentized marginal residual takes the form

$$\widehat{\mathrm{r}}_l^{[\mathrm{marg}]} = \frac{\mathbf{d}_l^T\left[\mathbf{I} - \widetilde{\mathbf{H}}(\boldsymbol{\alpha})\right]\mathbf{y}}{\sqrt{\mathbf{d}_l^T\widehat{\mathrm{Var}}_{\mathrm{approx}}(\widehat{\mathbf{r}}^{[\mathrm{marg}]})\mathbf{d}_l}}\,, \tag{4.4}$$

where $\widehat{\mathrm{Var}}_{\mathrm{approx}}(\widehat{\mathbf{r}}^{[\mathrm{marg}]})$=$\mathrm{Var}_{\mathrm{approx}}(\widehat{\mathbf{r}}^{[\mathrm{marg}]})\big|_{\widehat{\boldsymbol{\theta}}}$, whit $\mathbf{d}_l$ denoting an $(M\times 1)$ vector with 1 at the *l*-th position and 0 elsewhere, for $l = 1,\ldots,M$.

Conditional Residual

It follows from equation (4.3) that the conditional residuals vector is the difference between the observed data vector and the predicted value of the vector of observed responses, this is,

$$\widehat{\mathbf{r}}^{[\mathrm{cond}]} = \mathbf{y} - \widehat{\mathbf{y}}^{[\mathrm{cond}]} = \big[\mathbf{I} - \mathbf{H}(\boldsymbol{\alpha})\big]\,\mathbf{y}\,.$$

Here the name conditional residual stems from the fact that $\boldsymbol{\mu}_i + \mathbf{Z}_i\mathbf{b}_i$ is the conditional mean of $\mathbf{y}_i$. Since that $\mathbf{H}(\boldsymbol{\alpha})$ is not a projection operator, this is, $\mathbf{H}^2(\boldsymbol{\alpha}) \neq \mathbf{H}(\boldsymbol{\alpha})$, we have that the approximate variance of the conditional residual vector is given by

$$\mathrm{Var}_{\mathrm{approx}}(\widehat{\mathbf{r}}^{[\mathrm{cond}]}) \;\; = \;\; \big[\mathbf{I} - \mathbf{H}(\boldsymbol{\alpha})\big]\,\mathrm{Var}(\mathbf{y})\big[\mathbf{I} - \mathbf{H}(\boldsymbol{\alpha})\big]^T.$$

Consequently, we have that the *l*-th studentized conditional residual takes the form

$$\widehat{\mathrm{r}}_l^{[\mathrm{cond}]} \;\; = \;\; \frac{\mathbf{d}_l^T\big[\mathbf{I} - \mathbf{H}(\boldsymbol{\alpha})\big]\,\mathbf{y}}{\sqrt{\mathbf{d}_l^T\widehat{\mathrm{Var}}_{\mathrm{approx}}(\widehat{\mathbf{r}}^{[\mathrm{cond}]})\mathbf{d}_l}}\,, \tag{4.5}$$

where $\widehat{\mathrm{Var}}_{\mathrm{approx}}(\widehat{\mathbf{r}}^{[\mathrm{cond}]})=\mathrm{Var}_{\mathrm{approx}}(\widehat{\mathbf{r}}^{[\mathrm{cond}]})\big|_{\widehat{\boldsymbol{\theta}}}$, with $\mathbf{d}_l$ denoting an $(M \times 1)$ vector with 1 at the *l*-th position and 0 elsewhere, for $l = 1, \ldots, M$. The first n_1 elements of the vector $\widehat{\mathbf{r}}^{[\mathrm{cond}]}$ (or $\widehat{\mathbf{r}}^{[\mathrm{marg}]}$) corresponding to the observations of the first cluster, and so on.

Scaled Residual

Let $\widehat{\mathbf{r}}_i^{[\mathrm{marg}]} = \mathbf{y}_i - \widehat{\mathbf{y}}_i^{[marg]}$, with $\widehat{\mathbf{y}}_i^{[marg]} = \boldsymbol{\mu}_i\big|_{\widehat{\boldsymbol{\theta}}}$, and $\widehat{\boldsymbol{\Sigma}}_i^{-\frac{1}{2}}$ the Cholesky factor such that $\widehat{\boldsymbol{\Sigma}}_i = \big(\widehat{\boldsymbol{\Sigma}}_i^{-\frac{1}{2}}\big)^T\widehat{\boldsymbol{\Sigma}}_i^{-\frac{1}{2}}$. Then, we can consider scaled residuals vector for the *i*-th cluster defined as

$$\widehat{\mathbf{r}}_i^{[\mathrm{scal}]} \;\; = \;\; \widehat{\boldsymbol{\Sigma}}_i^{-\frac{1}{2}}\widehat{\mathbf{r}}_i^{[\mathrm{marg}]}\,. \tag{4.6}$$

The scaled residuals have zero mean and are approximately uncorrelated, since $\boldsymbol{\Sigma}_i$ is an estimated matrix and this is not the variance of $\widehat{\mathbf{r}}_i^{[\mathrm{marg}]}$(Schabenberger , 2004). However, this type of residuals can be useful to diagnose whether the covariance structure of the elliptical semiparametric additive mixed model has been specified correctly (see Li, 2009).

REMARK 2: Since the distribution of the residuals (conditional, marginal or scaled) is not known, we can generate envelopes, as suggested by Atkinson (1981), in order to know the empirical distribution of r_i^* and thus to detect misspecifications of the error distribution as well as the presence of outlying observations.

4.2 Local influence method

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ be an $(n \times 1)$ vector of perturbations restricted to some open subset $\Omega \in \mathcal{R}^n$ and the logarithm of the perturbed penalized likelihood denoted by $L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{\omega})$. Suppose that there is a point $\boldsymbol{\omega}_0 \in \Omega$ that represents no perturbation of the data so that $L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{\omega}_0) = L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda})$. To assess the influence of minor perturbations on $\widehat{\boldsymbol{\theta}}$, we consider the likelihood displacement $LD(\boldsymbol{\omega}) = 2\Big[L_{\rm p}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\lambda}) - L_{\rm p}(\widehat{\boldsymbol{\theta}}_\omega, \boldsymbol{\lambda})\Big] \geq 0$, where $\widehat{\boldsymbol{\theta}}_\omega$ is the MPLE under $L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{\omega})$. The measure $LD(\boldsymbol{\omega})$ is useful for assessing the distance between $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_\omega$. Cook (1986) suggested studying the local behavior of $LD(\boldsymbol{\omega})$ around $\boldsymbol{\omega}_0$. The procedure consists in selecting a unit direction $\boldsymbol{\ell} \in \Omega$ ($\|\boldsymbol{\ell}\| = 1$), and then to consider the plot of $LD(\boldsymbol{\omega}_0 + a\boldsymbol{\ell})$ against a, where $a \in \mathcal{R}$. This plot is called lifted line. Each lifted line can be characterized by considering the Normal curvature $C_\ell(\boldsymbol{\theta})$ around $a = 0$. The suggestion is to consider the direction $\boldsymbol{\ell} = \boldsymbol{\ell}_{max}$ corresponding to the largest curvature $C_{\ell_{max}}(\boldsymbol{\theta})$. The index plot of $\boldsymbol{\ell}_{max}$ may reveal those observations that under small perturbations exercise notable influence on $LD(\boldsymbol{\omega})$. According to Cook (1986), the Normal curvature in the unitary direction $\boldsymbol{\ell}$ is given by $C_\ell(\boldsymbol{\theta}) = -2\{\boldsymbol{\ell}^T \boldsymbol{\Delta}_{\rm p}^T \mathbf{L}_{\rm p}^{-1} \boldsymbol{\Delta}_{\rm p} \boldsymbol{\ell}\}$, where $\mathbf{L}_{\rm p}$ is the Hessian matrix and $\boldsymbol{\Delta}_{\rm p}$ is the perturbation matrix.

4.3 Perturbation schemes

In this subsection we present the expressions of the elements of the $(p^* \times n)$ $\boldsymbol{\Delta}_{\rm p}$ matrix for case-weight, scale matrix and explanatory variable perturbation schemes.

Case-weight perturbation

Let us consider the attributed weights for the observations in the penalized log-likelihood function as

$$L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda} | \boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \, L_i(\boldsymbol{\theta}) - \sum_{k=1}^{s} \frac{\lambda_k}{2} \, \boldsymbol{\beta}_k^T \mathbf{K}_k \boldsymbol{\beta}_k \,, \tag{4.7}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ is the vector of weights, with $0 \leq \omega_i \leq 1$ $(i = 1, \ldots, n)$. In this case, $\boldsymbol{\omega}_0 = (1, \ldots, 1)^T$. Differentiating the term (4.7) with respect to the elements of $\boldsymbol{\theta}$ and ω_i, we obtain

the expressions

$$\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial\boldsymbol{\alpha}\partial\omega_i}\right|_{\widehat{\boldsymbol{\theta}},\,\boldsymbol{\omega}_0} = v_i\,\mathbf{Z}_i^T\,\boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i\Big|_{\widehat{\boldsymbol{\theta}}}\,,$$

$$\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial\boldsymbol{\beta}_k\partial\omega_i}\right|_{\widehat{\boldsymbol{\theta}},\,\boldsymbol{\omega}_0} = v_i\,\widetilde{\mathbf{N}}_{ki}^T\,\boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i\Big|_{\widehat{\boldsymbol{\theta}}} \qquad (k=1,\ldots,s)$$

and

$$\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial\tau_j\partial\omega_i}\right|_{\widehat{\boldsymbol{\theta}},\,\boldsymbol{\omega}_0} = -\frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial j}\Big) + \frac{1}{2}v_i\,\mathbf{r}_i^T\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial j}\,\boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i\Big|_{\widehat{\boldsymbol{\theta}}}\,,$$

for $i=1,\ldots,n$ and $j=0,1,\ldots,d$.

Scale perturbation

Under the scale parameter perturbation scheme we assume that

$$\mathbf{y}_i \sim \mathrm{El}_{m_i}\big(\boldsymbol{\mu}_i,\omega_i^{-1}\,\boldsymbol{\Sigma}_i\big)\,,$$

where $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_n)^T$ is the vector of perturbations, with $\omega_i>0$ $(i=1,\ldots,n)$. In this case, $\boldsymbol{\omega}_0=(1,\ldots,1)^T$ such that $L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega}_0)=L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda})$. Taking differentials of $L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})$ with respect to the elements of $\boldsymbol{\theta}$ and ω_i, simple algebra yields

$$\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial\boldsymbol{\beta}\partial\omega_i}\right|_{\widehat{\boldsymbol{\theta}},\,\boldsymbol{\omega}_0} = \big\{v_i'\,\delta_i+v_i\big\}\,\mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i\Big|_{\widehat{\boldsymbol{\theta}}}\,,$$

$$\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial\boldsymbol{\beta}_k\partial\omega_i}\right|_{\widehat{\boldsymbol{\theta}},\,\boldsymbol{\omega}_0} = \big\{v_i'\,\delta_i+v_i\big\}\,\widetilde{\mathbf{N}}_{ki}^T\boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i\Big|_{\widehat{\boldsymbol{\theta}}} \qquad (k=1,\ldots,s)$$

and

$$\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial\tau_j\partial\omega_i}\right|_{\widehat{\boldsymbol{\theta}},\,\boldsymbol{\omega}_0} = \frac{1}{2}\big\{v_i'\,\delta_i+v_i\big\}\,\mathbf{r}_i^T\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial j}\boldsymbol{\Sigma}_i^{-1}\,\mathbf{r}_i\Big|_{\widehat{\boldsymbol{\theta}}}\,,$$

for $i=1,\ldots,n$ and $j=0,1,\ldots,d$.

Explanatory variable perturbation

Here the *k*-th explanatory variable (assumed continuous) is perturbed by considering additive perturbation schemes, namely $\mathbf{z}_{ik\omega}=\mathbf{z}_{ik}+\boldsymbol{\omega}_i$ $(i=1,\ldots,n)$, where $\mathbf{z}_{ik}$ denotes the *k*-th column of the matrix $\mathbf{Z}_i$ and $\boldsymbol{\omega}_i=(\omega_1,\ldots,\omega_{m_i})^T$ is the vector of perturbations. In this

case, the design matrix $\mathbf{Z}_i$ associated to the parametric components of the model is replaced by $\mathbf{Z}_{i\omega} = \begin{pmatrix} \mathbf{z}_{i1} & \ldots & \mathbf{z}_{ik\omega} & \ldots & \mathbf{z}_{ip} \end{pmatrix}$. Thus, the perturbed penalized log-likelihood function is constructed from (2.6) with $\mathbf{Z}_i$ replaced by $\mathbf{Z}_{i\omega}$, that is,

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega}) = L(\boldsymbol{\theta}|\boldsymbol{\omega}) - \sum_{k=1}^{s} \frac{\lambda_k}{2}\, \boldsymbol{\beta}_k^T \mathbf{K}_k \boldsymbol{\beta}_k \,, \tag{4.8}$$

where $L(\cdot)$ is given by (2.5) and evaluated at $\delta_{i\omega} = \mathbf{r}_{i\omega}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_{i\omega}$, with $\mathbf{r}_{i\omega} = \mathbf{y}_i - \boldsymbol{\mu}_{i\omega}$ and $\boldsymbol{\mu}_i = \mathbf{Z}_i\boldsymbol{\alpha} + \sum_{k=1}^{s} \widetilde{\mathbf{N}}_{ki}\boldsymbol{\beta}_k$. In this case, we have $\boldsymbol{\omega}_0 = \mathbf{0} \in \mathcal{R}^{n^*}$, with $n^* = \sum_{i=1}^{n} m_i$. Differentiating the term (4.8) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\omega}_i$, we obtain

$$\begin{aligned}
\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\omega}_i^T}\right|_{\widehat{\boldsymbol{\theta}},\, \boldsymbol{\omega}_0} &= -2v_i' \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i \mathbf{r}_i^T \boldsymbol{\Sigma}_i^{-1} \alpha_t - v_i \big\{ \mathbf{Z}_i^T \alpha_t - \mathbf{c}_t \mathbf{r}_i^T \big\} \boldsymbol{\Sigma}_i^{-1} \Big|_{\widehat{\boldsymbol{\theta}}}, \\
\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\omega}_i^T}\right|_{\widehat{\boldsymbol{\theta}},\, \boldsymbol{\omega}_0} &= -\widetilde{\mathbf{N}}_{ki}^T \boldsymbol{\Sigma}_i^{-1} \big\{ 2v_i'\, \mathbf{r}_i \mathbf{r}_i^T + v_i \boldsymbol{\Sigma}_i \big\} \boldsymbol{\Sigma}_i^{-1} \alpha_t \Big|_{\widehat{\boldsymbol{\theta}}} \qquad (k = 1, \ldots, s)
\end{aligned}$$

and

$$\left.\frac{\partial^2 L_{\mathrm{p}_i}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \tau_j \partial \boldsymbol{\omega}_i^T}\right|_{\widehat{\boldsymbol{\theta}},\, \boldsymbol{\omega}_0} = -\mathbf{r}_i^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial j} \boldsymbol{\Sigma}_i^{-1} \big\{ v_i'\, \mathbf{r}_i \mathbf{r}_i^T + v_i' \boldsymbol{\Sigma}_i \big\}\, \boldsymbol{\Sigma}_i^{-1} \alpha_k \Big|_{\widehat{\boldsymbol{\theta}}},$$

for $i = 1, \ldots, n$ and $j = 0, 1, \ldots, d$. Here α_t denotes the *t*-th element of $\boldsymbol{\alpha}$ and $\mathbf{c}_t$ is a $(p \times 1)$ vector with 1 in the *t*-th position and zero elsewhere.

4.4 Confirmatory analysis

In this analysis, influential profiles are removed from the original dataset, and then the MPLEs are recalculated to quantify the change in the respective estimates and associated standard errors. The relative change, expressed as a percentage, is used as the metric for this analysis. These changes for $\hat{\alpha}_h$ are calculated from

$$RC(\hat{\alpha}_{h(i)}) = \left| \frac{\hat{\alpha}_h - \hat{\alpha}_{h(i)}}{\hat{\alpha}_h} \right| \times 100$$

$$RC(SE(\hat{\alpha}_{h(i)})) = \left| \frac{\hat{\alpha}_h - \hat{\alpha}_{h(i)}}{\hat{\alpha}_h} \right| \times 100$$

where $\hat{\alpha}_{h(i)}$ and SE($\hat{\alpha}_{h(i)}$) denote the MPLEs of α_h and its corresponding standard error, obtained after extracting the *i*-th profile, for $i = 1, 2, \ldots, n$ and $h = 1, 2, \ldots, p$, where p is the number of parameters in $\boldsymbol{\alpha}$. The relative change for $\hat{\boldsymbol{\beta}}_k$ is obtained in the same way by replacing $\hat{\alpha}_h$ and SE($\hat{\alpha}_h$) by $\hat{\beta}_{kl}$ and SE($\hat{\beta}_{kl}$) respectively. The relative change for $\hat{\phi}$, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ is obtained in the same way.

Chapter 5

Application to real data-set

The autism-related dataset analyzed in Chapter 1 will be considered in this application. In Figures 1.1 - 1.5 it could be observed that the VSAE score of the children can be seen as a nonlinear function of time whose behavior is different when considering the gender of the children or their diagnosed condition, furthermore it was concluded that the intercepts and slopes have a different behavior between profiles, thus, to model this data set the following mixed model with partially varying coefficients is suggested:

$$\mathrm{y}_{ij} = \mathbf{w}_{ij}^T\boldsymbol{\alpha} + \mathrm{x}_{1ij}\,\beta_1(\mathrm{t}_j) + \mathrm{x}_{2ij}\,\beta_2(\mathrm{t}_j) + \mathbf{z}_{ij}^T\mathbf{b}_i + \epsilon_{ij} \quad , \tag{5.1}$$

where y_{ij} represents the VSAE score obtained by the *i*-th child at time j ($i = 1, \ldots, 41$; $j = 1, \ldots, 5$), $\mathbf{w}_{ij}$ is a vector such that $\mathbf{w}_{ij} = (1, l_{ij})$, where l_{ij} represents the expressive language development of the *i*-th child at time j, however this variable is ordinal categorical, so it is necessary to construct as many artificial dichotomous variables as the number of categories of the variable minus one, in this case two artificial variables, then $\mathbf{w}_{ij} = (1, l_{ij}^{(1)}, l_{ij}^{(2)})$, where $l_{ij}^{(1)}$ is a value such that $l_{ij}^{(1)} = 1$ if the child's expressive language development is medium and $l_{ij}^{(1)} = 0$ if it is low or high, $l_{ij}^{(2)}$ is a value such that $l_{ij}^{(2)} = 1$ if the child's expressive language development is high and $l_{ij}^{(2)} = 0$ if it is low or medium, the remaining category is used as a base and is not included because it can cause multicollinearity problems between the artificial variables, α is a column vector of fixed parameters, $f_1(t_j)$ and $f_2(t_j)$ are smooth functions of time t_j.

Gender and diagnosed condition are nominal categorical variables and, as with the expressive language development, the corresponding artificial variables must be created, in this case an artificial variable for each case, then $\mathrm{x}_{1ij}^{(1)}$ is a value such that $\mathrm{x}_{1ij}^{(1)} = 0$ if the gender of the *i*-th profile is male and $\mathrm{x}_{1ij}^{(1)} = 1$ if the gender is female, $\mathrm{x}_{2ij}^{(1)}$ is a value such that $\mathrm{x}_{2ij}^{(1)} = 0$ if the diagnosed

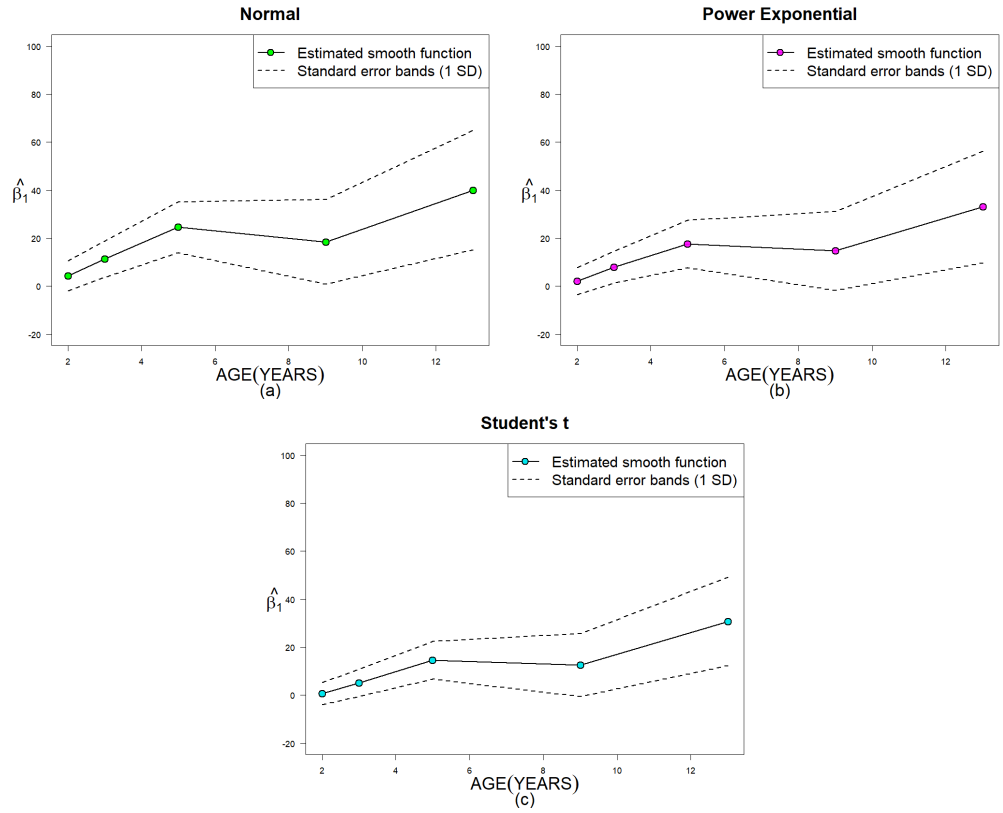


Figure 5.1: Estimated smooth function $\hat{\boldsymbol{\beta}}_1$ and standard errors bands $SE(\hat{\boldsymbol{\beta}}_1)$ for the female gender under different elliptical distributions.

	Normal	PE ($\rho = 0.75$)	Student's t ($\nu = 8$)
Parameter	Estimate	Estimate	Estimate
α_0	4.26 (2.89)	5.19 (2.73)	5.89 (2.16)
α_1	-2.52 (3.94)	-1.80 (3.72)	-1.25 (2.93)
α_2	-0.67 (4.22)	-0.71 (3.99)	-0.62 (3.15)
$\beta_1(t_1^0)$	4.27 (6.23)	2.06 (5.68)	0.73 (4.61)
$\beta_1(t_2^0)$	11.23 (7.50)	7.93 (6.70)	5.08 (5.57)
$\beta_1(t_3^0)$	24.56 (10.61)	17.63 (9.87)	14.53 (7.88)
$\beta_1(t_4^0)$	18.42 (17.59)	14.67 (16.49)	12.47 (13.09)
$\beta_1(t_5^0)$	39.99 (24.87)	32.99 (23.34)	30.76 (18.51)
$\beta_2(t_1^0)$	8.68 (4.07)	6.93 (3.81)	5.87 (3.07)
$\beta_2(t_2^0)$	11.92 (4.61)	9.95 (4.29)	8.85 (3.53)
$\beta_2(t_3^0)$	19.27 (6.30)	16.95 (5.91)	15.11 (4.71)
$\beta_2(t_4^0)$	41.96 (10.07)	36.73 (9.46)	32.67 (7.50)
$\beta_2(t_5^0)$	62.10 (14.08)	52.29 (13.22)	44.06 (10.48)
ϕ	80.07 (10.13)	25.96 (3.64)	35.19 (5.62)
γ_1	28.70 (19.25)	10.96 (6.86)	15.95 (9.96)
γ_2	17.36 (3.94)	5.88 (1.42)	8.34 (2.15)
$L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$	-834.08	-821.02	-806.44
AIC	1695.17	1667.31	1640.88
λ_1	0.0000073	0.0203	0.0000017
λ_2	0.03302	0.0733	0.000045

Table 5.1: MPLEs, in parentheses standard errors and smoothing parameters when fitting different elliptical distributions to the autism data set.

condition of the *i*-th profile is autism and $\mathrm{x}_{2ij}^{(1)} = 1$ if the diagnosed condition is PDD, $\mathbf{z}_{ij}$ is a vector such that $\mathbf{z}_{ij} = (1, t_j)$, b_i corresponds to the vector of effects associated with the *i*-th child, which contains the random intercept and random slope, ϵ_{ij} are the respective random errors of the model. Using the notation given in equation 2.2 the model in 5.1 is expressed as:

$$\mathbf{Y}_i = \mathbf{W}_i\boldsymbol{\alpha} + \mathbf{X}_{1i}^{(1)}\mathbf{N}_{1i}\boldsymbol{\beta}_1 + \mathbf{X}_{2i}^{(1)}\mathbf{N}_{2i}\boldsymbol{\beta}_2 + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad ,$$

where $\mathbf{Y}_i$ is a (5×1) vector containing the VSAE score responses of the *i*-th child, $\mathbf{W}_i$ is a matrix such that $\mathbf{W}_i = (\mathbf{1}, \mathbf{L}_i^{(1)}, \mathbf{L}_i^{(2)})$, with $\mathbf{1} = (1, \ldots, 1)^T$ an (5×1) vector of ones, $\mathbf{L}_i^{(1)} = (l_{i1}^{(1)}, \ldots, l_{i5}^{(1)})^T$ and $\mathbf{L}_i^{(2)} = (l_{i1}^{(2)}, \ldots, l_{i5}^{(2)})^T$ are (5×1) vectors, $\mathbf{T}_i$ a (5×1) vector containing the time values, such that $\mathbf{T}_i = (2, 3, 5, 9, 13)^T$, $\mathbf{X}_{1i}^{(1)} = \mathrm{diag}_{1\leq j\leq 5}\left(\mathrm{x}_{1ij}^{(1)}\right)$, $\mathbf{X}_{2i}^{(1)} = \mathrm{diag}_{1\leq j\leq 5}\left(\mathrm{x}_{2ij}^{(1)}\right)$,

$\mathbf{N}_{1i} = \mathbf{I}_5$ and $\mathbf{N}_{2i} = \mathbf{I}_5$, with $\mathbf{I}_5$ the identity matrix of order 5, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are (5×1) vectors whose elements correspond to the function β_1 and β_2 evaluated at the time values in the set $t^0 = \{t_1^0 = 2, t_2^0 = 3, t_3^0 = 5, t_4^0 = 9, t_5^0 = 13\}$. Finally, $\mathbf{Z}_i$ is a matrix such that $\mathbf{Z}_i = (\mathbf{1}, \mathbf{T}_i)$. Then,

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{b}_i \end{pmatrix} \sim \mathrm{El}_7 \left[\begin{pmatrix} \boldsymbol{\mu}_i \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_i} & \mathbf{Z}_i\mathbf{D} \\ \mathbf{D}\mathbf{Z}_i^T & \mathbf{D} \end{pmatrix}, \mathrm{g} \right],$$

where $\boldsymbol{\mu}_i = \mathbf{W}_i\boldsymbol{\alpha} + \mathbf{X}_i^{(1)}\mathbf{N}_{1i}\boldsymbol{\beta}_1 + \mathbf{X}_i^{(2)}\mathbf{N}_{2i}\boldsymbol{\beta}_2$, $\mathbf{D} = \begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}$, $\boldsymbol{\Sigma_i} = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \phi\mathbf{I}_5$, where ϕ is the variance of the random errors of the model. Then, the response vector associated with the *i*-th child has a distribution given by

$$\mathbf{y}_i \sim \mathrm{El}_5\left[\boldsymbol{\mu}_i, \boldsymbol{\Sigma_i}\right].$$

The MPLE of $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\tau}^T)^T$, with $\boldsymbol{\tau} = (\phi, \gamma_1, \gamma_2)^T$, the penalized log-likelihood function evaluated at $\hat{\boldsymbol{\theta}}$, and the Akaike Information Criterion (AIC), under different fitted elliptical distributions, are presented in Table 5.1. For the case of the Power Exponential distribution, it is assumed that the *i*-th shape parameter ρ_i is the same for all i, that is, $\rho_i = \rho$. This value is selected to minimize the AIC, which leads to $\rho = 0.75$ under this criterion. Similarly, for the Student's t distribution, a common number of degrees of freedom is assumed for all i, that is, $\nu_i = \nu$. Using the same selection criterion as above, an approximate value of $\nu = 8$ is obtained. In Table 5.1 it can be observed that the estimates of the parametric component of the model are similar between all the distributions adjusted to the autism data, however, the standard errors are smaller in the Power Exponential and Student's t distributions, the latter being the one with the lowest standard errors.

Regarding the non-parametric components, it can be observed that the estimates are similar between the distributions with heavier tails, but for the Normal model these estimates are considerably larger, including the respective standard errors. Regarding the variance components it can be observed that the estimates are closer under the Power Exponential and Student's t models while for the Normal model they are quite high, however, they are not comparable because they are measured on different scales.

On the other hand, in Table 5.1 it can be observed that the value of the penalized log-likelihood function, which produces the fit in the autism data, is higher when considering the Power Exponential and Student's t models compared to the Normal model, while the AIC value is lower for these heavier-tailed models, indicating that models with heavier tails have a better fit on the autism dataset in terms of this criterion.

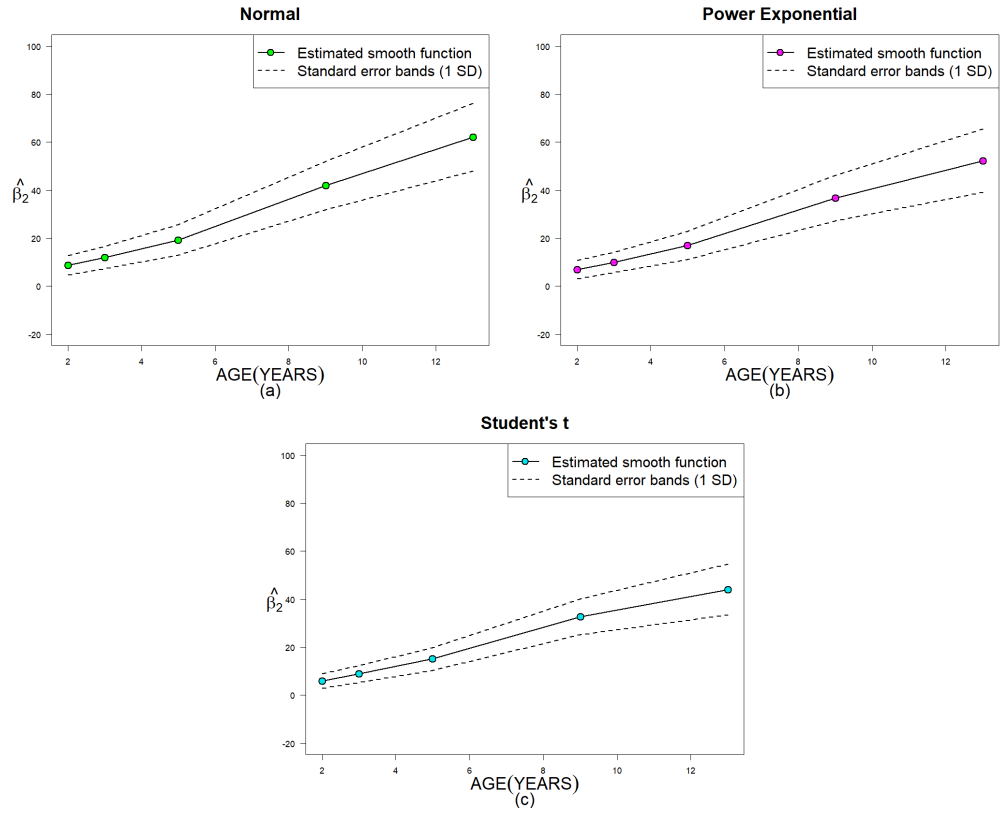


Figure 5.2: Estimated smooth function $\hat{\boldsymbol{\beta}}_2$ and standard errors bands $SE(\hat{\boldsymbol{\beta}}_2)$ for children diagnosed with PDD under different elliptical distributions.

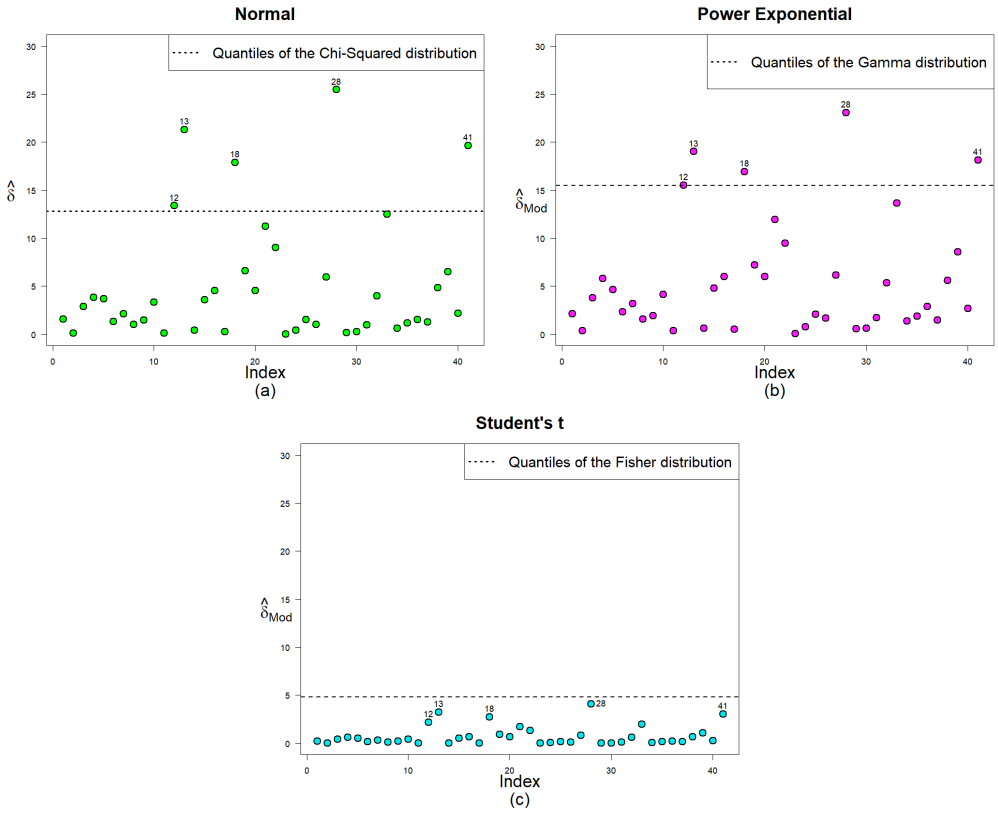


Figure 5.3: Mahalanobis distance (a), modified Mahalanobis distance (b) and (c), under different elliptical distributions.

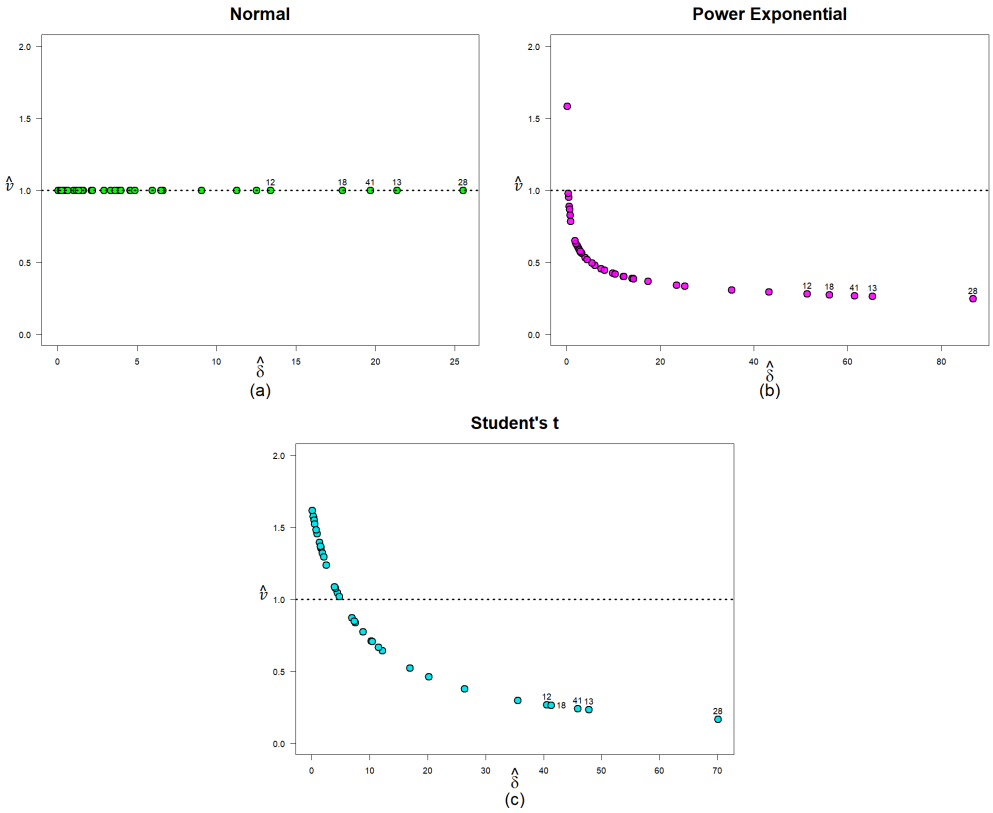


Figure 5.4: Scatter plot between the corresponding Mahalanobis distance and weights under different elliptical distributions.

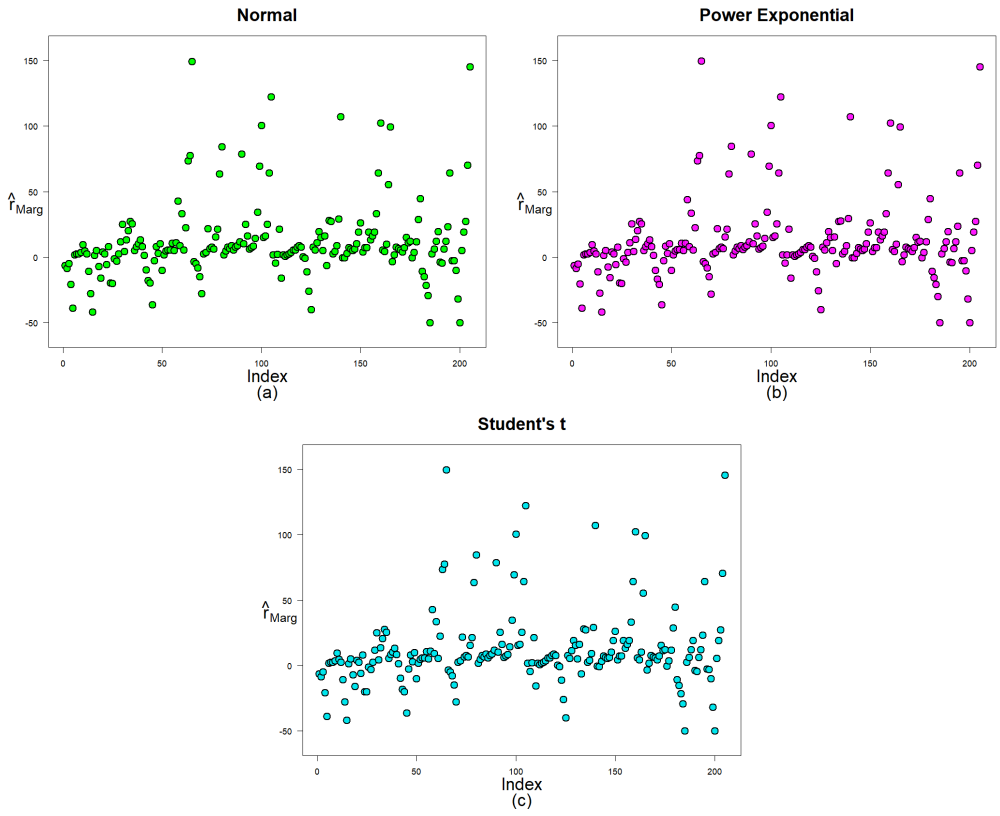


Figure 5.5: Index plot of marginal residuals for the fitted model under errors with different elliptical distributions.

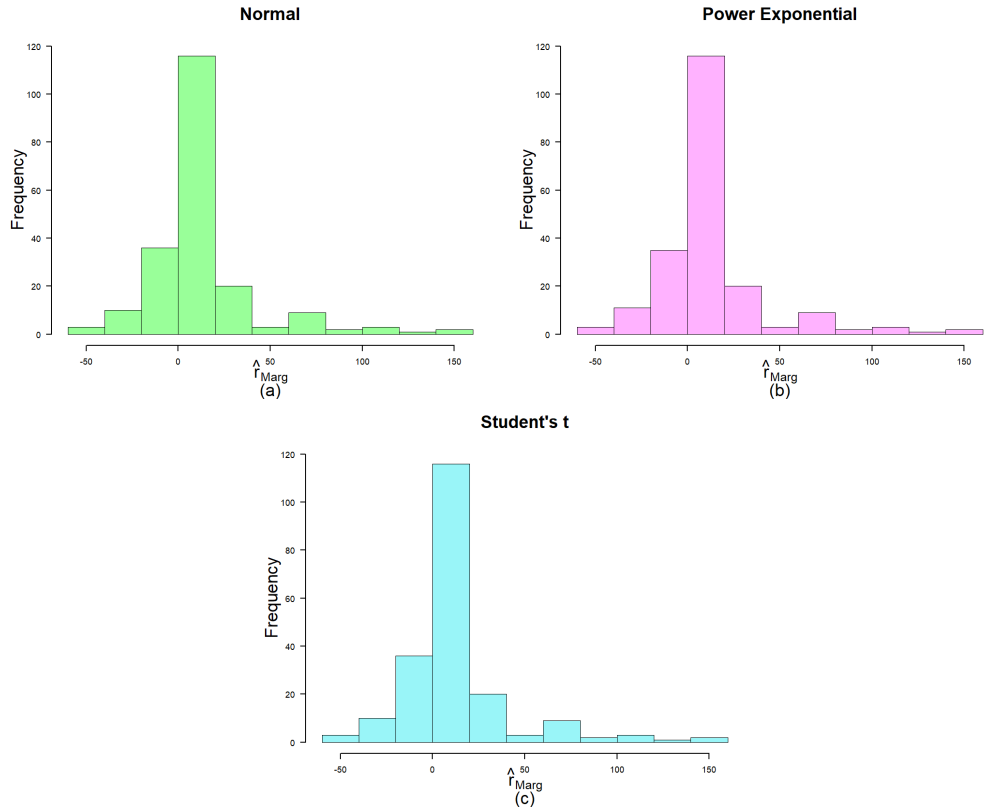


Figure 5.6: Histogram of marginal residuals for the fitted model under errors with different elliptical distributions.

Additionally, the estimated values for the smoothing parameters are selected through the data in such a way that a quantity equal to 5 degrees of freedom is obtained in the nonparametric functions considered.

Figure 5.1 shows the estimates of the smooth function β_1 with their respective standard errors under the different elliptical models. In this figure it can be observed how the standard errors are slightly smaller when adjusting the data considering the Student's t model compared to the Normal and Power Exponential. However, idependent of the elliptical model considered, the estimates reproduce well the trend in the VSAE scores of the female profiles and the dispersion by age is correctly modeled, since the bands defined by the standard errors are larger as age increases.

Figure 5.2 shows the estimates of the smooth function β_2 with their respective standard errors

under the different elliptical models. As in the previous case, the estimates present slightly smaller standard errors when adjusting the data under the Student's t model. In addition, the trend in the VSAE scores of children diagnosed with PDD is well reproduced and the dispersion by age is correctly modeled.

Mahalanobis distances are shown in Figure 5.3, however, for the Power Exponential and Student's t models the distances were modified as previously indicated. Figure 5.3a shows that the distances associated with profiles $\{12, 13, 18, 28, 41\}$ are detected as significant under the Normal and Power Exponential model, while under the Student's t model the same profiles are not detected as significant, This is most likely due to the fact that the Student's t distribution, with $\nu = 8$ degrees of freedom, has much heavier tails than the Normal distribution, and even heavier than the Power Exponential distribution, allowing us to consider those values that are very atypical under those models.

Figure 5.4 presents a scatter plot between the weights v_i and the Mahalanobis distances δ_i, in which it can be observed that those profiles detected as atypical, under the different elliptical models, are assigned a lower weight under the Power Exponential and Student's t models compared to the Normal case, where all profiles have the same weight, demonstrating the robustness of the estimation process under distributions with heavier tails than the Normal one. Theoretically, when considering a distribution with heavier tails than Normal, the Mahalanobis distances should be smaller because the outliers are accommodated in a better way probabilistically speaking, that is, their probability of occurrence is greater, however, when estimating these distances under the Power Exponential and Student's t models, these are affected by the estimates of the weights, since the most atypical values are assigned a lower weight within the estimation process and therefore the estimate of the Mahalanobis distance is greater than in the Normal case, this is because between v_i and δ_i there is an inverse relationship.

Figures 5.5 and 5.6 show an index diagram as well as a histogram of the marginal residuals, described in equation 4.4, for the PVCMM adjusted under errors with different elliptical distributions. In this case, it can be observed that the residuals are around 0. However, they present a rather particular behavior or pattern, in addition to having a positively asymmetric distribution under all elliptical models considered. This is because the response variable is conditioned to the *i*-th random effects vector and not considering it causes this rather particular behavior. Therefore, in order to perform a more precise analysis, it is necessary to consider the conditional residuals described in equation 4.5, which are shown in Figures 5.7 and 5.8 through an index diagram and histogram.

In it, it is possible to observe how the previously identified pattern disappears, since the behav-

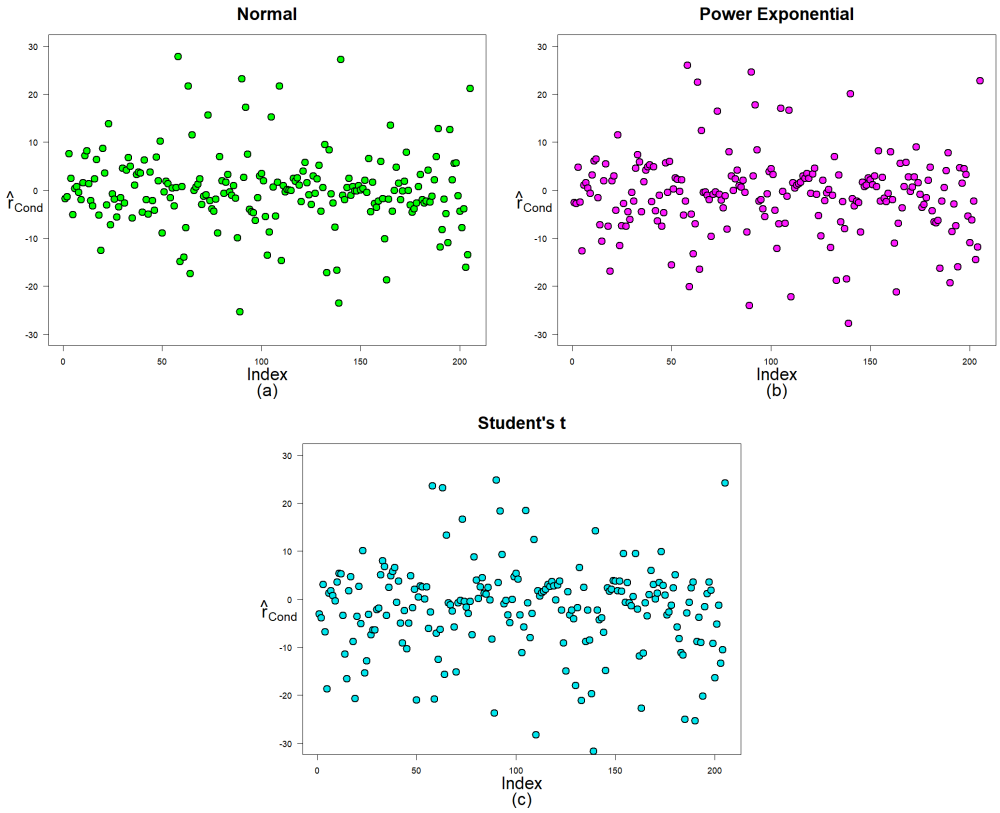


Figure 5.7: Index plot of conditional residuals for the fitted model under errors with different elliptical distributions.

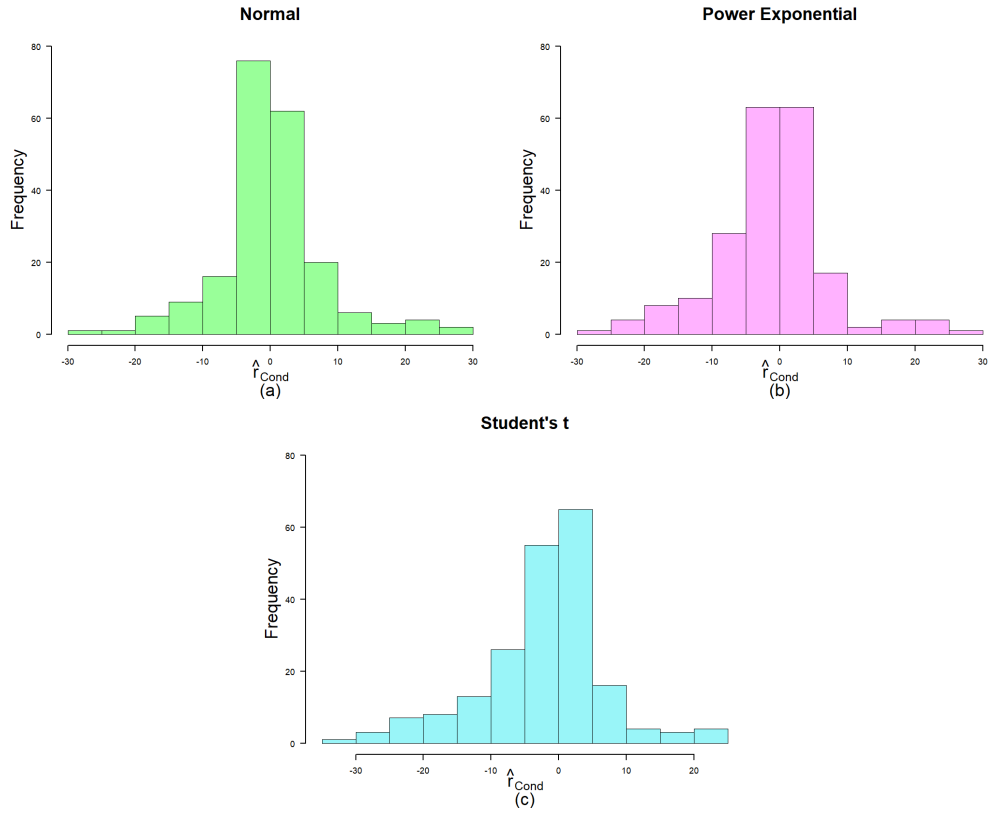


Figure 5.8: Histogram of conditional residuals for the fitted model under errors with different elliptical distributions.

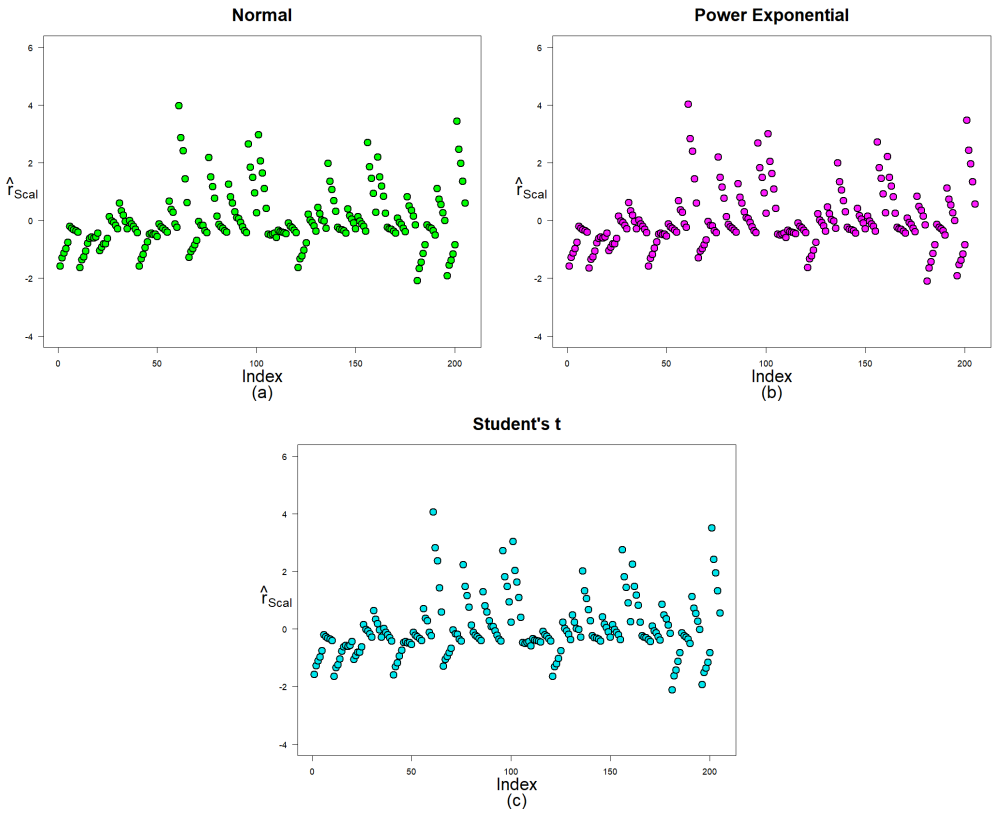


Figure 5.9: Index plot of scaled residuals for the fitted model under errors with different elliptical distributions.

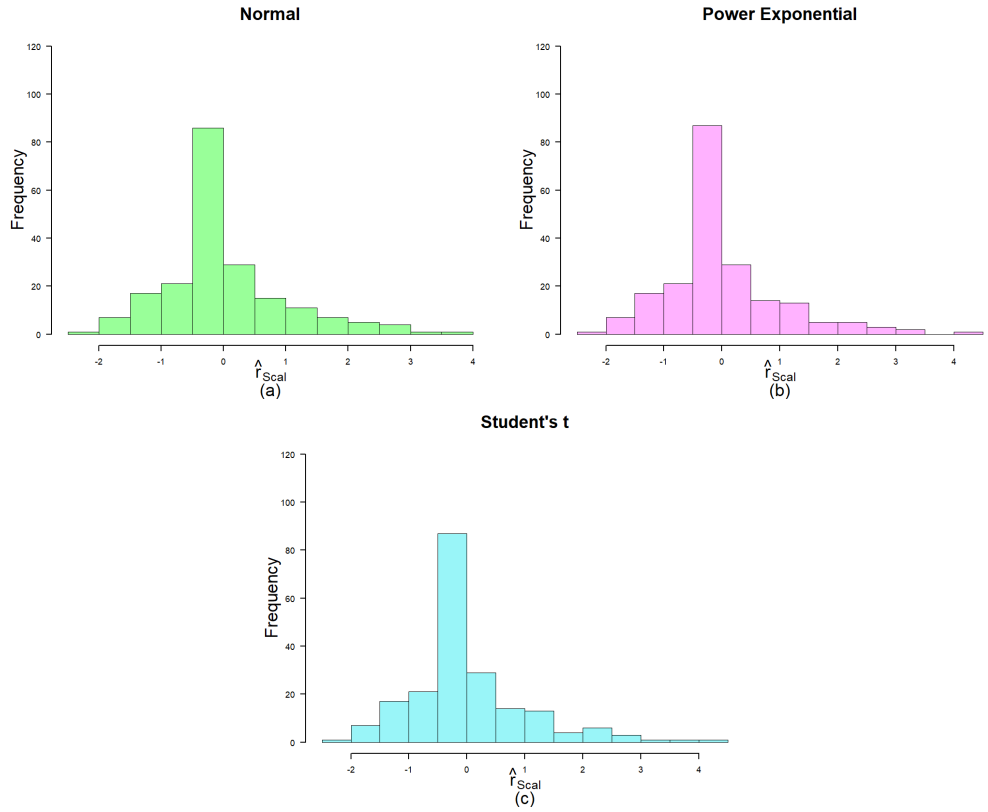


Figure 5.10: Histogram of scaled residuals for the fitted model under errors with different elliptical distributions.

ior of the residuals is more uniform throughout the observations. In addition, in Figure 5.8, it is observed that under all elliptical models, the degree of symmetry improves considerably. Additionally, the variance of the residuals shown in Figure 5.7 is considerably more uniform compared to the behavior observed in the marginal residuals.

Figures 5.9 and 5.10 show an index diagram as well as a histogram of the scaled residuals, described in equation 4.6, for the PVCMM adjusted under errors with different elliptical distributions. In this type of residuals, an evident structure can be presented by group, that is, patterns repeated over time, particularly every five observations. This is due to several reasons: the Σ_i matrix imposes a temporal correlation, the amount of data over time is small for each profile involved, and the distribution in question is elliptical and not Normal. Regardless of the shape, Figure 5.9 shows that the residuals are around 0 and the groups of five observations are not very dispersed, main-

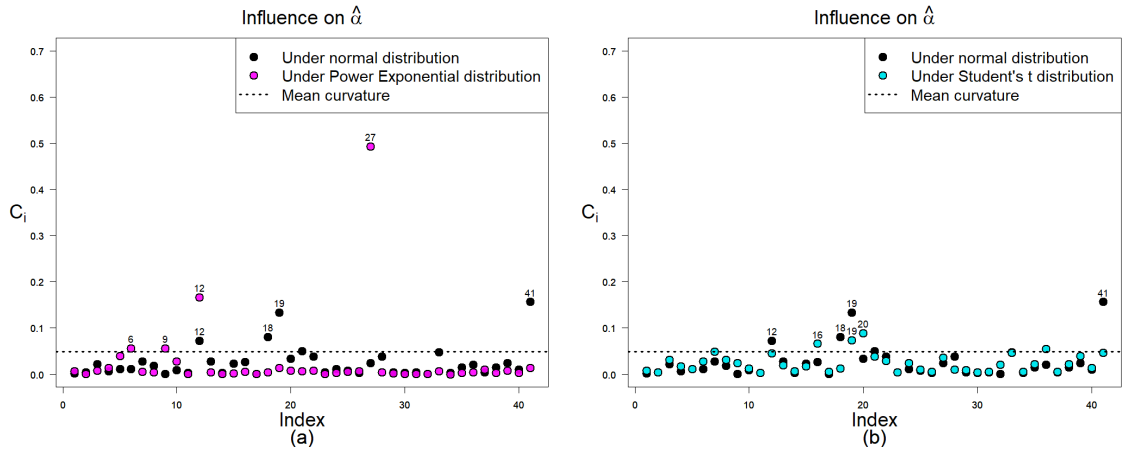


Figure 5.11: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\boldsymbol{\alpha}}$, under the case-weight perturbation scheme, when fitting different elliptical distributions to the autism data set.

taining overall uniformity, except for some groups that most likely correspond to those profiles or groups detected as atypical. This behavior is similar under the different elliptical models. Figure 5.10 shows similar behavior among models; in particular, the tails of the distribution appear to be heavy, and a slight positive asymmetry is observed.

Now, in order to identify potentially influential profiles under the different elliptical models fitted to the autism data, the local influence method proposed by Cook (1986) will be applied. Figures 5.11, 5.12, 5.13, 5.14, 5.15 and 5.16 show a index plot of the curvatures C_ℓ defined in Subsection 4.2, which allows detecting influential profiles in the parameter estimates under the case-weight perturbation scheme.

In Figure 5.11 it can be observed that the profiles $\{12, 18, 19, 41\}$ detected as significantly influential in the estimation of $\boldsymbol{\alpha}$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profile $\{12\}$, which is detected as more influential in the Power Exponential model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of $\boldsymbol{\alpha}$.

In Figure 5.12 it can be observed that the profiles $\{12, 27, 28\}$ detected as significantly influential in the estimation of $\boldsymbol{\beta}_1$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profile $\{27\}$, which is detected as

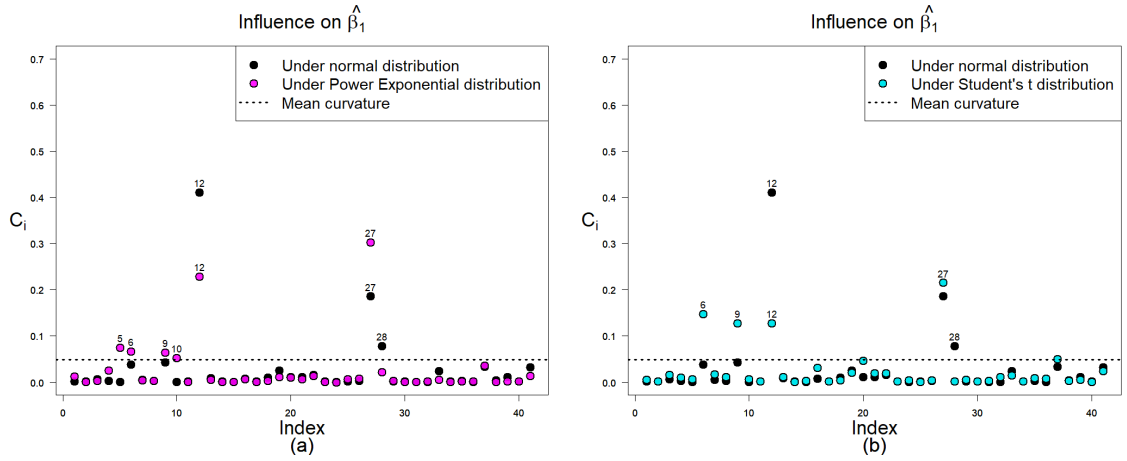


Figure 5.12: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\boldsymbol{\beta}}_1$, under the case-weight perturbation scheme, when fitting different elliptical distributions to the autism data set.

more influential in the elliptical models, furthermore, profile $\{12, 27\}$ were detected as the most influential in the estimation of $\boldsymbol{\beta}_1$.

In Figure 5.13 it can be observed that the profiles $\{4, 5, 10, 22, 28, 33, 39\}$ detected as significantly influential in the estimation of $\boldsymbol{\beta}_2$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profiles $\{4, 10\}$, which are detected as more influential in the Stundent's t model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of $\boldsymbol{\beta}_2$.

In Figure 5.14 it can be observed that the profiles $\{12, 13, 18, 28, 41\}$ detected as significantly influential in the estimation of ϕ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profiles $\{12\}$, which is detected as more influential in the Power Exponential model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of ϕ.

In Figure 5.15 it can be observed that the profiles $\{12, 19, 22, 28, 41\}$ detected as significantly influential in the estimation of γ_1 under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profile $\{12\}$, which is detected as more influential in the Power Exponential model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of γ_1.

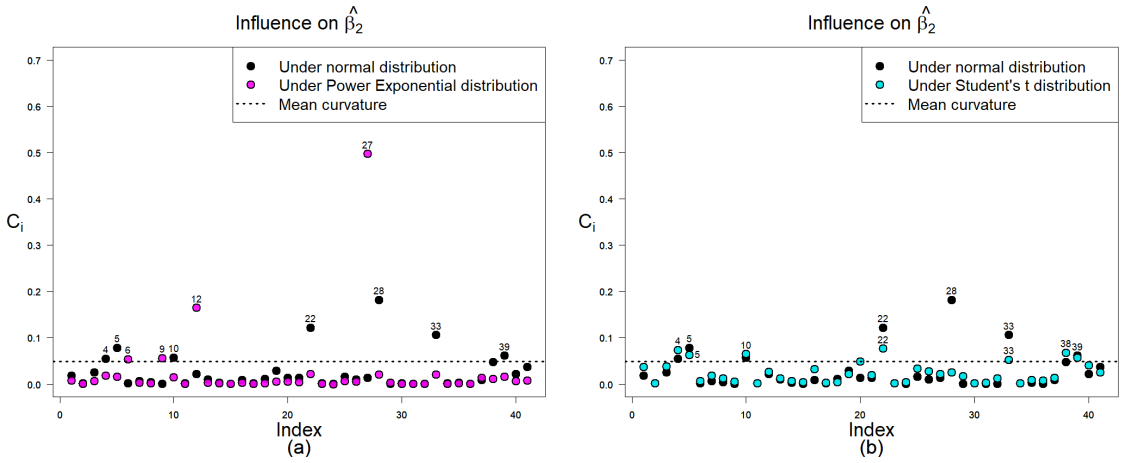


Figure 5.13: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\boldsymbol{\beta}}_2$, under the case-weight perturbation scheme, when fitting different elliptical distributions to the autism data set.

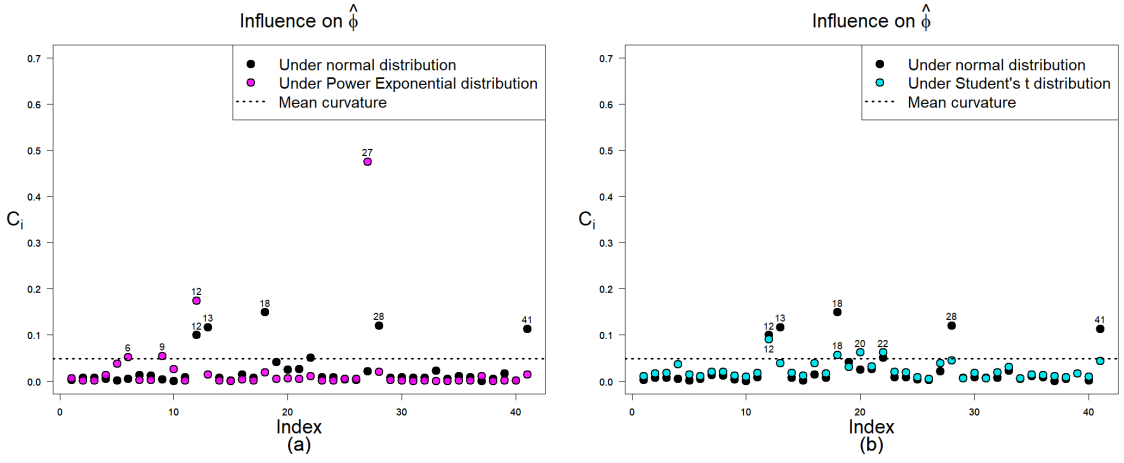


Figure 5.14: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\phi}$, under the case-weight perturbation scheme, when fitting different elliptical distributions to the autism data set.

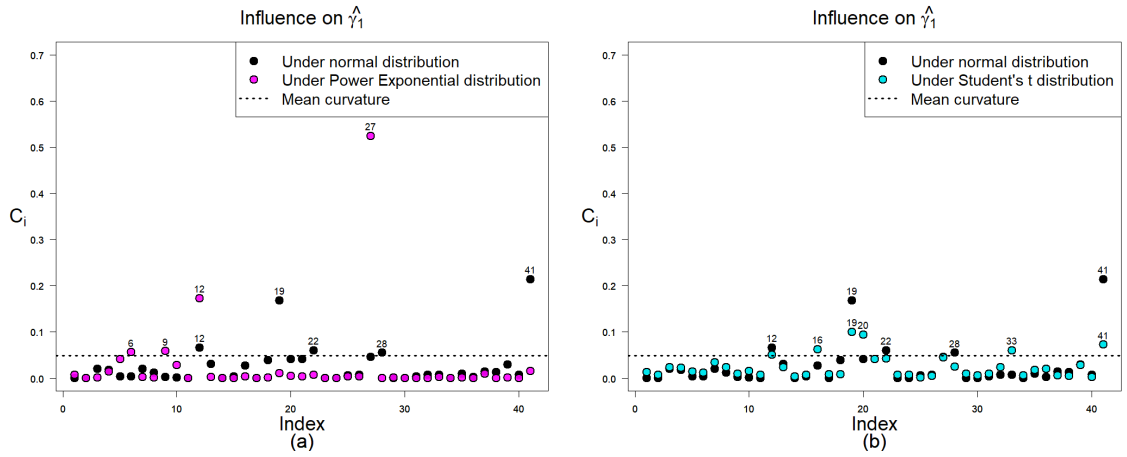


Figure 5.15: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\gamma}_1$, under the case-weight perturbation scheme, when fitting different elliptical distributions to the autism data set.

In Figure 5.16 it can be observed that the profiles $\{13, 19, 21, 28, 41\}$ detected as significantly influential in the estimation of $\boldsymbol{\gamma}_2$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of $\boldsymbol{\gamma}_2$.

Regarding the perturbation scheme of the scale parameter, Figures 5.17, 5.18, 5.19, 5.20, 5.21 and 5.22 are presented, in these the curvatures C_ℓ defined in Subsection 4.2 are shown, which allow to identify influential profiles in the parameter estimates when making small perturbations on the scale parameter.

In Figure 5.17 it can be observed that the profiles $\{12, 18, 19, 22, 28, 41\}$ detected as significantly influential in the estimation of $\boldsymbol{\alpha}$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profile $\{12\}$, which is detected as more influential in the Power Exponential model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of $\boldsymbol{\alpha}$.

In Figure 5.18 it can be observed that the profiles $\{12, 27, 28\}$ detected as significantly influential in the estimation of $\boldsymbol{\beta}_1$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profile $\{27\}$, which is detected as

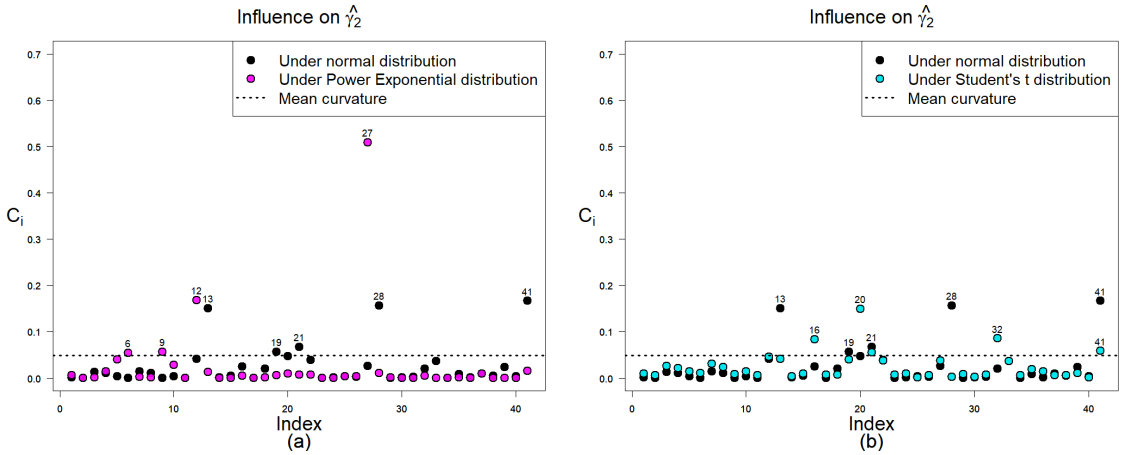


Figure 5.16: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\gamma}_2$, under the case-weight perturbation scheme, when fitting different elliptical distributions to the autism data set.

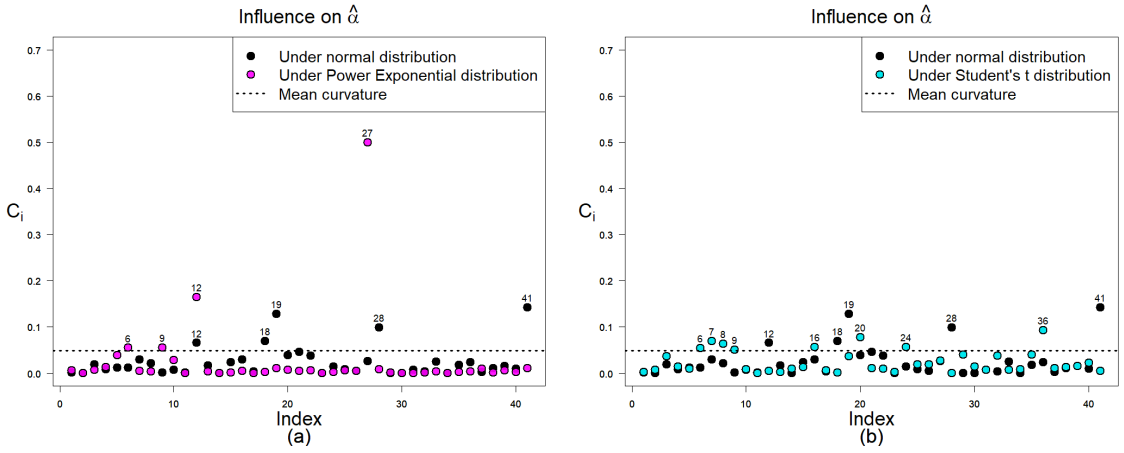


Figure 5.17: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\boldsymbol{\alpha}}$, under the scale perturbation scheme, when fitting different elliptic distributions to the autism data set.

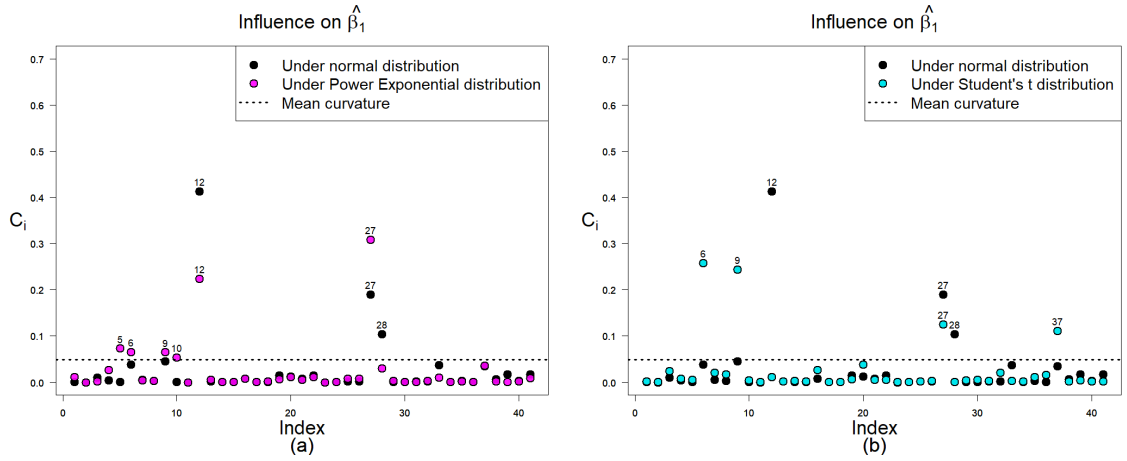


Figure 5.18: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\boldsymbol{\beta}}_1$, under the scale perturbation scheme, when fitting different elliptic distributions to the autism data set.

more influential in the Power Exponential model, furthermore, profile $\{12, 27\}$ were detected as the most influential in the estimation of $\boldsymbol{\beta}_1$.

In Figure 5.19 it can be observed that the profiles $\{4, 5, 10, 22, 28, 33, 38, 39\}$ detected as significantly influential in the estimation of $\boldsymbol{\beta}_2$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profile $\{10, 38\}$, which are detected as more influential in the Student's t model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of $\boldsymbol{\beta}_2$.

In Figure 5.20 it can be observed that the profiles $\{12, 13, 18, 22, 28, 41\}$ detected as significantly influential in the estimation of $\boldsymbol{\phi}$ under the Normal model, are no longer detected under the Power exponential and Student's t models, with the exception of the profile $\{12\}$, which is detected as more influential in the Power Exponential model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of $\boldsymbol{\phi}$.

In Figure 5.21 it can be observed that the profiles $\{12, 19, 22, 41\}$ detected as significantly influential in the estimation of $\boldsymbol{\gamma}_1$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profile $\{12\}$, which is detected as more influential in the Power Exponential model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of $\boldsymbol{\gamma}_1$.

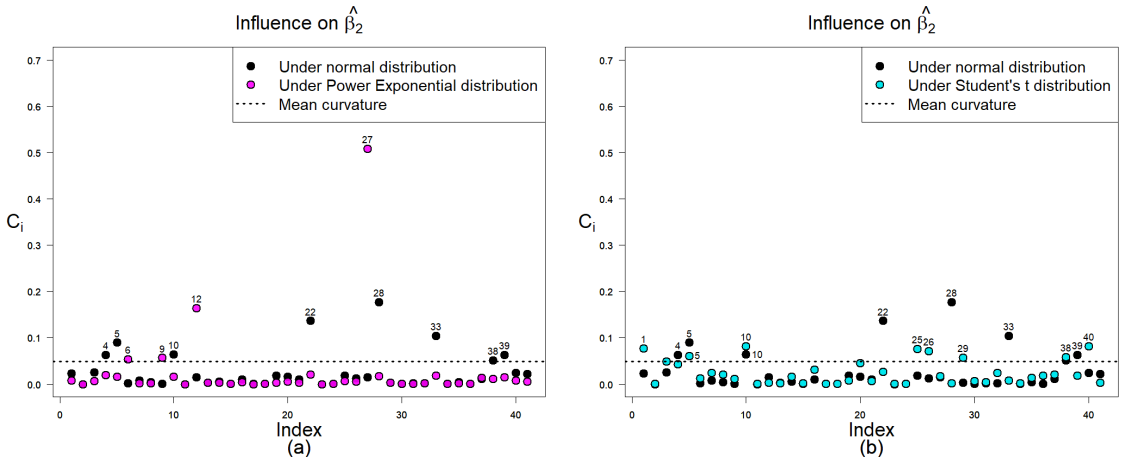


Figure 5.19: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\boldsymbol{\beta}}_2$, under the scale perturbation scheme, when fitting different elliptic distributions to the autism data set.

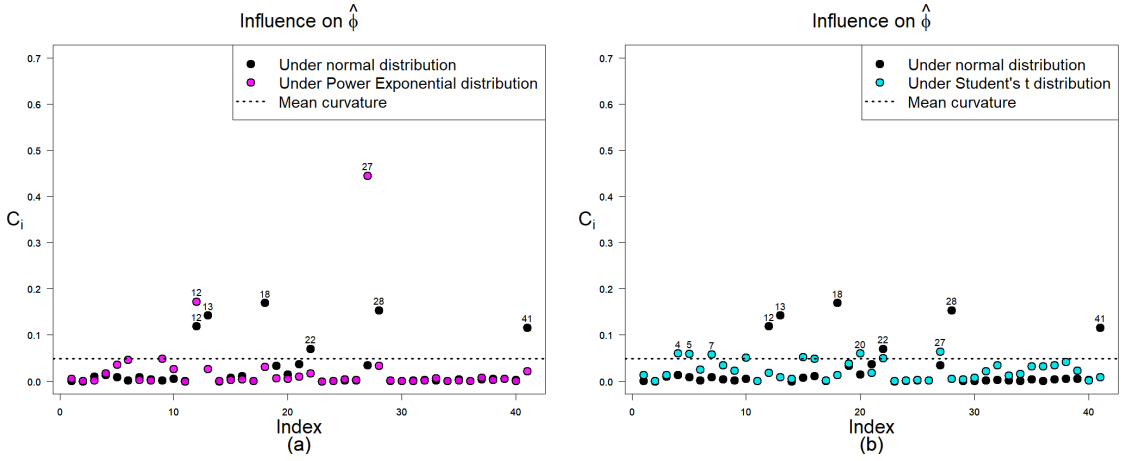


Figure 5.20: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\phi}$, under the scale perturbation scheme, when fitting different elliptic distributions to the autism data set.

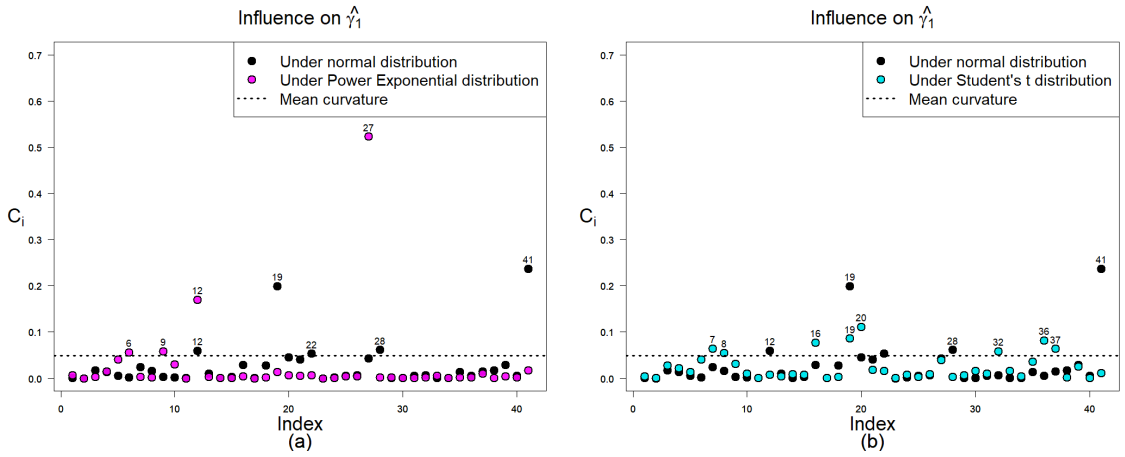


Figure 5.21: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\gamma}_1$, under the scale perturbation scheme, when fitting different elliptic distributions to the autism data set.

In Figure 5.22 it can be observed that the profiles $\{13, 20, 21, 28, 41\}$ detected as significantly influential in the estimation of $\boldsymbol{\gamma}_2$ under the Normal model, are no longer detected under the Power Exponential and Student's t models, with the exception of the profile $\{20\}$, which is detected as more influential in the Student's t model, furthermore, profile $\{27\}$ was detected as the most influential in the estimation of $\boldsymbol{\gamma}_2$.

It is important to highlight that, under the different perturbation schemes, the curvatures C_ℓ tend to be smaller when the adjustment is performed using the Power Exponential and Student's t models, although some exceptions were previously noted. Additionally, there are profiles that appear as influential exclusively under the heavier-tailed models. Some profiles identified as atypical by the Mahalanobis distance are also potentially influential in the estimation of $\boldsymbol{\theta}$. In particular, profile $\{27\}$ was detected as the most influential for the estimation of all parameters, except for $\boldsymbol{\beta}_1$, for which profile $\{12\}$ had slightly greater influence. On the other hand, profile $\{28\}$ exhibited less influence compared to the other two profiles, but it was the profile most frequently identified as influential. Therefore, those profiles highlighted, under the different probability models, will be studied in greater detail to know their characteristics and subsequently perform a confirmatory analysis on the influence of the most frequently detected profile on the parameter estimates. The characteristics of the profiles detected as influential, under the different perturbation schemes, are presented in Table 5.2. The highlighted profiles are those that were detected as atypical by the

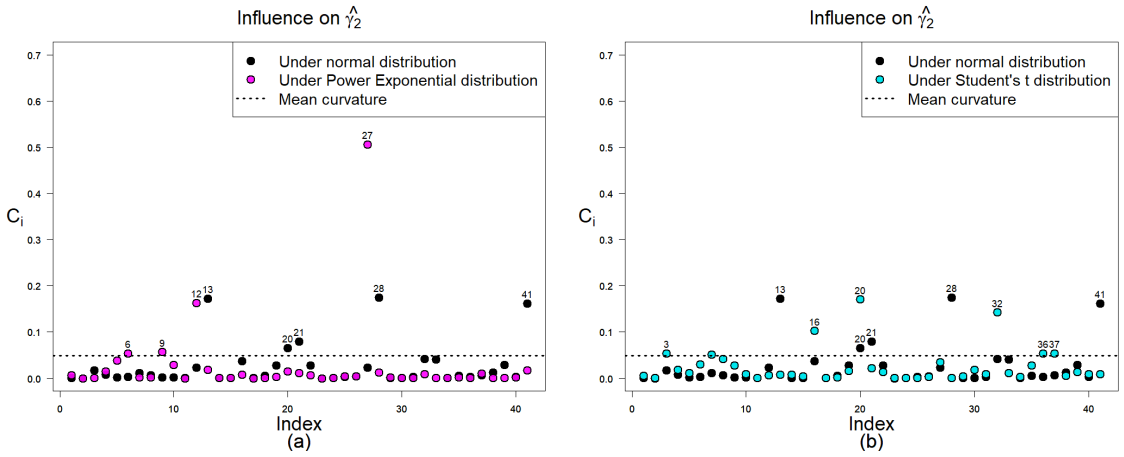


Figure 5.22: Index plot of the C_i curvatures to assess the local influence on the estimation of $\hat{\gamma}_2$, under the scale perturbation scheme, when fitting different elliptic distributions to the autism data set.

Mahalanobis distance.

In Table 5.2 it can be observed that the majority of the highlighted profiles correspond to the male gender whose diagnosed condition is autism, however, the diagnosed condition for profile $\{28\}$ is PDD and profile $\{12\}$ corresponds to the female gender. Additionally, the expressive language development is medium or high in children, except for profiles $\{9, 18\}$ in which it is low. Regarding profiles $\{12, 27\}$, it can be observed that the gender is female, the expressive language development score is high and the diagnosed condition is autism, which could mean that having that particular gender and score is a determining factor in the VSAE score obtained by the child.

Profiles $\{12, 27, 28\}$ were detected with the greatest influence on the parameter estimates when fitting the model under errors with heavier tailed distributions. Figure 5.23 shows the VSAE scores of those profiles detected. It can be observed that profile $\{28\}$ shows regular behavior considering the general trend of the scores, however, the score increases considerably from 9 years to 13 years. Profile $\{27\}$ shows a rather irregular behavior in the scores since at the ages of 2, 3 and 5 years it increases and then decreases, however, from the age of 5 years the score increases regularly. Profile $\{12\}$ shows a very pronounced behavior, as the score increases considerably at age 5, then decreases significantly at age 9, and then increases again to a high level. However, this profile seems to have a slightly more uniform behavior compared to the other two cases, where the score increases progressively with age.

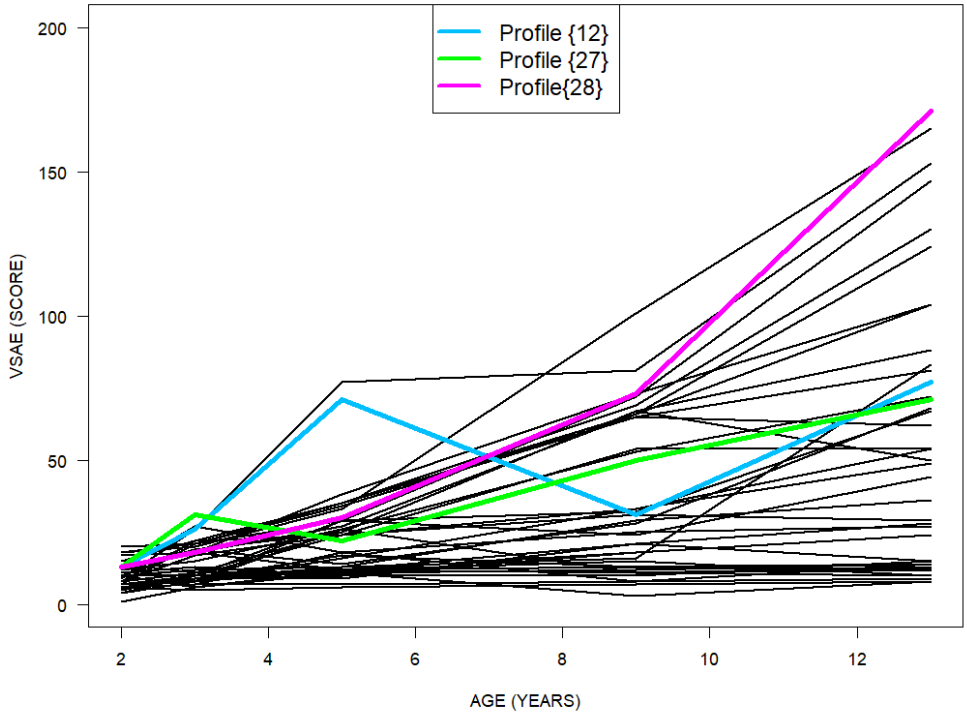


Figure 5.23: Individual VSAE score for profiles detected as potentially influential.

Profile ID	VSAE score by age					Expressive development	Gender	Diagnosed condition
5	13	8	29	24	44	Medium	Male	PDD
6	5	10	29	32	67	Medium	Female	Autism
9	10	6	11	3	8	Low	Female	Autism
12	13	26	71	31	77	High	Female	Autism
13	9	26	77	81	153	High	Male	Autism
18	10	12	13	16	83	Low	Male	Autism
19	12	27	18	8	9	Medium	Male	Autism
20	12	18	38	73	104	High	Male	Autism
22	14	11	25	67	50	High	Male	PDD
27	13	31	22	50	71	High	Female	Autism
28	13	18	30	73	171	Medium	Male	PDD
33	18	20	33	101	165	High	Male	PDD
41	7	21	29	72	147	Medium	Male	Autism

Table 5.2: Characteristics of the profiles detected as potentially influential under the different perturbation schemes and elliptical models (highlighted profiles correspond to those detected by the Mahalanobis Distance under the Normal model)

A confirmatory analysis was then performed to verify the influence of profile $\{27\}$ on the estimate of $\boldsymbol{\theta}$ when fitting the Power Exponential model. Table 5.3 reports the relative changes (RC) in the parameter estimates and their respective standard errors upon removing profile $\{27\}$ from the autism data set. From the table, it can be observed that, when this profile is removed and the model is fitted under the Power Exponential error distribution, the largest relative change occurs in the estimate of α_2, reaching 61%. Additionally, the estimate of β_1 also exhibits a substantial change of approximately 91%, and the corresponding standard errors show considerable variation compared to the rest. In contrast, the estimate of β_2 and its standard errors are affected to a much lesser extent. Among the dispersion parameters, the greatest change occurs in the estimate of γ_1. As shown in Table 5.2, profile $\{27\}$ corresponds to a female profile diagnosed with autism, whose score in expressive language development is high. This explains why, upon removing this profile, the relative changes are most pronounced in the estimates of α_2 and β_1, while changes in β_2 are relatively small. Specifically, α_2 is the coefficient associated with the artificial variable defined in terms of high expressive language scores, β_1 is the varying coefficient corresponding to the artificial variable defined by female gender, and β_2 is the varying coefficient associated with the artificial variable defined by the PDD diagnosis.

	PE ($\rho = 0.75$)	
Parameter	Estimate	Standard error
α_0	0.54	2.19
α_1	7.26	2.11
α_2	61.96	4.56
$\beta_1(t_1^0)$	91.25	11.86
$\beta_1(t_2^0)$	30.47	12.48
$\beta_1(t_3^0)$	1.59	13.24
$\beta_1(t_4^0)$	37.48	13.70
$\beta_1(t_5^0)$	18.08	13.83
$\beta_2(t_1^0)$	6.08	2.57
$\beta_2(t_2^0)$	4.04	2.66
$\beta_2(t_3^0)$	1.63	2.19
$\beta_2(t_4^0)$	1.89	2.01
$\beta_2(t_5^0)$	1.43	1.96
ϕ	4.17	2.89
γ_1	16.44	4.59
γ_2	3.60	4.78

Table 5.3: Relative change (%), after removing profile $\{27\}$, in MPLEs and standard errors when fitting Power Exponential model to autism data set.

Chapter 6

Concluding remarks

In this degree work, an iterative process was developed to obtain the MPLEs of the parameters, some aspects of statistical inference were reviewed, and an analysis of different types of residuals was carried out. Finally, the local influence method proposed by Cook (1986) was applied in the PVCMM under different perturbation schemes to identify potentially influential observations. The distributional assumption was made more flexible, allowing to consider elliptical distributions with different kurtosis coefficients, and the scale matrix $\boldsymbol{\Sigma}_i$ was assumed with a defined structure, particularly a structure that depends on the covariance matrix associated with the errors and random effects of the model. A real data set related to autism was studied using the entire methodology, first the analysis was performed considering errors with Normal distribution, where the presence of atypical values was detected, then it was analyzed under errors with Power Exponential, Student's t distribution, this with the purpose of having a more robust estimation process in the presence of atypical observations. The results showed that the estimates were similar under the different elliptical distributions, however, the standard errors were lower under the probability models with heavier tails than the Normal, including the AIC and the penalized log-likelihood function. The weights were estimated in a more robust way under the models with heavier tails, that is, less weight was assigned in the estimation process to those profiles that were detected as more atypical compared to the rest. The residual analysis showed that under the conditional residuals the model is better diagnosed, this is because the i-th response vector is conditioned to the i-th random effects vector. In general, there is no major violation of the classic assumptions, even the variance of the conditional residuals is quite uniform even though in the descriptive analysis a non-constant variance can be observed in the VSAE scores as the age of the children increases, this modeling of a non-constant variance is thanks to the structure given to the $\boldsymbol{\Sigma}_i$ matrix, allowing it to be different for each response vector. The local influence analysis showed that under the probability

models with heavier tails the profiles tend to be less influential independent of the perturbation scheme, except for some that are detected as more influential, in addition these profiles detected as influential showed a more irregular than atypical behavior, except for profile number $\{28\}$, which apart from being influential is quite atypical, even being detected by the Mahalanobis distance. It was also shown that when removing profile $\{27\}$ and performing the MPLEs of the parameters, relatively high changes occur, confirming its power of influence in the estimates, however, it is not highly influential in the estimation of all the parameters, but only in those that could somehow be affected given the characteristics observed in this profile.

Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Akademiai Kiado, Budapest.

[2] Atkinson, A. C. (1981). Plotting residuals to assess model adequacy. *Applied Statistics*, **30**(3), 242–261.

[3] Beckman, R. J. (1987). Local influence in generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **49**(3), 293-299.

[4] Berhane, K., & Tibshirani, R. (1998). Generalized additive models for longitudinal data. *Biometrika*, **85**(2), 413–423.

[5] Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, **17**(2), 453–510.

[6] Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1998). Measurement error in nonlinear models. *Journal of the American Statistical Association*, **93**(441), 131–142.

[7] Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **48**(2), 133-148.

[8] Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, **31**(4), 377–403.

[9] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**(1), 1–38.

[10] Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**(2), 89–121.

[11] Eubank, R. L. (1984). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

[12] Eubank, R. L., Huang, C., Muñoz Maldonado, Y., Wang, N., Wang, S., & Buchanan, R. J. (2004). Smoothing spline estimation in varying-coefficient models. *Journal of the Royal Statistical Society B*, **66**(4), 653-667.

[13] Fang, K.-T., Kotz, S., & Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London.

[14] Fan, J., & Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**(5), 1491–1518.

[15] Fung, W.-K., Zhu, Z., Wei, B.-C., & He, X. (2002). Influence diagnostics and outlier tests for semiparametric models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 565-579.

[16] Galea, E., O'Brien, J. R., & McCulloch, S. B. (1977). Some properties of linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **39**(3), 323-330.

[17] Gourieroux, C., & Monfort, A. (1995). *Statistics and Econometric Models*. Cambridge University Press, Cambridge.

[18] Green, P. J. (1987). *Non-linear Models for Longitudinal Data*. Chapman & Hall, London.

[19] Green, P. J., & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London.

[20] Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London. ISBN 0412343908.

[21] Hastie, T., & Tibshirani, R. (1993). Varying-coefficients models. *Journal of the Royal Statistical Society B*, **55**(4), 757-796.

[22] Hurvich, C. M., Simonoff, J. S., & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B*, **60**(2), 271–293.

[23] Ibacache-Pulgar, G., Paula, G., & Cysneiros, F. (2013). Semiparametric additive models under symmetric distributions. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, **22**(1), 103–121.

[24] Kano, Y., Berkane, M., & Bentler, P. M. (1993). Statistical inference based on pseudo maximum likelihood estimators in elliptical populations. *Journal of the American Statistical Association*, **88**(423), 135–143.

[25] Kim, C., Park, B. U., & Kim, W. (2002). Influence diagnostics in semiparametric regression models. *Journal of Statistical Planning and Inference*, **60**(1), 49-58.

[26] Laird, N. M., & Ware, J. M. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**(4), 963-974.

[27] Lange, K. L., Little, R. J., & Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, **84**(408), 881-896.

[28] Lesaffre, E., & Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**(2), 570–582.

[29] Li, Z., Xu, W., & Zhu, L. (2009). Influence diagnostics and outlier tests for varying coefficient mixed models. *Journal of Multivariate Analysis*, **100**(10), 2002-2017.

[30] Lindsey, J. K. (1999). *Models for repeated measurements*. Oxford: Oxford University Press.

[31] Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B*, **61**(2), 381–400.

[32] Liu, J. S. (2002). Local influence in linear models. *Journal of Statistical Planning and Inference*, **110**(1), 147-161.

[33] Opsomer, J. D. (1999). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **71**(1), 45–75.

[34] Oti, R., Anderson, D., Risi, S., Pickles, A., & Lord, C. (2006). Social trajectories among individuals with autism spectrum disorders [Data set]. University of Michigan.

[35] Pinheiro, J. C., Liu, C., & Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, **10**(2), 249–276.

[36] Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(3), 507–554.

[37] Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

[38] Savalli, C., Paula, G., & Cysneiros, F. (2006). Assessment of variance components in elliptical linear mixed models. *Statistical Modelling*, **6**(1), 59-76.

[39] Schabenberger, O. (2004). Mixed Model Influence Diagnostics. Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference, Paper 189-29, SAS Institute Inc., Cary, NC.

[40] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

[41] Segal, M. R., Bacchetti, P., & Jewell, N. P. (1994). Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**(2), 345–352.

[42] Silverman, B. W. (1985). *Some aspects of the spline smoothing approach to non-parametric regression curve fitting*. Journal of the Royal Statistical Society: Series B, **47**(1), 1–52.

[43] Simonoff, J. S., & Tsai, C.-L. (1999). Smoothing parameter selection in nonparametric regression. *Computational Statistics & Data Analysis*, **31**(2), 211–227.

[44] Staudenmayer, J., Lake, E., & Wand, M. (2009). Robustness for general design mixed models using the t-distribution. *Centre for Statistical & Survey Methodology Working Paper Series*, **9**.

[45] Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Series B*, **45**(1), 133–150.

[46] Wei, B. C., Hu, Y. Q., & Fung, W. K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, **25**(1), 25–37.

[47] Wu, H., & Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. Wiley, New York.

[48] Zhang, D. (1998). Inference in semiparametric mixed-effects models. *Biometrika*, **85**(1), 1–16.

[49] Zhang, W., Lee, S. Y., & Song, X. (2007). Local polynomial fitting in semivarying coefficient mixed-effects models. *Journal of Multivariate Analysis*, **98**(6), 1140–1156.

[50] Zhu, Z., Fung, W.-K., & Wei, B.-C. (2003). Measures of local influence in partial linear models and semiparametric mixed models. *Journal of Statistical Planning and Inference*, **113**(2), 319-335.

[51] Zhu, H., & Wei, B. (2007). Diagnostic analysis for semiparametric models. *Journal of Statistical Planning and Inference*, **137**(6), 1900–1915.