Universidad de Valparaíso
CHILE
Instituto de Estadística
Facultad de Ciencias

Doctorate in Statistics

Thin-plate spline varying-coefficient mixed model under elliptical distributions: estimation and influence diagnostic.

by

MAGALY SOFÍA MORAGA CÁRDENAS

Doctoral thesis

ADVISOR: **Dr. Germán Ibacache Pulgar**
CO-ADVISOR: **Dra. Orietta Nicolis**

Valparaíso
2018

Magaly Sofía Moraga Cárdenas

Thin-plate spline varying coefficient mixed model under elliptical distributions: estimation and influence diagnostic.

Doctoral thesis submitted to the Graduate Program in Statistics, Institute of Statistics, Universidad de Valparaíso, as a requirement to obtain a PhD in Statistics.

During the development of this work, the author received financial support from Universidad Austral de Chile, Valdivia.

Valparaíso
2018

"To Kary and Joaco"

Acknowledgements

To my advisors, Dr. Germán Ibacache Pulgar and Dra. Orietta Nicolis, for their constant guidance, dedication, patience and good advises, both academically and personally.

To the Professors of the PhD program in Statistics of the Institute of Statistics of the Universidad de Valparaíso, for their contributions in my doctoral training.

Finally, thank the Universidad Austral de Chile for support financial in the period of my PhD studies, without whose support I would not have been able to carry out these studies.

Abstract

The main idea of this work is to propose a semiparametric model under the hypothesis of elliptical distributions, called thin-plate spline varying-coefficients mixed model (TPSVCMM). The model can be considered an extension of the elliptical semiparametric model by incorporating a non-parametric dynamic characteristic to the regression coefficients and a thin-plate spline smoothing function. Also, we will develop the appropriate statistical inference based on penalized likelihood function to obtain robust estimates in the sense of the Mahalanobis distance. In addition, we will present a process of successive maximizations and a back-fitting algorithm to attain the estimates by using smoothing splines. Then, we will propose local influence curvatures and diagnostic graphics in order to study the sensitivity of the penalized estimates under some usual perturbation schemes in the model or data. Finally, a discussions on degrees of freedom an smoothing parameter estimation will be presented.

Keywords: Nonparametric models, Maximum penalized likelihood estimates, Robust estimates, Sensitivity analysis.

Resumen

La idea principal de este trabajo es proponer un modelo semiparamétrico bajo la hipótesis de distribuciones elípticas, denominado *thin-plate spline varying coefficients mixed model* (TPSVCMM). El modelo puede considerarse una extensión del modelo semiparamétrico elíptico al incorporar una característica dinámica no paramétrica a los coeficientes de regresión y una función de suavizado de thin-plate spline. Además, desarrollaremos la inferencia estadística apropiada basada en la función de verosimilitud penalizada para obtener estimaciones robustas en el sentido de la distancia de Mahalanobis. Además, presentaremos un proceso de maximizaciones sucesivas y un algoritmo de ajuste posterior para alcanzar las estimaciones mediante el uso de splines suavizados. Luego, propondremos curvaturas de influencia local y gráficos de diagnóstico para estudiar la sensibilidad de las estimaciones penalizadas bajo algunos esquemas de perturbación habituales en el modelo o los datos. Finalmente, se presentarán discusiones sobre grados de libertad y una estimación de parámetros de suavizado.

Palabras claves: Modelos no-paramétricos, estimación máxima verosimilitud penalizada, estimación robusta, análisis de sensibilidad.

Contents

List of Acronyms

AIC	Akaike Information Criterion.
BIC	Bayesian Information Criterion.
CVM	Varying-coefficient model.
df	degrees of freedom.
EM	
GCV	Generalized cross-validation.
LM	linear model.
MPLE	Maximum penalized likelihood estimate.
P-spline	Penalized splines.
SAMs	Semiparametric additive models.
SIAM	Society for Industrial and Applied Mathematics.
TPSPVCM	Thin-plate spline partially varying-coefficient model.
TPSVCMM	Thin-plate spline varying coefficient mixed model.

List of Figures

List of Tables

1

Preliminary

1.1 Introduction and bibliographical review

The last decades, there has been an upsurge of interest and effort in nonparametric models as researchers have realized that parametric models are inadequate in capturing the relationship between the response variable and its associated covariates in many practical situations.

Parametric models are, in general, parsimonious and parameters in these models often have meaningful interpretations. On the other hand, based on minimal assumtions about the relationship between variables, nonparametric models are flexible. However, nonparametric models lose the advantage of having interpretable parameters and may suffer from the curse of dimensionality. Combining both parametric and nonparametric components, a semiparametric model can overcome limitations in parametric and nonparametric models while maintaining advantages of having interpretable parameters and flexibility.

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, what includes all symmetric continuous distributions such as normal, Student-t, power exponential, logistics I and II, contaminated normal, among others. In the last years various works on regression models under elliptical errors were developed, including on the class of semiparametric mixed models.

Verbeke and Molenberghs presented in detail an extensive theory for longitudinal data analysis in linear mixed models for the normal case and in particular attention to the problem of missing data. Zhu and Lee [2003] describe a method for assessing model inadequacy in maximum likelihood estimation of a generalized linear mixed model with normal errors, by developing local influence analysis for penalized estimators and proposing a method for detecting influential observations in some perturbation schemes. Savalli et al. [2006] proposed a new class of models called elliptical linear mixed models, where the marginal model is also elliptical. This proposal brings many advantages, for example, development of estimation procedures, diagnostic and testing methodologies for the variance components. An advantage of these models is that when errors have distributions with heavier tails than the normal distribution, the maximum likelihood estimates of parameters involved are more robust in the presence of outliers in the sense of Mahalanobis distances.

Osorio et al. [2007] developed normal curvatures of local influence of disturbance to various schemes for elliptical linear mixed models. Russo et al. [2009] extended the class given by Savalli et al. [2006] replacing linear fixed effects by a nonlinear fixed effect, creating the class of models called partially nonlinear elliptical mixed models, for which estimation procedures and diagnostic methods are developed. Russo et al. [2012] developed for the same class of models test for the variance components through a statistical type score. Furthermore, Russo et al. [2012] proposed a general structure for the variance-covariance matrices of random errors and effects for partially elliptical nonlinear mixed models, including structures autoregressive and heteroscedastic as particular cases. They also developed estimation procedures and diagnostic methodologies.

Ibacache-Pulgar and Paula [2011] presented a study on the existence and uniqueness of the estimated maximum likelihood, in semiparametric t-Student models. Ibacache-Pulgar et al. [2012] developed influence diagnostics for elliptical semiparametric mixed models, where it is assumed that the non- parametric component is of type cubic spline. A back-fitting type procedure is developed for the estimation of the parameters involved and verifies that the estimates, including non-parametric component, are robust against outliers as in the parametric case. Curvature of local influence were developed to some perturbation schemes, such as perturbation for case-weight, perturbation in the scale matrix, in the explanatory variable and the response variable. Some applications were submitted. Ibacache-Pulgar et al. [2013] developed semiparametric additive model under symmetric distributions.

1.2 Background

1.2.1 Elliptical distributions

The class of elliptical distributions are characterized by symmetrical structure as the normal distribution, but with lighter or heavier tails. The normal distribution is a particular case of this class. The development of these class of distributions starts from 1970 and have been extensively studied by several authors, for instance,Kelker [1970], Kai-Tai and Yao-Ting [1990] and Cambanis et al. [1981], among others.

Elliptical distributions share many properties with normal distributions: marginal and conditional distributions are also elliptical, and conditioned means are linear function of the conditioning variables. However the normal distribution has the property that if the covariance matrix is diagonal, all components are independent random variables.

For the purpose of introducing TPSVCMM under elliptical distributions, some definitions and properties of elliptical distributions that will be needed for the development of this work are presented in this section.

Definition 1: Let $\mathbf{y} \sim El_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}$, with $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Sigma} > 0$. Then, the density function of $\mathbf{y}$ is given by (Kai-Tai and Yao-Ting [1990])

$$f(\mathbf{y}) = |\boldsymbol{\Sigma}|^{-1/2} h(u), \tag{1.1}$$

where $u = (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ is the Mahalanobis distance, and $h(\cdot)$ is a function of $\mathbb{R} \to [0, \infty]$ such that

$$\int_0^\infty u^{(-1/2)} h(u) du \quad < \infty.$$

The function $h(\cdot)$ is known as the density generator function. If $\mathbf{y}$ has elliptical density function given by 1.1, we use the notation $\mathbf{y} \sim El_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, h)$.

Property 1: Suppose $\mathbf{y} \sim El_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, h)$ with $rank(\boldsymbol{\Sigma}) = k$. Let $\boldsymbol{A}$ a $(n \times m)$ matrix and $\boldsymbol{\gamma}$ a $(m \times 1)$ vector. Then,

$$\boldsymbol{\gamma} + \boldsymbol{A}^T \mathbf{y} \sim El_m(\boldsymbol{\gamma} + \boldsymbol{A}^T \boldsymbol{\mu}, \boldsymbol{A}^T \boldsymbol{\Sigma} \boldsymbol{A}, h).$$

In addition, it is possible to show that the marginal distribution of a subvector of $\mathbf{y}$ also follows an elliptical distribution. For example, if consider the partition of $\mathbf{y}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, given by

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{y}^{(1)} \in \mathbb{R}^m$, $\boldsymbol{\mu}^{(1)} \in \mathbb{R}^m$ and $\boldsymbol{\Sigma}^{(1)} \in \mathbb{R}^{m\times m}$, then the marginal distribution of $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are, respectively,

(a) $\mathbf{y}^{(1)} \sim El_m(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_{11}, h)$ and,

(b) $\mathbf{y}^{(2)} \sim El_{n-m}(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}, h)$.

That is, a linear transformation of a random vector with an elliptical distribution also follows an elliptical distribution and each element of the random vector has an elliptical marginal distribution.

Property 2: Suppose $\mathbf{y} \sim El_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, h)$ and $\boldsymbol{\Sigma} \geq 0$ to consider the partition given in the property 1, then

$$(\mathbf{y}^{(1)}|\mathbf{y}_0^{(2)}) \sim El_m(\boldsymbol{\mu}^{(1.2)}, \boldsymbol{\Sigma}_{11.2}; h_{q(\mathbf{y}_0^{(2)})}),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{(1.2)} &= \boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_0^{(2)} - \boldsymbol{\mu}^{(2)}), \\
\boldsymbol{\Sigma}_{(11.2)} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}
\end{aligned}$$

and

$$q(\mathbf{y}_0^{(2)}) = (\mathbf{y}_0^{(2)} - \boldsymbol{\mu}^{(2)})^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_0^{(2)} - \boldsymbol{\mu}^{(2)}).$$

Similarly, for $\mathbf{y}^{(2)}|\mathbf{y}_0^{(1)}$. This means that the conditional distribution of a random vector that follows an elliptical distribution, given another random vector which also follows an elliptical distribution, also is elliptical.

Property 3: Let $\mathbf{y} \sim El_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, h)$ with $rank(\boldsymbol{\Sigma}) = r < n$. If the expected value and variance of the random vector $\mathbf{y}$ exist, then

(a) $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$ and,

(b) $\mathrm{Var}(\mathbf{y}) = \kappa\boldsymbol{\Sigma}$, where the constant $\kappa > 0$ is given by

$$\kappa = -2\varphi'(0) = -2\frac{d\varphi(u)}{du}\bigg|_{u=0},$$

where $\varphi(\cdot)$ is the characteristic function.

The table 1.1 Shows some examples of distributions pertaining to the elliptical class.

Distribution	$h(u)$	
Cauchy	$c(1+\frac{u}{s})^{-(v+1)/2}$	$s>0$
Power exponential	$c\ \exp(-u^s/2)$	
Logistic	$c\ \exp(-u)[1+\exp(-u)]^{-2}$	$u \geq 0$
Multivariate normal	$c\ \exp(-u/2)$	$u \geq 0$
Pearson type II	$c(1-u)^{-m}$	$m>0$
Pearson type VII	$c(1+\frac{u}{s})^{-m}$	$m>\frac{n}{2},\ s>0$
Slash	$\nu(2\pi)^{\frac{-m}{2}}\int_0^1 t^{\frac{m}{2}+\nu-1}\exp(-tu/2)dt$	$u \geq 0$
Kotz type	$cu^{m-1}\exp(ru^s)$	$r, s>0,\ 2m+n>2$
Student-t	$c(1+\frac{u}{s})^{-(\nu+m)/2}$	$m>0$

$c>0$ denotes the normalizing constant.

Tab. 1.1: *Elliptical distributions*

1.2.2 Varying coefficients model

A varying coefficient (Hastie and Tibshirani [1993]) assumes that

$$y_i = \alpha + \Sigma_{k=1}^{s}\beta_k(r_{ik})x_{ik} + \epsilon_i \qquad i = 1, \ldots, n, \tag{1.2}$$

where x_k and r_k are independent variables, α is a parameter, and β_k are nonparametric functions.

The varying coefficient regression models are very useful tools for analysing the relation between a response and a group of covariates. Their structure and interpretability are similar to those for the traditional linear regression model, but they are more flexible because of the infinite dimensionality of the corresponding parameter spaces. These models were introduced by Cleveland et al. [1992], and discussed by Hastie and Tibshirani [1993] in more detail, showing connections to other published models and giving some general estimation procedure.

Varying coefficient terms were introduced by Hastie and Tibshirani [1993] to accommodate a special type of interaction between explanatory variables. This interaction takes the form of $\beta(r)x$; that is, the linear coefficient of the explanatory variable x changes smoothly according to another explanatory variable r. More generally, r should be a continuous variable, while x can be either continuous or categorical.

The varying coefficient models have received much attention recently. They are a natural alternative to the additive model introduced by Breiman and Friedman [1985] and have greatly widened the scope of application by allowing the model coefficients may vary over certain covariates, such as time and temperature, see [Hastie and Tibshirani, 1993] and Fan and Zhang [2008]. Park et al. [2015] give a review of varying coefficient models.

1.2.3 Smoothing splines

Research on smoothing spline models has attracted a great deal of attention in recent years, and the methodology has been widely used in many areas. Most of the current smoothing techniques have their source in the 1980s Eilers and Marx [1996]. The connection between nonparametric regression and mixed models is discussed in the early 90s. Explanation about local regression models based on kernels is given in Härdle [1990]. Wand and Jones [1999] provided information on kernel smoothing. For local likelihood method in smoothing see Tibshirani and Hastie [1987] and Staniswalis [1989]. Wahba [1990] gives a rigorous treatment of smoothing splines. Penalization and spline models in a variety of settings are discussed in Green and Silverman [1994]. Hastie and Tibshirani [1990] discuss the procedures concerning linear smoothers in detail. For piecewise polynomials and splines, smoothing splines, mulidimensional splines and wavelet smoothing see Hastie et al. [2009]. Wand and Ormerod [2008] present semiparametric regression with O'Sullivan penalized splines. Wang [2011] explores smoothing splines methods and applications.

Other authors such as Verbyla et al. [1999], developed in more depth the issue of mixed smoothing models (in the context of cubic splines). Wand [2003] developed smoothing in mixed models considering truncated polynomial basis. A representation of the P-splines with base B-spline in mixed model is in Durbán et al. [2003]. Durbán et al. [2005] presented a simple semiparametric model using P-splines in mixed models for longitudinal data. Antoniadis et al. [2012] considered an estimator P-spline smoothing and variable selection in models with varying coefficient and demonstrates that the P-spline estimator is consistent when the number of knots increases with the number of individuals.

An obvious further generalization is to allow multiple smooth terms in a model, so that model fitting must balance goodness of fit against multiple penalties. Thin plate splines provide a way of modeling a single response to multiple covariates; data may be observations of linear functionals of the spline; models that are a mixture of parametric and spline terms can be produced, and all these extensions can be employed in a generalized linear modeling framework. A detailed account of thin plate smoothing splines, and various generalizations, may be found in Wahba [1990], Green and Silverman [1994], Wood [2003] and Wood [2006].

Thin-plate splines

Thin plate splines are a type of smoothing spline used for the visualization of complex relationships between continuous predictors and response variables. Thin plate smoothing splines are commonly applied to smooth multivariate interpolation of irregularly scattered noisy data. The theoretical foundations for the thin-plate spline were laid by Duchon [1975], Duchon [1977] and Meinguet [1979], and some further results and applications to meteorological problems were given in Wahba and Wendelberger [1980].

Definition 2: A function $g(\boldsymbol{t})$ is a thin plate spline on the data set $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$ if and only if g is of the form

$$g(\boldsymbol{t}) = \Sigma_{i=1}^{n} \delta_i \eta(\| \boldsymbol{t} - \boldsymbol{t}_i \|) + \Sigma_{j=1}^{3} a_j \phi_j(\boldsymbol{t})$$

for suitable constants δ_i and a_j, the function $\eta(r)$ is defined by

$$\eta(r) = \frac{1}{16\pi} r^2 \log r^2, \quad for\, r > 0$$

and $\mathbf{T}$ is $3 \times n$ matrix with elements $\mathbf{T}_{jk} = \phi_j(\boldsymbol{t}_k)$, so that

$$\mathbf{T} \quad = \quad \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \boldsymbol{t}_1 & \boldsymbol{t}_2 & \ldots & \boldsymbol{t}_n \end{pmatrix}.$$

If the vector $\boldsymbol{\delta}$ of coefficients δ_i satisfies

$$\mathbf{T}\boldsymbol{\delta} = 0$$

then g is said to be a natural thin-plate spline.

Given any smooth surface g, we wish to define a functional $J(g)$ that measured the overall roughness or "*wiggliness*" of g, in an analogous way to the integrated squared second derivative in one dimension. In two dimensions ($d = 2, m = 2$), the thin-plate penalty functional is given by

$$J_2^2(g) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (g_{t_1 t_2}^2 + 2g_{t_1 t_2}^2 + g_{t_2 t_2}^2)^2 dt_1 dt_2$$

and, in general,

$$J_m^d(g) = \sum_{\nu=0}^{m} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \binom{m}{\nu} \left(\frac{\partial^m g}{\partial t_1^\nu \partial t_2^{m-\nu}} \right)^2 dt_1 dt_2$$

The penalty function J(g) is always non-negative, and is zero if g is linear. In addition, thin plate penalty functional is rotationally invariant and are well suited to explanatory variables which are spatial coordinates in two (or three) dimensions. See Wahba [1990], Gu and Wahba [1993] or Green and Silverman [1994] for further information about thin-plate splines. Suppose, that g is a natural thin plate spline and let $\mathbf{y}$ be the vector with components y_i, Green and Silverman [1994] define the penalized residual sum of squares of a surface g by

$$S(g) = (\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}^T\boldsymbol{a})^T(\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}^T\boldsymbol{a}) + \lambda\boldsymbol{\delta}^T\mathbf{E}\boldsymbol{\delta}\ ,$$

where $\mathbf{a}$ and $\boldsymbol{\delta}$ are vectors with components a_i and δ_i, $\mathbf{E}$ is an $(n \times n)$ matrix defined by

$$E_{ij} = \frac{1}{16\pi}\|\mathbf{t}_i - \mathbf{t}_j\|^2 \log \|\mathbf{t}_i - \mathbf{t}_j\|^2\ ,$$

with $E_{ii} = 0$ for each i and the parameter $\lambda > 0$ is a smoothing parameter.
The minimization of this expression for $S(g)$ in matrix form, is given by

$$S(g) = \begin{pmatrix} \boldsymbol{\delta}^T & \mathbf{a}^T \end{pmatrix} \begin{bmatrix} \mathbf{E}^2 + \lambda\mathbf{E} & \mathbf{E}\mathbf{T}^T \\ \mathbf{T}\mathbf{E} & \mathbf{T}\mathbf{T}^T \end{bmatrix} \begin{pmatrix} \boldsymbol{\delta} \\ \boldsymbol{a} \end{pmatrix} - 2\begin{pmatrix} \boldsymbol{\delta}^T & \mathbf{a}^T \end{pmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{T} \end{bmatrix} \mathbf{y} + \mathbf{y}^T\mathbf{y}\,.$$

The smoothing problem as set out has a unique solution, that can be found by setting up and solving the system of equations

$$\begin{bmatrix} \mathbf{E} + \lambda\mathbf{I} & \mathbf{T}^T \\ \mathbf{T} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \widehat{\boldsymbol{\delta}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

1.2.4 Local influence

Diagnostics methods are techniques for exploring problems that compromise a regression analysis and for determining whether certain assumptions appear reasonable. This is essential to assess the sensitivity of the results with a set of available data, since the outliers (aberrant or influential) can distort the estimates of the parameters, leading in some cases to erroneous decisions.

There are several alternatives for assessing the influence of disturbances in data or assumptions of the model on the estimates of the parameters of interest (see, for example, Cook and Weisberg [1982] and Galea et al. [2000]). Among the most used techniques is the removal of cases to assess the effects of an observation about the process of estimation and hypothesis testing. This is an analysis of global influence [Cook, 1977].

Alternatively, Cook [1986] proposed an interesting method, called local influence, to evaluate the effect of small perturbations in the data or assumptions of the statistical model, on the maximum likelihood estimates, without removing observations.

Diagnostic methods have had a great development for linear mixed models under normal errors. Beckman et al. [1987] applied the approach to detect influential observations in normal Linear mixed models with emphasis on studying the influence of single observations. Verbeke and Lesaffre [1996] extended the local influence methodology to normal linear mixed models in repeated-measurement context and under the case-weight perturbation scheme. Ouwens et al. [2001] applied the local influence approach in generalized linear mixed models. Kim et al. [2002] presented some diagnostic measures of influence, as functions of residuals and deviations for the estimates of parametric and non-parametric components.

In the context of nonparametric and semiparametric regression models, the local influence diagnostics are almost rare. In particular, Thomas [1991] constructed local influence diagnostics for the smoothing parameter, Fung et al. [2002] extended the measures based diagnostic residuals and Cook distances for the penalized maximum likelihood estimation of a linear semiparametric mixed model with normal effect, and Zhu and Lee [2003] extended the works by Cook [1986] to provide local influence measures under different perturbation schemes in normal partially linear models. Ibacache-Pulgar [2009] extended influence local for semiparametric additive mixed models with elliptic contours and Ibacache-Pulgar and Paula [2011] for t-Student partially linear models.

On the other hand, the local influence methodology was extended to elliptic linear models that include distributions as Student-t, power exponential, Cauchy, normal contaminated, among others, with tails more or less heavy than the normal distribution. In Galea et al. [1997] and Liu [2000] applied the local influence approach in multivariate elliptical linear models under various perturbation schemes. Díaz-García et al. [2003] extended the local influence methodology to elliptical linear regression model. Galea et al. [2005] applied the local influence methods in functional and structural comparative calibration models under elliptical t-distributions. Paula et al. [2003] developed local influence for symmetrical nonlinear models. The development of local influence methods in the context of mixed effects models and longitudinal structure can be found in the work of Osorio [2006] and Osorio et al. [2007]. These authors developed normal curvatures of local influence of disturbance to various schemes for elliptical linear mixed models.

Recently, Zhang et al. [2007] extended the work of Cook [1986] to provide influence diagnostics in partially varying coefficient models and Li et al. [2009] applied influence diagnostics and outlier test for varying coefficient mixed models.

1.3 Motivation of the thesis

According to our review of literature discussed in Section 1.1 and the background presented in Section 1.2, we have the following motivations to develop this thesis:

- The need to model those aspects of the statistical literature in which it is suggested that the multivariate normal distribution is not appropriate when the data come, for example, from heavy tail distributions. In this context, several works recent, specially on normal varying coefficients mixed models, suggest to replace the multivariate normal distributions by elliptical distributions.

- The cases deletion and local influence diagnostic method have been applied en several areas of statistics, specially, in normal varying coefficients mixed model. In this context, there are few works where local influence is applied in varying-coefficients mixed models with elliptical errors and null for TPSVCMMs

In this circumstances, we propose the following hypothesis:

Hypothesis 1: It is possible adjust the TPSVCMM by using elliptical distributions.

Hypothesis 2: It is possible to apply the local influence method for assessing the influence of data and model perturbations in elliptical TPSVCMM.

Hypothesis 3: It is possible to estimate the smoothing parameters in the TPSVCMM under elliptical distributions using methods for choosing the smoothing parameters.

Hypothesis 4: It is possible to apply the results to real data sets from different research areas.

1.4 Objectives of the thesis

1.4.1 Objectives

The general objective of this thesis is study the TPSVCMM under elliptical distributions.

The specific objectives of this project are:

1. Study the statistical inference problem in TPSVCMM under elliptical distributions.

   1.1 Obtain the maximum penalized likelihood estimates.

1.2 Obtain the expected information matrix.

1.3 Obtain large sample test for hypothesis of interest.

1.4 Illustrate the methodology to a real data sets.

2. Apply the local influence diagnostic method to TPSVCMM under elliptical distributions.

2.1 Develop the method of local influence under this context.

2.2 Illustrate the methodology to a real data sets.

3. Estimate the smoothing parameters in TPSVCMM under elliptical distributions.

3.1 Derive the Generalized Cross Validation Score.

3.2 Derive the Akaike Information Criterion.

3.3 Obtain the estimates of the smoothing parameters.

1.5 Products of the thesis

This thesis led to the following products:

(1) **Moraga Cárdenas, M.S.**, **Ibacache Pulgar, Germán** and **Nicolis, Orietta** (2018) Thin-plate spline partially varying-coefficient model under elliptical distributions. Submitted.

(2) **Moraga Cárdenas, M.S.**, **Ibacache Pulgar, Germán** and **Nicolis, Orietta** (2017) Project DID $S2017-32$ of the Universidad Austral de Chile, Valdivia, Chile.

1.6 Organization of the thesis

This thesis contains six chapters taking into account this introduction chapter. In Chapter 2, we propose the Thin-plate spline varying coefficients mixed model, under elliptical errors. In Chapter 3, we develop the methodology based maximum penalized likelihood method that allows estimating the parameters involved in TPSVCMM, under elliptical erros. In Chapter 4 we developed methods of local influence for TPSVCMM under elliptical errors, considering several disturbance schemes. In the Chapter 5, we an application for TPSVCMM under elliptical errors, considering data of the house prices of area Boston. Finally, in Chapter 6, we present general conclusions for this thesis and propose some topics for future work.

2

Thin-plate spline varying coefficients mixed model

TPSVCMMs emerges as a powerful tool in statistical modeling because of its flexibility to model explanatory variables effects that can contribute to the parametric part (constant coefficients) and that varying with a given factor, and their ability to model the structure variance-covariance among observations through random effects. Moreover, this class of models incorporate thin-plate splines, which are the natural generalization of cubic splines to any number of dimensions and almost any order of wiggliness penalty, to model, for example, the effect of the geographical location of the observations under study.

2.1 Model specification

The TPSVCMM assume that the relationship between the response variable and the explanatory variables can be represented as

$$y_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\alpha} + \sum_{k=1}^{s} x_{ij}^{(k)} \beta_k(\mathrm{r}_{k_{ij}}) + \mathbf{u}_{ij}^T \mathbf{b}_i + g(\mathbf{t}_i) + \boldsymbol{\varepsilon}_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i)\,, \quad (2.1)$$

where y_{ij} denotes the jth measure associated with the ith cluster at time point $\mathrm{r}_{k_{ij}}$, $\mathbf{z}_{ij}$ and $\mathbf{u}_{ij}$ are $(p \times 1)$ and $(q \times 1)$ vectors of explanatory variable values, $\boldsymbol{\alpha}$ is a $(p \times 1)$ fixed parameter vector, $\beta_k(\cdot)$ $(k = 1, \ldots, s)$ are unknown smooth arbitrary functions of r_k associated with the covariates $x_{ij}^{(k)}$, $\mathbf{b}_i$ denotes the $(q \times 1)$ vector of random effects, $g(\cdot)$ is a smooth thin-plate spline that depends of the vector $\mathbf{t_i}$ of coordinates for smooth surface o well corresponds the interaction of other variables $\in \mathbb{R}^2$, and $\boldsymbol{\varepsilon}_{ij}$ is a random error.

TPSVCMM is an extension of others models that have been proposed in literature. For example:

1. If $\beta_k(\cdot) = 0 \quad (k = 1, \ldots, s)$ and $g(\mathbf{t}_i) = 0$, the model is reduced to mixed model proposed by Laird and Ware [1982] given by

$$y_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\alpha} + \mathbf{u}_{ij}^T \mathbf{b}_i + \boldsymbol{\varepsilon}_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i)\,.$$

2. If $\boldsymbol{\alpha} = \mathbf{0}$, $\mathbf{b}_i = \mathbf{0}$ $(i = 1, \ldots, n)$, $g(\mathbf{t}_i) = 0$ and $x_{ij}^{(k)} = 1$, then the model is reduced to the additive model discussed in Buja et al. [1989],

$$y_{ij} = \sum_{k=1}^{s} \beta_k(\mathrm{r}_{k_{ij}}) + \boldsymbol{\varepsilon}_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i) .$$

3. If $x_{ij} = 1$, $g(\mathbf{t}_i) = 0$, $s = 1$ and $\boldsymbol{\varepsilon}_{ij} = v_i(\mathrm{r}_{k_{ij}}) + \epsilon_{ij}$, with $v_i(t)$ being an independent stochastic process ϵ_{ij}, the model corresponds to semiparametric mixed model presented by Zhang et al. [1998], and has the form

$$y_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\alpha} + \mathbf{u}_{ij}^T \mathbf{b}_i + \beta(\mathrm{r}_{k_{ij}}) + v_i(\mathrm{r}_{k_{ij}}) + \epsilon_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i) .$$

4. If $\boldsymbol{\alpha} = \mathbf{0}$, $x_{ij} = 1$, $g(\mathbf{t}_i) = 0$ and $s = 1$, then the model is reduced to the nonparametric mixed model proposed by Wang [1998] and has the form

$$y_{ij} = \mathbf{u}_{ij}^T \mathbf{b}_i + \beta(\mathrm{r}_{k_{ij}}) + \boldsymbol{\varepsilon}_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i) .$$

5. If $\boldsymbol{\alpha} = \mathbf{0}$, $\mathbf{b}_i = \mathbf{0}$ and $\beta_k(\cdot) = 0$ $(k = 1, \ldots, s)$, the model is reduced to thin-plate spline model referenced in Green and Silverman [1994],

$$y_{ij} = g(\mathbf{t}_i) + \boldsymbol{\varepsilon}_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i) .$$

6. If $\mathbf{z} = \mathbf{1}$, $\mathbf{b}_i = \mathbf{0}$ and $\beta_k(\cdot) = 0$ $(k = 1, \ldots, s)$, the model is reduced to thin-plate spline model referenced in Yanosky et al. [2014],

$$y_{ij} = \boldsymbol{\alpha} + g(\mathbf{t}_i) + \boldsymbol{\varepsilon}_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i) .$$

7. If $\boldsymbol{\alpha} = \mathbf{0}$ and $g(\mathbf{t}_i) = 0$, the model is reduced to mixed model with varying coefficients proposed by Li et al. [2009] given by

$$y_{ij} = \sum_{k=1}^{s} x_{ij}^{(k)} \beta_k(\mathrm{r}_{k_{ij}}) + \mathbf{u}_{ij}^T \mathbf{b}_i + \boldsymbol{\varepsilon}_{ij} \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i) .$$

2.1.1 Matrix representation

In order to write model (2.1) in a matrix form, first consider the following one-to-one linear transformation of the vector $\mathbf{g}$ suggested, for example, by Green and Silverman

[1994]:

$$\mathbf{g} = \begin{pmatrix} g(\mathbf{t}_1) \\ \vdots \\ g(\mathbf{t}_n) \end{pmatrix} = \mathbf{E}\boldsymbol{\delta} + \mathbf{T}^T\mathbf{a}\,,$$

where $\mathbf{a}$ and $\boldsymbol{\delta}$ are vectors with components a_i and δ_i, $\mathbf{E}$ is an $(n \times n)$ matrix defined by

$$E_{ij} = \frac{1}{16\pi}\|\mathbf{t}_i - \mathbf{t}_j\|^2 \log \|\mathbf{t}_i - \mathbf{t}_j\|^2\,,$$

with $E_{ii} = 0$ for each i, and $\mathbf{T}$ is an $(2 \times n)$ matrix given by

$$\mathbf{T} = \begin{pmatrix} \mathbf{t}_1 & \mathbf{t}_2 & \ldots & \mathbf{t}_n \end{pmatrix}\,.$$

Thus, the model (2.1) take the form

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\alpha} + \sum_{k=1}^{s} \widetilde{\mathbf{N}}_{ki}\boldsymbol{\beta}_k + \mathbf{U}_i\mathbf{b}_i + \widetilde{\mathbf{E}}_i\boldsymbol{\delta} + \widetilde{\mathbf{T}}_i\mathbf{a} + \boldsymbol{\epsilon}_i \qquad (i = 1, \ldots, n; j = 1, \ldots, m_i)\,, \quad (2.2)$$

where $\mathbf{y}_i$ is a $(m_i \times 1)$ random vector of observed responses from the ith cluster, $\mathbf{Z}_i$ is a $(m_i \times p)$ design matrix with rows $\mathbf{z}_{ij}^T$, $\widetilde{\mathbf{N}}_{ki} = \mathbf{X}_i^{(k)}\mathbf{N}_{ki}$, $\mathbf{X}_i = \text{diag}(\mathrm{x}_{i1}, \ldots, \mathrm{x}_{im_i})$, $\mathbf{N}_{ki}$ is an $(m_i \times c)$ incidence matrix with the (j, l)th element equal to the indicator $I(\mathrm{r}_{ij} = \mathrm{r}_l^0)$ $(j = 1, \ldots, m_i)$, where r_l^0 $(l = 1, \ldots, c\,)$ denotes the distinct and ordered values of the explanatory variable r_{ij}, $\boldsymbol{\beta}_k = (\psi_{k_1}, \ldots, \psi_{k_c})^T$ is a $(c \times 1)$ vector of parameters with $\psi_{k_l} = \beta_k(\mathrm{r}_1^0)$, for $l = 1, \ldots, c$, $\mathbf{U}_i$ is a $(m_i \times q)$ design matrix with rows $\mathbf{u}_{ij}^T$, $\widetilde{\mathbf{E}}_i = \mathbf{F}_i\mathbf{E}$ $\widetilde{\mathbf{T}}_i = \mathbf{F}_i\mathbf{T}^T$, $\mathbf{F}_i$ is an $(m_i \times n)$ matrix with an $(m_i \times 1)$ vector of ones in the ith column and zeros in the remanning positions and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{m_i})^T$ is an $(m_i \times 1)$ vector of within-cluster errors that follows a elliptical distribution with location parameter $\mathbf{0}$ and scale matrix $\boldsymbol{\Sigma_i}$ (see, for example, Fang, KT and Kotz, 1990.

2.1.2 Distribution assumption

In general, in TPSVCMM under elliptical distributions two approaches can be adopted for obtaining the marginal distribution. On the assumption that the random effects and the errors follow the elliptical distributions, we have the following hierarchical formulation (see, Pinheiro et al. [2001]) for the model given in (2.1)

$$\mathbf{y}_i|\mathbf{b}_i \overset{\text{ind}}{\sim} El_{m_i}(\boldsymbol{\mu}_i, \boldsymbol{V}_i)$$

$$\mathbf{b}_i \overset{\text{ind}}{\sim} El_q(\mathbf{0}, \boldsymbol{D}) \qquad \text{and}$$

$$\boldsymbol{\varepsilon}_i \overset{\text{ind}}{\sim} El_{m_i}(\mathbf{0}, \boldsymbol{V}_i),$$

where $\boldsymbol{\mu}_i = \boldsymbol{Z}_i\boldsymbol{\alpha} + \sum_{k=1}^{s} \tilde{\boldsymbol{N}}_{ki}\boldsymbol{\beta}_k + \boldsymbol{F}_i\boldsymbol{g}$, $\boldsymbol{D}$ and $\boldsymbol{V}_i$ are $(q \times q)$ and $(m_i \times m_i)$ scale matrices, respectively. In this case, the joint distributions of $(\mathbf{y}_i^T, \mathbf{b}_i^T)^T$ is not necessarily elliptical and it becomes difficult to obtain the marginal distribution of the response $\mathbf{y}_i$.

Alternatively, we can consider the following marginal formulation described by Lange et al. [1989]:

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{b}_i \\ \boldsymbol{\varepsilon}_i \end{pmatrix} \sim El_{m_i+q+m_i} \left\{ \begin{pmatrix} \boldsymbol{\mu}_i \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_i & \boldsymbol{U}_i\boldsymbol{D} & \boldsymbol{V}_i \\ \boldsymbol{D}\boldsymbol{U}_i^T & \boldsymbol{D} & \mathbf{0} \\ \boldsymbol{V}_i & \mathbf{0} & \boldsymbol{V}_i \end{pmatrix} \right\},$$

where $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ are uncorrelated, but not necessarily independent, except for the normal case, and the matrices $\boldsymbol{\Sigma}_i = \boldsymbol{U}_i\boldsymbol{D}\boldsymbol{U}_i^T + \boldsymbol{V}_i$, $\boldsymbol{D}$, $\boldsymbol{U}_i\boldsymbol{D}$ and $\boldsymbol{V}_i$ are proportional to the variance-covariance matrices of $\mathbf{y}_i$, $\mathbf{b}_i$, $(\mathbf{y}_i^T, \mathbf{b}_i)^T$ and $\boldsymbol{\varepsilon}_i$, respectively, by a quantify k.

Using properties of elliptical distributions described in (1.2.1), we will assume that $\boldsymbol{\epsilon}_i$ follows a elliptical distribution with location parameter $\mathbf{0}$ and scale matrix $\boldsymbol{\Sigma_i}$ (see, for example, Kai-Tai and Yao-Ting [1990]). Consequently, the distribution of $\mathbf{y}_i$ is given by

$$\mathbf{y}_i \sim \mathrm{El}_{m_i}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma_i}\right),$$

where

$$\boldsymbol{\mu}_i = \mathbf{Z}_i\boldsymbol{\alpha} + \sum_{k=1}^{s} \widetilde{\mathbf{N}}_{ki}\boldsymbol{\beta}_k + \widetilde{\mathbf{E}}_i\boldsymbol{\delta} + \widetilde{\mathbf{T}}_i\boldsymbol{a}\,.$$

In order to ensure that the random vector $\mathbf{y}_i$ admits a density for all $\mathbf{y}_i \in \mathcal{R}^{m_i}$ with respect to the Lebesgue measure, we will assume the matrix $\boldsymbol{\Sigma}_i$ positive-definite with structure given by $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\boldsymbol{\tau})$, where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)^T$. Then, the density function of the random vector of observed responses $\mathbf{y}_i$, defined in 1.1, is given by

$$f(\mathbf{y}_i) = |\boldsymbol{\Sigma}_i|^{-1/2}\ h(u_i)\,, (i = 1, \ldots, n)\,,$$

where $u_i = \boldsymbol{\epsilon}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\epsilon}_i$ is the Mahalanobis distance, $\boldsymbol{\epsilon}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$, and $h(\cdot)$ is a function of $\mathcal{R} \rightarrow [0, \infty]$ known as the density generator function such that

$$\int_0^{\infty} u^{m/2-1}h(u)du < \infty.$$

From the analytical point of view, this formulation is quite convenient, but it has the disadvantage that random effects do not have the same interpretation of fixed effects as in the hierarchical case.

An important aspect to consider is the identifiability of the parameters or the model. It is said that parameters or models are unidentifiable if two different sets of parameters carry the same probability distribution [Demidenko, 2013].

Definition 3: Let a statistical model be defined by family of distributions for $\boldsymbol{y}$ parameterized by the vector $\boldsymbol{\theta}$, $\boldsymbol{P_\theta}, \boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the parameter space and $\boldsymbol{P_\theta}$ denotes the distribution associated with $\boldsymbol{\theta}$. We say that the model is identifiable on $\boldsymbol{\Theta}$ if $\boldsymbol{P_{\theta_1}} = \boldsymbol{P_{\theta_2}}$ implies that $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$.

According to Demidenko [2013], for regression models with normal distribution the condition

$$E_{\boldsymbol{\theta_1}}(\boldsymbol{y}) = E_{\boldsymbol{\theta_2}}(\boldsymbol{y}) \qquad \text{and} \qquad cov_{\boldsymbol{\theta_1}} = cov_{\boldsymbol{\theta_2}} \qquad \text{imply} \qquad \boldsymbol{\theta_1} = \boldsymbol{\theta_2},$$

is a necessary and sufficient condition to ensure identifiability, since normal distribution is specified only for the first two moments.

For model 2.1, the problem of identifiability can be generated by the overdispersion of non-parametric functions together with the other parameters. In practice, we can obtain the identifiability of the model (a) imposing conditions on functions $\beta_k (k = 1, \ldots, s)$, (b) incorporating conditions the distribution of random effects, or the combination of (a) and (b), (see, for example, Ke, Chunlei and Wang, Yuedong, 2001).

On the other hand, a deficiency that may occur with 2.1 is the sensitivity of estimates of maximum likelihood to aberrant points. In some situations there may be evidence (through residue analysis) that errors have lighter or heavier tails than normal errors. In these cases, distributions for errors that flexibilize kurtosis can be assumed, and in particular, in the case of error with elliptic distributions (Fang, KT and Kotz, S 1990) with heavy tails, the maximum likelihood estimates are robust against aberrant observations. Another alternative is the application of robust methods on models with elliptical errors with heavier tails than normal errors. There is, however, no guarantee that these methods will lead to greater protection against extreme observances than the maximum likelihood method, in addition to computational cost being more costly.

2.1.3 Penalized function

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_s^T, \boldsymbol{\delta}^T, \mathbf{a}^T, \boldsymbol{\tau}^T)^T$ the vector of parameters to be estimated. Thus, the log-likelihood function associated to , is given by

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} L_i(\boldsymbol{\theta}), \tag{2.3}$$

where $L_i(\boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log h(u_i)$ represents the individual contribution of the ith observation.

It is known fact that maximizing the log-likelihood function without imposing restrictions over the nonparametric functions may cause overfitting and non identification of $\boldsymbol{\alpha}$. A well known procedure that can solve this problem is based on the idea of log-likelihood penalization and consists in incorporating a penalty function over each function β_k and g, such that

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda_1, \ldots, \lambda_s, \lambda_g) \quad = \quad L(\boldsymbol{\theta}) + \sum_{k=1}^{s} \lambda_k^* J(\beta_k) + \lambda_g^* J(g)\ , \tag{2.4}$$

where $J(\beta_k)$ and $J(g)$ denotes the penalty functions over β_k and g, respectively, and $\lambda_k^* = \lambda^*(\lambda_k)$ and $\lambda_g^*(\lambda_g)$ are constants that depends on the parameter $\lambda_k \geq 0$ and $\lambda_g \geq 0$. In this work, we will consider penalty functions of type

$$J(\beta_k) \quad = \quad \int_{a_k}^{b_k} [\beta_k^{\imath}(\mathrm{r}_k)]^2 d\mathrm{t}_k,$$

and

$$J_m^d(g) = \sum_{\upsilon_1+\ldots+\upsilon_d=m} \frac{m!}{\upsilon_1!\ldots\upsilon_d!} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left(\frac{\partial^m g}{\partial t_1^{\alpha_1}\ldots\partial t_d^{\alpha_d}}\right)^2 \prod_{j=1}^{d} dt_j$$

where $\beta_k^{(\imath)}(\mathrm{r}_k) = \frac{\mathrm{d}^{\imath}}{\mathrm{dr}^{\imath}}\beta(\mathrm{r}_k)$, $\mathrm{r}_{k_l}^0 \in [a_k, b_k]$, and the functions β_k's belongs to the Sobolev function space

$$\mathcal{W}_2^{(\imath)} = \{\beta_k : \beta_k, \beta_k^{(1)}, \ldots, \beta_k^{(\imath-1)}\text{abs. cont.}, \beta_k^{(\imath)} \in \mathcal{L}^2[a_k, b_k]\}\ ,$$

and g belongs to the functions space whose partial derivatives of total order m are in Hilbert space $\mathcal{L}^2[E^d]$ of square integrable functions on Euclidean d-space. It is important mention that for $\imath = 2$, the estimation of β_k leads to a natural cubic spline with knots at the points $\mathrm{r}_{k_l}^0$, for $l = 1, \ldots, r_k$. On the other hand, for $d = 2, m = 2$ and $g = g(t_1, t_2)$, this is,

$$J(g) \quad = \quad \int\int_{\mathcal{R}^2} \left\{ \left(\frac{\partial^2 g}{\partial t_1^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial t_1 \partial t_2} \right)^2 + \left(\frac{\partial^2 g}{\partial t_2^2} \right)^2 \right\} dt_1 dt_2 \,,$$

the estimation of g leads to a natural thin-plate spline. According to Green and Silverman [1994], we may express the penalty functional as

$$J(\beta_k) = \boldsymbol{\beta}_k^T \mathbf{K}_k \boldsymbol{\beta}_k \qquad \text{and} \qquad J(g) = \boldsymbol{\delta}^T \mathbf{E} \boldsymbol{\delta},$$

where $\mathbf{K}_k$ is a $(q_k \times q_k)$ non-negative definite smoothing matrix associated with the kth explanatory variable that depends only on the knots. Then, if we consider $\lambda_k^* = -\lambda_k/2$ and $\lambda_g^* = -\lambda_g/2$, the penalized log-likelihood function (2.4) can be expressed as

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad = \quad L(\boldsymbol{\theta}) - \sum_{k=1}^{s} \frac{\lambda_k}{2} \boldsymbol{\beta}_k^T \mathbf{K}_k \boldsymbol{\beta}_k - \frac{\lambda_g}{2} \boldsymbol{\delta}^T \mathbf{E} \boldsymbol{\delta}, \tag{2.5}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s, \lambda_g)^T$ denotes a $((s+1) \times 1)$ vector of smoothing parameters that controls the tradeoff between goodness of fit and the smoothness estimated functions. Therefore, the determination of the parameters λ's and λ_g is a crucial part in the estimation process and different choice methods are available in the literature. For smoothing splines, for example, one may use the generalized cross-validation method or AIC criterion.

3

Parameter estimation

In this chapter we discuss the problem of estimating the parameters under TPSVCM model, based on penalized log-likelihood. Specifically, we propose an iterative process based on the Fisher score and backfitting algorithms to estimate the regression coefficient, the non-parametric functions and the smooth surface, and their respective standard errors from the penalized Fisher information matrix. In addition, we present a discussion about the calculation of effective degrees of freedom, the selection of models and the estimation of smoothing parameters.

3.1 Score function

Assuming that (2.4) is regular with respect to all elements of $\boldsymbol{\theta}$, we have that the penalized score function of $\boldsymbol{\theta}$ under elliptical TPSVCMM 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

$$\begin{aligned}
\frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\alpha}} &= \frac{\partial L_i(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} \\
&= \frac{g'(\delta_i)}{g(\delta_i)} \frac{\partial \epsilon_i^T}{\partial \alpha} \frac{\partial [\epsilon_i^T \Sigma_i^{-1} \epsilon_i]}{\partial \epsilon_i} \\
&= \upsilon(\delta_i) \mathbf{Z}^{T_i} \boldsymbol{\Sigma}_i^{-1} \epsilon_i \\
&= \mathbf{Z}^T \mathbf{W}_{\upsilon} \boldsymbol{\epsilon},
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}_k} &= \frac{\partial L_i(\boldsymbol{\theta})}{\partial \beta_k} - \frac{\lambda_k}{2n}\frac{\partial[\boldsymbol{\beta}_k^T \mathbf{K}_k \boldsymbol{\beta}_k]}{\partial \boldsymbol{\beta}_k} \\
&= \frac{g'(\delta_i)}{g(\delta_i)}\frac{\partial \epsilon_i^T}{\partial \boldsymbol{\beta}_k}\frac{\partial[\epsilon_i^T \boldsymbol{\Sigma}_i^{-1}\epsilon_i]}{\partial \epsilon_i} - \frac{\lambda_k}{n}\mathbf{K}_k\boldsymbol{\beta}_k \\
&= \upsilon(\delta_i)\widetilde{\mathbf{N}}_k^T \boldsymbol{\Sigma}_i^{-1}\epsilon_i - \frac{\lambda_k}{n}\mathbf{K}_k\boldsymbol{\beta}_k \\
&= \widetilde{\mathbf{N}}_k^T \mathbf{W}_\upsilon \boldsymbol{\epsilon} - \lambda_k \mathbf{K}_k \boldsymbol{\beta}_k \qquad (k = 1,\ldots,s),
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial \boldsymbol{\delta}} &= \frac{\partial L_i(\boldsymbol{\theta})}{\partial \delta} - \frac{\lambda_g}{2}\mathbf{E}\boldsymbol{\delta} \\
&= \frac{g'(\delta_i)}{g(\delta_i)}\frac{\partial \epsilon_i^T}{\partial \boldsymbol{\delta}}\frac{\partial[\epsilon_i^T \boldsymbol{\Sigma}_i^{-1}\epsilon_i]}{\partial \epsilon_i} - \frac{\lambda_g}{2}\mathbf{E}\boldsymbol{\delta} \\
&= \upsilon(\delta_i)\widetilde{\mathbf{E}}_i^T \boldsymbol{\Sigma}_i^{-1}\epsilon_i - \frac{\lambda_g}{2}\mathbf{E}\boldsymbol{\delta} \\
&= \widetilde{\mathbf{E}}\mathbf{W}_\upsilon \boldsymbol{\epsilon} - \lambda_g \mathbf{E}\boldsymbol{\delta},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial \boldsymbol{a}} &= \frac{\partial L_i(\boldsymbol{\theta})}{\partial \mathbf{a}} \\
&= \frac{g'(\delta_i)}{g(\delta_i)}\frac{\partial \epsilon_i^T}{\partial \mathbf{a}}\frac{\partial[\epsilon_i^T \boldsymbol{\Sigma}_i^{-1}\epsilon_i]}{\partial \epsilon_i} \\
&= \upsilon(\delta_i)\widetilde{\mathbf{T}}_i^T \boldsymbol{\Sigma}_i^{-1}\epsilon_i \\
&= \widetilde{\mathbf{T}}\mathbf{W}_\upsilon \boldsymbol{\epsilon},
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial \tau_\ell} &= \frac{\partial L_i(\boldsymbol{\theta})}{\partial \tau_\ell} \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left[\mathrm{tr}\Big(\boldsymbol{\Sigma}_i^{-1}\frac{\partial \boldsymbol{\Sigma}_i}{\partial \ell}\Big) - \upsilon_i\,\boldsymbol{\epsilon}_i^T\,\boldsymbol{\Sigma}_i^{-1}\frac{\partial \boldsymbol{\Sigma}_i}{\partial \ell}\,\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\epsilon}_i\right],
\end{aligned}
$$

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

Since under non-Gaussian semiparametric models the estimating equations are typically nonlinear requiring iterative solutions, various methods have been studied in the literature. For example, the Fisher-scoring or Newton-Raphson methods (see, for example, Green [1987] and Rigby and Stasinopoulos [2005] and the EM algorithm Dempster et al. [1977].

3.2 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

$$\mathbf{L}_{\rm p} = \sum_{i=1}^{n} \frac{\partial^2 L_{{\rm p}_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial\theta_{j^*}\theta_{\ell^*}} \,,$$

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 $\boldsymbol{\Psi}_{1i} = \upsilon^{'}_i \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma}_i^{-1}$, $\upsilon^{'}_i = \frac{{\rm d}\,\upsilon_i}{{\rm d}\delta_i}$ and $\boldsymbol{\Psi}_{2i} = \upsilon_i \boldsymbol{\Sigma}_i^{-1}$. In addition, let $\boldsymbol{\Psi} = {\rm diag}_{1\le i\le n}\big(\boldsymbol{\Psi}_i\big)$ and $\boldsymbol{\Omega} = {\rm diag}_{1\le i\le n}\big(\boldsymbol{\Omega}_i\big)$, with $\boldsymbol{\Omega}_i = \boldsymbol{\Psi}^*_i \frac{\partial \boldsymbol{\Sigma}_i}{\partial j} \boldsymbol{\Sigma}_i^{-1}$. After some algebraic manipulations we find

$$\mathbf{L}_{\rm p}(\boldsymbol{\theta}) \;=\; \frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \begin{pmatrix} \mathbf{L}_{\rm p}^{\alpha\alpha}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\alpha\beta}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\alpha\gamma}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\alpha\tau}(\boldsymbol{\theta}) \\ \mathbf{L}_{\rm p}^{\alpha\beta^T}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\beta\beta}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\beta\gamma}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\beta\tau}(\boldsymbol{\theta}) \\ \mathbf{L}_{\rm p}^{\alpha\gamma^T}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\beta\tau^T}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\gamma\gamma}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\gamma\tau}(\boldsymbol{\theta}) \\ \mathbf{L}_{\rm p}^{\alpha\tau^T}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\beta\tau^T}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\gamma\tau^T}(\boldsymbol{\theta}) & \mathbf{L}_{\rm p}^{\tau\tau}(\boldsymbol{\theta}) \end{pmatrix}$$

$$\begin{aligned} \frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} &= \frac{\partial[\upsilon(\delta_i)\mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1}\epsilon_i]}{\partial\boldsymbol{\alpha}^T} \\ &= \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \left[\epsilon_i \frac{\partial\upsilon(\delta_i)}{\partial\boldsymbol{\alpha}^T} + \upsilon(\delta_i)\frac{\partial\epsilon_i}{\partial\boldsymbol{\alpha}^T}\right] \\ &= \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \left[-2\upsilon'(\delta_i)\epsilon_i\epsilon_i^T \boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i - \upsilon(\delta_i)\mathbf{Z}_i\right] \\ &= -\mathbf{Z}^T \boldsymbol{\Psi} \mathbf{Z} \,, \end{aligned}$$

$$\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial\boldsymbol{\beta}_k\partial\boldsymbol{\beta}_{k'}^T} = \begin{cases} -\widetilde{\mathbf{N}}_k^T \boldsymbol{\Psi} \widetilde{\mathbf{N}}_k - \lambda_k\, \mathbf{K}_k & k = k' \\ \\ -\widetilde{\mathbf{N}}_k^T \boldsymbol{\Psi} \widetilde{\mathbf{N}}_{k'} & k \neq k', \end{cases}$$

$$\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial\boldsymbol{\beta}_k\partial\boldsymbol{\gamma}} = \begin{cases} -\widetilde{\mathbf{N}}_k^T \boldsymbol{\Psi} \widetilde{\mathbf{E}} & \boldsymbol{\gamma} = \boldsymbol{\delta} \\ \\ -\widetilde{\mathbf{N}}_k^T \boldsymbol{\Psi} \widetilde{\mathbf{T}}^T & \boldsymbol{\gamma} = \mathbf{a}, \end{cases}$$

$$\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} = \begin{cases} -\widetilde{\mathbf{E}}^T\boldsymbol{\Psi}\widetilde{\mathbf{E}} - \lambda_g\,\mathbf{E} & \boldsymbol{\gamma}=\boldsymbol{\delta} \\ \\ -\widetilde{\mathbf{T}}^T\boldsymbol{\Psi}\widetilde{\mathbf{T}} & \boldsymbol{\gamma}=\mathbf{a}, \end{cases}$$

$$\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{'}} = -\widetilde{\mathbf{Z}}^T\boldsymbol{\Psi}\widetilde{\mathbf{T}} \qquad \boldsymbol{\gamma}=\mathbf{a},\quad \boldsymbol{\gamma}^{'}=\delta$$

$$\begin{aligned}\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}_k^T} &= \frac{\partial[\upsilon(\delta_i)\mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\epsilon_i]}{\partial\boldsymbol{\beta}_k^T} \\ &= \mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\left[\epsilon_i\frac{\partial\upsilon(\delta_i)}{\partial\boldsymbol{\beta}_k^T} + \upsilon(\delta_i)\frac{\partial\epsilon_i}{\partial\boldsymbol{\beta}_k^T}\right] \\ &= \mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\left[-2\upsilon'(\delta_i)\epsilon_i\epsilon_i^T\boldsymbol{\Sigma}_i^{-1}\widetilde{\mathbf{N}}_k - \upsilon(\delta_i)\widetilde{\mathbf{N}}_k\right] \\ &= -\mathbf{Z}^T\boldsymbol{\Psi}\widetilde{\mathbf{N}}_k\,,\end{aligned}$$

$$\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\gamma}^T} = \begin{cases} -\mathbf{Z}^T\boldsymbol{\Psi}\widetilde{\mathbf{E}} & \boldsymbol{\gamma}=\boldsymbol{\delta} \\ \\ -\mathbf{Z}^T\boldsymbol{\Psi}\widetilde{\mathbf{T}} & \boldsymbol{\gamma}=\mathbf{a}, \end{cases}$$

$$\begin{aligned}\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\alpha}\partial\tau_\jmath} &= \frac{\partial[\upsilon(\delta_i)\mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\epsilon_i]}{\partial\tau_\jmath} \\ &= -\mathbf{Z}_i^T\left[\upsilon_i'\epsilon_i\epsilon_i^T\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\tau_\jmath}\boldsymbol{\Sigma}_i^{-1} + \upsilon_i\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\tau_\jmath}\boldsymbol{\Sigma}_i^{-1}\right]\epsilon_i \\ &= -\mathbf{Z}^T\boldsymbol{\Omega}\boldsymbol{\epsilon}\,,\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}_k\partial\tau_\jmath} &= \frac{\partial[\upsilon(\delta_i)\widetilde{\mathbf{N}}_k^T\boldsymbol{\Sigma}_i^{-1}\epsilon_i - \frac{\lambda_k}{2}\mathbf{K}_k\boldsymbol{\beta}_k]}{\partial\tau_\jmath} \\ &= -\widetilde{\mathbf{N}}_k^T\left[\upsilon_i'\epsilon_i\epsilon_i^T\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\tau_\jmath}\boldsymbol{\Sigma}_i^{-1} + \upsilon_i\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\tau_\jmath}\boldsymbol{\Sigma}_i^{-1}\right]\epsilon_i \\ &= -\widetilde{\mathbf{N}}_k^T\boldsymbol{\Omega}\boldsymbol{\epsilon}\end{aligned}$$

$$\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\gamma}\partial\tau_\jmath} = \begin{cases} -\widetilde{\mathbf{E}}^T\boldsymbol{\Omega}\boldsymbol{\epsilon} & \boldsymbol{\gamma}=\boldsymbol{\delta} \\ \\ -\widetilde{\mathbf{T}}^T\boldsymbol{\Omega}\boldsymbol{\epsilon} & \boldsymbol{\gamma}=\mathbf{a}, \end{cases}$$

and

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

3.3 Expected information matrix

Let $d_{gi} = \mathrm{E}(\zeta_g^2(\delta_i)\delta_i)$ and $f_{gi} = \mathrm{E}(\zeta_g^2(\delta_i)\delta_i^2)$, 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 $\mathbf{W}^* = \mathrm{blockdiag}_{1\le i\le n}(\frac{4d_{gi}}{m_i}\mathbf{W}_i)$. For simplicity, let $\widetilde{\mathbf{Z}}_i = \begin{pmatrix} \mathbf{Z}_i & \widetilde{\mathbf{T}}_i \end{pmatrix}$ is an $(m_i \times (p+2))$ design matrix and $\widetilde{\boldsymbol{\alpha}}^T = \begin{pmatrix} \boldsymbol{\alpha}^T & \mathbf{a}^T \end{pmatrix}$, by calculating the expectation of the matrix $-\mathbf{L}_\mathrm{p}$, we obtain the $(p^* \times p^*)$ penalized expected information matrix given by

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

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

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

After some algebraic manipulations we find that the $\boldsymbol{\mathcal{I}}_\mathrm{p}(\boldsymbol{\theta})$ matrix have a block-diagonal structure of the form

$$\boldsymbol{\mathcal{I}}_\mathrm{p}(\boldsymbol{\theta}) = \mathrm{blockdiag}\Big(\boldsymbol{\mathcal{I}}_\mathrm{p}^{11}(\boldsymbol{\theta}),\boldsymbol{\mathcal{I}}_\mathrm{p}^{22}(\boldsymbol{\theta})\Big),$$

where

$$\boldsymbol{\mathcal{I}}_\mathrm{p}^{11}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\mathcal{I}}_\mathrm{p}^{\widetilde{\alpha}\widetilde{\alpha}}(\boldsymbol{\theta}) & \boldsymbol{\mathcal{I}}_\mathrm{p}^{\widetilde{\alpha}\beta}(\boldsymbol{\theta}) & \boldsymbol{\mathcal{I}}_\mathrm{p}^{\widetilde{\alpha}\delta}(\boldsymbol{\theta}) \\ \boldsymbol{\mathcal{I}}_\mathrm{p}^{\beta\widetilde{\alpha}}(\boldsymbol{\theta}) & \boldsymbol{\mathcal{I}}_\mathrm{p}^{\beta\beta}(\boldsymbol{\theta}) & \boldsymbol{\mathcal{I}}_\mathrm{p}^{\beta\delta}(\boldsymbol{\theta}) \\ \boldsymbol{\mathcal{I}}_\mathrm{p}^{\delta\widetilde{\alpha}}(\boldsymbol{\theta}) & \boldsymbol{\mathcal{I}}_\mathrm{p}^{\delta\beta}(\boldsymbol{\theta}) & \boldsymbol{\mathcal{I}}_\mathrm{p}^{\delta\delta}(\boldsymbol{\theta}) \end{pmatrix},$$

whose elements of the matrix are given by

$$\boldsymbol{\mathcal{I}}_\mathrm{p}^{\widetilde{\alpha}\widetilde{\alpha}}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{Z}^T\mathbf{W}^*\mathbf{Z} & \mathbf{Z}^T\mathbf{W}^*\widetilde{\mathbf{E}} \\ \widetilde{\mathbf{E}}^T\mathbf{W}^*\mathbf{Z} & \widetilde{\mathbf{T}}^T\mathbf{W}^*\widetilde{\mathbf{T}} \end{pmatrix}, \quad \boldsymbol{\mathcal{I}}_\mathrm{p}^{\widetilde{\alpha}\beta}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{Z}^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 & \dots & \mathbf{Z}^T\mathbf{W}^*\widetilde{\mathbf{N}}_s \\ \widetilde{\mathbf{T}}\mathbf{W}^*\widetilde{\mathbf{N}}_1 & \dots & \widetilde{\mathbf{T}}\mathbf{W}^*\widetilde{\mathbf{N}}_s \end{pmatrix},$$

$$\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\widetilde{\alpha}\delta}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{Z}^T\mathbf{W}^*\widetilde{\mathbf{E}} \\ \mathbf{Z}^T\mathbf{W}^*\widetilde{\mathbf{T}} \end{pmatrix}, \quad \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\delta\delta}(\boldsymbol{\theta}) = \widetilde{\mathbf{E}}^T\mathbf{W}^*\widetilde{\mathbf{E}} + \lambda_g\,\widetilde{\mathbf{E}}\,,$$

$$\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\beta}(\boldsymbol{\theta}) = \begin{pmatrix} \widetilde{\mathbf{N}}_1^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 + \lambda_1\mathbf{K}_1 & \dots & \widetilde{\mathbf{N}}_1^T\mathbf{W}^*\widetilde{\mathbf{N}}_s \\ \widetilde{\mathbf{N}}_2^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 & \dots & \widetilde{\mathbf{N}}_2^T\mathbf{W}^*\widetilde{\mathbf{N}}_s \\ \vdots & \ddots & \vdots \\ \widetilde{\mathbf{N}}_s^T\mathbf{W}^*\widetilde{\mathbf{N}}_1 & \dots & \widetilde{\mathbf{N}}_s^T\mathbf{W}^*\widetilde{\mathbf{N}}_s + \lambda_s\mathbf{K}_s \end{pmatrix}, \quad \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{\beta\delta}(\boldsymbol{\theta}) = \begin{pmatrix} \widetilde{\mathbf{N}}_1^T\mathbf{W}^*\widetilde{\mathbf{E}} \\ \widetilde{\mathbf{N}}_2^T\mathbf{W}^*\widetilde{\mathbf{E}} \\ \vdots \\ \widetilde{\mathbf{N}}_s^T\mathbf{W}^*\widetilde{\mathbf{E}} \end{pmatrix}$$

and

$$\boldsymbol{\mathcal{I}}_p^{22} = \sum_{i=1}^{n} \boldsymbol{\mathcal{I}}_{\mathrm{p}_i}^{\tau\tau}\,,$$

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

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

where $b_{i_{\jmath\ell}} = \mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\jmath}\big)\mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\ell}\big)$. It is important to note that, for both the observed and expected information matrix, the matrices that are under the main diagonal are obtained as the transpose of the matrices that are on the main diagonal.

3.4 Maximizing the penalized log-likelihood function

Because β_k's are an infinite-dimensional parameters we consider the maximum penalized likelihood estimate of $\boldsymbol{\theta}$, which leads to a natural cubic spline estimate of β_k's and natural thin-plate spline of g. Specifically, the value of $\boldsymbol{\theta}$ that maximizes $L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda})$ over $\boldsymbol{\Theta}$, denoted by $\widehat{\boldsymbol{\theta}}$, is named maximum penalized likelihood estimate (PMLE) and satisfies

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

The determination of the maximum penalized likelihood estimate $\widehat{\boldsymbol{\theta}}$ can be performed by considering successive maximizations as described, for instance, in Gourieroux and Monfort [1995], Chapter 7. Specifically, for $\boldsymbol{\lambda}$ fixed and $s=2$, for example, the solution $\widehat{\boldsymbol{\theta}}$ to the maximization problem

$$\max_{\theta\in\Theta} L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \max_{\alpha,\beta_1,\beta_2,a,\delta,\tau} L_{\mathrm{p}}(\boldsymbol{\alpha},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{a},\boldsymbol{\delta},\boldsymbol{\tau},\boldsymbol{\lambda})\,,$$

can be obtained via the following four-step procedure (see Ibacache-Pulgar et al. [2012]):

(a) Firstly, we maximize the function $L_{\mathrm{p}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ over $\boldsymbol{\alpha}$ by remaining fixed the parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}$ and $\boldsymbol{\tau}$. The maximum value, $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$, is attained for values of $\boldsymbol{\alpha}$ in a set $\mathcal{B}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$ depending on the parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \hat{\mathbf{a}}$ and $\boldsymbol{\tau}$. Thus, if $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$, the penalized log-likelihood value is

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

Here L_{p}^c is called the concentred penalized log-likelihood in $\boldsymbol{\alpha}$.

(b) Then, we maximize $L_{\mathrm{p}}^c(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda}) = L_{\mathrm{p}}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ over $\boldsymbol{\beta}_1$ by remaining $\boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}$ and $\boldsymbol{\tau}$ fixed. The maximum value, $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_1(\boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$, is attained for values of $\boldsymbol{\beta}_1$ in a set $\mathcal{F}_1(\boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$ depending on the parameter $\boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}$ and $\boldsymbol{\tau}$. Therefore, if $\boldsymbol{\beta}_1 \in \mathcal{F}_1(\boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$, the penalized log-likelihood value is

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

Here L_{p}^c is called the concentrated penalized log-likelihood in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_1$.

(c) Now, we maximize $L_{\mathrm{p}}^c(\boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda}) = L_{\mathrm{p}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ over $\boldsymbol{\beta}_2$ by remaining $\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}$ fixed. The maximum value, $\widehat{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_2(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$, is attained for values of $\boldsymbol{\beta}_2$ in a set $\mathcal{F}_2(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$ depending on the parameter $\boldsymbol{a}, \boldsymbol{\delta}$ and $\boldsymbol{\tau}$. Therefore, if $\boldsymbol{\beta}_2 \in \mathcal{F}_2(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau})$, the penalized log-likelihood value is

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

Here L_{p}^c is called the concentrated penalized log-likelihood in $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. We proceed in the same way for maximize $L_{\mathrm{p}}^c(\boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda}$ and $L_{\mathrm{p}}^c(\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$.

(d) Finally, we maximize $L_{\mathrm{p}}^c(\boldsymbol{\tau}, \boldsymbol{\lambda}) = L_{\mathrm{p}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\delta}}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ over $\boldsymbol{\tau}$. The maximum value, $\widehat{\boldsymbol{\tau}}$, is attained on a set $\mathcal{C}$ of $\boldsymbol{\tau}$ values.

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

3.5 Existence of the MPLE

To ensure the existence of the MPLE we have to study the concavity of the log-likelihood function $L_{\mathrm{p}}(\boldsymbol{\alpha},\, \boldsymbol{\beta}_1,\, \boldsymbol{\beta}_2,\, \boldsymbol{a},\, \boldsymbol{\delta},\, \boldsymbol{\tau},\, \boldsymbol{\lambda})$ in $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{a}$, $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$.
In effect, let $\boldsymbol{\Psi} = \mathrm{blockdiag}_{1\leq i\leq n}\big(\boldsymbol{\Psi}_i\big)$, with $\boldsymbol{\Psi}_i$ defined in 3.2.

(a') In the step (a), the concavity (in $\boldsymbol{\alpha}$) of $L_{\mathrm{p}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ is guaranteed if only if the matrix $\mathbf{L}_{\mathrm{p}}^{\alpha\alpha} = -\mathbf{Z}^T\boldsymbol{\Psi}\mathbf{Z} \leq 0$ (negative semidefinite) or, equivalently, if only if $-\mathbf{L}_{\mathrm{p}}^{\alpha\alpha} \geq 0$ (positive semidefinite). One has $-\mathbf{L}_{\mathrm{p}}^{\alpha\alpha} \geq 0$ if the matrix $\boldsymbol{\Psi} \geq 0$, that is, if $\boldsymbol{\Psi}_i \geq 0 \; \forall \; i = 1, \ldots, n$. This latter is guaranteed if only if $\boldsymbol{\Psi}_{1i} \geq 0$ or, equivalently, if $v_i' \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \geq 0 \; \forall \; i = 1, \ldots, n$.

(b') Then, in the step (b), one has concavity (in $\boldsymbol{\beta}_1$) of $L_{\mathrm{p}}^c(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ if only if the matrix $\mathbf{L}_{\mathrm{p}}^{\beta_1\beta_1} = -\big(\widetilde{\mathbf{N}}_1^T\boldsymbol{\Psi}\widetilde{\mathbf{N}}_1 + \lambda_1\mathbf{K}_1\big) \leq 0$ or, equivalently, if only if $-\mathbf{L}_{\mathrm{p}}^{\beta_1\beta_1} \geq 0$. Consequently, $-\mathbf{L}_{\mathrm{p}}^{\beta_1\beta_1} \geq 0$ if only 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 $\mathbf{K}_1 \geq 0$, we have that $\lambda_1\mathbf{K}_1 \geq 0$. On the other hand, $\widetilde{\mathbf{N}}_1^T\boldsymbol{\Psi}\widetilde{\mathbf{N}}_1 \geq 0$ if $\boldsymbol{\Psi} \geq 0$, that is, if $\boldsymbol{\Psi}_i \geq 0 \; \forall \; i = 1, \ldots, n$. This latter is guaranteed if only if $\boldsymbol{\Psi}_{1i} \geq 0$ or, equivalently, if $v_i' \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \geq 0 \; \forall \; i = 1, \ldots, n$.

(c') Analogously, in the step (c), one has concavity (in $\boldsymbol{\beta}_2$) of $L_{\mathrm{p}}^c(\boldsymbol{\beta}_2, \boldsymbol{a}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\lambda})$ if only if the matrix $\mathbf{L}_{\mathrm{p}}^{\beta_2\beta_2} = -\big(\widetilde{\mathbf{N}}_2^T\boldsymbol{\Psi}\widetilde{\mathbf{N}}_2 + \lambda_2\mathbf{K}_2\big) \leq 0$ or, equivalently, if only if $-\mathbf{L}_{\mathrm{p}}^{\beta_2\beta_2} \geq 0$. Consequently, $-\mathbf{L}_{\mathrm{p}}^{\beta_2\beta_2} \geq 0$ if only if $\widetilde{\mathbf{N}}_2^T\boldsymbol{\Psi}\widetilde{\mathbf{N}}_2 \geq 0$ and $\lambda_2\mathbf{K}_2 \geq 0$. Since λ_2 is a positive scalar and $\mathbf{K}_2 \geq 0$, we have that $\lambda_2\mathbf{K}_2 \geq 0$. On the other hand, $\widetilde{\mathbf{N}}_2^T\boldsymbol{\Psi}\widetilde{\mathbf{N}}_2 \geq 0$ if $\boldsymbol{\Psi} \geq 0$, that is, if $\boldsymbol{\Psi}_i \geq 0 \; \forall \; i = 1, \ldots, n$. This latter is guaranteed if only if $\boldsymbol{\Psi}_{1i} \geq 0$ or, equivalently, if $v_i' \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \geq 0 \; \forall \; i = 1, \ldots, n$.

(d') Finally, in the step (d), the concavity (in $\boldsymbol{\tau}$) of $L_{\mathrm{p}}^c(\boldsymbol{\tau}, \boldsymbol{\lambda})$ is guaranteed if only if $L_{\mathrm{p}}^{\tau\tau} = \partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}\;, \boldsymbol{\lambda})/\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T < 0$.

Therefore, if $\widehat{\boldsymbol{\theta}}$ is the solution of the procedure (a)-(d) then $\widehat{\boldsymbol{\theta}}$ is the MPLE of $\boldsymbol{\theta}$ if only if $\boldsymbol{\Psi}_{1i} \geq 0$ or, equivalently, if $v_i' \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \geq 0 \; \forall \; i = 1, \ldots, n$. In particular, we have that under the normal distribution, $v_i' = 0$, $v_i = 1$ and $\boldsymbol{\Sigma}_i > 0$, hence the concavity is guaranteed.

3.6 Derivations of the Fisher score and weighted back-fitting algorithms

Let $\boldsymbol{\beta}_0 = (\boldsymbol{\alpha}^T \boldsymbol{a}^T)$ and $\mathbf{N}_0 = \widetilde{\mathbf{Z}}$, whit $\widetilde{\mathbf{Z}}$ defined above, and consider for simplicity $\boldsymbol{\lambda}$ and $\mathbf{W}^*$ fixed. According to Berhane and Tibshirani [1998], the four-step procedure (a)-(d) described in section 3.4 can be solved, for $1 \leq k \leq s$, by using the following Fisher scoring algorithm:

$$\begin{pmatrix} \mathbf{I} & \mathbf{S}_0^{(u)}\widetilde{\mathbf{N}}_1 & \ldots & \mathbf{S}_0^{(u)}\widetilde{\mathbf{N}}_s & \mathbf{S}_0^{(u)}\widetilde{\mathbf{E}} \\ \mathbf{S}_1^{(u)}\widetilde{\mathbf{N}}_0 & \mathbf{I} & \ldots & \mathbf{S}_1^{(u)}\widetilde{\mathbf{N}}_s & \mathbf{S}_1^{(u)}\widetilde{\mathbf{E}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{S}_s^{(u)}\widetilde{\mathbf{N}}_0 & \mathbf{S}_s^{(u)}\widetilde{\mathbf{N}}_1 & \ldots & \mathbf{I} & \mathbf{S}_s^{(u)}\widetilde{\mathbf{E}} \\ \mathbf{S}_\delta^{(u)}\widetilde{\mathbf{N}}_0 & \mathbf{S}_\delta^{(u)}\widetilde{\mathbf{N}}_1 & \ldots & \mathbf{S}_\delta^{(u)}\widetilde{\mathbf{N}}_s & \mathbf{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_0^{(u+1)} \\ \boldsymbol{\beta}_1^{(u+1)} \\ \vdots \\ \boldsymbol{\beta}_s^{(u+1)} \\ \boldsymbol{\delta}^{(u+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_0^{(u)}\boldsymbol{\eta}^{(u)} \\ \mathbf{S}_1^{(u)}\boldsymbol{\eta}^{(u)} \\ \vdots \\ \mathbf{S}_s^{(u)}\boldsymbol{\eta}^{(u)} \\ \mathbf{S}_\delta^{(u)}\boldsymbol{\eta}^{(u)} \end{pmatrix}, \qquad (3.1)$$

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

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

and

$$\mathbf{S}_\delta^{(u)} = (\widetilde{\mathbf{E}}^T\mathbf{W}^*\widetilde{\mathbf{E}} + \lambda_g\widetilde{\mathbf{E}})^{-1}\widetilde{\mathbf{E}}^T\mathbf{W}^*.$$

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

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

and

$$\boldsymbol{\delta}^{(u+1)} = \mathbf{S}_\delta^{(u)}\left(\boldsymbol{\eta}^{(u)} - \sum_{l=0, l\neq k}^{s}\widetilde{\mathbf{N}}_l\boldsymbol{\beta}_l^{(u)}\right). \quad (3.2)$$

Note that the back-fitting algorithm (3.2) depend on the elliptical distribution through the weight matrizes $\mathbf{W}^*$. In particular, under normal distribution for which $v_i = 1$ and $d_{g_i} = \frac{m_i}{4}$, we have that $\mathbf{W}_v = \mathbf{W}^* = \text{blockdiag}_{1\leq i\leq n}\big(\mathbf{W}_i\big)$ and, therefore, $\boldsymbol{\eta} = \mathbf{y}$. In addition, the system of equations (3.1) is consistent and the back-fitting algorithm (3.2) converges to a solution for any starting values if the weight matriz involved is symmetric and defined positive (see, for instance, Berhane and Tibshirani [1998]). Additionally, we have that this solution is unique when there not concurvity in the data. In particular, for model (2.1) with smooth terms β_1 and β_2 but without the constant terms, $\boldsymbol{\alpha}$, we have follows considerations:

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 $(r_k \times 1)$

vector of ones, and defined as

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

3.7 Joint iterative process

The solution of the estimating equation system (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 maximum likelihood estimation of the parameter $\boldsymbol{\tau}$, which is equivalent to the following iterative process:

(i) Initialize:

(a) Fitting a TPSPVCM under normal errors to get $\boldsymbol{\beta}_\jmath^{(0)}$ ($\jmath = 0, 1, \ldots, s$) and $\boldsymbol{\delta}_0$.

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

(c) From the current value $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}_0^{(0)^T}, \boldsymbol{\beta}_1^{(0)^T}, \ldots, \boldsymbol{\beta}_s^{(0)^T}, \boldsymbol{\tau}^{(0)})^T$ 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 $\mathbf{W}_v^{(0)} = \text{blockdiag}_{1\le i\le n}\big(v_i^{(0)}\mathbf{W}_i^{(0)}\big)$, with $\mathbf{W}_i^{(0)} = \boldsymbol{\Sigma}_i^{(0)-1}$. Then, obtain

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

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

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

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

(iii) Step 2: For current values $\boldsymbol{\beta}_{\jmath}^{(u+1)}$ $(\jmath = 0, 1, \ldots, s)$ and $\boldsymbol{\delta}^{(u+1)}$, 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 $\boldsymbol{\beta}_{\jmath}^{(0)}$ $(\jmath = 0, 1, \ldots, s)$, $\boldsymbol{\delta}^{(0)}$ and $\boldsymbol{\tau}^{(0)}$ by $\boldsymbol{\beta}_{\jmath}^{(u+1)}$, $\boldsymbol{\delta}^{(u+1)}$ and $\boldsymbol{\tau}^{(u+1)}$, respectively, until convergence.

It is important to note that in the interactive process above the parameter estimates depend on the smoothing matrices $\mathbf{S}_k$'s and $\mathbf{S}_{\delta}$, the modified variable $\boldsymbol{\eta}$ and the partial residuals (see Equation 3.2). In addition, the weights $v_i's$ have an influence on the estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_k$'s and $\boldsymbol{\delta}$, and it can be shown that for some distributions of elliptical contours these are inversely proportional to the Mahalanobis distance, as for example for the t-Student distribution; (df=ν_i) the current weight $v_i^{(u)} = (\nu_i + m_i)/(\nu_i + \delta_i^{(u)})$, with $\delta_i^{(u)} = \delta_i \Big|_{\boldsymbol{\theta}^{(u)}}$, is inversely proportional to the distance between the observed value $\mathbf{y}_i$ and its current predicted value $\boldsymbol{\mu}_i^{(u)}$, so that outlying observations tend to have small weights in the estimation process (see, for instance, Ibacache-Pulgar and Paula [2011]). [0.25cm]

3.8 Inference on the parameters

In this section we discuss some aspects of inference in elliptical TPSPCVMs. In the first sections we describe a procedure for calculating the approximate standard errors of the parameter estimates and we derive one approximate standard error bands for the coefficients functions. In the following sections we present a discussion on degrees of freedom estimation and smoothing parameters selection.

3.8.1 Estimation of the surface g

In Section 2 we represent the surface $\mathbf{g}$ as a linear combination of the coefficient vectors $\boldsymbol{\delta}$ and $\mathbf{a}$. Considering the MPLEs obtained through the iterative process described above, this is, $\widehat{\boldsymbol{\delta}}$ and $\widehat{\mathbf{a}}$, we have that the MPLE $\widehat{\mathbf{g}}$ can be obtained as

$$\widehat{\mathbf{g}} = \mathbf{E}\widehat{\boldsymbol{\delta}} + \mathbf{T}^{T}\widehat{\mathbf{a}} .$$

Consequently, the estimator of the surface g is a natural thin- plate spline. Details of the conditions that guarantee this result are given, for example, in Green and Silverman [1994].

3.8.2 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

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}}) \approx \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{-1}(\boldsymbol{\theta})\big|_{\widehat{\boldsymbol{\theta}}} = \mathrm{blockdiag}\left(\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{11^{-1}}(\widehat{\boldsymbol{\theta}}), \boldsymbol{\mathcal{I}}_{\mathrm{p}}^{22^{-1}}(\widehat{\boldsymbol{\theta}})\right). \tag{3.3}$$

In particular, if we are interested in drawing inferences for $\boldsymbol{\alpha}$ and $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s)$, the approximate covariance matrices can be estimated by using the corresponding block-diagonal matrices obtained from $\boldsymbol{\mathcal{I}}_{\mathrm{p}}^{-1}$, 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

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

By using variance-covariance matrix (3.3) 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]:

$$\mathrm{SEB}_{\mathrm{approx}}(\beta_k(\mathrm{t}_l^0)) = \widehat{\beta}_k(\mathrm{t}_l^0) \pm 2\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k(\mathrm{t}_l^0))},$$

where $\mathrm{Var}(\widehat{\beta}_k(\mathrm{t}_l))$ is the lth principal diagonal element of the matrix (3.3), for $l = 1, \ldots, r$.

3.8.3 On degrees of freedom

Numerous attempts have been made in the literature to extend familiar notion and practices in parametric statistics to nonparametric estimation. One popular notion of such is the so-called "degrees of freedom" as a model complexity index in nonparametric regression.

In parametric statistics, the degrees of freedom code the dimensions of the prospective model space. The notion is not defined through the trace of any matrix, and in many setting here is no matrix to talk about yet degrees of freedom are indispensable in inference. The fact that the trace o the hat matrix in linear regression models matches the dimension of the model space is conceptually a coincidence. In the contex of nonparametric regression, model complexity depends on a variety of factors including the structure of the smoothing matrix (see, for instance, Hastie and Tibshirani [1990] and Gu [2013]). In the literature concerning additive models there are different definitions for the

degrees of freedom (df), depending on the context in which they are used. Similarly to the classic linear case, the degree of freedom associated with a given term is defined in terms of the trace of the smoother matrix. In the elliptical TPSPVCM, the degree of freedom df associated with the kth nonparametric component is given by (see, for instance, Hastie and Tibshirani [1990]

$$\begin{aligned}\mathrm{df}(\lambda_k) &= \mathrm{tr}\big(\widetilde{\mathbf{N}}_k\mathbf{S}_k\big)\\ &= \mathrm{tr}\Big\{\widetilde{\mathbf{N}}_k\Big(\widetilde{\mathbf{N}}_k^T\mathbf{W}^*\widetilde{\mathbf{N}}_k + \lambda_k\mathbf{K}_k\Big)^{-1}\widetilde{\mathbf{N}}_k^T\mathbf{W}^*\Big\}\\ &= \mathrm{tr}\Big\{\widetilde{\mathbf{N}}_k^T\mathbf{W}^*\widetilde{\mathbf{N}}_k\Big(\widetilde{\mathbf{N}}_k^T\mathbf{W}^*\widetilde{\mathbf{N}}_k + \lambda_k\mathbf{K}_k\Big)^{-1}\Big\}.\end{aligned}$$

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 $\mathrm{rank}(\widetilde{\mathbf{N}}_k^T) \le r_k$, then $\mathbf{Q}_{\widetilde{\mathbf{N}}_k} \ge 0$. Therefore, there exists a matrix $\mathbf{Q}^{1/2}_{\widetilde{\mathbf{N}}_k} \ge 0$ such that $\mathbf{Q}_{\widetilde{\mathbf{N}}_k} = \mathbf{Q}^{1/2}_{\widetilde{\mathbf{N}}_k}\mathbf{Q}^{1/2}_{\widetilde{\mathbf{N}}_k}$. Thus, we can write $\mathrm{tr}\big\{\widetilde{\mathbf{S}}_k\big\}$ as (Eilers and Marx, 1996)

$$\mathrm{tr}\big\{\widetilde{\mathbf{S}}_k\big\} = \sum_{j=1}^{r_k}\frac{1}{1+\lambda_k\ell_j}, \tag{3.4}$$

where ℓ_j, for $j = 1, \ldots, r_k$, are the eigenvalues of the matrix $\mathbf{Q}^{-1/2}_{\widetilde{\mathbf{N}}_k}\mathbf{Q}_{\lambda_k}\mathbf{Q}^{-1/2}_{\widetilde{\mathbf{N}}_k}$, for $k = 1, \ldots, s$.

On the other hand, the degree of freedom associated to thin-plate spline is given by

$$\begin{aligned}\mathrm{df}(\lambda_g) &= \mathrm{tr}\big(\widetilde{\mathbf{E}}\,\mathbf{S}_\delta\big)\\ &= \mathrm{tr}\Big\{\widetilde{\mathbf{E}}\Big(\widetilde{\mathbf{E}}^T\mathbf{W}^*\widetilde{\mathbf{E}} + \lambda_g\,\widetilde{\mathbf{E}}\Big)^{-1}\widetilde{\mathbf{E}}^T\mathbf{W}^*\Big\}\\ &= \mathrm{tr}\Big\{\widetilde{\mathbf{E}}^T\mathbf{W}^*\widetilde{\mathbf{E}}\Big(\widetilde{\mathbf{E}}^T\mathbf{W}^*\widetilde{\mathbf{E}} + \lambda_g\,\widetilde{\mathbf{E}}\Big)^{-1}\Big\}.\end{aligned}$$

Analogous to the previous case, considering $\mathbf{Q}_{\widetilde{\mathbf{E}}} = \widetilde{\mathbf{E}}^T\mathbf{W}^*\widetilde{\mathbf{E}}$ and $\mathbf{Q}_{\lambda_g} = \lambda_g\mathbf{E}$, and since $\mathbf{W}^* > 0$ and $\mathrm{rank}(\widetilde{\mathbf{E}}^T) \le n$, then $\mathbf{Q}_{\widetilde{\mathbf{E}}} \ge 0$. Therefore, there exists a matrix $\mathbf{Q}^{1/2}_{\widetilde{\mathbf{E}}} \ge 0$ such that $\mathbf{Q}_{\widetilde{\mathbf{E}}} = \mathbf{Q}_{\widetilde{\mathbf{E}}}^{1/2}\mathbf{Q}_{\widetilde{\mathbf{E}}}^{1/2}$. Thus, we can write $\mathrm{tr}\{\widetilde{\mathbf{S}}_\delta\}$ as

$$\mathrm{tr}\{\widetilde{\mathbf{S}}_\delta\} = \sum_{j=1}^{n} \frac{1}{1+\lambda_g \ell_j}, \tag{3.5}$$

where ℓ_j, for $j = 1, \ldots, n$, are the eigenvalues of the matrix $\mathbf{Q}_{\widetilde{\mathbf{E}}}^{-1/2}\mathbf{Q}_{\lambda_g}\mathbf{Q}_{\widetilde{\mathbf{E}}}^{-1/2}$. It is important to note that both $\mathrm{df}(\lambda_k)$ and $\mathrm{df}(\lambda_g)$ are inversely proportional to λ_k and λ_g, respectively.

3.8.4 Selecting an appropriate model

Under the elliptical TPSVCMM one has a total of $p + d + \mathrm{df}(\boldsymbol{\lambda})$ parameters to be estimated, with $\mathrm{df}(\boldsymbol{\lambda}) = \mathrm{df}(\lambda_g) + \sum_{k=1}^{s} \mathrm{df}(\lambda_k)$ denoting approximately the number of effective parameters involved in modelling of the nonparametric effects and thin-plate spline. In this case, the Akaike information criterion (AIC) [Akaike, 1973] or the Bayesian information criterion (BIC) [Schwarz et al., 1978] can be used for selecting an appropriate model. The idea is to minimize the function

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

where $L_{\mathrm{p}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\lambda})$ denotes the penalized log-likelihood function available at $\widehat{\boldsymbol{\theta}}$ for a fixed $\boldsymbol{\lambda}$. It is important mention that AIC is based on information theory and is useful for selecting an appropriate model given data with adequate sample size. An alternative version of the AIC criterion, denoted by AICc, was proposed by Hurvich et al. [1998] in the context of parametric linear regression and autoregressive time series. Recently, **?** adapted this criterion for the partially linear model with first-order autoregressive symmetric errors. Considering such proposals, we propose the AICc criterion as an alternative for the selection of models under the TPSVCM model, which is given by

$$\mathrm{AIC_c}(\boldsymbol{\lambda}) = \log\left\{\frac{\parallel \sqrt{\widehat{\mathbf{W}}_v}(\mathbf{y} - \widehat{\mathbf{y}}) \parallel^2}{n}\right\} + \frac{2\Big[\mathrm{tr}(\widehat{\mathbf{H}(\boldsymbol{\lambda})}) + 1\Big]}{n - \mathrm{tr}(\widehat{\mathbf{H}(\boldsymbol{\lambda})}) - 2} + 1,$$

where $\widehat{\mathbf{H}(\boldsymbol{\lambda})}$ is commonly called smoother matrix and is equivalent to the *hat* matrix defined in the class of parametric regression models which satisfies $\widehat{\mathbf{y}} = \widehat{\mathbf{H}(\boldsymbol{\lambda})}\mathbf{y}$.

3.8.5 Choosing the smoothing parameters

The determination of the parameters λ's and λ_g is a crucial part in the estimation process and different choice methods are available in the literature. In practice it is desirable to select the smoothing parameter using an objective method rather than visual inspection. In a sense, a data-driven choice of λ allows data to speak for themselves. Thus, it is not exaggerating to say that the choice o λ is the spirit and sou of nonparametric regression. Large values of smoothing parameters produce smoother curves (surface) while smaller values produce more wiggly curves (surface). In non-parametric regression, for example, when $\lambda \to \infty$, the penalty term dominates, forcing the second derivative of the function equal to zero everywhere and thus the solution is the least-squares line. At the other extreme, as $\lambda = 0$, the penalty term becomes unimportant and the solution tends to an interpolating twice-differentiable function. Similarly, in P-spline regression,where the problem of determining the number of knots and their location becomes less important. The choose of the smoothing parameters it is crucial to control the tradeoff between data fitting and smoothness of surface in the thin-plate spline context.

A measure that is popular in the spline smoothing literature is the generalized cross-validation measure GCV developed by Craven and Wahba [1978]. It was originally developed as a simpler version of the cross validation procedure that avoided the need to re-smooth n times. But it also has been found to be rather more reliable than cross-validation in the sense of having less of a tendency to under-smooth.

Some methods that can be used in the context of TPSVCMMs are the following

Generalized Cross Validation. Following to Green and Silverman [1994], we have that the criterion of generalized cross validation under TPSVCMM is given by

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{\| \sqrt{\widehat{\mathbf{W}_v}}(\mathbf{y} - \widehat{\mathbf{y}}) \|^2}{[1 - n^{-1}\mathrm{tr}(\widehat{\mathbf{H}(\boldsymbol{\lambda})})]}. \tag{3.6}$$

In this case, $\boldsymbol{\lambda}$ should be obtained by minimizing $\mathrm{GCV}(\boldsymbol{\lambda})$ for a grid of $\boldsymbol{\lambda}$ values.

Akaike Information Criterion. Another selection method is the Akaike Information Criterion (AIC); see, for instance, Akaike [1973], Buja et al. [1989] and Durbán et al. [2003]. Under our model, the AIC takes the form

$$AIC(\boldsymbol{\lambda}) = -L_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \frac{2(tr(\widehat{\boldsymbol{H}(\boldsymbol{\lambda})}))}{n}.$$

Alternatively, Hurvich et al. [1998] proposed a modified version of the AIC designed to avoid overfitting in nonparametric regression. In our case, the AIC is given by

$$AIC_c(\boldsymbol{\lambda}) = L_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) + 1 + \frac{2(tr(\boldsymbol{H}(\boldsymbol{\lambda})) + 1)}{n - tr(\widehat{\boldsymbol{H}(\boldsymbol{\lambda})}) - 2}.$$

For the three methods describes above, the idea is to minimize criteria GCV, AIC and AIC_c.

In the previous sections the smoothing parameters λ_k's were assumed fixed. However, in practice situations the smoothing parameters should be selected from the data. When a smoothing spline is used, for example, it is usual to consider the cross-validation method or the generalized cross-validation method [Craven and Wahba, 1978]. Alternatively, these parameters may be selected by applying the Akaike information criterion (AIC) [Akaike, 1973] or the Bayesian information criterion (BIC) [Schwarz et al., 1978]. More details can be found, for example, in Hurvich et al. [1998], Buja et al. [1989], Simonoff and Tsai [1999] and Rigby and Stasinopoulos [2005].

In particular, we can consider the AIC($\boldsymbol{\lambda}$) or AIC$_c$($\boldsymbol{\lambda}$) criteria defined in the previous section, and use the effective degrees of freedom involved in nonparametric modeling to select appropriate smoothing parameters. Specifically, we will apply 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 m' values $u_{g_\ell} \in (0,1)$, and obtain the smoothing parameter values $\lambda_{k_\ell} = u_{k_\ell}/(1-u_{k_\ell})$ and $\lambda_{g_\ell} = u_{g_\ell}/(1-u_{g_\ell})$, for $\ell = 1,\dots,m$.
- (i.2) From the equations (3.4) and (3.5) obtain $\mathrm{df}_{k_\ell} = \mathrm{df}(\lambda_{k_\ell})$ and $\mathrm{df}_{g_\ell} = \mathrm{df}(\lambda_{g_\ell})$, and perform a dispersion graph between λ_{k_ℓ} and df_{k_ℓ} (λ_{g_ℓ} and df_{g_ℓ}).

(ii) Select a range for the smoothing parameters.

- (ii.1) Obtain an appropriate regression obtaining the fitted equation $\widehat{\lambda}_{k_\ell} = \psi(\mathrm{df}_{k_\ell})$ and $\widehat{\lambda}_{g_\ell} = \psi(\mathrm{df}_{g_\ell})$, where $\psi(\cdot)$ denotes the regression function.
- (ii.2) Since the relationship between λ_k and df_k (λ_g and df_g) is monotonically decreasing we may obtain from the fitted regression a range $[\lambda_k^{L_k}, \lambda_k^{U_k}]$ for λ_k and $[\lambda_g^{L_g}, \lambda_g^{U_g}]$ for λ_g 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 $\lambda_k^{L_k} = \psi(2)$. Analogously, we can consider the range $[2,16]$ we have that $\lambda_g^{U_g} = \psi(16)$ and $\lambda_g^{L_g} = \psi(2)$.

(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 $[\lambda_g^{L_g}, \lambda_g^{U_g}]$, and choosing the smoothing parameters values λ_k and λ_g that minimizes

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

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

3.8.6 Residual Analysis

In this work we propose a standardized residual which can be used to detect error distribution misspecification as well as the presence of outlying observations. We define the vector of fitted values as $\widehat{\mathbf{y}} = \widehat{\mathbf{H}(\boldsymbol{\lambda})}\mathbf{y}$, where the principal diagonal elements of $\mathbf{H}(\boldsymbol{\lambda})$ obtained in the last iteration of the iterative process, denoted by $h_{ii}(\boldsymbol{\lambda})$, are called leverage points and play an important role in the construction of diagnostic techniques. Consequently, the vector of ordinary residuals is given by $\mathbf{r} = \big\{\mathbf{I} - \mathbf{H}(\boldsymbol{\lambda})\big\}\mathbf{y}$, where the ith ordinary residual takes the form $\mathrm{r}_i = \{1 - h_{ii}(\boldsymbol{\lambda})\}\, \mathrm{y}_i$, whose approximate variance is given by $\mathrm{Var}_{\mathrm{approx}}(\mathrm{r}_i) = \xi\phi\,\{1 - h_{ii}(\boldsymbol{\alpha})\}^2$. Since the conditional residuals r_i's have different variances it is necessary to define a standardized version that allows a comparison between them. A natural definition is to divide each conditional residual by its approximate standard error. Thus, the standardized version of the residual r_i is given by

$$\mathrm{r}_i^* = \frac{\mathrm{r}_i}{\sqrt{\mathrm{Var}_{\mathrm{approx}}(\mathrm{r}_i)}}. \tag{3.7}$$

Because the distribution of the residuals r_i^*'s is not known, we can generate envelopes, as suggested by Atkinson (1981), in order to know the empirical distribution of r_i^* and thus detect misspecifications of the error distribution as well as the presence of outlying observations. Further details on the analysis of residuals in the semiparametric context can be found, for example, in Ibacache-Pulgar et al. [2013].

4

Local influence measure

In this chapter we present the local influence method and derive the perturbation matrix for different perturbation schemes under TPSCVMM.

4.1 The method

The interest of the local influence method is to investigate the behavior of some influence measure when perturbations are made in the model (or data).

Let $\boldsymbol{\omega}$ be a vector $(n \times 1)$ of perturbations restricted to some open subset $\Omega \in \mathbb{R}^n$ and the perturbed log-likelihood function is denoted by $L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})$. It is assumed that exist $\boldsymbol{\omega}_0 \in \Omega$, a vector of no perturbation, such that $L_p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega}_0) = L_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$. To assess the influence of minor perturbations on the maximum penalized likelihood estimate $\widehat{\boldsymbol{\theta}}$, one may considerer the likelihood displacement

$$LD(\boldsymbol{\omega}) = 2[L_p(\widehat{\boldsymbol{\theta}}, \boldsymbol{\lambda}|\boldsymbol{\omega}) - L_p(\widehat{\boldsymbol{\theta}}_\omega, \boldsymbol{\lambda})] \geq 0,$$

where $\hat{\boldsymbol{\theta}}_\omega$ denotes MPLE under $\boldsymbol{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})$. 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 behaviour 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 $a \in \mathbb{R}$. 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 $C_{\ell_{max}}$. The index plot of $\boldsymbol{\ell}_{max}$ may reveal those observations that under small perturbations exercise influence on $LD(\boldsymbol{\omega})$.

According to [Cook, 1986], the normal curvature in the unitary direction $\boldsymbol{\ell}$ is given by $C_\ell(\boldsymbol{\theta}) = -2\{\boldsymbol{\ell}^T \boldsymbol{\Delta}_p^T \boldsymbol{L}_p^{-1} \boldsymbol{\Delta}_p \boldsymbol{\ell}\}$, where

$$\boldsymbol{L}_p = \frac{\partial^2 L_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \; |_{\widehat{\boldsymbol{\theta}}}$$

and

$$\boldsymbol{\Delta}_p = \frac{\partial^2 L_p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\omega})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^T} \,|_{\widehat{\boldsymbol{\theta}}, \boldsymbol{\omega}_0}.$$

$C_\ell(\boldsymbol{\theta})$ denotes the local influence on the estimate $\widehat{\boldsymbol{\theta}}$ after perturbing the model or data. Escobar and Meeker Jr [1992] proposed to study the normal curvature in the direction $\boldsymbol{\ell} = \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is an n-dimensional vector with ones in the ith position and zeros in the remaining positions. In this case, the normal curvature, called total local influence of the ith individual, takes the form $C_{e_i}(\boldsymbol{\theta}) = 2|C_{ii}|$ $(i = 1, \ldots, n)$, where c_{ii} is the ith principal diagonal element of the matrix $C = \boldsymbol{\Delta}_p^T \boldsymbol{L}_p^{-1} \boldsymbol{\Delta}_p$. Verbeke and Molenberghs proposed as cutoff criteria to discriminate whether an observation is influential or not, $\boldsymbol{C}_i > 2\bar{\mathcal{C}}$, where $\bar{\mathcal{C}}$ is mean of $\mathcal{C} = \{C_i = C_{e_i}(\boldsymbol{\theta}) : i = 1, \ldots, n\}$.

Finally, consider the partition of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are subvectors of dimensions s and $(p^* - s)$, respectively. From Cook [1986] the normal curvature for $\boldsymbol{\theta}_1$ in the unitary direction $\boldsymbol{\ell}$ is given by

$$C_\ell(\boldsymbol{\theta}_1) = -2\{\boldsymbol{\ell}^T \boldsymbol{\Delta}_p^T (\boldsymbol{L}_p^{-1} - \boldsymbol{G}_{22}) \boldsymbol{\Delta}_p \boldsymbol{\ell}\},$$

where

$$\boldsymbol{G}_{22} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{L}_{p22}^{-1} \end{pmatrix}$$

with $\boldsymbol{L}_{p_{22}}$ obtained from the partition of $\boldsymbol{L}_p$ according to the partition of $\boldsymbol{\theta}$. In this case, the index plot of the eigenvector $\boldsymbol{\ell} = \boldsymbol{\ell}_{max}$ which corresponds to the largest absolute eigenvalue of the matrix $\boldsymbol{G} = \boldsymbol{\Delta}_p^T (\boldsymbol{L}_p^{-1} - \boldsymbol{G}_{22}) \boldsymbol{\Delta}_p$, may indicate those observations with large influence on $\widehat{\boldsymbol{\theta}}_1$. Alternatively, we can inspect the normal curvature $\boldsymbol{C}_\ell(\boldsymbol{\theta}_1)$ in the direction $\boldsymbol{\ell} = \boldsymbol{e}_i$.

4.1.1 Conformal normal curvature

In order to have a curvature invariant under uniform change of scale Poon and Poon [1999] proposed the conformal normal curvature defined as

$$B_\ell(\boldsymbol{\theta}) = \frac{C_\ell(\boldsymbol{\theta})}{2\sqrt{\mathrm{tr}(\boldsymbol{\Delta}_\mathrm{p}^T \mathbf{L}_\mathrm{p}^{-1} \boldsymbol{\Delta}_\mathrm{p})^2}} = -\frac{\boldsymbol{\ell}^T \boldsymbol{\Delta}_\mathrm{p}^T \mathbf{L}_\mathrm{p}^{-1} \boldsymbol{\Delta}_\mathrm{p} \boldsymbol{\ell}}{\sqrt{\mathrm{tr}(\boldsymbol{\Delta}_\mathrm{p}^T \mathbf{L}_\mathrm{p}^{-1} \boldsymbol{\Delta}_\mathrm{p})^2}}\ .$$

This curvature is characterized to allow for any unitary direction $\boldsymbol{\ell}$ that $0 \leq B_\ell(\theta) \leq 1$. A suggestion is to consider the direction $\boldsymbol{\ell} = \boldsymbol{\ell}_{max}$ corresponding to the largest curvature $B_{\ell_{max}}(\boldsymbol{\theta})$ or, alternatively, evaluating the normal curvature at the direction $\boldsymbol{\ell} = \mathbf{e}_i$ and observing the index plot of $B_{\mathbf{e}_i}(\boldsymbol{\theta})$.

4.1.2 Normal curvature derivation

In this subsection we present the expressions of the elements of the $(p^* \times n)$ $\boldsymbol{\Delta}_{\mathrm{p}}$ 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

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \, L_i(\boldsymbol{\theta}) - \sum_{k=1}^{s} \frac{\lambda_k}{2} \, \boldsymbol{\beta}_k^T \mathbf{K}_k \boldsymbol{\beta}_k \, , \tag{4.1}$$

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 (4.1) with respect to the elements of $\boldsymbol{\theta}$ and ω_i, we obtain the expressions

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

and

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

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

Scale perturbation

Under the scale parameter perturbation scheme we assume that

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

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

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

and

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

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

Explanatory variable perturbation

Here the kth 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 kth 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 (2.4) with $\mathbf{Z}_i$ replaced by $\mathbf{Z}_{i\omega}$, that is,

$$L_{\mathrm{P}}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})=L(\boldsymbol{\theta}|\boldsymbol{\omega})-\sum_{k=1}^{s}\frac{\lambda_k}{2}\,\boldsymbol{\beta}_k^T\mathbf{K}_k\boldsymbol{\beta}_k\,, \tag{4.2}$$

where $L(\cdot)$ is given by (**??**) 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 $\boldsymbol{\mu}_i=\mathbf{Z}_{i\omega}\boldsymbol{\alpha}+\sum_{k=1}^{s}\widetilde{\mathbf{N}}_{ki}\boldsymbol{\beta}_k$. In this case, we have $\boldsymbol{\omega}_0=\mathbf{0}\in\mathcal{R}^{n^*}$, with $n^*=\sum_{i=1}^{n}m_i$. Differentiating (4.2) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\omega}_i$, we obtain

$$\left.\frac{\partial^2 L_{\mathrm{P}_i}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\omega}_i^T}\right|_{\widehat{\boldsymbol{\theta}},\,\boldsymbol{\omega}_0} = -2v_i'\mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i\mathbf{r}_i^T\boldsymbol{\Sigma}_i^{-1}\alpha_t-v_i\left\{\mathbf{Z}_i^T\alpha_t-\mathbf{c}_t\mathbf{r}_i^T\right\}\boldsymbol{\Sigma}_i^{-1}\Big|_{\widehat{\boldsymbol{\theta}},\,\boldsymbol{\omega}_0},$$

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

and

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

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

5

Application

5.1 Application

In our application we will consider the house prices of Boston area reported by Harrison Jr and Rubinfeld [1978] and analyzed by many authors (see, for example, Belsley et al. [1980]; Ibacache-Pulgar et al. [2013]; and Ibacache-Pulgar and Reyes [2017]). This data set contains a sample of 506 observations collected by the U.S Census Service concerning housing in the area of Boston. The sample has variable attributes as housing values, variable of neighborhood, accessibility variables, besides an air pollution variable. A brief description of each variable is presented in 5.1. More details can be found in Harrison Jr and Rubinfeld [1978].

In this application we consider the variable LMV is related with four explanatory variables: LSTAT, ROOM, CRIM and TAX and the geographical coordinates expressed in longitude y latitude are added. Figure (5.1) represents the distribution of the spatial distribution of the LMV variable. In particular, from Figure (5.1) (right) we note that the lowest prices are concentrated between the latitudes 42.2 and 42.25 and longitudes between -71.0 and -71.1, while the highest prices are in the north part of the town.

The relationship between LMV and the explanatory variables is described in Figure 5.2. While the TAX seems linearly correlated with LMV Figure 5.2 a) the LSTAT variable shows a non linear behavior Figure 5.2 b).

On other hand, Figures 5.2 c) and Figure 5.2 d) suggest that the explanatory variables ROOM and CRIM might be interacting with the variable LSTAT in nonlinear fashion.

All this features suggest the application of a semiparametric model including non linear trends and the spatial variability. Specifically, we will assume the following thin-plate spline partially varying-coefficient model (TPSPVCM):

$$\mathrm{y}_i = \alpha_0 + \alpha_1 \, \mathrm{z}_i + \beta_1(r_i) \, \mathrm{x}_i^{(1)} + \beta_2(r_i) \, \mathrm{x}_i^{(2)} + g(\mathbf{t}_i) + \epsilon_i \qquad (i = 1, \ldots, 506) \, , \quad (5.1)$$

Symbol	Definition
LMV	logarithm of the median value of owner-occupied homes
CRIM	per capita crime rate by town
ZN	proportion of a townâs residential land zoned for lots greater than 25.000 square feet
INDUS	proportion of nonretail business acres per town
CHAS	Charles River dummy variable with value 1 if tract bounds on the Charles River
NOXSQ	nitrogen oxide concentration (parts per hundred million) squared
ROOM	average number of rooms squared
AGE	proportion of owner-occupied units built prior to 1940
DIST	logarithm of the weighted distances to five employment centers in the Boston region
RAD	logarithm of index of accessibility to radial highways
TAX	full-value property-tax rate (per 10.000 USD)
PTRATIO	pupil-teacher ratio by town
BLACK	$(Bk - 0.63)^2$ where Bk is the proportion of blacks in the population
LSTAT	logarithm of the proportion of the population that is lower status

Tab. 5.1: *definition of variables to analyze census data of Boston*

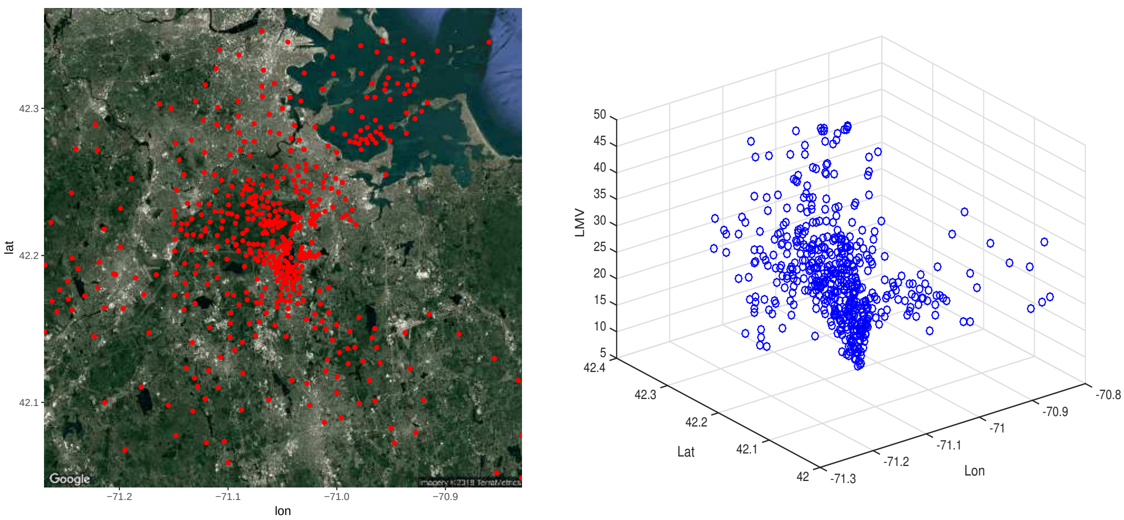

Fig. 5.1: *(Left) Google map of the Boston province. Red circles indicate the spatial distribution of the house prices data. (Right) Distributions of the LMV respect the longitude and the latitude.*

where y_i denotes the value of LMV in USD 1000, z_i denotes the value of TAX, $x_i^{(1)}$ denotes the value of CRIM, $x_i^{(2)}$ denotes the value of ROOM and r_i denotes the value of LSTAT from the ith experimental unit, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$ denotes the parameters vector associated to parametric component, $\beta_k(\cdot)$ $(k = 1, 2)$ are unknown functions, $g(\cdot)$ is a smooth surface that depends of the vector of coordinates $\mathbf{t}_i = (t_{1_i}, t_{2_i}) \in \mathcal{R}^2$, and ϵ_i are independent random errors that follow a symmetric distribution whit location parameter 0, scale pa-

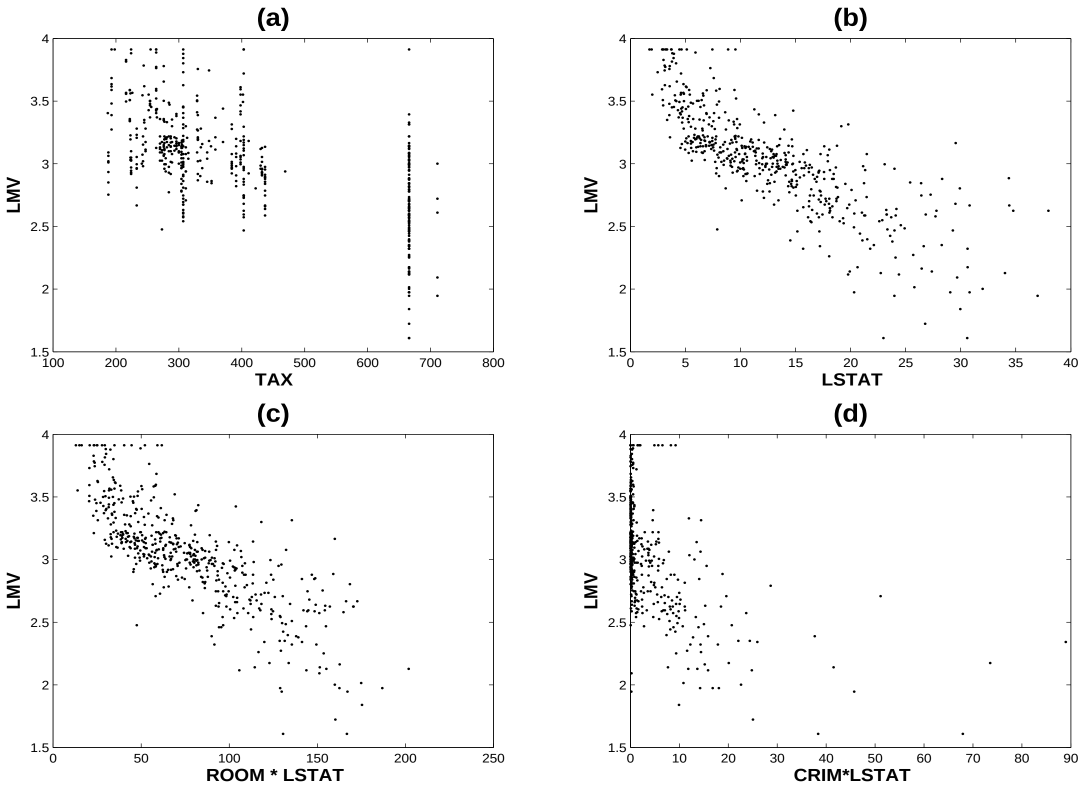

Fig. 5.2: *Scatter plots: (a) LMV versus TAX, (b) LMV versus LSTAT, (c) LMV versus ROOM × LSTAT, (d) and CRIM × LSTAT.*

rameter ϕ and density generator function h.

5.1.1 Fitting the models

We will compare in the sequel the fits based on normal and Student-t errors. The degrees of freedom ν for the Student-t model was selected by Akaike information criterion (AIC), this is, by defining a grid of values for ν and choosing the one that maximizes the AIC. Figure 5.3 shows the graph of AIC values for different degrees of freedom. We can see that this criterion is minimized for a value of $\nu = 4$. The MPLE estimates,

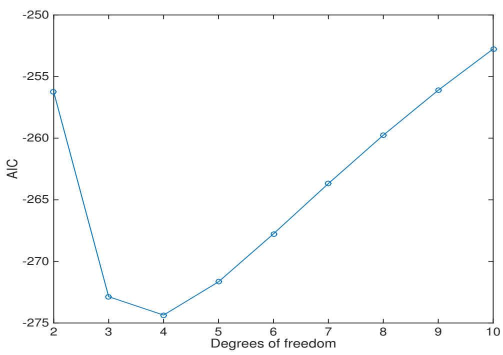

Fig. 5.3: *AIC values for different degrees of freedom ($\nu = 2, \ldots, 10$)*

estimated standard errors and the corresponding AIC for the model (5.1) under normal and Student-t distributions are presented in Table 5.2. Comparing these results, we may

Tab. 5.2: *Maximum penalized likelihood estimates, estimated standard errors and AIC values under normal and Student-t ($\nu = 4$) models fitted to house prices data.*

	Normal		Student-t	
	Estimate	SE	Estimate	SE
α_1	3.0668	0.1145	3.0637	0.0964
α_2	-0.0003	0.0001	-0.0002	0.0001
ϕ	0.0344	0.0022	0.0172	0.0372
AIC	-218.7586		-274.6546	

notice a similarity between the estimates $\widehat{\boldsymbol{\alpha}}$ under both models, but the standard error for $\widehat{\alpha}_1$ appears to be smaller under the Student-t model. On the other hand, it can be seen that the scale parameters are different for the two fitted models, but the estimates are not comparable since they are on different scales. Additionally, we may notice that the AIC value under the Student-t model is smaller than the one under the normal model, indicating that the models with longer-than-normal tails seem to better fit the data, a fact that is also confirmed through the QQ-plot presented in figure 5.4 (left panels). The standarized residuals plot provide in Figure 5.4 (right panels) is used to verify if there are outlying observations. The estimated coefficients functions β_1 and β_2 are computed

using the smoothing parameters obtained by the method described in Section 3.4. Figure 5.5 shows the estimated coefficient functions under both models and their corresponding approximate standard error band (dashed curves). The figures suggest clearly that the coefficient curves vary with the explanatory variable LSTAT. In addition, it can be seen that the functions estimated under the normal model have a higher smoothness compared to those obtained from the Student-t model.

It is important to remember that in this work we have incorporated the spatial variability of the data in the modeling process. Comparing with the results obtained by Ibacache-Pulgar and Reyes (2018), we can notice that the TPSPVCM model significantly improves the quality of the adjustment compared with the PVCM model. For example, for normal TPSPVCM model, the AIC value is -218.7586, while that for normal PVCM model the AIC value is -139.4998. Analogously, under Student-t TPSPVCM model with 4 degrees of freedom, the AIC value is -274.6546, while that under Student-t PVCM model with 5 degrees of freedom, the AIC value is -188.3909. In addition, we can notice that for the normal model, the estimated functions differ significantly, while under the Student-t model, they retain the same tendency but with a greater degree of smoothness. Figure 5.6

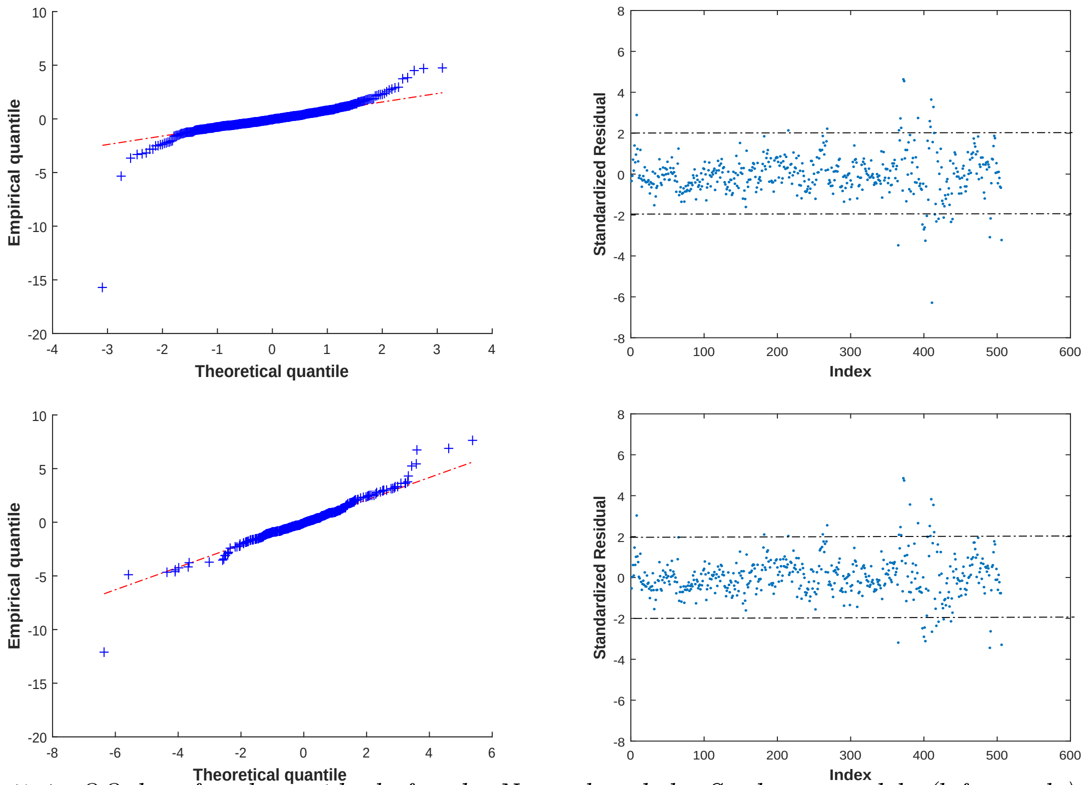

Fig. 5.4: *QQplots for the residuals for the Normal and the Student-t models (left panels). Plots standarized residual for the Normal and the Student-t models (right panels).*

displays the graphics of the LMV versus the fitted LMV from the two models. Although these plots indicate suitable fits for both models.

Figure (5.5) displays the graphics of the LMV versus the fitted LMV from the two models. Although these plots indicate suitable fits for both models, the Q-Q plots represented in Figure (5.6) show a better fitting for the t-Student model.

5.2 Concluding remarks

In this work a real data set previously analyzed under normal errors is reanalyzed under Student-t errors by assuming the the variability of the parameters and including a smoothing function for the spatial variability. By comparing the AIC values and graphics Q-Q plots for residuals en figure (5.6) of the two models, the t-Student showed the better fitting. Thus, we can recommend Student-t TPSPVCMs as an option to fit symmetric data sets with partially varying-coefficient and indications of heavy tails. The codes in MATLAB used in the application may be obtained from the authors by request.

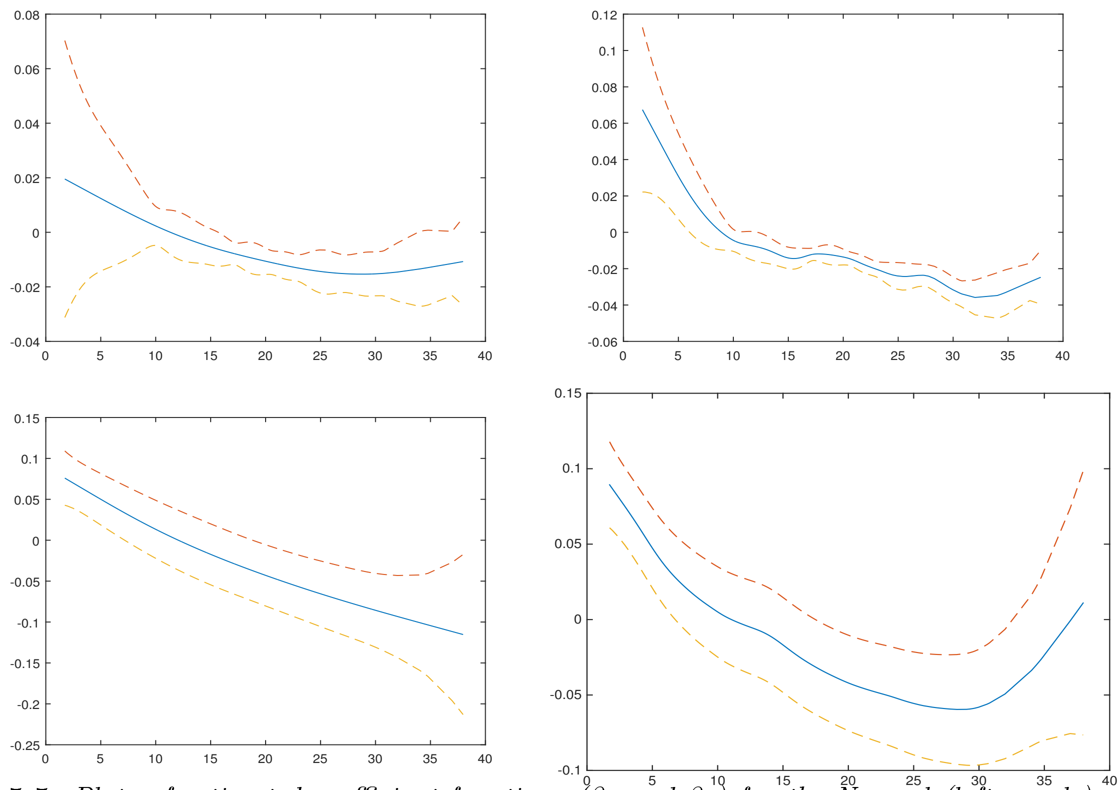

Fig. 5.5: *Plots of estimated coefficient functions (β_1 and β_2) for the Normal (left panels) and t-Student (right panels) models, and its approximate pointwise standard error band denoted by the dashed lines.*

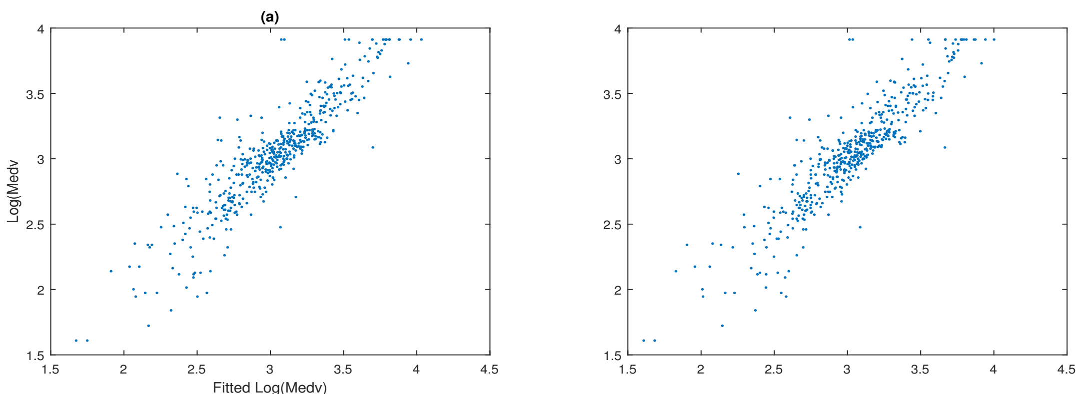

Fig. 5.6: *Scatter plots of LMV versus fitted LMV: normal (a) and Student-t models (b).*

6

Conclusions and future research

6.1 Conclusions

In this thesis we discuss parameter estimation and some statistical diagnostics for TPSVCMM under elliptical errors. Local influence approaches for the proposed model under case-weight, scale parameter and explanatory variable perturbations are developed. Closed form expressions are obtained for the penalized observed and expected information matrices. A real data set previously analyzed under normal errors is reanalyzed under Student-t errors by assuming the smoothing parameter fixed and by applying the Akaike information criteria to choose a degrees of freedom parameter estimate. The study provides evidences on the robust aspects of the MPLEs from Student-t TPSPVCM with small degrees of freedom against outlying observations, as pointed out by Ibacache-Pulgar et al. (2013) in the context of symmetric semiparametric additive models. However, these robust aspects do not seem to be extended to all perturbation schemes of the local influence approach, indicating the usefulness of the normal curvatures derived in this thesis for assessing the sensitivity of the MPLEs from the elliptical TPSPVCMs. Thus, we can recommend Student-t TPSPVCMs as an option to fit symmetric data sets with partially varying-coefficient and indications of heavy tails. The codes in MATLAB used in the application may be obtained from the authors by request.

6.2 Future research

We are considering to study some new aspects related to this thesis in a future work. For example,

(i) A future work perspective consists of developing other inferential aspects for the proposed model, for example, hypothesis testing and confidence intervals, as well as ANOVA.

(ii) For thin plate spline varying coefficient mixed model, derive the methodology based on other structures of data to apply local influence methods.

All the possible future works provide to us challenging aspects to be studied.

Bibliography

Akaike, H. (1973). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723. 32, 33, 34

Antoniadis, A., Gijbels, I., and Verhasselt, A. (2012). Variable selection in varying-coefficient models using p-splines. *Journal of Computational and Graphical Statistics*, 21(3):638–661. 6

Beckman, R. J., Nachtsheim, C. J., and Cook, R. D. (1987). Diagnostics for mixed–model analysis of variance. *Technometrics*, 29(4):413–426. 9

Berhane, K. and Tibshirani, R. J. (1998). Generalized additive models for longitudinal data. *Canadian Journal of Statistics*, 26(4):517–535. 26, 27

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598. 5

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510. 13, 33, 34

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

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. *Statistical models in S*, 2:309–376. 5

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18. 8

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–169. 9, 36, 37

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression.* New York: Chapman and Hall. 8

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403. 33, 34

Demidenko, E. (2013). *Mixed models: theory and applications with R.* John Wiley & Sons. 16

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38. 20

Díaz-García, J. A., Galea Rojas, M., and Leiva-Sánchez, V. (2003). Influence diagnostics for elliptical multivariate linear regression models. *Communications in statistics-Theory and Methods*, 32(3):625–641. 9

Duchon, J. (1975). Fonctions splines et vecteurs aleatoires. *Séminaire dÁnalyse Numérique, Univ. Sci. Médicale, Grenoble, France, Tech. Rep*, 213. 7

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. *Constructive Theory of Functions of Several Variables*, pages 85–100. 7

Durbán, M., Hackett, C. A., McNicol, J. W., Newton, A. C., Thomas, W. T., and Currie, I. D. (2003). The practical use of semiparametric models in field trials. *Journal of agricultural, biological, and environmental statistics*, 8(1):48–66. 6, 33

Durbán, M., Harezlak, J., Wand, M., and Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in medicine*, 24(8):1153–1167. 6

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102. 6

Escobar, L. A. and Meeker Jr, W. Q. (1992). Assessing influence in regression analysis with censored data. *Biometrics*, pages 507–528. 37

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179. 5

Fung, W.-K., Zhu, Z.-Y., Wei, B.-C., and He, X. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):565–579. 9

Galea, M., Bolfarine, H., and Vilca, F. (2005). Local influence in comparative calibration models under elliptical t-distributions. *Biometrical journal*, 47(5):691–706. 9

Galea, M., Paula, G. A., and Bolfarine, H. (1997). Local influence in elliptical linear regression models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(1):71–79. 9

Galea, M., Riquelme, M., and Paula, G. A. (2000). Diagnostic methods in elliptical linear regression models. *Brazilian Journal of Probability and Statistics*, pages 167–184. 8

Gourieroux, C. and Monfort, A. (1995). *Statistics and econometric models*, volume 1. Cambridge University Press. 24

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, pages 245–259. 20

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall, London. 6, 8, 13, 18, 30, 33

Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media. 31

Gu, C. and Wahba, G. (1993). Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 353–368. 8

Härdle, W. (1990). *Smoothing techniques: with implementation in S*. Springer-Verlag, New York. 6

Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102. 40

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Wiley Online Library. 6, 29, 30, 31

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796. 5

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction*, volume 1. Springer,New York, 2nd edition. 6

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293. 32, 33, 34

Ibacache-Pulgar, G. (2009). *Semiparametric additive mixed models with elliptic contours*. PhD thesis, IME-USP, Sao Paulo. 9

Ibacache-Pulgar, G. and Paula, G. A. (2011). Local influence for student-t partially linear models. *Computational Statistics & Data Analysis*, 55(3):1462–1478. 2, 9, 29

Ibacache-Pulgar, G., Paula, G. A., and Cysneiros, F. J. A. (2013). Semiparametric additive models under symmetric distributions. *Test*, 22(1):103–121. 2, 34, 35, 40

Ibacache-Pulgar, G., Paula, G. A., and Galea, M. (2012). Influence diagnostics for elliptical semiparametric mixed models. *Statistical Modelling*, 12(2):165–193. 2, 25

Ibacache-Pulgar, G. and Reyes, S. (2017). Local influence for elliptical partially varying-coefficient model. *Statistical Modelling*, page 1471082X17728770. 40

Kai-Tai, F. and Yao-Ting, Z. (1990). *Generalized multivariate analysis*. Science Press Beijing and Springer-Verlag, Berlin. 3, 15

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 419–430. 3

Kim, C., Park, B. U., and Kim, W. (2002). Influence diagnostics in semiparametric regression models. *Statistics & probability letters*, 60(1):49–58. 9

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974. 12

Lange, K. L., Little, R. J., and Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896. 15, 23

Li, Z., Xu, W., and Zhu, L. (2009). Influence diagnostics and outlier tests for varying coefficient mixed models. *Journal of multivariate analysis*, 100(9):2002–2017. 9, 13

Liu, S. (2000). On local influence for elliptical linear models. *Statistical Papers*, 41(2):211–224. 9

Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 30(2):292–304. 7

Osorio, F. (2006). *Influence diagnostics in elliptical mixed-effect models.* PhD thesis, IME-USP. 9

Osorio, F., Paula, G. A., and Galea, M. (2007). Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics & Data Analysis*, 51(9):4354–4368. 2, 9

Ouwens, M. J., Tan, F. E., and Berger, M. P. (2001). Local influence to detect influential data structures for generalized linear mixed models. *Biometrics*, 57(4):1166–1172. 9

Park, B. U., Mammen, E., Lee, Y. K., and Lee, E. R. (2015). Varying coefficient regression models: a review and new developments. *International Statistical Review*, 83(1):36–64. 5

Paula, G. A., Cysneiros, F. J. d. A., and Galea, M. (2003). Local influence and leverage in elliptical nonlinear regression models. *Proceedings of the 18th InternationalWorkshop on Statistical Modelling, Verbeke, G., Molenberghs, G., Aerts, A. and Fieuws, S.(Eds). Leuven: Katholieke Universiteit Leuven*, pages 361–365. 9

Pinheiro, J. C., Liu, C., and 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. 14

Poon, W.-Y. and Poon, Y. S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):51–61. 37

Rigby, R. A. and 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. 20, 34

Russo, C. M., Aoki, R., and Paula, G. A. (2012). Assessment of variance components in nonlinear mixed-effects elliptical models. *Test*, pages 1–27. 2

Russo, C. M., Paula, G. A., and Aoki, R. (2009). Influence diagnostics in nonlinear mixed-effects elliptical models. *Computational Statistics & Data Analysis*, 53(12):4143–4156. 2

Savalli, C., Paula, G. A., and Cysneiros, F. J. (2006). Assessment of variance components in elliptical linear mixed models. *Statistical Modelling*, 6(1):59–76. 2

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464. 32, 34

Segal, M. R., Bacchetti, P., and Jewell, N. P. (1994). Variances for maximum penalized likelihood estimates obtained via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 345–352. 30

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 (Methodological)*, pages 1–52. 30

Simonoff, J. S. and Tsai, C.-L. (1999). Semiparametric and additive model selection using an improved akaike information criterion. *Journal of Computational and Graphical Statistics*, 8(1):22–40. 34

Staniswalis, J. G. (1989). Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association*, 84(405):284–288. 6

Thomas, W. (1991). Influence diagnostics for the cross-validated smoothing parameter in spline smoothing. *Journal of the American Statistical Association*, 86(415):693–698. 9

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567. 6

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221. 9

Verbeke, G. and Molenberghs, G. Linear mixed models for longitudinal data.. 2000 new york. *NY Springer*. 1, 37

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):269–311. 6

Wahba, G. (1983). Bayesian" confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–150. 30

Wahba, G. (1990). *Spline models for observational data*. SIAM,Philadelphia, Pennsylvania. 6, 8

Wahba, G. and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly weather review*, 108(8):1122–1143. 7

Wand, M. and Ormerod, J. (2008). On semiparametric regression with o'sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50(2):179–198. 6

Wand, M. P. (2003). Smoothing and mixed models. *Computational statistics*, 18(2):223–249. 6

Wand, M. P. and Jones, M. C. (1999). *Kernel smoothing.* Chapman & Hall, Essen, Germany. 6

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the royal statistical society: Series b (statistical methodology)*, 60(1):159–174. 13

Wang, Y. (2011). *Smoothing splines: methods and applications.* CRC Press. 6

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114. 6

Wood, S. N. (2006). *Generalized additive models: an introduction with R.* Chapman & Hall. 6

Yanosky, J. D., Paciorek, C. J., Laden, F., Hart, J. E., Puett, R. C., Liao, D., and Suh, H. H. (2014). Spatio-temporal modeling of particulate air pollution in the conterminous united states using geographic and meteorological predictors. *Environmental Health*, 13(1):63. 13

Zhang, C.-x., Mei, C.-l., and Zhang, J.-s. (2007). Influence diagnostics in partially varying-coefficient models. *Acta Mathematicae Applicatae Sinica (English Series)*, 23(4):619–628. 9

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93(442):710–719. 13

Zhu, H.-T. and Lee, S.-Y. (2003). Local influence for generalized linear mixed models. *Canadian Journal of Statistics*, 31(3):293–309. 1, 9