High breakdown inference for mixed linear models

Copt, Samuel

HIGH BREAKDOWN INFERENCE FOR MIXED LINEAR MODELS

Samuel Copt

Submitted for the degree of Ph.D in Econometrics and Statistics

Faculty of Economics and Social Sciences

University of Geneva

Switzerland

Accepted on the recommendation of :

Prof. Gérard Antille, University of Geneva,

Prof. Chris Field, Dalhousie University, Canada,

Prof. Elvezio Ronchetti, University of Geneva,

Prof. Maria-Pia Victoria-Feser, University of Geneva, supervisor.

Thesis number 563

Geneva, 2004

La Faculté des sciences économiques et sociales,
sur préavis du jury, a autorisé l'impression de la présente thèse,
sans entendre, par là, émettre aucune opinion sur les propositions
qui s'y trouvent énoncées et qui n'engagent que la responsabilité
de leur auteur.

Genève, le 2 juillet 2004

Le Doyen :

Pierre ALLAN

Abstract

Mixed linear models are used to analyze data in many settings. These models have in most cases a multivariate normal formulation. The maximum likelihood estimator (MLE) or the residual MLE ($REML$) are usually chosen to estimate the parameters. However, the latter are based on the strong assumption of exact multivariate normality. Welsh and Richardson (1997) have shown that these estimators are not robust to small deviations from the multivariate normality. This means, in practice, that a small proportion of data (even only one) can drive the value of the estimates on their own.

We present some of the most used models in the analysis of variance. We introduce the mixed linear model formulation and see that in most cases it is possible to extract independent subvectors of observation. The structure of the covariance matrix is derived for a great variety of models. Since the model is multivariate, we propose in this thesis a high breakdown multivariate robust estimator for very general mixed linear models, that include, for example, covariates. This robust estimator belongs to the class of S-estimators (Rousseeuw and Yohai 1984) from which we can derive the asymptotic properties for inference. We also use it as a diagnostic tool to detect outlying subjects. We derive the estimating equation defining the high breakdown estimator and we describe how it can be computed via a simple iterative algorithm. We study the behavior of the robust estimator through an extensive simulation study. It is compared to the maximum likelihood estimator under a great variety of configuration implying different models, different contamination patterns and different samples size. We also discuss the advantages of this estimator and illustrate its performance with the analysis of four datasets.

We also consider robust inference for multivariate hypotheses as an alternative to the classical F-test by using a robust score type test statistic proposed by Heritier and Ronchetti (1994) and study its properties by means of simulations and real data analysis.

Résumé

Les modeles lin. mixtes sont utilisés pour analyser les données issues de recherches dans de nombreux domaines des sciences humaines et sociales. Pour estimer ces modèles et tester des hypothèses, nous proposons dans ce travail une approche dite robuste qui protège les analyses de bias potentiels dus à la présence dans les échantillons d'une minorité de données atypiques. Nous partons d'une formalisation multivariee de ces modeles et proposons un estimateur robuste appartenant à la classe des estimateurs S (Rousseeuw et Yohai 1984) dont nous dérivons les fonctions de scores ainsi que les propriétés asymptotiques pour l'inférence. Nous pouvons également l'utiliser comme outil de diagnostique afin de pouvoir détecter de potentielles valeurs extrêmes. Les avantages de cet estimateur ainsi que son comportement sont illustrés au travers d'une étude de simulation ainsi que par l'analyse de quatre jeux de données réels.

Basée sur cet estimateur, une partie inférence robuste est également développée. Nous proposons une alternative robuste au test classique de Fisher. Il s'agit d'un test du score robuste proposé par Hériter et Ronchetti (1994). Le comportement de ce test est étudié aux moyens de simulations ainsi que de jeux de donnée réels.

Aknowledgments

I would like to express my gratitude to Prof. Maria-Pia Victoria-Feser for her valuable suggestions and her generous guidance throughout the course of this research. Her confidence and her encouragements have made this work possible.

I am also grateful to Prof. Gérard Antille, Prof. Chris Field and Prof. Elvezio Ronchetti for their helpful comments and for taking the time and effort to ensure my work was of quality. Their valuable suggestions improved the clarity of presentation.

My thanks also go to my friends and colleagues of the Faculty of Economics and Social Sciences of the University of Geneva for their stimulating discussions and their moral support.
Finally, my thanks go to my wife Sonia, who deserves an award for her patience and understanding during my PhD study and the writing of this thesis.

To my parents,

Contents

1 Introduction **4**

2 Models formulation **9**

2.1 mixed linear models 9

2.2 One factor within-subject ANOVA model 12

2.2.1 Model I 12

2.2.2 Model II 14

2.3 Within-subject ANOVA models with more than one factor 15

2.4 Multilevel models 19

2.5 Hierarchical models 19

2.6 ANCOVA Models 23

3 Estimation of mixed linear models **24**

3.1 Classical estimators 24

3.1.1 Maximum likelihood estimator 25

3.1.2 REML in the general mixed linear models 27

3.2 Robust estimators 29

3.2.1 Estimation by maximizing a robustified likelihood 30

3.2.2 Estimating equations approach 32

3.2.3 B-optimal estimator 33

3.2.4 Robust versions of the restricted likelihood estimator 34

3.2.5 Rocke's method 35

3.3 Some remarks . . . 35

4 High breakdown multivariate estimator of location and constrained scale 36

4.1 Introduction . . . 36
4.1.1 Constrained S-estimator . . . 37
4.1.2 The maximum likelihood estimator for constrained covariance matrix . . . 44
4.1.3 Starting point for the constrained robust estimator . . . 45
4.1.4 An algorithm procedure for the constrained S-estimator . . . 50
4.1.5 Concluding remarks . . . 51

5 Simulation study for the $CTBS$ estimator 53

5.1 One factor ANOVA model with repeated measures . . . 54
5.2 Two factors within-subject ANOVA model . . . 57
5.3 Hierarchical models . . . 63
5.4 Breakdown analysis . . . 64
5.5 Small sample behavior study . . . 68

6 Robust Inference 76

6.1 Robust Inference for the univariate case . . . 77
6.2 Robust multivariate tests . . . 79

7 Simulation study for the robust score type test 83

7.1 One factor within-subject ANOVA model . . . 84
7.2 Two factors within-subject ANOVA model . . . 87
7.3 Hierarchical model . . . 93

8 Data analysis 96

8.1 Electrode resistance data . . . 96
8.2 Semantic priming data . . . 102
8.3 Metallic oxide data . . . 109
8.4 Orthodontic data . . . 111

9 Extensions **116**

9.1 Incomplete Design 118

10 Conclusion **122**

Chapter 1

Introduction

Statistics is concerned with the variability that is inevitably present in any sets of data. The traditional method for isolating the sources of variability in a set of measurements is known as the analysis of variance. It has long been used in statistics and provides a simple and optimal description of a complete data set by means of models characterized by a certain number of parameters. Its purpose is to determine the extent to which the effect of an independent variable, often called a factor, is a major component of the variability. The partition of the variability is summarized in what is known as an ANOVA table.

Historically, the original work of Fisher (1925) is at the beginning of the development of the analysis of variance. Fisher's description of the analysis of variance methodology was based on sums of squares of differences among observed means. In the recent decades, the trend has been to present many of the ideas behind the analysis of variance in terms of what are called nowadays linear models and more particularly that class of linear models which is called fixed effects models. Variations among data can be studied through different classes of linear models, those known as random effects models and also those called mixed effects models, which are models that have mixed features of fixed and random models. In general, data analysis using mixed linear models is closely related to the classical analysis of variance, but this is not always true for linear and random effects models. To have a thorough history of the development of the analysis of variance or to have a good documented bibliography on the subject, consider the excellent survey of Sahai, Khuri, and Kapadia (1985).

Considering a more standard terminology, an ANOVA model is also defined through an experimental design. An experimental design refers to a plan for assigning subjects to experimental conditions. Such designs involve a certain number of specifications such as the determination of the experimental conditions (independent variable) to be used, the measurements (dependent variables) to be recorded and the nuisance variable that must be controlled. Of course, those designs are also defined through the statistical hypothesis that are made about the parameters and the population from which the subjects are drawn. Given the choices made about the design and the statistical hypothesis, the model will be referred to as a fixed, random or mixed effects model. In this work, the terms model and design will be used indifferently. For a review of the various designs used in the statistical literature see e.g. Kirk (1982).

In this work we consider more particularly designs that involve repeated measurement of the same group of individuals. Such design is called a repeated measures design. More specifically, when separate groups of individuals are studied, the number of individuals required is some multiple of the combined levels of each factor that is introduced. With repeated measures, rather than randomly assigning the individuals to the various conditions, the same individuals can be measured at all levels. The advantage of such studies is that differences in measures cannot be attributed to individual characteristics (for example, motivation, intelligence). Another advantage is that we need fewer individuals, the same individual being measured at all levels. In either case, the design is termed as repeated measures, because the same individuals are measured on a number of occasions corresponding to each treatment level. It is also referred to as randomized block design, with each individual designated as a "block".

Repeated measures studies have been introduced in virtually all behavioral and social sciences: psychology, medicine, education, sociology, political science, economics, business and industry. There are many research hypotheses that can be tested using repeated measures designs, such as hypotheses that compare the same subjects under different treatments, or those that follow performance over time. Repeated measures designs are quite versatile and are called by many different names. For example, a one-way repeated measures ANOVA model may be known

as a one factor within-subjects ANOVA model, a treatments-by-subjects ANOVA model, or a randomized blocks ANOVA model. A two-way repeated measures ANOVA model may be referred to as a two factors within-subjects ANOVA model, a two-way ANOVA model with repeated measures on both factors, a multiple treatments-by-subjects ANOVA model. In fact, repeated measures designs are special cases of the randomized complete block design wherein each subject is considered to be a block and is observed under all treatment levels.

Actually, the first insight of this work was to develop a robust estimator for designs involving repeated measurements only. We will see that the results of this research are applicable not only to repeated measures models but also to other kind of models. As long as it is possible to write the corresponding models as a multivariate normal model with constrained covariance matrix, we can apply the robust method we propose. These models belong to the class of mixed linear models and include, for example, hierarchical or multilevel models (random nested models), longitudinal data (repeated measures) and others. The aim of this research is thus to investigate robust procedures for ANOVA models which can be described as multivariate normal with constrained covariance matrix and to propose new robust inference tools. In terms of robustness, most of the proposals in the field of ANOVA (or mixed linear models) are based on a weighted version of the corresponding log-likelihood function. Welsh and Richardson (1997) present a survey of the various robust methods developed so far.

Here we propose a different approach that we hope will improve the performance of the robust estimators especially in terms of breakdown point (see later). It starts from a reformulation of the models as multivariate normal distributions. For example, consider the one factor within-subject ANOVA model given by the following structural equation

$$y_{ij} = \mu + \lambda_j + s_i + \varepsilon_{ij} \qquad i = 1, .., n \quad j = 1, ..., l$$

where y_{ij} is an observation for subject i at treatment level j, μ is the grand mean, λ_j is a fixed effect for the jth treatment levels and $\Sigma_{j=1}^{l}\lambda_j = 0$, s_i is a random variable that explains the random effect of the ith subject on the response variable y_{ij} and ε_{ij} is a residual error

term. We suppose that the unobservable random variables s_i and ε_{ij} are independent and have independent $N(0, \sigma_s^2)$ and $N(0, \sigma_\varepsilon^2)$ respectively. An equivalent multivariate formulation of the model is given by

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{e}_l s_i + \boldsymbol{\varepsilon}_i \quad i = 1, ..., n$$

where $\mathbf{y}_i = [y_{i1,...,}y_{il}]^T$, $\boldsymbol{\mu} = [\mu + \lambda_1, \mu + \lambda_2 \ldots, \mu + \lambda_l]^T$,$\mathbf{e}_l$ is a vector of ones of length l and $\boldsymbol{\varepsilon}_i = [\varepsilon_{i1}, ...\varepsilon_{il}]^T$. Given the assumptions on s_i and ε_{ij}, the $\mathbf{y}_i$ are independent multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variables with

$$\boldsymbol{\Sigma} = \sigma_\varepsilon^2 \mathbf{I}_l + \sigma_s^2 \mathbf{J}_l$$

with $\mathbf{J}_l$ being the $l \times l$ matrix of ones. For other models, the covariance matrix $\boldsymbol{\Sigma}$ has always a particular structure (see Chapter 2). The problem is then reduced to the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and our aim is to propose a robust estimator for a structured or constrained $\boldsymbol{\Sigma}$. We can say that the covariance matrix $\boldsymbol{\Sigma}$ is constrained to have a particular structure given the assumptions on the structural model and we will use this structure in the definition of our robust estimator.

Estimating the parameters in mixed linear models is only a first step into inference. Therefore in a second part of our research, we will also consider robust testing for comparing treatment levels or in other words the elements of the mean vector. Fisher (1925) developed the widely used F-test. Unfortunately, the F-test is known as being not robust to small deviations such as extreme observations (see for example Braun and McNeil, 1981). It is true however that the robustness of the classical F-test is still a subject of research. While it can remain stable in presence of misspecified models, it can break completely in some other cases. As we will see later, there are two main reasons that affect the robust properties of the F-test. To our knowledge, no formal robustness study has been done in the repeated measures settings and we will show that small amounts of extreme observations can considerably bias the decisions taken on the basis of the F-test. With this idea in mind, we adapt a robust testing procedure based on the results of Heritier and Ronchetti (1994), who propose robust versions of the classical Wald test, score type test and likelihood ratio test for general parametric models. Since those results are asymptotically true and since the data sets encountered in practice are usually rather

small, we will also investigate the small sample properties of the robust tests.

This work is organized as follows. In Chapter 2 we present some of the most used models in the analysis of variance. We introduce the mixed linear models formulation and see that in most cases it is possible to extract independent subvectors of observation. The structure of the covariance matrix is derived for a great variety of models. In Chapter 3 we review both the classical and robust methods of estimation that exist so far. The estimating equation defining the high breakdown estimator are derived in Chapter 4. We describe how it can be computed via a simple iterative algorithm. In Chapter 5, we study the behavior of the robust estimator through an extensive simulation study. It is compared to the maximum likelihood estimator under a great variety of configuration implying different models, different contamination patterns and different samples size. Robust inference is developed in Chapter 6 and analyzed through a simulation experiment in Chapter 7. Chapter 8 analyzes four data sets using classical and robust estimators. In Chapter 9, we summarize the questions that have been left open during the course of this work and suggest different directions for further research. Finally, Chapter 10 concludes.

Chapter 2

Models formulation

In this chapter, we review some of the most used designs in analysis of variance through the mixed linear models formulation. Indeed, the mixed linear models formulation has the advantage of being very flexible for complex ANOVA designs. We present the formulation of mixed linear models as multivariate normal distribution with constrained covariance matrices. We explore different kinds of models and see that in each case it is always possible to extract random independent subvectors that are multivariate normal.

2.1 mixed linear models

mixed linear models can be expressed generally by the regression type equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

where $\mathbf{y}$ is the $N-$ vector of all measurements (observations), $\mathbf{X}$ is a $N \times q_0$ design matrix for the fixed effects component $\boldsymbol{\alpha}$, a q_0-vector of unknown fixed effects, $\mathbf{Z}$ is the $N \times q$ design matrix for the random effect vector $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ is the N-vector of independent residual errors. $\boldsymbol{\beta}$ can actually be partitioned into a series of r sub-vectors,

$$\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, ..., \boldsymbol{\beta}_r^T]^T \tag{2.2}$$

We have that

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\alpha}$$

$$var(\boldsymbol{\varepsilon}) = \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N \tag{2.3}$$

We suppose that for each $\boldsymbol{\beta}_j$

$$var(\boldsymbol{\beta}_j) = \sigma^2_j\mathbf{I}_{q_j} \tag{2.4}$$

with q_j being the number of elements in $\boldsymbol{\beta}_j$. Moreover

$$cov(\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j^T}) = 0 \quad j \neq j^T \tag{2.5}$$

and similarly

$$cov(\boldsymbol{\beta}, \boldsymbol{\varepsilon}) = 0 \tag{2.6}$$

Using (2.4)-(2.6), the covariance structure of $\boldsymbol{\beta}$ is

$$\mathbf{D} = var(\boldsymbol{\beta}) = \begin{bmatrix} \sigma^2_1\mathbf{I}_{q_1} & & & \\ & \sigma^2_2\mathbf{I}_{q_2} & & \\ & & \ddots & \\ & & & \sigma^2_r\mathbf{I}_{q_r} \end{bmatrix} \tag{2.7}$$

Then partitioning $\mathbf{Z}$ in submatrices as in (2.2) i.e.

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_r]$$

with each submatrix $\mathbf{Z}_j$ of dimension $N \times q_j$, (2.1) becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \sum_{j=1}^{r} \mathbf{Z}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon} \tag{2.8}$$

Hence,

$$\mathbf{V} = var(\mathbf{y}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N = \sum_{j=1}^{r} \sigma^2_j \mathbf{Z}_j \mathbf{Z}_j^T + \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N$$

A useful extension of this formulation is to treat $\boldsymbol{\varepsilon}$ just as another $\boldsymbol{\beta}_j$, say $\boldsymbol{\beta}_0$, and incorporate it into (2.8) by defining

$$\boldsymbol{\beta}_0 = \boldsymbol{\varepsilon}, \quad \mathbf{Z}_0 = \mathbf{I}_N, \quad \text{and} \quad \sigma^2_0 = \sigma^2_{\boldsymbol{\varepsilon}}$$

and so we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \sum_{j=0}^{r} \mathbf{Z}_j \boldsymbol{\beta}_j \tag{2.9}$$

and

$$\mathbf{V} = \sum_{j=0}^{r} \sigma^2_j \mathbf{Z}_j \mathbf{Z}_j^T \tag{2.10}$$

We assume that all the $q_0 + r + 1$ effects are identifiable and concentrate on models for which we can write

$$\mathbf{V} = \text{diag}(\boldsymbol{\Sigma}_i) \tag{2.11}$$

with $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}, \forall i = 1, \ldots, n$. For such models, we have an equivalent multivariate formulation for (2.1) which is

$$\mathbf{y}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.12}$$

with $\mathbf{y}_i$ the p-vector of independent observations obtained by partitioning $\mathbf{y}$ according to the covariance structure in (2.11) and $\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\alpha}$ with $\mathbf{x}$ a $p \times q_0$ matrix obtained by partitioning $\mathbf{X}$ according to the covariance structure in (2.11). The case in which $\boldsymbol{\mu} = \boldsymbol{\mu}_i = \mathbf{x}_i\boldsymbol{\alpha}$, i.e. with the presence in the model of covariates, will also be discussed. We can actually write

$$\boldsymbol{\Sigma} = \sum_{j=0}^{r} \sigma^2_j \mathbf{z}_j \mathbf{z}_j^T \tag{2.13}$$

with $\mathbf{z}_j$ being the design matrices that define the structure of each block of $\mathbf{V}$.

Most of the well-known models can actually be written as in (2.12) with a covariance matrix defined with (2.13) and we now present in detail some of them.

2.2 One factor within-subject ANOVA model

We begin with a simple repeated measures model with only one experimental factor that is usually termed as one factor within-subject ANOVA model. It is not of great interest in practice but has the advantage of being simple, easy to describe and theoretically interesting. There are actually two possible models and we will define them by means of structural equations. The first model, which is called Model I, expresses the response variable as a function of the fixed experimental factor and a random factor "subject". For the second model, Model II, the response variable is a function of the fixed experimental factor, a random factor "subject" and an interaction random factor which represents the interaction between the subject and the experimental factor. We treat here the case where the model is balanced i.e. all y_{ij} for all i and j are observed.

2.2.1 Model I

The one factor within-subject ANOVA model can be written as

$$y_{ij} = \mu_j + s_i + \varepsilon_{ij} \qquad i = 1, .., n \quad j = 1, ..., l \tag{2.14}$$

where y_{ij} is the response of subject i at treatment level j. There are l levels for the experimental factor (in this case $p = l$) and l measurements on n subjects are taken on the response variable. μ_j is a fixed effect for each level j of the experimental factor and s_i is a variable that represents the random effect of the ith subject on the response variable. The unobservable random variables s_i and ε_{ij} are supposed independent and have independent $N(0, \sigma_s^2)$ and $N(0, \sigma_\varepsilon^2)$ respectively $\forall ij$.

Usually, in the literature, an alternative formulation of model (2.14) is proposed. This model can be written as

$$y_{ij} = \mu + \lambda_j + s_i + \varepsilon_{ij} \tag{2.15}$$

where μ is the overall mean and λ_j is the fixed effect deviation from the overall mean μ for level (or treatment) j. We need $\Sigma_{j=1}^{l}\lambda_j = 0$. The mean μ_j in the structural model (2.14) is related to μ and λ_j with

$$\mu_j = \mu + \lambda_j$$

Here we will use formulation (2.15). Given the assumption on s_i and ε_{ij}, it follows that the y_{ij} are jointly normally distributed with mean $\mu + \lambda_j$ and

$$cov(y_{ij}, y_{lk}) = \begin{cases} \sigma_\varepsilon^2 + \sigma_s^2 & i = l, j = k \\ \sigma_s^2 & i = l, j \neq k \\ 0 & \text{otherwise} \end{cases}$$

The equivalent multivariate formulation is obtained by making n l-vectors of observations $\mathbf{y}_i$ which we can write as

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{e}_l s_i + \boldsymbol{\varepsilon}_i \quad i = 1, ..., n \tag{2.16}$$

with $\boldsymbol{\mu} = vec(\mu + \lambda_j)$. The $\mathbf{y}_i$ are then independent multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variables with

$$\begin{aligned} \boldsymbol{\Sigma} &= var(\mathbf{y}_i) = var(\mathbf{e}_l s_i + \boldsymbol{\varepsilon}_i) = \mathbf{e}_l \mathbf{e}_l^T \sigma_s^2 + \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}_l \\ &= \sigma_s^2 \mathbf{J}_l + \sigma_\varepsilon^2 \mathbf{I}_l = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_s^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \sigma_s^2 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_s^2 & \cdots & \cdots & \sigma_\varepsilon^2 + \sigma_s^2 \end{bmatrix} \end{aligned} \tag{2.17}$$

To match the notation of (2.13) we can write

$$\mathbf{\Sigma} = \sum_{j=0}^{r} \sigma_j^2 \mathbf{z}_j \mathbf{z}_j^T = \sigma_\varepsilon^2 \mathbf{I}_l + \sigma_s^2 \mathbf{J}_l$$

with $\mathbf{z}_1 = \mathbf{e}_l$ and $\mathbf{z}_0 = \mathbf{I}_l$. Note that $\boldsymbol{\mu}$ can also be expressed as $\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\alpha}$. Using the constraint $\Sigma_{j=1}^{l} \lambda_j = 0$, the fixed effect vector $\boldsymbol{\alpha}$ is defined as

$$\boldsymbol{\alpha} = [\mu, \lambda_1, \lambda_2, ..., \lambda_{l-1}]^T \tag{2.18}$$

and hence the design matrix $\mathbf{x}$ is

$$\mathbf{x} = \begin{bmatrix} \mathbf{e}_{l-1} & \mathbf{I}_{l-1} \\ 1 & -\mathbf{e}_{l-1}^T \end{bmatrix} \tag{2.19}$$

Finally, in the mixed linear models formulation, we have

$$\begin{aligned} \mathbf{y} &= (\mathbf{e}_n \otimes \mathbf{x})\boldsymbol{\alpha} + (\mathbf{I}_n \otimes \mathbf{e}_l)\boldsymbol{\beta} + \varepsilon \\ &= \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} \end{aligned} \tag{2.20}$$

where $\mathbf{y}$ is a nl-vectors of responses, $\boldsymbol{\varepsilon}$ is a nl-vectors of residual errors, and $\otimes$ is the Kronecker product. We have $\boldsymbol{\beta}_1 = (s_1, \ldots, s_n)^T$, $\mathbf{Z}_1 = \mathbf{I}_n \otimes \mathbf{e}_l$, so that $\mathbf{Z}_1\mathbf{Z}_1^T = (\mathbf{I}_n \otimes \mathbf{e}_l)(\mathbf{I}_n \otimes \mathbf{e}_l)^T$. It follows that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\alpha}$, and that

$$\begin{aligned} \mathbf{V} &= var(\mathbf{y}) = (\mathbf{I}_n \otimes \mathbf{1}_l)\sigma_s^2 \mathbf{I}_l (\mathbf{I}_n \otimes \mathbf{1}_l)^T + \sigma_\varepsilon^2 \mathbf{I}_{nl} \\ &= \sigma_s^2(\mathbf{I}_n \otimes \mathbf{J}_l) + \sigma_\varepsilon^2(\mathbf{I}_n \otimes \mathbf{I}_l) = \mathbf{I}_n \otimes (\sigma_s^2 \mathbf{J}_l + \sigma_\varepsilon^2 \mathbf{I}_l) \end{aligned}$$

2.2.2 Model II

Suppose now that we have to following structural model

$$y_{ij} = \mu + \lambda_j + s_i + (\lambda s)_{ij} + \varepsilon_{ij} \qquad i = 1, .., n \quad j = 1, ..., l \tag{2.21}$$

where y_{ij}, μ, λ_j and s_i are defined as before. We suppose here that there is an interaction between the random effect of the model (i.e. the subject) and the fixed experimental factor. A new random variable $(\lambda s)_{ij}$ is thus included in the model. The unobservable random variables s_i, ε_{ij}, $(\lambda s)_{ij}$ are supposed independent and have independent $N(0, \sigma_s^2)$, $N(0, \sigma_\varepsilon^2)$ and $N(0, \sigma_{\lambda s}^2)$ respectively.

As before we can use a multivariate formulation to get

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{e}_l s_i + (\boldsymbol{\lambda}\mathbf{s})_i + \boldsymbol{\varepsilon}_i \quad i = 1, ..., n \tag{2.22}$$

and $(\boldsymbol{\lambda}\mathbf{s})_i = [(\lambda s)_{ij}, ..., (\lambda s)_{il}]^T$. The $\mathbf{y}_i$ are then independent multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variables with

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_s^2 + \sigma_{\lambda s}^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \sigma_s^2 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_s^2 & \cdots & \cdots & \sigma_\varepsilon^2 + \sigma_s^2 + \sigma_{\lambda s}^2 \end{bmatrix} \tag{2.23}$$

Note that it is clear from (2.23) that $\sigma_{\lambda s}^2$ cannot be separated form $(\sigma_s^2 + \sigma_{\lambda s}^2)$, so that model II cannot be estimated. In other words, the interaction effect is confounded with the residual errors. This is not necessary the case with models that include more fixed effects.

2.3 Within-subject ANOVA models with more than one factor

As we want to use the particular covariance structure of the different models (either model I or model II) we now enlarge the analysis to more complex models. Suppose we have a structural equation following the model II but this time with two fixed effects factors λ and γ with respectively l and g levels, that is

$$y_{ijk} = \mu + \lambda_j + \gamma_k + (\lambda\gamma)_{jk} + s_i + (\lambda s)_{ij} + (\gamma s)_{ik} + \varepsilon_{ijk} \tag{2.24}$$

with $i = 1, .., n, \quad j = 1, ..., l$ and $k = 1, ..., g$ (hence in this case $p = gl$). For the parameters to be identifiable, we need

$$\Sigma_{j=1}^{l}\lambda_j = 0,\ \Sigma_{k=1}^{g}\gamma_k = 0 \ \text{ and } \ \Sigma_{j=1}^{l}\Sigma_{k=1}^{g}(\lambda\gamma)_{jk} = 0.$$

Note that an alternative model is given by (2.24) in which the random interaction $(\lambda s)_{ij}$ and $(\gamma s)_{ik}$ are omitted. The unobservable random variables s_i, ε_{ij}, $(\lambda s)_{ij}$ and $(\gamma s)_{ik}$ are independent and have independent $N(0, \sigma_s^2)$, $N(0, \sigma_\varepsilon^2)$, $N(0, \sigma_{\lambda s}^2)$ and $N(0, \sigma_{\gamma s}^2)$ respectively. Using the formulation given in (2.12) we have that the $\mathbf{y}_i$ are multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = vec(\mu + \lambda_j + \gamma_k + (\lambda\gamma)_{jk})\ \forall i$$

$\boldsymbol{\mu}$ can also be defined as $\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\alpha}$. An example of the structure of $\mathbf{x}$ and $\boldsymbol{\alpha}$ is given below.
For the covariance matrix, define

$$\begin{aligned}
\boldsymbol{\beta}_1 &= (s_1, \ldots, s_n)^T & \mathbf{Z}_1 &= \mathbf{I}_n \otimes \mathbf{e}_{gl} \\
\boldsymbol{\beta}_2 &= ((\lambda s)_{11}, (\lambda s)_{12}, \ldots, (\lambda s)_{nl})^T & \mathbf{Z}_2 &= \mathbf{I}_n \otimes \mathbf{I}_l \otimes \mathbf{e}_g \\
\boldsymbol{\beta}_3 &= ((\gamma s)_{11}, (\gamma s)_{12}, \ldots, (\gamma s)_{ng})^T & \mathbf{Z}_3 &= \mathbf{I}_n \otimes \mathbf{e}_l \otimes \mathbf{I}_g
\end{aligned}$$

so that $\mathbf{Z}_2\mathbf{Z}_2^T = \mathbf{I}_n \otimes \mathbf{I}_l \otimes \mathbf{J}_g$ and $\mathbf{Z}_3\mathbf{Z}_3^T = \mathbf{I}_n \otimes \mathbf{J}_l \otimes \mathbf{I}_g$, and therefore

$$\boldsymbol{\Sigma} = \sigma_s^2\mathbf{J}_{gl} + \sigma_{\lambda s}^2(\mathbf{I}_l \otimes \mathbf{J}_g) + \sigma_{\gamma s}^2(\mathbf{J}_l \otimes \mathbf{I}_g) + \sigma_\varepsilon^2\mathbf{I}_{gl}$$

with $\mathbf{z}_1 = \mathbf{e}_{gl}$, $\mathbf{z}_2 = \mathbf{I}_l \otimes \mathbf{e}_g$ and $\mathbf{z}_3 = \mathbf{e}_l \otimes \mathbf{I}_g$.
To get an idea of the structure of the covariance matrix $\boldsymbol{\Sigma}$ and $\mathbf{x}\boldsymbol{\alpha}$, suppose now that $l = 2$ and $g = 2$. Then $\boldsymbol{\Sigma}$ would have the form

$$\boldsymbol{\Sigma} = \begin{bmatrix} [\Sigma_1] & [\Sigma_2] \\ [\Sigma_2] & [\Sigma_1] \end{bmatrix}$$

with

$$\Sigma_1 = \begin{bmatrix} \sigma_1^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \sigma_3^2 & \sigma_4^2 \\ \sigma_4^2 & \sigma_3^2 \end{bmatrix}$$

and

$$\begin{aligned} \sigma_1^2 &= \sigma_\varepsilon^2 + \sigma_s^2 + \sigma_{\lambda s}^2 + \sigma_{\gamma s}^2 \\ \sigma_2^2 &= \sigma_s^2 + \sigma_{\gamma s}^2 \\ \sigma_3^2 &= \sigma_s^2 + \sigma_{\lambda s}^2 \\ \sigma_4^2 &= \sigma_s^2 \end{aligned}$$

For the fixed effect vector $\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\alpha}$, one would get

$$\boldsymbol{\alpha} = \left[\mu, \lambda_1, \gamma_1, (\lambda\gamma)_{11}\right]^T$$

and

$$\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

Note that in this case the interaction variance $\sigma_{\lambda s}^2$ and $\sigma_{\gamma s}^2$ can be separated from the residual variance.

The general structure of the covariance matrix remains actually the same when the number of fixed effects increases. Take a model with three fixed effect factors λ , γ and ζ. One would get for example the (rather undigest) structural equation,

$$y_{ijkm} = \mu + \lambda_j + \gamma_k + \zeta_m + (\lambda\gamma)_{jk} + (\lambda\zeta)_{jm} + (\gamma\zeta)_{km} + (\lambda\gamma\zeta)_{jkm} + s_i + (\lambda s)_{ij} + (\gamma s)_{ik} + (\zeta s)_{im} + \varepsilon_{ijkm}$$

with $i = 1, .., n \quad j = 1, ..., l$ and $k = 1, ..., g$, $m = 1, ..., h$ and we deduce that

$$\boldsymbol{\mu} = vec(\mu + \lambda_j + \gamma_k + \zeta_m + (\lambda\gamma)_{jk} + (\lambda\zeta)_{jm} + (\gamma\zeta)_{km} + (\lambda\gamma\zeta)_{jkm}) \ \forall i$$

$\boldsymbol{\mu}$ can also be written as $\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\alpha}$ with the structure of $\mathbf{x}$ and $\boldsymbol{\alpha}$ depending of l, g and m. The covariance matrix has the form

$$\boldsymbol{\Sigma} = \sigma_s^2\mathbf{J}_{glq} + \sigma_{\lambda s}^2 (\mathbf{I}_l \otimes \mathbf{J}_{gq}) + \sigma_{\gamma s}^2(\mathbf{J}_l \otimes \mathbf{I}_g \otimes \mathbf{J}_q) + \sigma_{\zeta s}^2(\mathbf{J}_{gl} \otimes \mathbf{I}_q) + \sigma_\varepsilon^2\mathbf{I}_{glq}$$

Suppose that each factor has two levels, one gets the following covariance matrix $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} [\Sigma_1] & [\ \Sigma_2] & [\Sigma_3] & [\Sigma_4] \\ [\Sigma_2] & [\Sigma_1] & [\Sigma_4] & [\Sigma_3] \\ [\Sigma_3] & [\Sigma_4] & [\Sigma_1] & [\Sigma_2] \\ [\Sigma_4] & [\Sigma_3] & [\Sigma_2] & [\Sigma_1] \end{bmatrix}$$

with

$$\Sigma_1 = \begin{bmatrix} \sigma_1^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \sigma_3^2 & \sigma_4^2 \\ \sigma_4^2 & \sigma_3^2 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} \sigma_5^2 & \sigma_6^2 \\ \sigma_6^2 & \sigma_5^2 \end{bmatrix}, \Sigma_4 = \begin{bmatrix} \sigma_7^2 & \sigma_8^2 \\ \sigma_8^2 & \sigma_7^2 \end{bmatrix}$$

and

$$\begin{aligned} \sigma_1^2 &= \sigma_\varepsilon^2 + \sigma_s^2 + \sigma_{\lambda s}^2 + \sigma_{\gamma s}^2 + \sigma_{\zeta s}^2 & \sigma_5^2 &= \sigma_s^2 + \sigma_{\gamma s}^2 + \sigma_{\zeta s}^2 \\ \sigma_2^2 &= \sigma_s^2 + \sigma_{\lambda s}^2 + \sigma_{\gamma s}^2 & \sigma_6^2 &= \sigma_s^2 + \sigma_{\gamma s}^2 \\ \sigma_3^2 &= \sigma_s^2 + \sigma_{\lambda s}^2 + \sigma_{\zeta s}^2 & \sigma_7^2 &= \sigma_s^2 + \sigma_{\zeta s}^2 \\ \sigma_4^2 &= \sigma_s^2 + \sigma_{\lambda s}^2 & \sigma_8^2 &= \sigma_s^2 \end{aligned}$$

We can see that for classical repeated measures models (having only crossed within-subject factors), the block structure of the covariance matrix $\boldsymbol{\Sigma}$ remains always the same.

2.4 Multilevel models

In a multilevel model or random nested model, the random factors are only nested in random ones, leading to models of the type

$$y_{ijk} = \mu + \lambda_j + s_i + \gamma_{i(k)} + \varepsilon_{i(k)j}$$

with $\mu+\lambda_j, j = 1, \ldots, l$ the fixed effect, and $s_i, i = 1, \ldots, n$ and $\gamma_{i(k)}, k = 1, \ldots, g \cdot n$ the random effects. The distributional hypothesis on s_i, $\gamma_{i(k)}$ and $\varepsilon_{i(k)j}$ are $N(0, \sigma_s^2)$, $N(0, \sigma_\gamma^2)$ and $N(0, \sigma_\varepsilon^2)$ respectively. A simple example is the case in which g measures are taken on each subject i and each experimental condition j. With this model, we have that $\boldsymbol{\mu} = \mathbf{e}_g \otimes \text{vec}(\mu + \lambda_j),\ \ j = 1, ..., l$, $\mathbf{Z}_1 = \mathbf{I}_n \otimes \mathbf{e}_{gl}$, $\mathbf{Z}_2 = \mathbf{I}_n \otimes \mathbf{e}_l \otimes \mathbf{I}_g$, so that

$$\boldsymbol{\Sigma} = \sigma_s^2 \mathbf{J}_{gl} + \sigma_\gamma^2 \left(\mathbf{J}_l \otimes \mathbf{I}_g\right) + \sigma_\varepsilon^2 \mathbf{I}_{gl}.$$

For example, suppose that $j = 2$ and $g = 2$ then $\boldsymbol{\Sigma}$ would have the form

$$\boldsymbol{\Sigma} = \begin{bmatrix} [\Sigma_1] & [\Sigma_2] \\ [\Sigma_2] & [\Sigma_1] \end{bmatrix}$$

with

$$\Sigma_1 = \begin{bmatrix} \sigma_s^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2 & \sigma_s^2 + \sigma_\gamma^2 \\ \sigma_s^2 + \sigma_\gamma^2 & \sigma_s^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 \end{bmatrix}$$

2.5 Hierarchical models

Until now, we have only presented models in which each level of a factor is combined with every level of another factor. Hierarchical models are models where only some levels of a factor are combined with the levels of another factor. More formally, suppose that we have two treatments λ and γ with respectively l and g levels. In the language of experimental design, if each level of treatment γ appears with only *one* level of treatment λ, γ is said to be nested in λ. Experimental designs with one or more nested treatments are particularly well suited for research in the behavioral and medical sciences, education and industry. Consider an example in education

in which two types of instruction materials (treatment levels λ_1 and λ_2) are to be evaluated using students in four classes (treatment levels γ_1,γ_2,γ_3 and γ_4). Two classrooms are randomly assigned to one type of programmed material. For obvious reasons, all children in a particular classroom are subject to the same type of material. We assume that each classroom contains the same number of children. Each classroom γ_k appears with only one level of instruction material. Thus treatment γ is nested in λ.

One can also extend the models so as to include covariates. For example, we have the typical experiment in which a measure is taken from n_1 samples of type $j = 1$ and n_2 samples of type $j = 2$, and in each sample the measure is taken on g "objects". For example, the "objects" can be rats, the samples cages, n_1 of which are given treatment $j = 1$ and treatment $j = 2$ to the n_2 others. This type of design is called a nested design. The covariate is here a dummy variable for the type of treatment. The corresponding model can be written as

$$y_{ijk} = \mu + \lambda J_i(j) + \gamma_{j(i)} + \varepsilon_{j(i(k))}$$

with

$$J_i(j) = \begin{cases} 0 & j = 1 \\ 1 & j = 2 \end{cases}$$

$\mu + \lambda J_i(j)$ the fixed effect and $\gamma_{j(i)}, i = 1, \ldots n$ $(n = n_1 + n_2)$, with $k = 1, \ldots, g \cdot n$. We then have

$$\boldsymbol{\mu}_i = \mathbf{e}_g \left(\mu + \lambda J_i(j)\right) = \mathbf{e}_g \otimes \left(1, J_i(j)\right) \left(\mu, \lambda\right)^T = \mathbf{x}_i \boldsymbol{\alpha}$$

$\boldsymbol{\beta}_1 = (\gamma_{1(1)}, \ldots, \gamma_{1(n_1)}, \gamma_{2(n_1+1)}, \ldots, \gamma_{2(n)})^T$, $\mathbf{Z}_1 = \mathbf{I}_n \otimes \mathbf{e}_g$ and therefore

$$\boldsymbol{\Sigma} = \sigma^2_\gamma \mathbf{J}_g + \sigma^2_\varepsilon \mathbf{I}_g$$

Note that the structure of $\boldsymbol{\Sigma}$ is the same as with the one factor within-subject ANOVA model, the difference lies in the mean that depends on the "sample" which plays here the role of observation.

A more complicated example is a design reported by Fellner (1986) and also analyzed by Richardson and Welsh (1995) on the content of one type ($j = 1$) of metallic oxide measured in $n_1 = 18$ lots and another type of metallic oxide ($j = 2$) measured in $n_2 = 13$ other lots. Two samples were drawn from each lot and duplicate analyses were then performed by each of two chemists randomly selected for each sample. The model is

$$y_{ijklm} = \mu + \lambda J_i(j) + \gamma_{j(i)} + \delta_{j(i(k))} + \xi_{j(i(k(l)))} + \varepsilon_{j(i(k(l(m))))} \tag{2.25}$$

with $\mu + \lambda J_i(j)$ the fixed effect and $\gamma_{j(i)}, i = 1, \ldots n$ $(n = n_1 + n_2)$ the random effect due to the lot, $\delta_{j(i(k))}, k = 1, \ldots, 2n$, the random effect due to the sample and $\xi_{j(i(k(l)))}, l = 1, \ldots, 4n$, the random effect due to the chemist. We then have

$$\boldsymbol{\mu}_i = \mathbf{e}_8 \left(\mu + \lambda J_i(j)\right) = \mathbf{e}_8 \otimes \left(1, J_i(j)\right) \left(\mu, \lambda\right)^T = \mathbf{x}_i \boldsymbol{\alpha}$$

and $\mathbf{Z}_1 = \mathbf{I}_n \otimes \mathbf{e}_8$, $\mathbf{Z}_2 = \mathbf{I}_n \otimes \mathbf{I}_2 \otimes \mathbf{e}_4$, $\mathbf{Z}_3 = \mathbf{I}_n \otimes \mathbf{I}_4 \otimes \mathbf{e}_2$, so that

$$\boldsymbol{\Sigma} = \sigma^2_\gamma \mathbf{J}_8 + \sigma^2_\lambda \mathbf{I}_2 \otimes \mathbf{J}_4 + \sigma^2_\delta \mathbf{I}_4 \otimes \mathbf{J}_2 + \sigma^2_\varepsilon \mathbf{I}_8$$

Thus the parameters to be estimated are the means for each type of metallic oxide and the variances respectively associated with lots, samples and chemists. Suppose that we have 248 observations. We then can make n =31 independent sub-vectors y_i of size 8. To give a better idea of the structure of each independent subvector $y_i = [y_{i111}, ..., y_{i222}]^T$, the sub-vector y_1 is defined as

$$\begin{bmatrix} y_{1111} \\ y_{1112} \\ y_{1121} \\ y_{1122} \\ y_{1211} \\ y_{1212} \\ y_{1221} \\ y_{1222} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \lambda \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \gamma_1 + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_{11} \\ \delta_{12} \end{bmatrix} +$$
$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \xi_{111} \\ \xi_{112} \\ \xi_{121} \\ \xi_{122} \end{bmatrix} + I_8 \begin{bmatrix} \varepsilon_{1111} \\ \varepsilon_{1112} \\ \varepsilon_{1121} \\ \varepsilon_{1122} \\ \varepsilon_{1211} \\ \varepsilon_{1212} \\ \varepsilon_{1221} \\ \varepsilon_{1222} \end{bmatrix}$$

The covariance matrix $\mathbf{\Sigma}$ is given by

$$var(\mathbf{y}_i) = \sum_{r=1}^{3} \sigma_r^2 \mathbf{z}_r \mathbf{z}_r^T + \sigma_\varepsilon^2 \mathbf{I}_8 = \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{bmatrix}$$

with

$$\Sigma_1 = \begin{bmatrix} \sigma_1^2 & \sigma_2^2 & \sigma_3^2 & \sigma_3^2 \\ \sigma_2^2 & \sigma_1^2 & \sigma_3^2 & \sigma_3^2 \\ \sigma_3^2 & \sigma_3^2 & \sigma_1^2 & \sigma_2^2 \\ \sigma_3^2 & \sigma_3^2 & \sigma_2^2 & \sigma_1^2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} \sigma_4^2 & \sigma_4^2 & \sigma_4^2 & \sigma_4^2 \\ \sigma_4^2 & \sigma_4^2 & \sigma_4^2 & \sigma_4^2 \\ \sigma_4^2 & \sigma_4^2 & \sigma_4^2 & \sigma_4^2 \\ \sigma_4^2 & \sigma_4^2 & \sigma_4^2 & \sigma_4^2 \end{bmatrix}$$

$$
\begin{aligned}
\sigma_1^2 &= \sigma_\gamma^2 + \sigma_\lambda^2 + \sigma_\delta^2 + \sigma_e^2 \\
\sigma_2^2 &= \sigma_\gamma^2 + \sigma_\lambda^2 + \sigma_\delta^2 \\
\sigma_3^2 &= \sigma_\gamma^2 + \sigma_\lambda^2 \\
\sigma_4^2 &= \sigma_\gamma^2
\end{aligned}
$$

2.6 ANCOVA Models

In psychology, ANCOVA models denote ANOVA models with covariates (like a pre-measurement). As an example consider the model in which a covariate is added to a within-subject ANOVA model as in (2.15) for each level of the within-subject factor, i.e.

$$y_{ij} = \mu + \lambda_j x_{ij} + \gamma_j + s_i + \varepsilon_{ij} \tag{2.26}$$

with $\mu + \lambda_j x_{ij} + \gamma_j$ the fixed effect and s_i the random effect. The parameter γ_j is added when it is supposed that the intercepts of the regression lines are different. Similarly, one could add the constraint of equal regression slopes with $\lambda_j = \lambda \; \forall j$. We then have

$$
\begin{aligned}
\boldsymbol{\mu}_i &= vec\left(\mu + \lambda_j \mathbf{x}_{ij} + \gamma_j\right) = \left[\mathbf{e}_l, \operatorname{diag}\left(x_{ij}\right), \mathbf{I}_l\right] \left(\mu, \lambda_1, \ldots, \lambda_l, \gamma_1, \ldots, \gamma_l\right)^T \\
&= \mathbf{x}_i \boldsymbol{\alpha},
\end{aligned}
$$

and $\mathbf{Z}_1 = \mathbf{I}_n \otimes \mathbf{e}_l$, so that

$$\boldsymbol{\Sigma} = \sigma_s^2 \mathbf{J}_l + \sigma_\varepsilon^2 \mathbf{I}_l.$$

Here, the primary goal of this research is to define a robust estimator for mixed linear models. In this chapter we have seen that it is possible to extract independent multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ vectors of observations $\mathbf{y}_i$ for some of the well-known mixed linear models, so that a robust estimator based on the multivariate normal model $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with constrained covariance matrix can be used for the estimation of the parameters.

Chapter 3

Estimation of mixed linear models

3.1 Classical estimators

The purpose of analysis of variance is to test differences in means for statistical significance. This is accomplished by analyzing the variance, that is, by partitioning the total variance into the component that is due to the random error and the components that are due to differences between means. Another classical approach to the estimation of the variance components is the maximum likelihood approach.

Suppose that we observe a response variable y_{ij} described through the structural equation (2.15). The analysis of sources of variability between observations and the grand mean proceeds with estimation of each component of the structural equation. Hence, the grand mean is estimated by $\bar{y}_{..} = (ln)^{-1}\Sigma_{i=1}^{n}\Sigma_{j=1}^{l}y_{ij}$. The effect of each level of the treatment factor λ_j is estimated by the difference between each treatment's level mean and the grand mean $\bar{y}_{.j} - \bar{y}_{..}$ with $\bar{y}_{.j} = n^{-1}\Sigma_{i=1}^{n}y_{ij}$ .The effect of each subject s_i is estimated by the difference between each subject's mean and the grand mean $\bar{y}_{i.} - \bar{y}_{..}$ with $\bar{y}_{i.} = l^{-1}\Sigma_{j=1}^{l}y_{ij}$. This component, *between-subjects*, reflects differences among the subjects. Random errors, ε_{ij}, would be estimated by $y_{ij} - \bar{y}_{i.}$, the difference between a score and the subject's mean. This component, *within-subjects*, reflects variability within each subject. The simple estimators of σ_{ε}^{2} and σ_{s}^{2} are obtained by

$$\hat{\sigma}_s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 - \frac{1}{l}\hat{\sigma}_\varepsilon^2, \tag{3.1}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n(l-1)}\sum_{i=1}^{n}\sum_{j=1}^{l}(y_{ij} - \bar{y}_{i\cdot})^2 \tag{3.2}$$

These estimators are unbiased estimators of the variance components. For other types of structural equation it is always possible to proceed with the classical computation of the various sums of squares that can become quite tedious when the complexity of the design increases. We will not pursue this issue here, and consider the maximum likelihood to derive estimates for the various parameters.

3.1.1 Maximum likelihood estimator

As we have said, when the models become more complicated, it is more convenient to use the multivariate formulation used in mixed linear models. Recall that we can write a mixed linear models as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} \tag{3.3}$$

with $\mathbf{Z}\boldsymbol{\beta}$ partitioned as

$$\mathbf{Z}\boldsymbol{\beta} = [\mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_r]\begin{bmatrix}\boldsymbol{\beta}_0\\ \boldsymbol{\beta}_1\\ \vdots\\ \boldsymbol{\beta}_r\end{bmatrix} = \sum_{j=0}^{r}\mathbf{Z}_j\boldsymbol{\beta}_j \tag{3.4}$$

where $\boldsymbol{\beta}_j$ is the vector for random factor j, and $\boldsymbol{\beta}_0 = \boldsymbol{\varepsilon}$ and $\mathbf{Z}_0 = \mathbf{I}_N$. The number of level for random factor j is denoted by q_j. Recall that the random effects $\boldsymbol{\beta}_j$ have the properties

$$E(\boldsymbol{\beta}_j) = 0 \quad \text{and} \quad var(\boldsymbol{\beta}_j) = \sigma_i^2\mathbf{I}_{q_i}$$

with $\sigma_0^2 = \sigma_\varepsilon^2$ and

$$cov(\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j^T}) = 0 \quad \text{for } j \neq j^T$$

Thus

$$var(\boldsymbol{\beta}) = \mathbf{D}$$

with $\mathbf{D}$ defined in (2.7).

Using these assumptions we know that

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\alpha}$$

and

$$\mathbf{V} = var(\mathbf{y}) = \sum_{j=0}^{r} \sigma_j^2 \mathbf{Z}_j \mathbf{Z}_j^T \tag{3.5}$$

Given that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\alpha}, \mathbf{V})$, the likelihood function is

$$L = L(\boldsymbol{\alpha}, \mathbf{V}\,|\mathbf{y}\,) = \frac{\exp^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})^T \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})}}{(2\pi)^{\frac{1}{2}N}\,|\mathbf{V}|^{\frac{1}{2}}} \tag{3.6}$$

where $|\mathbf{V}|$ stands for the determinant of $\mathbf{V}$. The log-likelihood

$$\log L = l(\boldsymbol{\alpha}, \mathbf{V}\,|\mathbf{y}\,) = -\frac{1}{2} N \log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}) \tag{3.7}$$

Recall that the unknown parameters are $\boldsymbol{\alpha}$ and $\sigma_\varepsilon^2, \sigma_1^2, ..., \sigma_r^2$ included in $\mathbf{V}$. To maximize $l(\boldsymbol{\alpha}, \mathbf{V}\,|\mathbf{y}\,)$, we differentiate (3.7), first with respect to $\boldsymbol{\alpha}$ which yields (for rules of matrix differentiation, see Graybill, 1983)

$$\frac{\delta l(\boldsymbol{\alpha}, \mathbf{V}\,|\mathbf{y}\,)}{\delta \boldsymbol{\alpha}} = \mathbf{X}^T \mathbf{V}^{-1}\mathbf{y} - \mathbf{X}^T \mathbf{V}^{-1}\mathbf{X}\boldsymbol{\alpha} \tag{3.8}$$

Second, differentiating (3.7) with respect to σ_j^2 with

$$\frac{\delta \mathbf{V}}{\delta \sigma_j^2} = \mathbf{Z}_j \mathbf{Z}_j^T$$

gives for $j = 0, ..., r$

$$\frac{\delta l(\boldsymbol{\alpha}, \mathbf{V} \,|\mathbf{y}\,)}{\delta \sigma_j^2} = -\frac{1}{2} tr(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j^T) + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})^T \mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}) \tag{3.9}$$

Equation (3.9) is a complicated function of the variance components. Hence expressions for the solutions have to be obtained numerically, usually using either an iteratively reweighted least squares or a Newton-Raphson procedure. In the mixed linear theory this problem has been extensively studied and numerous algorithms have been developed (see e.g. Jennrich and Schluchter, 1986 or Lindstrom and Bates, 1988).

3.1.2 REML in the general mixed linear models

Patterson and Thompson (1971) introduced restricted maximum likelihood estimation (REML) as a method of estimating variance components in the context of unbalanced incomplete block designs. REML is often preferred to maximum likelihood estimation because it takes into account the loss of degrees of freedom in estimating the mean and hence produces unbiased estimating equations for the variance parameters. Alternative and more general derivations of REML are given by Harville (1977).

As for the maximum likelihood estimator we first define the REML in the case of the general mixed linear models. Rather than using the data vector $\mathbf{y}$ directly, REML is based on linear combinations of elements of $\mathbf{y}$, chosen in such a way that the resulting model does not contain any fixed elements. This arises from starting with a set of values $k^T\mathbf{y}$ where the vectors k^T of size $1 \times N$ are chosen so that $k^T\mathbf{y} = k^T\mathbf{X}\boldsymbol{\alpha} + k^T\mathbf{Z}\mathbf{u}$ contains no term in $\boldsymbol{\alpha}$, i.e. so that

$$k^T\mathbf{X}\boldsymbol{\alpha} = 0 \tag{3.10}$$

Hence

The differentiation with respect to σ_j^2 gives for $j = 0, ..., r$

$$\frac{\delta l_R(\boldsymbol{\alpha}, \mathbf{V} \,|\mathbf{y}\,)}{\delta \sigma_j^2} = -\frac{1}{2} tr(\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j^T) + \frac{1}{2}\mathbf{y}^T\, \mathbf{P}\mathbf{Z}_j\mathbf{Z}_j^T\mathbf{V}^{-1}\mathbf{P}\mathbf{y} \tag{3.13}$$

with (see Kathri, 1966),

$$\mathbf{P} = \mathbf{K}(\mathbf{K}^T\mathbf{V}\mathbf{K})^{-1}\mathbf{K}^T$$

The transformation of $\mathbf{y}$ into $\mathbf{K}^T\mathbf{y}$ is not suitable for a multivariate normal formulation. Indeed, for example, in the one factor within-subject ANOVA model, $\mathbf{X} = \mathbf{e}_n \otimes \mathbf{I}_l$ ($\boldsymbol{\alpha} = \boldsymbol{\mu}$) then $\mathbf{M} = \mathbf{I}_N - (\mathbf{e}_n \otimes \mathbf{I}_l)\,(\mathbf{e}_n \otimes \mathbf{I}_l)^+$ and $(\cdot)^+$ is the Moore-Penrose inverse. It is clear that it is not possible to recover a structure like in (2.12) from the REML formulation, so that we won't pursue this route here. In terms of robustness, Richardson and Welsh (1994) have shown that the REML is not robust and hence have proposed a robust version (see later).

3.2 Robust estimators

Robust statistics is an extension of parametric statistics, taking into account that parametric models are only approximation of the reality. It is concerned with the behavior of statistical procedures (tests, estimators,...) under small model deviations. Moreover, robust statistics must provide statistical procedures which are reliable and reasonably efficient under deviations from the assumed parametric models. We define a few basic concepts developed in robust statistics that will be used in the present work. We first define a local concept, namely the influence function. It measures the asymptotic bias caused to an estimator by an infinitesimal amount of contamination at some particular point. This local concept is complemented by a global notion, the breakdown point which measures the maximal proportion of contamination that an estimator can tolerate without taking arbitrarily large values. For example, Maronna (1976) showed that robust estimator based on a weighting scheme that is not redescending (no weight of zero) fails to be robust in high dimension. This happens because for such estimators, their breakdown point is at most $1/(p+1)$, p being the dimension of the data. When working in high dimension it is therefore crucial to consider high

breakdown estimators. An estimator can be robust in the infinitesimal sense i.e. to have a bounded influence function but not in the global one i.e. to have a low breakdown point. For more details and formulae on robustness measures, see e.g. Hampel et al. (1986).

Most of the work about robustness in the variance components estimation and/or repeated measures analysis has been done through the mixed linear models. There is an extensive literature about robust analysis of variance in mixed linear models. A review of the various methods can be found in Stahel and Welsh (1997) and in Welsh and Richardson (1997). Several different robust procedures are proposed but most of them are based on the maximum likelihood principle. In fact, most of these methods propose the use of a robustified log-likelihood instead of the classical log-likelihood objective function. To robustify the log-likelihood, one can for example replace the quadratic function in the log-likelihood by a slower growing one, see Huggins (1993a, 1993b) and Huggins and Staudte (1994). Another solution is to modify the estimating equations rather than the likelihood itself, see Richardson (1997). Richardson and Welsh (1995) propose two robust versions of the restricted maximum likelihood estimator in the general mixed linear models. All these methods are based on the fact that variance component models are based on an additive decomposition of variability into several components. Rocke (1983) states that this decomposition is a particular property of the variance which is not shared by other measures of spread. This is natural when the estimators are derived by modifying definitions of this form but other methods of defining estimators can be used as well. In particular, an estimator can be defined by the algorithm used to compute it and by modifying the algorithmic definition of such estimators to construct robust estimator. Rocke (1983) proposes an estimator of this type.

3.2.1 Estimation by maximizing a robustified likelihood

A general method for obtaining robust estimators is to define a robustified likelihood by replacing the quadratic function in (3.7) by a slower growing one in the sense that it has bounded derivatives. In the context of mixed linear models, Huggins (1993a) and Huggins and Staudte (1994) propose using

$$k^{T}\mathbf{X} = 0 \tag{3.11}$$

Harville (1977) refers to $k^{T}y$ for k^{T} satisfying (3.11) as being an "error contrast": its expected value is zero,

$$E(k^{T}\mathbf{y}) = k^{T}\mathbf{X}\boldsymbol{\alpha} = 0$$

The maximum possible number of linearly independent error contrasts k^{T} is $N - v$ where $v = rank(\mathbf{X})$. For example, a particular set of $N - v$ linearly independent error contrast is given by $\mathbf{K}^{T}\mathbf{y}$ where $\mathbf{K}^{T}$ is a $(N-v)\times N$ matrix whose rows are any $N-v$ linearly independent rows k^{T} of the matrix $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^{T}\mathbf{X})^{-}\mathbf{X}^{T}$.

For $\mathbf{K}^{T}\mathbf{X} = 0$, we then have the model

$$\mathbf{K}^{T}\mathbf{y} \sim N(0, \mathbf{K}^{T}\mathbf{V}\mathbf{K})$$

The REML equations can be derived from the ML equations (3.9), in which the terms are transformed. Indeed by making use of the transformation by $\mathbf{K}^{T}$, we replace

$$\begin{array}{ll} \mathbf{y} \text{ by } \mathbf{K}^{T}\mathbf{y} & \mathbf{Z} \text{ by } \mathbf{K}^{T}\mathbf{Z} \\ \mathbf{X} \text{ by } \mathbf{K}^{T}\mathbf{X} = 0 & \mathbf{V} \text{ by } \mathbf{K}^{T}\mathbf{V}\mathbf{K} \end{array}$$

so that $tr(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j^{T})$ in (3.9) becomes

$$tr[(\mathbf{K}^{T}\mathbf{V}\mathbf{K})^{-1}\mathbf{K}^{T}\mathbf{Z}_j\mathbf{Z}_j^{T}\mathbf{K}] = \mathbf{y}^{T}\mathbf{K}(\mathbf{K}^{\mathbf{T}}\mathbf{V}\mathbf{K})^{-1}\mathbf{K}^{T}\mathbf{Z}_j\mathbf{Z}_j^{T}\mathbf{K}(\mathbf{K}^{\mathbf{T}}\mathbf{V}\mathbf{K})^{-1}\mathbf{K}^{T}\mathbf{y} \tag{3.12}$$

Finally, let L_R be the likelihood function of $\mathbf{K}^{T}\mathbf{y}$ and define the log-likelihood l_R as

$$l_R(\boldsymbol{\alpha}, \mathbf{V}\,|\mathbf{y}) = \log L_R\,(\boldsymbol{\alpha}, \mathbf{V}\,|\mathbf{y}) = -\frac{1}{2}(N-r)\log(2\pi) - \frac{1}{2}\log\left|\mathbf{K}^{\mathbf{T}}\mathbf{V}\mathbf{K}\right| - \frac{1}{2}(\mathbf{y}^{T}\mathbf{K}(\mathbf{K}^{T}\mathbf{V}\mathbf{K})^{-1}\mathbf{K}^{T}\mathbf{y})$$

$$\rho(\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}))+(k_1/2)\log|\mathbf{V}| \tag{3.14}$$

instead of (3.7), where ρ is suitable non-negative function, $k_1 = E\psi(z)z$ with $z \sim N(0,1)$ and $\psi(z) = \rho^{'}(z)$. Welsh and Richardson (1997) refer to estimators obtained this way as Robust ML I estimators. The robustness properties of ρ-function depend on their derivatives ψ. In the mixed linear models context, both $\psi(z)$ and $z\psi(z)$ need to be bounded. Function ψ that satisfy this condition are said to be redescending. Example of such ρ functions are the Tukey biweight

$$\psi(z) = \begin{cases} z(1-(z/\kappa)^2)^2 & |z| \leq \kappa \\ 0 \text{ otherwise} \end{cases}$$

or the Hampel piecewise linear

$$\psi(z) = \begin{cases} z & |z| \leq \kappa_1 \\ \kappa_1 & \kappa_1 \leq \ |z| \leq \kappa_2 \\ (\kappa_3 - z)\kappa_1/(\kappa_3 - \kappa_2) & \kappa_2 \leq \ |z| \leq \kappa_3 \\ 0 & |z| \geq \kappa_3 \end{cases}$$

as well as others (see Chapter 4).

From (3.14), we deduce that the robust MLI estimators satisfies the equations

$$\mathbf{X}^T\mathbf{V}^{-1/2}\psi(\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})) = 0 \tag{3.15}$$

$$\begin{aligned} -k_1\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i^T)+\frac{1}{2}\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})^T\mathbf{V}^{-1/2}\mathbf{Z}_j\mathbf{Z}_j^T\mathbf{V}^{-1/2}\psi(\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})) &= 0 \quad (3.16) \\ \text{for } j &= 0,...,r \end{aligned}$$

However, Welsh and Richardson (1997) showed that the influence function of the robust ML I estimator is bounded in $\boldsymbol{\alpha}$ if ψ is a bounded redescending function but not in $\mathbf{X}$.

3.2.2 Estimating equations approach

While it is attractive to construct robust procedures by modifying likelihoods, much greater flexibility can be achieved if the estimating equations rather than the likelihood are modified. Welsh and Richardson (1997) define a very general class of estimators which they call the estimating equation estimator as the solution of

$$\mathbf{X}\mathbf{V}^{-1/2}\psi_0(\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})^T)=0 \tag{3.17}$$

and

$$-k_2\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j^T)+\frac{1}{2}\psi_1(\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}))\mathbf{V}^{-1/2}\mathbf{Z}_j\mathbf{Z}_j^T\mathbf{V}^{-1/2}\psi_2(\mathbf{V}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}))^T=0 \tag{3.18}$$

where $k_2 = E[\psi_1(z)\psi_2(z)]$ with $z \sim N(0,1)$. The estimating equation estimator with $\psi_0 = \psi_1 = \psi_2$ is called Robust ML II in Richardson and Welsh (1995). If $\psi_0 = \psi_2$ and $\psi_1(z) = z$, we obtain the Robust ML I estimator. In general there is no likelihood function which has (3.17)-(3.18) as its derivatives with respect to $\boldsymbol{\alpha}$ and the σ^2, but this is outweighed by the increased flexibility in constructing the estimating equations. For example, Richardson (1997) introduced bounded influence versions of the estimating equation estimators that are analogous to the bounded influence estimators of Mallows, Andrews, Hill and Ryand and Schweppe for the linear regression model (see e.g. Hampel et al., 1986, p. 347). The bounded influence estimating equation estimators are defined by the estimating equations

$$\mathbf{X}\mathbf{W}_0\mathbf{V}^{-1/2}\psi_0(\mathbf{V}^{-1/2}\mathbf{U}_0(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})^T)=0 \tag{3.19}$$

and

$$-k_2\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j^T)+\frac{1}{2}\psi_1(\mathbf{V}^{-1/2}\mathbf{U}_1(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}))\mathbf{W}_1\mathbf{V}^{-1/2}\mathbf{Z}_j\mathbf{Z}_j^T\mathbf{V}^{-1/2}\mathbf{W}_1\psi_2(\mathbf{V}^{-1/2}\mathbf{U}_1(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}))^T=0$$

Richardson's estimators (1997) have $\psi_0 = \psi_1 = \psi_2$, $\mathbf{W}_0 = \mathbf{W}_1$ and $\mathbf{U}_0 = \mathbf{U}_1$. The definition for $\mathbf{W}_0$, $\mathbf{W}_1$, $\mathbf{U}_0$ and $\mathbf{U}_1$ depends on the choice of the estimator. For example, for Mallows estimator $\mathbf{U}_0 = \mathbf{U}_1 = \mathbf{I}$ and for Andrews estimator, Richardson (1997) set $\mathbf{W}_0 = \mathbf{W}_1 = \mathbf{I}$.

3.2.3 B-optimal estimator

An alternative to the approach based on estimating equations is to construct B-optimal estimators as defined by Hampel et al. (1986). These estimators have minimum variance subject to a bound on the sensitivity at the core model and an attractive optimality property. Stahel and Welsh (1997) developed B-optimal estimation for the one factor ANOVA model (2.14) but with μ constant over all level l of the fixed effect. This model has the important simplifying characteristic that the parameters μ, σ_ε^2 and $\tau_s \ = \sigma_s^2 + \sigma_\varepsilon^2/l$ are orthogonal. It turns out that B-optimal estimating equations (treating each parameters separately) for σ_ε^2, τ_s and μ are :

$$\sum_{i=1}^{n} \psi\left(\frac{(y_i - \mu)^T (y_i - \mu)/\sigma_\varepsilon^2 - (l-1)}{\sqrt{(2(l-1))}} - k_\varepsilon\right) = 0$$

$$\sum_{i=1}^{n} \psi\left(\frac{(y_i - \mu)^2/\tau_s \ - 1)}{\sqrt{2}} - k_s\right) = 0$$

$$\sum_{i=1}^{n} \psi((\bar{y}_i - \mu)^2/\tau_s \ ) = 0$$

where ψ is the univariate Huber ψ function with tuning constant κ, $\bar{y}_i$ is the mean of the i^{th} subvector of y, k_ε and k_s are suitable consistency corrections. Stahel and Welsh (1997) also discussed a modification of B-optimal estimator that let the tuning constant κ in the function ψ tend to zero. This has the effect of replacing means with medians and the resulting estimators may be termed "most B-robust". It is more complicated to obtain B-robust estimators for more complex models, as it is not necessarily possible to find sets of orthogonal parameters. However, Welsh and Richardson (1997) propose a general formulation for this type of estimator based on the general procedure given in Hampel et al. (1986).

3.2.4 Robust versions of the restricted likelihood estimator

Richardson and Welsh (1995) present two modifications of the REML. Starting from

$$tr(\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j^T) = (\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})^T\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}) \text{ for } j = 1, .., r \tag{3.20}$$

they propose a two-stage robustification procedure. The first stage of robustification is to replace $\boldsymbol{\alpha}$ in (3.20) with $\boldsymbol{\alpha}_\rho$, a robust estimate of $\boldsymbol{\alpha}$. They suggest for the estimation of $\boldsymbol{\alpha}_\rho$ to use a robust weighted least squares estimate, which minimizes

$$\rho[\mathbf{V}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}_\rho)] \tag{3.21}$$

i.e. the solution of

$$\mathbf{X}^T\mathbf{V}^{-1/2}\psi[\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}_\rho)] = 0 \tag{3.22}$$

The second stage is to apply ψ functions and consistency corrections on (3.20). Richardson and Welsh (1995) give two different proposals :

Robust REML proposal I solve for $j = 0, ..., r$:

$$\frac{1}{2}\left\{\mathbf{V}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}_\rho)\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j^T\mathbf{V}^{-1/2}\psi(\mathbf{V}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}_\rho)) - k_1 tr(\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j^T)\right\} = 0 \tag{3.23}$$

Robust REML proposal II solve for $j = 0, ..., r$:

$$\frac{1}{2}\left\{\psi(\mathbf{V}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}_\rho))^T\mathbf{V}^{-1/2}\ \mathbf{Z}_j\mathbf{Z}_j^T\mathbf{V}^{-1/2}\psi(\mathbf{V}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}_\rho)) - k_2 tr(\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j^T)\right\} = 0 \tag{3.24}$$

As for the method by solving estimating equation, there is no associated robust restricted likelihood for (3.23) and (3.24). Richardson and Welsh (1995) argue that it is not a drawback since we have gained the flexibility of using separate estimating equations for the fixed effects and variance components.

3.2.5 Rocke's method

All of the above procedures are presented in terms of an optimization problem or a fixed point equation which needs to be solved. Rocke (1983) constructed robust estimators of the variance components by constructing robust pseudo observations. Consider the model

$$y_{ij} = \mu + \lambda_j \ + s_i + \varepsilon_{ij} \ \text{ for } i = 1,..,n, \quad j = 1,..,l$$

Rocke proposed to construct robust pseudo observations $\tilde{y}_{ij}$ which are used in place of the original observations to compute robust analysis of variance estimates σ_s^2 and σ_ε^2. The robust pseudo observations are constructed from an M-estimator of μ, λ_j and s_i (treating s_i as a fixed effect) and σ_ε^2 which are denoted $\tilde{\mu}$, $\tilde{\lambda}_j$, $\tilde{s}_i$ and $\tilde{\sigma}_\varepsilon^2$. The algorithm computes robust estimates $\tilde{s}$ and $\tilde{\sigma}_s^2$ of the location and scale of $\tilde{s}_i$ and defines pseudo observations $\tilde{y}_{ij}$ by

$$\tilde{y}_{ij} = \tilde{\mu} + \tilde{\lambda}_j + \tilde{s} + k_s\sigma_s\psi((\tilde{s}_i - \tilde{s})/\tilde{\sigma}_b) + k_e\sigma_e\psi(\tilde{\varepsilon}_{ij}/\tilde{\sigma}_\varepsilon)$$

where $\tilde{\varepsilon}_{ij} = y_{ij} - \tilde{\mu} - \tilde{\lambda}_j - \tilde{s}_i$ and k_s and k_ε are correction factors associated with the M-estimator of s and ε respectively. ψ is a bounded function on, for example the Huber function $\psi_c(x) = x\min(1, c/|x|)$ with a suitable constant c. For k_s and k_ε, Rocke used the constants given by Huber (1981, p.174). Rocke's algorithm and some modifications of it are discussed in Stahel and Welsh (1997). But as Welsh and Richardson (1997) argue, there are considerable difficulties in extending the method to more complex models.

3.3 Some remarks

Recall that an estimator can be robust in the sense to have a bounded influence function (as it is the case for the robust estimators of Richardson and Welsh) but not in the global sense if the amount of contamination is greater than its breakdown point. For example, Welsh and Richardson (1997) state that their procedures are likely to have very low breakdown point. It is therefore important to also consider robust estimators that have a high breakdown point especially when working in high dimensions. We will propose a robust estimator based on an S-estimator of multivariate location and scale that is known to have a high breakdown point.

Chapter 4

High breakdown multivariate estimator of location and constrained scale

4.1 Introduction

As presented in section 3.2, robust estimators for mixed linear models proposed until now are based on robustified likelihood functions or estimating equations. We propose instead to follow another approach, based on considering the data as multivariate normal with a covariance matrix constrained through the model. Hence, our robust estimators and the resulting tests are an extension of robust estimation of multivariate location and scale in which the structure of the covariance matrix is constrained.

In the unconstrained case, i.e. when the covariance matrix is not patterned, the statistical literature proposes several robust estimators of multivariate location and scale such as the minimum covariance determinant (MCD) (Rousseeuw and Leroy, 1987), S-estimator (Rousseeuw and Yohai,1984) or Huber's M-estimator (Kent and Tyler, 1991). All of these estimators have good robustness properties, especially concerning their breakdown point. To our knowledge, nothing has been done about the robust estimation of covariance matrices when they are constrained

to have a particular structure.

Here we propose to define an S-estimator when the covariance matrix is structured. S-estimators were introduced in the framework of multiple regression and where shown to have good robustness properties in high dimension. Davies (1987) investigated existence, consistency, asymptotic normality and breakdown point of multivariate S-estimators. It is also known (Lopuhaä, 1989) that an S-estimator of multivariate location and scale can be found by an iterative procedure. To ensure that the estimator does not yields estimates that are biased by data contamination, it is also necessary to start the iterative procedure of the S-estimator with a good robust starting point (Woodruff and Rocke, 1994). We suggest adapting a high breakdown estimator proposed by Maronna and Zamar (2002) as starting point.

4.1.1 Constrained S-estimator

An S-estimator of location and scale is defined through a ρ function having the following properties (Rousseeuw and Yohai, 1984)

(1) ρ is symmetric, continuously differentiable and $\rho(0) = 0$

(2) there exists $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$

(3) $E_\Phi[\rho]/\rho(c) = \epsilon^*$ where Φ is the standard normal and ϵ^* is the desired breakdown point (usually set to 1/2)

An S-estimator of multivariate location and scale is defined as the solution for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which minimize $|\boldsymbol{\Sigma}|$ subject to

$$n^{-1}\sum_{i=1}^{n}\rho\left[\sqrt{(\mathbf{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})}\right] = n^{-1}\sum_{i=1}^{n}\rho(d_i) = b_0 \tag{4.1}$$

where ρ is a function with the above properties and which usually satisfy

$$E_{\chi^2_p}[\rho(d_i)] = b_0$$

b_0 is the parameter chosen in order to determine the desired breakdown point. The breakdown point is given by the ratio of b_0 to the maximum of ρ (see Lopuhaä and Rousseeuw, 1991).

Therefore b_0 is usually computed for a chosen breakdown point ϵ^* and a ρ function by means of

$$b_0 = \epsilon^* \max_d \rho(d_i)$$

We propose here to derive expressions for the location and scale estimating equations of the S-estimator for mixed linear models. Let $\mathbf{y}_i$ be the response variable following a multivariate normal distribution with $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose that the covariance matrix $\boldsymbol{\Sigma}$ is of the form

$$\boldsymbol{\Sigma} = \sum_{j=0}^{r} \sigma_j^2 \mathbf{z}_j \mathbf{z}_j^T \quad \text{with } \mathbf{z}_0 = \mathbf{I}_N \quad \text{and } \sigma_0^2 = \sigma_\varepsilon^2 \tag{4.2}$$

where the $\mathbf{z}_j$ are the design matrix associated with each random effect.

Now recall that a multivariate S-estimator for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is defined through the following minimization problem : minimize $|\boldsymbol{\Sigma}|$ under the constraint given in (4.1). We can then write the Lagrangian Γ as

$$\Gamma = \log(|\boldsymbol{\Sigma}|) + \lambda[\frac{1}{n} \sum_{i=1}^{n} \rho(d_i) - b_0]$$

Solving the Lagrangian yields the following equations (using $\frac{\partial}{\partial \boldsymbol{\mu}}(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) = -2\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$ and $\frac{\partial}{\partial \boldsymbol{\sigma}_j^2} \log(|\boldsymbol{\Sigma}|) = tr(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\sigma}_j^2})$)

$$\frac{\partial \Gamma}{\partial \boldsymbol{\mu}} = \frac{2\lambda}{n} \sum_{i=1}^{n} u(d_i) \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) = 0 \tag{4.3}$$

$$\frac{\partial \Gamma}{\partial \sigma_j^2} = tr(\boldsymbol{\Sigma}^{-1} \mathbf{z}_j \mathbf{z}_j^T) - \frac{\lambda}{2n} \sum_{i=1}^{n} u(d_i)(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \mathbf{z}_j \mathbf{z}_j^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) = 0 \tag{4.4}$$

with $u(d_i)$ defined as

$$u(d_i) = \frac{\partial}{\partial d_i} \rho(d_i)/d_i = \psi(d_i)/d_i$$

From (4.3) we get

$$\frac{1}{n} \sum_{i=1}^{n} u(d_i)(\mathbf{y}_i - \boldsymbol{\mu}) = 0$$

so that

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{n} u(d_i)\mathbf{y}_i}{\sum_{i=1}^{n} u(d_i)} \tag{4.5}$$

Alternatively, if $\boldsymbol{\mu}_i = \mathbf{x}_i\boldsymbol{\alpha}$, (4.3) becomes

$$\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\alpha}}\right)^T \frac{\partial L}{\partial \boldsymbol{\mu}} = -\frac{2\lambda}{n}\sum_{i=1}^{n} u(d_i)\boldsymbol{\Sigma}^{-1}\mathbf{x}_i^T(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\alpha}) = \mathbf{0} \tag{4.6}$$

which reduces to

$$\boldsymbol{\alpha} = \left[\sum_{i=1}^{n} u(d_i)\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right]^{-1}\sum_{i=1}^{n} u(d_i)\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{y}_i \tag{4.7}$$

The following step is to find an expression for the variance components σ_j^2. To do so, we first isolate λ in equation (4.4). Using the properties of the trace,

$$\begin{aligned}
\sum_{j=0}^{r}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T) - \frac{\lambda}{2n}\sum_{j=0}^{r}\sum_{i=1}^{n} u(d_i)(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}) &= 0\\
\sum_{j=0}^{r}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\sigma_j^2) - \frac{\lambda}{2n}\sum_{j=0}^{r}\sum_{i=1}^{n} u(d_i)(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\sigma_j^2\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}) &= 0\\
\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\sum_{j=0}^{r}\mathbf{z}_j\mathbf{z}_j^T\sigma_j^2) - \frac{\lambda}{2n}\sum_{i=1}^{n} u(d_i)(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\sum_{j=0}^{r}\mathbf{z}_j\mathbf{z}_j^T\sigma_j^2\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}) &= 0\\
\mathrm{tr}(\mathbf{I}_p) - \frac{\lambda}{2n}\sum_{i=1}^{n} u(d_i)(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}) &= 0\\
p - \frac{\lambda}{2n}\sum_{i=1}^{n} u(d_i)d_i^2 &= 0
\end{aligned}$$

so that

$$\lambda = 2p[\frac{1}{n}\sum u(d_i)d_i^2]^{-1} \tag{4.8}$$

Using (4.8) in (4.4) yields

$$\frac{\partial \Gamma}{\partial \sigma_j^2} = \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T) - [\frac{1}{n}\sum u(d_i)d_i^2]^{-1}\left[\frac{1}{n}\sum pu(d_i)(\mathbf{y}_i - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\mathbf{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right] = 0$$

so that

$$tr(\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T) = \left[\frac{1}{n}\sum u(d_i)d_i^2\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} pu(d_i)(\mathbf{y}_i - \boldsymbol{\mu})\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\mathbf{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right] \tag{4.9}$$

In order to be able to solve equation (4.9) through an iterative procedure, we have to sort out the variance components σ_j^2. We rewrite the left part of equation (4.9) as

$$\begin{aligned}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T) &= \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\mathbf{\Sigma}^{-1}\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\mathbf{\Sigma}^{-1}\sum_{k=0}^{r}\sigma_k^2\mathbf{z}_k\mathbf{z}_k^T) \\ &= \sum_{k=0}^{r}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\mathbf{\Sigma}^{-1}\mathbf{z}_k\mathbf{z}_k^T)\sigma_k^2 \end{aligned} \tag{4.10}$$

Using an alternative formulation we can rewrite equation (4.9). Indeed, define

$$\mathbf{M}_j = \mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^{T} \quad \text{and} \quad \mathbf{S}_0 = [\sigma_0^2, ..., \sigma_r^2]^T$$

and let

$$\mathbf{Q} = [tr(\mathbf{M_j}\mathbf{M_k})]_{j,k=0,...,r} \tag{4.11}$$

so that equation (4.10) becomes

$$\begin{aligned}\mathbf{Q}S_0 &= \left[\frac{1}{n}\sum u(d_i)d_i^2\right]^{-1}\left[\frac{1}{n}\sum pu(d_i)(\mathbf{y}_i - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\mathbf{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right]_{j=0,...,r} \\ &= \left[\frac{1}{n}\sum u(d_i)d_i^2\right]^{-1}\mathbf{U}\end{aligned}$$

$$\frac{\partial}{\partial d}\rho(d;c,M) = \psi(d;c,M)\begin{cases} d & 0 \le d < M \\ d(1-(\frac{d-M}{c})^2)^2 & M \le d \le M+c \\ 0 & d > M+c \end{cases}$$

$$\psi(d;c,M)/d = u(d;c,M)\begin{cases} 1 & 0 \le d < M \\ (1-(\frac{d-M}{c})^2)^2 & M \le d \le M+c \\ 0 & d > M+c \end{cases} \tag{4.13}$$

where the parameter c and M can be chosen to achieve the desired breakdown point and ARP. When $M = 0$, the translated biweight ρ-function reduces to the Tukey's biweight. When $c = 0$, one gets the least winsorized squares (see Rocke 1996) given by

$$\rho(d;c) = \begin{cases} d^2/2 & 0 \le d < c \\ c^2/2 & d > c \end{cases}$$

$$\frac{\partial}{\partial d}\rho(d;c,M) = \psi(d;c,M) = \begin{cases} d & 0 \le d < c \\ 0 & d > c \end{cases}$$

$$\psi(d;c,M)/d = u(d;c,M) = \begin{cases} 1 & 0 \le d < c \\ 0 & d > c \end{cases}$$

We can now define a rather general algorithm to compute the constrained S-estimator. Given a starting point $\boldsymbol{\alpha}$ and $\mathbf{S_0}$, a chosen breakdown point ϵ^* and an ARP value π, the algorithm is as follows. First the constants c and M are found by solving

$$\begin{aligned} \epsilon^* \max_d \rho(d;c,M) &= E_{\chi^2_p}[\rho(d;c,M)] \ , \\ M + c &= \sqrt{(\chi^2_p)^{-1}(1-\pi)} \ ; \end{aligned}$$

with

$$\mathbf{U} = \left[\frac{1}{n} \sum pu(d_i)(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \mathbf{z}_j \mathbf{z}_j^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right]_{j=0,\ldots,r}$$

We can now define an expression for the variance components σ_j^2 given by

$$\mathbf{S_0} = \left[\frac{1}{n} \sum_{i=1}^{n} u(d_i) d_i^2 \right]^{-1} \mathbf{Q}^{-1} \mathbf{U} \tag{4.12}$$

A usual choice for the function ρ is Tukey's biweight (Beaton and Tukey 1974). Rocke (1996) argues that Tukey's ρ function fails to downweight outliers with large distances in high dimensions. This is measured by using the concept of asymptotic rejection probability (ARP) which can be interpreted as the probability of an estimator, in large samples under a reference distribution, to give a null (or nearly null) weight. Although the ARP should be small for sake of efficiency, it is useful to be able to downweight points that are very improbable under the null model. Rocke (1996) shows that the ARP of the S-estimator based on the biweight function tends to 0 as the dimension p rises. This means that points lying far away from the center of the data are not downweighted when p is large. Therefore he proposes a modified biweight estimator, namely the TBS (for translated biweight S-estimator), defined through,

$$\rho(d; c, M) = \begin{cases} d^2/2 & 0 \leq d < M \\ M^2/2 & \\ -M^2(M4 - 5M^2c^2 + 15c4)/30c^4 & \\ +d^2(0.5 + M^4/2c^4 - M^2/c^2) & \\ +d^3(4M/3c^2 - 4M^3/3c^4) & M \leq d \leq M + c \\ +d^4(3M^2/2c^4 - 1/2c^2) & \\ -4Md^5/5c^4 + d^6/6c^4 & \\ & \\ M^2/2 + c(5c + 16M)/30 & d > M + c \end{cases}$$

The constrained S-estimator of $\boldsymbol{\alpha}$ and $\mathbf{S}_0$ is computed through an iterative procedure by simultaneously solving

$$\begin{aligned}
\boldsymbol{\alpha}^{(t+1)} &= \left[\sum_{i=1}^{n} u(d_i^{(t)})\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right]^{-1}\sum_{i=1}^{n} u(d_i^{(t)})\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{y}_i \\
\mathbf{S}_{\mathbf{0}}^{(t+1)} &= \mathbf{Q}^{(t)^{-1}}\left[\frac{1}{n}\sum_{i=1}^{n} u(d_i^{(t)})d_i^{(t)}\right]^{-1}\mathbf{U}^{(t)}
\end{aligned}$$

We will call the resulting constrained S-estimator using the TBS, the $CTBS$. Finally, note that to compute the $CTBS$, Rocke (1996) actually rescales the distances by a factor k, i.e. d_i/k, with

$$k = \frac{d_{(q)}}{\sqrt{(\chi_p^2)^{-1}(q/(n+1))}} \tag{4.14}$$

where $d_{(q)}$ denotes the q-th ordered distance, and $q = [(n+p+1)/2]$. By doing so, Rocke (1996) scales the Mahalanobis distances so that since k varies with the dimension p, the desired breakdown point is maintained.

Note that to compute the d_i and hence the weights $u(d_i)$, one needs to reconstruct from $\mathbf{S_0}$, the covariance matrix $\boldsymbol{\Sigma}$ by using $\boldsymbol{\Sigma} = \sum_{j=0}^{r}\sigma_j^2\mathbf{z}_j\mathbf{z}_j^T$ and from $\boldsymbol{\alpha}$, the mean vector $\boldsymbol{\mu}$ by using $\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\alpha}$. Finally, we can use the mahalanobis distances as a diagnostic tool to detect outlying observations.

It should be noted that Wellmann (2000) defines an S-estimator for what he called a "one way random effects model" (similar to the model with structural equation (2.14) but where the $\boldsymbol{\mu}$'s are constant over all levels l) analogous to an S-estimator for multivariate location and scale. Wellman (2000) uses two useful results of Searle et al. (1992, p. 443). If the covariance matrix $\boldsymbol{\Sigma}$ is of the form $\sigma_\varepsilon^2\mathbf{I}_l + \sigma_s^2\mathbf{J}_l$ then

$$\begin{aligned}
|\boldsymbol{\Sigma}| &= \sigma_\varepsilon^{2l-1}(\sigma_\varepsilon^2 + l\sigma_s^2) \\
\boldsymbol{\Sigma}^{-1} &= \frac{1}{\sigma_\varepsilon^2}\left(\mathbf{I}_l - \frac{\sigma_s^2}{\sigma_\varepsilon^2 + l\sigma_s^2}\mathbf{J}_l\right)
\end{aligned}$$

Using these two results, he restates the minimization problem (4.1) as an S-estimation of multivariate location and scale which minimize $\sigma_\varepsilon^{2n-1}(\sigma_\varepsilon^2 + l\sigma_s^2)$ subject to

$$n^{-1}\sum_{i=1}^{n}\rho\left[l\frac{(\mathbf{y}_i-\boldsymbol{\mu})^T(\mathbf{y}_i-\boldsymbol{\mu})}{\sigma_\varepsilon^2}+\frac{l(\mathbf{y}_i-\boldsymbol{\mu})^2}{\sigma_\varepsilon^2+l\sigma_s^2}\right] = b_0 \tag{4.15}$$

We cannot use such a procedure because there is no explicit generalized definition for the determinant and the inverse of the various constrained covariances $\boldsymbol{\Sigma}$ when the models become more complicated.

4.1.2 The maximum likelihood estimator for constrained covariance matrix

To compute to maximum likelihood estimator for constrained covariance matrix, we can also use an analogous procedure as for the $CTBS$. Let l be the log-likelihood defined as

$$l = \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{n}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}) \tag{4.16}$$

Solving (4.16) with respect to $\boldsymbol{\mu}$ and σ_j^2 gives

$$\frac{\partial L}{\partial \boldsymbol{\mu}} = n\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}) = 0 \tag{4.17}$$

$$\frac{\partial L}{\partial \sigma_j^2} = \frac{n}{2}tr(\Sigma^{-1}\mathbf{z}_j\mathbf{z}_j^{T}) - \frac{n}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}) = 0 \tag{4.18}$$

From (4.17) we get an expression for the vector of fixed effect

$$\boldsymbol{\mu}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i \tag{4.19}$$

When considering $\boldsymbol{\mu}_i = \mathbf{x}_i\boldsymbol{\alpha}$, an alternative to (4.19) is given by

$$\boldsymbol{\alpha} = \left[\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{y}_i$$

and, using the same arguments as in the robust case, (4.18) leads to

$$\sum_{k=0}^{r} tr(\boldsymbol{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{z}_k\mathbf{z}_k^T)\sigma_k^2 = \quad [\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})]_{j=0,\ldots,r} = 0$$

Let $\mathbf{Q}$ be defined as in (4.11) and

$$\mathbf{U}_{MLE}= \left[\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{z}_j\mathbf{z}_j^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right]_{j=0,\ldots,r} \tag{4.20}$$

so that an iterative expression for the vector of random effect variances is given by

$$\mathbf{S}_{MLE} = \mathbf{Q}^{-1}\mathbf{U}_{MLE} \tag{4.21}$$

4.1.3 Starting point for the constrained robust estimator

It is important that the starting point of the iterative procedure of the $CTBS$ be also robust. To do this, one usually uses another robust high breakdown estimator as starting point. In this research, we choose to modify the Minimum Covariance Determinant, MCD of Rousseeuw and Leroy (1987) and the orthogonalized Gnanadesikan-Kettenring (OGK) estimator proposed by Maronna and Zamar (2002) to constrained covariance matrix. We first present the two estimators and then how we constraint them.

A natural choice is to use the MCD (Rousseeuw and Leroy 1987) estimator as starting point for the S-estimator. The objective of the MCD estimator is to find h observations (out of n) whose covariance matrix has the lowest determinant. The MCD mean estimator is then the sample mean of those h points, and the MCD covariance estimator is their sample covariance matrix. One difficulty is that the MCD requires a decision on h. One way is to choose a value of h that provides the maximum breakdown point. In this case, the minimal value of h is given by Rousseeuw and Leroy (1987) :

$$h^* = n - \lfloor\frac{n+p+1}{2}\rfloor$$

where $\lfloor x\rfloor$ denotes the integer part of x. But this is also the choice that gives the largest efficiency loss. In terms of computation of the MCD estimator, several algorithms have been

proposed. One of the most recent is the Fast-MCD of Rousseeuw and Van Driessen (1999). Here we propose to adapt it when the covariance matrix is constrained.

A key idea of the Fast-MCD algorithm is the fact that starting from any approximation to the MCD, it is possible to find an approximation with a lower determinant. Indeed, Rousseeuw and Van Driessen (1999) observed that from a subset H_k of size h in which μ, Σ and the Mahalanobis distances are computed, one can create a subset H_{k+1} by taking among the n observations the h ones with the smallest Mahalanobis distances with the property that the determinant of Σ based on H_{k+1} is smaller. Each step is called a C-step and can be defined more formally as follows. Consider the subsample H_k of size h and put

$$\begin{aligned} \boldsymbol{\mu}_k &= \frac{1}{h}\sum_{i\in H_k}\mathbf{y}_i \\ \boldsymbol{\Sigma}_k &= \frac{1}{h}\sum_{i\in H_k}(\mathbf{y}_i-\boldsymbol{\mu}_k)(\mathbf{y}_i-\boldsymbol{\mu}_k)^T \end{aligned}$$

If $|\Sigma_k| \neq 0$, calculate the distances

$$d_k(i) = \sqrt{(\mathbf{y}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_k)}$$

and form a new subset H_{k+1} by choosing the h point with the smallest distances d_k. Then $|\Sigma_{k+1}| \leq |\Sigma_k|$with equality only if $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_{k+1} = \boldsymbol{\Sigma}_k$.

Another robust estimator of multivariate location and scale has been proposed by Maronna and Zamar (2002) called OGK. Maronna and Zamar (2002) base their OGK on the robust estimator for covariances σ_{jk} proposed by Gnanadesikan and Kettenring (1972) which is very simple to compute and much more faster than the MCD estimator.

Define a multivariate variable $Y = (Y_1, ..., Y_p)$ that has supposedly generated the sample $\mathbf{y}_i$, $i = 1, .., n$ at hand. For a pair of random variables (i.e. $p = 2$) the OGK estimator is defined

as

$$\frac{1}{4}\left(\sigma(Y_j + Y_k)^2 - \sigma(Y_j - Y_k)^2\right)$$

where $\sigma()$ is a standard deviation function applied on its argument. A robust estimator for σ_{jk} is obtained when $\sigma()$ is a robust function. When $p > 2$, the covariance matrix $\mathbf{\Sigma}$ is estimated by replacing all its elements by all pairwise estimates. It is known that such an estimator may produce non positive definite matrices. To overcome the lack of positive definitness, Maronna and Zamar (2002) propose an estimator defined by the following four steps:

1. Let $\mathbf{D} = \text{diag}\left(\sigma\left(Y_j\right)\right)|_{j=1,\ldots,p}$ and define $\mathbf{x}_i = \mathbf{D}^{-1}\mathbf{y}_i, i = 1, \ldots, n$, i.e. realizations from $X = (X_1, \ldots, X_p)$

2. Compute the matrix $\mathbf{G} = (u_{jk})$ with

$$u_{jk} = \begin{cases} \frac{1}{4}\left(\sigma(X_j + X_k)^2 - \sigma(X_j - X_k)^2\right) & j \neq k \\ 1 & j = k \end{cases} \tag{4.22}$$

3. Decompose $\mathbf{G}$ as $\mathbf{G} = \mathbf{E}\Lambda\mathbf{E}^T$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$

4. Define $\mathbf{z}_i = \mathbf{E}^T\mathbf{x}_i$, i.e. realizations from $Z = (Z_1, \ldots, Z_p)$ and $\mathbf{A} = \mathbf{DE}$. The estimator of $\mathbf{\Sigma}$ is $\mathbf{A}\Gamma\mathbf{A}^T$ with $\Gamma = \text{diag}\left(\sigma\left(Z_j\right)^2\right)\Big|_{j=1,\ldots,p}$.

A location estimator for $\boldsymbol{\mu}$ is given by $\mathbf{A}\nu$ with $\nu = \left(m\left(Z_j\right)\right)|_{j=1,\ldots,p}$, $m()$ being a (robust) mean function. The procedure can be iterated by replacing $\mathbf{G}$ in step 2 by $\mathbf{E}\Gamma\mathbf{E}^T$ until convergence. For $\sigma()$ and $m()$, Maronna and Zamar (2002) propose the following functions

$$m(Y) = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \tag{4.23}$$

and

$$\sigma\left(Y\right)^2 = \frac{\sigma_0\left(Y\right)}{n}\sum_{i=1}^n \rho_{c2}\left(\frac{y_i - m(Y)}{\sigma_0\left(Y\right)}\right) \tag{4.24}$$

with

$$w_i = W_{c1}\left(\frac{y_i - m(Y)}{\sigma_0\left(Y\right)}\right)$$

and

$$\sigma(Y)_0 = MAD(Y)$$

$$W_c(x) = \begin{cases} \left(1 - \left(\frac{x}{c}\right)^2\right)^2 & |x| \leq c \\ 0 & \text{otherwise} \end{cases}$$

and

$$\rho_c(x) = \min\left(x^2, c^2\right)$$

Maronna and Zamar (2002) propose to use the values of $c1 = 4.5$ and $c2 = 3$. Moreover, they argue that to improve the efficiency of the OGK, one could use it as a hard rejection tool in that a reweighted estimator is used with

$$w_i = \begin{cases} 1 & \left(\mathbf{y}_i - \boldsymbol{\mu}_{\text{OGK}}\right)^T \boldsymbol{\Sigma}_{\text{OGK}}^{-1} \left(\mathbf{y}_i - \boldsymbol{\mu}_{\text{OGK}}\right) \leq \chi_p^2(.9) \\ 0 & \text{otherwise} \end{cases}$$

Note that this strategy is also used most of the times with the MCD but with the quantile 0.975 (instead of 0.9) of the χ_p^2.

To constraint the OGK and the MCD estimator we propose a rather "ad hoc" procedure. Let $\boldsymbol{\Sigma}_{okg}$ (respectively $\boldsymbol{\Sigma}_{mcd}$) be the robust unconstrained covariance matrix obtained with the OGK or the MCD. A constrained covariance matrix $\boldsymbol{\Sigma}$ (4.2) is obtained by taking the mean of the equal elements in the covariance matrix $\boldsymbol{\Sigma}_{okg}$ (respectively $\boldsymbol{\Sigma}_{mcd}$). For example, with model (2.14), the covariance matrix is patterned as in (2.17). In this case, the estimator for σ_1^2 will be the mean of the diagonal elements of $\boldsymbol{\Sigma}_{ogk}$ (respectively $\boldsymbol{\Sigma}_{mcd}$) and the estimator for σ_2^2 the mean of the off-diagonal elements. For the maximum likelihood estimator (MLE) we proceed in the same way. Note that for the MLE this procedure produces the same estimates as (4.20) and (4.21) when there is no missing data. This may not be the best method for constraining a covariance matrix but recall here that we are interested in estimating a starting point for the more general constrained S-estimator. More research is needed to propose a better starting point.

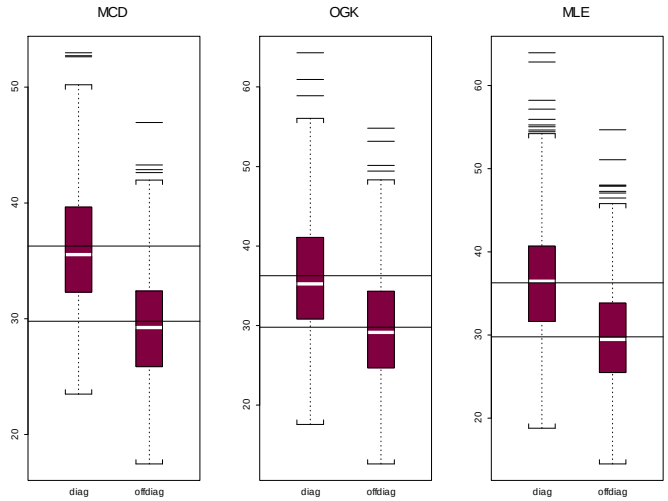

Figure 4-1: constrained estimators without contamination

To compare the estimates for $\mathbf{\Sigma}$ given by the OKG, MCD and MLE we performed a small simulation study. Suppose that the true underlying distribution of the data is normal multivariate with (the values for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are taken from a real example given in Griden (1992))

$$\boldsymbol{\mu} = \begin{bmatrix} 24 \\ 24 \\ 24 \\ 24 \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} 36.28 & 29.78 & 29.78 & 29.78 \\ 29.78 & 36.28 & 29.78 & 29.78 \\ 29.78 & 29.78 & 36.28 & 29.78 \\ 29.78 & 29.78 & 29.78 & 36.28 \end{bmatrix} \tag{4.25}$$

We generated 500 samples of size $n = 200$. We created small model deviations by generating $(1-\epsilon)\%$ of the data from the multivariate normal distribution with parameters given in (4.25), and $\epsilon\%$ from a multivariate normal distribution with the same covariance matrix, but with a shifted mean $\mu = [48\ 24\ 24\ 24]$. This type of model deviation produces so-called shift outliers (see e.g. Woodruff and Rocke, 1994). Figure 4-1 presents the estimates distribution for the diagonal and off-diagonal element of the constrained covariance matrix $\mathbf{\Sigma}$ when there is no contamination. We see that all three estimators are not biased.

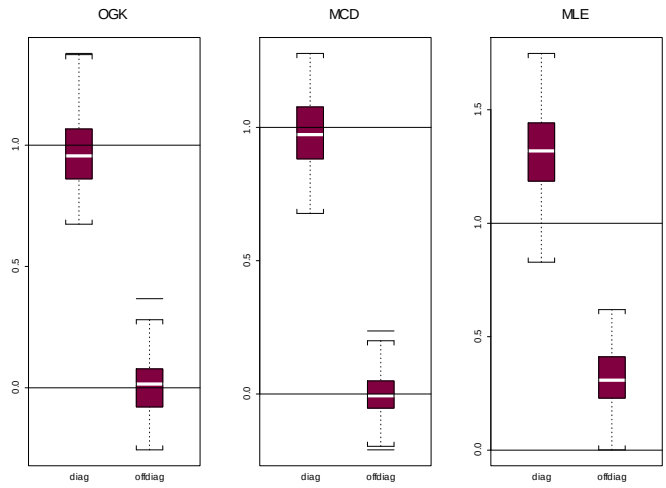

Figure 4-2: constrained estimators with 5% contamination

Figure 4-2 corresponds to the simulation with 5% of data contamination. We see that with contamination, the constrained MCD estimator and the constrained OGK estimator are not biased, whereas the maximum likelihood estimator is biased for the diagonal element of $\mathbf{\Sigma}$.

4.1.4 An algorithm procedure for the constrained S-estimator

The solutions of the constrained S-estimator are found via an iterative procedure as described in the above section. Below is the pseudo-code of the algorithm used to compute the constrained estimator.

- Given a model, define the design matrices $\mathbf{z}_j$ to get the structure of the covariance matrix and the matrices $\mathbf{x}_i$ that define the mean vectors $\boldsymbol{\mu}_i$, so that

$$\mathbf{\Sigma} = \sum_{j=0}^{r} \sigma_j^2 \mathbf{z}_j \mathbf{z}_j^T \text{ and } \boldsymbol{\mu}_i = \mathbf{x}_i \boldsymbol{\alpha}$$

- Compute the starting point of the constrained estimator either with the constrained MCD estimator or the OGK estimator, that is

$$\boldsymbol{\mu}_{i\ start} \text{ and } \boldsymbol{\Sigma}_{start}$$

- Compute the constrained estimator through the following iterative procedure :

  1 Compute the mahalanobis distances $d_i^{(1)} = \sqrt{(\mathbf{y}_i - \boldsymbol{\mu}_{start})^T \boldsymbol{\Sigma}_{start}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_{start})}$

  2 Compute the different weights $u(d_i^{(1)})$ and $u(d_i^{(1)})(d_i^{(1)})^2$

  3 Compute the fixed effects parameters

$$\boldsymbol{\alpha}^{(1)} = \left[\sum_{i=1}^{n} u(d_i^{(1)}) \mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i \right]^{-1} \sum_{i=1}^{n} u(d_i^{(1)}) \mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{y}_i$$

  4 Compute the random components vector

$$\mathbf{S_0}^{(1)} = \left[\frac{1}{n} \sum u(d_i^{(1)}) (d_i^{(1)})^2 \right]^{-1} \mathbf{Q}^{-1} \mathbf{U} = \left[(\sigma_0^{(1)})^2, (\sigma_1^{(1)})^2, ..., (\sigma_r^{(1)})^2 \right]$$

  5 Using the design matrices $\mathbf{z}_j$, build up the constrained matrix

$$\boldsymbol{\Sigma}^{(1)} = \sum_{j=0}^{r} \sigma_j^{2(1)} \mathbf{z}_j \mathbf{z}_j^T$$

  6 Compute the mean vectors $\boldsymbol{\mu}_i$

$$\boldsymbol{\mu}_i^{(1)} = \mathbf{x}_i \boldsymbol{\alpha}^{(1)}$$

  7 Compute some convergence criterion. If the conditions of the criterion are met, stop, otherwise put $\boldsymbol{\mu}_{i\ start} = \boldsymbol{\mu}_i^{(1)}$, $\boldsymbol{\Sigma}_{start} = \boldsymbol{\Sigma}^{(1)}$ and start again at point 1 by computing $d_i^{(2)}$, the weights $u(d_i^{(2)})$ and $u(d_i^{(2)})(d_i^{(2)})^2$ then $\boldsymbol{\mu}^{(2)}$, $\mathbf{S_0}^{(2)}$, $\boldsymbol{\Sigma}^{(2)}$. Repeat the procedure until convergence.

4.1.5 Concluding remarks

Finally, it should be stressed that although the framework of this paper is mixed linear models, the *CTBS* can actually be used for any model in which the covariance matrix can be written

as in (2.13). In fact the matrices $\mathbf{z}_j\mathbf{z}_j^T$ are simply the derivative of $\mathbf{\Sigma}$ with respect to one of the σ_j^2's. For example, if the only constraint imposed on the covariance matrix $\mathbf{\Sigma}$ is the constraint of symmetry then we will get $p(p+1)/2$ $\mathbf{z}_j\mathbf{z}_j^T$ design matrices representing the various derivatives of $\mathbf{\Sigma}$ with respect to the $p(p+1)/2$ non redundant σ_j^2. In this case, applying the algorithm for the $CTBS$ with these design matrices will yield the same matrix $\mathbf{\Sigma}$ as the one computed using the formula described in Lopuhaä (1989), that is

$$\mathbf{\Sigma} = \frac{\sum u(d_i)(\mathbf{y}_i - \mu)(\mathbf{y}_i - \mu)^T}{\sum u(d_i)d_i^2}$$

Of course with unconstrained matrices there is no real interest in using the $CTBS$ algorithm since it is more computational intensive than the simple iterative procedure described in Rocke (1996).

Chapter 5

Simulation study for the $CTBS$ estimator

The aim of this simulation study is to compare the behavior of the $CTBS$ against the maximum likelihood estimator under different kinds of models and data contamination. We first analyze the behavior of the $CTBS$: Estimation of the mean vector and of the various variance components will be analyzed thoroughly. This will be done by displaying the bias distribution (using boxplots) of the different estimates. We also study the behavior of the robust estimator in terms of breakdown point properties. Finally, we investigate how the robust estimator behaves in small sample situations.

We tried different amounts of contamination such as 2%, 5% and 10% but we present only the 2% case. As it will be shown below, the MLE is already seriously biased with as little as 2% of contamination, so that higher percentages are not informative. On the other hand, the $CTBS$ is not biased with as much as 10% of data contamination as the empirical breakdown study will show. We note that the $CTBS$ estimator was always computed using the constrained OGK estimator as starting point. The values of π and ϵ^* for the $CTBS$ estimator are 0.01 and 0.5.

5.1 One factor ANOVA model with repeated measures

The first model we are interested in is the simple one factor within-subject model defined through the following structural equation

$$y_{ij} = \mu + \lambda_j + s_i + \varepsilon_{ij} \qquad i = 1, .., n \quad j = 1, ..., l$$

Recall that we need $\Sigma_{j=1}^{l} \lambda_j = 0$ and that the random variables s_i and ε_{ij} are independent and have independent $N(0, \sigma_s^2)$ and $N(0, \sigma_\varepsilon^2)$ respectively $\forall ij$. We showed in Chapter 1 that we can rewrite this model in a multivariate formulation and that the $\mathbf{y}_i$ are then multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = vec(\mu + \lambda_j)\ \forall i$ and with a covariance matrix of the form $\boldsymbol{\Sigma} = \sigma_s^2 \mathbf{J}_l + \sigma_\varepsilon^2 \mathbf{I}_l$. That means that besides the mean vector, we have two variance components to estimate σ_s^2 and σ_ε^2. For our simulation, suppose that we want to study the reaction of subjects in four experimental situations ($l = 4$). Each of the n subject is observed under each experimental situation and, for example, a reaction time to some task is recorded. We also suppose that sufficient time is allowed between the experimental situation to avoid any interaction effect.

The values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\boldsymbol{\mu} = \begin{bmatrix} 24 \\ 24 \\ 24 \\ 24 \end{bmatrix}, \ \boldsymbol{\Sigma} = \begin{bmatrix} 49.6 & 9.4 & 9.4 & 9.4 \\ 9.4 & 49.6 & 9.4 & 9.4 \\ 9.4 & 9.4 & 49.6 & 9.4 \\ 9.4 & 9.4 & 9.4 & 49.6 \end{bmatrix} \tag{5.1}$$

so that $\mu = 24$, $\lambda_j = 0\ \forall j = 1, \ldots 4$, $\sigma_s^2 = 9.4$ and $\sigma_\varepsilon^2 = 49.6 - 9.4 = 40.2$. 1000 samples of size $n = 100$ were generated . Figure 5-1 presents the bias' distribution of the mean and variance components estimators ($CTBS$ and MLE) without model contamination. One can see that the $CTBS$ is not biased and appears to be, as expected, slightly less efficient than the MLE.

To create a small model deviation $(1 - \epsilon)\%$ of the data were generated from the multivariate normal distribution with parameters given in (5.1), and $\epsilon\%$ from a multivariate normal distribution with the same covariance matrix, but with a shifted mean $\boldsymbol{\mu} = [48\ 24\ 24\ 24]$. In other

words, $\mu + \lambda_1 = 48$ so that, $\mu = 30$, $\lambda_1 = 18$ and $\lambda_1 = \lambda_2 = \lambda_3 = -6$.

This type of model deviation produces so-called shift outliers (see e.g. Woodruff and Rocke, 1994). In Figure 5-2 are presented the bias' distribution of the mean and variance components estimators ($CTBS$ and MLE) with $\epsilon = 2\%$ of model contamination. Whereas the $CTBS$ is not affected by the model deviation, the MLE for the corresponding mean and the error variance are completely biased by only 2% of the data.

If the shift of the mean is applied on more than one mean component, the MLE of the subject variance is also affected. In Figure 5-3 are presented the bias' distribution of the mean and variance components estimators ($CTBS$ and MLE) with $\epsilon = 2\%$ and contaminated model with $\mu + \lambda_1 = \mu + \lambda_3 = 48$.

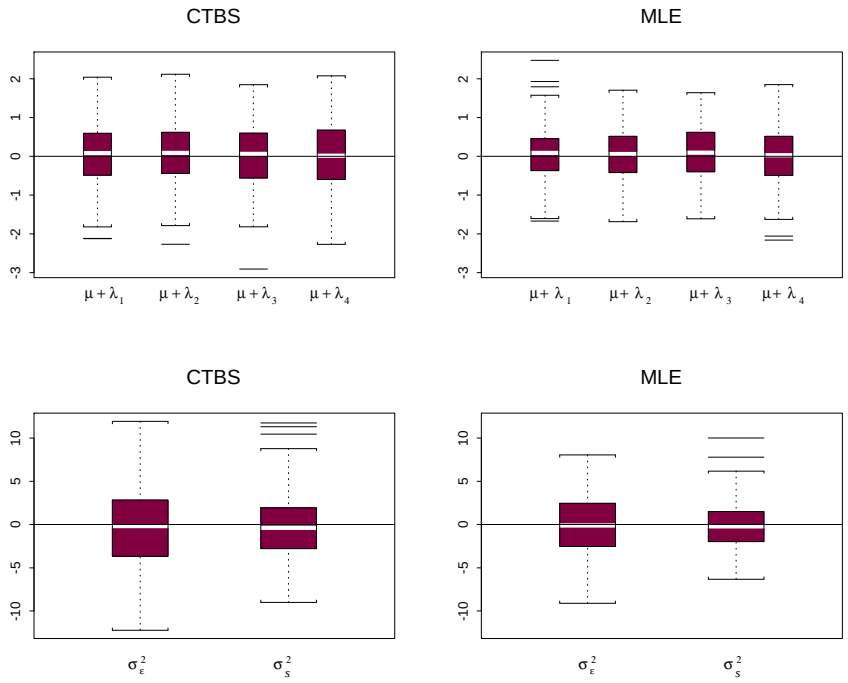

Figure 5-1: Bias distribution of the parameter's estimators for the one factor within-subject model, without data contamination

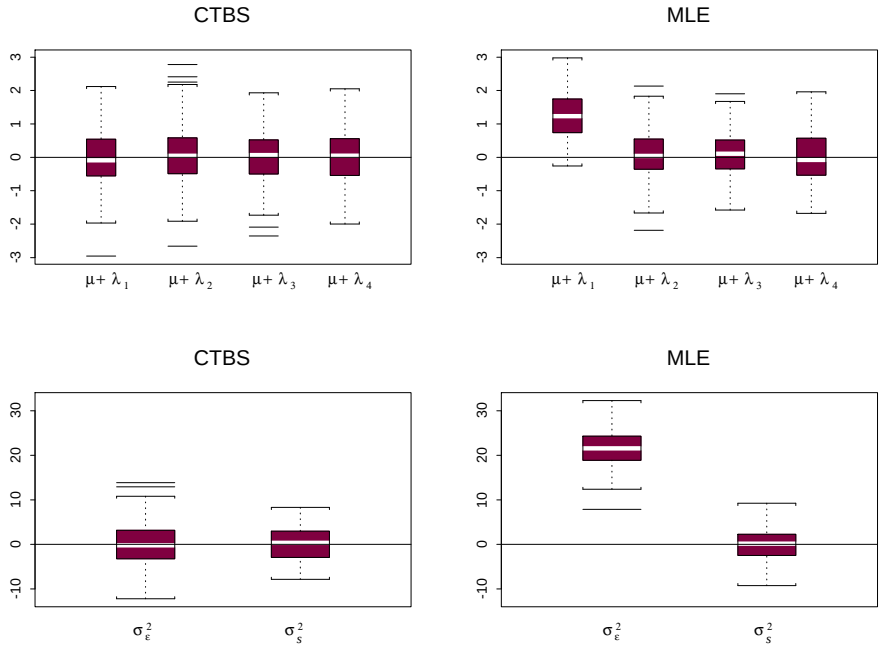

Figure 5-2: Bias distribution of the parameter's estimators for the one factor within-subject model, with 2% of data contamination on one mean ($\mu + \lambda_1 = 48$)

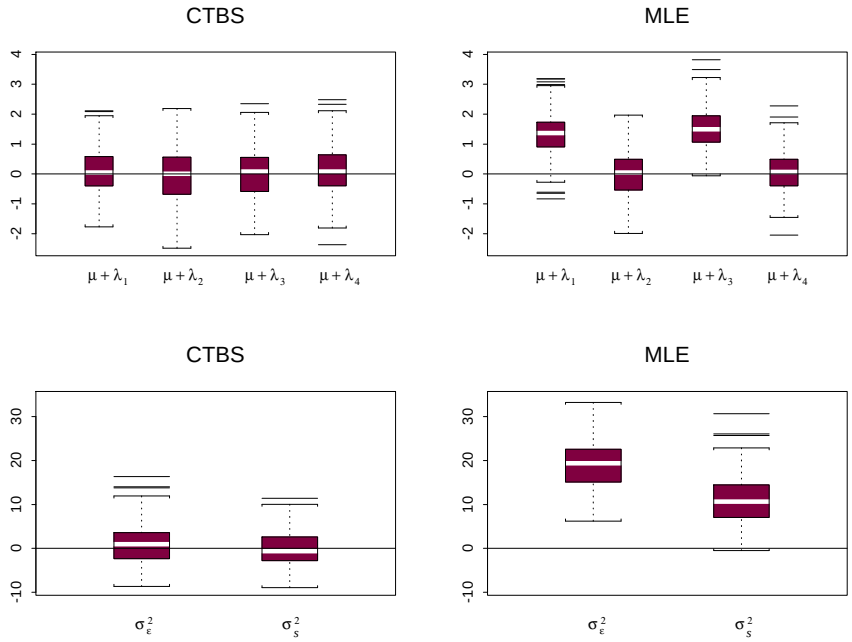

Figure 5-3: Bias distribution of the parameter's estimators for the one factor within-subject model, with 2% of data contamination on two means ($\mu + \lambda_1 = \mu + \lambda_3 = 48$)

5.2 Two factors within-subject ANOVA model

Here we are interested in a model of the form

$$y_{ijk} = \mu + \lambda_j + \gamma_k + (\lambda\gamma)_{jk} + s_i + (\lambda s)_{ij} + (\gamma s)_{ik} + \varepsilon_{ijk}$$

with $i = 1, .., n \quad j = 1, ..., l$ and $k = 1, ..., g$. We need $\Sigma_{j=1}^{l}\lambda_j = 0$, $\Sigma_{k=1}^{g}\gamma_k = 0$ and $\Sigma_{j=1}^{l}\Sigma_{k=1}^{g}(\lambda\gamma)_{jk}$. The unobservable random variables s_i, ε_{ij}, $(\lambda s)_{ij}$ and $(\gamma s)_{ik}$ are independent and have independent $N(0, \sigma_s^2)$, $N(0, \sigma_{\varepsilon}^2)$, $N(0, \sigma_{\lambda s}^2)$ and $N(0, \sigma_{\gamma s}^2)$ respectively. Using the formulation given in (2.12) we have that the $\mathbf{y}_i$ of size gl are multivariate normal with

$$\boldsymbol{\mu} = vec(\mu + \lambda_j + \gamma_k + (\lambda\gamma)_{jk})\forall i$$

and

$$\boldsymbol{\Sigma} = \sigma_s^2 \mathbf{J}_{gl} + \sigma_{\lambda s}^2 \left(\mathbf{I}_l \otimes \mathbf{J}_g\right) + \sigma_{\gamma s}^2 \left(\mathbf{J}_l \otimes \mathbf{I}_g\right) + \sigma_{\varepsilon}^2 \mathbf{I}_{gl}$$

1000 samples of size $n = 100$ were generated from the multivariate normal distribution with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} 24 \\ 24 \\ 24 \\ 24 \end{bmatrix}, \ \boldsymbol{\Sigma} = \begin{bmatrix} 70 & 16 & 14 & 10 \\ 16 & 70 & 10 & 14 \\ 14 & 10 & 70 & 16 \\ 10 & 14 & 16 & 70 \end{bmatrix}$$

so that $\mu = 24$, $\lambda_j = \gamma_k = (\lambda\gamma)_{jk} = 0 \forall j = 1, 2, k = 1, 2$, $\sigma_s^2 = 10$, $\sigma_{\lambda s}^2 = 6$, $\sigma_{\gamma s}^2 = 4$ and $\sigma_{\varepsilon}^2 = 50$. Figure 5-4 presents the bias' distribution of the mean and variance components estimators ($CTBS$ and MLE) without model contamination. Figures 5-5, 5-6 and 5-7 show the bias' distribution of the mean and variance components estimators ($CTBS$ and MLE) with 2% of model contamination, with respectively $\mu + \lambda_1 = 48$, $\mu + \lambda_1 = \mu + \lambda_2 = 48$ and $\mu + \lambda_1 = \mu + \gamma_1 = 48$ as contamination models. As with the previous model, we can conclude the following: the $CTBS$ is unbiased for all parameters with or without contamination, the MLE of the means are biased for the corresponding contaminated means, and depending on the type of contaminations, one or some of the variances are also seriously biased.

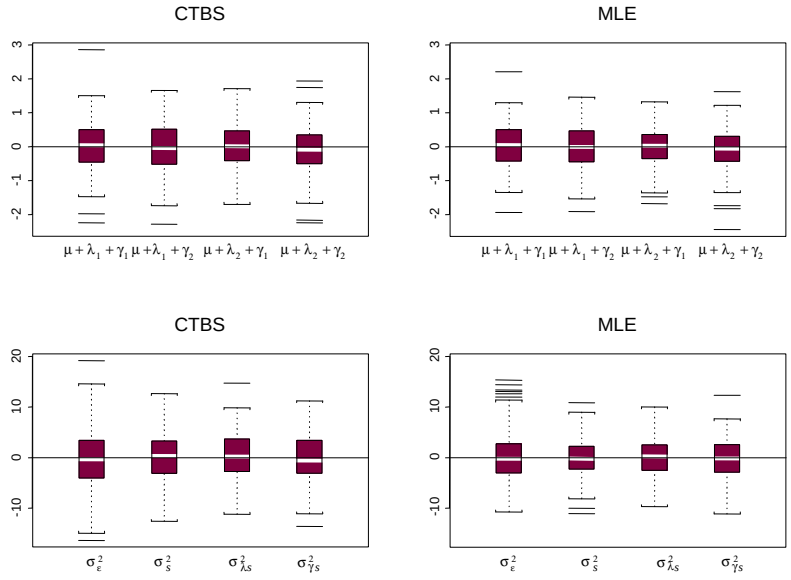

Figure 5-4: Bias distribution of the parameter's estimators for the two factors within-subject ANOVA model without data contamination

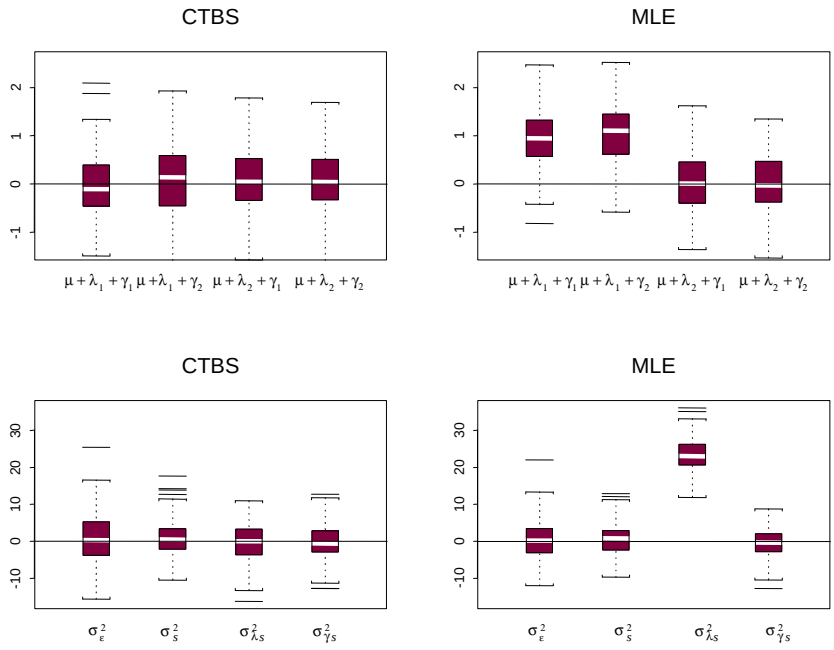

Figure 5-5: Bias distribution of the parameter's estimators for the two factors within-subject ANOVA model with 2% of data contamination ($\mu + \lambda_1 = 48$)

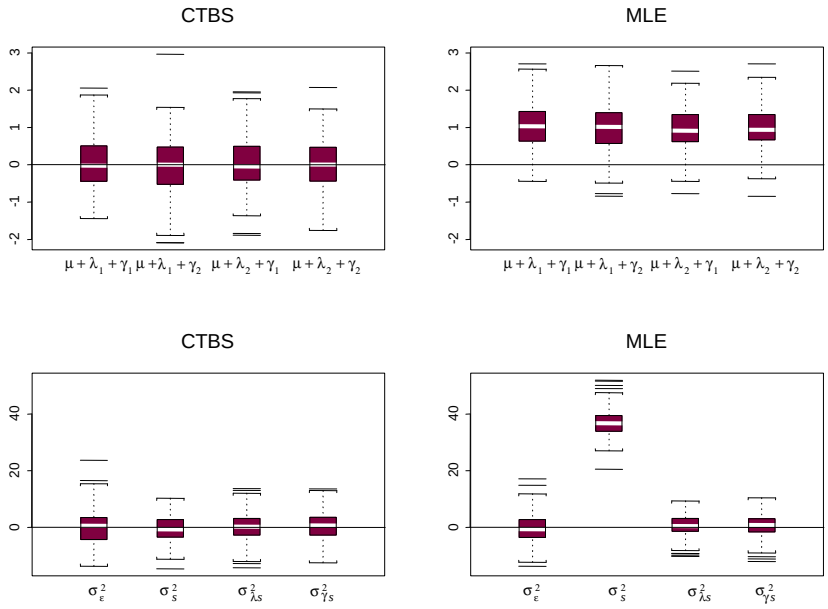

Figure 5-6: Bias distribution of the parameter's estimators for the two factors within-subject ANOVA model with repeated measures with 2% of data contamination ($\mu + \lambda_1 = \mu + \lambda_2 = 48$)

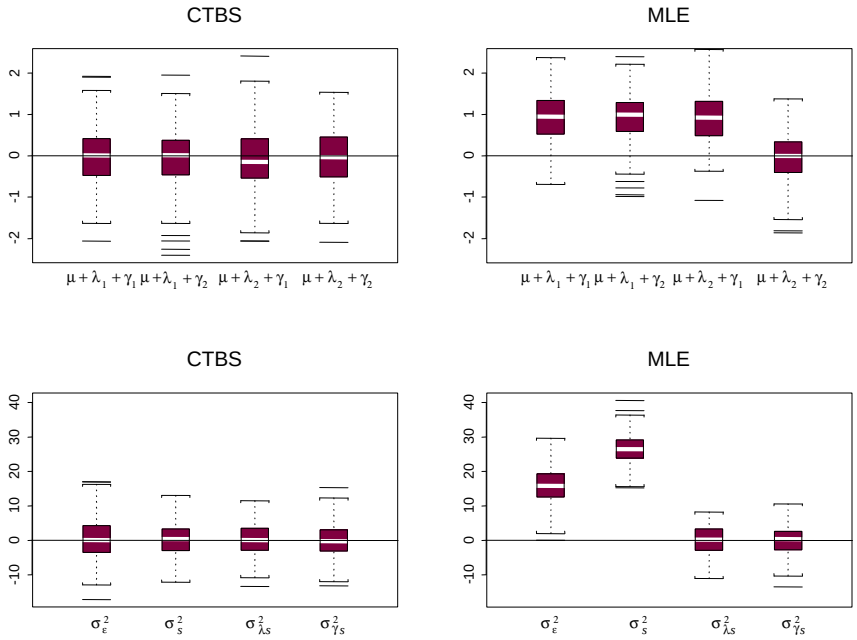

Figure 5-7: Bias distribution of the parameter's estimators for the two factors within-subject ANOVA model with 2% of data contamination ($\mu + \lambda_1 = \mu + \gamma_1 = 48$)

5.3 Hierarchical models

We consider here a hierarchical model with three nested random factors similar to the one presented in section 2.5

$$y_{ijklm} = \mu + \lambda J_i(j) + \gamma_{j(i)} + \delta_{j(i(k))} + \xi_{j(i(k(l)))} + \varepsilon_{j(i(k(l(m))))}$$

with the $\mathbf{y}_i$ multivariate normal $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu}_i = \mathbf{e}_8 \left(\mu + \lambda J_i(j)\right) = \mathbf{e}_8 \otimes \left(1, J_i(j)\right) \left(\mu, \lambda\right)^T = \mathbf{x}_i \boldsymbol{\alpha}$$

and

$$\boldsymbol{\Sigma} = \sigma^2_\gamma \mathbf{J}_8 + \sigma^2_\delta \mathbf{I}_2 \otimes \mathbf{J}_4 + \sigma^2_\xi \mathbf{I}_4 \otimes \mathbf{J}_2 + \sigma^2_\varepsilon \mathbf{I}_8$$

1000 samples of size $n = 100$ were generated from the multivariate normal distribution with the following parameters $\mu = 15$, $\lambda = 0$, $\sigma^2_\gamma = 2$, $\sigma^2_\delta = 2$, $\sigma^2_\xi = 1$, $\sigma^2_\varepsilon = 3$ and

$$J_i(j) = \begin{cases} 0 & j = 1 \\ 1 & j = 2 \end{cases}$$

Figure 5-8 presents the bias' distribution of the mean and variance components estimators ($CTBS$ and MLE) without model contamination. Figure 5-9 shows the bias' distribution of the mean and variance components estimators ($CTBS$ and MLE) with 2% of model contamination with $\lambda = 15$. Again, the $CTBS$ is unbiased for all parameters with or without contamination, the MLE of the means are biased for the corresponding contaminated means and the subject variance is also seriously biased.

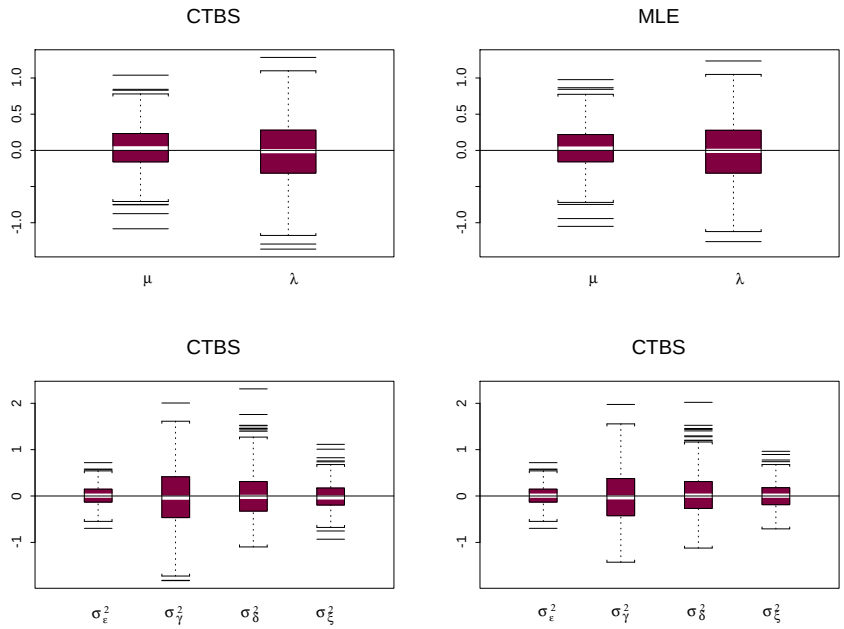

Figure 5-8: Bias distribution of the parameter's estimators for a hierarchical model with four random factors, without data contamination

5.4 Breakdown analysis

In this subsection, we study the breakdown properties of the $CTBS$, for the different models and the different model contaminations analyzed previously. We study the bias of the $CTBS$ when the amount of contamination increases up to 15%. Figure 5-10 shows the bias' distribution of the mean and variance components $CTBS$ estimator for the one factor within-subjects ANOVA model, with model contamination $\mu + \lambda_1 = \mu + \lambda_3 = 48$. Figure 5-11 shows the bias' distribution of the mean and variance components $CTBS$ estimator for the two factors within-subject ANOVA model, with model contamination $\mu + \lambda_1 = \mu + \gamma_1 = 48$. For the nested model and other model contaminations, we found similar results. One can see that the $CTBS$ can withstand up to 10% of contamination for the subject error, and even more for the other parameters. This doesn't prove that the empirical breakdown point is 10% since not all types of contamination have been considered, but the simulation study shows that the behavior of the $CTBS$ is very satisfactory: for a very small efficiency loss, the $CTBS$ can withstand up to 10% of model deviation.

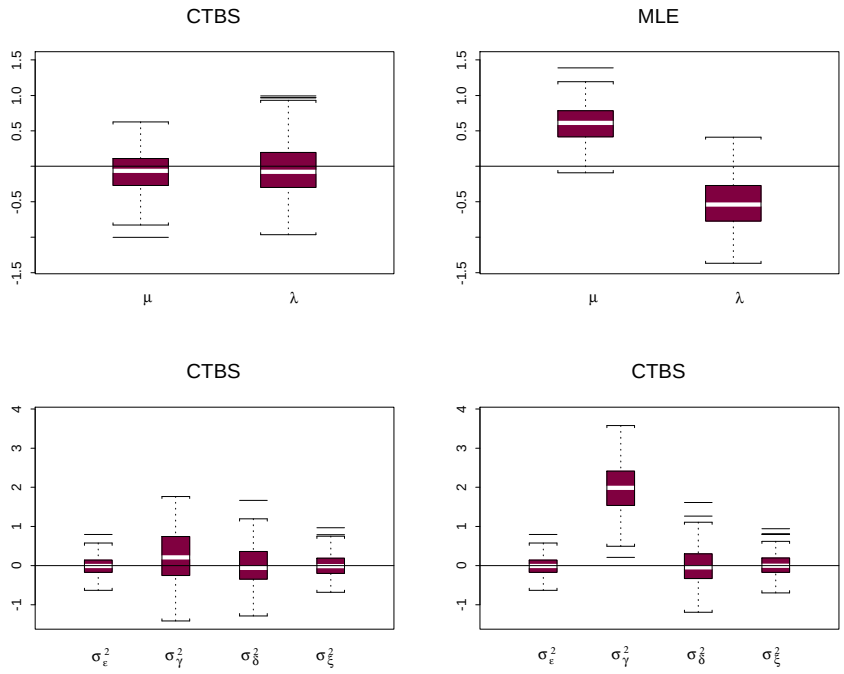

Figure 5-9: Bias distribution of the parameter's estimators for a hierarchical model with four random factors, with 2% of data contamination ($\lambda = 15$)

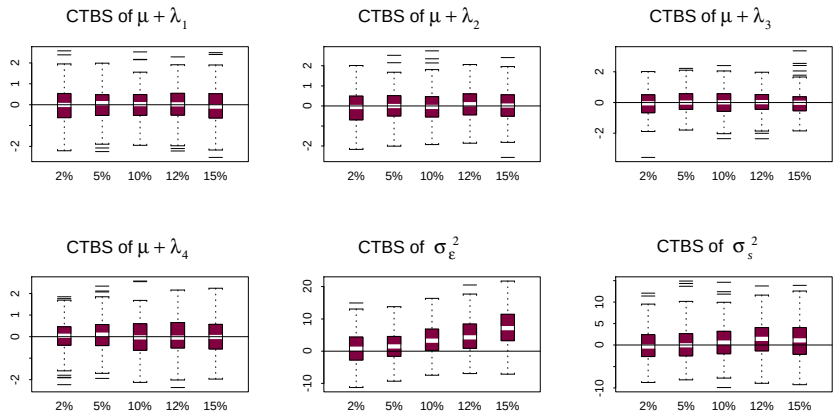

Figure 5-10: Bias distribution of the $CTBS$ for the one factor within-subject ANOVA model for different amounts of data contamination of type $\mu + \lambda_1 = \mu + \lambda_3 = 48$.

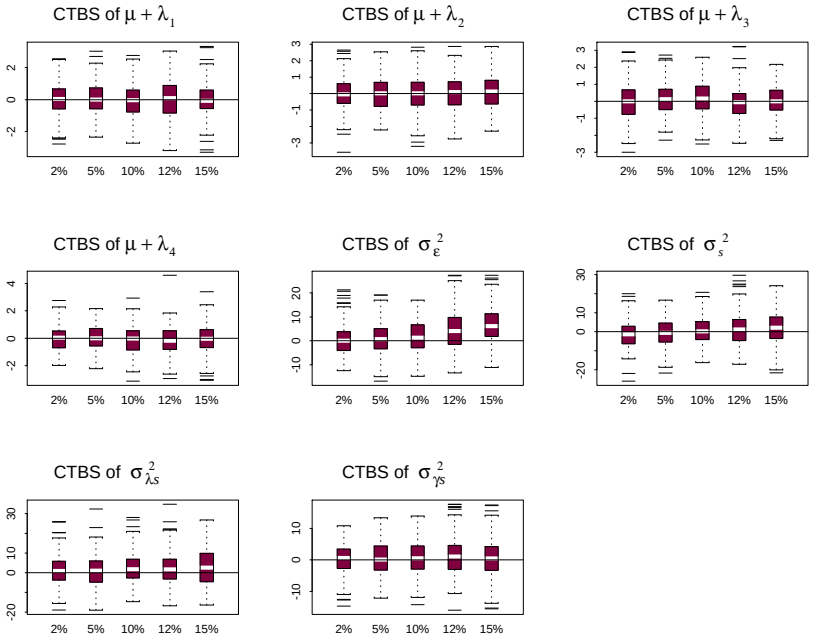

Figure 5-11: Bias distribution of the $CTBS$ for the two factors within-subject ANOVA model for different amounts of data contamination of type contamination $\mu + \lambda_1 = \mu + \gamma_1 = 48$.

5.5 Small sample behavior study

Until now the size of the simulation's samples were all set to $n = 100$. The choice of a relatively large size was motivated by the fact that we were primarily interested in the robustness aspect of the estimator and not in any asymptotic effect that might affect the results. However, a large size may be suitable from a theoretical point of view but in practice, one often encounters data sets of much smaller size.

Therefore, we explore the behavior of $CTBS$ in small sample situation. The idea is to detect if there is any potential bias when working in small samples. The data sets are not contaminated since it might interfere with the potential bias due to size. Again we analyze the one factor within-subject ANOVA model, the two factors within-subject ANOVA model and a hierarchical model. The chosen size are $n = 20$ and $n = 10$. In the graphics we compare the results with the maximum likelihood for information purpose only. The number of simulated data sets in each case is 500.

The first model is the one factor within-subject ANOVA model. For $n = 20$, the results are displayed in Figure 5-12. The mean vector is clearly not biased. The variance of the mean vector is greater with $n = 10$ than with $n = 100$ (see Figure 5-1). For the two variance components, results between the classical and robust estimator are of the same order and both estimators estimate correctly the variance components. Dividing the sample size by two doesn't change the results (see Figure 5-13) and there is no real difference between $n = 20$ and $n = 10$.

To continue to illustrate the behavior of the $CTBS$ estimator in small samples, we choose to simulate under the two factors within-subject ANOVA model. Again the sample sizes are set to $n = 10$ and $n = 20$. The aim is here to see if the number of parameters to estimate has an impact on the potential bias. Figures 5-14 and 5-15 show the bias distribution for $n = 20$ and $n = 10$ respectively. The same commentary we made for the one factor within-subject ANOVA model are applicable here. The mean vector is clearly unbiased for the $CTBS$ estimator and there is more variability for the variance components.

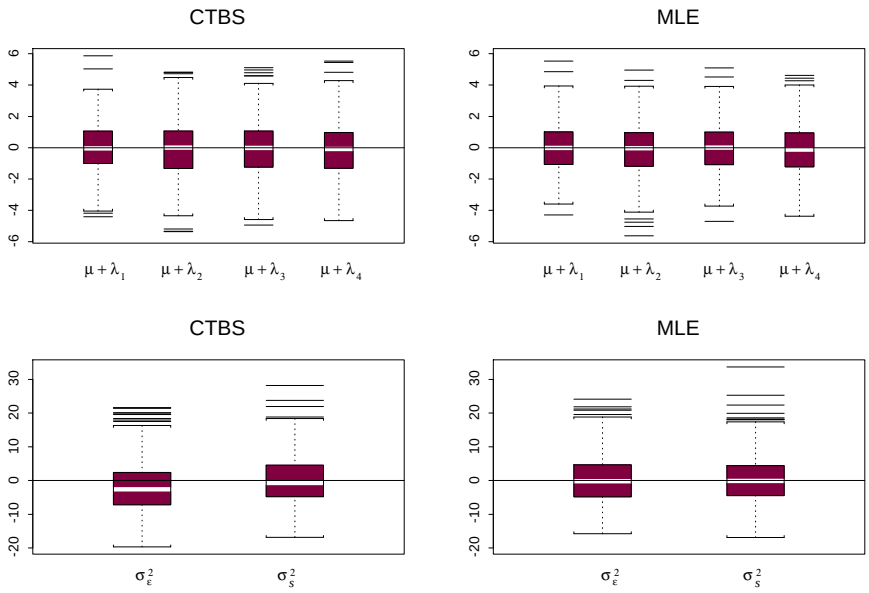

Figure 5-12: Bias distribution of the one factor within-subjet ANOVA model with $n = 20$.

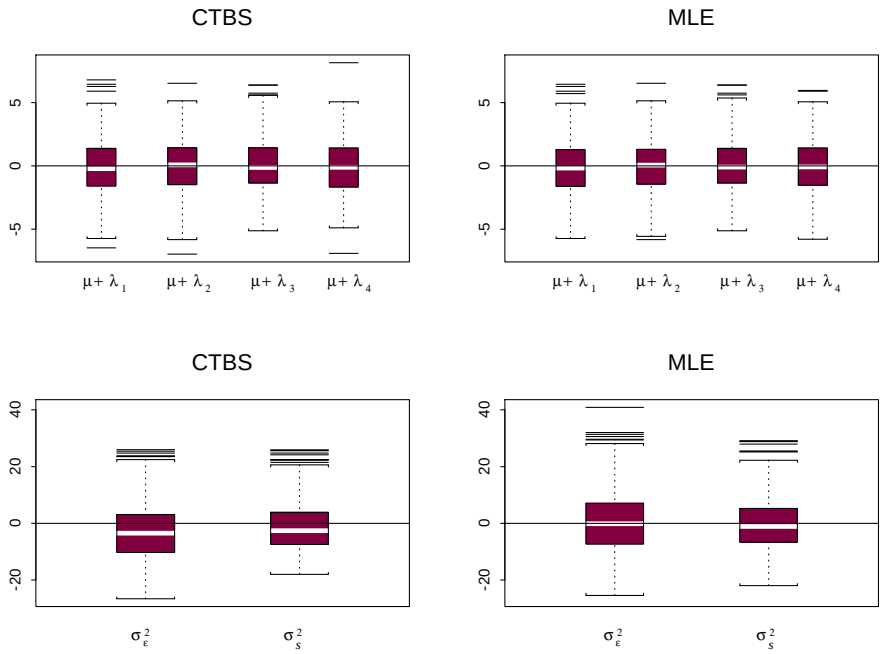

Figure 5-13: Bias distribution of the one factor within-subjet ANOVA model with $n = 10$.

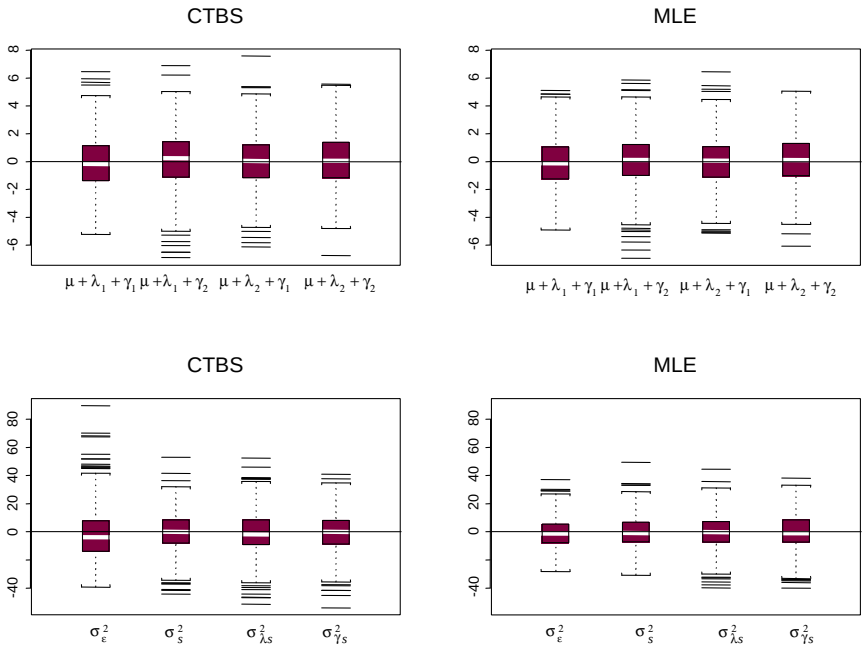

Figure 5-14: Bias distribution of the two-factors within subject ANOVA model with $n = 20$.

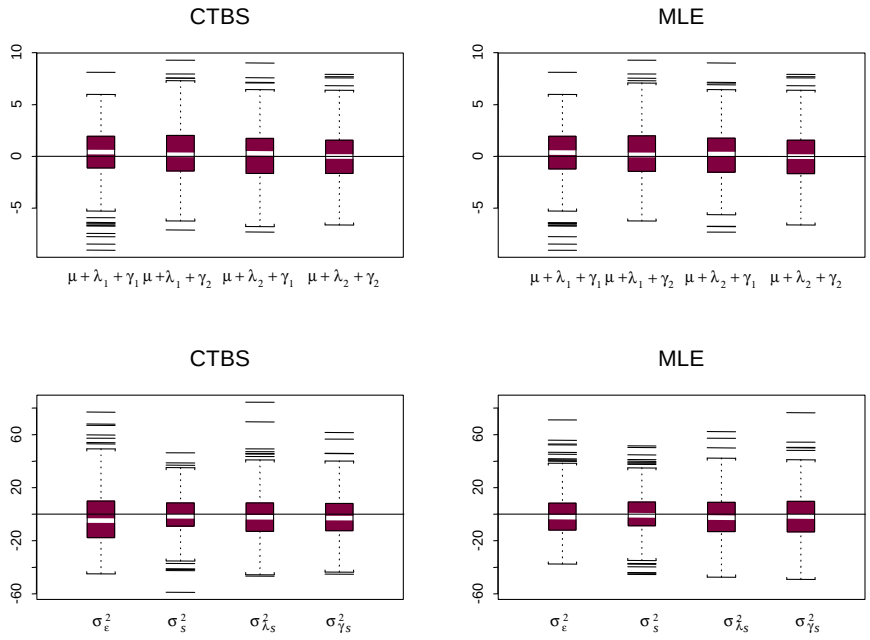

Figure 5-15: Bias distribution of the two-factors within subject ANOVA model with $n = 10$.

Finally, the last model we consider is a hierarchical model with 4 random factors. With this model we encountered numerical problems when computing the $CTBS$ estimator in small samples. The first problem was a convergence one in that the algorithm fails to converge after a very large number of iterations and the second one was some inversion matrix computational problems (in fact, during the iterative procedure, the covariance matrix is not positive definite anymore). To get an idea of the frequencies of this type of numerical problem, the $CTBS$ converged 500 times over 542 with data sets of size $n = 20$ and about 500 times over 612 with $n = 10$. This means roughly that for 8% and 20% of the data sets, the algorithm to compute the $CTBS$ did not converge. In fact this problem arises only with a hierarchical model with 4 random factors. When we lower the number of random factors to 3 or 2, these computational problems disappear.

Figure 5-16 displays the bias distribution for $n = 20$. The fixed effects μ and λ are not biased and the variability of their distribution is comparable between estimates.The same can be said about the estimates of the variance components. Figure 5-17 gives the bias distribution for $n = 10$. The interpretation remains the same as with $n = 20$.
In conclusion, with this small sample simulation study, we have seen that the $CTBS$ estimator remains stable when working in small dimensions. It is not biased down to a size of n =10 and the loss of efficiency is acceptable when compared to the maximum likelihood estimator. However, there is still some computational (or programming ?) problems that have to be addressed in further research.

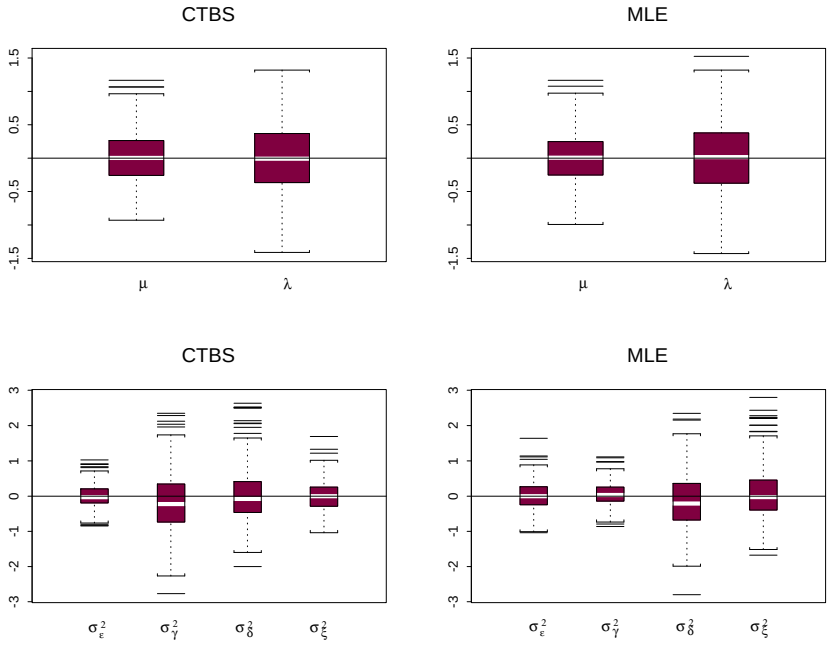

Figure 5-16: Bias distribution of the hierarchical model with 4 random factors, $n = 20$.

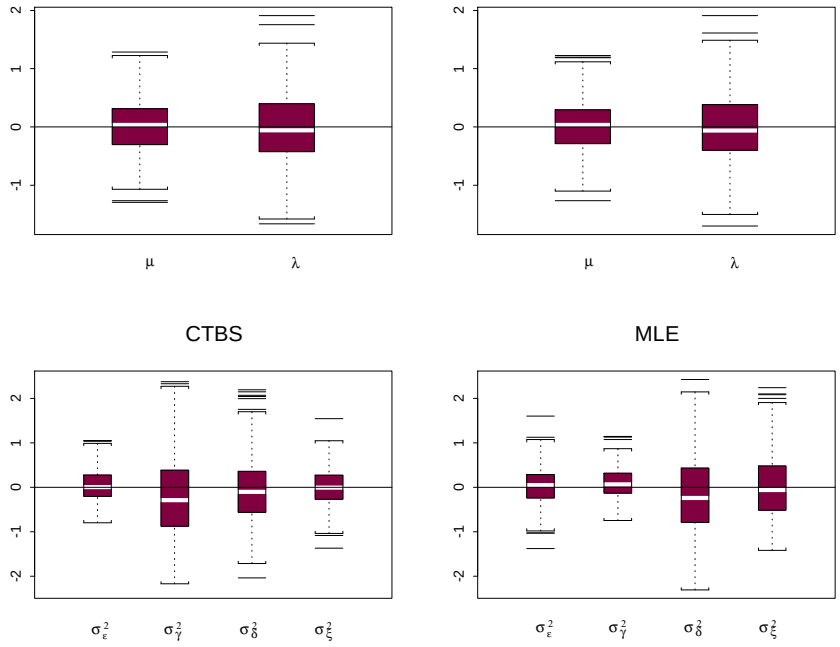

Figure 5-17: Bias distribution of the hierarchical model with 4 random factors, $n = 10$.

Chapter 6

Robust Inference

In classical ANOVA (with repeated measures or not) the primary goal is to test for the parameters. This is usually done by stating an hypothesis that descibes the test we are considering. Two cases can be considered. In the first case, we are interested in hypotheses involving only two parameters (or equivalently one sign of equality). This is what is called univariate hypothesis. To test such hyptoheses, one uses the t-test. However, the application of the t-test is limited and in the multivariate case, i.e. when considering hypotheses involving more than two parameters, we consider a generalization of the t-test, the F-test. As for the t-test with the student distribution, one uses the F distribution to test an hypothesis about the parameters of the population. This is usually done by creating an F-ratio. This ratio is formed by the various variance components and its form depends on the structural equation that defines the model.

In our case the need for robust testing is obvious. We cannot robustly estimate the parameters of a model and leave unchanged the usual procedures to test hypothesis about the parameters. Here we will use the results of Davies (1987) and Heriter and Ronchetti (1994). The latter propose robust versions of the Wald, score and likelihood ratio type tests for general parametric models and study their properties (asymptotic distributions, influence functions,...). The robust tests they propose are the natural counterpart of the corresponding classical tests ; see for instance Silvey (1975) or Rao (1973).

In their article, Heriter and Ronchetti (1994) suggest that the purpose of robust testing is

twofold. First, the level of a test should be stable under small, arbitrary departures from the null hypothesis. This is what they call robustness of validity. Second, the test should still have a good power under small arbitrary departures from specified alternative (i.e. robustness of efficiency). Heriter and Ronchetti (1994) explored the effect of contamination on the level and power of tests and showed that these effects are functions of the influence function of the underlying estimator. It means that a robust test, as a robust estimator, is able to deals with slightly misspecified models. When the data have not been generated exactly from the assumed models, robust testing procedures are by definition not (or very little) influenced by these model deviations.

6.1 Robust Inference for the univariate case

We propose here a univariate robust inference for the fixed effects. Such hypotheses can be described by means of contrasts. A contrast $\mathbf{c}$ is a linear combination of parameters $\boldsymbol{\alpha}$ such that $\mathbf{c}^T\boldsymbol{\alpha} = 0$. For example, suppose that we have a one factor within-subject ANOVA model with three levels. Our goal is to test for differences among elements of the mean vector $\boldsymbol{\mu}$. In terms of hypothesis, it is given by

$$\begin{aligned} H_0 &: \lambda_1 - \lambda_2 = 0 \\ H_1 &: \lambda_1 - \lambda_2 \neq 0 \end{aligned} \tag{6.1}$$

We know that for the one factor within-subject ANOVA model, the fixed effect vector $\boldsymbol{\alpha}$ is defined as $\boldsymbol{\alpha} = [\mu, \lambda_1, \lambda_2]^T$. A contrast $\mathbf{c}$ for hypothesis (6.1) is thus given by $\mathbf{c} = [0, 1, -1]$ so that $\mathbf{c}^T\ \boldsymbol{\alpha} = 0$.

Simple robust inference for the contrasts can be performed using the asymptotic results on MLE and S-estimators with the multivariate normal model. Indeed, for the MLE, we have that $\text{var}(\widehat{\boldsymbol{\mu}}) = \frac{1}{n}\boldsymbol{\Sigma}$, and for an S-estimator, we have that (see Davies, 1987 or Lopühaa, 1989)

$\text{var}(\widehat{\boldsymbol{\mu}}) = \frac{e_1}{ne_2^2}\boldsymbol{\Sigma}$ where

$$\begin{aligned} e_1 &= \frac{1}{p}E_\Phi\left[\left(\rho'(d)\right)^2\right] \\ e_2 &= E_\Phi\left[\left(1-\frac{1}{p}\right)u(d) + \frac{1}{p}\rho''(d)\right] \end{aligned}$$

with p the dimension of $\boldsymbol{\mu}$, ρ' (respectively ρ'') the derivative (respectively the second derivative) of ρ with respect to d, and the expectations are taken at the standardized p-variate normal distribution Φ. The values e_1 and e_2 depend on ρ (and the value of its parameters) and p. They can be obtained by Monte-Carlo simulations. Table 6.1 presents the values for $e_1/e_2^2 = e$ for different p, when ρ is the translated biweight with $\epsilon^* = 0.50$ and $\pi = 0.01$.

Table 6.1: e_1/e_2^2 for the translated biweight with $\epsilon^* = 0.50$ and $\pi = 0.01$ with 1000000 replicates for the Monte-Carlo simulations

p	2	3	4	5	6	7	8	9	10
e_1/e_2^2	1.5204	1.3920	1.2706	1.1994	1.1515	1.1164	1.0933	1.0739	1.0610

For both estimators, the covariance can be estimated by using the corresponding estimate for $\boldsymbol{\Sigma}$. In the case were $\boldsymbol{\mu} = \mathbf{x}_i\boldsymbol{\alpha}$ as in (2.25), we have that

$$\frac{e}{n}\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\text{var}\left(\mathbf{x}_i\widehat{\boldsymbol{\alpha}}\right) = \frac{1}{n}\mathbf{X}^T\text{var}\left(\widehat{\boldsymbol{\alpha}}\right)\mathbf{X}$$

with $\mathbf{X}^T = [\mathbf{x}_i]_{i=1,\ldots,n}$. We then deduce that

$$\begin{aligned} \text{var}\left(\widehat{\boldsymbol{\alpha}}\right) &= e\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1} \\ &= e\left(\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\left(\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i\right)^{-1} \end{aligned} \tag{6.2}$$

Note that for the MLE, $e = 1$ and (6.2) is equal to

$$\left(\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)^{-1}$$

We can then use contrasts to compute p-values. For example, a p-value associated to the hypoth-

esis $H_0 : \lambda_1 = \lambda_2$ in the one factor within-subject ANOVA model (2.15) with $\boldsymbol{\alpha} = [\mu, \lambda_1, \lambda_2, \ldots, \lambda_{l-1}]^T$ (and hence $\mathbf{x}_i = \begin{bmatrix} \mathbf{e}_{l-1} & \mathbf{I}_{l-1} \\ 1 & -\mathbf{e}_{l-1}^T \end{bmatrix}$) is obtained from the normal distribution with mean 0 and variance estimated by

$$e\mathbf{c}^T \left(\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i^T \widehat{\boldsymbol{\Sigma}} \mathbf{x}_i \left(\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \right)^{-1} \mathbf{c}$$

with the contrast defined as $\mathbf{c} = (0, 1, -1, 0, 0, ...)^T$. In Chapter 8, we will use these results to test contrasts with real data sets.

6.2 Robust multivariate tests

More generally, we are interested in testing the null hypothesis that q ($< p$) linearly estimable functions of the vector of parameters $\boldsymbol{\alpha}$ are zero. Denote by $\boldsymbol{\alpha}^T = (\boldsymbol{\alpha}_{(1)}^T, \boldsymbol{\alpha}_{(2)}^T)$ the partition of the vector $\boldsymbol{\alpha}$ into $p - q$ and q components and by $A_{(ij)}, i, j = 1, 2$ the corresponding partition of $p \times p$ matrices. Through a linear transformation for the parameters, this problem can be reformulated in testing

$$\begin{aligned} H_0 &: \boldsymbol{\alpha} = \boldsymbol{\alpha}_0 \text{ where } \boldsymbol{\alpha}_{0(2)} = 0, \boldsymbol{\alpha}_{0(1)} \text{ unspecified} \\ H_1 &: \boldsymbol{\alpha}_{0(2)} \neq 0, \boldsymbol{\alpha}_{(1)} \text{ unspecified} \end{aligned}$$

The tests that Heriter and Ronchetti (1994) consider rely on M-estimators defined as

$$\sum_{i=1}^{n} \psi(\mathbf{y}_i, \boldsymbol{\alpha}) = 0 \tag{6.3}$$

with asymptotic covariance matrix

$$\mathbf{H} = \mathbf{M}^{-1} \mathbf{K} \mathbf{M}^{-T}$$

where

$$\mathbf{M} = -\int \frac{\partial \psi}{\partial \boldsymbol{\alpha}}(\mathbf{y}, \boldsymbol{\alpha}) dF_{\boldsymbol{\alpha}}(\mathbf{y}) \tag{6.4}$$

$$\mathbf{K} = \int \psi(\mathbf{y}, \boldsymbol{\alpha}) \psi(\mathbf{y}, \boldsymbol{\alpha})^T dF_{\boldsymbol{\alpha}}(\mathbf{y}) \tag{6.5}$$

(for validity conditions, see Heriter and Ronchetti,1994). Note that for consistent M-estimates, one can write (6.4) as

$$\mathbf{M} = \int \psi(\mathbf{y}, \boldsymbol{\alpha}) s^T(\mathbf{y}, \boldsymbol{\alpha}) dF_{\boldsymbol{\alpha}}(\mathbf{y})$$

where $s(\mathbf{y}, \boldsymbol{\alpha})$ is the score function of model $F_{\boldsymbol{\alpha}}(\mathbf{y})$.

In practice, the asymptotic covariance matrix H can be estimated with

$$\mathbf{M} = \frac{1}{n} \sum_{i=1}^{n} \psi(\mathbf{y}_i, \boldsymbol{\alpha}) s^T(\mathbf{y}_i, \boldsymbol{\alpha}) \tag{6.6}$$

$$\mathbf{K} = \frac{1}{n} \sum_{i=1}^{n} \psi(\mathbf{y}_i, \boldsymbol{\alpha}) \psi^T(\mathbf{y}_i, \boldsymbol{\alpha}) \tag{6.7}$$

In the context of mixed linear models, we are interested in testing multivariate hypothesis in the fixed effects $\boldsymbol{\alpha}$. The $CTBS$ estimator of $\boldsymbol{\alpha}$ (ignoring the estimator for $\boldsymbol{\Sigma}$), can also be seen as M-estimators with ψ defined as

$$\psi(\mathbf{y}, \boldsymbol{\alpha}) = u(d) \left(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} \boldsymbol{\alpha} - \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} \right). \tag{6.8}$$

Note that in Lopühaa (1989) it is shown that S-estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ are asymptotically independent.

As in classical likelihood inference, there are three classes of tests that can be considered : Wald type tests, score (or Rao) type tests and likelihood ratio type test. Here we consider score and likelihood ratio type tests.

For an M-estimator, a score (or Rao) type test is based on the test statistic

$$R_n^2 = Z_n^T \mathbf{C}^{-1} Z_n$$

where

$$Z_n = \frac{1}{n}\sum_{i=1}^{n} \psi(\mathbf{y}_i, \boldsymbol{\alpha}^{\omega})_{(2)}$$

$\boldsymbol{\alpha}^{\omega}$ is the M-estimator in the reduced model, i.e. the solution of the equation

$$\sum_{i=1}^{n} \psi(\mathbf{y}_i, \boldsymbol{\alpha}^{\omega})_{(1)} = 0 \text{ with } \boldsymbol{\alpha}^{\omega}_{(2)} = 0$$

$$\mathbf{C} = \mathbf{M}_{(22.1)}\mathbf{H}_{(22)}\mathbf{M}^{T}_{(22.1)} \tag{6.9}$$

is a $q \times q$ positive definite matrix,

$$\mathbf{M}_{(22.1)} = \mathbf{M}_{(22)} - \mathbf{M}_{(21)}\mathbf{M}^{-1}_{(11)}\mathbf{M}_{(12)} \tag{6.10}$$

and $\mathbf{M}$ and $\mathbf{H}$ are computed using $\boldsymbol{\alpha}^{\omega}$.
Heritier and Ronchetti (1994) show that, under H_0, the statistic nR_n^2 is asymptotically χ^2_q. When $\psi(\mathbf{y}, \boldsymbol{\alpha}) = s(\mathbf{y}, \boldsymbol{\alpha})$ the test is the classical score or Rao test.

A likelihood ratio test is defined by a test statistic of the form

$$S_n^2 = \frac{2}{n}\sum_{i=1}^{n} [\rho(\mathbf{y}_i, \boldsymbol{\alpha}) - \rho(\mathbf{y}_i, \boldsymbol{\alpha}^{\omega})] \tag{6.11}$$

$\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^{\omega}$ are the M-estimator in respectively the full and reduced model. With S-estimators however, the likelihood ratio test cannot be used. Indeed, the S-estimator is the solution that minimize the determinant of the covariance matrix under the constraint that

$$\frac{1}{n}\sum_{i=1}^{n} \rho[\sqrt{(d_i/k}] = b_0 \tag{6.12}$$

The constraint (6.12) does not depend on the model (full or reduced) and it implies that the statistic S_n^2 will always be equal to zero. We will therefore consider here only the score type test.

Since the above results are only asymptotically true, the size of the data sets can play a con-

siderable role in terms of accuracy of the distribution of the robust test statistics. As a size too small can cause the distribution of the statistic to shift from the asymptotic one and hence yield to the wrong decision. We will therefore also investigate the effect of small samples in testing in a simulation study (see Chapter 7).
Consider now the following example. Suppose that in the one factor within-subject ANOVA model with $\boldsymbol{\alpha} = [\mu, \lambda_1, ..., \lambda_{l-1}]^T$, we want to test globally the fixed effect factor, i.e.

$$\begin{aligned} H_0 &: \lambda_1 = ... = \lambda_{l-1} = 0, \ \mu \text{ unspecified} \\ H_1 &: \text{one of the } \lambda_j \neq 0, \ \mu \text{ unspecified} \end{aligned}$$

To compute the test statistics R_n^2, one can use the following iterative steps. At the ith iteration,

1. Compute the vector of fixed effects

$$\hat{\boldsymbol{\alpha}} = \left[\sum_{i=1}^{n} u(d_i)\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right]^{-1} \sum_{i=1}^{n} u(d_i)\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{y}_i$$

2. Constraint $\hat{\boldsymbol{\alpha}}$ by replacing the corresponding elements in the vector by zero (e.g. $\hat{\boldsymbol{\alpha}}^{\omega} = [\hat{\mu}, 0, ..., 0]^T$)

3. Compute the Mahalanobis distances d_i^{ω} and the weights $u(d_i^{\omega})$ using $\hat{\boldsymbol{\mu}}^{\omega} = \mathbf{x}\hat{\boldsymbol{\alpha}}^{\omega}$

4. Using the new weights $u(d_i^{\omega})$, start again at step one and iter until convergence.

5. Compute

$$\mathbf{Z}_n = \frac{1}{n}\sum_{i=1}^{n} u(d_i)\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\hat{\boldsymbol{\alpha}}^{\omega} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right)_{(2)}$$

6. Compute

$$R_n^2 = Z_n^T\mathbf{C}^{-1}Z_n$$

The matrix $\mathbf{C}$ is the asymptotic covariance matrix of Z_n given in (6.9). In the following Chapter, we will study the behavior of the robust score test under various models.

Chapter 7

Simulation study for the robust score type test

In this Chapter we study the robust version of the score test in the setting of mixed linear models. We propose the application of the robust and non-robust score type test to different models. The interest here is twofold. First we want to study the robustness (or non-robustness) of the tests when the data are contaminated. We will see that in the case of contaminated data sets, the robust score test outperforms its classical counterpart. Second, the effect of the sample size will also be investigated. Indeed, it is well known (see for example Gallant, 1987) that the asymptotic χ^2_q distribution for score type test does not really hold for small or moderate sample sizes.

The idea here is to compare the theoretical levels of the tests, which are fixed a priori, and the experimental ones given by the simulations. If the test behaves well, one can expect small differences between those two levels. To be fair in our comparison, we also included the classical F-test. This is by far the most used test in the mixed models or ANOVA settings and it is known to be even exact under normality. The models we studied are the simple one factor within-subject ANOVA model which is easy to compute and gives a good appreciation of the robust score type test. Second we propose to study a two factors within-subject ANOVA model and finally a model with data generated from a hierarchical model with one covariate.

7.1 One factor within-subject ANOVA model

Recall that such model is given by the equation

$$y_{ij} = \mu + \lambda_j + s_i + \varepsilon_{ij}$$

Values for the model's parameters are $\mu = 24$, $\lambda_j = 0\ \forall j = 1, \ldots 4$, $\sigma_s^2 = 9.4$ and $\sigma_\varepsilon^2 = 40.2$. In this case $\boldsymbol{\alpha} = [\mu, \lambda_1, \lambda_2, \lambda_3]^T$ and the hypothesis is stated as

$$\begin{aligned} H_0 &: \lambda_1, \lambda_2, \lambda_3 = 0, \mu \text{ unspecified} \\ H_1 &: \lambda_1, \lambda_2, \lambda_3 \neq 0, \mu \text{ unspecified} \end{aligned}$$

For this model the F-test is given by the ratio

$$\frac{n \sum_{j=1}^{4} \hat{\lambda}_j^2 / (l-1)}{\hat{\sigma}_\varepsilon^2}$$

where $\hat{\lambda}_j$ and $\hat{\sigma}_\varepsilon^2$ are the restricted maximum likelihood estimates. The ratio follows a Fisher distribution with $(l-1)$ and $(l-1) \times (n-1)$ degrees of freedom.

10000 samples of size 100, 20 and 10 were generated under the null hypothesis. The percentages of contamination ϵ are 0%, 2%, 5% and 10%. For each test we recorded how many times the tests were rejected at the theoretical levels of 1%, 5% and 10% and reported those values in the tables below.

When there is no contamination the only parameter of interest is the size of the data set. For a value of n relatively large, one can expect a good behavior of those three tests. Indeed, as it can be seen in Table 7.1, when the size is set to $n = 100$, all tests perform well and are very close to each other.

However, as shown in Table 7.2, when we lower the sample size to $n = 20$, both the classical

and robust score tests fail to approach the theoretical level of 1 percent with values equal to 0.28 and 0.25 percent respectively. With 5% and 10% levels, the χ^2_q distribution still holds for both tests, being slightly too conservative at 5%. On the contrary, we see clearly that as expected the classical F-test is not influenced by the size of the data set. With a size of $n = 10$, clearly the χ^2_q distribution is not a good approximation for the score type tests. At levels of 1% and 5% the tests are too conservative and completely fail to approximate the theoretical levels (see Table 7.3). The classical F-test performs well.

We now proceed with contaminated data sets. The contamination method is the same as the one used in Chapter 5. To create a small model deviation, $(1 - \epsilon)\%$ of the data were generated from the multivariate normal distribution with parameters given in (5.1), and $\epsilon\%$ from a multivariate normal distribution with the same covariance matrix, but with a shifted mean $\boldsymbol{\mu} = [48\ 24\ 24\ 24]$. More precisely, the shape of the mean vector is obtained by letting $\mu+\lambda_1 = 48$ so that $\boldsymbol{\alpha} = [30, 18, -6, -6]^T$. The results presented here are obtained with $n = 100$. With different samples sizes, the results are similar. Table 7.4 contains the results we get when the contamination level is set at 2%. The results clearly show that the robust version of the score type test is not influenced by data contamination. It outperforms the classical score test which is very sensitive even to a small amount of contamination. Even if it fails to accurately approximate to true theoretical level, the F-test gives better approximation than the classical score test but is also clearly influenced by a small contamination. We also tried with 5% percent of contamination (see Table 7.5). The robust score test is not influenced by outliers. The classical score test and the F-test are clearly influenced. At levels of 5% and 10%, both empirical levels are very far away from the theoretical levels.

With this simple model we have seen that the robust score test clearly outperform the classical one in that it is resistant to small levels of data contamination. When contaminating more than one mean, the results of the two non-robust test statistics worsen. For example, with 5% of contamination and two means contaminated, the F-test reject 39.5 percent of the time the null hypothesis at the 1% level.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.8%	4.8%	10.3%
robust scores	0.8%	4.9%	9.8%
F-test	0.8%	5.4%	10.6%

Table 7.1: Proportion of times H_0 is rejected at different levels with $n = 100$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.2%	4.3%	10.3%
robust scores	0.2%	4.3%	10.4%
F-test	1%	5.2%	9.4%

Table 7.2: Proportion of times H_0 is rejected at different levels with $n = 20$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0%	2.2%	9.9%
robust scores	0%	2.1%	9.9%
F-test	1.1%	5%	9.9%

Table 7.3: Proportion of times H_0 is rejected at different levels with $n = 10$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	3.1%	14%	22.4%
robust scores	1.2%	5.2%	10.9%
F-test	1.9%	9.4%	17.4%

Table 7.4: Proportion of times H_0 is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = 48$.

$\epsilon = 5\%$	actual levels		
	1%	5%	10%
classical scores	6.8%	47%	54.9%
robust scores	1.5%	5.1%	11.1%
F-test	3.3%	37.9%	49.5%

Table 7.5: Proportion of times H_0 is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = 48$.

7.2 Two factors within-subject ANOVA model

The second model we propose to study is a model that implies two fixed factors (see 2.24). This kind of model is more interesting than the one factor within-subject ANOVA model since it allows us to test for more than one factor and to test for the interaction. We choose each factor λ and γ to have 3 levels. Thus the vector $\boldsymbol{\alpha}$ is of the form

$$\boldsymbol{\alpha} = (\mu, \lambda_1, \lambda_2, \gamma_1, \gamma_2, (\lambda\gamma)_{11}, (\lambda\gamma)_{12}, (\lambda\gamma)_{21}, (\lambda\gamma)_{22})^T$$

As before, we test using the classical and robust score type tests and the F-test. The form of F-test depends on the hypothesis we want to test. For

$$\begin{aligned} H_0 &: \lambda_1, \lambda_2 = 0, \mu, \gamma_1, \gamma_2, (\lambda\gamma)_{11}, (\lambda\gamma)_{12}, (\lambda\gamma)_{21}, (\lambda\gamma)_{22} \text{ unspecified} \\ H_1 &: \lambda_1, \lambda_2 \neq 0, \mu\ \gamma_1, \gamma_2, (\lambda\gamma)_{11}, (\lambda\gamma)_{12}, (\lambda\gamma)_{21}, (\lambda\gamma)_{22} \text{ unspecified} \end{aligned} \tag{7.1}$$

the F-statistic is

$$\frac{n\sum \hat{\lambda}_j^2/(l-1) + g\hat{\sigma}^2_{\lambda s}}{\hat{\sigma}^2_\varepsilon}$$

for

$$\begin{aligned} H_0 &: \gamma_1, \gamma_2 = 0, \mu, \lambda_1, \lambda_2, (\lambda\gamma)_{11}, (\lambda\gamma)_{12}, (\lambda\gamma)_{21}, (\lambda\gamma)_{22} \text{ unspecified} \\ H_1 &: \gamma_1, \gamma_2 \neq 0, \mu\ \lambda_1, \lambda_2, (\lambda\gamma)_{11}, (\lambda\gamma)_{12}, (\lambda\gamma)_{21}, (\lambda\gamma)_{22} \text{ unspecified} \end{aligned} \tag{7.2}$$

the F-statistic is

$$\frac{n\sum \hat{\gamma}_j^2/(g-1) + l\hat{\sigma}^2_{\gamma s}}{\hat{\sigma}^2_\varepsilon}$$

and for

$$
\begin{aligned}
H_0 &: (\lambda\gamma)_{11}, (\lambda\gamma)_{12}, (\lambda\gamma)_{21}, (\lambda\gamma)_{22} = 0,\ \mu, \lambda_1, \lambda_2, \gamma_1, \gamma_2 \text{ unspecified} \\
H_1 &: (\lambda\gamma)_{11}, (\lambda\gamma)_{12}, (\lambda\gamma)_{21}, (\lambda\gamma)_{22} \neq 0,\ \mu\ \lambda_1, \lambda_2, \gamma_1, \gamma_2 \text{ unspecified}
\end{aligned} \tag{7.3}
$$

the F-statistic is

$$\frac{n\sum(\hat{\lambda}\hat{\gamma})_j^2/(l-1)(g-1)}{\hat{\sigma}_\varepsilon^2}$$

in which the estimates are the $REML$ estimates. The procedure of simulation is the same as the one with the one factor within-subject ANOVA model. We generated 10000 samples under the null hypothesis and recorded the proportion of times the null hypothesis was rejected at different levels. We considered samples of size $n = 100$ and $n = 20$. The parameters for the multivariate normal distribution are $\mu = 24$, $\lambda_j = \gamma_k = (\lambda\gamma)_{jk} = 0\ \forall j = 1, 3, k = 1, 3$, $\sigma_s^2 = 5$, $\sigma_{\lambda s}^2 = 3$, $\sigma_{\gamma s}^2 = 4$ and $\sigma_\varepsilon^2 = 2$. We first present the results without contamination for the three hypotheses. Then in a second part we will study the influence of data contamination on the level of the tests.

For the first hypothesis (see 7.1) and without contamination all three tests behave nicely (see Table 7.6) when the sample size is relatively large ($n = 100$). For smaller data sets ($n = 20$), the results for the score type tests are slightly influenced by the sample size, especially at a low level (see Table 7.7). Since there are more parameters to estimate than in the one factor within-subject ANOVA model, the results for the score tests are slightly worse.

Table 7.6: Proportion of times H_0 (7.1) is rejected at different levels with $n = 100$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.9%	4.8%	10.2%
robust scores	0.9%	4.9%	9.8%
F-test	1.1%	5%	10.1%

However with $n = 10$ both score tests are clearly influenced by the sample size, especially at 1% and 5%. We tried different values for n between 10 and 20 and we found that for $n = 15$

the score tests remain acceptable at a level of 5%. Results are shown in Table 7.8. For the two remaining hypotheses we only present the results for $n = 100$. With $n = 20$ and 15 we found similar results as with hypothesis (7.1). Results for the second hypothesis (7.2) are presented in Table 7.9. They are very similar as the ones from the first hypothesis. When n is sufficiently large the χ^2_q distribution is a good approximation for the score type tests but not really when n gets smaller. Table 7.10 presents the results for the third hypothesis (7.3). The conclusions remain the same.

We now study the consequences of contaminated data on the tests. As we will see, the results of the tests differ under different contamination patterns. Given a certain pattern of contamination, the different tests will be more or less influenced. To illustrate this, we contaminated the models with respectively $\mu + \lambda_1 = 48$, $\mu + \lambda_1 = \mu + \lambda_2 = 48$ and $\mu + \lambda_1 = \mu + \gamma_1 = 48$. Like before the percentages of contamination are 2%, 5% and 10%. Only the results with 2% are presented here. Higher percentages do not change the results (for all three tests) but only deepen the difference between the robust and classical tests.

Tables 7.11- 7.13 present the results with $\mu+\lambda_1 = 48$. As expected, for null hypothesis (7.1) the classical tests are clearly influenced by this kind of data contamination. The robust score test remains stable at all theoretical levels whereas the two non-robust tests are not. On the other hand, for hypotheses (7.2) and (7.3) no test is influenced by this type of data contamination. With $\mu+\lambda_1 = \mu+\lambda_2 = 48$ as model contamination, things are slightly different (see Tables 7.14-7.16). Once again the robust score test is not influenced by the presence of data contamination. The novelty is that for hypothesis (7.2), the classical tests are also influenced and in the same proportion as hypothesis (7.1). On the other hand, the classical tests for hypothesis (7.3) are not influenced. For contamination type $\mu+\lambda_1 = \mu+\gamma_1 = 48$, the results are resumed in Tables 7.17 through 7.19. As with the two other contamination models, the robust score test is not influenced. However, for the three hypotheses (7.1), (7.2) and (7.3), the classical score test or the F-test are influenced. The difference with the other type of data contamination is that with this kind of model contamination, classical tests are also influenced for hypothesis (7.3).

Table 7.7: Proportion of times H_0 (7.1) is rejected at different levels with $n = 20$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.2%	3.9%	10.1%
robust scores	0.4%	4.1%	9.9%
F-test	1.1%	5.1%	9.5%

Table 7.8: Proportion of times H_0 (7.1) is rejected at different levels with $n = 15$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.9%	4.1%	9.9%
robust scores	0.9%	5.1%	9.8%
F-test	1%	5%	10%

Table 7.9: Proportion of times H_0 (7.2) is rejected at different levels with $n = 100$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.9%	5.1%	9.9%
robust scores	0.9%	5.1%	9.8%
F-test	1%	5%	10%

Table 7.10: Proportion of times H_0 is (7.3) rejected at different levels with $n = 100$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.9%	5.1%	9.9%
robust scores	0.9%	5.1%	9.8%
F-test	1%	5%	10%

Table 7.11: Proportion of times H_0 (7.1) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	3.1%	11.7%	22.9%
robust scores	1.1%	4.9%	9.9%
F-test	2.1%	9.2%	19%

Table 7.12: Proportion of times H_0 (7.2) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	0.9%	4.7%	10.2%
robust scores	1%	5%	9.8%
F-test	1.2%	5%	10.4%

Table 7.13: Proportion of times H_0 (7.3) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	0.8%	4.8%	9.8%
robust scores	0.8%	4.9%	10.1%
F-test	1.1%	5.1%	10%

Table 7.14: Proportion of times H_0 (7.1) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = \mu + \lambda_2 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	8.1%	20.4%	30.3%
robust scores	0.9%	5.1%	9.8%
F-test	6.5%	18.1%	28.2%

Table 7.15: Proportion of times H_0 (7.2) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = \mu + \lambda_2 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	0.9%	21%	31.1%
robust scores	0.8%	5.2%	9.9%
F-test	6.3%	17.4%	27.5%

Table 7.16: Proportion of times H_0 (7.3) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = \mu + \lambda_2 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	0.9%	4.8%	9.9%
robust scores	1%	4.9%	10%
F-test	1.1%	4.9%	9.9%

Table 7.17: Proportion of times H_0 (7.1) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = \mu + \gamma_1 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	3.1%	15%	24.9%
robust scores	0.9%	5.1%	10.3%
F-test	2.8%	14.3%	21.8%

Table 7.18: Proportion of times H_0 (7.2) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = \mu + \gamma_2 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	4,7%	18.2%	24.9%
robust scores	0.8%	5.2%	9.8%
F-test	2.4%	13.4%	23.9%

Table 7.19: Proportion of times H_0 (7.3) is rejected at different levels with $n = 100$ and contamination $\mu + \lambda_1 = \mu + \gamma_2 = 48$.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	5.1%	19.8%	27.4%
robust scores	0.9%	5.1%	9.4%
F-test	4.2%	13.1%	25.8%

The comments we can make about robust testing in the two factors within-subject ANOVA model are twofold. First, as for the one factor within-subject ANOVA model, the score type tests (both the classical and the robust) are quite sensitive to the size of the data sets. One has to be cautious about the fact that as the complexity of the model rises the number of parameters to estimate increases and one needs a suitable size to get a good approximation of the score test distribution. Second, when we concentrate only on the robustness aspect of the tests, the robust score type test perform well under any kind of contamination. This is not the case for the two other tests.

7.3 Hierarchical model

As for the simulation study for the $CTBS$, the model we consider is a model with four random factors. Recall that the parametrization of this model is given by

$$\begin{aligned} \boldsymbol{\mu}_i &= \mathbf{e}_8 \left(\mu + \lambda J_i(j)\right) = \mathbf{e}_8 \otimes (1, J_i(j)) \, (\mu, \lambda)^T = \mathbf{x}_i \boldsymbol{\alpha} \\ \boldsymbol{\Sigma} &= \sigma^2_\gamma \mathbf{J}_8 + \sigma^2_\lambda \mathbf{I}_2 \otimes \mathbf{J}_4 + \sigma^2_\delta \mathbf{I}_4 \otimes \mathbf{J}_2 + \sigma^2_\varepsilon \mathbf{I}_8 \end{aligned}$$

with $\mu = 15$, $\lambda = 0$, $\sigma^2_\gamma = 2$, $\sigma^2_\delta = 2$, $\sigma^2_\xi = 1$ and $\sigma^2_\varepsilon = 3$. In terms of fixed effect, the only possible test we can do for this type of model is given by

$$\begin{aligned} H_0 &: \lambda = 0, \mu = \text{unspecified} \\ H_1 &: \lambda \neq 0, \mu = \text{unspecified} \end{aligned} \tag{7.4}$$

It is worth mentioning that hypothesis (7.4) will always have this form regardless of the model complexity. Indeed, for hierarchical models, there will always be only one fixed effect, the others being nested in it and considered random. It is clear that one can use a simple t-test for hypothesis (7.4). The idea here is also to propose the robust score type test as a robust alternative to the t-test. The simulation parameters are the same as for the two previous models, that is $n = 10$, 20 and 100 and $\epsilon = 0\%$, 2%, 5% and 10%. As before, we first investigate the size's effect without contamination. The results are very similar to those for the one factor within-subject

ANOVA model. With small samples, the score type tests (both classical and robust) do not behave very well. The results are presented in Tables 7.20-7.22.

Table 7.20: Proportion of times H_0 (7.4) is rejected at different levels with $n = 100$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.9%	4.9%	9.9%
robust scores	0.8%	5.1%	10%
F-test	1%	5%	9.9%

Table 7.21: Proportion of times H_0 (7.4) is rejected at different levels with $n = 100$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0.3%	4%	9.9%
robust scores	0.4%	4.1%	10.1%
F-test	0.9%	4.7%	9.8%

Table 7.22: Proportion of times H_0 (7.4) is rejected at different levels with $n = 100$.

$\epsilon = 0\%$	actual levels		
	1%	5%	10%
classical scores	0%	2%	9.9%
robust scores	0%	2.1%	9.8%
F-test	1%	4.8%	9.8%

As expected, when the data are contaminated with model contamination $\lambda = 15$, the robust score type test is not influenced whereas the classical score type test and the F-test are. Table 7.23 presents the results with 2% of contamination. Raising the percentage of contamination until 10% doesn't change the results for the robust test.

The simulations performed on this model do not bring any new informations. The robust score type test is robust to small model deviations but n must be sufficiently large for the χ^2_q distribution to be accurate.

$\epsilon = 2\%$	actual levels		
	1%	5%	10%
classical scores	2.8%	13.1%	21.8%
robust scores	0.9%	4.8%	10.1%
F-test	1.9%	8.9%	18.1%

Table 7.23: Proportion of times H_0 (7.4) is rejected at different levels with $n = 100$ and contamination $\lambda = 15$.

In fact with this model the interest lies elsewhere. We know that for any hierarchical model, there always will be only one fixed effect. However, the number of nested random effects can vary considerably and the possibility to test these effects is more than legitime since they are often the parameters of interest. We have already said that we don't consider applying testing procedures to the random effects as a good solution. In conclusion, this legitimates the need to pursue other ways of research to propose a mean to test for the random effects. In the last part of this thesis, we suggest that this can be done by applying robust model selection techniques.

Chapter 8

Data analysis

In this Chapter we analyze 4 data sets in order to study the impact on the analysis of the use of robust methods when compared to classical ones. For each data set, we perform a graphical data analysis to detect if there are any outliers or influential observations. In each case, we compute the MLE and the $CTBS$ and corresponding standard errors for the fixed effects. The $REML$ gave similar results which are not presented here. We also could provide standard errors for the variance component estimates, but since the samples are relatively small, the asymptotic normal theory is not suitable in these cases. We compute the robust score type test and compare it to the F-test. As a diagnostic tool, we also provide a scatter plot of the Mahalanobis distances d_i estimated using respectively the MLE and the $CTBS$. These graphics will reveal which observations are considered as outliers and/or influencial observations.

8.1 Electrode resistance data

Berry (1987) analyzed the data from an experiment in which five types of electrodes were applied to the arms of 16 subjects and their resistance measured. The experiment was designed to see whether all five electrode types performed similarly. The model is the one factor within-subject ANOVA model (2.15) with $l = 5$. If we look at the interaction plot (see Figure 8-1) of the 16 readings we notice that two measures (electrodes of type 2 and 3) taken on subject 15 are much larger than the others. The experimenter decided that the reason for the two large readings on a subject (subject 15) was the excessive amount of hair on the subject's arm (see Berry, 1987) .

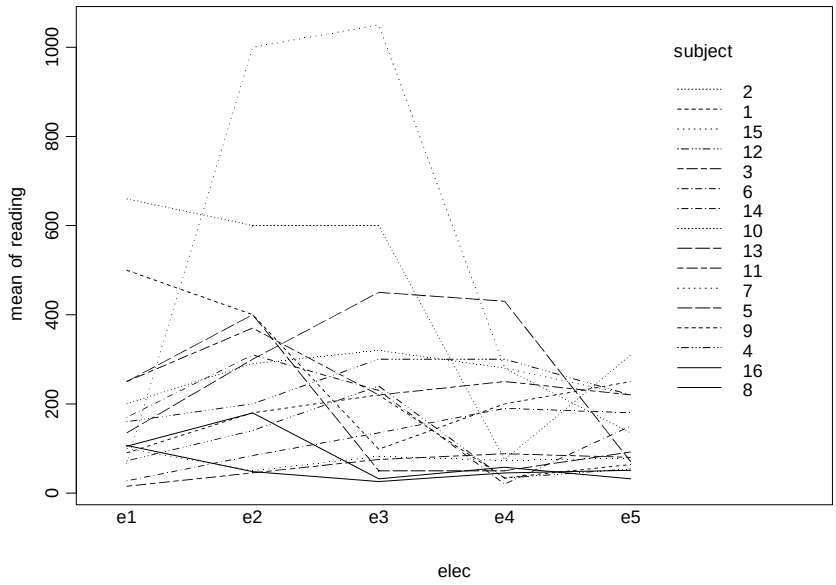

Figure 8-1: Figure 16 : Interaction plot for the skin resistance data

Table 8.1 presents the MLE and $CTBS$ estimates and standard errors for the raw data (divided by 100). The mean vector estimate is $\hat{\boldsymbol{\mu}} = [1.81, 2.87, 2.58, 1.50, 1.37]$ for the MLE and $\hat{\boldsymbol{\mu}} = [1.55, 2.14, 1.77, 1.46, 1.19]$ for the $CTBS$. One notices that the mean resistance estimate is larger for the MLE in general and in particular for electrodes of type 2 and 3. This difference is certainly due to the two extreme measurements observed in the interaction plot. The latter also probably explain the large difference in the residual error variance estimate $\hat{\sigma}^2_{\varepsilon}$.

In Figure 8-2 is given the scatter plot of the Mahalanobis distances computed with the MLE and the $CTBS$ respectively. The horizontal and vertical dotted lines correspond to the 0.975 quantile on the χ^2_2 distribution, to detect outlying observations. When both the MLE and $CTBS$ detect an extreme observation, it is considered as an outlier and when only the $CTBS$ detects an extreme observation, it is also considered as influencial. On the interaction plot 8.1, observation number 2 corresponds the the second highest response. Here the $CTBS$ estimator detects observations 15 as an extreme outlier and observation 2 as an influential outlier.

We turn now to testing the following multivariate hypothesis of equality of means resistances given by

$$\begin{aligned} H_0 &: \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0 \quad \mu \text{ unspecified} \\ H_1 &: \lambda_1 \neq \lambda_2 \neq \lambda_3 \neq \lambda_4 \neq 0 \quad \mu \text{ unspecified} \end{aligned} \tag{8.1}$$

The classical F-ratio test statistic is 3.1455. When compared to an F distribution with degrees of freedom 4 and 60, we find a p-value of 0.020 so that the test is significant. We can conclude that there is a difference between the 5 electrode types. The robust test statistics R^2 is 7.396. When compared to a χ^2 with 4 degrees of freedom, we find a p-value of 0.116. The test is hence not significant. Observations 15 and 2 seem to have an influence on the MLE's estimates and consequently on the F-test.

To limit the influence of potential outlying observations, Berry (1987) actually proposed to use

	MLE		*CTBS*	
Parameter	Estimate	S.E.	Estimate	S.E.
μ	2.03	0.3306	1.628	0.2638
λ_1	−0.21	0.3239	−0.068	0.2553
λ_2	0.84	0.3239	0.512	0.2553
λ_3	0.55	0.3239	0.142	0.2553
λ_4	−0.53	0.3239	−0.158	0.2553
σ_s^2	1.329		1.087	
σ_ε^2	2.098		0.711	

Table 8.1: Estimates and standard errors for the MLE and the $CTBS$ (bold values are for significant parameters at the 5% level) for the skin resistance data.

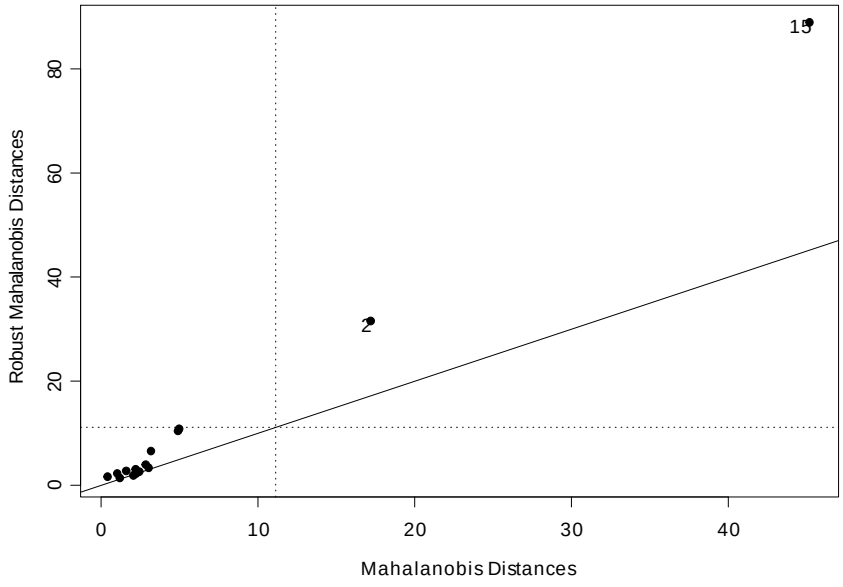

Figure 8-2: MLE and $CTBS$ Mahalanobis distances for the skin resistance data

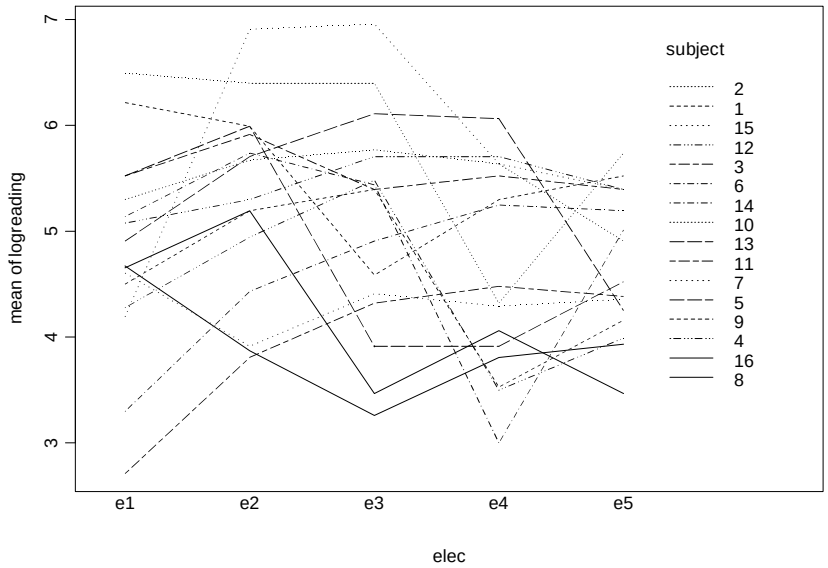

Figure 8-3: Interaction plot for the skin resistance data (with $\log(y + 32)$ transformation).

a $\log(y + c)$ ($c = 32$) transformation of the data. An interaction plot of the transformed data is presented in Figure 8-3. Graphically, the log transformation has the effect of minimizing the potential outliers. In the scatter plot of the Mahalanobis distances (see Figure 8-4) the $CTBS$ detects observation no 15 as an extreme outlier but not observation 2. We analyzed the transformed data and the results are presented in Table 8.2. The overall mean, the contrasts and the variance components are found smaller when using the $CTBS$. We obtained an F-ratio test statistic of 2.87 corresponding to a p-value of 0.030 and a robust score type test statistic R^2 of 6.94, corresponding to a p-value of 0.139. In conclusion, reading 15 seems to have a considerable influence. Berry (1987) analyzed this dataset with subject 15 deleted. In this case, the F-test is significant when performed on the original data (p-value of 0.044) and it is not significant when performed on the transformed data (p-value of 0.1).

	MLE		*CTBS*	
Parameter	Estimate	S.E.	Estimate	S.E.
μ	**5.1925**	0.1269	**5.1415**	0.1259
λ_1	−0.0719	0.1166	0.0148	0.1136
λ_2	**0.3156**	0.1166	**0.2712**	0.1136
λ_3	0.1535	0.1166	0.0915	0.1136
λ_4	−0.2176	0.1166	−0.2178	0.1136
σ_s^2	0.2033		0.1683	
σ_ε^2	0.2719		0.2150	

Table 8.2: Estimates and standard errors for the MLE and the $CTBS$ (bold values are for significant parameters at the 5% level) for the skin resistance data (with $\log(y + 32)$ transformation)

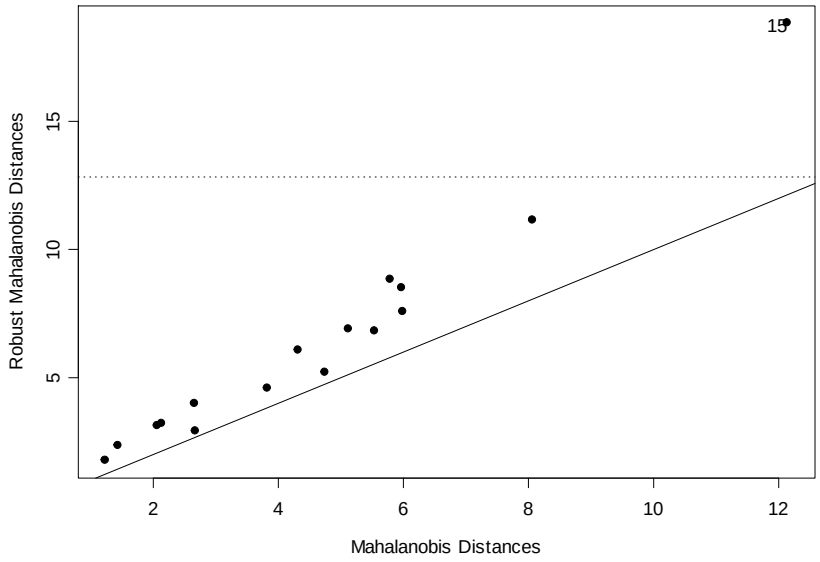

Figure 8-4: MLE and $CTBS$ Mahalanobis distances for the skin resistance data (with $\log(y + 32)$ transformation).

8.2 Semantic priming data

The study of semantic and associative priming in picture naming is well known in psychology (see e.g. Alario, Segui and Ferrand, 2000, Holcomb and McPherson, 1994). The data we have come from an experiment in which 16 young subjects had to decide as quickly as possible if a target (object's drawing), which appeared after a prime (action of a pantomime), was a real object or not. There were 3 different real objects of interest (a hammer, a saw, a screwdriver) and 3 different types of item (related, neutral, unrelated). For each combination of real object and type of item, five measures (time to decide if the object was real or not) were taken, of which the first one (trial) and the errors (wrong object decision) were discarded , and the means of the remaining were taken as the response variable. The underlying hypothesis is that the reaction time is shorter when there is a link between the priming and the object (i.e. in the related item). The data have been collected at the university of Geneva (see Moy and Mounoud, 2003). The model used to analyze these data is given in (2.24) with $\lambda_j, j = 1, 2, 3$ the object effect and $\gamma_k, k = 1, 2, 3$ the item type. Interaction plots for each factor and the interaction are given in Figures 8-5, 8-6 and 8-7. Figure 8-5 suggests that observations 3 and 16 seem higher than the others. It means that the subject variance for these two observations will probably be also higher. The same comments can be made about Figure 8-6. Again, measure for observation 3 and 16 are higher that the others. Finally, Figure 8-7 suggests that there is an interaction between the two factors.

Table 8.3 gives the estimates for the MLE and the $CTBS$ and the standard errors for the fixed effects (bold values are for significant fixed effects at the 5% level). In Figure 8-8 is given the scatter plot of the Mahalanobis distances computed with the MLE and the $CTBS$ respectively. One can see that the $CTBS$ and the MLE detect one outlier and the $CTBS$ detects two influencial observations. The influence on the estimates is quite substantial. Indeed, the mean reaction times are generally smaller and the error and subject variances are also a lot smaller with the robust approach. The consequence is that the effect of the object is not found significant with a classical approach, whereas it is found significant with a robust approach. In this case, because of the presence of influencial outliers, the variance components are overestimated and bias inference on the contrasts. Note that with both analyses the effect of the type

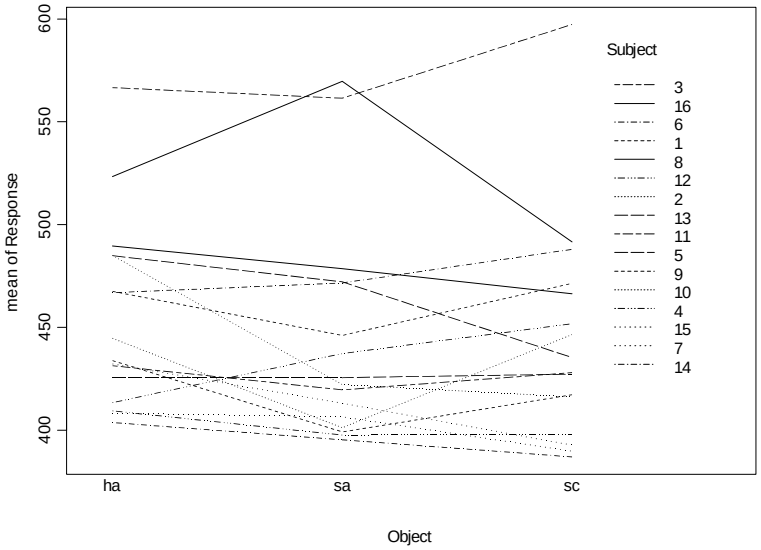

Figure 8-5: Interaction plot for the factor Object for the semantic priming data set

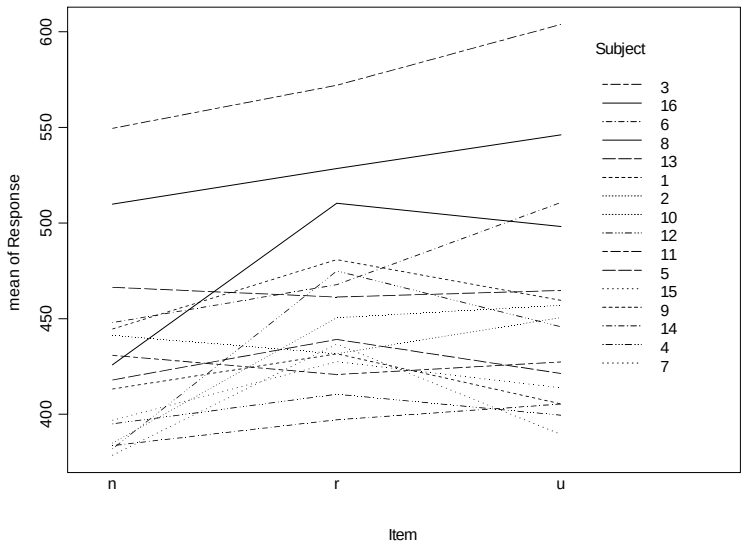

Figure 8-6: Interaction plot for the factor Item for the semantic priming data set

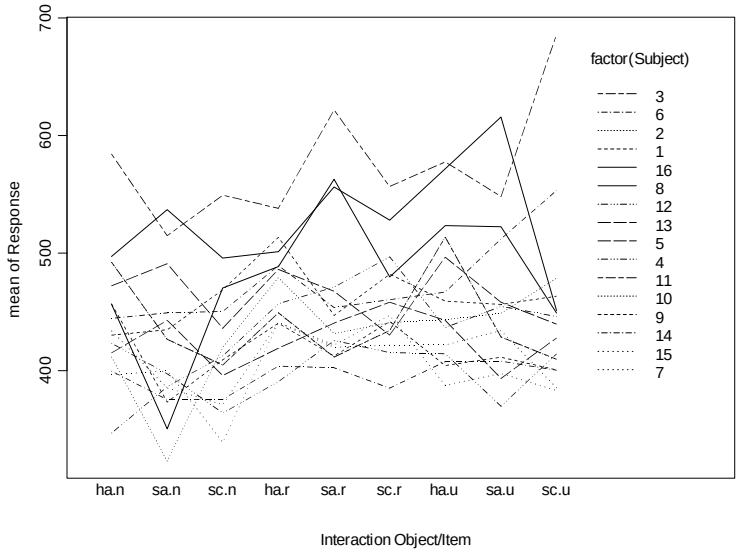

Figure 8-7: Interaction plot for the interaction for the semantic priming data set

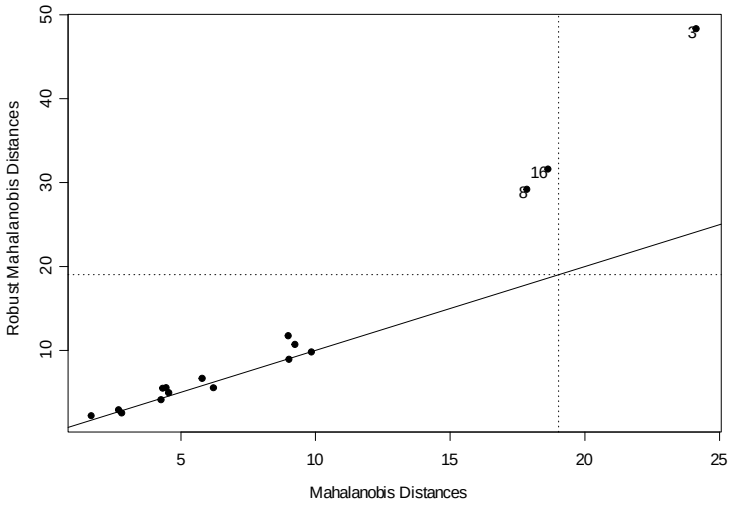

Figure 8-8: MLE and $CTBS$ Mahalanobis distances for the semantic priming data.

	MLE		$CTBS$	
Parameter	Estimate	S.E.	Estimate	S.E.
μ	**448**	12.2	**430**	7.6
λ_1	7	4.0	**9**	4.2
λ_2	-3	4.0	-6	4.2
γ_1	**-19**	4.0	**-16**	4.8
γ_2	**11**	4.0	**11**	4.8
$\lambda\gamma_{11}$	7	5.4	5	4.7
$\lambda\gamma_{12}$	-7	5.4	-2	4.7
$\lambda\gamma_{21}$	-7	5.4	0	4.7
$\lambda\gamma_{22}$	5	5.4	-2	4.7
σ^2_s	2222		649	
$\sigma^2_{\lambda s}$	43		158	
$\sigma^2_{\gamma s}$	49		261	
σ^2_ε	1035		733	

Table 8.3: Estimates and standard errors for the MLE and the $CTBS$ (bold values are for significant parameters at the 5% level) for the semantic priming data.

of item is significant.

We also tested for each factor and for the interaction, using the F-test and the robust score type test. Results are presented in Table 8.4

The classical F-test and the robust score type test give similar results for the three hypotheses. In this example, the presence of possible outlier doesn't seem to influence the results of the tests.

With this type of data, one can also consider a log transformation, although in this domain one usually prefers the original scale, mainly for interpretation reasons. Interaction plots for the transformed data are given in Figures 8-9, 8-10 and 8-7. Comments are the same as for the raw data. In that case, the estimates for the MLE and the $CTBS$ and the standard errors for the contrasts are given in table 8.5. The scatter plot of the Mahalanobis distances is given in Figure 8-12. One can see that there are 3 outliers detected only by the $CTBS$, but they do not seem to be influencial since the estimates are quite similar with the two analyses. The tests on the factors and the interaction for the classical and robust tests give the same result as with the raw data (not presented here).

Classical F-test				
	F-stat	df1	df2	p-val
Item	1.20	2	30	0.30
Object	1.63	2	30	0.19
Item:Object	0.48	4	60	0.74
Robust score type test				
	F-stat	df		p-val
Item	1.63	2		0.44
Object	3.32	2		0.19
Item:Object	1.70	4		0.79

Table 8.4: Results for the classical F-test and the robust score type test for the semantic priming data.

	MLE		*CTBS*	
Parameter	Estimate	S.E.	Estimate	S.E.
μ	**6.097**	0.0247	**6.060**	0.0167
λ_1	**0.020**	0.0084	**0.023**	0.0094
λ_2	−0.010	0.0084	−0.013	0.0094
γ_1	**−0.043**	0.0089	**−0.037**	0.0105
γ_2	**0.027**	0.0089	**0.027**	0.0105
$\lambda\gamma_{11}$	0.017	0.0112	0.013	0.0105
$\lambda\gamma_{12}$	−0.013	0.0112	−0.010	0.0105
$\lambda\gamma_{21}$	−0.013	0.0112	0.000	0.0105
$\lambda\gamma_{22}$	0.007	0.0112	−0.003	0.0105
σ^2_s	0.0091		0.0031	
$\sigma^2_{\lambda s}$	0.0002		0.0007	
$\sigma^2_{\gamma s}$	0.0004		0.0012	
σ^2_ε	0.0045		0.0037	

Table 8.5: Estimates and standard errors for the MLE and the $CTBS$ (bold values are for significant parameters at the 5% level) for the semantic priming data (log transformed)

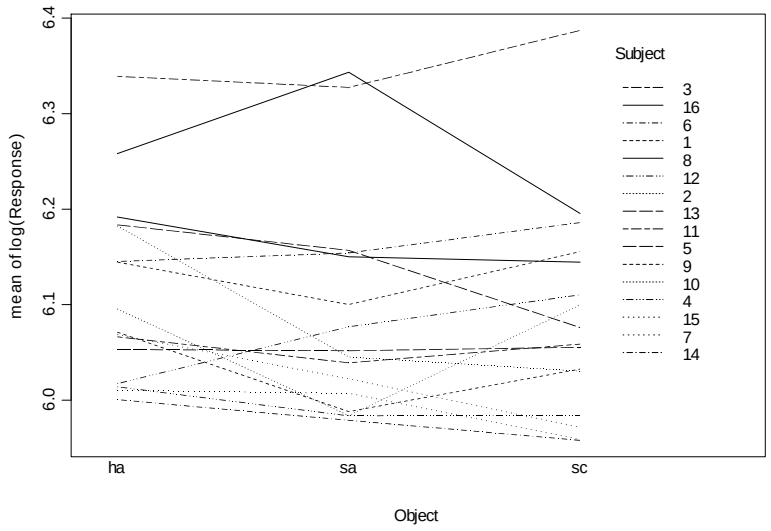

Figure 8-9: Interaction plot for the factor Object for the semantic priming data (log-transformed).

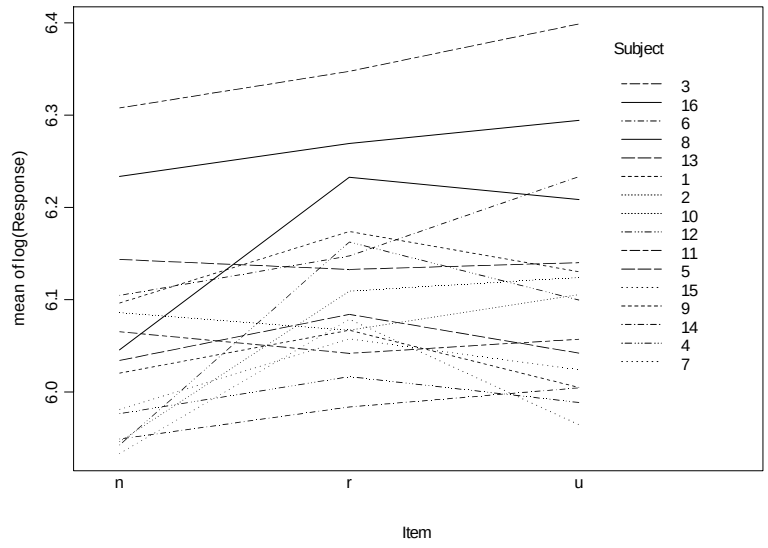

Figure 8-10: Interaction plot for the factor Item for the semantic priming data (log-transformed).

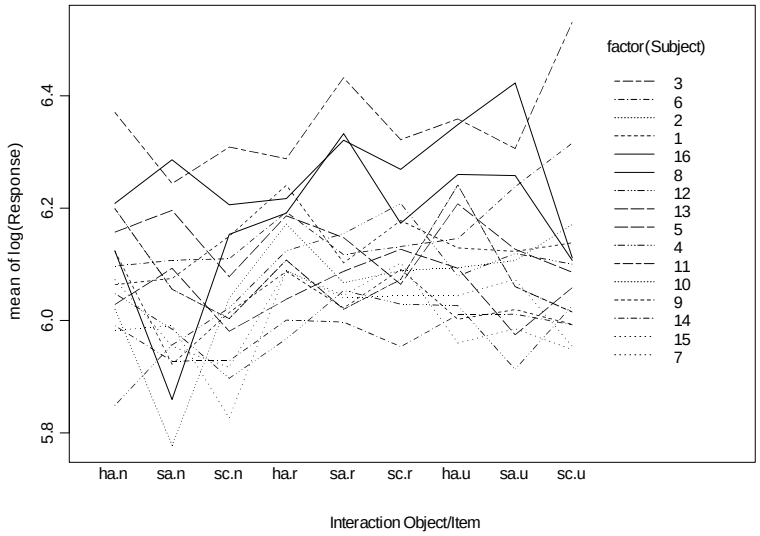

Figure 8-11: Interaction plot for the interaction for the semantic priming data (log-transformed).

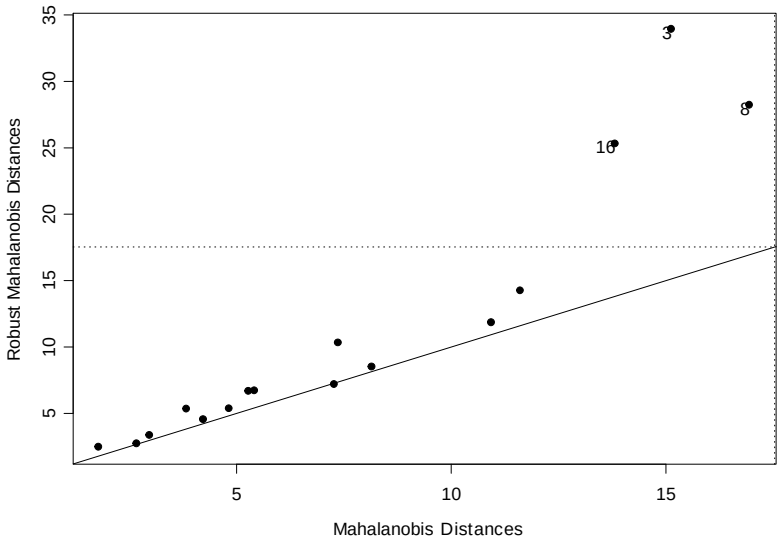

Figure 8-12: MLE and $CTBS$ Mahalanobis distances for the semantic priming data (log transformed data).

8.3 Metallic oxide data

This data set originate from a sampling study designed to explore the effects of process and measurement variation on properties of lots of metallic oxide (Bennett, 1954). Two samples were drawn from each lot. Duplicate analyses were then performed by each of two chemists, with a pair of chemists randomly selected for each sample. This model corresponds to a hierarchical model similar to the one presented in (2.25). Type is considered fixed with type 1 ($j = 1$) measured in $n_1 = 18$ lots and type 2 ($j = 2$) measured in $n_2 = 13$ other lots. Lot, sample and chemist are considered random. Thus the parameter to estimate are the two means corresponding to type 1 and type 2 and the variances corresponding respectively to lots, samples, chemists and residual. This data set has already been considered by Fellner (1986) and analyzed by Richardson and Welsh (1995). An important assumption here is that the two types of metallic oxide have the same covariance matrix (i.e. the same variance components) though possibly different means. Table 8.6 presents the estimates and standard errors for the MLE and the $CTBS$. The $CTBS$ gives higher estimates for the mean components (i.e. μ and $\mu + \lambda$) than the MLE. For the variance components, the $CTBS$ provides generally smaller estimates than the MLE. The consequence is that the standard errors for the fixed effects are smaller with the $CTBS$. However, the effect of the type of metallic oxide is found significant with both analyses. In Figure 8-13 is given the scatter plot of the Mahalanobis distances computed with the MLE and the $CTBS$ respectively. Both the MLE and $CTBS$ detect observations 24 and 30 (and maybe 17) as outlying, but the $CTBS$ also detects observation number 25 as influencial. The difference between the MLE and $CTBS$ estimates is then certainly due to this last influencial observation. It should be stressed that Richardson and Welsh (1995) found not the same, but similar results with their robust method.

Method	Estimate (S.E.)					
	μ	λ	σ^2_{lot}	σ^2_{sample}	$\sigma^2_{\text{chemist}}$	σ^2_{ε}
MLE	3.85625	−0.79375	.565	.043	.032	.043
	(0.1826328)	(0.2820253)				
$CTBS$	3.9318973	−0.4017356	.097	.012	.040	.036
	(0.09964693)	(0.15387677)				

Table 8.6: Estimates and standard errors for the MLE and the $CTBS$ (bold values are for significant parameters at the 5% level) for the metallic oxide data

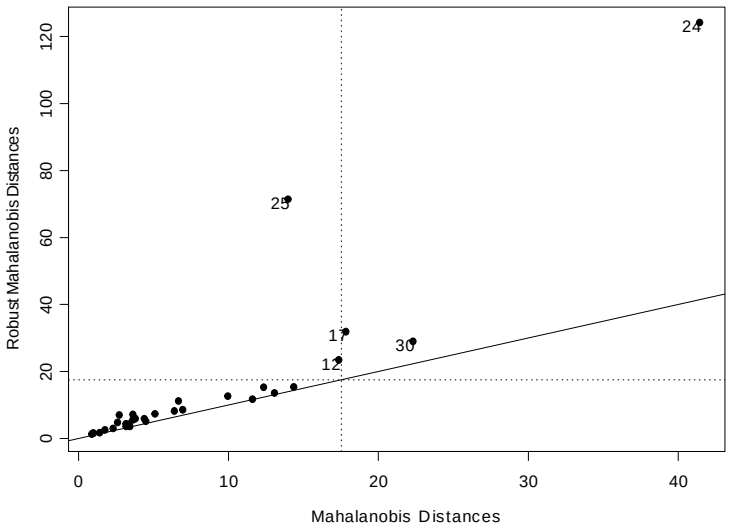

Figure 8-13: MLE and $CTBS$ Mahalanobis distances for the metallic oxide data

8.4 Orthodontic data

The data come from an othodontic study on 16 boys and 11 girls between the ages of 8 and 14 years and were originally reported by Pothoff and Roy (1964). Pinheiro, Liu and Wu (2001) also analyzed these data using a robust estimator based on the multivariate t distribution (see also Welsh and Richardson, 1997 and Lang, Little and Taylor, 1989). The response variable is the distance (in millimeters) between the pituitary and the pterygomaxillary fissure, which was measured at 8, 10, 12 and 14 years for each boy and girl. Figure 8-14 presents the scatter plots of the data with individual least squares fits for the simple linear regression model. It reveals that slope for subject M13 is larger that the other estimated slopes. Overall, it seems that the responses for the boys vary more than those for the girls. Moreover, the plot suggests that two observations on subject M09 are outliers. These potential outliers are also detected in Figure 8-15 that present the residuals plots by gender.

Because both intercept and slope seem to be different between subjects and the within-subject variation is larger among boys than girls, Pinheiro et al. (2001) suggest that a potential model for these data is

$$y_{ijt} = \beta_0 + \beta_1 t + \left(\beta_{0g} + \beta_{1g} t\right) J_i(j) + \gamma_{0i} + \gamma_{1i} t + \varepsilon_{ijt}$$

with y_{ijt} the response for the ith subject ($i = 1, \ldots, 27$) of sex j ($j = 1$ for boys and $j = 2$ for girls) at age $t = 8.10, 12, 14$,

$$J_i(j) = \begin{cases} 0 & j = 1 \\ 1 & j = 2 \end{cases}$$

a dummy variable for sex. There $\boldsymbol{\alpha} = \left(\beta_0, \beta_1, \beta_{0g}, \beta_{1g}\right)^T$ are the fixed effects and $\gamma_{0i}, \gamma_{1i}, \varepsilon_{ijt}$ the random effects with zero mean and respective variances of $\sigma^2_{\gamma_0}, \sigma^2_{\gamma_1}, \sigma^2_{\varepsilon}$. This model is actually a random slope and intercept model. The multivariate formulation is given by $\boldsymbol{\mu}_i = \mathbf{x}_i \boldsymbol{\alpha}$ with $\mathbf{x}_i = (\mathbf{e}_4, \mathbf{e}_4 J_i(j), \mathbf{t}, \mathbf{t} J_i(j))$ and $\mathbf{t} = (8, 10, 12, 14)^T$, and covariance $\boldsymbol{\Sigma} = \sigma^2_{\gamma_0} J_4 + \sigma^2_{\gamma_1} \mathbf{t}\mathbf{t}^T + \sigma^2_{\varepsilon} \mathbf{I}_4$. Table 8.7 presents the MLE and $CTBS$ estimates and standard errors for the parameters.

The estimates show that the slopes and the difference in slopes between boys and girls are

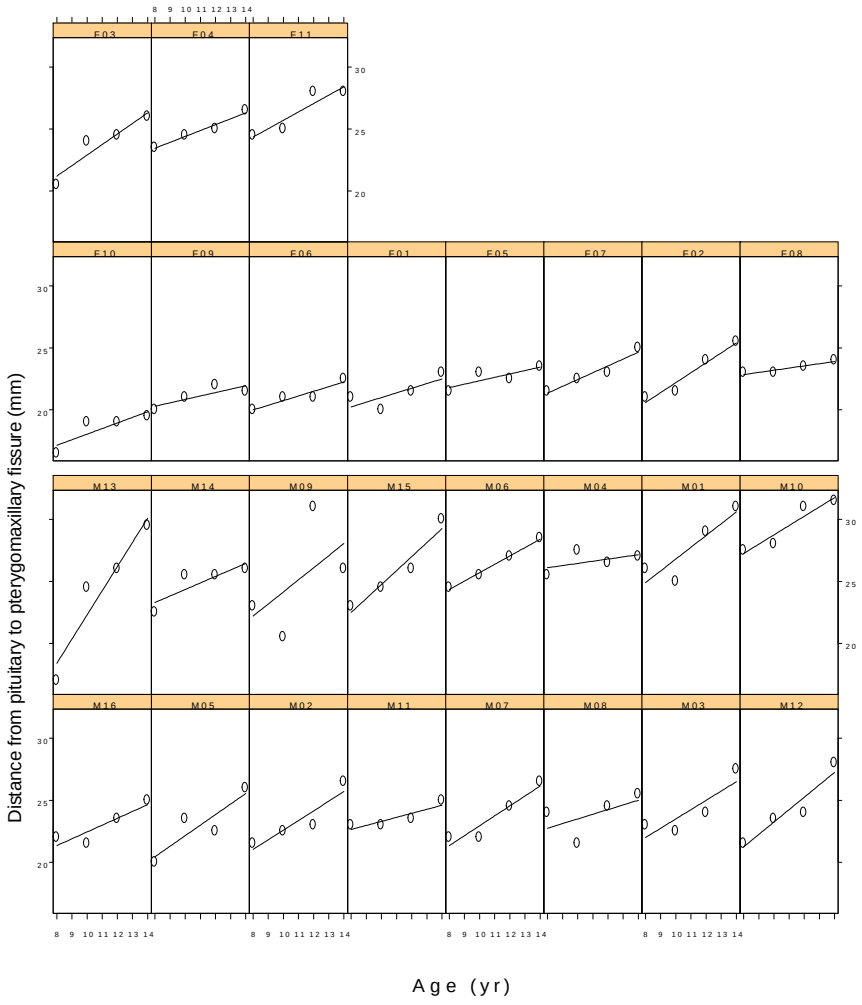

Figure 8-14: Orthodontic growth patterns in 16 boys (M) and 11 girls (F). Lines represent the individual least square fits of the simple regression model.

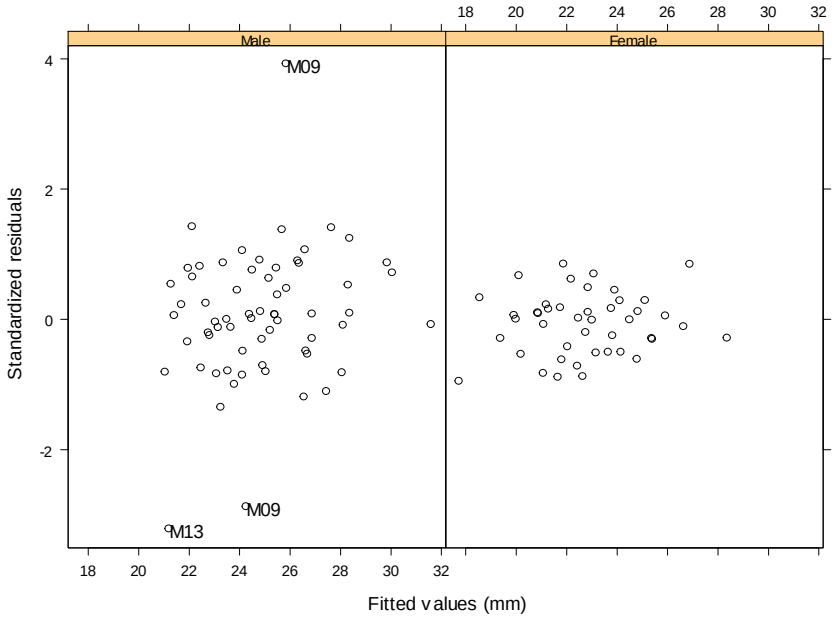

Figure 8-15: Residuals versus fitted values by gender corresponding to individual least square fits

smaller with a robust approach. With both analyses there is no significant difference in the intercept between boys and girls. The error variance with a robust approach is smaller, whereas the intercept and slope variances are larger. In the scatter plot of the Mahalanobis distances (Figure 8-16) the $CTBS$ and the MLE detect observations no 20 and 24 (corresponding to the 9th and 13th boy) as outlying. These observations are certainly the cause of the difference found in the estimates. It should be noted, that Pinheiro et al. (2001) also find the same outliers.

	MLE		$CTBS$	
Parameter	Estimate	S.E.	Estimate	S.E.
β_0	**16.34**	1.139	**17.24**	1.037
β_{0g}	1.03	1.480	0.14	1.347
β_1	**0.78**	0.096	**0.70**	0.086
β_{1g}	**−0.30**	0.125	**−0.22**	0.112
$\sigma^2_{\gamma_0}$	2.336		2.812	
$\sigma^2_{\gamma_1}$	0.007		0.012	
σ^2_{ε}	1.894		1.032	

Table 8.7: Estimates and standard errors for the MLE and the $CTBS$ (bold values are for significant parameters at the 5% level) for the orthodontic data

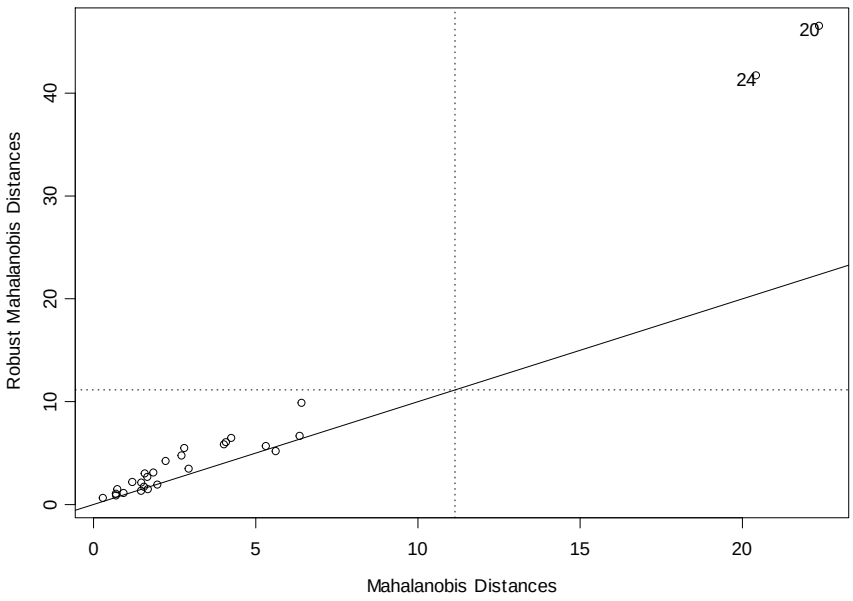

Figure 8-16: MLE and $CTBS$ Mahalanobis distances for the orthodontic data

Chapter 9

Extensions

In this Chapter, we briefly explore some other research fields that are of interest in the context of robust estimation and testing in the mixed models. This is by no means an exhaustive list but we consider that those domains are the more interesting and important ones.

The first question is how to deal with missing data or, as it is often referred in the mixed model literature, incomplete design. In many cases the classical analytical techniques cannot be used, either because some observations are missing or because the design is unbalanced for some other reason, for example, the presence of time varying covariates. There is a large literature both in ANOVA and in mixed linear models about how to handle the problem (see e.g. Jennrich and Schluchter, 1986 and Laird, Lang and Stram, 1987) and numerous specific solutions have been developed. Laird, Lange, and Stram (1987) consider the EM algorithm (Dempster, Laird and Rubin, 1977) for both maximum likelihood and restricted maximum likelihood estimation in a general repeated measures setting using a multivariate normal model with linear covariance structure. Because the EM algorithm is a very general iterative method for computing maximum likelihood estimates with incomplete data, it has often been used in this particular setting. For example Jennrich and Schluchter (1986) describe generalized EM algorithms for computing restricted and unrestricted maximum likelihood estimates and Newton-Raphson and Fisher scoring algorithms for computing maximum likelihood estimates. Our idea is to adapt the work of Cheng and Victoria-Feser (2002) and Copt and Victoria-Feser (2004) which propose an adaptation of the EM algorithm to the S-estimator defined by Rocke

(1996).

We considered in this work general testing procedures for the fixed effects but neglected testing for the variance components. A reasonable alternative would be to rely on model selection techniques. Some well-known classical model selection procedure are Mallow's C_p (Mallows, 1973) and Akaike Information Criterion (Akaike, 1973). Model selection is very important in statistics and is considered as a key component in any analysis. Classical model selection procedures are based on classical estimators and tests and it is obvious that when dealing with robust estimators, one wants to use robust model selection techniques, see, for example Ronchetti and Staudte (1994), Sommer and Huggins (1996) and Ronchetti (1997). For the robust Akaike criterion of Ronchetti (1997) we will most likely face the same problem encountered when we tried to implement the robust procedure of testing based on the likelihood ratio test. Another possibility could be to adapt the robust version of the Mallow's C_p (Ronchetti and Staudte, 1994). Another direction is the use of cross-validation and other resampling techniques as model selection procedures. Ronchetti, Field, and Blanchard (1997) propose to robustify a least squares selection procedure based on cross-validation proposed by Shao (1993).

For testing fixed effects, we have used score type tests that are valid asymptotically. However when working with small sample size, the χ^2_q distribution fails to approximate the true distribution of the test statistic yielding confidence levels far more conservative than they should be. An idea could be to obtain test for small samples based on saddlepoint approximation. Even if the saddlepoint approximation is derived from an asymptotic expansion, it is often very accurate in small sample. It is known that saddlepoint techniques are useful tools to derive very accurate approximation of densities and tail probability. More particularly Hampel (1973) described these techniques as small asymptotic approximation. Since their introduction in statistics (see Daniels 1954), they have been used successfully in a great variety of fields. Recent examples that are of interest here can be found in Ronchetti and Welsh (1994), Ronchetti, Field, and Blanchard (1997) and more particularly in Robinson, Ronchetti, and Young (2002). In this paper, they consider test statistics suggested by the saddlepoint approximation and show that this statistic is analogous to the likelihood ratio test in the parametric case.

The next extension concerns generalized linear models (see for example McCullagh and Nelder, 1989 or Breslow and Clayton, 1993). In these models the unobserved effects are modelled by the inclusion of random effect in the linear predictor of the generalized linear model. Generalized linear mixed models are a natural combination of both mixed linear models as presented in this work and generalized linear models. As such, they are of considerable interest in practice due to their great applicability. They enable the accommodation of nonnormality distributed responses and specification of a possibly non linear link between the mean of the response and the predictors. Unfortunately it is not possible to apply the procedure proposed here to the generalized linear context. In fact, the assumption of normality is a necessary condition for the $CTBS$ estimator and hence restrains its use to normal data.

In what follows, we present more in detail what could be the extensions for the $CTBS$ estimator to deal with missing data.

9.1 Incomplete Design

We have seen that the statistical literature contains several proposals for high breakdown estimators of the mean and covariance in multivariate data when it is suspected that the data contain outliers or extreme observations. When there are missing data, Little and Smith (1987), Cheng and Victoria-Feser (2002) and Copt and Victoria-Feser (2004) propose different solutions. The ER algorithm of Little and Smith (1987) is based on a combination of the EM algorithm for missing data and a robust estimation step based on an M-estimator. However, the ER algorithm as originally proposed suffers from a lack of robustness in some cases, especially in high dimensions. Cheng and Victoria-Feser (2002) propose two alternatives to avoid this problem. The first one is to provide the ER algorithm with a high breakdown estimator as starting point of the iterative procedure. In this case, the chosen high breakdown estimator is the MCD estimator, modified to handle missing data. The other solution is to base the estimation step of the ER algorithm on a high breakdown estimator. They consider an adaptation of the translated biweight S-estimator to the case of incomplete data. More recently Copt and Victoria-Feser

(2004) adapted the OGK estimator proposed by Maronna and Zamar (2002) to the case of missing data and used it as a starting point for an adapted S-estimator.

Besides the fact that researchers are often confronted with missing data in practice, the need for an estimator able to deal with missing data in the mixed model setting is important. Consider the two factors within-subject ANOVA model. If for any reason, a subject does not respond to one (or more) factor's combinations, the entire observation has to be deleted. For example, the original semantic priming data has 24 observations of which 9 had one or more missing values and had to be removed.

Technically speaking, let $\mathbf{y}_i = [\mathbf{y}_{[oi]}^T, \mathbf{y}_{[mi]}^T]^T$ so that a distinction is made between the observed (oi) and the missing (mi) data. We suppose that the data are missing at random (see Rubin, 1976), a sufficient condition for correct likelihood-based inferences. Most known estimators of mean and covariance with missing data fall in the class proposed by Cheng and Victoria-Feser (2002), i.e.

$$\frac{1}{n}\sum_{i=1}^{n} w_i^{\mu}(\boldsymbol{\mu} - \hat{\mathbf{y}}_i) = 0 \tag{9.1}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[w_i^{\delta}\boldsymbol{\Sigma} - w_i^{\eta}((\hat{\mathbf{y}}_i - \boldsymbol{\mu})(\hat{\mathbf{y}}_i - \boldsymbol{\mu})^T - \mathbf{C}_i)\right] = 0 \tag{9.2}$$

where

$$\begin{aligned}\hat{y}_{ij} &= \begin{cases} y_{ij} & \text{if } y_{ij} \text{ is observed} \\ E\left[y_{ij} | \, \mathbf{y}_{[oi]}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right] & \text{if } y_{ij} \text{ is missing} \end{cases} \\ &= \begin{cases} y_{ij} & \text{if } y_{ij} \text{ is observed} \\ \boldsymbol{\mu}_{[mi]} + \boldsymbol{\Sigma}_{[moi]}\boldsymbol{\Sigma}_{[ooi]}^{-1}(\mathbf{y}_{[oi]} - \boldsymbol{\mu}_{[oi]}) & \text{if } y_{ij} \text{ is missing} \end{cases}, \end{aligned} \tag{9.3}$$

and

$$C_{ijk} = \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ \text{cov}\left[y_{ij}, y_{ik} | \, \mathbf{y}_{[oi]}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right] & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases}$$

$$= \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ \boldsymbol{\Sigma}_{[mmi]} - \boldsymbol{\Sigma}_{[moi]}\boldsymbol{\Sigma}_{[ooi]}^{-1}\boldsymbol{\Sigma}_{[omi]} & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases} . \quad (9.4)$$

where for example $\boldsymbol{\Sigma}_{[ooi]}$ denotes the partition of $\boldsymbol{\Sigma}$ corresponding to the observed part of $\mathbf{y}_i$, etc. The different estimators are actually defined through the data weighting system given by w_i^{μ}, w_i^{δ} and w_i^{η} in (9.1) which in turn also depends on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (see below). To compute the estimators, one can use an iterative procedure in which, given the current values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the $\hat{\mathbf{y}}_i$, $\mathbf{C}_i$ and the weights are first computed by (9.3) and (9.4), and then updated by

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} w_i^{\mu}\hat{\mathbf{y}}_i \Big/ \frac{1}{n}\sum_{i=1}^{n} w_i^{\mu} \quad (9.5)$$

$$\boldsymbol{\Sigma} = \left[\frac{1}{n}\sum_{i=1}^{n} w_i^{\eta}((\hat{\mathbf{y}}_i - \boldsymbol{\mu})(\hat{\mathbf{y}}_i - \boldsymbol{\mu})^T - \mathbf{C}_i)\right] \Big/ \left[\frac{1}{n}\sum_{i=1}^{n} w_i^{\delta}\right] \quad (9.6)$$

For example, the classical MLE is obtained when $w_i^{\mu} = w_i^{\eta} = w_i^{\delta} = 1 \; \forall i$, and (9.5) and (9.6) define the EM algorithm. To construct a high breakdown estimator of mean and covariance matrix when some are missing, Cheng and Victoria-Feser (2002) propose two strategies to adapt the S-estimator of Rocke (1996) to incomplete data. The resulting estimator which is called the $ERTBS$ is then defined through (9.5) and (9.6) with

$$w_i^{\mu} = \psi(d_{oi}/k)/\left(d_{oi}/k\right) , \quad (9.7)$$

$$w_i^{\eta} = p w_i^{\mu} , \quad (9.8)$$

$$w_i^{\delta} = (d_{oi}/k)^2 w_i^{\mu} \quad (9.9)$$

where

$$d_{oi}^2 = \left(\mathbf{y}_{[oi]} - \boldsymbol{\mu}_{[oi]}\right)^T \boldsymbol{\Sigma}_{[ooi]}^{-1} \left(\mathbf{y}_{[oi]} - \boldsymbol{\mu}_{[oi]}\right) \quad (9.10)$$

is the squared Mahalanobis distance corresponding to the observed part of $\mathbf{y}_i$. ψ is the function

that defines the translated biweight S-estimator proposed by Rocke (1996) given in (4.13) and k is given in (4.14). Moreover, if the case i is uncontaminated, the data are normal and missing values are missing at random, then (9.10) is asymptotically $\chi^2_{p_i}$. p_i is the number of variables present for obervation i. The Wilson-Hilferty transformation of the chi-squared distribution yields $(d_i^2/p_i)^{1/3} \sim N(1-2/(9p_i), 2/(9p_i))$. Little and Smith (1987) therefore also proposed a probability plot of

$$Z_i = \frac{(d_i^2/p_i)^{1/3} - 1 + 2/(9p_i)}{\sqrt{2/(9p_i)}} \tag{9.11}$$

versus standard normal order statistics, that should reveal atypical observation.

To modify the MCD for missing values, Cheng and Victoria-Feser (2002) propose the following procedure. Recall that the Fast-MCD algorithm search through a certain number of subset H_k the one with the smallest determinant. If for any subset H_k there are missing values, one compute μ_k and Σ_k with the EM algorithm. The Mahalanobis distances are also changed as in (9.10) and standardized using the Wilson-Hilferty transformation. To extend the OGK to the case of missing data, Copt and Victoria-Feser (2004) propose to impute the missing values by means of the $\widehat{y}$ in (9.3) obtained by the EM algorithm, i.e. with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ estimated by (9.1) where all weights are equal to 1. The reason is that the EM algorithm is very fast, and although it leads to biased estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and therefore of the imputed values $\widehat{y}$, this shouldn't affect the resulting OGK. Indeed, the OGK downweights extreme observations in (4.23) and (4.24), and these observations can be extreme because of either the observed values or the imputed ones.

What we propose to do is to adapt the robust estimator to the case of missing data using the procedure introduced by Cheng and Victoria-Feser (2002). For the MCD and the OGK with missing data, simply compute these estimators using their modified version and then constraint them using the procedure described on page 34. For the $CTBS$ itself an analogous idea as the one proposed by Cheng and Victoria-Feser (2002) could be to use the modified version (9.10) of the mahalanobis distances and to add an E-step for computing $\hat{\mathbf{y}}_i$ and $\mathbf{C}_i$.

Chapter 10

Conclusion

In this thesis we have proposed a high breakdown estimator for very general mixed models, derived the asymptotic variances of the possible contrasts and applied a more general robust procedure of testing. This procedure was then successfully applied through an extensive simulations study to various models, some including covariates and real data issued from various research fields. We showed that a robust analysis can give another insight to the analysis of real data. As many before, we advocate the use of robust statistical procedures because real data are frequently plagued by extreme observations and the parametric models used are only approximately valid. We want to use and promote robust methods in mixed linear models because we believe that they constitute safer alternative than the classical ones (maximum likelihood, generalized least squares,...). The procedures presented here offer the advantage of taking into account the possible model misspecifications and/or the occurrence of atypical observations.

We started by presenting some of the most used models. To do so, we used the mixed linear models formulation which is more complete and flexible than the one used with traditional ANOVA. In particular we saw that in most cases, it is always possible to extract independent subvector of observation $\mathbf{y}_i$. We also derived the structure of the covariance matrix $\mathbf{\Sigma}$ for a great variety of models.

We then defined a high breakdown estimator for constrained covariance matrices. We developed a multivariate S-estimator of location and scale, namely the $CTBS$ for translated biweight S-

estimator for constrained covariance matrices, and derived the estimating equations so that it can be implemented in a relative straightforward manner. We proposed two ad-hoc modifications for the constrained MCD and the constrained OGK estimator used as starting point for the iterative procedure of the $CTBS'$s algorithm. In future research we would like to improve the constrained starting point. We also stressed the fact that the $CTBS$ can actually be used for any model in which the covariance matrix can be written as in (2.13).

We then focused on the problem of robust testing. We proposed an application of the robust score type test of Heritier and Ronchetti (1994) and derived its form in the context of mixed linear models. We have seen that robust likelihood ratio test cannot be proposed in this case, due to the shape of the objective function the S-estimator has to minimize but further work remains to be done in this domain to find a way to overcome this difficulty. We also derived the asymptotic variances of the possible contrasts, applicable to simple robust inference for the contrasts.

To illustrate the performance of our robust estimator and testing procedure, we performed an extensive simulation study. We compared the behavior of the $CTBS$ estimator against the maximum likelihood estimator with different kinds of models, some including covariates. To produce small model deviations, we created shift-outliers and explored the estimates's distributions of the mean vector and the various variance components. We have shown that the $CTBS$ is robust to data contamination at the cost of a small efficiency loss compared to the maximum likelihood estimator. We also studied the breakdown properties of the $CTBS$ for the different models and the different types of contamination. Moreover, we performed a simulation experiment to study the behavior of the $CTBS$ in small samples and saw that in this case, it also remains unbiased.

The behavior of the robust score type test for the $CTBS$ estimator was also analyzed through an extensive simulation study. As benchmark, we also computed the F statistic, which is commonly used in practice and to be fair in our comparison with its natural counterpart, the classical score type test. We noticed that for n relatively large, the robust score type test

performed well in that the empirical levels matched the true ones. The classical tests are, on the other hand, far too conservative in the presence of small amounts of data contamination. We also noticed that with small sample sizes the score type tests (robust and classical) are too conservative.

Finally, we illustrated the performance of classical and robust estimation and testing on different real data sets. We analyzed four data sets in order to see the impact on the analysis of the robust method when compared to the classical one. We showed that a robust analysis can give another insight to the analysis of the data. A diagnostic tool was also provided by means of a scatter plot of the Mahalanobis distances estimated using the MLE and the $CTBS$ respectively. These graphics help reveal atypical observations.

This concludes this thesis. We hope that this work will be helpful to many researchers and convince them that robust methods do not only remain theoretical concepts but are of real and valuable interest in practice.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and B. Csaki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Academiai Kiado.

Alario, F., J. S. and L. Ferrand (2000). Semantic and associative priming in picture naming. *The Quaterly Journal of Experimental Psychology 53*, 741–764.

Beaton, A. E. and J. W. Tukey (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics 16*, 147–185.

Bennet, C. A. (1954). Effect on measurement error on chemical process control. *Industrial Quality Control 11*, 17–20.

Berry, D. A. (1987). Logarithmic transformations in ANOVA. *Biometrics 43*, 439–456.

Braun, H. and D. McNeil (1981). Testing in robust ANOVA. *Communications in Statistics, Part A - Theory and Methods 10*, 149–165.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

Cheng, T.-C. and M.-P. Victoria-Feser (2002). High-breakdown estimation of multivariate mean and covariance with missing observations. *The British Journal of Mathematical and Statistical Psychology 55*(2), 317–335.

Copt, S. and M.-P. Victoria-Feser (2004). Fast algorithms for computing high breakdown covariance matrices with missing data. In G. P. In A. S. M. Hubert and S. V. A. (Eds.) (Eds.), *Theory and Applications of Recent Robust Methods*, Basel. Brikhauser.

Daniels, H. E. (1954). Saddlepoint approximation in statistics. *Annals of Mathematical Statistics] 25*, 631–650.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics 15*, 1269–1292.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (C/R: p22-37). *Journal of the Royal Statistical Society, Series B, Methodological 39*, 1–22.

Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics 28*, 51–60.

Fisher, R. (1925). *Statistical Methods for Research Workers (1st edn.)*. Oliver and Boyd.

Gallant, A. R. (1987). *Nonlinear statistical models*. John Wiley.

Girden, E. R. (1992). *ANOVA: repeated measures*. Sage Publications Inc.

Gnanadesikan, R. and J. R. Kettenring (1972). Robust estimates, residuals, and outlier detection with multiresponse data (Corr: V28 p1142). *Biometrics 28*, 81–124.

Graybill, F. A. (1983). *Matrices with applications in statistics*. Duxbury Press.

Hampel, F. R. (1973). Some small sample asymptotics. In J. H. (Ed.) (Ed.), *Proceedings Prague Symp. Asymptotic Statistics*, pp. 109–126. Prague: Charles University.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust statistics: The approach based on influence functions*. John Wiley and Sons.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*, 320–338.

Heritier, S. and E. Ronchetti (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association 89*, 897–904.

Holcomb, P. and W. McPherson (1994). Event-related brain potentials reflect semantic priming in an object decision task. *Brain and Cognition 24*, 259–276.

Huber, P. J. (1981). *Robust statistics*. John Wiley and Sons.

Huggins, R. M. (1993a). On the robust analysis of variance components models for pedigree data. *The Australian Journal of Statistics 35*, 43–57.

Huggins, R. M. and R. G. Staudte (1994). Variance components models for dependent cell populations. *Journal of the American Statistical Association 89*, 19–29.

Jennrich, R. I. and M. D. Schluchter (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics 42*, 805–820.

Kent, J. T. and D. E. Tyler (1991). Redescending M-estimates of multivariate location and scatter. *The Annals of Statistics 19*, 2102–2119.

Khatri, C. G. (1966). A note on a manova model applied to problem in growth curves. *The Annals of Mathematical Statistics 18*, 75–78.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences (Second edition)*. Brooks/Cole Publishing Co.

Laird, N., N. Lange, and D. Stram (1987). Maximum likelihood computations with repeated measures: Applications of the EM algorithm. In *ASA Proceedings of the Statistical Computing Section*, pp. 34–42.

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

Lindstrom, M. J. and D. M. Bates (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association 83*, 1014–1022.

Little, R. J. A. and P. J. Smith (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association 82*, 58–68.

Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics 17*, 1662–1683.

Lopuhaä, H. P. and P. J. Rousseeuw (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics 19*, 229–248.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics 15*, 661–675.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics 4*, 51–67.

Maronna, R. A. and R. H. Zamar (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics 44*(4), 307–317.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models (Second edition)*. Chapman and Hall Ltd.

Moy, G. and P. Mounoud (2003). Object recognition in young adult : Is priming with pantomimes possible ? In *Catalogue des abstracts : 8ème congrès de la société suisse de psychologie (SSP)*, Bern, Switzerland.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

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

Potthoff, R. F. and S. N. Roy (1964). A generalized multivariate analysis of variance model useful especially for growth curve problem. *Biometrika 51*, 313–326.

Rao, C. R. (1973). *Linear statistical inference and its applications*. John Wiley and Sons.

Richardson, A. M. (1997). Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association 92*, 154–161.

Richardson, A. M. and A. H. Welsh (1994). Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *The Australian Journal of Statistics 36*, 31–43.

Richardson, A. M. and A. H. Welsh (1995). Robust restricted maximum likelihood in mixed linear models. *Biometrics 51*, 1429–1439.

Robinson, J., E. Ronchetti, and G. A. Young (2002). Saddlepoint approximations and tests based on multivariate m-estimates. *The Annals of statistics 24*, 1154–1169.

Rocke, D. M. (1983). Robust statistical analysis of interlaboratory studies. *Biometrika 70*, 421–431.

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of Statistics 24*, 1327–1345.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica 7*, 327–338.

Ronchetti, E., C. Field, and W. Blanchard (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association 92*, 1017–1023.

Ronchetti, E. and R. G. Staudte (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association 89*, 550–559.

Ronchetti, E. and A. H. Welsh (1994). Empirical saddlepoint approximations for multivariate M-estimators. *Journal of the Royal Statistical Society, Series B, Methodological 56*, 313–326.

Rousseeuw, P. and V. Yohai (1984). Robust regression by means of S-estimators. In W. H. J. Franke and D. Martin (Eds.), *Robust and Nonlinear Time Series Analysis*, New-York, pp. 256–272. Springer-Verlag.

Rousseeuw, P. J. and A. M. Leroy (1987). *Robust regression and outlier detection*. John Wiley and Sons.

Rousseeuw, P. J. and K. Van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*, 212–223.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–590.

Sahai, H., A. I. Khuri, and C. H. Kapadia (1985). A second bibliography on variance components. *Communications in Statistics, Part A – Theory and Methods 14*, 63–115.

Searle, S. R., G. Casella, and C. E. McCulloch (1992). *Variance components*. John Wiley and Sons.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association 88*, 486–494.

Silvey, S. D. (1975). *Statistical inference*. John Wiley and Sons.

Sommer, S. and R. M. Huggins (1996). Variables selection using the Wald test and a robust C_p. *Applied Statistics 45*, 15–29.

Stahel, W. A. and A. Welsh (1997). Approaches to robust estimation in the simplest variance components model. *Journal of Statistical Planning and Inference 57*, 295–319.

Wellmann, J. (2000). Robustness of an S-estimator in the one-way random effects model. *Biometrical Journal. Journal of Mathematical Methods in Biosciences 42*(2), 215–221.

Welsh, A. and A. Richardson (1997). Approaches to the robust estimation of mixed models. In G. Maddala and C. Rao (Eds.), *Handbook of Statistics Vol. 15*, pp. 343–385. Holland: Elsevier Science B.V.

Woodruff, D. L. and D. M. Rocke (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association 89*, 888–896.