UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

A robust lasso regression for linear mixed-effects models with diagnostic analysis

Rafael Rocha de Oliveira Garcia

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

Rafael Rocha de Oliveira Garcia

A robust lasso regression for linear mixed-effects models with diagnostic analysis

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Cibele Maria Russo Novelli

USP – São Carlos
December 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

G216r

Garcia, Rafael Rocha de Oliveira
A robust lasso regression for linear mixed-effects models with diagnostic analysis / Rafael Rocha de Oliveira Garcia; orientadora Cibele Maria Russo Novelli. -- São Carlos, 2021.
77 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Mixed models. 2. lasso. 3. Robust models. 4. Diagnostics. 5. Regression analysis. I. Novelli, Cibele Maria Russo, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Rafael Rocha de Oliveira Garcia

Regressão lasso robusta para modelos lineares de efeitos mistos com análise de diagnóstico

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Cibele Maria Russo Novelli

USP – São Carlos
Dezembro de 2021

ACKNOWLEDGEMENTS

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Profa. Dra. Cibele Maria Russo Novelli, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for funding this research under the project number 88882.461700/2019-01.

I would like to extend my thanks to the Dissertation Oral Defense Committee Profa. Dra. Camila Borelli Zeller and Prof. Dr. Juvêncio Santos Nobre.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear, and also my sister. You are always there for me. Finally, I could not have completed this dissertation without the support of my friends and colleagues, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

“Inductive inference is the only process known to us by which essentially new knowledge comes into the world.”

(Sir Ronald A. Fisher, The Design of Experiments, 1935)

“There is nothing like looking, if you want to find something.”

(J.R.R. Tolkien, The Hobbit, or There and Back Again, 1937)

“Not all those who wander are lost.”

(J.R.R. Tolkien, The Lord of the Rings, 1954)

RESUMO

GARCIA, R. R. O. **Regressão lasso robusta para modelos lineares de efeitos mistos com análise de diagnóstico**. 2021. 75 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Seleção de variáveis é um tópico de elevada importância para o processo de modelagem. A escolha do melhor conjunto de variáveis explicativas pode ser feita com o intuito de melhorar uma previsão ou facilitar a interpretação dos resultados. Contudo, os métodos para seleção de variáveis nem sempre são triviais, principalmente no contexto de modelos lineares de efeitos mistos. A seleção para esses modelos deve ser feita para os efeitos fixos, que estão relacionados a uma média global, e para os efeitos aleatórios, relacionados à variância a nível individual nesse contexto. São dois os tipos de abordagens para a seleção de variáveis em modelos de efeitos mistos: conjunta ou em dois estágios, havendo na literatura existente o processo de seleção conjunta via lasso para modelos lineares de efeitos-mistos normais. Outro tópico de elevada importância, é a análise de diagnóstico e resíduos. Enquanto as análises de resíduos são feitas para investigar problemas com o modelo ajustado e identificação de observações atípicas, uma análise de diagnóstico é feita assumindo o modelo como correto, e investigando a robustez das conclusões a pequenas perturbações dos dados e/ou no modelo. Para lidar com essas observações, são várias as alternativas. Uma delas, é a utilização de modelos robustos, os quais seriam ditos robustos a perturbações nos dados. Isto é, modelos que melhor se ajustam a conjuntos de dados que possuem pontos considerados como sendo outliers e/ou alavanca. Este trabalho tem como objetivo utilizar o método robusto para seleção de variáveis em modelos lineares de efeitos mistos e compará-lo com o método normal através de análise de diagnóstico.

Palavras-chave: Modelos mistos, lasso, Modelos robustos, Diagnóstico, Análise de regressão.

ABSTRACT

GARCIA, R. R. O. **A robust lasso regression for linear mixed-effects models with diagnostic analysis**. 2021. 75 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Variable selection has been a topic of great interest for statisticians and researchers alike. The choice of the best subset of predictors may be carried out with the objective of improving prediction or for easier interpretation of results. However, such methods are not always straightforward, mainly in the context of linear mixed-effects models. Variable selection for such models must be carried out for both fixed and random effects, the first being related to the global mean of data and the second to subject-level variance. There are two possible approaches when selecting variables for mixed-effects models: joint or two-stage procedures. In existing literature on the topic of variable selection for linear mixed-effects model, there is a method of joint selection via lasso for linear mixed-effects models under a normal distribution. Another topic of remarkable importance, is diagnostics and residual analysis. While residual analyses are carried out to assess issues with the fitted model and identification of atypical observations, diagnostic analyses are carried out assuming the model as correct and, assessing its conclusions robustness to small disturbances in the data and/or the model. There are many possible ways to deal with such observations. One is using robust models, which are said to be robust to disturbances in the data. That is, models that are better fit to data sets that possess observations considered to be as outliers and/or leverage. This work aims to use the robust method for variable selection in linear mixed-effects model and compare it with the normal method using diagnostic analysis.

Keywords: Mixed models, lasso, Robust models, Diagnostics, Regression analysis.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CASTNet	Clean Air Status and Trends Network
CV	coefficient of variation
GCV	generalized cross-validation
GLMM	generalized linear mixed models
GLS	generalized least squares
LASSO	Least Absolute Shrinkage and Selection Operator
LME	Linear Mixed-Effects
LV	Lesaffre-Verbeke
MAD	median absolute deviance (or deviation)
MLE	maximum likelihood estimator
MSPE	mean squared prediction error
U.S. EPA	Unites States Environmental Protection Agency

CONTENTS

CHAPTER

1

INTRODUCTION

Linear Mixed-Effects (LME) models may be used to explain both global and local changes, that is, if a data set consists of a repeated measures experiment, the researcher would be able to model both a global mean and subject-level variation. Aside from the interest of modeling the mean, interest may rely in selecting variables from the set of covariates. For instance, consider the data set in Bondell, Krishna and Ghosh (2010) which describes the association between the total nitrate concentration in the atmosphere and a set of measured predictors collected from 2000 to 2004. The data were obtained from fifteen of the Unites States Environmental Protection Agency (U.S. EPA) Clean Air Status and Trends Network (CASTNet) sites. The referred data set is unbalanced and consists of repeated measures of pollution from each of those sites and its characteristics suggest that a linear mixed-effects models may be adequate, where the random effects arise from the repeated sites measures.

Variable selection has been a topic of interest for statisticians for a long time. It can be used either as a dimensionality reduction tool or as a method to improve precision for prediction, as it is supposed that the subset of selected variables represent those with higher impact on the response (MILLER, 2002). There are two possible ways to tackle the problem of variable selection in LME models: joint selection and two-stage procedures. Bondell, Krishna and Ghosh (2010) points out that the usual methods for selecting variables in LME models work under the assumption that one of the effects is considered observed (for example, the fixed effects) and the selection is carried out for the other (the random effects, for example). Bondell, Krishna and Ghosh (2010) proposes an approach of joint variable selection based on the lasso method (TIBSHIRANI, 1996; JAMES *et al.*, 2013) adapted with weights (ZOU, 2006) and argue that the joint procedure is preferred because "changing the structure of one set of effects can lead to different choices of variables". The adaptive lasso is then used with a constrained expectation-maximization algorithm for estimation and selection of the parameters of interest. Pan and Shang (2017) points out that the main difference between one-stage and two-stage procedures is that the latter is more effective and stable, but an incorrect selection in the first step may lead to

sub-sequential errors. For an intensive review in variable selection in LME models, refer to Buscemi and Plaia (2019). Cruz (2020) also presents a thorough review on information criteria and LME model selection.

Throughout the process of fitting a regression model, the researcher may come across a few interesting observations that differ from the bulk of the data. According to Seber (2012), there are two kinds of observation points that should receive attention: those whose residuals are large and those whose explanatory variable values are far from the rest of the data. The latter are called *leverage* points and the former, *outliers*. Observation points which are both *high-leverage* and *outliers* are candidates to be *influential points* Seber (2012, Sec. 9.4). There are a few ways for dealing with such points (SEBER, 2012, Sec. 10.6), but the interest of this dissertation relies on robust methods such that models produced using these methods are less sensitive to outlying, leverage and influential observations. Sinha (2004) presents a robustified method for fitting generalized linear mixed models (GLMM), in which the author uses a function of the Mahalanobis distance with robust estimates of the location and scale of the explanatory variable in order to decrease the impact of possible leverage points. As for outliers, Fan, Qin and Zhu (2014) introduce a modification to the response by adding to it a function of the studentized residuals.

In the light of variable selection and robustness, Fan, Qin and Zhu (2014) proposes a robust method for joint variable selection in LME models. The authors, extend the joint method proposed by Bondell, Krishna and Ghosh (2010), along with the robustified log-likelihood proposed by Sinha (2004) to deal with leverage and outlier points.

The main objective of this dissertation is to present residual analysis and diagnostic techniques for lasso regression for Linear Mixed-Effects models. To illustrate that, a comparison of both fitting procedures proposed by Bondell, Krishna and Ghosh (2010) and Fan, Qin and Zhu (2014) will be carried out using a combination of residual analysis (NOBRE; SINGER, 2007; SINGER; ROCHA; NOBRE, 2017; SINGER; NOBRE; ROCHA, 2018) and diagnostics techniques for LME (SINGER; NOBRE; ROCHA, 2018) that will be modified to account for the lasso method (KIM *et al.*, 2015; RAJARATNAM *et al.*, 2019). It is worth noting that an expression for the Cook's distance that assess the lasso for LME models will be defined, following a modification from the original work, and an extension of Theorem 10.1 from Seber (2012, Sec. 10.2) is also presented, to account for the lasso regression for LME models.

In order to do that, an exploratory analysis of the said data set is presented, both methods proposed by Bondell, Krishna and Ghosh (2010) and Fan, Qin and Zhu (2014) are fitted to the data and, using residual analysis and diagnostics techniques, observations that may not be properly fitted by the methods will be sought out, such as outliers and leverage points, in order to determine which of the models better explains the data.

Notice that diagnostics and residual analysis in lasso regression is a relatively recent topic of research. Considering that, the main contribution of this dissertation is to present the

adapted diagnostics and residual analysis techniques to be used in the context of using lasso as a tool for estimating and selecting effects in LME models.

1.1 Motivation - exploratory analysis

In this section, an exploratory analysis of the data set for the total nitrate concentration and the logarithm of the total nitrate concentration is presented. The data set is unbalanced, as each of the sites have a different number of observations and consists of one response, ten observed explanatory variables and six other artificial variables as function of time to account for effects of time and seasonality, that are presented in Chart 1. Note that the values in Table 1 are for the original values of the response variable and the ones presented in Table 2 refer to the transformed response. Previous analyses of this data set have used the log transformed data in order to correct skewness (GHOSH *et al.*, 2010; BONDELL; KRISHNA; GHOSH, 2010).

Chart 1 – Variables

Y	$LOG(TNO)_3$, log of total nitrate concentration $(\mu mol/m^3)$	`log.nitrate`
x_1	$(SO)_4$ sulphate concentration $(\mu mol/m^3)$	`sulphate`
x_2	$(NH)_4$ ammonia concentration $(\mu mol/m^3)$	`ammonia`
x_3	$(O)_3$ maximum ozone (ppb, parts per billion)	`ozone`
x_4	(T) average temperature (^oC)	`atemp`
x_5	$(T)_d$ average dew point temperature (^oC)	`adptemp`
x_6	RH relative humidity $(\%)$	`humidity`
x_7	SR average solar radiation (W/m^2)	`radiation`
x_8	WS average wind speed (m/s)	`windspeed`
x_9	P total precipitation $(mm/month)$	`precipitation`
$l(t)$	Time of measurement in months $(1, \ldots, 60)$ from 2000 to 2004	`time.in.months`
$s_j(t)$	$\sin\left(\frac{2\pi jt}{12}\right)$, where $j = 1,2,3$	`s1, s2, s3`
$c_j(t)$	$\cos\left(\frac{2\pi jt}{12}\right)$, where $j = 1,2,3$	`c1, c2, c3`

Recall that the measurements were made between the years of 2000 and 2004. The sites, overall, seem to be symmetrical around each individual mean, and they do not seem to be overly dispersed. Note that observation six (site COW137), in both tables, seems to have the most different value, as it has the smallest mean in both the original and transformed data set. It is worth noting that site CDR119 presents the largest variance and standard deviation values.

Along with the summary statistics in Table 1, the graph on the left of Figure 1 also aids visualizing the response location and dispersion. Note that site COW137 draws attention, as it has a smaller mean than the other sites, and also smaller dispersion, when compared to the other sites. There are a few other observations that should be highlighted: ANA115 (observation one), CDR119 (observation three), DCP114 (observation eight), GAS153 (observation ten), PNF126 (observation twelve), SNH418 (observation 14) and VPI120 (observation fifteen).

Table 1 – Descriptive statistics for response variable for each site.

	Site	Mean	Median	SD	Var	CV	Min	Max	n
1	ANA115	3,807113	3,442840	1,462255	2,138189	38,408485	2,072776	8,089750	54
2	BEL116	3,147775	3,095050	0,625606	0,391383	19,874555	1,949033	4,488425	55
3	CDR119	1,671963	1,549075	0,742280	0,550979	44,395716	0,613790	3,763450	50
4	CKT136	2,967542	2,988785	0,917651	0,842083	30,922924	1,360867	5,410180	50
5	CND125	2,591234	2,525380	0,707815	0,501002	27,315752	1,379650	4,408120	59
6	**COW137**	**0,922525**	**0,929091**	0,296677	0,088017	32,159237	0,435395	1,757625	58
7	CTH110	2,808046	2,651607	0,887970	0,788491	31,622353	1,375870	5,397100	54
8	DCP114	4,434197	4,352800	1,001717	1,003436	22,590715	2,511325	6,967480	56
9	ESP127	2,208814	2,134960	0,819548	0,671659	37,103540	1,092524	3,925120	59
10	GAS153	2,426614	2,375138	0,530833	0,281784	21,875464	1,367700	3,706600	54
11	MKG113	3,308561	3,182104	0,893070	0,797574	26,992708	1,731775	6,401960	58
12	PNF126	1,932381	1,650075	1,242488	1,543776	64,298280	0,935090	8,382633	57
13	PSU106	3,530280	3,461240	0,929202	0,863416	26,320913	2,106875	6,283100	55
14	SHN418	2,814966	2,767135	0,611062	0,373396	21,707604	1,498850	4,611940	48
15	VPI120	2,666810	2,602520	0,690689	0,477052	25,899453	1,369990	4,960600	59
	Total	2,744613	2,648504	1,210237	1,464674	44,094997	0,435395	8,382633	826

Source: Research data.

Table 2 – Descriptive statistics for log transformed response variable for each site - transformed data.

	Site	Mean	Median	SD	Var	CV	Min	Max	n
1	ANA115	0,373612	0,332883	0,335950	0,112863	89,919511	-0,174521	1,187188	54
2	BEL116	0,223849	0,226394	0,199569	0,039828	89,153674	-0,236077	0,598092	55
3	**CDR119**	-0,483751	-0,465789	**0,441139**	**0,194604**	-91,191357	-1,391513	0,421926	50
4	CKT136	0,136019	0,191457	0,318747	0,101600	234,339578	-0,595288	0,784872	50
5	CND125	0,011371	0,022981	0,278285	0,077442	2447,415862	-0,581580	0,580038	59
6	**COW137**	-1,037164	-0,977304	0,334839	0,112117	-32,284092	-1,734912	-0,339447	58
7	CTH110	0,081804	0,071716	0,309984	0,096090	378,934809	-0,584324	0,782452	54
8	**DCP114**	0,561370	0,567355	0,223735	0,050057	39,855068	0,017400	1,037844	56
9	ESP127	-0,178053	-0,144962	0,369892	0,136820	-207,742694	-0,814919	0,463987	59
10	GAS153	-0,040538	-0,038372	0,220921	0,048806	-544,974210	-0,590280	0,406705	54
11	MKG113	0,258419	0,254132	0,265895	0,070700	102,892932	-0,354263	0,953194	58
12	**PNF126**	-0,342578	-0,402589	0,384887	0,148138	-112,350166	-0,970523	1,222752	57
13	PSU106	0,325815	0,338217	0,253869	0,064449	77,918004	-0,158204	0,934453	55
14	SHN418	0,107596	0,114308	0,224980	0,050616	209,096693	-0,498712	0,625239	48
15	VPI120	0,045402	0,053070	0,256233	0,065655	564,363794	-0,588607	0,698117	59
	Total	0,000000	0,070585	0,487734	0,237885	-	-1,734912	1,222752	826

Source: Research data.

As of Table 2 and the graph on the right of Figure 1, the summary statistics and the box-plots for the logarithm of the response for each site are presented. The transformed data will be the one used to fit both the normal and robust selection methods. A word of caution when analyzing the coefficient of variation (CV) column in Table 2: the mean values of all the sites are close to zero, which causes the CV values to be larger. Other than that, the values in this table may be used to better understand the response variable. Site COW137 once again draws attention; notice that site CDR119 stands out in a sense that it has larger variance, when compared to the

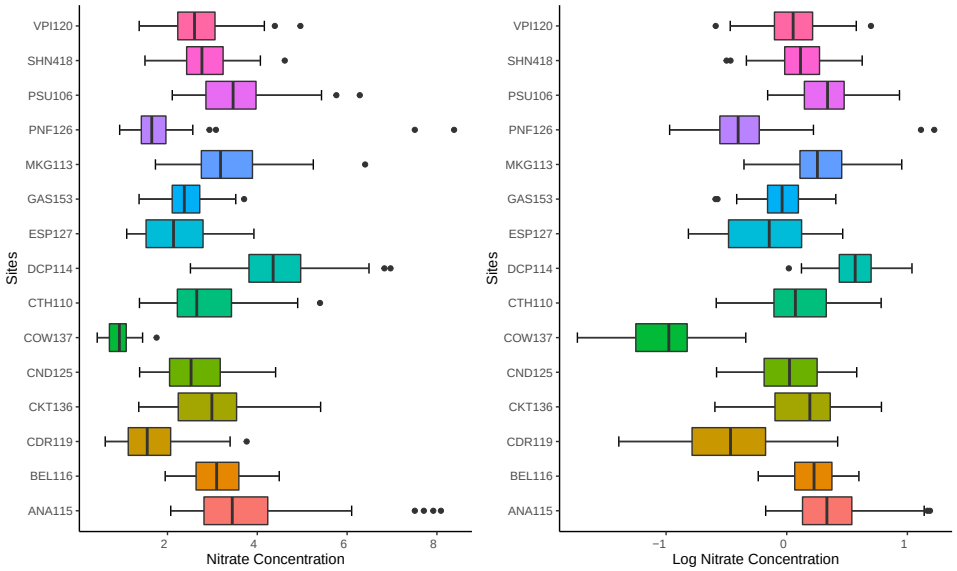

Figure 1 – Box-plot of response variable by subject.

Source: Research data.

other sites. After this first analysis of the response variable, some of the observations suspected as being outliers are: CDR119, COW137, DCP114 and PNF126. They are only suspected as being outliers; the residual analysis along with the diagnostics techniques later presented on the text will provide enough evidence in order to confirm whether or not they are in fact outliers.

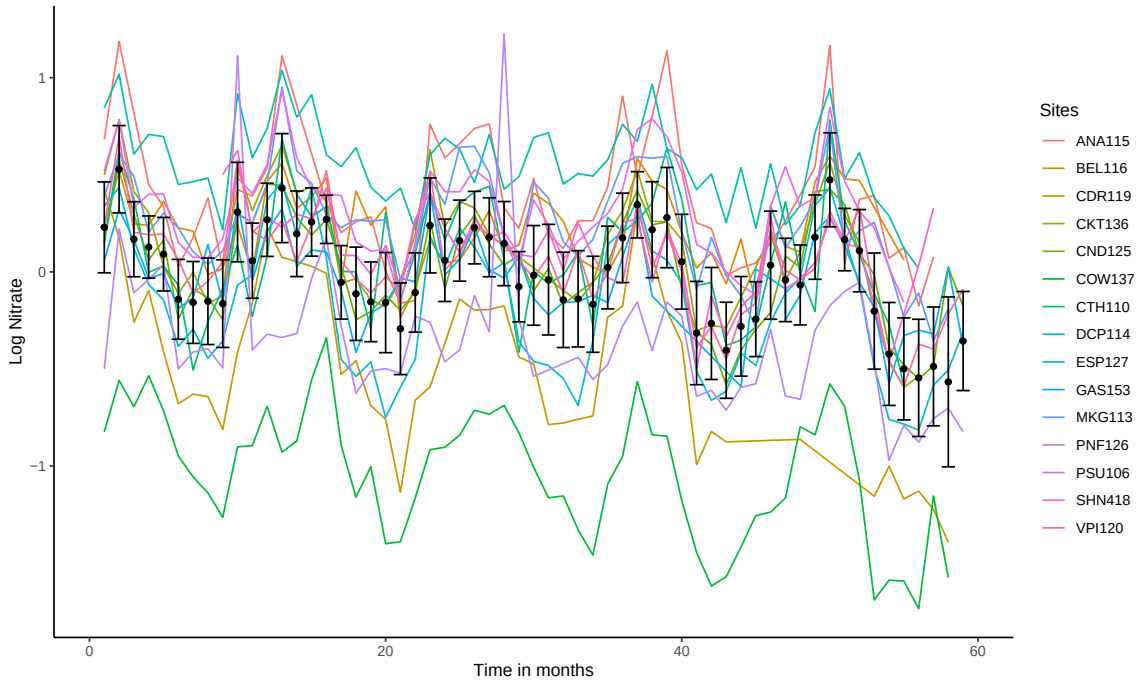

Figure 2 – Site profile plot over time.

Source: Research data.

The graphs in Figure 2 and Figure 3 present another point of view of the data, as they are

individual profile plots of the response variable over time. Figure 2 provides insights as how each site is dispersed around the global mean, as a well as the seasonal trend over time. The black dots represent the overall mean over time; the bars that accompany them are ± 2 standard-errors. Once again, the site COW137 draws attention, as its profile plot is lower than the other sites, further from the mean and not even in the interval constructed around the sample mean and the sample standard-deviation. In Figure 3 the same information presented as in the previous graph, but this time each site is separated, which provides easier visualization of each individual behaviour.

Other sites also draw attention upon analyzing both graphs, CDR119 for having a decreasing behavior and PNF126 for having two abnormal observations, for example. Site COW137, however, is the one that is consistently outside the intervals for each time.

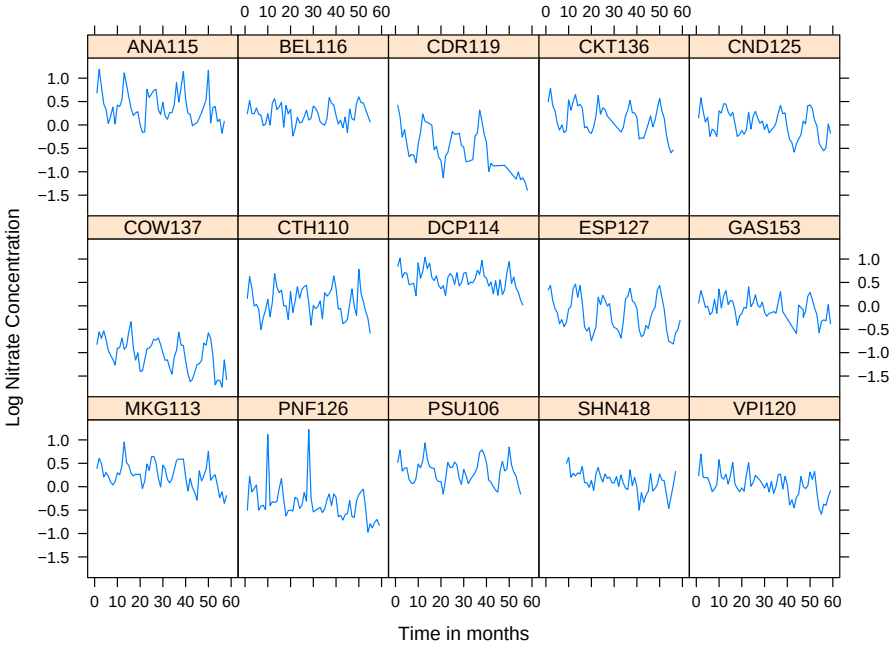

Figure 3 – Site profile plot over time by sites.

Source: Research data.

The data set is also composed of a set of explanatory variables and it is of interest to investigate the relationship they hold with the response. For example, Figure 4 presents a heat correlation plot, that aids better understanding the linear relationship between the response and the explanatory variables. It is expected that those variables with least linear correlation with the response to be removed from the final model and also those which appear to be linear correlated in the set of explanatory variables.

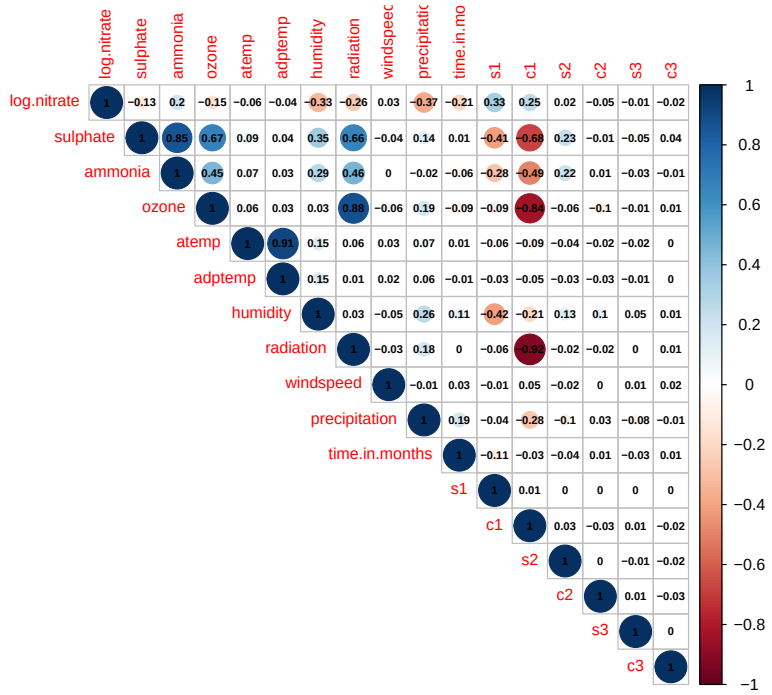

Figure 4 – Correlation plot.

Source: Research data.

CHAPTER

2

LASSO FOR LME

2.1 Preliminaries

2.1.1 Maximum Likelihood Estimation for LME Models

Suppose that a researcher is analyzing data from a repeated measures experiment, as the one described in Chapter 1. In order to model the relationship between response and covariates, suppose a model for $\mathbf{Y}_i$ (SINGER; NOBRE; ROCHA, 2018), the i-th subject as

$$\mathbf{Y}_i = g(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{b}_i^*) + \mathbf{e}_i,\ i = 1, \ldots, m \tag{2.1}$$

where $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ $(n_i \times 1)$ is the vector which includes the observed response variables for the i-th subject, $\boldsymbol{\beta}(p \times 1)$ is the vector of fixed unknown parameters to be estimated, $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$ $(n_i \times p)$ is the known full rank design matrix for the fixed effects and $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijn_i})^\top$ $(n_i \times 1)$ is the vector which contains the values of the j-th covariate $(j = 1, \ldots, p)$ for the i-th subject, $\mathbf{b}_i^*$ $(q \times 1)$ is a vector of latent variables, known as the random effects which represent the subject-level behavior of the i-th subject, $\mathbf{Z}_i$ $(n_i \times q)$ is the known full rank design matrix for the random effects, g is a twice-differentiable function and $\mathbf{e}_i$ $(n_i \times 1)$ is the vector of random errors.

Assume that $\mathbf{b}_i^* \sim N(\mathbf{0}, \sigma^2\mathbf{G})$ and $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2\mathbf{R}_i)$ where $\mathbf{G}(q \times q)$ and $\mathbf{R}_i(n_i \times n_i)$ are both symmetric positive definite matrices, $\mathbf{b}_i^*$ and $\mathbf{e}_i$ are independent random variables. Suppose that it is reasonable to assume that g is a linear function of the fixed effects $\boldsymbol{\beta}$ and the random effects $\mathbf{b}_i^*$, then Equation 2.1 may be expressed as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i^* + \mathbf{e}_i,\ i = 1, \ldots, m. \tag{2.2}$$

The expected value and variance of $\mathbf{Y}_i$ are

$$E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$$
$$Var(\mathbf{Y}_i) = \boldsymbol{\Omega}_i = \sigma^2\left(\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^\top + \mathbf{R}_i\right).$$

Note that the variance $\boldsymbol{\Omega}_i$ can be rewritten as a decomposition of the individual profile dispersion around the response ($\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^\top$) and the response dispersion around the individual profiles ($\mathbf{R}_i$). Both matrices $\mathbf{G}$ and $\mathbf{R}_i$ are known functions of t_1 and t_2 unknown parameters respectively, that is, $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$ and $\mathbf{R}_i = \mathbf{R}_i(\boldsymbol{\theta})$. Thus the covariance matrix for the i-th subject $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}_i(\boldsymbol{\theta})$ also depends in the vector $\boldsymbol{\theta}$ $(t \times 1)$, $t = t_1 + t_2$.

A special case of Equation 2.2 is the homoscedastic[1] conditional independence model, where $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$. This notation shows that the n_i observations from the i-th subject are conditionally independent given $\mathbf{b}_i^*$.

Equation 2.2 can be rewritten in a stacked notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}^* + \mathbf{e}, \tag{2.3}$$

$\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_m^\top)^\top$ $(N \times 1)$ is the response vector where $N = \sum_{i=1}^m n_i$, $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_m^\top)^\top$ $(N \times p)$ is the fixed effects design matrix, $\mathbf{Z} = \oplus_{i=1}^m \mathbf{Z}_i$ $(N \times mq)$ is the random effects design matrix and $\oplus$ denotes the direct sum of matrices (SEARLE, 2017, Sec. 4.10), $\mathbf{b}^* = (\mathbf{b}_1^{*\top}, \ldots, \mathbf{b}_m^{*\top})^\top$ $(mq \times 1)$ is the random effects vector and $\mathbf{e} = (\mathbf{e}_1^\top, \ldots, \mathbf{e}_m^\top)^\top$ $(N \times 1)$ is the random errors vector. Using this notation, $\mathbf{b}^* \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Gamma}(\boldsymbol{\theta}))$, where $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \mathbf{I}_m \otimes \mathbf{G}(\boldsymbol{\theta})$ and $\otimes$ denotes the direct (or Kronecker) product of matrices (SEARLE, 2017, Sec. 4.11), $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{R}(\boldsymbol{\theta}))$, where $\mathbf{R} = \oplus_{i=1}^m \mathbf{R}_i(\boldsymbol{\theta})$, with $\mathbf{b}^*$ and $\mathbf{e}$ independents. Thus, $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}(\boldsymbol{\theta}))$, where $\boldsymbol{\Omega}(\boldsymbol{\theta}) = \sigma^2\left(\mathbf{Z}\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{Z}^\top + \mathbf{R}(\boldsymbol{\theta})\right)$.

In Singer, Nobre and Rocha (2018) there are many different references on the estimation of linear mixed-effects models such as Equation 2.3; the authors also present different possible structures for the covariance matrix. In order to begin with the estimation process, assume that $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ and $\mathbf{R}(\boldsymbol{\theta})$ are known. The aim is to find estimators for the unknown parameters in a way that they maximize the likelihood function or, equivalently, the log-likelihood function.

The log-likelihood function for the model where $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}(\boldsymbol{\theta}))$ is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Omega}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top [\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.4}$$

alternatively,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^m n_i \log 2\pi - \frac{1}{2}\sum_{i=1}^m \log|\boldsymbol{\Omega}_i(\boldsymbol{\theta})| - \frac{1}{2}\sum_{i=1}^m (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top [\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{2.5}$$

Maximization of $\ell(\boldsymbol{\beta}, \boldsymbol{\theta})$ using Equation 2.5 may be carried out differentiating the right side with respect to $\boldsymbol{\beta}$ and equaling the result to zero so that

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left[\sum_{i=1}^m \mathbf{X}_i^\top (\boldsymbol{\Omega}_i(\boldsymbol{\theta}))^{-1}\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^m \mathbf{X}_i^\top (\boldsymbol{\Omega}_i(\boldsymbol{\theta}))^{-1}\mathbf{Y}_i\right], \tag{2.6}$$

[1] Either spellings *homoscedastic* or *homoskedastic* (MCCULLOCH, 1985) are frequently used, but the first spelling was preferred for this work.

the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ (as a function of $\boldsymbol{\theta}$), which is the same as the weighted least squares estimator. Replacing the previous estimator in Equation 2.5, the profile log-likelihood function $\ell(\widehat{\boldsymbol{\beta}}, \boldsymbol{\theta})$ is defined, which may be differentiated with respect to $\boldsymbol{\theta}$, equal to zero and obtain the following system of equations

$$-\frac{1}{2}\sum_{i=1}^{m} tr\left[\left[\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\theta}})\right]^{\top}\dot{\boldsymbol{\Omega}}_i(\widehat{\boldsymbol{\theta}})\right] - \frac{1}{2}\sum_{i=1}^{m}\left[\frac{\partial Q_i(\boldsymbol{\theta})}{\partial \theta_j}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right] = \mathbf{0}, \tag{2.7}$$

$j = 1, \ldots, t$, where

$$\dot{\boldsymbol{\Omega}}_i(\widehat{\boldsymbol{\theta}}) = \left[\frac{\partial \boldsymbol{\Omega}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^{\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

and

$$Q_i(\boldsymbol{\theta}) = (\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))^{\top}\left[\boldsymbol{\Omega}_i(\boldsymbol{\theta})\right]^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})).$$

Therefore, the solution for Equation 2.7, $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. Replacing $\widehat{\boldsymbol{\theta}}$ in Equation 2.6, leads to the maximum likelihood estimator of $\boldsymbol{\beta}$. Details on the previous matrices derivatives can be found in Singer, Nobre and Rocha (2018, Appx. A.5).

According to Diggle (2002), Singer, Nobre and Rocha (2018), the maximum likelihood method provides unbiased estimators only for the fixed effects; whereas for the random effects this methodology does not account for the loss in degrees of freedom in the estimation of the covariance matrix due to the estimation of the fixed effects. The authors suggest that one should use the restricted maximum likelihood method (PATTERSON; THOMPSON, 1971). This methodology requires an orthogonal transformation like $\mathbf{Y}^{\dagger} = \mathbf{U}\mathbf{Y}$, such that $E(\mathbf{Y}^{\dagger}) = 0$, $Var(\mathbf{Y}^{\dagger}) = \mathbf{U}\boldsymbol{\Omega}(\boldsymbol{\theta})\mathbf{U}^{\top}$ and $\mathbf{U}^{\top}\mathbf{X} = \mathbf{0}$. Although this method is invariant under the choice of $\mathbf{U}$, one usually chooses $\mathbf{U} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}$, the weighted least squares residual projection matrix. In the case which observations are i.i.d.'s, set $\boldsymbol{\Omega} = \mathbf{I}$. As $\mathbf{U}$ has rank $N - p$, then $\mathbf{Y}^{\dagger} \sim N_{N-p}(\mathbf{0}, \mathbf{U}\boldsymbol{\Omega}(\boldsymbol{\theta})\mathbf{U}^{\top})$. The maximization process under this approach is similar to the one already presented.

Predictors for the random effects are derived from the joint distribution of the random effects $\mathbf{b}^*$ and observations $\mathbf{Y}$

$$f(\mathbf{y}, \mathbf{b}^*) = f(\mathbf{y}|\mathbf{b}^*)f(\mathbf{b}^*),$$

where $\mathbf{Y}|\mathbf{b}^* \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}^*, \sigma^2\mathbf{R})$ and $\mathbf{b}^* \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Gamma})$. Singer, Nobre and Rocha (2018) presents a few comments on the best linear unbiased estimator for $\boldsymbol{\beta}$ and best linear unbiased predictor for $\mathbf{b}^*$.

In order to solve the systems in Equation 2.6 and Equation 2.7, one must use use an iterative method such as Newton-Raphson, Fisher's scoring or EM algorithm. There are particular cases for the covariance matrix where the researcher can find an explicit solution for the estimators. Due to the approach presented at Bondell, Krishna and Ghosh (2010), the EM algorithm will be presented.

The EM algorithm (DEMPSTER; LAIRD; RUBIN, 1977) was first proposed to deal with missing data during estimation. However, it can be used to obtain MLEs of the parameters for models based in longitudinal data. This algorithm relies on the data augmented likelihood function $\mathbf{Y}_c = (\mathbf{Y}^\top, \nu^\top)^\top$, where $\mathbf{Y}$ denotes the vector of observations and ν the vector of omitted observations (when dealing with missing data) or latent variables ($\nu = \mathbf{b}$, when dealing with random effects) of mixed models. Let $\mathbf{s}$ be the vector of sufficient statistics for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Then, the l-th iteration of the algorithm is

E-step: Conditional expected value of the sufficient statistics given the observations and the updated values of the parameter of the previous iteration

$$\mathbf{s}^{(l)} = E\left[\mathbf{s}(\mathbf{Y}_c)|\mathbf{y}, \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}\right]$$

M-step: Solve with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$

$$\mathbf{s}^{(l)} = E\left[\mathbf{s}(\mathbf{Y}_c)|\boldsymbol{\beta}, \boldsymbol{\theta}\right].$$

For the homoscedastic conditional independence model Singer, Nobre and Rocha (2018) presents the sufficient statistics that are necessary for the E-step, as well as other details pertaining the EM algorithm for LME models.

2.1.2 Variable selection

According to Hastie, Tibshirani and Friedman (2009), there are two main reasons to perform a search for the best subset of explanatory variables in a data set:

- Prediction accuracy: the least squares estimates (or MLE) has large variance, which can be reduced by shrinking or setting some coefficients to zero;
- Interpretation: one would like to determine which of the explanatory variables has a greater impact on explaining the response, and a subset of these variables may be better suited than the whole set of predictors.

There are a few methods for shrinking and variable selection: best subset selection, forward/backward/stepwise selection, ridge regression and lasso, for example. Miller (2002) and Hastie, Tibshirani and Friedman (2009) bring an intensive review of these methods.

2.1.2.1 Ridge regression

Before introducing the lasso regression method the ridge regression will be presented, as it will be useful when constructing the diagnostics measures.

The main objective of the ridge regression is to shrink the regression coefficients by imposing a penalty term. Suppose that $Var(\mathbf{Y}) = \boldsymbol{\Omega} = \text{diag}(\omega_{11}, \ldots, \omega_{nn})$. Then, the ridge regression estimate is obtained by minimizing the penalized sum of squares

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \omega_{ii}^{-1} \left(y_i - \beta_0 - x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{2.8}$$

The penalty term $\sum_{j=1}^{p} \beta_j^2$ is called the L_2 penalty.

The following is an equivalent way to present the ridge regression minimization problem

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{ridge} &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \omega_{ii}^{-1} \left(y_i - \beta_0 - x_{ij}\beta_j\right)^2 \right\}, \\ &\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t. \end{aligned} \tag{2.9}$$

As pointed out by Hastie, Tibshirani and Friedman (2009), there is a one-to-one relationship between λ in Equation 2.8 and t in Equation 2.9. The authors also argue that the ridge regression estimates are not equivariant under scaling of the inputs, which leads the researcher to standardize the explanatory variables. Note that the intercept has been left out of the penalty term. Penalizing the intercept would make the procedure depend on the origin chosen for Y, the authors emphasize. From now on, assume that the variables in the design matrix $\mathbf{X}$ have been standardized.

Equation 2.8, considering any positive-definite $\boldsymbol{\Omega}$, can be rewritten in matrix form as

$$RSS(\lambda) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\top} \boldsymbol{\Omega}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{\top}\boldsymbol{\beta}. \tag{2.10}$$

Differentiating Equation 2.10 with respect to $\boldsymbol{\beta}$ and setting it to zero,

$$\begin{aligned} \mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{Y} + \mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}^{ridge} + \lambda\hat{\boldsymbol{\beta}}^{ridge} &= \mathbf{0} \\ \mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{Y} + \left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{ridge} &= \mathbf{0} \\ \hat{\boldsymbol{\beta}}^{ridge} &= \left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{Y} \end{aligned} \tag{2.11}$$

Note that even if $\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}$ was singular, a positive constant is added to its diagonal. Then, the inverse matrix in Equation 2.11 would still exist.

The variance of the estimator is given by

$$Var(\hat{\boldsymbol{\beta}}^{ridge}) = \left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}. \tag{2.12}$$

The expression in Equation 2.12 will be useful when constructing the diagnostics measures.

In order to find the best value for λ, cross-validation methods can be used, for example James *et al.* (2013, Sec. 3.8.5) .

2.1.2.2 The lasso

Recall that, for the ridge regression, the solutions are not equivariant under scaling of the inputs, which leads the researcher to standardize the explanatory variables. A similar rationale can be made for the lasso estimates. Remember that $\mathbf{X}$ is assumed to be standardized, unless told otherwise.

The lasso (also known as Least Absolute Shrinkage and Selection Operator (LASSO) [2]) regression method (TIBSHIRANI, 1996; JAMES *et al.*, 2013) aims, primarily, for the shrinking of the estimated parameter. However, due to the nature of the L_1 penalty some of the estimated parameters will the exactly zero. This means that the lasso regression can be used for selection of predictors. The lasso estimator is given by

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{m} \omega_{ii}^{-1} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.13}$$

if $\boldsymbol{\Omega}$ is diagonal, or

$$RSS(\lambda) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\top} \boldsymbol{\Omega}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.14}$$

if $\boldsymbol{\Omega}$ is any positive-definite matrix. Notice that both equations are simply restricted least squares estimation, where λ is the regularization (or tuning) parameter that controls the shrinkage. The best value for λ is obtained by some criteria, for example, one chooses a value for λ such that it minimizes some sort of cross validation score.

2.2 Reparametrizing the model and likelihood function

Assume that a linear mixed-effects model of the form presented in Equation 2.3 is to be fitted to a repeated measures data set (longitudinal data, for example). Furthermore, a variable selection procedure is to be performed in the predictors, both for fixed and random effects. Under this scope, Bondell, Krishna and Ghosh (2010) recalls that previous methods for selecting variables assumed that one of the effects was observed, whereas the selection was performed for the other. For the proposed selection method proposed which performs a selection procedure for both the fixed and random effects at the same time, Bondell, Krishna and Ghosh (2010) reparametrizes the LME model via modified Cholesky decomposition, which allows for selection on the random effects.

Consider the model defined by Equation 2.2 where the error term is assumed to be distributed as $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$ (the homoscedastic conditional independence model), the random effects are distributed as $\mathbf{b}_i^* \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Gamma})$ and consider the factorization of the covariance matrix

[2] Although there is doubt on whether one should write LASSO or lasso, the author in Tibshirani (1996) himself makes use of the lowercase version, and that is the version used throughout this text.

by $\boldsymbol{\Gamma} = \mathbf{D}\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top\mathbf{D}$, where $\mathbf{D} = diag(d_1, d_2, \ldots, d_q)$ is a diagonal matrix and $\boldsymbol{\Upsilon}$ is a $q \times q$ matrix, with 1's on its diagonal, whose (l,r)-th element is denoted by υ_{lr}. This decomposition is unique and leads to a non-negative definite matrix. Given the reparametrization, Equation 2.2 is rewritten as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{D}\boldsymbol{\Upsilon}\mathbf{b}_i + \mathbf{e}_i, \tag{2.15}$$

where $\mathbf{Y}_i$ is assumed mean centered, that is $E[\mathbf{Y}_i] = \mathbf{0}$, with the predictors standardized such that both $\mathbf{X}_i^\top\mathbf{X}_i$ and $\mathbf{Z}_i^\top\mathbf{Z}_i$ represent correlation matrices, $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})^\top$ is a $q \times 1$ vector distributed as $N(0, \sigma^2\mathbf{I}_q)$. $\mathbf{b}_i^*$'s covariance matrix is expressed as a function of $\mathbf{d} = (d_1, d_2, \ldots, d_q)^\top$, and the $q(q-1)/2$ free elements of $\boldsymbol{\Upsilon}$ denoted by the vector $\boldsymbol{\upsilon} = (\upsilon_{lr} : l = 1, \ldots, q : r = l+1, \ldots, q)^\top$. Denote $\boldsymbol{\phi} = \left(\boldsymbol{\beta}^\top, \mathbf{d}^\top, \boldsymbol{\upsilon}^\top\right)^\top$ as the $k \times 1$ vector of unknown parameters to be estimated, where $k = p + q(q+1)/2$.

The response variable $\mathbf{Y}_i$ for the reparametrized model in Equation 2.15, conditioning on $\mathbf{X}_i$ and $\mathbf{Z}_i$, is $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}_i)$, where $\boldsymbol{\Omega}_i = \sigma^2(\mathbf{Z}_i\mathbf{D}\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top\mathbf{D}\mathbf{Z}_i^\top + \mathbf{I}_{n_i})$. Thus, the log-likelihood function for the model Equation 2.3, dropping constant terms is given by

$$\ell(\boldsymbol{\phi}) = -\frac{1}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.16}$$

where $\boldsymbol{\Omega} = Diag(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_m)$ is a block-diagonal matrix with elements $\boldsymbol{\Omega}_i$ in the main diagonal, and $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_m^\top)^\top$, $\mathbf{X} = \left[\mathbf{X}_1^\top, \ldots, \mathbf{X}_m^\top\right]$.

2.2.1 Penalized selection and estimation for the reparametrized model using the constrained EM algorithm

Treating $\mathbf{b}$ as observed and dropping constant terms, the log-likelihood function for the augmented data is given by

$$\ell_c(\boldsymbol{\phi}|\mathbf{y}, \mathbf{b}) = -\frac{N + mq}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(||\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Upsilon}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}||^2 + \mathbf{b}^\top\mathbf{b}\right) \tag{2.17}$$

where $\widetilde{\mathbf{D}} = \mathbf{I}_m \otimes \mathbf{D}$ and $\widetilde{\boldsymbol{\Upsilon}} = \mathbf{I}_m \otimes \boldsymbol{\Upsilon}$.

Including a penalty as function of $\mathbf{d}$ and $\boldsymbol{\beta}$ to Equation 2.17, it is then possible to decide whether to include or not predictors involving the parameters by maximizing the conditional expectation of Equation 2.17 and, equivalently, minimizing the quadratic form $||\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Upsilon}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}||^2$.

Bondell, Krishna and Ghosh (2010) suggests using the adaptive lasso Zou (2006) for variable selection in the model from Equation 2.15. The adaptive lasso is defined as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda_t\sum_{i=1}^{p}\bar{w}_j|\beta_j|\right\}, \tag{2.18}$$

where λ_t is the regularization parameter, $\bar{w}_j$ are adaptive weights, that may be $\bar{w}_j = 1/|\bar{\beta}_j|$ where $\bar{\beta}_j$ is the generalized least squares estimate for the j-th coefficient. As λ_t increases, the coefficients are continuously shrunk to zero and, due to the L_1 penalty, some can be exactly zero.

Thus, given the reparametrized model in Equation 2.15 and the log-likelihood function for the augmented data in Equation 2.17, the joint criterion for estimation and selection penalized by the adaptive weights is given by

$$Q_c(\boldsymbol{\phi}|\mathbf{y},\mathbf{b}) = ||\mathbf{y} - \mathbf{Z}\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Upsilon}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda_t \left(\sum_{j=1}^{p} \frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|\bar{d}_j|} \right), \tag{2.19}$$

where $\bar{\boldsymbol{\beta}}$ is the generalized least squares estimates for $\boldsymbol{\beta}$, and $\bar{\mathbf{d}}$ is obtained by decomposing the covariance matrix estimated via unpenalized restrict maximum likelihood. Rearranging the terms of Equation 2.19,

$$Q_c(\boldsymbol{\phi}|\mathbf{y},\mathbf{b}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}Diag(\widetilde{\boldsymbol{\Upsilon}}\mathbf{b})(\mathbf{1}_q \otimes \mathbf{I}_m)\mathbf{d}||^2 + \lambda_t \left(\sum_{j=1}^{p} \frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|\bar{d}_j|} \right), \tag{2.20}$$

where $\mathbf{1}_q$ is a $q \times 1$ vector of 1s. Note that Equation 2.20 is a quadratic form in $(\boldsymbol{\beta}^\top, \mathbf{d}^\top)^\top$.

In order to find the estimates, the EM algorithm will be used. For the E-step, compute the conditional expectation of Equation 2.20 and next, for the M-step, minimize the conditional expectation with respect to the parameters of interest. Iterate until convergence.

The conditional distribution of $\mathbf{b}$ given $\boldsymbol{\phi}$ and $\mathbf{y}$ using Equation 2.17 is $\mathbf{b}|\mathbf{y},\boldsymbol{\phi} \sim N(\boldsymbol{\tau}, \sigma^2\mathbf{V})$, with mean and variance given by

$$\begin{aligned}
\boldsymbol{\tau}^{(l)} &= \left(\widetilde{\boldsymbol{\Upsilon}}^{\top(l)} \widetilde{\mathbf{D}}^{(l)} \mathbf{Z}^\top \mathbf{Z} \widetilde{\mathbf{D}}^{(l)} \widetilde{\boldsymbol{\Upsilon}}^{(l)} + \mathbf{I} \right)^{-1} \left(\mathbf{Z}\widetilde{\mathbf{D}}^{(l)}\widetilde{\boldsymbol{\Upsilon}}^{(l)} \right)^\top \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(l)} \right) \\
\mathbf{V}^{(l)} &= \left(\widetilde{\boldsymbol{\Upsilon}}^{\top(l)} \widetilde{\mathbf{D}}^{(l)} \mathbf{Z}^\top \mathbf{Z} \widetilde{\mathbf{D}}^{(l)} \widetilde{\boldsymbol{\Upsilon}}^{(l)} + \mathbf{I} \right)^{-1},
\end{aligned}$$

respectively. The l indexes the iteration, and $l = 0$ corresponds to the initial values given by the restricted maximum likelihood estimates. The updated estimate for σ^2 on the l-th iteration is given by

$$\sigma^{2(l)} = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(l)} \right)^\top \left(\mathbf{Z}\widetilde{\mathbf{D}}^{(l)}\widetilde{\boldsymbol{\Upsilon}}^{(l)}\widetilde{\boldsymbol{\Upsilon}}^{\top(l)}\widetilde{\mathbf{D}}^{(l)}\mathbf{Z}^\top + \mathbf{I}_N \right)^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(l)} \right) / N.$$

Let $\boldsymbol{\phi}^{(l)}$ be the estimate of $\boldsymbol{\phi}$ on the l-th iteration. The E-step is obtained by taking the conditional expectation of $Q_c(\boldsymbol{\phi}|\mathbf{y},\mathbf{b})$,

$$g(\boldsymbol{\phi}|\boldsymbol{\phi}^{(l)}) = E_{\mathbf{b}|\mathbf{y},\boldsymbol{\phi}^{(l)}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}Diag(\widetilde{\boldsymbol{\Upsilon}}\mathbf{b})(\mathbf{1}_q \otimes \mathbf{I}_m)\mathbf{d}||^2 \right\} + \lambda_t \left(\sum_{j=1}^{p} \frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|\bar{d}_j|} \right). \tag{2.21}$$

As for the M-step, the objective function Equation 2.21 is minimized with respect to $\left(\boldsymbol{\beta}^\top, \mathbf{d}^\top, \boldsymbol{\upsilon}^\top\right)^\top$. The optimization inside the M-step is made by iterating between $\boldsymbol{\upsilon}$ and $\left(\boldsymbol{\beta}^\top, \mathbf{d}^\top\right)^\top$. The iteration with respect to $\boldsymbol{\upsilon}$ has closed form (BONDELL; KRISHNA; GHOSH,

2010, Appendix), whereas the iteration with respect to $\left(\boldsymbol{\beta}^\top, \mathbf{d}^\top\right)^\top$ will be a quadratic programming problem. Upon convergence, the final estimates are defined as $\widehat{\boldsymbol{\phi}} = \left(\widehat{\boldsymbol{\beta}}^\top, \widehat{\mathbf{d}}^\top, \widehat{\boldsymbol{\upsilon}}\right)^\top$.

In order to find the best tuning parameter λ_t, the EM algorithm described above is applied to a grid of fixed possible values for λ_t. The final value of λ_t is the one that minimizes a criterion such as AIC, BIC, GIC (generalized information criteria), generalized cross validation or k-fold cross validation. Bondell, Krishna and Ghosh (2010) suggests that BIC is used, as it is consistent for model selection under general conditions (SCHWARZ, 1978; SHAO, 1997). The BIC-type criterion is given by

$$BIC_{\lambda_t} = -2\ell(\widehat{\boldsymbol{\phi}}) + \log(N) \times (df_{\lambda_t}), \tag{2.22}$$

where $\ell(\widehat{\boldsymbol{\phi}})$ is the log-likelihood evaluated at the estimate of $\boldsymbol{\phi}$ for that specific value of λ_t. The degrees of freedom df_{λ_t} are defined as the number of non-zero coefficients in $\boldsymbol{\phi}$. Given a set of values for λ_t, choose λ_t such that it minimizes the criterion BIC_{λ_t}. Bondell, Krishna and Ghosh (2010) also presents asymptotic properties for the estimators.

2.3 Robust approach

Consider the case where the joint criterion in Equation 2.19 is used in a data set but possible outliers and leverage points were identified. Fan, Qin and Zhu (2014) proposes an extension of the previous method in order to take into account inadequacies in the fitted model. The authors propose two main changes in the method

- Robustified likelihood by a function of the Mahalanobis distance to reduce the impact of outliers in the covariates;
- Adding a function of the studentized residuals to the response to reduce the influence of outliers in the response.

The robust likelihood used was proposed by Sinha (2004) in the context of GLMM, but is suited to the purposes of this work.

Thus, suppose that the vector of covariates $\mathbf{x}_{ij}$ is an outlier, that is, $\mathbf{x}_{ij}$ is a leverage observation. Then, its impact may be reduced by introducing a weight w_{ij}. The weight w_{ij} is defined as

$$w_{ij} = \min\left\{1, \left[\frac{d_0}{\left(\mathbf{x}_{ij}-\mathbf{m}_x\right)^\top S_x^{-1}\left(\mathbf{x}_{ij}-\mathbf{m}_x\right)}\right]^{\delta/2}, \left[\frac{b_0}{\left(\mathbf{z}_{ij}-\mathbf{m}_z\right)^\top S_z^{-1}\left(\mathbf{z}_{ij}-\mathbf{m}_z\right)}\right]^{\delta/2}\right\}, \tag{2.23}$$

where $\delta \geq 1$, d_0 is chosen as the 95th percentile of the chi-square distribution with the degrees of freedom equal to the dimension of $\mathbf{x}_{ij}$, b_0 is chosen as the 95th percentile of the chi-square distribution with the degrees of freedom equal to the dimension of $\mathbf{z}_{ij}$, $\mathbf{m}_x$ equals to the median

of $\mathbf{x}_{ij}$, $\mathbf{m}_z$ equals to the median of $\mathbf{z}_{ij}$, S_x denotes the median absolute deviance of $\mathbf{x}_{ij}$ and S_z denotes the median absolute deviance of $\mathbf{z}_{ij}$. The median absolute deviance (or deviation) (MAD) for a sample $X_1,\ldots,X_n$ from the random variable X is defined as

$$MAD = median(|X_i - \tilde{X}|),$$

where $\tilde{X} = median(X)$. To reduce the influence of outliers in the response, the authors propose to add a quantity v_{ij} to y_{ij}, that is defined as

$$v_{ij} = sign(r_{ij})(|r_{ij}| - c)\sigma \mathbb{1}_{(|r_{ij}|>c)}, \tag{2.24}$$

where $\mathbb{1}_{(.)}$ is the indicator function and $r_{ij} = (y_{ij} - \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} - \mathbf{z}_{ji}^\top \mathbf{D}\boldsymbol{\Upsilon}\hat{\mathbf{b}}_i)/\hat{\sigma}$ is the conditional residual. This quantity is added to the response and a new variable $y_{ij}^* = y_{ij} - v_{ij}$ is created, so that the studentized residual associated with the modified variable is limited to the interval $[-c,c]$. The constant c is similar to a tuning parameter and can be chosen by considering the balance of robustness and estimation efficiency, according to the authors.

The robustified log-likelihood for the i-th subject is then written as

$$\ell_i^R(\boldsymbol{\phi}|\mathbf{y}_i) = \log \int \prod_j^{n_i} (\sigma^2)^{-1/2} \left\{ \exp\left[-\frac{1}{2\sigma^2}\left(y_{ij}^* - \mathbf{x}_{ij}^\top\boldsymbol{\beta} - \mathbf{z}_{ij}^\top \mathbf{D}\boldsymbol{\Upsilon}\mathbf{b}_i \right) \right] \right\}^{w_{ij}} \times$$
$$\times \left(\sigma^2\right)^{-q/2} \exp\left\{ -\frac{1}{2\sigma^2}\mathbf{b}_i^\top \mathbf{b}_i \right\} d\mathbf{b}_i. \tag{2.25}$$

The robustified log-likelihood for the augmented data $\mathbf{y}_{ic} = (\mathbf{y}_i^\top, \mathbf{b}_i^\top)^\top$ is

$$\ell_i^R(\boldsymbol{\phi}|\mathbf{y}_{ic}) = -\frac{n_i+q}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n_i}\left\{ w_{ij}\left(y_{ij}^* - \mathbf{x}_{ij}^\top\boldsymbol{\beta} - \mathbf{z}_{ij}^\top\mathbf{D}\boldsymbol{\Upsilon}\mathbf{b}_i \right)^2 + \mathbf{b}_i^\top\mathbf{b}_i \right\}, \tag{2.26}$$

and the complete robustified log-likelihood is

$$\begin{aligned}
\ell^R(\boldsymbol{\phi}|\mathbf{y}_c) &= \sum_{i=1}^m \ell_i^R(\boldsymbol{\phi}|\mathbf{y}_{ic}) \\
&= -\frac{N+mq}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^m\sum_{j=1}^{n_i}\left\{ w_{ij}\left(y_{ij}^* - \mathbf{x}_{ij}^\top\boldsymbol{\beta} - \mathbf{z}_{ij}^\top\mathbf{D}\boldsymbol{\Upsilon}\mathbf{b}_i \right)^2 + \mathbf{b}_i^\top\mathbf{b}_i \right\} \\
&= -\frac{N+mq}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\left(\mathbf{y}^* - \mathbf{Z}\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Upsilon}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta} \right)^\top \mathbf{W}\left(\mathbf{y}^* - \mathbf{Z}\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Upsilon}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta} \right) + \mathbf{b}^\top\mathbf{b} \right]
\end{aligned} \tag{2.27}$$

where $\widetilde{\mathbf{D}}$ and $\widetilde{\boldsymbol{\Upsilon}}$ are the same as defined in section 2.2, $\mathbf{W} = diag(\mathbf{W}_1,\ldots,\mathbf{W}_m)$, $\mathbf{W}_i = diag(w_{i1},\ldots,w_{in_i})$ and $\boldsymbol{\phi} = \left(\boldsymbol{\beta}^\top, \mathbf{d}^\top, \boldsymbol{\upsilon}^\top\right)^\top$. The objective function to be minimized via adaptive lasso, that is, the joint variable selection criteria for the robustified model, is defined as

$$Q_c(\boldsymbol{\phi}) = -\ell(\boldsymbol{\phi}|\mathbf{y},\mathbf{b}) + \lambda_t\left(\sum_{j=1}^p \frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^q \frac{|d_j|}{|\bar{d}_j|} \right). \tag{2.28}$$

Once again, the constrained EM algorithm will be used as the one presented at subsection 2.2.1. For the E-step, the posterior distribution for the random effects must be defined. In a similar way than for the previous case, $\mathbf{b}^{(l)}|\mathbf{y},\boldsymbol{\phi}^{(l)} \sim N(\boldsymbol{\tau}^{(l)}, \sigma^{2(l)}\mathbf{V}^{(l)})$, where the mean and variance are given by

$$\begin{aligned}\boldsymbol{\tau}^{(l)} &= \left(\widetilde{\boldsymbol{\Upsilon}}^{\top(l)}\widetilde{\mathbf{D}}^{(l)}\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z}\widetilde{\mathbf{D}}^{(l)}\widetilde{\boldsymbol{\Upsilon}}^{(l)} + \mathbf{I}\right)^{-1}\left(\mathbf{Z}\widetilde{\mathbf{D}}^{(l)}\widetilde{\boldsymbol{\Upsilon}}^{(l)}\right)^{\top}\mathbf{W}\left(\mathbf{y}^{*(l)} - \mathbf{X}\boldsymbol{\beta}^{(l)}\right)\\ \mathbf{V}^{(l)} &= \left(\widetilde{\boldsymbol{\Upsilon}}^{\top(l)}\widetilde{\mathbf{D}}^{(l)}\mathbf{Z}^{\top}\mathbf{W}(\boldsymbol{\varphi}^{(l)})\mathbf{Z}\widetilde{\mathbf{D}}^{(l)}\widetilde{\boldsymbol{\Upsilon}}^{(l)} + \mathbf{I}\right)^{-1}.\end{aligned} \tag{2.29}$$

$\sigma^{2(l)}$ is the current median absolute deviation estimator for σ^2 (ROUSSEEUW; CROUX, 1993). During the E-step, the corrected response y^*_{ij} is updated based on $\boldsymbol{\phi}^{(l)}$.

For the M-step, $g(\boldsymbol{\phi}|\boldsymbol{\phi}^{(l)})$ will be minimized, where $\boldsymbol{\phi}^{(l)}$ is the updated estimate in the l-th iteration for $\boldsymbol{\phi}$. Upon convergence, the final estimate $\widehat{\boldsymbol{\phi}}$ is attained. For the choice of the tuning parameter, λ_t, proceed as the previous case, in which λ_t is chosen such that it minimizes the BIC-type criterion given by

$$BIC_{\lambda_t} = -2\ell(\widehat{\boldsymbol{\phi}}) + \log(N) \times (df_{\lambda_t}),$$

where $\ell(\widehat{\boldsymbol{\phi}})$ is the log-likelihood evaluated at the estimate of $\boldsymbol{\phi}$ for that specific value of λ_t. The degrees of freedom df_{λ_t} are defined as the number of non-zero coefficients in $\boldsymbol{\phi}$.

CHAPTER

3

RESIDUAL ANALYSIS AND DIAGNOSTICS FOR LME LASSO

In this Chapter, diagnostic techniques for the ridge regression and lasso method will be first introduced. In addition, some residual analysis techniques for LME models will also be presented. Finally, some existing diagnostic techniques for LME models will be modified so that they can suit the lasso regression context. Throughout this Chapter, $\mathbf{y}_i$, $\mathbf{x}_i$ and $\mathbf{z}_i$ will be used to denote the set of rows associated with the i-th subject in $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{Z}$, respectively; and the subscript (i) denotes deletion of the rows associated with the i-th subject.

3.1 Cook's Distance for ridge and lasso regression

Consider the linear regression model given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\mathbf{Y}(N \times 1)$ is the vector of responses, $\mathbf{X}(N \times p)$ is the known full-rank design matrix and $\boldsymbol{\varepsilon}(N \times 1)$ is a vector of random variables with expected value $\mathbf{0}$ and variance $\boldsymbol{\Omega}$. If $\boldsymbol{\Omega}$ is known, the generalized least squares (GLS) estimator (also known as weighted estimator) $\hat{\boldsymbol{\beta}}^{GLS}$ of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{GLS} = \left(\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{Y}. \tag{3.2}$$

If $\boldsymbol{\Omega}$ is unknown, a suitable estimate $\hat{\boldsymbol{\Omega}}$ may be used instead, for example the one obtained using one of the methods proposed in Chapter 2. Using this expression, along with Equation 2.11 the weighted ridge regression estimator (HOLLAND, 1973) of $\boldsymbol{\beta}$ is as it was previously defined

$$\hat{\boldsymbol{\beta}}_{\phi}^{ridge} = \left(\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{I}\right)^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{Y}. \tag{3.3}$$

One possible criterion to determine the best value of ϕ is the generalized cross-validation (GCV) criterion that is defined as

$$GCV(\phi) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - \hat{y}_{ij,\phi}\right)^2}{\left\{1 - tr(\mathbf{H}_\phi)\right\}^2}, \tag{3.4}$$

where $\mathbf{H}_\phi = \mathbf{X}\left(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{I}\right)^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}$ is the ridge hat matrix, $y_{ij,\phi}$ is the ij-th entry of the vector of fitted values $\mathbf{Y}_\phi = \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{ridge} = \mathbf{H}_\lambda\mathbf{Y}$ and $tr(\cdot)$ is the trace of a matrix.

Given the value of ϕ that minimizes Equation 3.4, Kim *et al.* (2015) defines the Cook's distance for the ridge regression for the i-th subject as

$$C_i^{ridge} = \frac{1}{p}\left(\hat{\boldsymbol{\beta}}_\phi^{ridge} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{ridge}\right)^\top Cov\left(\hat{\boldsymbol{\beta}}_\phi^{ridge}\right)^{-1}\left(\hat{\boldsymbol{\beta}}_\phi^{ridge} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{ridge}\right), \tag{3.5}$$

where $\hat{\boldsymbol{\beta}}_{\phi(i)}^{ridge}$ is the weighted ridge estimator without the observations associated with the i-th subject, and $Cov\left(\hat{\boldsymbol{\beta}}_\phi^{ridge}\right)$ is as defined in Equation 2.12, if $\boldsymbol{\Omega}$ is known. If $\boldsymbol{\Omega}$ is unknown, its estimate $\hat{\boldsymbol{\Omega}}$ can be a suitable substitute. Kim *et al.* (2015) shows the expressions for the basic building blocks for both when a single observation is deleted from the data set (equivalent to removing the (ij)-th observation for a LME model) and when a set of observations is removed from the data set (equivalent to removing the observations associated to the i-th subject) for the case which the hat matrix $\mathbf{H}$ is derived from a ridge regression model assuming the observations are independent and identically distributed. However, due to the presence of the random effects, that is not the case presented in this dissertation.

So, Equation 3.5 can be rewritten in terms of basic building blocks, such that it accounts for the random effects and even a general structure for the model error. First, write (SEBER, 2012)

$$\left(\mathbf{X}_{(i)}^\top\boldsymbol{\Omega}_{(i)}^{-1}\mathbf{X}_{(i)} + \phi\mathbf{I}\right)^{-1} = \left[\left(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{I}\right) - \mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i\right]^{-1}. \tag{3.6}$$

Using the *Sherman-Morrison-Woodbury formula* (SEBER, 2012, Appx. A.9.3),

$$\left(\mathbf{X}_{(i)}^\top\boldsymbol{\Omega}_{(i)}^{-1}\mathbf{X}_{(i)} + \phi\mathbf{I}\right)^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\left[\boldsymbol{\Omega}_i^{-1} - \boldsymbol{\Omega}_i^{-1}\mathbf{H}_{\phi,i}\right]^{-1}\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i\mathbf{A}^{-1} \tag{3.7}$$

where $\mathbf{H}_{\phi,i} = \mathbf{X}_i\left(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{I}\right)^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}$ and $\mathbf{A} = \mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{I}$. Then, write

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\phi(i)}^{ridge} &= \left(\mathbf{X}_{(i)}^\top\boldsymbol{\Omega}_{(i)}^{-1}\mathbf{X}_{(i)} + \phi\mathbf{I}\right)^{-1}\mathbf{X}_{(i)}^\top\boldsymbol{\Omega}_{(i)}^{-1}\mathbf{Y}_{(i)} \\
&= \left\{\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\left[\boldsymbol{\Omega}_i^{-1} - \boldsymbol{\Omega}_i^{-1}\mathbf{H}_{\phi,i}\right]^{-1}\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i\mathbf{A}^{-1}\right\}\left(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{Y} - \mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{Y}_i\right) \\
&= \left\{\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\left[\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right]^{-1}\boldsymbol{\Omega}_i\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i\mathbf{A}^{-1}\right\}\left(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{Y} - \mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{Y}_i\right) \\
&= \left\{\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\left[\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right]^{-1}\mathbf{X}_i\mathbf{A}^{-1}\right\}\left(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{Y} - \mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{Y}_i\right) \\
&= \hat{\boldsymbol{\beta}}_\phi^{ridge} + \mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\left[\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right]^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}}_\phi^{ridge} - \mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{Y}_i - \\
&\quad \mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\left[\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right]^{-1}\mathbf{X}_i\mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{Y}_i.
\end{aligned} \tag{3.8}$$

The difference $\hat{\boldsymbol{\beta}}_{\phi}^{ridge} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{ridge}$ is written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\phi}^{ridge} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{ridge} &= \mathbf{A}^{-1}\mathbf{X}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{Y}_i + \mathbf{A}^{-1}\mathbf{X}_i^{\top}\boldsymbol{\Omega}_i^{-1}\left[\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right]^{-1}\mathbf{H}_{\phi,i}\mathbf{Y}_i - \\ &\quad \mathbf{A}^{-1}\mathbf{X}_i^{\top}\boldsymbol{\Omega}_i^{-1}\left[\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right]^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}}_{\phi}^{ridge} \\ &= \mathbf{A}^{-1}\mathbf{X}_i^{\top}\boldsymbol{\Omega}_i^{-1}\left\{\left[\mathbf{I}_{n_i} + \left(\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right)^{-1}\right]\mathbf{H}_{\phi,i}\mathbf{Y}_i - \left(\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right)^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}}_{\phi}^{ridge}\right\} \\ &= \mathbf{A}^{-1}\mathbf{X}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i. \end{aligned} \tag{3.9}$$

Finally, Cook's Distance may be written as

$$\begin{aligned}C_i^{ridge} &= \frac{1}{p}\left(\hat{\boldsymbol{\beta}}_{\phi}^{ridge} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{ridge}\right)^{\top} Cov\left(\hat{\boldsymbol{\beta}}_{\phi}^{ridge}\right)^{-1}\left(\hat{\boldsymbol{\beta}}_{\phi}^{ridge} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{ridge}\right) \\ &= \frac{1}{p\sigma^2}\mathbf{B}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i\mathbf{A}^{-1}\mathbf{A}\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{X}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i \\ &= \frac{1}{p\sigma^2}\mathbf{B}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{H}_i\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i, \end{aligned} \tag{3.10}$$

where $\mathbf{B}_i = \left[\mathbf{I}_{n_i} + \left(\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right)^{-1}\right]\mathbf{H}_{\phi,i}\mathbf{Y}_i - \left(\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right)^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}}_{\phi}^{ridge}$ and $\mathbf{H}_i = \mathbf{X}_i\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}_i^{\top}$.

Cook's Distance for lasso regression associated with the i-th subject is defined as

$$C_i^{lasso} = \frac{1}{p}\left(\hat{\boldsymbol{\beta}}_{\phi}^{lasso} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{lasso}\right)^{\top} Cov\left(\hat{\boldsymbol{\beta}}_{\phi}^{lasso}\right)^{-1}\left(\hat{\boldsymbol{\beta}}_{\phi}^{lasso} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{lasso}\right). \tag{3.11}$$

As there is no analytic expression for the lasso estimate, $\hat{\boldsymbol{\beta}}_{\phi}^{lasso}$ is used as an approximate solution (TIBSHIRANI, 1996) of the form

$$\hat{\boldsymbol{\beta}}_{\phi}^{lasso} = \left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{K}^{-}\right)^{-1}\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{Y}. \tag{3.12}$$

where $\mathbf{K} = diag(\hat{\boldsymbol{\beta}}_1^{lasso}, \ldots, \hat{\boldsymbol{\beta}}_p^{lasso})$ is a diagonal matrix, $\mathbf{K}^{-}$ is the generalized inverse matrix of $\mathbf{K}$ and ϕ is the value that minimizes the GCV in Equation 3.4. The covariance of $\hat{\boldsymbol{\beta}}_{\phi}^{lasso}$ is defined as

$$Cov(\hat{\boldsymbol{\beta}}^{lasso}) = \sigma^2\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{K}^{-}\right)^{-1}\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{K}^{-}\right)^{-1}. \tag{3.13}$$

Following a similar rationale to the ridge regression case, Cook's Distance for the lasso regression is given by

$$\begin{aligned}C_i^{lasso} &= \frac{1}{p}\left(\hat{\boldsymbol{\beta}}_{\phi}^{lasso} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{lasso}\right)^{\top} Cov\left(\hat{\boldsymbol{\beta}}_{\phi}^{lasso}\right)^{-1}\left(\hat{\boldsymbol{\beta}}_{\phi}^{lasso} - \hat{\boldsymbol{\beta}}_{\phi(i)}^{lasso}\right) \\ &= \frac{1}{p\sigma^2}\mathbf{B}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i\mathbf{A}^{-1}\mathbf{A}\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{X}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i \\ &= \frac{1}{p\sigma^2}\mathbf{B}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{H}_i\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i, \end{aligned} \tag{3.14}$$

where $\mathbf{A} = \mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{K}^{-}$, $\mathbf{H}_i = \mathbf{X}_i\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}_i^{\top}$, $\mathbf{B}_i = \left[\mathbf{I}_{n_i} + \left(\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right)^{-1}\right]\mathbf{H}_{\phi,i}\mathbf{Y}_i - \left(\mathbf{I}_{n_i} - \mathbf{H}_{\phi,i}\right)^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}}_{\phi}^{lasso}$ and $\mathbf{H}_{\phi,i} = \mathbf{X}_i\left(\mathbf{X}^{\top}\boldsymbol{\Omega}^{-1}\mathbf{X} + \phi\mathbf{I}\right)^{-1}\mathbf{X}_i^{\top}\boldsymbol{\Omega}_i^{-1}$.

3.2 Df-Model and Df-lambda

While Cook's Distance may be used to identify influential subjects on the coefficients estimates, in the lasso context influential subjects on the model selection procedure may also be assessed. Next to that, the impact on the regularization parameter λ by removing a subject may also be quantified. Both these measures are proposed in the work by Rajaratnam *et al.* (2019).

Df-model is defined as

$$\text{df-model}(i) = \frac{\delta(i) - E[\delta(i)]}{\sqrt{var[\delta(i)]}}, \tag{3.15}$$

where $\delta(i) = \sum_{j=1}^{p} \left| \mathbb{1}\left(\hat{\boldsymbol{\beta}}_{\phi}^{lasso}\right) - \mathbb{1}\left(\hat{\boldsymbol{\beta}}_{\phi(i)}^{lasso}\right) \right|$, that is, df-model quantifies the model change when the set of observations associated with the i-th subject is removed. Rajaratnam *et al.* (2019) argues that "df-model is a scaled measure of the number of changes in the selected predictor variables that occur in the lasso solution when an observation (or a set of observations) is removed." $m+1$ models have to be fitted to calculate the values of $\delta(i)$ and the sample mean and sample variance can be used as estimates of $E[\delta(i)]$ and $var[\delta(i)]$, respectively. The cut-offs for the df-model may be set to ± 2.

The next measure to be defined is df-lambda, that measures the change in the optimal value of the regularization parameter λ in the lasso regression. This influence measure is defined as

$$\text{df-lambda}(i) = \frac{\hat{\lambda} - \hat{\lambda}(i) - E[\hat{\lambda} - \hat{\lambda}(i)]}{\sqrt{var[\lambda - \hat{\lambda}(i)]}}, \tag{3.16}$$

and the authors describe df-lambda as "a scaled measure of the difference between the optimal value of λ based on the entire data set $\hat{\lambda}$ and the value when the i-th subject is removed $\hat{\lambda}(i)$." The cut-offs for the df-lambda may be set to ± 2. Rajaratnam *et al.* (2019) justifies the cut-offs values for both measures.

For both df-model and df-lambda measures, the sites were removed one by one the model refitted to the data.

3.3 Generalized leverage matrices

Nobre and Singer (2011) defines the marginal generalized leverage matrix for the LME normal model as the derivative of the fitted values with respect to the observed response variable, that is

$$\begin{aligned} \mathbf{L}_1 &= \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}^\top} = \frac{\partial \mathbf{X}\hat{\boldsymbol{\beta}}}{\partial \mathbf{y}^\top} = \frac{\partial \mathbf{X}\left(\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}\right)\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{y}}{\partial \mathbf{y}^\top} \\ &= \mathbf{X}\left(\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X}\right)\mathbf{X}^\top \boldsymbol{\Omega}^{-1}. \end{aligned} \tag{3.17}$$

Note that Equation 3.17 assumes that $\hat{\boldsymbol{\beta}}$ has an explicit expression, which is not the case in the lasso context. For this, the approximation previously used in Equation 3.12 may be of assistance. Thus,

$$\begin{aligned}\mathbf{L}_{1\phi} &= \frac{\partial \hat{\mathbf{y}}_\phi}{\partial \mathbf{y}^{*\top}} = \frac{\partial \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso}}{\partial \mathbf{y}^{*\top}} = \frac{\partial \mathbf{X}\left(\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X} + \phi \mathbf{K}^-\right)^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{y}^*}{\partial \mathbf{y}^{*\top}} \\ &= \mathbf{X}\left(\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X} + \phi \mathbf{K}^-\right)^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}. \end{aligned} \tag{3.18}$$

Consider the vector of conditional fitted values $\hat{\mathbf{y}}_\phi^c = \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} + \mathbf{Z}\widehat{\mathbf{D}}\widehat{\boldsymbol{\Upsilon}}\hat{\mathbf{b}}_\phi = \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} + \mathbf{Z}^*\hat{\mathbf{b}}_\phi$, where $\hat{\mathbf{b}}_\phi = \left[\mathbf{Z}^{*\top}\mathbf{W}\mathbf{Z}^* + \mathbf{I}_{mq}\right]^{-1}\mathbf{Z}^{*\top}\mathbf{W}(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso}) = \mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso})$ as in Equation 2.29. Then,

$$\begin{aligned}\hat{\mathbf{y}}_\phi^c &= \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} + \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso}) \\ &= \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} + \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}\mathbf{y}^* - \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} \\ &= \mathbf{X}\left(\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X} + \phi \mathbf{K}^-\right)^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{y}^* + \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}\mathbf{y}^* \\ &\quad - \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}\mathbf{X}\left(\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{X} + \phi \mathbf{K}^-\right)^{-1}\mathbf{X}^\top \boldsymbol{\Omega}^{-1}\mathbf{y}^*. \end{aligned} \tag{3.19}$$

Differentiating Equation 3.19 with respect to $\mathbf{y}^*$ leads to

$$\begin{aligned}\mathbf{L}_\phi &= \mathbf{L}_{1\phi} + \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W} - \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}\mathbf{L}_{1\phi} \\ &= \mathbf{L}_{1\phi} + \mathbf{L}_2\mathbf{W} - \mathbf{L}_2\mathbf{W}\mathbf{L}_{1\phi} \qquad (3.20) \\ &= \mathbf{L}_{1\phi} + \mathbf{L}_2\left(\mathbf{W} - \mathbf{W}\mathbf{L}_{1\phi}\right). \qquad (3.21) \end{aligned}$$

Thus, $\mathbf{L}_{1\phi}$ is the generalized leverage matrix associated with the fixed effects and $\mathbf{L}_2$ is the generalized leverage matrix associated with the random effects. One may argue that $\mathbf{L}_{2\phi} = \mathbf{L}_2\left(\mathbf{W} - \mathbf{W}\mathbf{L}_{1\phi}\right)$ should be used as the generalized leverage matrix for the random effects. Note, however, that this quantity is a function of $\mathbf{L}_{1\phi}$ and Nobre and Singer (2011) notes that this dependence on the fixed effects could mask the leverage associated with the random effects.

The expressions above were differentiated with respect to $\mathbf{y}^*$ instead of $\mathbf{y}$. That was a convenience choice, as differentiation with respect to $\mathbf{y}$ would lead to derivatives of sign and indicator functions. And although such derivatives are zero for all the domain points, except in the discontinuity points, we opted to avoid possible problems.

For both fixed and random effects leverage matrices, $tr(\mathbf{L}_{ti}/n_i)$ was plotted, in order to illustrate the leverage measure the i-th subject exerts on the fitted model. As a cut-off for the fixed, $2p/N$ (NOBRE; SINGER, 2011), was chosen; as for the random effects, $2q/N$.

3.4 Residual analysis

Nobre and Singer (2007) claims that there are three possible types of residuals in LME models

1. Marginal residuals: $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ that predicts the marginal errors;
2. Conditional residuals: $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}$ that predicts the conditional errors;
3. BLUP: $\mathbf{Z}\hat{\mathbf{b}}$ that predicts the random effects.

Where $\hat{\boldsymbol{\beta}}$ is the lasso estimate (either normal or robust) for LME model. Each type of residual may be used to verify one assumption of the model.

Hilden-Minton (1995) states that a residual is considered to be *pure* for a specific type of error, if it only depends on the "the unseen disturbance and known, fixed quantities", that is, a function only of the fixed effects, observed matrices and the error that it is supposed to predict. If a residual does not meet those requirements, it is then called a *confounded* residual. In each of the subsections below, it will be shown whether a residual is pure or confounded.

3.4.1 Marginal residuals

The first marginal residual presented is the studentized marginal residual. Nobre and Singer (2007) suggests that $LV_i = ||\mathbf{I}_{n_i} - \mathscr{R}_i\mathscr{R}_i^\top||^2$ (LESAFFRE; VERBEKE, 1998) should be plotted versus subject indices, with $\mathscr{R}_i = \hat{\boldsymbol{\Omega}}_i^{-1/2}\hat{\boldsymbol{\varepsilon}}_i$, where $\hat{\boldsymbol{\Omega}}_i^{-1/2}$ is the estimate of the covariance matrix of the response variable, to further investigate whether the structure chosen for the covariance matrix is suited to the i-th subject, that is, the closer this measure is to zero, better fitted is the chosen structure to i-th subject. Singer, Rocha and Nobre (2017) suggests that $\widehat{Var}(\hat{\boldsymbol{\varepsilon}}_i)$ should be used in place of $\hat{\boldsymbol{\Omega}}_i$. Note, however, if $\widehat{Var}(\hat{\boldsymbol{\varepsilon}}_i)$ was to be used, an expression for $Var(\hat{\boldsymbol{\beta}})$ would be necessary and its estimate $\widehat{Var(\hat{\boldsymbol{\beta}})}$, which cannot be derived in the lasso context unless the approximated form in Equation 3.12 is used. Thus, given ϕ, the marginal residual is written as

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}_\phi &= \mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} \\ &= \mathbf{y}^* - \mathbf{H}_\phi\mathbf{y}^* \\ &= \left(\mathbf{I} - \mathbf{H}_\phi\right)\mathbf{y}^* \\ &= \boldsymbol{\Omega}\mathbf{Q}_\phi\mathbf{y}^*\end{aligned} \tag{3.22}$$

where $\mathbf{Q}_\phi = \boldsymbol{\Omega}^{-1}\left(\mathbf{I} - \mathbf{H}_\phi\right)$ and $\hat{\boldsymbol{\varepsilon}}_\phi$ variance is given by $Var(\hat{\boldsymbol{\varepsilon}}_\phi) = \sigma^2\left(\mathbf{I} - \mathbf{H}_\phi\right)\boldsymbol{\Omega}\left(\mathbf{I} - \mathbf{H}_\phi\right)^\top$ that can be estimated if σ^2 and $\boldsymbol{\Omega}$ are replaced by suited estimates. Finally,

$$LV_{i\phi} = ||\mathbf{I}_{n_i} - \mathscr{R}_{i\phi}\mathscr{R}_{i\phi}^\top||^2, \tag{3.23}$$

where $\mathscr{R}_{i\phi} = \left[\widehat{Var}(\hat{\boldsymbol{\varepsilon}}_\phi)\right]^{-1/2}\hat{\boldsymbol{\varepsilon}}_\phi$. $LV_{i\phi}$ can be further standardized by using $LV_{i\phi}^* = \sqrt{LV_{i\phi}}/n_i$.

Note that $\mathscr{R}_{i\phi}$ is the studentized residual associated with the i-th subject. Singer, Nobre and Rocha (2018) suggests this quantity should be plotted versus intra-units observations to further investigate for outlying observations.

Recall that $\hat{\boldsymbol{\varepsilon}}_\phi$ is an estimate for $\boldsymbol{\varepsilon}$, the marginal error. In order to show that $\hat{\boldsymbol{\varepsilon}}_\phi$ is a pure residual, consider the quantity $\hat{\boldsymbol{\varepsilon}}_\phi - \boldsymbol{\varepsilon}$ and notice that

$$\begin{aligned}
\hat{\boldsymbol{\varepsilon}}_\phi - \boldsymbol{\varepsilon} &= \boldsymbol{\Omega}\mathbf{Q}_\phi \mathbf{y}^* - \boldsymbol{\varepsilon} \\
&= \left(\mathbf{I} - \mathbf{H}_\phi\right)\mathbf{y}^* - \boldsymbol{\varepsilon} \\
&= \mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} - \mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} \\
&= -\mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} + \mathbf{X}\boldsymbol{\beta} \\
&= -\mathbf{H}_\phi \mathbf{y}^* + \mathbf{X}\boldsymbol{\beta}
\end{aligned}$$

which only depends on known quantities and on the fixed effects. Due to the lasso approximation, it cannot be shown that it depends only on the marginal error, as shown in Hilden-Minton (1995) for the best linear unbiased estimator of $\boldsymbol{\beta}$.

3.4.2 Conditional residuals

Nobre and Singer (2007) suggests that $\hat{\mathbf{e}}/\hat{\sigma}$ should be plotted versus the fitted values. In the case presented in this work, $\hat{\mathbf{e}}_\phi/\hat{\sigma}$ will be plotted versus $\hat{\mathbf{y}}_\phi = \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} + \mathbf{Z}^*\hat{\mathbf{b}}_\phi$, where $\hat{\mathbf{e}}_\phi = \mathbf{y} - \hat{\mathbf{y}}_\phi$ and $\hat{\sigma}$ is a suitable estimate for the standard deviation σ. QQ-plots for checking normality and homocedasticity of those residuals could also be produced. As the elements of $\hat{\mathbf{e}}_\phi$ may have different variances, Nobre and Singer (2007) suggests that the standardization to be

$$\hat{\mathbf{e}}_{i\phi}^* = \frac{\hat{\mathbf{e}}_{i\phi}}{\hat{\sigma}\sqrt{\hat{q}_{ii}}}, \tag{3.24}$$

where $\hat{q}_{ii}$ is an estimate of the i-th principal diagonal element of $\mathbf{Q}_\phi$, under the homoscedastic conditional independence model assumption. To check for normality, one should be aware that confounding is present in $\hat{\mathbf{e}}_\phi$. Hilden-Minton (1995) suggests that one should use a linear transformation of the conditional residuals, say $\mathbf{J}\hat{\mathbf{e}}_\phi$, such that it has a minimal *fraction of confounding*. To find such fraction, first, $\hat{\mathbf{e}}_\phi$ must be shown to be confounded. It follows,

$$\begin{aligned}
\hat{\mathbf{e}}_\phi &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} - \mathbf{Z}^*\hat{\mathbf{b}}_\phi \\
&= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso} - \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso}) \\
&= \left[\mathbf{I}_N - \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}\right]\left(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\phi^{lasso}\right) \\
&= \left[\mathbf{I}_N - \mathbf{Z}^*\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W}\right]\boldsymbol{\Omega}\mathbf{Q}_\phi\mathbf{y}^* \\
&= \left[\mathbf{I}_N - \mathbf{Z}^*\mathbf{Z}^{*\top}\boldsymbol{\Omega}^{-1}\right]\boldsymbol{\Omega}\mathbf{Q}_\phi\mathbf{y}^* \quad (*) \\
&= \left[\boldsymbol{\Omega} - \mathbf{Z}^*\mathbf{Z}^{*\top}\right]\mathbf{Q}_\phi\mathbf{y}^* \\
&= \left[\mathbf{Z}^*\mathbf{Z}^{*\top} + \mathbf{I}_N - \mathbf{Z}^*\mathbf{Z}^{*\top}\right]\mathbf{Q}_\phi\mathbf{y}^* \\
&= \mathbf{Q}_\phi\mathbf{y}^*,
\end{aligned} \tag{3.25}$$

where $(*)$ is due to the fact that $\mathbf{V}^{-1}\mathbf{Z}^{*\top}\mathbf{W} = \mathbf{Z}^{*\top}\boldsymbol{\Omega}^{-1}$. Using Equation 3.25,

$$\begin{aligned}\hat{\mathbf{e}}_\phi - \mathbf{e} &= \mathbf{Q}_\phi \mathbf{y}^* - \mathbf{e} \\ &= \mathbf{Q}_\phi \mathbf{y}^* - \mathbf{Q}_\phi \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}_\phi \mathbf{X}\boldsymbol{\beta} - \mathbf{Q}_\phi \mathbf{Z}^*\mathbf{b} + \mathbf{Q}_\phi \mathbf{Z}^*\mathbf{b} - \mathbf{e} \\ &= \mathbf{Q}_\phi \left(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}^*\mathbf{b}\right) - \mathbf{e} + \mathbf{Q}_\phi \mathbf{Z}^*\mathbf{b} + \mathbf{Q}_\phi \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Q}_\phi \mathbf{e} - \mathbf{e} + \mathbf{Q}_\phi \mathbf{Z}^*\mathbf{b} + \mathbf{Q}_\phi \mathbf{X}\boldsymbol{\beta} \\ &= \left(\mathbf{Q}_\phi - \mathbf{I}_N\right)\mathbf{e} + \mathbf{Q}_\phi \mathbf{Z}^*\mathbf{b} + \mathbf{Q}_\phi \mathbf{X}\boldsymbol{\beta}, \end{aligned} \tag{3.26}$$

thus, a confounded residual. Furthermore, note that

$$\hat{\mathbf{e}}_\phi = \mathbf{Q}_\phi \mathbf{e} + \mathbf{Q}_\phi \mathbf{Z}^*\mathbf{b} + \mathbf{Q}_\phi \mathbf{X}\boldsymbol{\beta}, \tag{3.27}$$

thus,

$$Var\left[\hat{\mathbf{e}}_\phi\right] = \sigma^2 \mathbf{Q}_\phi \boldsymbol{\Omega} \mathbf{Q}_\phi^\top + \sigma^2 \mathbf{Q}_\phi \mathbf{Z}^* \mathbf{Z}^{*\top} \mathbf{Q}_\phi^\top, \tag{3.28}$$

where $\mathbf{u}_i^\top$ is the i-th row of the identity matrix. Note that as the variance related to the confounded amount increases, compared to the variance related to the pure residual amount, the ability to check for normality decreases (NOBRE; SINGER, 2007). This motivates the definition of the *fraction of confounding* for the i-th conditional residual $\hat{\mathbf{e}}_{i\phi}$ as

$$CF(\hat{\mathbf{e}}_{i\phi}) = \frac{\mathbf{u}_i^\top \mathbf{Q}_\phi \mathbf{Z}^* \mathbf{Z}^{*\top} \mathbf{Q}_\phi^\top \mathbf{u}_i}{\mathbf{u}_i^\top \mathbf{Q}_\phi \boldsymbol{\Omega} \mathbf{Q}_\phi^\top \mathbf{u}_i} = 1 - \frac{\mathbf{u}_i^\top \mathbf{Q}_\phi \mathbf{Q}_\phi^\top \mathbf{u}_i}{\mathbf{u}_i^\top \mathbf{Q}_\phi \boldsymbol{\Omega} \mathbf{Q}_\phi^\top \mathbf{u}_i}. \tag{3.29}$$

To obtain the minimal fraction of confounding, the quantity

$$\frac{\mathbf{u}_i^\top \mathbf{Q}_\phi \mathbf{Q}_\phi^\top \mathbf{u}_i}{\mathbf{u}_i^\top \mathbf{Q}_\phi \boldsymbol{\Omega} \mathbf{Q}_\phi^\top \mathbf{u}_i}$$

should be maximized. In order to do that, consider the linear transformation $\mathbf{J}\hat{\mathbf{e}}$ such that $\mathbf{J}\hat{\mathbf{e}}$ is least confounded, that is, the rows of $\mathbf{J}$ are those $\mathbf{j}_i$ such that $\mathbf{j}_i^\top \hat{\mathbf{e}}_\phi$ is least confounded. Then, one should maximize the quantity

$$\xi = \frac{\mathbf{j}^\top \mathbf{Q}_\phi \mathbf{Q}_\phi^\top \mathbf{j}}{\mathbf{j}^\top \mathbf{Q}_\phi \boldsymbol{\Omega} \mathbf{Q}_\phi^\top \mathbf{j}}$$

The maximization problem is solved using the method described in Ghojogh, Karray and Crowley (2019, Sec. 7).

3.4.3 BLUP

The random effects may also be used to identify outlying observations, as pointed out by Nobre and Singer (2007) "$\mathbf{Z}^*\hat{\mathbf{b}}_i$ reflects the difference between the predicted response for the ith subject and the population average." One possible way to make use of the predicted random effects as a tool for outlying detection, is to plot $\hat{\mathbf{b}}_i$ versus subject indices. An alternative to this measure, is the Mahalanobis distance

$$M_i = \hat{\mathbf{b}}_i^\top \left[\widehat{Var}\left(\hat{\mathbf{b}}_i - \mathbf{b}_i\right)\right]^{-1} \hat{\mathbf{b}}_i, \tag{3.30}$$

where $\hat{\mathbf{b}}_i$ is the posterior mean as in Equation 2.29. This measure can either be plotted versus subject indices to identify the outlying observations, or as QQ plot based on a χ^2_q distribution.

Once again, however, due to the non-linear nature of the lasso solution, an explicit expression for the variance is impossible to be obtained. On the other hand, given ϕ, it is possible to use the approximate solution for the fixed effects in Equation 3.12 to approximate the random effects by

$$\begin{aligned}\hat{\mathbf{b}}_{i\phi} &= \left[\mathbf{Z}_i^{*\top}\mathbf{W}_i\mathbf{Z}_i^* + \mathbf{I}_q\right]^{-1}\mathbf{Z}_i^{*\top}\mathbf{W}_i(\mathbf{y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{\phi}^{lasso}) \\ &= \mathbf{V}_i^{-1}\mathbf{Z}_i^{*\top}\mathbf{W}_i(\mathbf{y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{\phi}^{lasso}) \\ &= \mathbf{V}_i^{-1}\mathbf{Z}_i^{*\top}\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{i\phi}\right)\mathbf{y}_i^*, \end{aligned} \tag{3.31}$$

and the variance (LAIRD; WARE, 1982) $\widehat{Var}\left(\hat{\mathbf{b}}_i - \mathbf{b}_i\right)$ may be approximated by

$$\widehat{Var}\left(\hat{\mathbf{b}}_{i\phi} - \mathbf{b}_i\right) = \hat{\sigma}^2\left\{\mathbf{I}_q - \mathbf{V}_i^{-1}\mathbf{Z}_i^{*\top}\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{i\phi}\right)\hat{\boldsymbol{\Omega}}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{i\phi}\right)^{\top}\mathbf{W}_i\mathbf{Z}_i^*\mathbf{V}_i^{-1}\right\}. \tag{3.32}$$

Thus we can define

$$M_{i\phi} = \hat{\mathbf{b}}_{i\phi}^{\top}\left[\widehat{Var}\left(\hat{\mathbf{b}}_{i\phi} - \mathbf{b}_i\right)\right]^{-1}\hat{\mathbf{b}}_{i\phi}, \tag{3.33}$$

as the approximated Mahalanobis distance for the random effects.

CHAPTER

4

APPLICATION TO A REAL DATA SET

The purpose of this chapter is to fit both normal and robust models to the CASTNet data set and compare their performance using the techniques presented in Chapter 3.

4.1 Approaches

For both the normal and robust approaches, a LME model was fitted to the CASTNet data set, as follows

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i^* + \boldsymbol{\varepsilon}_i \quad i = 1,\ldots,m, \tag{4.1}$$

where $\mathbf{y}_i$ denotes the $(n_i \times 1)$ response vector for the i-th site, that is the log of the total nitrate concentration for the i-th site, the matrix $\mathbf{X}_i$ denotes the $(n_i \times p)$ design matrix of fixed effects without the intercept, as presented in Chart 1, the vector $\boldsymbol{\beta}$ denotes the $(p \times 1)$ vector of unknown fixed effects that will be estimated, $\mathbf{Z}_i$ denotes the $(q \times 1)$ random effects design matrix for the i-th and its values are the same as the fixed effects including an intercept, that is $\mathbf{Z}_i = [\mathbf{1}_i \quad \mathbf{X}_i]$, $\mathbf{b}_i^*$ denotes the $(q \times 1)$ vector of random effects, assumed to be distributed as $\mathbf{b}_i^* \sim N_q(\mathbf{0}, \sigma^2\mathbf{G})$, and $\mathbf{G}$ will be decomposed using the method presented in section 2.2, lastly, the error term $\boldsymbol{\varepsilon}_i$ is a $(n_i \times 1)$ vector assumed to be distributed as $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, that is, the homoscedastic conditional independence model.

Equation 4.1 presents the model that will be fitted using the approaches presented back in Chapter 2. Both approaches will be, then, compared using the techniques presented in Chapter 3. It is worth noting that all the techniques presented already exist either in the context of maximum likelihood or least squares estimation of LME models, or in the context of normal multiple regression models, or in the lasso context under multiple regression models. The aim of this work, is to either extend or combine such techniques in a way that models fitted using the methods such as the ones proposed by Bondell, Krishna and Ghosh (2010) and Fan, Qin and Zhu (2014) could be further analyzed.

4.2 Fitted models

Both the normal and robust approaches were fitted to the data set, and the resulting estimated fixed effects are presented, along with its standard errors (obtained from Equation 3.13) and relative change, in Table 3. The relative changes (RC) were calculated as follows

$$RC_i = 100 * \frac{\beta_i^{new} - \beta_i^{ref}}{\beta_i^{ref}},$$

where β_i^{ref} is the value taken as reference (the i-th estimate of the normal fit, for example) and β_i^{new} is the value that is to be compared (the i-th estimate of the robust fit, for example).

Table 3 – Estimates of the fixed effects and standard errors under normal and robust approaches.

	Normal	**Robust**	**Relative change (%)**
`sulphate`	-0,026255 (0,199290)	-0,036743 (0,000198)	39,94668
`ammonia`	0,142092 (0,224992)	0,196646 (0,001206)	38,39344
`ozone`	0,097812 (0,120888)	0,079932 (0,000271)	-18,27997
`atemp`	0 (0,089332)	0 (0,086880)	-
`adptemp`	0 (0,089904)	0 (0,086885)	-
`humidity`	-0,030033 (0,064965)	-0,016239 (0,000118)	-45,92948
`radiation`	0 (0,163125)	0 (0,035799)	-
`windspeed`	0 (0,036370)	0 (0,035595)	-
`precipitation`	-0,021885 (0,045337)	-0,027328 (0,000208)	24,87092
`time.in.months`	-0,002284 (0,003334)	-0,001406 (0,000179)	-38,44133
`s1`	0,234566 (0,163718)	0,281805 (0,001661)	20,13889
`c1`	0,323886 (0,287495)	0,356987 (0,000682)	10,21995
`s2`	-0,015338 (0,059650)	**0** (0,049089)	-
`c2`	0 (0,053244)	0 (0,049536)	-
`s3`	0 (0,050643)	0 (0,048963)	-
`c3`	0 (0,050685)	0 (0,049578)	-

Source: Elaborated by the author.

Table 4 presents the estimates for the square root of the diagonal of the estimated random effects covariance matrix for the normal and robust approaches, respectively, along with relative change.

Note that in both tables, if a coefficient has an estimate of 0, it means that it was removed by the lasso procedure. Bold values in both tables indicate that the associated coefficient is either removed or included, compared to the normal approach.

It is worth noticing that as there are changes between the models fitted by the two approaches, mainly the inclusion or exclusion or variables, the conclusions from each of the models are different. For example, variable `s2` for the fixed effects is removed in the robust approach, which means that it does not impact the response under this approach. As for the

Table 4 – Predictors of the random effects under normal and robust approaches.

	Normal	**Robust**	**Relative change (%)**
`int`	0,256995719	0,30679025	19,375627
`sulphate`	0,086267984	**0**	-
`ammonia`	0,112850831	**0**	-
`ozone`	0	0	-
`atemp`	0	0	-
`adptemp`	0	0	-
`humidity`	0	0	-
`radiation`	0	**0,01581101**	-
`windspeed`	0	0	-
`precipitation`	0	0	-
`time.in.months`	0,001245839	**0**	-
`s1`	0,074284367	0,06792454	-8,561465
`c1`	0,087790862	**0**	-
`s2`	0	0	-
`c2`	0	0	-
`s3`	0	0	-
`c3`	0	0	-

Source: Elaborated by the author.

random effects, removing or including a variable affects the estimate for the covariance matrix, which impacts mainly the construction of the diagnostic and residual measures that depend on the covariance matrix.

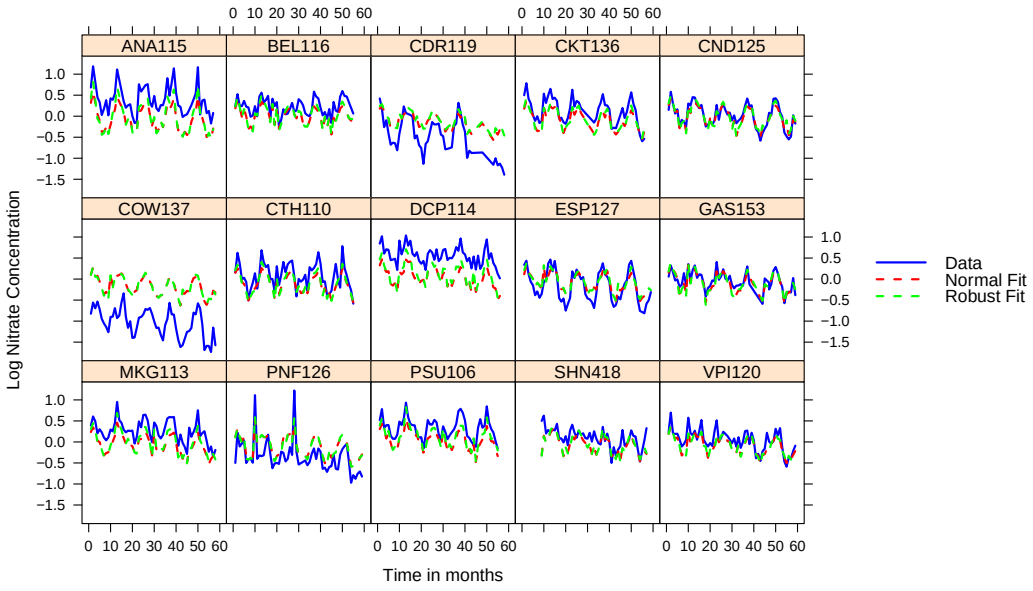

Figure 5 – Profile plot of response variable and fitted values over time for normal and robust fits.

Source: Elaborated by the author.

Profile plots of the observed response and fitted values over time are presented in Figure 5 for both the normal (dashed red line) and robust (dashed green line) approaches. These figures indicate the behavior of the residuals that will be calculated. For instance, for both approaches, site CDR119 does not appear to have a proper fit for all observations observations; site COW137 fitted values are far from the observed data; and site DCP114 also seems to have fitted values far from the observed response.

4.3 A naive residual analysis

To begin with the comparison between the fitted models, the standardized marginal residuals, $\mathscr{R}_{i\phi} = \left[\widehat{Var}(\hat{\boldsymbol{\varepsilon}}_\phi)\right]^{-1/2} \hat{\boldsymbol{\varepsilon}}_\phi$, will be first analyzed. Figure 6 and Figure 7 present the scatter-plot for the standardized marginal residuals versus observations indices for both normal and robust fit, respectively. In both figures, a local smoother (LOWESS) is fitted but only for exploratory purposes, that is, to highlight possible trends in the residuals.

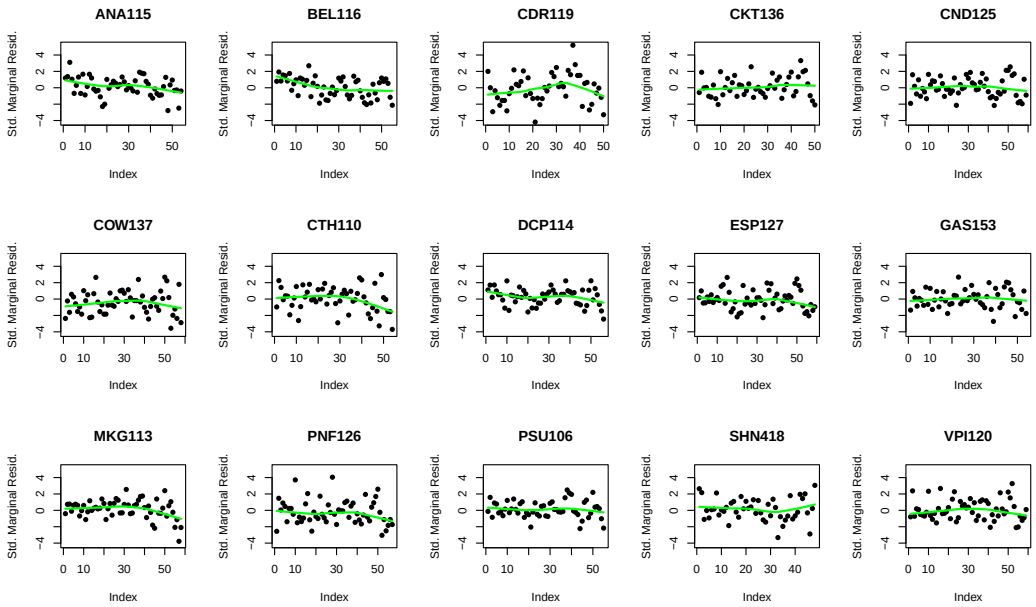

Figure 6 – Standardized marginal residuals for the normal method.

Source: Elaborated by the author.

Note – The green line is a local smoother (LOWESS) that aids visualizing possible trends in the residuals.

In Figure 6, the residuals associated with each of the sites seem to be, overall, randomly dispersed around 0. Sites CDR119, COW137, CTH110, DCP114, GAS153, MKG113, PNF126, SHN418 draws attention as some values for their marginal residuals seem to be large, which could indicate outlying observations. However, it is not possible to attest that with only this analysis. In the next section, the conditional residuals will be used to that end.

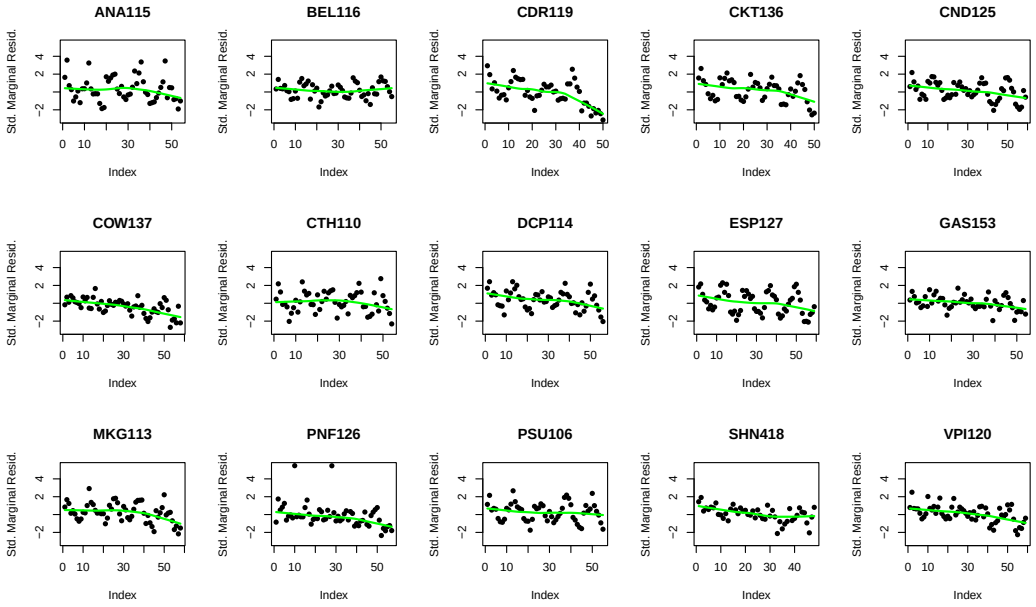

Figure 7 – Standardized marginal residuals for the robust method.

Source: Elaborated by the author.

Note – The green line is a local smoother (LOWESS) that aids visualizing possible trends in the residuals.

Figure 7 present the standardized marginal residuals versus observation indices for the robust fit. In this case, the residuals, overall, present trends, which could be an indicative of misspecification of the model. As for possible outlying observations, sites ANA115 and PNF126 are the ones that draws attention. Once again, in the next section, appropriate measures for detecting misspecification and outlying observations will be used to further analyze sites and observations.

A more thorough analysis of the residuals should be carried out, in order to assess the goodness of the fitted models to the data.

4.4 Further investigating the residuals

Using the standardized marginal residual for both models, the standardized Lesaffre-Verbeke (LV) measure defined in Equation 3.23. Figure 8 presents the graphs for the standardized LV measure, normal fit on the left and robust fit on the right. As a cutoff point was not defined, the analysis of this measure is subjective. Recall that this measure indicates whether the chosen covariance structure is adequate to the individual, or site in this case.

The graph in Figure 8 seems to indicate that the chosen covariance structure for the normal fit is not appropriate for site CDR119 as it stands out from the other sites. Both graphs are on the same scale, and overall the robust fit covariance seems to be more adequate, probably due to the introduction of the weight matrix in the estimation procedure.

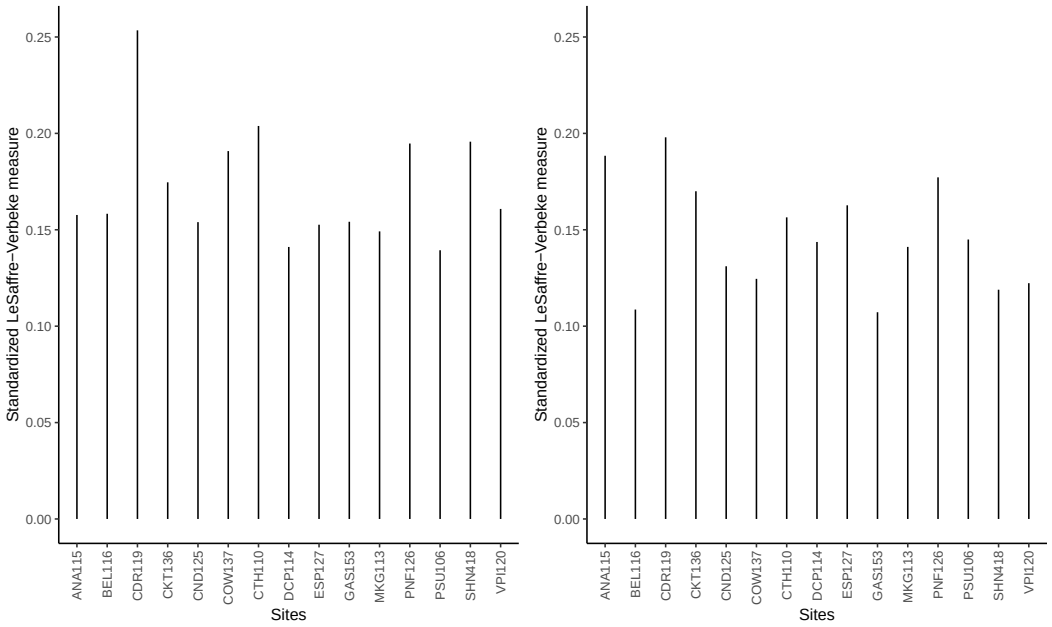

Figure 8 – Lesaffre-Verbekke measure for the normal (left) and robust (right) fit.

Source: Elaborated by the author.

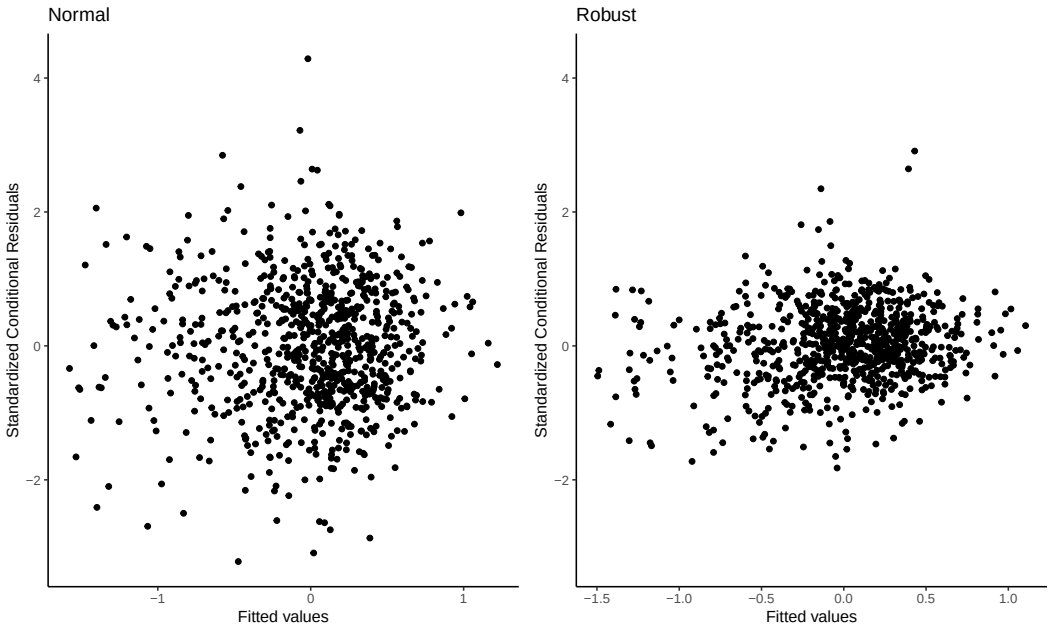

Figure 9 – Standardized conditional residuals versus conditional fitted for the normal (left) and robust (right) fits.

Source: Elaborated by the author.

Figure 9 presents the plot of the standardized conditional residuals versus the conditional fitted values for both normal (left) and robust (right) fits. Note that both graphs do not present an underlying structure, which could indicate that the model chosen, in both cases, is adequate. Also, this graph assists on checking for possible outliers.

Note that the normal approach seems to present more possible outliers than the robust fit, as there are more observations above and below ± 2 lines on the y-axis of the graphs. If one chooses to be more conservative in the search of outliers and assume cutoffs of ± 3, the robust approach does not present any outliers, whereas the normal fit presents a few outlying observations.

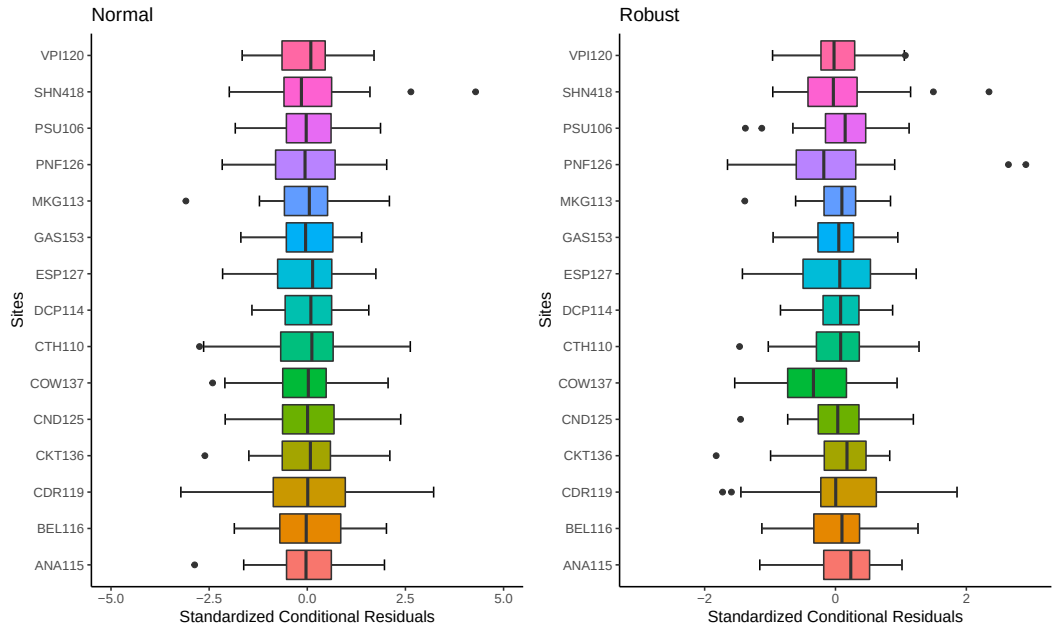

Figure 10 – Boxplots for the standardized conditional residuals for the normal (left) and robust (right) fits.

Source: Elaborated by the author.

The boxplots of the standardized conditional residuals by sites in Figure 10 aids to the search for outlying observations, as it presents the distribution of such residuals in each site, and which observations are outliers inside each site. And also, using an overall cutoff of, say, ± 2 for both the normal and robust fits, it is possible to identify which sites present outlying observations in each of the fits.

Before analyzing the QQ-plots for the least confounded residuals, it is worth assessing whether the transformed residuals are uncorrelated. In order to do that, line plots for the residuals associated to each of the sites are presented in Figure 11. Notice that, for the normal fit on the left, some sites present trend in their residuals, and also several of the sites present large residuals. As for the robust fit, the residuals associated to each of sites overall do not seem to present neither trend nor large values.

Analyzing trend and outlying values alone do not properly address the fact of whether the least confounded residuals are in fact uncorrelated. To assess the correlation, the Box-Pierce test (BOX; PIERCE, 1970) was used, via the `Box.test` function from the `stats` R package (R Core Team, 2021). Table 5 presents the test statistics as well as the associated p-value for each of the sites least confounded residuals. All of the residuals associated to each of the sites reject

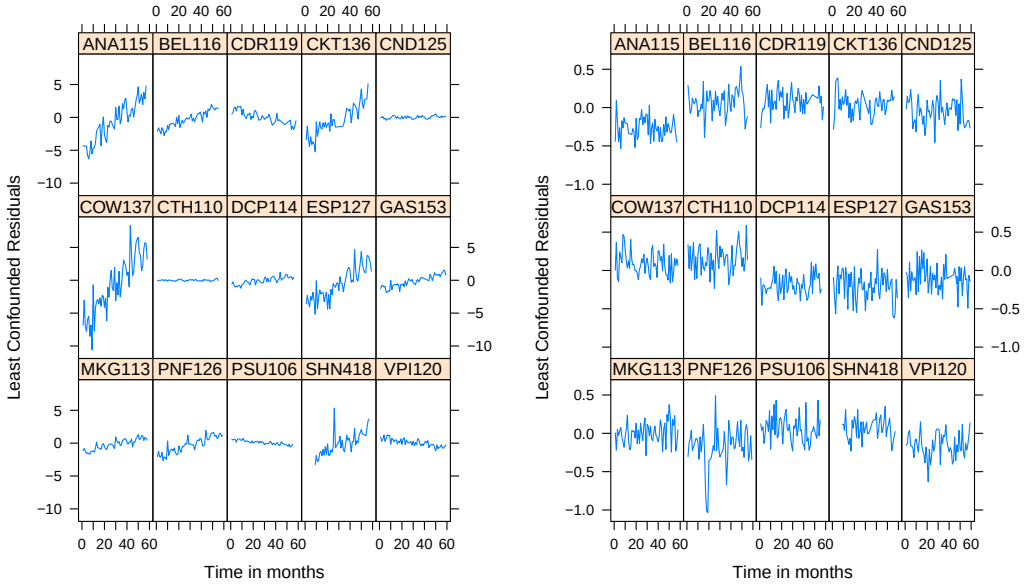

Figure 11 – Line plot for the least confounded residuals - normal (left) and robust (right).

Source: Elaborated by the author.

Table 5 – Box-Pierce correlation test.

	Normal		Robust	
Site	Statistic	p-value	Statistic	p-value
ANA115	36,417	$< 10^{-8}$	1,683	0,194
BEL116	36,106	$< 10^{-8}$	3,199	0,074
CDR119	30,031	$< 10^{-7}$	0,056	0,813
CKT136	27,423	$< 10^{-6}$	3,669	0,055
CND125	12,154	$< 10^{-3}$	1,120	0,290
COW137	30,214	$< 10^{-7}$	0,819	0,365
CTH110	6,335	0,012	0,042	0,838
DCP114	26,408	$< 10^{-6}$	0,006	0,938
ESP127	30,868	$< 10^{-7}$	0,094	0,759
GAS153	32,286	$< 10^{-7}$	6,929	**0,008**
MKG113	36,500	$< 10^{-8}$	0,146	0,702
PNF126	33,158	$< 10^{-8}$	10.788	**0,001**
PSU106	29,601	$< 10^{-7}$	0,293	0,588
SHN418	10,156	0,001	0,104	0,747
VPI120	15,831	$< 10^{-4}$	3,921	0,048

Source: Elaborated by the author.

the null hypothesis of independence at a 5% significance level. As for the robust fit, only two of the sites (boldfaced) reject the null hypothesis at a 5% significance level.

Although the least confounded residuals of the normal present correlation, QQ-plots

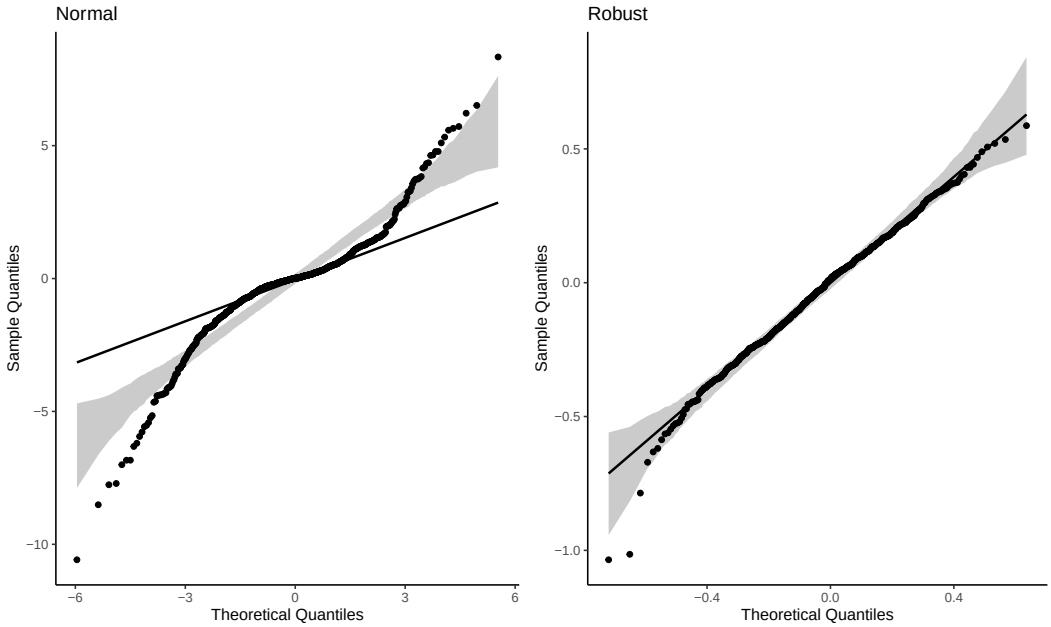

Figure 12 – QQ-plots LC Res.

Source: Elaborated by the author.

for both the normal and robust approaches will be constructed in order to compare the fitting procedures. Figure 12 present the QQ-plots for both the normal (left) and robust (right) fits; the confidence bands were obtained via bootstrap option from the `qqplor` package (ALMEIDA; LOY; HOFMANN, 2018), with confidence level of 0,99. These graphs are used to verify the normality of the conditional errors. This figure suggests that the normal fit residuals are not normally distributed, whereas the robust fit is normally distributed, according to the QQ-plots. In this sense, the robust fit could be considered as a far more adequate model to be fitted to the CASTNet data. Along with the lack of normality of the normal fit, the trends observed in Figure 11 and hypothesis test from Table 5, could be further evidence for either the misspecification of the covariance structure or that a different error distribution should be used.

One last measure based on the residuals is presented in Figure 13, the EBLUP measure. Recall that this measure indicates the presence of outlying subjects or, in this case, sites. In order to make these graphs comparable, the values were standardized, following the standardized LV measure, that is $EBLUP_i^* = \sqrt{EBLUP_i}/n_i$. Comparing these two graphs, the normal fit seems to have more outlying sites than the robust fit. However, one can also analyze each fit on its own.

Then, for the normal fit, sites ANA115, COW137 and DCP114 draws attention, as they have the larger values for the standardized EBLUP measure. As for the robust fit, sites CDR119, COW137 and DCP114 are the ones that stand above the others.

Notice that throughout the analysis proposed in this Chapter, up to this point, sites ANA115, CDR119, COW137, DCP114 and PNF126 were recurrent during the search of outlying observation and sites, misspecification of the covariance structure.

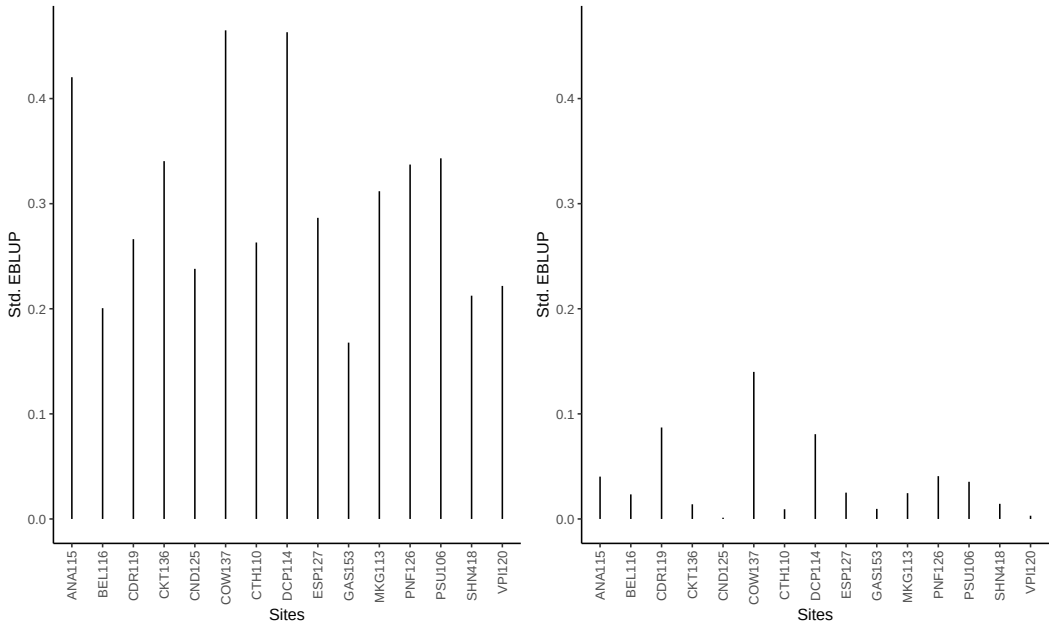

Figure 13 – EBLUP for normal (left) and robust (right) fits.

Source: Elaborated by the author.

4.5 Diagnostic and influential analysis

This section is dedicated to the diagnostic and influential analysis that the sites have on the estimation process. The different measures presented throughout this section aim to asses whether a site affects the selection procedure or the coefficients. Also, it is presented leverage measures, in order to identify the outliers with respect to the explanatory variables.

The first influence measure presented here is the Df-model, which assess whether an observation or a set of observations has influence over the model selection procedure. If an observation, in the case of this work a set of observations as the sites will be assessed, is outside the cutoffs of ± 2, then it is considered influential on the selection procedure.

Figure 14 presents such measure. Notice that for the normal fit (left) none of the sites is considered to be influential on the selection procedure. As for the robust fit (right), site PSU106 is considered influential on the model selection procedure.

The next measure analyzed is the Df-lambda, which asses whether a site is influential on the choice of the tuning parameter. Figure 15 presents such measure for both the normal (left) and robust (right) fits. Neither of the fits present sites above the cutoffs of ± 2, indicating that none of the sites is influential on the choice of the tuning parameter.

Next, the Cook's distance will be assessed. This measure assess whether a site is influential on estimation of the parameters. Figure 16 presents this measure for both approaches, normal on the left and robust on the right. Site SNH418 is highly influential on the normal fit; site CDR119 also seems to be influential, according to the graph. None of the sites seem to be

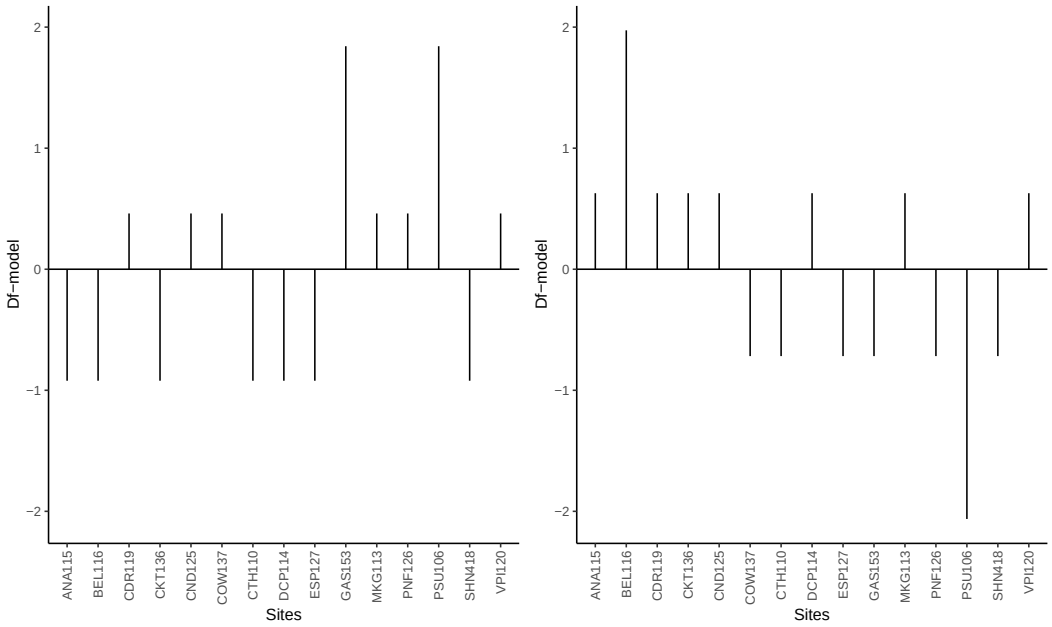

Figure 14 – Df-model for normal (left) and robust (right) fit.

Source: Elaborated by the author.

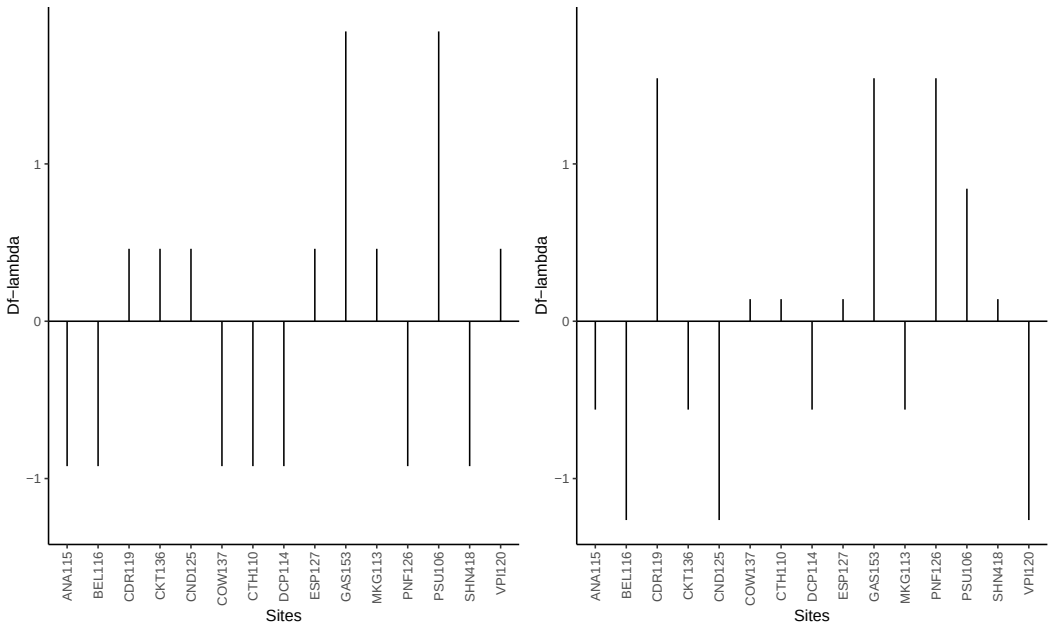

Figure 15 – Df-lambda for normal (left) and robust (right) fit.

Source: Elaborated by the author.

influential for the robust fit.

It is worth noting that the Cook's distance used in this work measures an "average influence", in a sense that the fixed effects and random effects are confounded (similar to the least confounded residuals), which could explain why site SNH418 stands out from the other sites in the normal fit. It was not possible yet to find the expression for the conditional Cook's

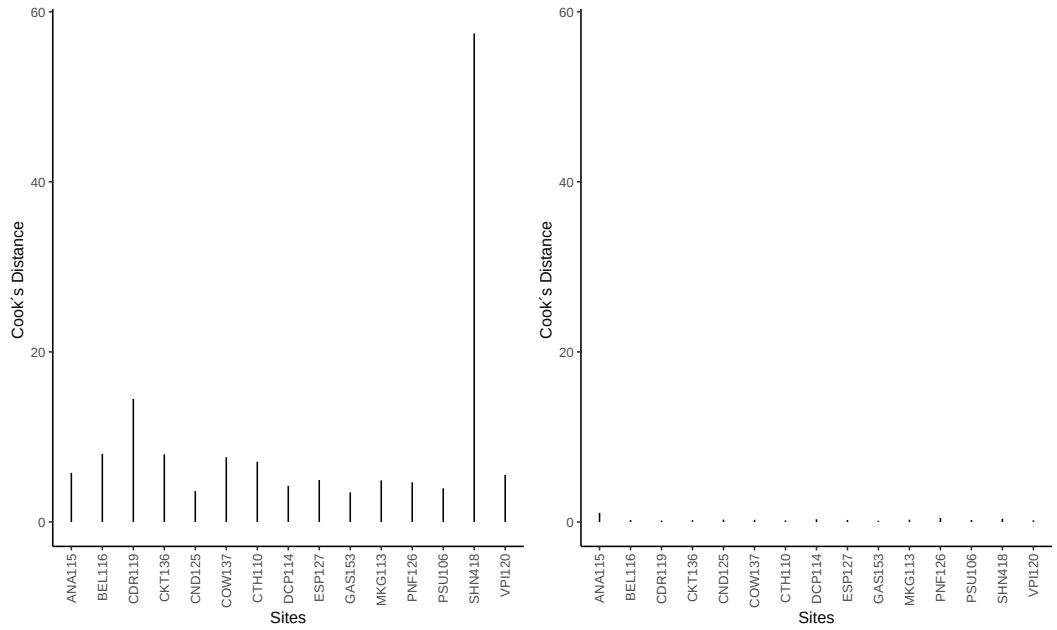

Figure 16 – Cook's distance for normal (left) and robust (right) fit.

Source: Elaborated by the author.

distance (TAN; OUWENS; BERGER, 2001; PINHO; NOBRE; SINGER, 2015), which deals with this "confounding" effect.

Lastly, a leverage analysis of the sites will be presented. Figure 17 presents the approximated leverage measure for the fixed effects for the normal (left) and robust (right) fits. For the normal fit, sites BEL116, CDN125, GAS153, PNF126, PSU106 and SNH418 are high-leverage points, as they cross the indicated cutoff. As for the robust fit, none of the sites were considered high-leverage points. This could be an indicative of the aim of the robustness, that in the case of this work, is to better fit outlying observations. And as leverage points could also be interpreted as outliers with respect to the explanatory observations, this means that the correction that the weight matrix used in the robust process is effective, somehow, in order to decrease the effect that the high-leverage site could have on the estimation of the fixed effects.

Following with the leverage point analysis, Figure 18 presents the leverage measure for the random effects. Notice that in both fits, all observations could be considered as high-leverage points, but notice that the main difference between the graphs is that the sites on the robust fit (right) all have approximately the same value for this leverage measure, whereas for the normal fit (left) there are a few sites that stand out form the others. Namely, the sites that have the highest values for this measure are ANA115, DCP114 and PSU106.

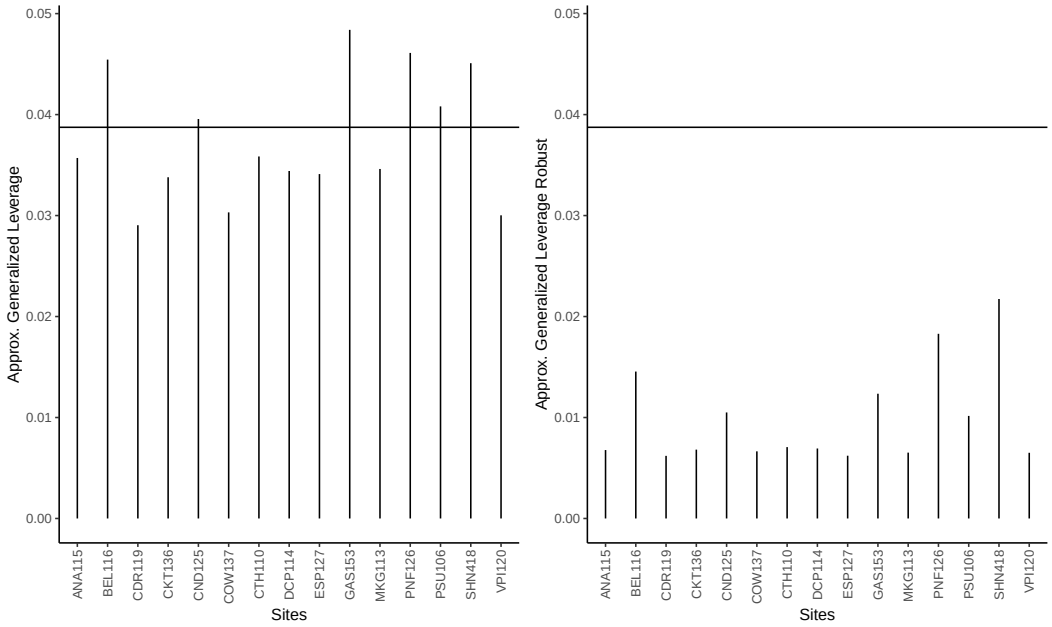

Figure 17 – Approximated generalized leverage values for fixed effects.

Source: Elaborated by the author.

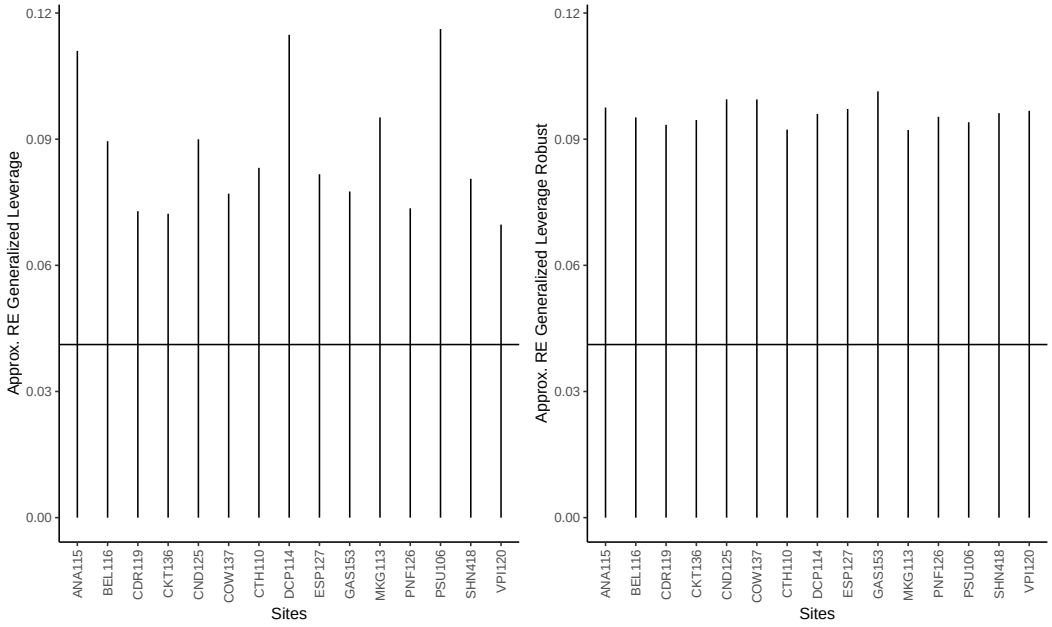

Figure 18 – Approximated generalized leverage values for random effects.

Source: Elaborated by the author.

4.6 Removing specific observations to assess their influence

A few of the sites that were recurrent on the analysis on the previous sections were chosen to have their influence assessed. The sites chosen were CDR119, COW137, DCP114 and PNF126.

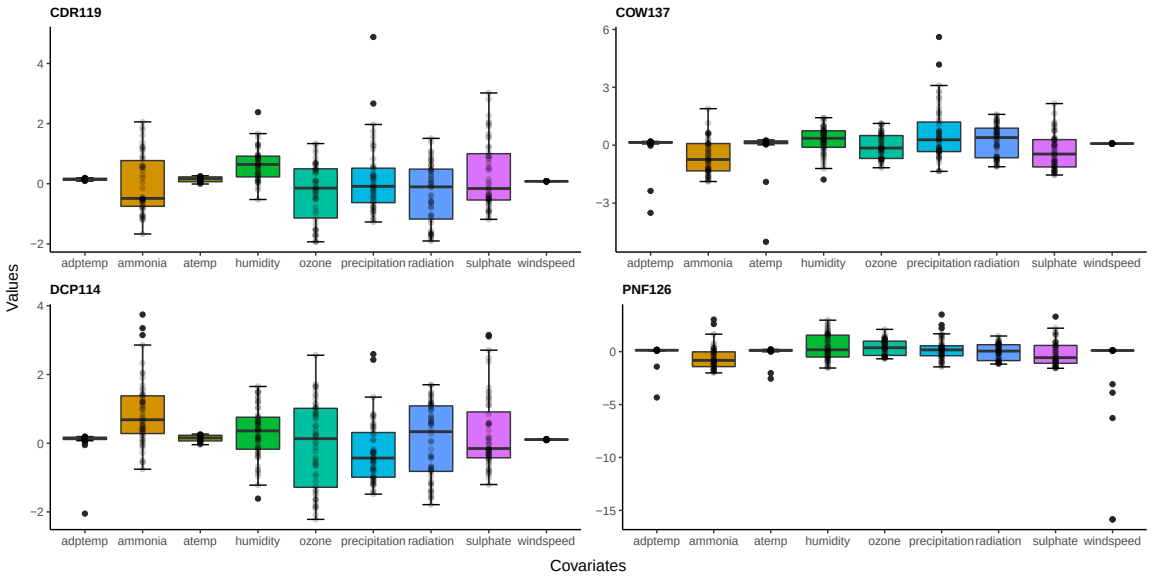

Figure 19 – Boxplots for the explanatory variables of specific sites.

Source: Elaborated by the author.

Figure 19 presents boxplots for the explanatory variables of such sites. Recall that the design matrices for both the fixed and random effects were standardized, as a requirement for the lasso procedure. As these variables were standardized, it is expected that they are symmetrically distributed around 0, which is not the case, as some variables are skewed. For example, the `ammonia` variable for all sites is displaced from 0. It is worth noticing that most of these variables have outlying obsrevations, for example, `windspeed` for site PNF126 has large values for this specific variables, when compared to the other sites; `precipitation` for site COW137 also seems to have large values than other sites.

The four sites above were removed and estimation was carried out without them, each one at a time. Table 6 and Table 7 present the estimates for the fixed for the model with all sites, along with estimates for the fits without each of the four sites. The last four columns in these tables are the percentage relative change of the estimates when compared to the model with all sites. Bold values indicate whether the associated coefficient was removed or included, compared to the model with all sites. A similar analysis is carried out for the random effects in Table 8 and Table 9.

These tables are similar to Table 3 and Table 4, in a sense that a model is taken as a

reference (full model) and as the observations associated to the sites are removed from the data set, conclusions drawn from such models differ from those related to the reference model.

For instance, Table 6 presents this analysis under the normal approach. Note that removing sites 3 and 6 causes a variable to be deleted and a new one to be added, respectively. A similar situation is seen in Table 7. Removing site 3 causes two variables to be removed from the model, removing site 6 adds a variable, removing site 8 adds two new variables and removing site 12 causes a variable to be removed. For Table 8 and Table 7, removing or including variables, when compared to the model with all the observations, changes the covariance matrix estimates.

Table 6 – Estimates of fixed effects - Normal.

	Full	**W/o Obs. 3**	**W/o Obs. 6**	**W/o Obs. 8**	**W/o Obs. 12**	$\mathbf{RC}_3$	$\mathbf{RC}_6$	$\mathbf{RC}_8$	$\mathbf{RC}_{12}$
sulphate	-0,0263 (0,1993)	**0** (0,0764)	-0,0733 (0,0016)	-0,0282 (0,0074)	-0,0733 (0,2674)	-100	179,0326	7,4272	-100
ammonia	0,1421 (0,2250)	0,1036 (0,0016)	0,2150 (0,0037)	0,1443 (0,0349)	0,2150 (0,0237)	-27,0719	51,2999	1,5814	-12,7241
ozone	0,0979 (0,1209)	0,1008 (0,0034)	0,1053 (0,0032)	0,0926 (0,0079)	0,1053 (0,0595)	3,0436	7,6821	-5,3623	-12,3063
atemp	0 (0,0893)	0 (0,0895)	0 (0,0902)	0 (0,0924)	0 (0,3451)	0	0	0	0
adptemp	0 (0,0899)	0 (0,0899)	0 (0,0898)	0 (0,0938)	0 (0,6652)	0	0	0	0
humidity	-0,0300 (0,0650)	-0,0301 (0,0019)	-0,0222 (0,0007)	-0,0349 (0,0037)	-0,0222 (0,1348)	0,2830	-26,0380	16,3387	16,2321
radiation	0 (0,1631)	0 (0,0893)	$-\mathbf{0,0165}$ (0,0005)	0 (0,1970)	-0,0165 (2,0496)	0	-	0	0
windspeed	0 (0,0364)	0 (0,0371)	0 (0,0362)	0 (0,0377)	0 (0,0989)	0	0	0	0
precipitation	-0,0219 (0,0453)	-0,0164 (0,0014)	-0,0228 (0,0008)	-0,0215 (0,0029)	-0,0228 (0,0104)	-25,2365	4,3866	-1,5307	0,7539
time.in.months	-0,0023 (0,0033)	-0,0022 (0,0057)	-0,0027 (0,0010)	-0,0023 (0,0431)	-0,0027 (0,4609)	-2,9772	19,3958	2,7583	18,6515
s1	0,2346 (0,1637)	0,2194 (0,0062)	0,2647 (0,0049)	0,2299 (0,0095)	0,2647 (0,1842)	-6,4728	12,8582	-2,0024	4,0790
c1	0,3239 (0,2875)	0,3051 (0,0050)	0,3547 (0,0067)	0,3267 (0,0099)	0,3547 (0,2133)	-5,7996	9,5166	0,8824	0,3260
s2	-0,0153 (0,0596)	-0,0080 (0,0005)	-0,0000 (0)	-0,0122 (0,0022)	-0,0000 (0,0023)	-47,4638	-99,9087	-20,3677	17,7272
c2	0 (0,0532)	0 (0,0595)	0 (0,0522)	0 (0,0619)	0 (0,7024)	0	0	0	0
s3	0 (0,0506)	0 (0,0513)	0 (0,0511)	0 (0,0571)	0 (0,4966)	0	0	0	0
c3	0 (0,0507)	0 (0,0518)	0 (0,0517)	0 (0,0535)	0 (0,1153)	0	0	0	0
$\hat{\sigma}^2$	0,0146	0,0137	0,0146	0,0151	0,0141				
λ	0,3000	0,2500	0,3000	0,3000	0,3000				
BIC	-772,4134	-776,5118	-676,2743	-685,4313	-696,3553				

Source: Elaborated by the author.

Note – RC_i refers to the percentage relative change in the fixed effect estimates without the i-th subject.

Table 7 – Estimates of fixed effects - Robust.

	Full	**W/o Obs. 3**	**W/o Obs. 6**	**W/o Obs. 8**	**W/o Obs. 12**	**RC$_3$**	**RC$_6$**	**RC$_8$**	**RC$_{12}$**
sulphate	-0,0367 (0,0002)	**0** (0,0556)	-0,1350 (0,0057)	-0,0655 (0,0003)	**0** (0,0569)	-100	267,4931	78,3088	-100
ammonia	0,1966 (0,0012)	0,1535 (0,0039)	0,2963 (0,0112)	0,2316 (0,0010)	0,1251 (0,00001)	-21,9572	50,6789	17,7995	-36,3613
ozone	0,0799 (0,0003)	0,0797 (0,0024)	0,1168 (0,0054)	0,0976 (0,0004)	0,0750 (0,000007)	-0,2414	46,0829	22,1351	-6,2040
atemp	0 (0,0869)	0 (0,0872)	0 (0,0876)	0 (0,0878)	0 (0,0881)	0	0	0	0
adptemp	0 (0,0869)	0 (0,0870)	0 (0,0871)	0 (0,0884)	0 (0,0884)	0	0	0	0
humidity	-0,0162 (0,0001)	-0,0259 (0,0015)	-0,0147 (0,0007)	-0,0360 (0,0001)	-0,0183 (0,000003)	59,4310	-9,3971	121,5284	12,6978
radiation	0 (0,0358)	0 (0,0530)	$-\mathbf{0,0691}$ (0,0032)	$-\mathbf{0,0684}$ (0,0003)	0 (0,0527)	0	-	-	0
windspeed	0 (0,0356)	0 (0,0357)	0 (0,0356)	0 (0,0355)	0 (0,0632)	0	0	0	0
precipitation	-0,0273 (0,0002)	**0** (0,0392)	-0,0341 (0,0018)	-0,0394 (0,0002)	-0,007949 (0,000002)	-100	24,7036	44,0866	-70,9126
time.in.months	-0,0014 (0,0002)	-0,0004 (0,0007)	-0,0005 (0,0008)	-0,0020 (0,0001)	-0,000813 (0,000003)	-70,9815	-60,7397	41,1095	-42,1764
s1	0,2818 (0,0017)	0,2476 (0,0093)	0,2764 (0,0093)	0,2680 (0,0009)	0,25935 (0,00003)	-12,1414	-1,9276	-4,9013	-7,9683
c1	0,3570 (0,0007)	0,3489 (0,0055)	0,3055 (0,0096)	0,2816 (0,0009)	0,34517 (0,00002)	-2,2631	-14,4148	-21,1038	-3,3377
s2	0 (0,0491)	0 (0,0542)	0 (0,0510)	$-\mathbf{0,01387}$ (0,00004)	0 (0,05461)	0	0	-	0
c2	0 (0,0495)	0 (0,0512)	0 (0,0513)	0 (0,0513)	0 (0,0515)	0	0	0	0
s3	0 (0,0490)	0 (0,0509)	0 (0,0508)	0 (0,0507)	0 (0,0510)	0	0	0	0
c3	0 (0,0496)	0 (0,0512)	0 (0,0514)	0 (0,0514)	0 (0,0514)	0	0	0	0
$\hat{\sigma}^2$	0,0740	0,0659	0,0427	0,0596	0,0709				
λ	0,2000	0,1500	0,2500	0,3000	0,1500				
BIC	-194,5579	-259,6946	-425,3721	-260,0390	-163,2266				

Source: Elaborated by the author.

Note – RC_i refers to the percentage relative change in the fixed effect estimates without the i-th subject.

Table 8 – Predictors of random effects - Normal.

	Full	**W/o Obs. 3**	**W/o Obs. 6**	**W/o Obs. 8**	**W/o Obs. 12**	**RC$_3$**	**RC$_6$**	**RC$_8$**	**RC$_{12}$**
int	0,2570	0,2535	0,1726	0,2620	0,2779	-1,3387	-32,8512	1,9376	8,1357
sulphate	0,0863	0,0985	0,0603	0,0810	0,0869	14,1651	-30,0551	-6,0543	0,7157
ammonia	0,1128	0,1223	0,0695	0,1056	0,0964	8,3546	-38,3892	-6,4380	-14,5696
ozone	0	0	0	0	0	0	0	0	0
atemp	0	0	0	0	0	0	0	0	0
adptemp	0	0	0	0	0	0	0	0	0
humidity	0	0	0	0	0,0253	0	0	0	-
radiation	0	0	**0,0399**	0	0	0	-	0	0
windspeed	0	0	0	0	0	0	0	0	0
precipitation	0	0	0	0	0	0	0	0	0
time.in.months	0,0012	**0**	0,0011	0,0014	0,0016	-100	-8,1556	11,9382	24,4378
s1	0,0743	0,0644	0,0770	0,0694	0,0750	-13,3127	3,6930	-6,5718	0,9816
c1	0,0878	0,0802	0,0773	0,0919	0,0719	-8,6870	-11,9544	4,6342	-18,1373
s2	0	0	0	0	0	0	0	0	0
c2	0	**0,0206**	0	0	0	-	0	0	0
s3	0	0	0	0	0	0	0	0	0
c3	0	0	0	0	0	0	0	0	0

Source: Elaborated by the author.

Note – RC_i refers to the percentage relative change in the fixed effect estimates without the i-th subject.

Table 9 – Predictors of random Effects - Robust.

	Full	**W/o Obs. 3**	**W/o Obs. 6**	**W/o Obs. 8**	**W/o Obs. 12**	**RC**$_3$	**RC**$_6$	**RC**$_8$	**RC**$_{12}$
int	0,3068	0,3035	0,2173	0,2427	0,3195	-1,0664	-29,1541	-20,9045	4,1415
sulphate	0	0	0	0	0	0	0	0	0
ammonia	0	0	**0,0355**	0	**0,0098**	0	-	0	-
ozone	0	0	0	0	0	0	0	0	0
atemp	0	0	0	0	0	0	0	0	0
adptemp	0	0	0	0	0	0	0	0	0
humidity	0	**0,0362**	0	0	**0,0394**	-	0	0	-
radiation	0,0158	**0**	0,0493	0,0174	0,0086	-100	211,9768	10,1384	-45,7003
windspeed	0	0	0	0	0	0	0	0	0
precipitation	0	0	0	0	0	0	0	0	0
time.in.months	0	0	0	**0,0015**	0	0	0	-	0
s1	0,0679	**0**	0,0826	0,0760	**0**	-100	21,6661	11,8624	-100
c1	0	0	0	0	0	0	0	0	0
s2	0	0	0	0	0	0	0	0	0
c2	0	0	0	0	0	0	0	0	0
s3	0	0	0	0	0	0	0	0	0
c3	0	0	0	0	0	0	0	0	0

Source: Elaborated by the author.

Note – RC_i refers to the percentage relative change in the fixed effect estimates without the i-th subject.

4.7 Cross-validation

For this section, 10-fold cross-validation (HASTIE; TIBSHIRANI; FRIEDMAN, 2009, Sec. 7.10) was performed in order to further compare the performance of both the normal and robust methods. One of the objectives of using cross-validation is to estimate the prediction error (JAMES *et al.*, 2013, Sec. 5.1). There are a few methods to estimate the prediction error, such as the validation set approach, the leave-one-out cross-validation and the k-fold cross-validation, which was the one chosen in this work.

Table 10 presents the mean squared prediction error (MSPE) for the 10-fold cross-validation procedure. The robust fit has a smaller MSPE value, which indicates that it has a better predictive power.

Table 10 – Mean squared prediction error for the 10-fold cross-validation.

	Normal	Robust
MSPE	0,14451	0,12613

Source: Elaborated by the author.

Table 11 presents the cross-validation confidence intervals for the fixed effects coefficients. Note that if the lower and upper bounds are both zero, the variable was not selected in none of 10 folds.

Table 11 – Cross-validation confidence intervals for the fixed effects.

	Normal		Robust	
	Lower-bound	Upper-bound	Lower-bound	Upper-bound
sulphate	-0,1013775723	0,0310857723	-0,175781939	0,0374823395
ammonia	0,0634405756	0,2382048244	0,138282882	0,3376761179
ozone	0,0706260255	0,1261111745	0,032180161	0,1554194393
atemp	-0,0004929426	0,0003583426	0	0
adptemp	0	0	0	0
humidity	-0,0446989123	-0,0183564877	-0,049551588	-0,0021804124
radiation	-0,0098944683	0,0065422683	-0,092826863	0,0306720628
windspeed	0	0	0	0
precipitation	-0,0323512921	-0,0104523079	-0,055501999	-0,0030538011
time.in.months	-0,0027979241	-0,0017412759	-0,001674195	-0,0009064053
s1	0,1968494773	0,2617469227	0,241606026	0,2803917741
c1	0,2737175754	0,3549400246	0,284181269	0,3723077309
s2	-0,0216213966	0,0025255966	-0,044432795	0,0176667953
c2	0	0	0	0
s3	-0,0042792053	0,0029608053	-0,004136909	0,0030073088
c3	0	0	0	0

CHAPTER

5

FINAL CONSIDERATIONS

The main objective of this master's dissertation was to present residual analysis and diagnostic techniques for the lasso regression for LME models. It arose from the question whether the normal approach by Bondell, Krishna and Ghosh (2010) and the robust approach by Fan, Qin and Zhu (2014) were comparable beyond the BIC value, as the normal model is a particular case of the robust model, assuming $v_{ij} = 0$ and $w_{ij} = 1$ for all observations. Although the BIC-type measure was used to select the best tuning parameter for each of the approaches, it was unclear whether the normal approach, despite having a smaller BIC value for the final model, was indeed better fitted to the data than the robust model. Those BIC values, being $BIC_{normal} = -772,4134$ and $BIC_{robust} = -194,5579$, can be found in Table 6 for the normal method and in Table 7 for the robust approach.

Usual diagnostic techniques and residual analysis for LME models were not usable, as they rely in the closed forms of the estimators for the fixed and random effects, and recall that lasso estimates do not possess explicit expressions. Thus, it was necessary to adapt such measures using the recently developed lasso diagnostic measures (KIM *et al.*, 2015; RAJARATNAM *et al.*, 2019) in order to take into account the caveats of LME regression models. Even though the lasso diagnostic measures are based on an approximated expression for the coefficients vector, in this case the fixed effects vector, they provided useful insights to the fitted models.

For example, it was possible to identify observations and sites that were not properly fitted by the models chosen, either being outliers (that were dealt better by the robust approach) or misspecification of the covariance structure (neither of the approaches seem to be properly fitted). It is important noticing that the robust approach produced normally distributed residuals, which confirms that the initial assumption of normality for the errors was correct for this approach. Despite the residuals not being normally distributed for the normal approach, the conditional residuals did not present any underlying structure, which indicates that the LME model chosen is somewhat well fitted to the data, and perhaps a change in the covariance structure would probably produce normally distributed conditional residuals.

Removing some of the sites was important to assess the impact that these sites had on the estimates, notice from Table 6 to Table 9 that removing specific sites caused some coefficients to be removed from the final model and also cause changes in the estimates, as it is seen on the final four columns of these tables.

Notice back on Figure 19 that all four sites had outlying observations for the explanatory variables. This must have had an impact on the normal fit estimates, and also recall that site PNF126 is also a high-leverage point for the fixed effects and also presents large values for `adptemp` and `windspeed`. Treating the data, or even removal of such observations throughout the data set could improve the effectiveness of the normal fit, however, to highlight the resourcefulness of the robust fit, their original values were kept on the data set.

As future work, studies presented in this dissertation might be extended as simulation studies, for example, in order to assess the behaviour of the diagnostic measures.

It is also possible to change the underlying distribution for both approaches to check whether heavy-tailed or even skewed distributions would better fit the data.

The penalized selection algorithm for the normal approach was obtained directly from the authors of Bondell, Krishna and Ghosh (2010) website [1]. The authors also provided the data set that we used in this work. For the robust approach, the original implementation from the normal approach was adapted in order to account for the modifications. The residuals and diagnostics measures were all implemented by the author if this dissertation. The routines could surely be improved as well, as they usually take 20-40 minutes to fit the models.

All the results were obtained using the software R (R Development Core Team, 2010).

[1] <https://blogs.unimelb.edu.au/howard-bondell/>

BIBLIOGRAPHY

ALMEIDA, A.; LOY, A.; HOFMANN, H. **ggplot2 Compatible Quantile-Quantile Plots in R**. [S.l.], 2018. v. 10, n. 2, 248–261 p. Available: <https://doi.org/10.32614/RJ-2018-051>. Citation on page 59.

BONDELL, H. D.; KRISHNA, A.; GHOSH, S. K. Joint variable selection for fixed and random effects in linear mixed-effects models. **Biometrics**, Wiley Online Library, v. 66, n. 4, p. 1069–1077, 2010. Citations on pages 21, 22, 23, 31, 34, 35, 37, 51, 71, and 72.

BOX, G. E. P.; PIERCE, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. **Journal of the American Statistical Association**, Informa UK Limited, v. 65, n. 332, p. 1509–1526, dec 1970. Citation on page 57.

BUSCEMI, S.; PLAIA, A. Model selection in linear mixed-effect models. **AStA Advances in Statistical Analysis**, Springer Science and Business Media LLC, oct 2019. Citation on page 22.

CRUZ, R. M. da. **Critérios de informação e seleção de modelos lineares mistos**. Master's Thesis (Master's Thesis) — Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2020. Available: <https://doi.org/10.11606/D.45.2020.tde-17082020-100010>. Citation on page 22.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citation on page 32.

DIGGLE, P. **Analysis of longitudinal data**. Oxford New York: Oxford University Press, 2002. ISBN 9780198524847. Citation on page 31.

FAN, Y.; QIN, G.; ZHU, Z. Y. Robust variable selection in linear mixed models. **Communications in Statistics - Theory and Methods**, Informa UK Limited, v. 43, n. 21, p. 4566–4581, oct 2014. Citations on pages 22, 37, 51, and 71.

GHOJOGH, B.; KARRAY, F.; CROWLEY, M. **Eigenvalue and Generalized Eigenvalue Problems: Tutorial**. 2019. Available: <https://arxiv.org/abs/1903.11240>. Citation on page 48.

GHOSH, S. K.; BHAVE, P. V.; DAVIS, J. M.; LEE, H. Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. Informa UK Limited, v. 105, n. 490, p. 538–551, jun 2010. Citation on page 23.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. [S.l.]: Springer-Verlag New York Inc., 2009. ISBN 978-0387848570. Citations on pages 32, 33, and 70.

HILDEN-MINTON, J. A. **Multilevel diagnostics for mixed and hierarchical linear models**. Phd Thesis (PhD Thesis) — University of California, 1995. PhD thesis in Mathematics. Available: <https://hdl.handle.net/10568/81585>. Citations on pages 46 and 47.

HOLLAND, P. W. **Weighted ridge regression: Combining ridge and robust regression methods**. [S.l.], 1973. Citation on page 41.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning**. [S.l.]: Springer-Verlag GmbH, 2013. ISBN 978-1-4614-7138-7. Citations on pages 21, 33, 34, and 70.

KIM, C.; LEE, J.; YANG, H.; BAE, W. Case influence diagnostics in the lasso regression. **Journal of the Korean Statistical Society**, Springer Science and Business Media LLC, v. 44, n. 2, p. 271–279, jun 2015. Citations on pages 22, 42, and 71.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. **Biometrics**, [Wiley, International Biometric Society], v. 38, n. 4, p. 963–974, 1982. ISSN 0006341X, 15410420. Available: <http://www.jstor.org/stable/2529876>. Citation on page 49.

LESAFFRE, E.; VERBEKE, G. Local influence in linear mixed models. **Biometrics**, JSTOR, v. 54, n. 2, p. 570, jun 1998. Citation on page 46.

MCCULLOCH, J. H. Miscellanea: On heteros*edasticity. **Econometrica**, [Wiley, Econometric Society], v. 53, n. 2, p. 483–483, 1985. ISSN 00129682, 14680262. Available: <http://www.jstor.org/stable/1911250>. Citation on page 30.

MILLER, A. **Subset selection in regression**. [S.l.]: Chapman and Hall/CRC, 2002. Citations on pages 21 and 32.

NOBRE, J. S.; SINGER, J. da M. Residual analysis for linear mixed models. **Biometrical Journal**, Wiley, v. 49, n. 6, p. 863–875, jun 2007. Citations on pages 22, 45, 46, 47, and 48.

NOBRE, J. S.; SINGER, J. M. Leverage analysis for linear mixed models. **Journal of Applied Statistics**, Informa UK Limited, v. 38, n. 5, p. 1063–1072, may 2011. Citations on pages 44 and 45.

PAN, J.; SHANG, J. Adaptive LASSO for linear mixed model selection via profile log-likelihood. **Communications in Statistics - Theory and Methods**, Informa UK Limited, v. 47, n. 8, p. 1882–1900, oct 2017. Citation on page 21.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, Oxford University Press (OUP), v. 58, n. 3, p. 545–554, 1971. Available: <https://www.jstor.org/stable/2334389>. Citation on page 31.

PINHO, L. G. B.; NOBRE, J. S.; SINGER, J. M. Cook's distance for generalized linear mixed models. **Computational Statistics & Data Analysis**, v. 82, p. 126–136, 2015. ISSN 0167-9473. Available: <https://www.sciencedirect.com/science/article/pii/S0167947314002400>. Citation on page 62.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Available: <https://www.R-project.org/>. Citation on page 57.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2010. ISBN 3-900051-07-0. Available: <http://www.R-project.org>. Citation on page 72.

RAJARATNAM, B.; ROBERTS, S.; SPARKS, D.; YU, H. Influence diagnostics for high-dimensional lasso regression. **Journal of Computational and Graphical Statistics**, Informa UK Limited, v. 28, n. 4, p. 877–890, jun 2019. Citations on pages 22, 44, and 71.

ROUSSEEUW, P. J.; CROUX, C. Alternatives to the median absolute deviation. **Journal of the American Statistical Association**, Informa UK Limited, v. 88, n. 424, p. 1273–1283, dec 1993. Citation on page 39.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. ISSN 00905364. Available: <http://www.jstor.org/stable/2958889>. Citation on page 37.

SEARLE, S. R. **Matrix algebra useful for statistics**. Hoboken, N.J: Wiley-Interscience, 2017. ISBN 978-1-118-93514-9. Citation on page 30.

SEBER, A. J. L. G. A. F. **Linear Regression Analysis**. [S.l.]: John Wiley & Sons, 2012. ISBN 9781118274422. Citations on pages 22 and 42.

SHAO, J. An asymptotic theory for linear model selection. **Statistica Sinica**, Institute of Statistical Science, Academia Sinica, v. 7, n. 2, p. 221–242, 1997. ISSN 10170405, 19968507. Available: <http://www.jstor.org/stable/24306073>. Citation on page 37.

SINGER, J.; NOBRE, J.; ROCHA, F. Análise de dados longitudinais: versão parcial preliminar. **São Paulo:[sn]**, 2018. Citations on pages 22, 29, 30, 31, 32, and 46.

SINGER, J. M.; ROCHA, F. M.; NOBRE, J. S. Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. **International Statistical Review**, Wiley, v. 85, n. 2, p. 290–324, aug 2017. Available: <https://doi.org/10.1111/insr.12178>. Citations on pages 22 and 46.

SINHA, S. K. Robust analysis of generalized linear mixed models. **Journal of the American Statistical Association**, Informa UK Limited, v. 99, n. 466, p. 451–460, jun 2004. Citations on pages 22 and 37.

TAN, F. E. S.; OUWENS, M. J. N.; BERGER, M. P. F. Detection of influential observations in longitudinal mixed effects regression models. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley, v. 50, n. 3, p. 271–284, sep 2001. Citation on page 62.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Citations on pages 21, 34, and 43.

ZOU, H. The adaptive lasso and its oracle properties. **Journal of the American statistical association**, Taylor & Francis, v. 101, n. 476, p. 1418–1429, 2006. Citations on pages 21 and 35.


ICMC
USP
SÃO CARLOS