FACULTEIT ECONOMIE EN
BEDRIJFSWETENSCHAPPEN

KATHOLIEKE
UNIVERSITEIT
LEUVEN

ROBUST ESTIMATION AND MODEL SELECTION IN SEMIPARAMETRIC REGRESSION MODELS

Proefschrift voorgedragen tot
het behalen van de graad van
Doctor in de Toegepaste
Economische Wetenschappen

door

Kukatharmini THARMARATNAM

Nummer 366 2011

Committee

Prof. Dr. Gerda Claeskens (Promotor)	*Katholieke Universiteit Leuven*
Prof. Dr. Irène Gijbels (Co-promotor)	*Katholieke Universiteit Leuven*

Prof. Dr. Christophe Croux	*Katholieke Universiteit Leuven*
Prof. Dr. Paul Janssen	*Universiteit Hasselt*
Prof. Dr. Thomas Kneib	*Carl von Ossietzky Universität Oldenburg*
Prof. Dr. Martina Vandebroek	*Katholieke Universiteit Leuven*

Daar de proefschriften in de reeks van de Faculteit Economie en Bedrijfswetenschappen het persoonlijk werk zijn van hun auteurs, zijn alleen deze laatsten daarvoor verantwoordelijk.

Acknowledgements

"Learning is wealth none could destroy
nothing else gives genuine joy."
Thiruvalluvar - Thirukkural

Though it is impossible to express my gratitude to all individuals who supported me, I wish to extend my deepest appreciation to several special people who have helped and encouraged me in compiling this dissertation. I owe them my sincere thanks.

My earnest gratitude goes to the VLIR administration for granting the scholarship to obtain my Master's degree in Biostatistics at Universiteit Hasselt, which gave me the chance to come to Belgium. I extent my deepest gratitude to the lecturers at Universiteit Hasselt for their invaluable advises and great support that was always available and for the encouragement given to me. After my Master's degrees, I got the opportunity to work on a PhD at the Faculty of Business and Economics of K.U.Leuven.

It is with a great sense of gratitude and earnest appreciation that I acknowledge my promotor, Prof. Gerda Claeskens, for giving me her expert guidance and encouragement, constructive comments and necessary supervision that made this thesis a fruitful one. At times when I needed some encouragement, Gerda encouraged and supported me a lot. I would like to thank my co-promotor Prof. Irène Gijbels for giving me her comments and encouragements through my doctoral seminars.

I would also like to express my gratitude to the members of my doctoral committee: Prof. Christophe Croux, Prof. Martina Vandebroek,

Prof. Paul Janssen and Prof. Thomas Kneib for agreeing to be part of the committee and their insightful comments, questions and remarks. They improved the quality of my work. Moreover, I extend my deepest gratitude to the Faculty of Business and Economics, the Research Fund of K.U.Leuven and the GOA-project for their financial support.

A special word of thanks goes to my co-authors of my papers Prof. Gerda Claeskens, Prof. Christophe Croux and Prof. Matias Salibián-Barrera. It was a grateful learning experience to discuss and write a paper with them and they also helped me a lot to improve my writing skills and presentation skills. I would like to thank the reviewers of my papers for their insightful comments and questions, which helped to improve the quality of the papers.

Furthermore, I would like to thank all my colleagues of the faculty, specially all the members of ORSTAT. A special thank goes to my office mates Leen and Koen, for their company in both professional and private context. It was a pleasant and interesting experience to spend time with you. Many thanks to Annie, Nicole, Hilde, Hanneke and Barbara for helping me out with all kinds of administrative issues. I wish to thank my colleagues from Universiteit Hasselt, Betty, Chantal, Javier, Setio and Susan, also my friends Pushpike & family, Vishva and others from Sri lanka. They encouraged and supported to me during my PhD life.

Last but not least I would like to thank my family for their unconditional support to my studies in Belgium. My deep regret is that my father didn't live to see me complete my PhD, may he rest in peace and I know he would be proud. My special thanks to my mother, chithy and my sister for their constant source of love, patience, concern, support and strength during all these years. I wish to express my heart-felt gratitude to my uncles, my aunties and my cousins who have aided and encouraged me throughout this effort.

Kukatharmini Tharmaratnam Leuven, July 2011.

Summary

In the first part of this dissertation, we propose a robust estimation method for penalized regression splines based on S-estimators that can be used in the presence of outliers in the response variable. Second we study and propose a robust version of the model selection criterion AIC, Akaike's information criterion, for regression models where S- and MM-estimators are used for estimation. The last part of this dissertation presents the robust S-estimation method and a robust version of AIC for use in linear mixed models and in particular for additive semiparametric regression models.

Penalized regression splines are one of the popular methods for smoothing noisy data. The estimation methods used for fitting such a penalized regression spline model are usually based on least squares methods, which are known to be sensitive to outlying observations. The main objective of the second chapter is to extend the estimation method for penalized regression splines to that of S-estimation. We used the Tukey's biweight family of loss functions to estimate the S-estimates. We propose a computationally fast procedure for estimating penalized regression spline models via S-estimators. Simulated data and real data examples are used to illustrate the effectiveness of the procedure. The results of these examples indicate that S-estimates for penalized regression splines are more appropriate for data with outliers.

The third chapter is about robust model selection strategies for regression models. Model selection is a key component in any statistical analysis. We derive a model selection strategy in the style of Akaike's information criterion (AIC) based on S- and MM-estimators. We compare different robust AIC methods based on M-, S- and MM- estimators to the classical AIC method, that uses maximum likelihood estimators. In a simulation study we observe that the proposed AIC with S- and MM- estimators selects more appropriate models for data sets with a large contamination level of outliers in the response variables.

In the fourth chapter we study model selection strategies for semiparametric additive models fit with penalized regression splines. This estimation method is attractive because of its link to mixed models. We work specifically with outlier robust versions. In the context of mixed models there exist two different forms of AIC. The marginal AIC (MAIC) is used for selecting covariates in the model, and is based on the marginal likelihood. The conditional AIC (CAIC) is based on the conditional likelihood given the random effects. Our proposal leads to robust versions of the MAIC and CAIC that are based on S-estimators. We consider the robustness with respect to the outliers in the individual level and in the cluster level of the variables in the mixed models. Simulated data and real data examples are used to illustrate the effectiveness of the proposed method.

We discuss the computational issues using R software in the fifth chapter. We present and briefly illustrate the R-code for all statistical methods which we used in this dissertation. Finally, we discuss some general conclusions and prospectives for future research in the last chapter.

Table of contents

Chapter 1

Introduction

In the old days statistics was used by governments to keep record of births, deaths, population sizes etc. for administrative purpose and the scope was limited. The utility of statistics as a discipline has increased as the years went by. Nowadays statistical methods and techniques are used in data collection, in the presentation, the organization and the analysis and interpretations of data in many fields such as agriculture, economics, sociology, medicine, business management, etc, for different purposes.

In this dissertation we study a combination of three main topics which are of considerable interest in statistical modeling in different fields. These three topics are robust estimation methods, semiparametric regression models and Akaike's information criterion (AIC) for model selection. In this chapter, we present the definition, properties and some literature review of the statistical models, estimation methods and the model selection criterion AIC for regression models and for linear mixed models. Also we define robust estimation methods and their properties for data with outliers.

1.1 Linear regression models

Linear regression is a statistical modeling technique that relates the change in one variable to other variables. Linear regression models are used in many application in real life. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. A simple linear regression line has an equation of the form $Y = \alpha + \beta X + \varepsilon$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is β, α is the intercept, and ε is an error term.

Identifying a linear regression model requires first determining the dependent variable Y and the explanatory variables $X_1, \ldots, X_p$ that are to be included in the model. Coefficients are traditionally estimated by using ordinary least squares (OLS). This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the fitted regression line.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. A scatter plot can be a helpful tool in determining the kind of relationship between two variables.

If the association between the proposed explanatory and dependent variables appears not linear, then fitting a linear regression model to the data probably will not provide a useful model. In this case non linear regression models might be useful to fit the data.

Once a regression model has been fit to the data, one can investigate the validity of the modeling assumptions by examining the residuals (that is, the deviations from the fitted line to the observed values). Plotting the residuals against the explanatory variables reveals possible non-linear relationships among the variables and the residual plot might indicate the presence of outliers. For the situation of linear regression models where outliers might be present, we derive in Chapter 3 a robust version of Akaike's information criterion (AIC) for variable selection (see also Section 1.5).

1.2 Linear mixed models

In many applications in different fields, we need to use one of a collection of models for correlated data structures, for example, multivariate observations, clustered data, repeated measurements, longitudinal data and spatially correlated data. Often random effects are used to describe the correlation structure in clustered data, repeated measurements and longitudinal data. Models with both fixed and random effects are called mixed models.

1.2.1 Notation in mixed models

The general form of a linear mixed model for the ith subject $(i = 1, \dots, n)$ is given as follows,

$$Y_i = X_i\beta + \sum_{j=1}^{r} Z_{ij}u_{ij} + \varepsilon_i; \quad u_{ij} \sim N(0, G_j), \varepsilon_i \sim N(0, R_i), \tag{1.1}$$

where the vector Y_i has length m_i, X_i and Z_{ij} are, respectively, a $m_i \times p$ design matrix and a $m_i \times q_j$ design matrix of fixed and random effects. β is a p-vector of fixed effects and u_{ij} are the q_j-vectors of random effects. The variance matrix G_j is a $q_j \times q_j$ matrix and R_i is a $m_i \times m_i$ matrix. We assume that the random effects $\{u_{ij}; i = 1, \dots, n, j = 1, \dots, r\}$ and the set of error terms $\{\varepsilon_1, \dots, \varepsilon_n\}$ are independent. In matrix notation, $Y = X\beta + Zu + \varepsilon$. Here $Y = (Y_1, \dots, Y_n)^t$ has length $N = \sum_{i=1}^n m_i$, $X = (X_1^t, \dots, X_n^t)^t$ is a $N \times p$ design matrix of fixed effects, Z is a $N \times q$ block diagonal design matrix of random effects, $q = \sum_{j=1}^r q_j$, $u = (u_1^t, \dots, u_r^t)^t$ is a q-vector of random effects, $R = \text{diag}(R_1, \dots, R_n)$ is a $N \times N$ matrix and $G = \text{diag}(G_1, \dots, G_r)$ is a $q \times q$ block diagonal matrix.

1.2.2 The marginal likelihood for a mixed model

Consider the marginal model $Y \sim N(X\beta, V)$, where $V = (Z^tGZ + R)$. The framework of mixed models suggests the use of maximum likelihood estimation of β and V by minimizing the marginal log-likelihood (leaving out constants),

$$m\ell(\beta, V|Y) = -\frac{1}{2}\log|V| - \frac{1}{2}(Y - X\beta)^tV^{-1}(Y - X\beta). \tag{1.2}$$

This likelihood approach is computationally convenient and software already exists for longitudinal, hierarchical or other dependent data.

1.2.3 The conditional likelihood for a mixed model

The conditional distribution of $Y|u \sim N(X\beta + Zu, R)$ corresponding to the mixed model representation and the conditional log-likelihood (leaving out constants),

$$c\ell(Y|\beta, u, R) = -\frac{1}{2}\log|R| - \frac{1}{2}(Y - X\beta - Zu)^t R^{-1}(Y - X\beta - Zu). \quad (1.3)$$

The conditional likelihood of $Y|u$ has a mean that depends on u. The estimators for β, u and for the variance components that are contained in the matrix R, are obtained by maximizing the conditional likelihood. We study robust model selection in linear mixed models in Chapter 4.

1.3 Robust estimation methods

We consider the regression model

$$Y_i = \theta_0^t X_i + u_i, \qquad i = 1, \ldots, n, \quad (1.4)$$

where the response variables $Y_i \in \mathbb{R}$, the covariate vector $X_i \in \mathbb{R}^p$ with a corresponding coefficient vector $\theta_0 \in \mathbb{R}^p$ and the u_i are random errors independent from the explanatory variable X_i, with mean zero and constant variance σ^2. In the case that outliers are present in the data, only the majority of the data follows the above model (1.4). Extreme observations might occur in both the explanatory variables and the response. Model selection investigates the inclusion or exclusion of components of the covariate vector X. To handle this problem of outliers, in the model fitting procedure there exist several robust estimation methods. We give a brief overview of some of the robust estimators in this section, which are used in Chapters 2, 3 and 4.

1.3.1 M-estimators

A general M-estimator (Huber, 1964) is defined as the minimum with respect to θ of the objective function $\sum_{i=1}^n \rho(Y_i|x_i, \theta)$, for a given function

ρ that has the properties of being even, non-decreasing in $[0, \infty)$ and with $\rho(0) = 0$. Equivalently, when the response values $Y_1, \ldots, Y_n$ are independent, the M-estimator for θ solves the equation

$$\sum_{i=1}^{n} \psi(Y_i|x_i, \theta) = 0 \tag{1.5}$$

where $\psi(y|x, \theta) = \frac{\partial \rho(y|x,\theta)}{\partial \theta}$. Intuitively, to take care of outliers which result in large residuals when OLS estimation would be used, the function $\rho(\cdot)$ should increase at a slower rate than t^2, particularly for large residuals. A common choice for ρ is given by Huber's family with an unbounded loss function

$$\rho_c(t) = \begin{cases} t^2 & \text{if } |t| \le c \\ 2\,c\,|t| - c^2 & \text{if } |t| > c\,, \end{cases} \tag{1.6}$$

where $c > 0$ is a tuning constant that can be thought of as a threshold value such that observations with residuals larger than c have a reduced effect in the estimating equation (1.5). The plot of Huber's loss function (1.6) is given in Figure 1.1. A table with different values of c is given in Maronna et al. (2006, page 27), see Huber (2004) and Hampel et al. (1986) for details. A typical choice for c is $1.345\,\widehat{\sigma}_m$, where $\widehat{\sigma}_m$ is the median absolute deviation of the residuals. The formula for median absolute deviation (MAD) is $\text{MAD}(x_1, \ldots, x_n) = 1.4826\,\text{median}\{|x_i - \text{median}(x_1, \ldots, x_n)|, i = 1, \ldots, n\}$, where $1/\Phi^{-1}(3/4) = 1.4826$. The 95% asymptotic efficiency on the standard normal distribution is obtained with the constant 1.345. The M-estimator is computed with $\rho(y_i|x_i, \theta) = \rho_c\left(\frac{y_i - \theta^t x_i}{\widehat{\sigma}_m}\right)$. In practice, iteration is used between estimation of θ and estimation of the standard deviation σ until convergence.

1.3.2 S-estimators

S-estimators for linear regression were introduced by Rousseeuw and Yohai (1984) as an alternative to M-estimators that do not suffer that much from leverage points (which are outliers in the covariates) and at the same time have a high breakdown point and do not require an auxiliary scale estimator.

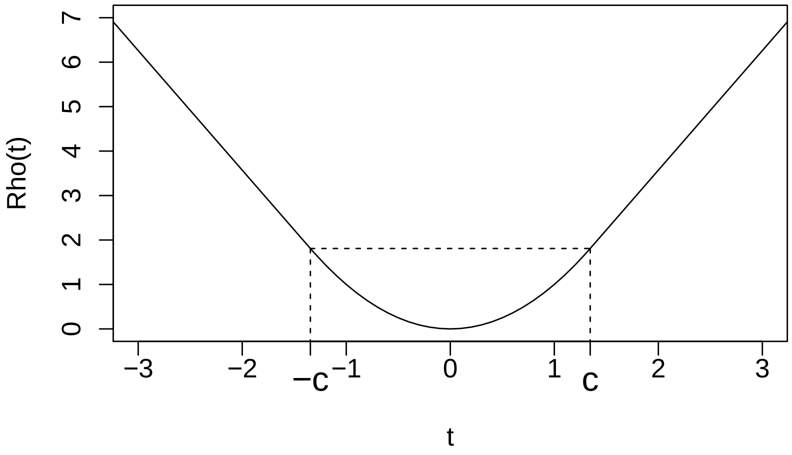


Figure 1.1: *Plot of Huber's loss function*

Let G_0 and F_0 be the cumulative distribution functions of X_i and u_i respectively. The cumulative distribution of (Y_i, X_i) under model (1.4) is then given by $H_0(y, x) = G_0(x)F_0(y - \theta_0^t x)$. In the presence of outliers, we make the assumption that the cumulative distribution function H of the data belongs to a contamination neighborhood of H_0 of size ϵ_0. More precisely,

$$H \in \mathcal{H}_{\epsilon_0} = \{(1 - \epsilon)H_0 + \epsilon H^*; \epsilon \in [0, \epsilon_0]\},$$

where H^* is an arbitrary cumulative distribution function and $\epsilon_0 < 0.5$.

The loss function ρ_0 is a function that is even, continuously differentiable, non-decreasing on $[0, \infty)$, satisfies that $\rho_0(0) = 0$ and is bounded from above by 1, that is, $\sup_{u \in \mathbb{R}} \rho_0(u) = 1$. We define $b = E_{F_0}[\rho_0(u)]$ to ensure consistency of the scale estimator under the central model F_0 and assume that $\epsilon_0 < b < 1 - \epsilon_0$. The notation E_{F_0} means that the expectation is computed with respect to F_0.

First we implicitly define the scale function $\widehat{\sigma}_n(\theta)$ by that function of

θ that satisfies the equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{Y_i-\theta^t x_i}{\widehat{\sigma}_n(\theta)}\right)=b,$$

with $\rho(y_i|x_i,\theta) = \rho_0\left(\frac{y_i-\theta^t x_i}{\widehat{\sigma}_n(\theta)}\right)$. The S-estimator $\widehat{\theta}_s$ minimizes the scale function, $\widehat{\theta}_s = \mathrm{argmin}_{\theta\in\mathbb{R}^p}\,\widehat{\sigma}_n(\theta)$, and the scale estimator itself is $\widehat{\sigma}_s = \widehat{\sigma}_n(\widehat{\theta}_s)$.

A commonly used family of loss functions ρ_0 is given by Tukey's bisquare family (Beaton and Tukey, 1974)

$$\rho_d(u)=\begin{cases}3\,(u/d)^2-3\,(u/d)^4+(u/d)^6 & \text{if } |u|\le d\,,\\ 1 & \text{if } |u|>d\,.\end{cases}\tag{1.7}$$

The plot of Tukey's bisquare loss function (1.7) is given in Figure 1.2.

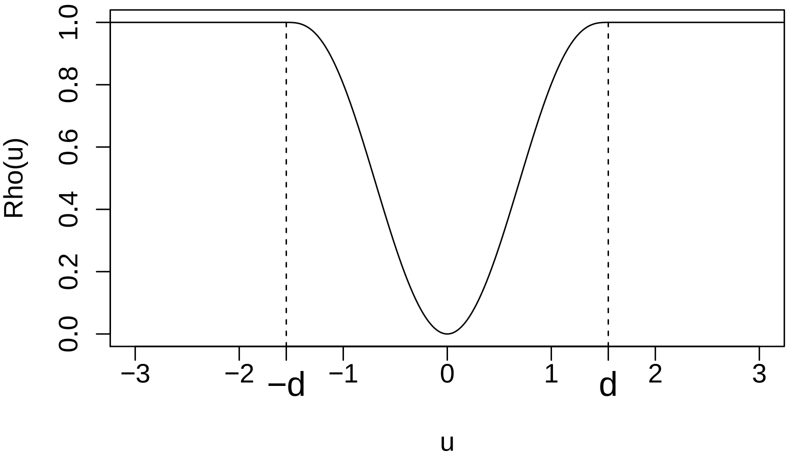


Figure 1.2: *Plot of Tukey's bisquare loss function*

The choice $d = 1.5476$ yields $b = E_\Phi\left[\rho_d\left(Z\right)\right] = 0.5$. The associated S-regression estimator has the maximal asymptotic breakdown point of 50%

(Rousseeuw and Yohai, 1984). Estimators with 30% breakdown point are gotten when $d = 2.5608$, resulting in a higher efficiency.

1.3.3 MM-estimators

A further step in robust estimation uses the S-scale estimator in an M-estimating equation. Let $\rho_1 : \mathbb{R} \to \mathbb{R}_+$ be a loss function such that $\rho_1(u) \leq \rho_0(u)$ for all $u \in \mathbb{R}$ and $\sup_u \rho_1(u) = \sup_u \rho_0(u)$. The MM-regression estimator $\widehat{\theta}_{mm}$ is defined as the global minimum of $f : \mathbb{R}^p \to \mathbb{R}_+$, with

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{Y_i - \theta^t x_i}{\widehat{\sigma}_s} \right).$$

We can write the MM-estimator as follows

$$\widehat{\theta}_{mm} = \underset{\|\theta\| \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{Y_i - \theta^t x_i}{\widehat{\sigma}_s} \right),$$

The MM-variance estimator is taken to be the S-scale estimator $\widehat{\sigma}_s$. In practice, often the choice ρ_1 is a re-scaled version of ρ_0 (Tukey's bi-square family loss function). Let $\rho_0(u) = \rho_d(u/d_0)$ and $\rho_1(u) = \rho_d(u/d_1)$ and to get $\rho_1(u) \leq \rho_0(u)$ we must have $d_1 \geq d_0$, the larger d_1 gives a higher efficiency at the normal distribution.

We illustrate the importance of the robust estimators in the presence of outliers in the data by some real data and by a simulated data example. We have used Hofstedt's highway data from the R library `alr3` as `data(highway)` (see also Weisberg, 2005). There are 39 observations on several highway related measurements in this dataset. The response variable is the accident rate per million vehicle miles in the year 1973 and there are 11 potential explanatory variables. The explanatory variable is the truck volume as a percentage of the total volume.

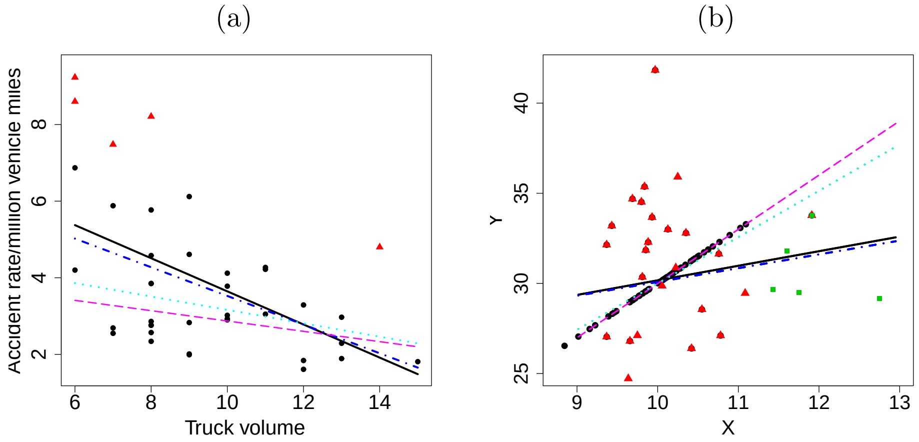


Figure 1.3: *Fitted values (a) Highway data and (b) Simulated data with y-outliers (solid triangle) and x-outliers (solid square). Fitted curves from least squares estimation (solid line), M-estimation (dot-dashed), a more relevant fit from S-estimation (dotted line) and MM-estimation (dashed).*

Figure 1.3 shows clearly the need of robust estimation methods in case outliers are present. Outliers in y are plotted as triangles, while outliers in the explanatory variable are plotted as squares. Non-outlying observations are represented by solid circles. Panel (a) for the highway data shows the effect of outliers on the response variable where robust S- and MM-estimation methods result in more relevant fits to the data. The M-estimator behaves here more in line with the least squares estimator. In the simulated example in panel (b) we have generated 25% outliers in y and 15% outliers on x. There are outliers on both the response variable and the explanatory variable. Also here, S- and MM-estimation leads to a good fit.

1.3.4 S-estimators for linear mixed models

In Maronna (1976), the robust estimators of the multivariate mean and covariance are built from weighted score functions. Their breakdown point becomes smaller as the dimension increases. In the linear mixed model, the dimension can be large and it is important to consider high-breakdown estimators. Copt and Victoria-Feser (2006) propose a constrained S-estimator for the multivariate mean and a constrained covariance estimator for mixed linear models as in (1.1). Consider the marginal log-likelihood as given in (1.2). The S-estimator of the multivariate mean and covariance is defined as the solution for β and V that minimizes $\det(V) = |V|$ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\sqrt{(Y_i - X_i\beta)^t V^{-1}(Y_i - X_i\beta)}\right) = b_0, \tag{1.8}$$

where $\rho(u)$ is a even function with the properties of being non-decreasing and being a bounded function, as given by Rousseeuw and Yohai (1984) and b_0 is a parameter chosen to determine the breakdown point. Generally b_0 is defined by $b_0 = E(\rho(\sqrt{U}))$, where U is a Chi-squared distribution with p degrees of freedom, p is number of parameters in the model. The Tukey biweight loss function is a usual choice in the univariate case. Rocke (1996) proposed a translated Tukey biweight function for multivariate data that can control the probability of an estimator giving a null weight to extreme observations, the latter which is called the asymptotic rejection probability (ARP). The translated Tukey biweight loss function is given by,

$$\rho(d; c.M) = \begin{cases} \frac{d^2}{2}, & 0 \le d \le M \\ \rho_{M\le d\le M+c}(d; c, M), & M \le d \le M+c \\ \frac{M^2}{2} + \frac{c(5c+16M)}{30}, & d > M+c, \end{cases} \tag{1.9}$$

with $M + c < \infty$ and

$$\begin{aligned} \rho_{M\le d\le M+c}(d; c, M) = {} & \frac{M^2}{2} - \frac{M^2(M^4 - 5M^2c^2 + 15c^4)}{30c^4} \\ & + d^2\left(0.5 + \frac{M^4}{2c^4} - \frac{M^2}{c^2}\right) + d^3\left(\frac{4M}{3c^2} - \frac{4M^3}{3c^4}\right) \\ & + d^4\left(\frac{3M^2}{2c^4} - \frac{1}{2c^2}\right) - \frac{4Md^5}{5c^4} + \frac{d^6}{6c^4}. \end{aligned}$$

The plot of the translated Tukey biweight loss function (1.9) is given in Figure 1.4.

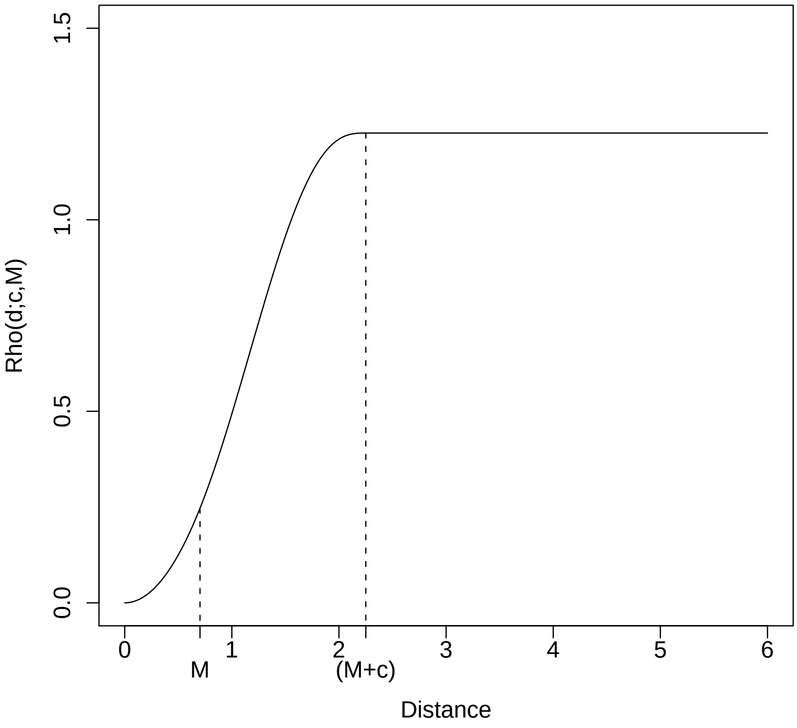


Figure 1.4: *Plot of the translated Tukey biweight loss function*

This translated Tukey biweight ρ function leads to the weight function

$$u(d;c,M) = \begin{cases} 1, & 0 \leq d \leq M \\ \left(1 - \left(\frac{d-M}{c}\right)^2\right)^2, & M \leq d \leq M + c \\ 0, & d > M + c, \end{cases} \tag{1.10}$$

where the constants c and M can be chosen to achieve the desired breakdown point and ARP. Copt and Victoria-Feser (2006) compute the S-estimator for the parameter estimators for fixed effects and variance components for the marginal model $Y \sim N(X\beta, V)$ given by,

$$\widehat{\beta} = \left(X^t \widehat{V}^{-1} X\right)^{-1} \frac{\sum_{i=1}^n u(d_i) X_i^t \widehat{V}_i^{-1} Y_i}{\sum_{i=1}^n u(d_i)}, \tag{1.11}$$

$$\widehat{S}_0 = \left[\frac{1}{n}\sum_{i=1}^n u(d_i) d_i^2\right]^{-1} Q^{-1} U, \tag{1.12}$$

$$\widehat{V} = Z^t \widehat{G} Z + \widehat{R}, \tag{1.13}$$

where $d_i = \sqrt{(Y_i - X_i\widehat{\beta})^t \widehat{V}_i^{-1} (Y_i - X_i\widehat{\beta})}$, $u(d_i) = \frac{\partial}{\partial d_i}\rho(d_i)/d_i$,

$\widehat{S}_0 = (\widehat{\sigma}_0^2, \ldots, \widehat{\sigma}_{K_0}^2)^t$, $K_0 = \sum_{j=1}^q K_j$, $\widehat{G} = (\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_{K_0}^2) I_{K_0}$, $\widehat{R} = \widehat{\sigma}_0^2 I_n$,
$Q = \left[\mathrm{Tr}\left(\widehat{V}^{-1} z_j z_j^t \widehat{V}^{-1} z_k z_k^t\right)\right]_{j,k=0,\ldots,K_0}$
and $U = \left[\frac{1}{n}\sum_{i=1}^n p u(d_i) \times (Y_i - X_i\widehat{\beta})^t \widehat{V}_i^{-1} z_j z_j^t \widehat{V}_i^{-1} (Y_i - X_i\widehat{\beta})\right]_{j=0,\ldots,K_0}$,
here X_i is the ith row of the design matrix X, V_i is the (i,i)th element of variance matrix V, z_j is the jth block matrix of the design matrix Z. With this procedure, we do not yet obtain predictions for the random effects. To construct the random effect predictions, we need to consider the conditional model. Details of the derivation of S-estimators for linear mixed models are given in Chapter 4.

1.4 Semiparametric regression models

Semiparametric regression models retain the virtues of both parametric and nonparametric modeling. Ruppert et al. (2003) presents various semiparametric regression models, their inference procedures and applications. Additive penalized regression spline models have found a lot of applications in the last few years. These models are easy to fit. They allow a flexible choice of the knots and in addition the smoothing parameter can be obtained in a data driven way. All this has made them a popular nonparametric smoothing method.

1.4.1 Mixed model representation of additive penalized regression splines models

Consider the regression model

$$Y_i = \sum_{j=0}^{p} \beta_j X_{ji} + \sum_{j=1}^{q} m_j(X_{p+j,i}) + \varepsilon_i, \qquad i = 1, \dots, n. \tag{1.14}$$

where X_{0i},...,X_{qi} are $q+1$ explanatory variables for observation i, $m_j : [a, b] \to \mathbb{R}$ are unknown but smooth regression functions for each of the explanatory variables and the random errors ε_i are independent from the explanatory variables, with mean zero and a constant variance σ^2. We are interested in estimating the parameters $\beta_0, \beta_1, \dots, \beta_p$ together with the functions $m_j(\cdot), j = 1, \dots, q$ based on a random sample $(Y_i, X_{ji}), i = 1, \dots, n$.

To fix ideas, we focus our presentation on truncated polynomial bases, but other choices can be used as well. More specifically, we take regression splines of degree s, K_j inner knots $a < \kappa_{j1} < \cdots < \kappa_{jK_j} < b$ and define

$$m_j(x; \beta_j, \kappa_j) = \beta_{j1} x + \cdots + \beta_{js} x^s + \sum_{k=1}^{K_j} u_{jk} \{\max(x - \kappa_{jk}, 0)\}^s.$$

Given a sample, this approach transforms the estimation of $m_j(\cdot)$ into a least squares problem. To reduce the influence of the spline coefficients $u_{jk}(k = 1, \dots, K_j)$, a penalty is introduced. Denote $(x - \kappa_k)_+ = \{\max(x - \kappa_k, 0)\}^s$. Define the design matrix $F = (X, Z)$ with

$$X = \begin{pmatrix} x_{11} & \dots & x_{p1} & x_{p+1,1} & \dots & x_{p+1,1}^s & \cdots & x_{q1} & \dots & x_{q1}^s \\ x_{12} & \dots & x_{p2} & x_{p+1,2} & \dots & x_{p+1,2}^s & \cdots & x_{q2} & \dots & x_{q2}^s \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{1n} & \dots & x_{pn} & x_{p+1,n} & \dots & x_{p+1,n}^s & \cdots & x_{qn} & \dots & x_{qn}^s \end{pmatrix}$$

$Z =$

$$\begin{pmatrix} (x_{p+1,1} - \kappa_{11})_+^s & \cdots & (x_{p+1,1} - \kappa_{1K_1})_+^s & \cdots & (x_{q1} - \kappa_{q1})_+^s & \cdots & (x_{q1} - \kappa_{qK_q})_+^s \\ & & \vdots & & & \vdots & \\ (x_{p+1,n} - \kappa_{11})_+^s & \cdots & (x_{p+1,n} - \kappa_{1K_1})_+^s & \cdots & (x_{qn} - \kappa_{q1})_+^s & \cdots & (x_{qn} - \kappa_{qK_q})_+^s \end{pmatrix}$$

Let $\beta = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{j1}, \ldots, \beta_{js})^t; j = 1, \ldots, q,\ u = (u_1, \ldots, u_q)^t$ with $u_j = (u_{j1}, \ldots, u_{jK_j})^t$. A traditional penalized least squares (PLS) estimator with smoothing parameter λ_j for $m_j(\cdot), j = 1, \ldots, q$ is

$$(\widehat{\beta}, \widehat{u})_{PLS} = \underset{\beta,u}{\operatorname{argmin}} \left[\parallel Y - F \begin{pmatrix} \beta \\ u \end{pmatrix} \parallel^2 + \sum_{j=1}^{q} \lambda_j \parallel u_j \parallel^2 \right].$$

The penalized least squares estimator of β and u are explicitly obtained as follows,

$$(\widehat{\beta}, \widehat{u})_{PLS} = (F^t F + D_\lambda)^{-1} F^t Y, \tag{1.15}$$

where $Y = (Y_1, \ldots, Y_n)^t$, $D_\lambda = \operatorname{diag}(0_{s_X}, \lambda_1 1_{K_1}, \cdots, \lambda_q 1_{K_q})$, s_X is the number of columns of the design matrix X, $s_X = p + 1 + s_1 + \cdots + s_q$, 0_s is a vector of 0 with length s, 1_s is a vector of 1 with length s.

1.4.2 Mixed model representation

There exists a convenient connection between penalized splines and mixed models (Brumback et al., 1999; Ruppert et al., 2003). Model (1.14) is re-written using the matrix notation

$$Y = X\beta + Zu + \varepsilon \tag{1.16}$$

where Y is a $n \times 1$ vector of response variables, β is a $(p + (q - p)s) \times 1$ vector of fixed effects, u is a $K_0 \times 1$ vector of random effects, $K_0 = \sum_{j=1}^{q} K_j$, X and Z are the $n \times (p + (q - p)s)$ and $n \times K_0$ design matrices for the fixed and random effects respectively, ε is the error term, a $n \times 1$ vector. We assume that u and ε are independent and normally distributed as $u \sim N(0, G)$, $\varepsilon \sim N(0, R)$ where $G = \sigma_u^2 I_{K_0}$, $R = \sigma^2 I_n$, u is considered to be a random variable. The Lagrange multiplier or penalty constant λ_j appears in this model as a ratio of the error variance to the random effects variance: $\lambda_j = \sigma^2/\sigma_{u_j}^2$.

The estimation of the parameters β and u entails minimizing the penalized least squares criterion

$$\parallel Y - X\beta - Zu \parallel^2 + u^t D_\lambda u, \tag{1.17}$$

where D_λ is a known $K_0 \times K_0$ penalty matrix. For a given smoothing parameter matrix D_λ, the penalized least squares estimators from (1.17) are

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{u} \end{pmatrix} = \begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ + D_\lambda \end{pmatrix}^{-1} \begin{pmatrix} X^t \\ Z^t \end{pmatrix} Y, \tag{1.18}$$

and the fitted values are $\widehat{Y} = X\widehat{\beta} + Z\widehat{u} = HY$, where H is the smoothing matrix given by

$$H = \begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ + D_\lambda \end{pmatrix}^{-1} \begin{pmatrix} X^t \\ Z^t \end{pmatrix}. \tag{1.19}$$

The trace of the smoothing matrix H has a monotone relationship with the smoothing parameters $\lambda_1, \cdots, \lambda_q$, and is often used to compute the generalized, or an effective, degrees of freedom. The mixed model representation of the semiparametric regression models is used in model selection in Chapter 4.

1.5 Model selection methods

Variable selection is fundamental in statistical modeling. In practice, a number of variables are available to include in an analysis, but many of them may not be relevant and should be excluded from the final model in order to increase the accuracy of the estimators and predictors based on this model.

The variable selection procedures need special care in the presence of outliers in the data. Since most of the classical procedures are likelihood-based, alternatives have been developed. Some of the main developments to make classical model selection procedures for linear models less sensitive to outlying observations are a robust version of Akaike's information criterion (AIC Akaike, 1973) based on M-estimators (Ronchetti, 1985). Other model selection methods based on M-estimators are a robust C_p (Ronchetti and Staudte, 1994; Sommer and Staudte, 1995) and a robust version of cross-validation (Ronchetti et al., 1997). A robust way of model selection using the concept of stochastic complexity is presented in Qian and Künsch (1998). Agostinelli (2002) deals with weighted versions of likelihood estimators and uses this method for model selection. Several of these model selection methods are described in Maronna et al. (2006, Sec. 5.12) and Claeskens and Hjort (2008, Ch. 2 and 4).

1.5.1 Akaike's Information Criterion (AIC)

There are different model selection strategies corresponding to different aims and uses associated with the selected model. One of the most important selection criteria is Akaike's information criterion (AIC). In practice, the maximum likelihood estimator $\widehat{\theta}$ is computed based on data. Suppose that $y = (y_1, \ldots, y_n)$ is a vector of observations, generated from a true underlying distribution with joint density $g(.)$ and that $f_\theta(.) = f(.;\theta)$ is a family of approximating models with unknown parameter θ. The Akaike information is defined as $-2E_y(E_z[\log f_\theta(z)])$, where E_z is the expectation with regard to the distribution of another realization z. A general formula of AIC for a candidate model M which contains a parameter vector θ is,

$$\text{AIC}(M) = -2 \log \text{likelihood}(\widehat{\theta}) + 2\,\text{length}(\theta) \tag{1.20}$$

where $\widehat{\theta}$ is the maximum likelihood estimator of θ. The AIC value is making a balance between a good fit and complexity. Akaike's method aims at finding models that in a sense have few parameters but nevertheless fit the data well. The model with the smallest AIC value is selected as a best choice of the model. In the next section we explain the connection between AIC and the Kullback-Leibler distance for regression models.

1.5.2 AIC and the Kullback-Leibler distance

Generally, we construct models for observations $Y = (Y_1, \ldots, Y_n)$, containing the parameters $\theta = (\theta_1, \ldots, \theta_p)^t$. In this model the joint density for Y is defined $f_{\text{joint}}(y;\theta)$. The likelihood function can be written as follows,

$$L_n(\theta) = f_{\text{joint}}(y, \theta).$$

We work with the log-likelihood function $\ell_n(\theta) = \log L_n(\theta)$ instead of with the likelihood itself. The maximum likelihood estimator of θ is the maximizer of $L_n(\theta)$. That is,

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}(L_n(\theta)) = \underset{\theta}{\operatorname{argmax}}(\ell_n(\theta)).$$

If the data Y are independent and identically distributed, the likelihood and log-likelihood functions can be written as

$$L_n(\theta) = \prod_{i=1}^{n} f(y_i;\theta) \text{ and } \ell_n(\theta) = \sum_{i=1}^{n} f(y_i,\theta),$$

in terms of the density $f(y;\theta)$ for an individual observation. We should make a distinction between the model $f(y,\theta)$ that we construct for the data, and the true density $g(y)$ of the data. The true density $g(y)$ is always unknown and this is called the data-generating density.

There are several ways of measuring closeness of a parametric approximation $f(.,\theta)$ to the true density g. The Kullback-Leibler (KL) distance, or discrepancy, is linked to the maximum likelihood method and the general definition is

$$\mathrm{KL}(g, f(.,\theta)) = \int g(y) \log \frac{g(y)}{f(y,\theta)} dy. \tag{1.21}$$

The maximum likelihood estimator $\widehat{\theta}$ that maximizes $\ell_n(\theta)$ will, under suitable conditions, tend to the minimizer θ_0 of the Kullback-Leibler discrepancy from the true model to the used model $f(y;\theta)$. Thus

$$\widehat{\theta} \to \theta_0 = \underset{\theta}{\operatorname{argmin}}\{\mathrm{KL}(g, f(.,\theta))\}.$$

Here θ_0 is the best approximating parameter value. The maximum likelihood estimator aims at providing the best approximation to the real density g inside the parametric set of density functions $f(.;\theta)$.

Let us consider the definition of AIC in (1.20). The AIC method is penalizing maximized log-likelihoods for complexity, but it is not clear why the penalty factor should take the form of (1.20). The maximum likelihood estimator $\widehat{\theta}$ minimizes the Kullback-Leibler distance (1.21). We can re-write the Kullback-Leibler distance as follows,

$$\begin{aligned}\mathrm{KL}(g, f(., \widehat{\theta})) &= \int g(y)\{\log g(y) - \log f(y, \widehat{\theta})\}dy \\ &= \int g(y) \log g(y)dy - R_n\end{aligned}$$

where $R_n = \int g(y) \log f(y, \widehat{\theta})dy$. The first term $\int g(y) \log g(y)dy$ is the same across models. Therefore, we need to study only R_n, it is dependent upon the data via the maximum likelihood estimator $\widehat{\theta}$ and also R_n is a random variable. Let us take the expected value of R_n with respect to maximum likelihood estimator, under the true density $g(.)$ for the response variable Y_i and denote this as Q_n,

$$Q_n = E_g R_n = E_g \int g(y) \log f(y, \widehat{\theta})dy. \qquad (1.22)$$

We can estimate Q_n from the data by

$$\widehat{Q}_n = n^{-1} \sum_{i=1}^{n} \log f(Y_i, \widehat{\theta}) = n^{-1}\ell_n(\widehat{\theta}). \qquad (1.23)$$

We can define the score function $u(y, \theta)$ and information matrix $I(y, \theta)$ for the situation of identically and independent distributed response variables,

$$u(y, \theta) = \frac{\partial \log f(y, \theta)}{\partial \theta} \quad \text{and} \quad I(y, \theta) = \frac{\partial^2 \log f(y, \theta)}{\partial \theta \partial \theta^t}.$$

We need to define $p \times p$ matrices J and K as,

$$J = \mathrm{E}_g I(Y, \theta) \quad \text{and} \quad K = \mathrm{Var}_g u(Y, \theta).$$

Under various and essentially rather mild regularity conditions, one may prove that

$$\widehat{\theta} = \theta + J^{-1}\overline{U_n} + o_P(n^{-1/2}),$$

where $\overline{U_n} = n^{-1} \sum_{i=1}^{n} u(Y_i, \theta)$. We can write the asymptotic distribution of the maximum likelihood estimator $\widehat{\theta}$ in the following form,

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} J^{-1}U' = N_p(0, J^{-1}KJ^{-1}). \qquad (1.24)$$

Let us denote $V_n = \sqrt{n}(\widehat{\theta} - \theta)$ and let $\overline{Z_n}$ be the average of the i.i.d zero mean variables $Z_i = \log f(Y_i, \theta) - Q_0$, where $Q_0 = \int g(y) \log f(y, \theta) dy$. We can get the result as,

$$\widehat{Q}_n - R_n = \overline{Z_n} + n^{-1} V_n^t J V_n + o_P(n^{-1}). \tag{1.25}$$

Consider this as a view of (1.24) and since $V_n^t J V_n \rightarrow_d W = (U')^t J^{-1} U'$, where $U' \sim N_q(0, K)$ and by equation (1.25), this leads to the approximation,

$$\mathrm{E}(\widehat{Q}_n - Q_n) \approx p^*/n, \;\text{ where }\; p^* = \mathrm{E}W = \mathrm{Tr}(J^{-1}K).$$

We can write $\widehat{Q}_n - p^*/n = n^{-1}\{\ell_n(\hat{\theta}) - p^*\}$ as the bias-corrected version of the estimator $\widehat{Q}_n$. Note that, if the model is correct, $g(y) = f(y, \theta)$, then $J = K$, and $p^* = \text{length}(\theta)$. Taking $p^* = p$, leads to the AIC formula (1.20). For more details and proofs, see Section 2.3 in Claeskens and Hjort (2008). This criterion is called as the TIC, as proposed in Takeuchi (1976) and is considered to be an AIC-type of model selection in the literature.

1.6 Marginal AIC for mixed models

Variable selection for the additive semiparametric models is challenging since it includes the selection of variables in the nonparametric component as well as the identification of variables in the parametric component. This will increase the computational efforts.

When writing the additive penalized regression spline model in its representation of a linear mixed model, we can use the variable selection methods for the corresponding linear mixed model. An AIC based on the marginal likelihood is generally used in linear mixed models (marginal AIC) and returned by standard statistical software.

Using the marginal likelihood (1.2) from the marginal model $Y \sim N(X\beta, V)$

$$\mathrm{mAIC} = -2\, m\ell(\widehat{\beta}, \widehat{V}|Y) + 2\,(p_X + v + 1)$$

where p_X is the number of columns of X, $v + 1$ is the number of variance components in (1.16) and $(\widehat{\beta}, \widehat{V})$ are the marginal maximum likelihood estimates based on maximizing (1.2). This is appropriate when the interest is in the fixed effects in a mixed model context.

1.7 Conditional AIC for mixed models

The paper by Vaida and Blanchard (2005) focuses on model selection for linear mixed models using the conditional Akaike information criterion and shows that the marginal AIC (Akaike, 1973) is not appropriate for conditional inference when both the fixed and the random parts of linear mixed models are of interest. Vaida and Blanchard (2005) propose the conditional Akaike information and the conditional AIC based on the likelihood for the conditional model $Y|u \sim N(X\beta + Zu, R)$. This approach is more appropriate when the focus is on the random effects. The penalty term in the conditional AIC is related to the effective number of parameters of a linear mixed model proposed by Hodges and Sargent (2001). In semi-parametric models, such as penalized regression spline models, model selection entails selecting among the explanatory variables, interactions between them and the random components. Smooth functions in penalized regression splines using the linear mixed model representation are parameterized by variance parameters ($\lambda_j = \sigma^2/\sigma^2_{u_j}$) and mean parameters ($\beta$).

The conditional Akaike information in this setting is defined as,

$$c\text{AI} \;\; = \;\; -2\, E_{y,u}(E_{\widetilde{Y}|u}[\log\{f(\widetilde{Y}|\widehat{\beta}(y), \widehat{u}(y))\}])$$

where $g(Y,u) = g_{Y|u}(Y|u)g_u(u)$ is the true joint distribution of Y and u and $\widetilde{Y}$ is a random variable with the same distribution as Y, though independent from Y. Assume that the variance components are known. Vaida and Blanchard (2005) show that an asymptotically unbiased estimator of cAI is their conditional AIC using the hat matrix H as given in (1.19),

$$\text{cAIC} = -2\, \log_{Y|u} f(Y|\,\widehat{\beta}, \widehat{u}, R) + 2(\text{Tr}(H) + 1) \tag{1.26}$$

where $\log_{Y|u} f(Y|\,\widehat{\beta}, \widehat{u}, R)$ is the conditional log-likelihood for Y, conditioning on u as given in (1.3). Vaida and Blanchard (2005) considered the case of a known variance component R and computed the cAIC in (1.26) by using

$$\begin{aligned}\log_{Y|u} f(Y|\,\widehat{\beta},\widehat{u},R) &= -\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(|R|)\\ &\quad -\frac{1}{2}(Y-X\widehat{\beta}-Z\widehat{u})^tR^{-1}(Y-X\widehat{\beta}-Z\widehat{u}).\end{aligned}$$

This conditional log-likelihood is computed at the estimated quantities $(\widehat{\beta},\widehat{u})$ based on maximum likelihood or restricted maximum likelihood estimation. The penalty includes the trace of smoothing matrix H, projecting Y onto $\widehat{Y}=X\widehat{\beta}+Z\widehat{u}$. The effective degrees of freedom $\mathrm{Tr}(H)$ lies between those of a linear model without random effects u and those of a linear model with fixed effects u as noted in Vaida and Blanchard (2005).

Liang et al. (2008) have proposed a corrected conditional AIC that accounts for the estimation of the variance components. Instead of $2(\mathrm{Tr}(H)+1)$ they used as a penalty term $2(\mathrm{Tr}(\partial\widehat{Y}/\partial Y)+1)$. Greven and Kneib (2010) study the theoretical properties of the corrected conditional AIC and they provide a computationally feasible penalty term using maximum likelihood or restricted maximum likelihood estimators.

We extend this model selection on both the fixed and the random effects in the linear mixed models to be used with S-estimators, in particular we derive the appropriate penalty terms in Chapter 4.

Chapter 2

S-Estimation for penalized regression splines

This chapter is based on the following publication:
Tharmaratnam, K., Claeskens, G., Croux, C. and Salibian-Barrera, M. (2010). S-estimation for penalized regression splines. *Journal of Computational and Graphical Statistics*, 19(3), 609-625.

Abstract

This chapter is about S-estimation for penalized regression splines. Penalized regression splines are one of the currently most used methods for smoothing noisy data. The estimation method used for fitting such a penalized regression spline model is mostly based on least squares methods, which are known to be sensitive to outlying observations. In real world applications, outliers are quite commonly observed. There are several robust estimation methods taking outlying observations into account. We define and study S-estimators for penalized regression spline models. Hereby we replace the least squares estimation method for penalized regression splines by a suitable S-estimation method. By keeping the modeling by means of splines and by keeping the penalty term, though using S-estimators instead of least squares estimators, we arrive at an estimation method that is both

robust and flexible enough to capture non-linear trends in the data. Simulated data and a real data example are used to illustrate the effectiveness of the procedure. Software code (for use with R) is available online.

2.1 Introduction

Penalized regression spline models have found a lot of applications in the last 10–15 years. Their ease of fitting and flexible choice of knots and smoothing parameter has made them a popular nonparametric smoothing method. The use of a combination of regression splines, which have a substantially smaller number of knots than the sample size, and the use of a penalty, dates back to at least O'Sullivan (1986) who used a cubic B-spline basis for estimation in inverse problems. Hybrid splines, which approximate the smoothing splines (the latter which have knots equal to the data points and a penalty for complexity) have been studied by Kelly and Rice (1990) and Besse et al. (1997). Eilers and Marx (1996) proposed the use of a difference penalty on the spline coefficients. For more explanation and examples on the class of penalized regression spline models, we refer to Ruppert et al. (2003). Theoretical aspects of penalized spline regression fitting are only recently starting to develop. We refer to Hall and Opsomer (2005) for a white noise representation of the model, Claeskens et al. (2009) for relating theoretical properties of penalized regression splines to those of regression splines (without a penalty) and smoothing splines, and Kauermann et al. (2009) for results in generalized penalized spline smoothing models.

The estimation method used for fitting such penalized regression spline models minimizes the sum of squared residuals subject to a bound on the norm of the spline regression coefficients. Alternatively, one can work with the equivalent penalized minimization problem, that has a closed-form expression for its solution. It is easy to see that this approach may be highly sensitive to the presence of a small proportion of atypical observations. One way to obtain a fit that is more resistant to the effect of atypical observations in the data is to replace the squared residuals by a slowly increasing loss function, as it is done for M-regression estimators (Huber, 1964). Early proposals dealing with M-type robust smoothing go back to Huber (1979) and Cox (1983) for the particular case of cubic regression splines. Other papers on the topic include Härdle and Gasser (1984), Silverman (1985) and Hall and Jones (1990). More recently Oh et al. (2004, 2007) used the "pseudo data" introduced in Cox (1983) to derive iterative algorithms for M-type cubic splines, while Lee and Oh (2007) applied this approach to M-penalized regression splines.

As already noted by Huber (1979) and Cox (1983), a serious difficulty with replacing the squared residuals by a slower-increasing loss function to obtain M-type smoothers is that one needs to either know or robustly estimate the residual scale. In principle, one can use simultaneous estimation of the regression and scale parameters (Huber's Proposal II (Huber, 1964)), as in Lee and Oh (2007). Unfortunately, our numerical experiments show that, as in the simple location/scale and linear regression models, simultaneous estimation of the regression coefficients and the residual scale may not have good robustness properties. In particular, the procedure may be seriously affected by a relatively small proportion of outliers.

The main purpose of this chapter is to propose robust penalized regression splines that are able to resist the potentially damaging effect of outliers in the sample, and that do not require the separate estimation of the residual scale. To achieve these goals we propose to compute penalized S-regression spline estimators. In the unpenalized case, these estimators are consistent, asymptotically normal, and have high-breakdown point regardless of the dimension of the vector of regression coefficients (Rousseeuw and Yohai, 1984).

First we show that the solution to the penalized S-regression spline problem can be written as the solution of a weighted penalized least squares problem. This representation naturally leads to an iterative algorithm to compute these estimators. We also study how to robustly select the penalty parameter when there may be outliers in the data. This was studied for M-cubic splines by Cantoni and Ronchetti (2001b). We propose a robust penalty parameter selection criteria based on generalized cross-validation that also borrows from the weighted penalized least squares representation of the penalized S-regression spline estimator. Extensive simulation studies show that our algorithm works well in practice and that the resulting regression function estimator is robust to the presence of outliers in the data. Furthermore, these estimators compare favorably to the penalized M-regression splines of Lee and Oh (2007).

The rest of this chapter is organized as follows. Section 2.2 introduces penalized S-regression spline estimators and an algorithm to compute them, while Section 2.3 reports the results of a simulation study that compared the performance of classical least-squares, penalized M- and S-regression spline estimators. A data set is analyzed in Section 2.4 and concluding remarks are included in Section 2.5.

2.2 Penalized S-regression splines

Consider the regression model

$$Y = m(x) + \varepsilon, \tag{2.1}$$

where $m : [a, b] \to \mathbb{R}$ is an unknown but smooth regression function and the random error ε is independent from the explanatory variable $x \in \mathbb{R}$, and has mean zero and constant variance σ^2. We are interested in estimating the function $m(x)$ based on a random sample (Y_i, x_i), $i = 1, \ldots, n$.

A widely used estimation method for $m(x)$ is to assume that

$$m(x) = \sum_{j=1}^{L} \beta_j \, f_j(x) \,,$$

for some basis $f_1(x), \ldots, f_L(x)$ and coefficients $\beta_j \in \mathbb{R}$. To fix ideas, we focus our presentation on truncated polynomial bases, but other choices can be used as well. More specifically, we take K inner knots $a < \kappa_1 < \cdots < \kappa_K < b$ and define

$$m(x; \beta) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{j=1}^{K} \beta_{p+j} \, (x - \kappa_j)_+^p \,, \tag{2.2}$$

where $a_+ = \max(a, 0)$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_{p+K})^t$. Given a sample $(Y_1, x_1), \ldots, (Y_n, x_n)$ this approach transforms the estimation of $m(\cdot)$ into a least squares problem, where we find the member of the class $m(x; \beta)$ that minimizes the sum of squared residuals. To avoid overfitting, we solve the problem subject to a bound on the size of the spline coefficients:

$$\min_{\beta \in \mathbb{R}^{p+K+1}} \sum_{i=1}^{n} (Y_i - m(x_i; \beta))^2 \quad \text{subject to} \quad \sum_{j=1}^{K} \beta_{p+j}^2 \leq C,$$

for some $C > 0$ as in Ruppert et al. (2003). If we let $F(x) = (1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p)^t \in \mathbb{R}^{p+K+1}$, it is easy to see that the penalized least squares regression spline estimator $\widehat{\beta}$ is the minimizer of

$$\sum_{i=1}^{n} \left(Y_i - F(x_i)^t \beta\right)^2 + \lambda \sum_{j=1}^{K} \beta_{p+j}^2, \tag{2.3}$$

for some penalty parameter $\lambda > 0$.

Denoting the spline design matrix $F = \{F(x_1)^t, \ldots, F(x_n)^t\}^t$, the vector of responses $Y = (Y_1, \ldots, Y_n)^t$ and $D_p = \text{diag}(0_{p+1}, 1_K)$ the matrix indicating that only the spline coefficients are to be penalized, the resulting estimator $\widehat{\beta}$ is given by the ridge regression formula

$$\widehat{\beta} = \left(F^t F + \lambda D\right)^{-1} F^t \, Y, \tag{2.4}$$

and the corresponding estimated vector $\widehat{m} = (\widehat{m}(x_1), \ldots, \widehat{m}(x_n))^t$:

$$\widehat{m} = F\widehat{\beta} = F \, \left(F^t F + \lambda D\right)^{-1} F^t \, Y. \tag{2.5}$$

2.2.1 Penalized S-regression spline estimation

It is easy to see that, as in unpenalized linear regression, the estimator defined by the minimum of (2.3) may be seriously affected by a small proportion of atypical observations. These “outliers” may be errors in the data, or, more interestingly, data points that follow a different model or random process. In what follows we will be concerned with estimating the regression function $m(x)$ in (2.1) that applies to the majority of the data.

A straightforward approach to obtain penalized regression estimators that are more resistant to outliers than those defined by (2.3) is to replace the squared residual loss function by a slowly increasing function ρ:

$$\sum_{i=1}^{n} \rho \left(Y_i - F(x_i)^t \beta \right) + \lambda \sum_{j=1}^{K} \beta_{p+j}^2 \,, \tag{2.6}$$

where ρ is even, non-decreasing in $[0, \infty)$ and $\rho(0) = 0$ (see also Lee and Oh, 2007). Intuitively, the function $\rho(t)$ should increase at a slower rate than t^2, particularly for large residuals. A common choice for ρ in (2.6) is given by Huber's family

$$\rho_c(t) = \begin{cases} t^2 & \text{if } |t| \leq c \\ 2\,c\,|t| - c^2 & \text{if } |t| > c\,, \end{cases} \tag{2.7}$$

where $c > 0$ is a tuning constant. The parameter c can be thought of as a threshold such that observations with residuals larger than c have a reduced effect on the estimating equation (2.6). Note that as c increases, the minimum of (2.6) approaches that of (2.3). In other words, the estimator downweights the influence of observations with large residual (i.e. larger than c).

To apply this method in practice, we need to select a value of c depending on σ, the standard deviation of the errors ε in (2.1). This can be easily done if a robust scale estimator $\widehat{\sigma}_n$ of σ is available. In this case we can compute our estimator using the standardized residuals:

$$\widehat{\beta}_n = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_c \left(\frac{Y_i - F(x_i)^t \beta}{\widehat{\sigma}_n} \right) + \lambda \sum_{j=1}^{K} \beta_{p+j}^2 . \tag{2.8}$$

Given a set of residuals $r_i = Y_i - F(x_i)^t \widehat{\beta}_n$, $i = 1, \ldots, n$, corresponding to an estimator $\widehat{\beta}_n$, a robust M-scale estimator $\widehat{\sigma}_n$ (Huber, 1964) satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{r_i}{\widehat{\sigma}_n} \right) = b, \tag{2.9}$$

where, $\rho : \mathbb{R} \to [0, \infty)$ is bounded and even and, to obtain consistency when the errors are normal, the constant b satisfies $b = E_\Phi \left[\rho\left(Z\right) \right]$, with Φ

the standard normal distribution. Note that if $\rho(t) = t^2$ and $b = 1$ then $\widehat{\sigma}_n = s_n$ the residual standard deviation.

Huber (1964) proposed to simultaneously solve the "regression" and "scale" equations, (2.8) and (2.9), respectively. In our context this is equivalent to finding the solutions $\widehat{\beta}_n$ and $\widehat{\sigma}_n$ to the following non-linear system of equations:

$$\frac{\partial}{\partial \beta} \left(\sum_{i=1}^{n} \rho_c \left(\frac{Y_i - F(x_i)^t \beta}{\widehat{\sigma}_n} \right) + \lambda \sum_{j=1}^{K} \beta_{p+j}^2 \, . \right) \Bigg|_{\beta = \widehat{\beta}_n} = \mathbf{0} \, ,$$

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{Y_i - F(x_i)^t \, \widehat{\beta}_n}{\widehat{\sigma}_n} \right) = b.$$

Finding $\widehat{\beta}_n$ and $\widehat{\sigma}_n$ generally requires using an iterative algorithm. This scheme is known in the robustness literature as Huber's Proposal II. Unfortunately, the robustness properties of the solution to this problem are not completely satisfactory. In particular, the resulting estimators may not be resistant to outliers, i.e. they have low breakdown point (see Donoho and Huber (1983) for a definition of breakdown point). This was shown by Maronna and Yohai (1991) for simultaneous general M-estimators of regression and scale.

S-estimators for linear regression were introduced by Rousseeuw and Yohai (1984). They can be tuned to have a high breakdown point and do not require an auxiliary residual scale estimator. The basic idea is to note that the least squares estimator is the vector of regression coefficients that produces residuals with minimal sample standard deviation.

A robust alternative is then obtained by finding the vector of regression coefficients β that produces residuals that minimize a robust scale estimator of the residuals, instead of the standard deviation. In other words, the S-estimators are defined by

$$\widehat{\beta}_n = \underset{\beta}{\operatorname{argmin}} \ \widehat{\sigma}_n \left(\beta \right) , \tag{2.10}$$

where $\widehat{\sigma}_n(\beta)$ is an M-scale that solves (2.9). It is easy to see that $\widehat{\sigma}_n = \widehat{\sigma}_n(\widehat{\beta})$ is also a consistent estimator of the scale σ of the errors. For linear regression models, Rousseeuw and Yohai (1984) and Davies (1990) showed that S-estimators are consistent and asymptotically normal when the distribution of the errors is symmetric.

Note that there is no explicit formula to compute $\widehat{\sigma}(\beta)$ for each β. Furthermore if ρ is bounded, then the function $\sigma(\beta)$ is non-convex, and may have several local minima. Solving (2.10) is a difficult numerical problem that involves finding the minimum of an implicitly defined non-convex function in several variables. A recently proposed algorithm for unpenalized S-regression estimators can be found in Salibian-Barrera and Yohai (2006).

One way to obtain robust penalized spline estimators is to replace the mean squared residuals in (2.3) by a robust estimator of the scale of residuals. In this chapter we consider using the S-scale, which can naturally be seen as a penalized S-regression spline estimator.

More specifically, we define $\widehat{\beta}_S$ as

$$\widehat{\beta}_S = \underset{\beta}{\operatorname{argmin}} \left[n\,\widehat{\sigma}_n^2\,(\beta) + \lambda\,\beta^t D \beta \right], \tag{2.11}$$

where, for each β, $\widehat{\sigma}_n(\beta)$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Y_i - F(x_i)^t \beta}{\widehat{\sigma}_n(\beta)}\right) = b, \tag{2.12}$$

the constant $b = E_\Phi\,[\rho\,(Z)]$, and Φ is the standard normal distribution (Maronna et al., 2006).

A commonly used family of loss functions ρ is given by Tukey's bi-square family (Beaton and Tukey, 1974)

$$\rho_d(u) = \begin{cases} 3\,(u/d)^2 - 3\,(u/d)^4 + (u/d)^6 & \text{if } |u| \leq d, \\ 1 & \text{if } |u| > d. \end{cases} \tag{2.13}$$

The choice $d = 1.5476$ yields $b = E_\Phi\,[\rho\,(Z)] = 0.50$. The associated unpenalized S-regression estimator has maximal asymptotic breakdown point 50% (Rousseeuw and Yohai, 1984). Tukey's bi-square ρ function is the standard choice for a bounded ρ function. Changing the ρ function will not significantly increase the efficiency of the estimator (see in Table 2.5, section 2.3.2). The use of the biweight loss function leads to an efficiency that comes close to the maximal value.

The next result shows that the critical points of the objective function in (2.11) can be written as the solution of a weighted penalized splines problem. This expression suggests an iterative procedure to compute the penalized S-regression spline estimators. A similar procedure holds for computing penalized MM-regression spline estimators.

Result 2.1. The penalized S-regression spline estimator for the regression spline model (2.1) can be written as $\widehat{m}_S = F\widehat{\beta}_S$ where

$$\widehat{\beta}_S = \left\{ F^t W(\widehat{\beta}_S) F + \frac{\lambda}{\tau(\widehat{\beta}_S)} D \right\}^{-1} F^t W(\widehat{\beta}_S) Y, \tag{2.14}$$

where $W(\beta) = \operatorname{diag}(W_i(\beta)) \in \mathbb{R}^{n\times n}$ with $W_i(\beta) = \rho'\left(\tilde{r}_i(\beta)\right)/\tilde{r}_i(\beta)$, $\tilde{r}_i(\beta) = (Y_i - F(x_i)^t\,\beta)/\widehat{\sigma}_n(\beta)$, and
$\tau(\beta) = n\,\widehat{\sigma}_n^2(\beta)\,/\,\left[(Y - F\beta)^t\, W(\beta)\,(Y - F\beta)\right]$.

Proof of Result 2.1: Taking the derivative with respect to β for $\widehat{\sigma}_n(\beta) \neq 0$ of the M-scale function in (2.12), we obtain

$$\sum_{i=1}^{n} \rho'\left(\frac{r_i(\beta)}{\widehat{\sigma}_n(\beta)}\right)\left(\frac{-F(x_i)\widehat{\sigma}_n(\beta) - r_i(\beta)\nabla\widehat{\sigma}_n(\beta)}{\widehat{\sigma}_n^2(\beta)}\right) = \mathbf{0},$$

where $\nabla\widehat{\sigma}_n(\beta) = \partial\widehat{\sigma}_n(\beta)/\partial\beta$. It follows that

$$\begin{aligned}
\nabla\widehat{\sigma}_n(\beta) &= -\sum_{i=1}^{n} \rho'\left(\frac{r_i(\beta)}{\widehat{\sigma}_n(\beta)}\right) F(x_i) \Bigg/ \left[\sum_{i=1}^{n} \rho'\left(\frac{r_i(\beta)}{\widehat{\sigma}_n(\beta)}\right)\left(\frac{r_i(\beta)}{\widehat{\sigma}_n(\beta)}\right)\right] \\
&= \left[-\widehat{\sigma}_n(\beta)\, F^t\, W(\beta)\, r(\beta)\right] \Big/ \left[r(\beta)^t\, W(\beta)\, r(\beta)\right],
\end{aligned} \tag{2.15}$$

where $r(\beta) = (Y - F^t\beta)$. At the minimum of (2.11) $\widehat{\beta}_S$ we have

$$2\,n\,\widehat{\sigma}_n(\widehat{\beta}_S)\,\nabla\widehat{\sigma}_n(\widehat{\beta}_S) + 2\,\lambda\, D\,\widehat{\beta}_S = \mathbf{0},$$

from which follows, using (2.15) that

$$-\tau(\widehat{\beta}_S)\, F^t\, W(\widehat{\beta}_S)\, r(\widehat{\beta}_S) + \lambda\, D\,\widehat{\beta}_S = \mathbf{0},$$

and thus equation (2.14) follows. □

Remark 2.1. Note that both the weights and the penalty parameter on the right-hand side of (2.14) depend on $\widehat{\beta}_S$ on the left of that equation. Although not useful for direct calculation of $\widehat{\beta}_S$, this representation naturally suggests iterations of the form

$$\widehat{\beta}_{S,k+1} = \left\{ F^t W(\widehat{\beta}_{S,k}) F \, + \, \lambda D \tau(\widehat{\beta}_{S,k})^{-1} \right\}^{-1} F^t W(\widehat{\beta}_{S,k}) Y, k = 0, 1, \ldots ,$$

to find critical points of (2.11). The corresponding algorithm is presented in the next section 2.2.2.

Remark 2.2. When $\rho(t) = t^2$ the M-scale estimator $\widehat{\sigma}_n$ reduces to the sample standard deviation. In this case we have $W(\beta) = 2\, I_n$, where I_n is the $n \times n$ identity matrix, and $\tau(\beta) = 1/2$. Hence, as expected, (2.14) reduces to the usual penalized least squares formula (2.4).

2.2.2 Algorithm

Although (2.14) suggests easily implementable iterations to calculate a critical point of (2.11), care should be taken as the function $\widehat{\sigma}_n : \mathbb{R}^p \to \mathbb{R}_+$ in (2.12) is generally non-convex. In other words, the objective function in (2.11) may have several critical points that only correspond to local minima. As a result, the iterations derived from Result 2.1 above may converge to different critical points (some of them non-optimal) depending on the starting value. As it is done for S-estimators for linear regression models, we propose to start the iterations from many initial points, and select the best resulting point (in terms of value of the objective function) as our approximate solution to the minimization problem (2.11).

Our algorithm can be described in the following steps:

Step (1) Let $\tilde{\beta}_1^{(0)}, \ldots, \tilde{\beta}_J^{(0)}$, be initial candidates. For each $\tilde{\beta}_j^{(0)}$:

(a) Compute $\widehat{\sigma}_n(\widehat{\beta}_j^{(0)})$, $\tau(\widehat{\beta}_j^{(0)})$, and $W(\widehat{\beta}_j^{(0)})$.

(b) Set $k = 0$. Iterate the following steps:

(i) Let$\widehat{\beta}_j^{(k+1)} =$
$\left\{ F^t W(\widehat{\beta}_j^{(k)}) F + \lambda D \tau^{-1}(\widehat{\beta}_j^{(k)}) \right\}^{-1} F^t W(\widehat{\beta}_j^{(k)}) Y.$

(ii) If either $k = maxit$ (maximum number of iterations) or $\|\widehat{\beta}_j^{(k)} - \widehat{\beta}_j^{(k+1)}\| < \epsilon \, \|\widehat{\beta}_j^{(k)}\|$ where $\epsilon > 0$ is a fixed small constant (the tolerance level) , then set $\widehat{\beta}_j^F = \widehat{\beta}_j^{(k)}$ and `break`.

(iii) Else, compute $\widehat{\sigma}_n(\widehat{\beta}_j^{(k+1)})$, $\tau(\widehat{\beta}_j^{(k+1)})$, $W(\widehat{\beta}_j^{(k+1)})$ and set $k \leftarrow k + 1$.

Step (2) Calculate the objective function for each $\widehat{\beta}_j^F$, $j = 1, 2, \ldots, J$, and select the one with the lowest value, i.e. let

$$\widehat{\beta}_S = \underset{1 \le j \le J}{\operatorname{argmin}} \left[n \, \widehat{\sigma}_n^2(\widehat{\beta}_j^F) + \lambda \, \widehat{\beta}_j^F \, D \, \widehat{\beta}_j^F \right].$$

The J initial candidates $\tilde{\beta}_j^{(0)}$ in Step 1 can be chosen in a number of ways. Intuitively we want them to correspond to different regions of the optimization domain. In linear regression problems, these initial points are generally chosen based on the sample. For example, if there are d covariates, J random subsamples of size $d + 1$ are selected from the data, and $\tilde{\beta}_j^{(0)}$ is set to the least squares fit of the j-th subsample. A similar approach can be applied here, where, to avoid ill-conditioned subsamples caused by the sparsity of the design matrix based on the spline basis in (2.2), we take subsamples of larger size, e.g. `floor`$(n/5)$. Note that this set of J initial candidates can also be extended to include the M- and classical penalized regression splines estimators at very little additional computational cost.

We have coded the above algorithm in `R` (R Development Core Team, 2008), and made it publicly available at `http://www.stat.ubc.ca/~matias/penalised`, as well as through the journal's supplemental materials facility. In our experience the above algorithm converges without problems in the vast majority of the cases. The algorithm with $\epsilon = 10^{-6}$ and $maxit = 500$ converges generally in less than 60 iterations. For all of our simulation experiments, see section 2.3.2, we have never encountered a situation where the algorithm for penalized S-regression spline estimation diverged.

2.2.3 Penalty parameter selection

To avoid overfitting the data, the penalty parameter λ in (2.11) is often chosen so as to minimize an estimator of the resulting mean squared prediction error. Such an estimator can be computed by leave-one-out cross-validation. More specifically, for each value of λ, let

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{m}(x_i)^{(-i)} \right)^2 ,$$

where $\widehat{m}(x)^{(-i)}$ is the regression estimator obtained without using the pair of observations (y_i, x_i). To evaluate $\mathrm{CV}(\lambda)$ above it is not necessary to re-compute the estimator $\widehat{m}(x)$ n times. It has been shown in Ruppert et al. (2003) that

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \widehat{m}(x_i))^2}{(1 - H(\lambda)_{i,i})^2}, \tag{2.16}$$

where $H(\lambda)_{i,i}$ denotes the i-th diagonal element of the "hat"-matrix

$$H(\lambda) = F \left(F^t \, F + \lambda \, D \right)^{-1} \, F^t,$$

with F and D as in (2.4). Furthermore, if one replaces each $(1 - H(\lambda)_{i,i})$ by their average $1 - \mathrm{trace}(H(\lambda))/n$, the generalized cross-validation criterion is obtained.

$$\begin{aligned} \mathrm{GCV}(\lambda) &= n \sum_{i=1}^{n} (y_i - \widehat{m}(x_i))^2 \,/\, (n - \mathrm{trace}(H(\lambda)))^2 \\ &= n \left\| Y - F\widehat{\beta} \right\|^2 / \, (n - \mathrm{trace}(H(\lambda)))^2 \, . \end{aligned} \tag{2.17}$$

See Craven and Whaba (1979) and Ruppert et al. (2003), among others, for more details.

Using these criteria to select a value of λ when the data may contain outliers is generally not recommended (see, for example, Cantoni and

Ronchetti (2001b) and references therein). Intuitively one can see that all observations y_i, $i = 1, \ldots, n$ in (2.17) are treated with equal importance. However, if, for some $1 \leq j \leq n$, the observation y_j is atypical, we would not want to fit it well. In other words, regardless of the robustness of the estimator $\widehat{m}(x)$, the criteria above may select a value of λ that results in an estimated $m(x_j)$ closer to y_j than desired. For the case of M-type smoothing splines, using the concept of pseudo-data of Cox (1983), Cantoni and Ronchetti (2001b) proposed to down-weight the terms in (2.16) according to their residuals. This resulted in their robust CV criterion. Define the scaled residuals of the M-estimator by $\tilde{r}_{i,M} = (y_i - \widehat{m}_M(x_i))/\widehat{\sigma}$, where $\widehat{\sigma}$ is the median absolute deviation of the residuals and $\widehat{m}_M(x_i)$ is the M-estimator of $m(x_i)$. With $\overline{\rho_c''} = n^{-1} \sum_{i=1}^n \rho_c''(\tilde{r}_{i,M})$, ρ_c'' denoting the second derivative of ρ_c and $KK = (I_n + (\lambda\widehat{\sigma}/\overline{\rho_c''})D_p)^{-1}$,

$$\mathrm{RCV} = \frac{1}{n}\left(\frac{\widehat{\sigma}}{\overline{\rho_c''}}\right)^2 \sum_{i=1}^n \frac{\{\rho_c'(\tilde{r}_{i,M})\}^2}{(1-K_{ii})^2}.$$

For penalized S-regression splines, Result 2.1 suggests that we can think of $\widehat{\beta}_S$ as the solution to

$$\min_{\beta} \left\| W(\widehat{\beta}_S)^{1/2}\,(Y - F\,\beta)\right\|^2 + (\lambda/\tau(\widehat{\beta}_S))\,\beta^t\,D\,\beta,$$

where $W(\widehat{\beta}_S)$ and $\tau(\widehat{\beta}_S)$ are given in Result 2.1. The above representations leads us to consider the GCV criterion in (2.17) with response variable $\tilde{Y} = W(\widehat{\beta}_S)^{1/2}\,Y$, predictors $\tilde{F} = W(\widehat{\beta}_S)^{1/2}\,F$ and penalty term $\lambda/\tau(\widehat{\beta}_S)$. Noting that some of the weights may be zero, we propose to select λ by minimizing

$$\mathrm{RGCV}(\lambda) = n_w \left\| W(\widehat{\beta})^{1/2}\left(Y - F\widehat{\beta}\right)\right\|^2 / \left(n_w - \mathrm{trace}(H_S(\lambda))\right)^2, \quad (2.18)$$

where n_w is the number of non-zero weights and

$$\begin{aligned} H_S(\lambda) &= \tilde{F}\left(\tilde{F}^t\tilde{F} + (\lambda/\tau(\widehat{\beta}_S))\,D\right)^{-1}\tilde{F}^t \\ &= W(\widehat{\beta}_S)^{1/2}\,F\left(F^tW(\widehat{\beta}_S)F + (\lambda/\tau(\widehat{\beta}_S))\,D\right)^{-1} F\,W(\widehat{\beta}_S)^{1/2}, \end{aligned}$$

2.3 Numerical results

2.3.1 Simulation settings

The settings for the simulation study are as follows. The observations for the design variable $x_1, \ldots, x_n$ are generated from the uniform distribution on the interval $[-1, 1]$, for various sample sizes. These values are kept fixed for all settings to reduce simulation variability. The sample sizes taken are $n = 25$, 100 and 250.

For the mean structure in (2.1) we have used the following functions, which represent a variety of shapes, $m_1(x) = \sin(\pi x)$, $m_2(x) = \sin(2\pi(1 - x)^2)$, $m_3(x) = x + x^2 + x^3 + x^4$, and $m_4(x) = -20 + e^{3x}$. Function m_2 is the same one used by Lee and Oh (2007) to facilitate a comparison with the results presented there.

For the error distribution we used five possibilities, ordered according to the heaviness of their tails, (i) uniform distribution(-1,1), (ii) normal distribution $N(0, 0.7^2)$, (iii) logistic distribution(0,1), (iv) slash distribution, defined as $N(0, 1)$/uniform$(0, 1)$, and (v) Cauchy distribution(0,1). Both the Cauchy and slash distribution are heavy-tailed.

We compare three penalized regression spline estimation methods in this simulation study: (A) the non-robust method for penalized regression spline estimation as in (2.5), using the method of penalized least squares (LS), (B) Penalized M-regression spline estimators as studied by Lee and Oh (2007). (C) the method proposed in this chapter, using penalized S-regression spline estimators, and employing the algorithm as described in section 2.2.2. For the proposed method using penalized S-regression spline estimators we use the Tukey's biweight family of loss function $\rho_d(u)$ as in (2.13) with $d = 1.547$. For the penalized M-regression spline estimators we use, as suggested in Lee and Oh (2007), $\rho_c(t)$ as in (2.7) with $c = 1.345\,\widehat{\sigma}$, where $\widehat{\sigma}$ is the median absolute deviation of residuals.

For all three methods, we use truncated cubic splines ($p = 3$) with $K = 6, 25$ or 35 knots (corresponding to sample sizes 25, 100 and 250), spread equally according to the quantiles of the data. We have tried with different choices of K as well (results shown in Table 2.6, section 2.3.2) and found similar results. The penalty parameter λ is chosen by minimizing the generalized cross validation (GCV) criterion for the LS-estimation method. Robust cross validation (RCV) defined in Cantoni and Ronchetti (2001b) is used for the M-regression spline estimation method. Robust generalized cross validation (RGCV) defined in section 2.2.3 is used for the S-estimation method.

For the proposed method of penalized S-regression spline estimation and the M-regression spline estimation method as proposed by Lee and Oh (2007) we set the tolerance level in the algorithm step(1) (b) (ii) to $\epsilon = 10^{-6}$. The maximum number of iterations was set to 500.

To investigate the robustness of the methods against outliers, we randomly generated different percentages of outliers (5%, 10%, 20%, 30% and 40%) for each of the simulated cases using either a normal distribution with mean 20 and standard deviation 20, to get scattered outliers, or with mean 20 and standard deviation 2 for a more concentrated cloud of outliers.

To give an impression on the variability of the obtained estimators, we plot in Figure 2.1, a scatter plot of one of the randomly generated data sets, together with the fitted values from the penalized LS-, M- and S-regression spline estimation methods. We used randomly generated data sets with mean function $m_1(x)$ and error distribution $N(0, 1)$ for sample size $n = 100$. Figure 2.1 (a) shows the situation without outliers, giving close correspondence between all three methods. In the situation of 30% of scattered outliers in Figure 2.1 (b), the drastic effects of the outliers are clearly visible for the penalized least squares method. A smaller effect is detected for the penalized M-regression spline estimation method. In contrast to both penalized LS- and M-regression spline estimator, the penalized S-regression spline estimator remains close to the true regression function, also in presence of outliers.

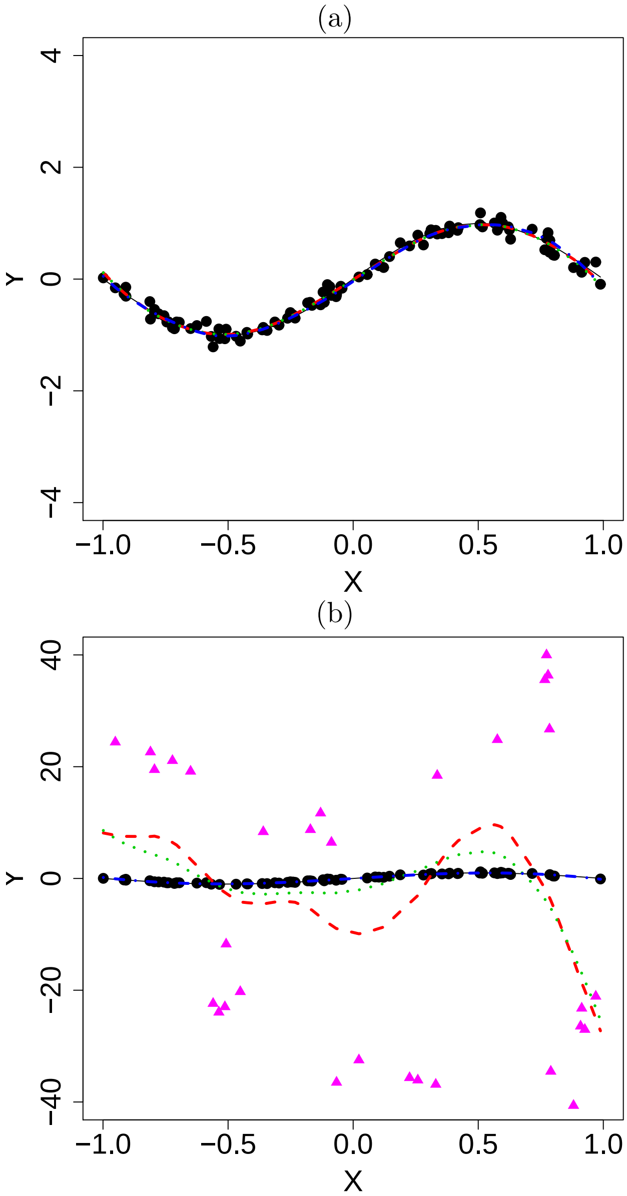


Figure 2.1: *Fitted values (a) without outliers and (b) with 30% of outliers from* $N(20, 2^2)$ *. True function* $\sin(\pi x)$ *(solid line); fitted curves from penalized LS-regression spline estimation (dashed); penalized M-regression spline estimation (dotted) and penalized S-regression spline estimation (dot-dashed).*

2.3.2 Simulation results

The goodness of fit of the estimated model is quantified by computing the median average squared error and median absolute deviation of average squared error. Denoting $\widehat{m}_j(x_i)$ the estimated value of $m(x_i)$ for simulation run j ($j = 1, \ldots, J = 1000$), the average squared error (ASE) is defined by

$$ASE_j = \frac{1}{n}\sum_{i=1}^{n}(m(x_i) - \widehat{m}_j(x_i))^2, \quad j = 1, 2, \ldots, J.$$

Table 2.1 presents summary values of the ASE (median and median absolute deviation) for the three estimation methods for the normal error distribution and with mean function m_1.

In all cases, the median ASE of the proposed method of penalized S-regression spline estimation is smaller than that of the other two methods for samples with more than 10% of outliers. Note that Lee and Oh's (2007) method of penalized M-regression spline estimation works better for samples with 5% and 10% of outliers.

For the penalized least squares and penalized M-regression spline estimators, the ASE is clearly increasing with the percentage of outliers increasing. For penalized S-regression spline estimation, the ASE values tend to be quite stable, only increasing near a high fraction of outliers ($> 40\%$). As expected, the goodness of fit as measured by the ASE values improves for larger sample sizes.

Table 2.1 clearly shows that the penalized least squares method may already break down with only 5% of outliers. For the proposed method of penalized S-regression spline estimation, the simulated ASE values are relatively small even with 40% of scattered outliers for sample sizes $n = 100$ and $n = 250$. For $n = 25$ a clearer increase (breakdown) is observed for the penalized S-regression spline estimation method when the presence of outliers reaches 40% of the sample size. For penalized M-regression spline estimation, the breakdown arrives earlier, showing the need for taking the scale into consideration in the fitting method and working with a bounded ρ-function.

Table 2.1: *Median and median absolute deviation (between parenthesis) of the average squared error ASE for penalized least squares (LS), penalized M (M) and penalized S (S) regression spline estimation for data generated with mean structure $m_1(x)$, error terms from a $N(0, 0.7^2)$ distribution, and for different sample sizes. We consider different percentages ε of outliers generated from $N(20, 2^2)$.*

ε	$n = 25$			$n = 100$			$n = 250$		
	LS	*M*	*S*	*LS*	*M*	*S*	*LS*	*M*	*S*
0%	0.07	0.08	0.18	0.02	0.03	0.07	0.01	0.01	0.04
	(0.05)	(0.05)	(0.13)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.02)
5%	2.31	0.09	0.21	1.57	0.03	0.08	1.35	0.02	0.04
	(3.25)	(0.07)	(0.17)	(1.19)	(0.02)	(0.05)	(0.69)	(0.01)	(0.02)
10%	7.48	0.12	0.21	5.12	0.06	0.07	4.56	0.04	0.03
	(7.07)	(0.09)	(0.17)	(2.84)	(0.03)	(0.05)	(1.73)	(0.02)	(0.02)
20%	22.9	0.44	0.24	18.5	0.20	0.06	16.9	0.17	0.03
	(16.4)	(0.42)	(0.22)	(7.01)	(0.09)	(0.04)	(4.18)	(0.05)	(0.02)
30%	45.0	3.95	0.35	38.8	0.97	0.05	37.3	0.79	0.02
	(24.5)	(5.47)	(0.42)	(12.0)	(0.47)	(0.03)	(7.06)	(0.20)	(0.01)
40%	75.6	70.2	32.5	66.8	36.4	0.07	66.0	7.62	0.02
	(34.7)	(18.2)	(48.0)	(16.8)	(40.6)	(0.06)	(9.74)	(4.69)	(0.02)

To give an impression on the variability of the obtained estimators, we plot the box plots of log scale of ASEs of the simulation samples from penalized least squares, penalized M- and S-regression spline estimation in Figures 2.2 and 2.3 for the data with outliers $N(20, 2^2)$ and $N(20, 20^2)$ respectively. These plots show that the ASEs of the penalized S-regression spline estimator remain stable as the proportion of contamination increases. Even though they become more variable for 40% of outliers, the median is still at the same level as before. The penalized LS-estimator's ASEs grow very rapidly. Similarly, the penalized M-regression spline estimator's ASEs grow rapidly after 10% of outliers. These results are confirming that the penalized M-regression spline estimation method works better with less than

10% of outliers, while the penalized S-regression spline estimation method works well for all considered percentages of outliers.

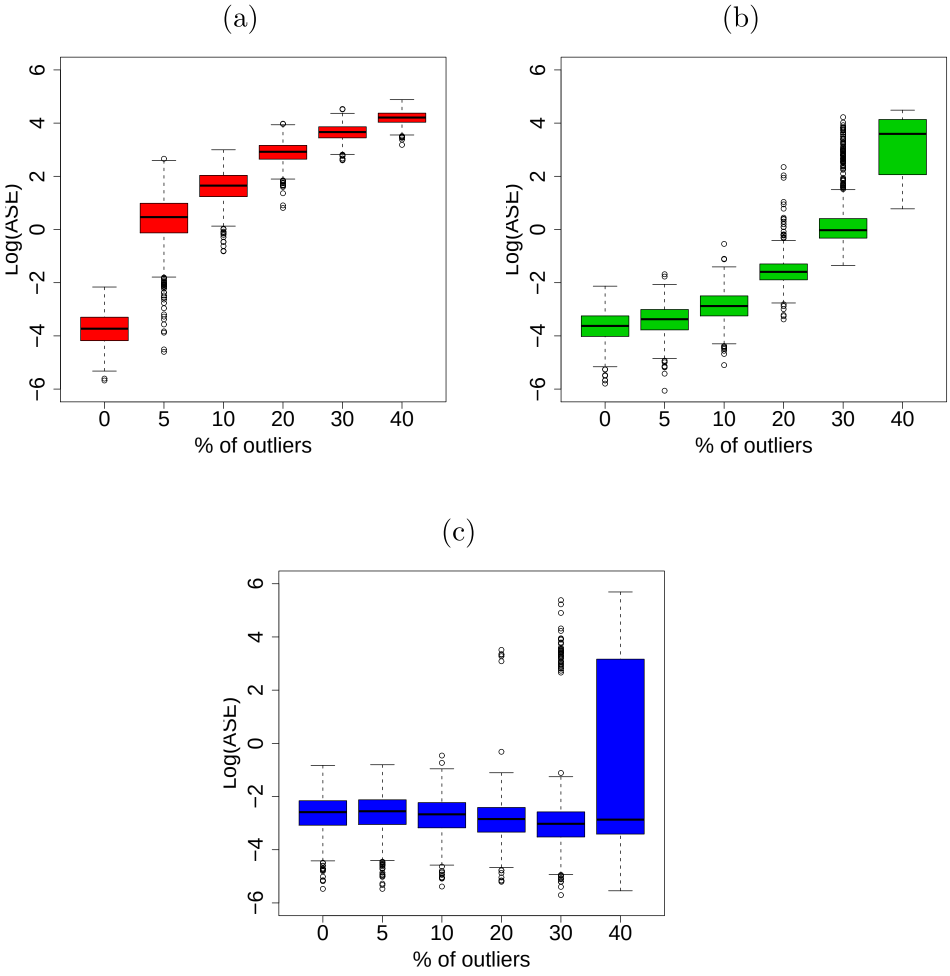


Figure 2.2: *Box plots of ASEs using (a) penalized LS-estimation, (b) penalized M-regression spline estimation and (c) penalized S-regression spline estimation for samples with mean structure $m_1(x)$, error distribution $N(0, 0.7^2)$ and outliers $N(20, 2^2)$, for sample size $n = 100$.*

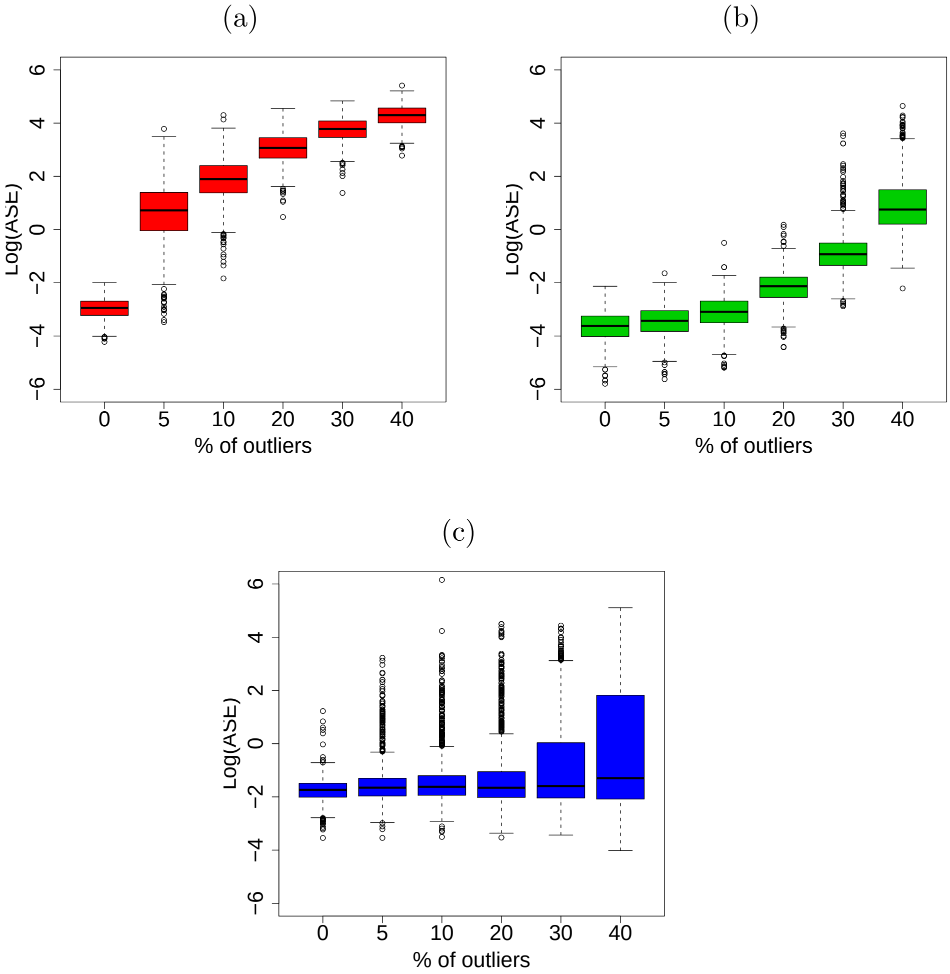


Figure 2.3: *Box plots of ASEs using (a) penalized LS-estimation, (b) penalized M-regression spline estimation and (c) penalized S-regression spline estimation for samples with mean structure $m_1(x)$, error distribution $N(0, 0.7^2)$ and scattered outliers $N(20, 20^2)$, for sample size $n = 100$.*

Table 2.2: *Median and median absolute deviation (between parenthesis) of the average squared error ASE for penalized least squares (LS), penalized M (M) and penalized S (S) regression spline estimation for data generated with mean structure $m_1(x)$, error terms from different distributions Uniform, Logistic, Slash and Cauchy, and for sample sizes $n = 100$ with different percentages ε of outliers generated from $N(20, 2^2)$.*

ε		0%	5%	10%	20%	30%	40%
Uniform	*LS*	0.02	1.62	5.27	18.4	39.4	68.3
		(0.0)	(1.2)	(2.9)	(7.0)	(11)	(15)
	M	0.02	0.03	0.04	0.16	0.78	30.9
		(0.0)	(0.0)	(0.0)	(0.1)	(0.3)	(38)
	S	0.15	0.15	0.13	0.10	0.06	0.07
		(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.1)
Logistic	*LS*	0.14	1.75	5.44	18.5	39.5	68.6
		(0.1)	(1.3)	(3.0)	(7.0)	(11)	(16)
	M	0.16	0.21	0.34	1.24	6.15	59.6
		(0.1)	(0.1)	(0.2)	(0.6)	(3.2)	(12)
	S	0.34	0.33	0.31	0.28	0.26	0.49
		(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.6)
Slash	*LS*	6.69	8.95	12.8	25.5	46.1	73.4
		(8.4)	(9.8)	(11)	(16)	(22)	(27)
	M	0.37	0.51	0.87	3.66	24.0	73.4
		(0.2)	(0.3)	(0.6)	(2.2)	(15.2)	(8)
	S	0.32	0.32	0.31	0.31	0.42	56.5
		(0.2)	(0.2)	(0.2)	(0.2)	(0.4)	(83)
Cauchy	*LS*	4.92	6.71	10.3	23.5	45.1	73.3
		(6.1)	(7.2)	(8.8)	(13)	(19)	(25)
	M	0.19	0.26	0.45	2.05	14.1	71.1
		(0.1)	(0.2)	(0.3)	(1.3)	(11)	(9)
	S	0.12	0.12	0.12	0.14	0.21	33.3
		(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(49)

Next we compare the effects of the different error distributions on the performance of the estimates. The results are shown in Table 2.2 for sample size $n = 100$ and true mean function m_1. The proposed method gives the smallest median ASE values for all considered error distributions if there are more than 20% outliers. Penalized M-regression spline estimation works better for the samples with 5% and 10% of outliers for uniform and logistic error distributions. For penalized LS- and M-regression spline estimation methods, the ASE values are relatively large for heavy-tailed distributions (Slash and Cauchy). Note that in absence of outliers (ε=0%) the method of penalized S-regression spline estimation works better than LS at heavy tailed distributions.

Table 2.3: *Median and median absolute deviation (between parenthesis) of the average squared error ASE for penalized least squares (LS), penalized M (M) regression spline and penalized S (S) regression spline estimation for data generated from functions m_2, m_3 and m_4 with error terms from $N(0, 0.7^2)$ for sample size $n = 100$ with different percentages ε of outliers generated from $N(20, 2^2)$.*

ε%		m_2			m_3			m_4	
	LS	*M*	*S*	*LS*	*M*	*S*	*LS*	*M*	*S*
0	0.25	0.27	0.38	0.02	0.03	0.07	0.04	0.04	0.14
	(0.02)	(0.02)	(0.11)	(0.01)	(0.02)	(0.05)	(0.02)	(0.02)	(0.07)
5	1.87	0.28	0.40	1.47	0.03	0.08	5.30	0.05	0.13
	(1.17)	(0.03)	(0.13)	(1.09)	(0.02)	(0.05)	(4.00)	(0.02)	(0.07)
10	5.44	0.32	0.37	4.83	0.06	0.07	17.3	0.07	0.12
	(2.82)	(0.05)	(0.13)	(2.73)	(0.03)	(0.05)	(9.88)	(0.03)	(0.07)
20	18.6	0.57	0.33	17.4	0.20	0.06	63.0	0.22	0.11
	(7.08)	(0.14)	(0.12)	(6.67)	(0.09)	(0.04)	(24.2)	(0.09)	(0.06)
30	38.5	1.89	0.31	36.9	0.96	0.05	133	1.02	0.11
	(11.8)	(0.85)	(0.12)	(11.2)	(0.46)	(0.04)	(38.0)	(0.45)	(0.06)
40	66.4	47.3	0.38	63.5	30.7	0.06	229	51.1	0.11
	(16.2)	(29.6)	(0.25)	(15.4)	(36.4)	(0.06)	(54.6)	(69.2)	(0.08)

We have further checked our proposed method with that of Lee and Oh (2007) using the same regression function m_2 as in their paper. We generated errors ε_i from a normal distribution, and included different percentages of outliers for sample size $n = 100$. For each of these settings we computed the ASE over 1000 simulation runs; the results are presented in Table 2.3. All previous findings are confirmed. The S-regression spline estimation method does a better job than penalized M-regression spline estimation when there are 20% of outliers or more. The penalized M-regression spline estimation method works better for the cases with 5% and 10% of outliers. This holds for the goniometric (m_2), the polynomial (m_3), and the exponential (m_4) mean functions.

For completeness we here present the results of an additional simulation study, we used an S-estimator with 25% breakdown point and we observed that the efficiency of the proposed method is higher in the absence of outliers, but it is lower than that of penalized LS- and M-regression spline estimators.

Table 2.4: *Median and median absolute deviation (between parenthesis) of the average squared error ASE for penalized least squares (LS), penalized M (M) regression spline estimation and penalized S (S) regression spline estimation with 50% and 25% breakdown point for data generated with mean structure* $\sin(\pi x)$*, error terms from a* $N(0, 0.7^2)$ *distribution and sample size* $n = 100$ *with different percentages* ε *of outliers generated from* $N(20, 2^2)$*.*

ε			50% Breakdown point	25% Breakdown point
	LS	*M*	*S*	*S*
0%	0.02	0.03	0.08	0.04
	(0.01)	(0.01)	(0.06)	(0.02)
5%	1.62	0.03	0.08	0.04
	(0.48)	(0.02)	(0.05)	(0.02)
10%	5.18	0.06	0.08	0.04
	(0.99)	(0.03)	(0.05)	(0.02)
20%	18.60	0.21	0.06	0.04
	(2.41)	(0.09)	(0.04)	(0.02)
30%	38.86	1.02	0.05	12.49
	(2.77)	(0.51)	(0.04)	(3.99)
40%	67.97	46.43	0.04	53.38
	(3.54)	(28.48)	(0.03)	(5.80)

Table 2.4 shows for one of the simulation settings the results for penalized S-regression estimation using Tukey's bi-square ρ function with first a

50% breakdown point ($d = 1.5476$) and next with a 25% breakdown point ($d = 2.937$).

Figure 2.4 shows the box plots of average squared error (ASE) for the same setting and estimators. As we expect, the average squared errors are lower for the case of the 25% breakdown point than for the 50% breakdown point case. That is, to increase the efficiency of the penalized S-regression spline estimator one needs to lower its breakdown point. The price to pay for this increase of efficiency in absence of outliers (by taking a lower breakdown point) is a decrease of the robustness. As can be seen from Table 2.4, the S-estimator with 25% breakdown point has an large bias if one has large amounts of outliers (30% or 40%).

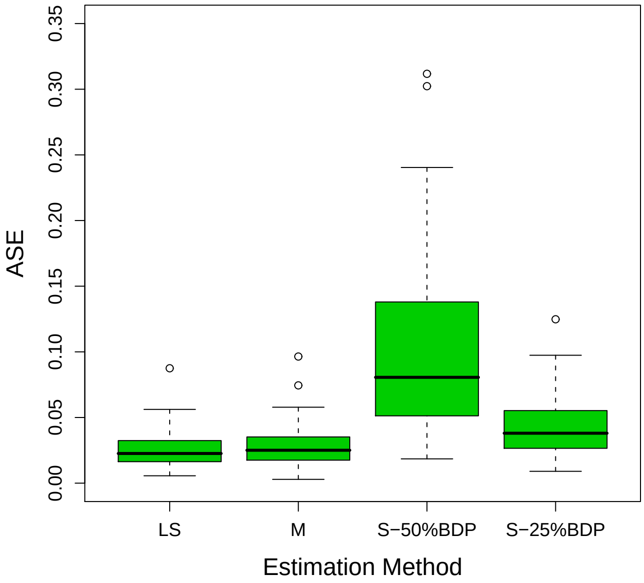


Figure 2.4: *ASE for penalized least squares (LS), penalized M (M) regression spline estimation and penalized S-regression spline estimation with 50% (S-50%BDP) and 25% (S-25%BDP) breakdown point for data with no outliers.*

In addition we did the simulations with a different ρ function. Table 2.5 shows one of the simulation setting results for penalized S-regression estimator with different ρ functions. We defined

$$\rho_1(u) = \begin{cases} 4\,(u/d)^2 - 3\,(u/d)^4 & \text{if } |u| \leq d\,, \\ 1 & \text{if } |u| > d. \end{cases} \tag{2.19}$$

The choice $d = 0.57735$ yields $b = E_\Phi\left[\rho_1\left(Z\right)\right] = 0.50$. We compared this to the results that we earlier obtained when using a ρ function from Tukey's bi-square family. Table 2.5 illustrates that the efficiency of the penalized S-regression spline estimator does not change significantly with respect to different ρ functions.

Table 2.5: *Median and median absolute deviation (between parenthesis) of the average squared error ASE for penalized least squares (LS), penalized M (M) regression spline estimation and penalized S (S) regression spline estimation using Tukey's bi-square family ρ-function and ρ_1-function with 50% breakdown point for data generated with mean structure* $\sin(\pi x)$*, error terms from a* $N(0, 0.7^2)$ *distribution and sample size* $n = 100$ *with different percentages* ε *of outliers generated from* $N(20, 2^2)$*.*

ε			Tukey's bi-square family ρ-function	ρ_1-function
	LS	*M*	*S*	*S*
0%	0.02	0.03	0.08	0.07
	(0.01)	(0.01)	(0.06)	(0.04)
5%	1.62	0.03	0.08	0.07
	(0.48)	(0.02)	(0.05)	(0.05)
10%	5.18	0.06	0.08	0.07
	(0.99)	(0.03)	(0.05)	(0.05)
20%	18.60	0.21	0.06	0.08
	(2.41)	(0.09)	(0.04)	(0.05)
30%	38.86	1.02	0.05	0.09
	(2.77)	(0.51)	(0.04)	(0.04)
40%	67.97	46.43	0.04	0.12
	(3.54)	(28.48)	(0.03)	(0.07)

In an additional simulation study we have tried with different choices of K as well. We selected the number of knots (num.knots) as follows: we took either the number of unique x-values divided by four (and rounded downwards), or 35, whichever value was the smallest. In case that number was smaller than 5, we would use 5 knots. We checked our results with a large number of knots (two times the number of knots) in a simulation study. We present in Table 2.6 the summary results from 100 simulation samples, the same as in the previous case, but with a double number of knots.

Table 2.6: *Median and Median absolute deviation (between parenthesis) of the average squared error ASE for penalized least squares (LS), penalized M (M) regression spline estimation and penalized S (S) regression spline estimation using different number of knots for data generated with mean structure* $\sin(\pi x)$*, error terms from a* $N(0, 0.7^2)$ *distribution and sample size* $n = 100$ *with different percentages* ε *of outliers generated from* $N(20, 2^2)$*.*

ε	num.knots			2*num.knots		
	LS	*M*	*S*	*LS*	*M*	*S*
0%	0.02	0.03	0.08	0.02	0.03	0.08
	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	0.06
5%	1.62	0.03	0.08	1.63	0.03	0.09
	(0.48)	(0.02)	(0.05)	(0.49)	(0.02)	(0.06)
10%	5.18	0.06	0.08	5.20	0.06	0.08
	(0.99)	(0.03)	(0.05)	(1.01)	(0.03)	0.05
20%	18.60	0.21	0.06	18.66	0.21	0.07
	(2.41)	(0.09)	(0.04)	(2.39)	(0.09)	(0.04)
30%	38.86	1.02	0.05	38.93	1.13	0.06
	(2.77)	(0.51)	(0.04)	(2.79)	(0.61)	(0.04)
40%	67.97	46.43	0.04	68.20	47.83	0.04
	(3.54)	(28.48)	(0.03)	(3.76)	(27.96)	(0.03)

Table 2.6 shows similar results for both cases, hardly any differences are observed. The penalized S-estimator is quite insensitive with respect to the number of knots.

2.4 Balloon data

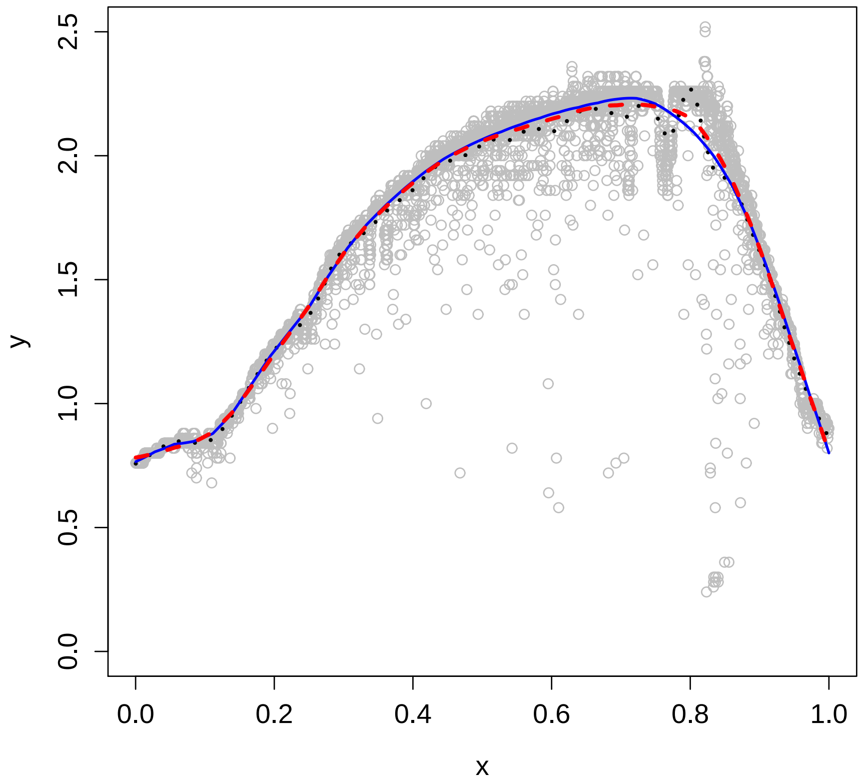


Figure 2.5: *Fitted values for the balloon data. Penalized LS-regression spline method (dotted), penalized S-regression spline method (solid) and penalized M-regression spline method (dashed).*

In this section, we have used the balloon data set from the R software's library `ftnonpar`. The data are radiation measurements from the sun, taken from a flight of a weather balloon. Due to the rotation of the balloon, or for some other reasons, outliers were introduced because the measuring device was occasionally blocked from the sun. The response variable Y is a radiation measurement and the explanatory variable x is the index of the measurement. The sample size equals 4,984. We took $K = 35$ knots spread equally, and scaled the value λ according to the GCV, RCV and RGCV methods, described in section 2.2.3. We obtained $\lambda = 0.04$ for penalized LS-estimation method and $\lambda = 0.1$ for penalized M- and S-regression spline estimation method.

Displayed in Figure 2.5 are regression estimates obtained by the penalized LS method, our proposed method of penalized S-regression spline estimation and penalized M-regression spline estimation. The non-robust curve suffers from the presence of the outliers, which is clearly visible around the value $x = 0.8$. That is, the estimated curve was pulled upwards, in the direction of the outliers. The robust methods do not suffer from this phenomenon.

2.5 Discussion

In this chapter a simple and effective method is proposed for robust fitting penalized regression spline models. Generally, smoothing methods may be influenced by outliers. The proposed method is easy to implement and fast to converge. Penalized S-regression spline estimators improve on penalized least squares regression splines and penalized M-regression spline estimators. The procedure performs very well in all of our numerical examples. The penalized M-regression spline estimation works better for the cases with a small percentage of contamination but penalized S-regression spline estimation works well for higher percentage of contamination too.

In the absence of outliers, the efficiency of the proposed method is not very high. This is the price to pay for a high robustness. To increase the efficiency of an S-estimator, we need to lower its breakdown point. In an additional simulation study (results shown in Table 2.4, section 2.3.2) we used an S-estimator with 25% breakdown point and we observed that the efficiency of the proposed method is higher in the absence of outliers, but it is lower than that of penalized LS- and M-regression spline estimators. Changing the ρ function will not significantly increase the efficiency. This is known from robust regression analysis (see Hössjer, 1992) where it has been shown that the highest possible Gaussian efficiency of an S-estimator with the highest possible value for the breakdown point is about 33%. The efficiency of the biweight loss function (leading to Tukey's bi-square ρ function) is close to this maximal value.

The asymptotic properties of penalized S-regression splines have not yet been studied, and are a topic of our further research. We expect that consistency and asymptotic normality still hold, under appropriate regularity conditions. These results would be useful in order to construct confidence bands for the curves, for example.

Chapter 3

A comparison of robust versions of the AIC based on M, S and MM-estimators

This chapter is based on the following publication:
Tharmaratnam, K. and Claeskens, G. (2011a). A comparison of robust versions of the AIC based on M, S and MM-estimators. *Statistics*, in press.

Abstract

Variable selection in the presence of outliers may be performed by using a robust version of Akaike's information criterion AIC. In this chapter explicit expressions are obtained for such criteria when S- and MM-estimators are used. The performance of these criteria is compared to the existing AIC based on M-estimators and to the classical non-robust AIC. In a simulation study and in data examples we observe that the proposed AIC with S- and MM-estimators selects more appropriate models in case outliers are present.

3.1 Introduction

It has been recognized that variable selection procedures need special care in the presence of outliers in the data. Since most of the classical procedures are likelihood-based, alternatives have been developed. Some of the main developments to make classical model selection procedures for linear models less sensitive to outlying observations are a robust version of Akaike's information criterion (AIC Akaike, 1973) based on M estimators (Ronchetti, 1985), a robust C_p (Ronchetti and Staudte, 1994; Sommer and Staudte, 1995), a robust version of cross-validation (Ronchetti et al., 1997), see also the survey presented in Ronchetti (1997). Qian and Künsch (1998) select models in a robust way using the concept of stochastic complexity, while Agostinelli (2002) rather deals with weighted versions of likelihood estimators. Several of these model selection methods are described in Maronna et al. (2006, Sec. 5.12) and Claeskens and Hjort (2008, Ch. 2 and 4). Müller and Welsh (2005) make use of the bootstrap to combine a robust penalized criterion with a robust conditional expected prediction loss function. Other use of the bootstrap for robust variable selection is made by Salibián-Barrera and Van Aelst (2008). Heritier et al. (2009, p. 159) present a form of the AIC based on robust quasilikelihood.

While the emphasis in the existing literature is mostly on M-estimation when it comes to variable selection methods, in this chapter we investigate whether improvements can be achieved when using S- or MM-estimators. The derivation of information criteria in the style of the AIC using these robust estimators is in the line with the generalized information criteria of Konishi and Kitagawa (1996). When applied to estimation in likelihood models free of outliers, this approach would lead to Takeuchi's information criterion (Takeuchi, 1976), which differs from the traditional AIC only in its penalty term.

The rest of this chapter is organized as follows. The formula for the robust version of AIC based on general M-estimators is derived in Section 3.2. In Section 3.3 we provide a version of the AIC for use with robust estimators of scale, which require separate attention. Some extensions to the use of MM-estimators and towards using uniform asymptotic expressions (Omelka and Salibián-Barrera, 2010) are contained in Section 3.4. Section 3.5 reports the results of a simulation study and data examples that compare the performance of classical AIC, AIC based on M, S and MM-estimation. Finally, Section 3.6 contains a discussion and concluding remarks. The appendix contains the R code that is used for the calculations.

3.2 AIC for use with robust M-estimation methods

3.2.1 AIC for linear regression models

We consider the linear regression model

$$Y_i = \theta_0^t X_i + u_i, \qquad i = 1, \dots, n, \tag{3.1}$$

where the response variables $Y_i \in \mathbb{R}$ $(i = 1, \dots, n)$ are independent, the covariate vector $X_i \in \mathbb{R}^p$ with a corresponding coefficient vector $\theta_0 \in \mathbb{R}^p$ and the u_i are random errors independent from the explanatory variable X_i, with mean zero and constant variance σ^2. For normal errors with standard deviation σ, the Akaike information criterion for variable selection is given by

$$\text{AIC} = 2n \log \widehat{\sigma} + 2(p+1) + \{n + n\log(2\pi)\}, \tag{3.2}$$

where the last term, $\{n + n\log(2\pi)\}$, may be omitted because it is independent of the choice of the variables in the model and where $\widehat{\sigma}$ is the maximum likelihood estimate of σ. The penalty takes the p regression coefficients θ_0 and the unknown error variance into account.

In general, Akaike's information criterion is in full likelihood models defined as AIC $= -2$ log-likelihood$(\widehat{\theta}) + 2\times$ length(θ), with length(θ) the number of parameters that are estimated in the model, and with $\widehat{\theta}$ the maximum likelihood estimator of the model parameters θ. The AIC arises as an estimator of the expected value of the Kullback-Leibler distance between the maximized density of the data implied by the model and the true density g, that is nearly always unknown,

$$\mathrm{KL}(g, f(.,\widehat{\theta})) = \int\int g(y|x)\log g(y|x)dydG(x) - R_n$$

where $R_n = \int\int g(y|x)\log f(y|x,\widehat{\theta})dydG(x)$ and G is the cumulative distribution function of X. A derivation of the traditional AIC can for example be found in Claeskens and Hjort (2008, Sec. 2.3).

Since the AIC is likelihood-based, and thus is sensitive to outlying observations in the data, we here search for more robust alternatives, in the spirit of the generalized information criterion of Konishi and Kitagawa (1996). In the case that outliers are present in the data, only the majority of the data follows the above model (3.1). Extreme observations might occur in both the explanatory variables and the response. It is in these circumstances that we wish to investigate the inclusion or exclusion of components of the covariate vector X.

3.2.2 M-estimators

As a robust alternative to maximum likelihood estimators, M-estimators are used. Huber (1964) defined a general M-estimator as the minimum with respect to θ of the objective function $\sum_{i=1}^{n}\rho(y_i|x_i,\theta)$, for a given function ρ that has the properties of being even, non-decreasing in $[0,\infty)$ and with $\rho(0)=0$. Equivalently, when the response values $Y_1,\ldots,Y_n$ are independent, the M-estimator for θ solves the equation

$$\sum_{i=1}^{n}\psi(Y_i|x_i,\theta) = 0 \tag{3.3}$$

where $\psi(y|x,\theta) = \frac{\partial\rho(y|x,\theta)}{\partial\theta}$. Intuitively, to take care of outliers which result in large residuals, the function $\rho(\cdot)$ should less increase than the squared function, particularly for large residuals. A common choice for ρ is given by Huber's family with an unbounded loss function

$$\rho_c(t) = \begin{cases} t^2 & \text{if } |t| \le c \\ 2\,c\,|t| - c^2 & \text{if } |t| > c\,, \end{cases} \tag{3.4}$$

where $c > 0$ is a tuning constant that can be thought of as a threshold value such that observations with residuals larger than c have a reduced effect in the estimating equation (3.3). A value of 95% asymptotic efficiency on the standard normal distribution is obtained when the constant equals 1.345 (Huber, 2004). In practice, a typical choice for c is $1.345\,\widehat{\sigma}_m$, with $\widehat{\sigma}_m$ the median absolute deviation (MAD) of the residuals, $\text{MAD}(r_1,\ldots,r_n) = 1.4826\,\text{median}_{i=1,\ldots,n}(|r_i|)$ (with the constant 1.4826 based on the normality assumption). The M-estimator is computed with $\rho(y_i|x_i,\theta) = \rho_c\left(\frac{y_i-\theta^t x_i}{\widehat{\sigma}_m}\right)$. The implementation of M-estimators uses an iteratively reweighted least squares algorithm.

3.2.3 Derivation of a robust AIC

Instead of working with the maximized likelihood function in the Kullback-Leibler distance, we use the loss function ρ and the corresponding robust estimator $\widehat{\theta}$ and consider as a good model one that minimizes the expected value of the following weighted Kullback-Leibler distance that involves the empirical distribution of the covariates,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\int g(y|x_i)\{\log g(y|x_i) + \rho(y|x_i,\widehat{\theta})\}dy\\&\quad = \frac{1}{n}\sum_{i=1}^{n}\int g(y|x_i)\log g(y|x_i)dy + R_n^{\rho},\end{aligned}\qquad (3.5)$$

where $R_n^{\rho} = \frac{1}{n}\sum_{i=1}^{n}\int g(y|x_i)\rho(y|x_i,\widehat{\theta})dy$. In the next section we make this more concrete for the different robust estimators. For M-estimators, such a robust AIC with the scale assumed to be known (and later estimated from the largest model) has been obtained by Ronchetti (1997).

Since the first term is independent of the model, the key quantity to study is R_n^{ρ}, which depends on the data through the robust estimator $\widehat{\theta}$. The expected value of R_n^{ρ} with respect to the robust estimator, under the true density g for the response variable Y_i given the covariate is equal to

$$Q_n = E(R_n^{\rho}) = \frac{1}{n}\sum_{i=1}^{n}E\left[\int g(y|x_i)\rho(y|x_i,\widehat{\theta})dy\right], \qquad (3.6)$$

which is estimated by replacing the true distribution functions by their empirical counterparts, leading to the estimator

$$\widehat{Q}_n = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i|x_i, \widehat{\theta}).$$

For maximum likelihood estimation, $\widehat{Q}_n$ corresponds to the minus log likelihood function, evaluated at the maximum likelihood estimator, divided by the sample size. To construct an AIC, we investigate the bias of $\widehat{Q}_n$ for estimation of Q_n, which will lead to an appropriate penalty term in the variable selection criterion.

Define by $\theta_{0,n}$ the least false parameter vector that corresponds to the empirical distribution of the covariates and thus maximizes $\mathcal{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\int g(y|x_i)\rho(y|x_i,\theta)dy$. Denote $Q_{0,n} = \mathcal{Q}_n(\theta_{0,n})$, $V_n = \sqrt{n}(\widehat{\theta} - \theta_{0,n})$ and $J_n = -\frac{1}{n}\sum_{i=1}^{n}\int g(y|x_i)I(y|x_i,\theta_{0,n})dy$, with information function $I(y|x,\theta) = -\frac{\partial^2\rho(y|x,\theta)}{\partial\theta\partial\theta^t}$. The score function is defined as $u(y|x,\theta) = -\frac{\partial\rho(y|x,\theta)}{\partial\theta}$, with variance $K_n = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\{u(Y|x_i,\theta_{0,n})\}$. The limit versions of J_n and K_n are denoted by J and K, respectively.

Result 3.1. Let $\bar{Z}_n$ be the average of the values $Z_i = -\rho(Y_i|x_i,\theta_{0,n}) + \int g(y|x_i)\rho(y|x_i,\theta_{0,n})dy$, assume that ρ is two times differentiable, and using the notation as defined above,

$$\widehat{Q}_n - R_n^{\rho} = -\bar{Z}_n - \frac{1}{n}V_n^t J V_n + o_p(1/n). \tag{3.7}$$

Proof. A Taylor expansion for R_n^{ρ} gives that

$$\begin{aligned} R_n^{\rho} &= \frac{1}{n}\sum_{i=1}^{n}\int \Big\{ g(y|x_i)\Big[\rho(y|x_i,\theta_{0,n}) - u(y|x_i,\theta_{0,n})(\widehat{\theta} - \theta_{0,n}) \\ &\quad -\frac{1}{2}(\widehat{\theta} - \theta_{0,n})^t I(y|x_i,\theta_{0,n})(\widehat{\theta} - \theta_{0,n}) + o_P(1/n)\Big]\Big\} dy \\ &= Q_{0,n} + \frac{1}{2n}V_n^t J_n V_n + o_P(1/n). \end{aligned}$$

In a similar fashion, a Taylor expansion for $\widehat{Q}_n$ results in

$$\begin{aligned}\widehat{Q}_n &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\rho(Y_i|x_i,\theta_{0,n}) - u(Y_i|x_i,\theta_{0,n})(\widehat{\theta}-\theta_{0,n}) \\ &\quad -\frac{1}{2}(\widehat{\theta}-\theta_{0,n})^t I(Y_i|x_i,\theta_{0,n})(\widehat{\theta}-\theta_{0,n})\Big\} \\ &\quad + o_P(1/n) = Q_{0,n} - \bar{Z}_n - \frac{1}{2n}V_n^t J_n V_n + o_P(1/n).\end{aligned}$$

Thus, it holds that $\widehat{Q}_n - R_n^\rho = -\bar{Z}_n - \frac{1}{n}V_n^t J_n V_n + o_P(1/n)$. □

From (3.7) and since for robust estimators it holds that $V_n \xrightarrow{d} N(0, J^{-1}KJ^{-1})$, it follows that $E(\widehat{Q}_n - Q_n)$ is approximately (leaving out remainder terms of smaller order) equal to $-\text{Trace}(J^{-1}K)/n$.

3.2.4 AIC for M-estimation

Based on the results of Section 3.2.3, a model selection criterion in the style of Akaike's information criterion is to compute $\widehat{Q}_n + \text{Trace}(J_n^{-1}K_n)/n$ for each candidate model, and then to select the model with the smallest such value. Equivalently, we define a robust AIC, specific to the loss function leading to different robust estimators,

$$\text{AIC}_\rho = 2\sum_{i=1}^{n}\rho(Y_i|x_i,\widehat{\theta}) + 2\ \text{Trace}(J_n^{-1}K_n) \qquad (3.8)$$

and select that model which has the smallest AIC_ρ value.

In the equation above, the vector θ represents *all* unknown parameters in the model, thus including the unknown σ. This implies that the information matrices J_n and K_n have dimension $(p+1)\times(p+1)$ and partial derivatives are computed with respect to all $p+1$ unknown parameters.

A slightly simpler version is presented by Ronchetti (1997), when considering the case of a known σ (and afterwards plugging in an estimate from the largest model). His robust AIC for M-estimators which fits within the form (3.8).

More in line with the application of the AIC for use with maximum likelihood estimation, all parameters are re-estimated in each model, which implies that the scale estimator also changes from model to model. This leads to defining

$$\mathrm{AIC}_\rho.\mathrm{M} = 2\sum_{i=1}^{n} \rho_c \left(\frac{Y_i - \widehat{\theta}_m^t x_i}{\widehat{\sigma}_m} \right) + 2\ \mathrm{Trace}(J_{m,n}^{-1} K_{m,n}), \tag{3.9}$$

where ρ_c is, for example, the Huber loss function as in (3.4). In this equation the empirical information matrices $J_{m,n}$ and $K_{m,n}$ both have dimension $p \times p$ and partial derivatives are only calculated with respect to the regression parameters in the location part of the model. As requested by a referee, we will use this simpler version of the AIC in the simulation study and data analysis. For a better comparison, we will hence make such a simplification for the other considered criteria as well. Using the full information matrices (with dimension $(p+1)\times(p+1)$) is computationally a bit more involved but turns out, at least for the considered datasets, not to make much difference with respect to variable selection. Simulation results are shown in Table 3.9 in section 3.5.2.

3.3 AIC for use with robust estimators for scale

3.3.1 S-estimators

S-estimators for linear regression were introduced by Rousseeuw and Yohai (1984) as an alternative to M-estimators that do not suffer that much from leverage points (which are outliers in the covariates) and at the same time have a high breakdown point and do not require an auxiliary scale estimator.

The S-estimator $\widehat{\theta}_s$ minimizes the scale function, that is, $\widehat{\theta}_s = \mathrm{argmin}_{\theta \in \mathbb{R}^p}\, \widehat{\sigma}_n(\theta)$, where the scale function $\widehat{\sigma}_n(\theta)$ is implicitly defined by that function of θ that satisfies the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0 \left(\frac{y_i - \theta^t x_i}{\widehat{\sigma}_n(\theta)} \right) = b, \tag{3.10}$$

with $\rho(y_i|x_i,\theta) = \rho_0\left(\frac{y_i-\theta^t x_i}{\widehat{\sigma}_n(\theta)}\right)$. The scale estimator is $\widehat{\sigma}_s = \widehat{\sigma}_n(\widehat{\theta}_s)$. The loss function ρ_0 is a function that is even, continuously differentiable, non-decreasing on $[0, \infty)$, satisfies that $\rho_0(0) = 0$ and has $\sup_{u \in \mathbb{R}} \rho_0(u) = 1$. We define $b = E_{F_0}[\rho_0(u_1)]$, with u_1 one of the error terms in model (3.1) with cumulative distribution function F_0, and assume that $0 < \epsilon_0 < b < 1 - \epsilon_0$ to ensure consistency of the scale estimator under the central model F_0. The notation E_{F_0} means that the expectation is computed with respect to F_0.

A commonly used family of loss functions ρ_0 is given by Tukey's bisquare family (Beaton and Tukey, 1974)

$$\rho(u; d) = \begin{cases} 3\,(u/d)^2 - 3\,(u/d)^4 + (u/d)^6 & \text{if } |u| \le d\,, \\ 1 & \text{if } |u| > d\,. \end{cases} \tag{3.11}$$

The choice $d = 1.5476$ yields $b = E_\Phi\left[\rho\left(Z; d\right)\right] = 0.5$, with Φ the standard normal cumulative distribution function and $Z \sim N(0, 1)$. The associated S-regression estimator has maximal asymptotic breakdown point 50% (Rousseeuw and Yohai, 1984). Estimators with 30% breakdown point are gotten when $d = 2.5608$, resulting in a higher efficiency. Both options are contrasted in the simulation study.

3.3.2 AIC for S-estimation

For S-estimators the above approach for obtaining an AIC as in (3.8) does not work because of the constraint (3.10). Indeed, when substituting S-estimators on the right hand side of (3.8) this gives as a first term $2\sum_{i=1}^{n} \rho(Y_i|x_i, \widehat{\theta}_s) = 2nb$, which is a constant for all models and thus does not differentiate between different models. Therefore, based on (3.2), we propose a robust AIC with respect to S-estimation of the following form

$$\text{AIC.S} = 2\,n \log(\widehat{\sigma}_s) + 2\ \text{Trace}(J_{s,n}^{-1} K_{s,n}). \tag{3.12}$$

In this criterion we use the robust S-scale estimator $\widehat{\sigma}_s$ and take possible model misspecification into account by the form of the penalty term (rather than just counting the number of parameters). The empirical information matrices $J_{s,n}$ and $K_{s,n}$ (when considering partial derivatives with respect to θ) are defined as follows,

$$J_{s,n} = \frac{1}{n}\sum_{i=1}^{n} \rho_d'' \left(\frac{y_i - \widehat{\theta}_s^t x_i}{\widehat{\sigma}_s} \right) \frac{x_i x_i^t}{\widehat{\sigma}_s^2} \text{ and } K_{s,n} = \frac{1}{n}\sum_{i=1}^{n} \rho_d'^2 \left(\frac{y_i - \widehat{\theta}_s^t x_i}{\widehat{\sigma}_s} \right) \frac{x_i x_i^t}{\widehat{\sigma}_s^2}.$$

Model selection proceeds by computing AIC.S for all models under consideration and by selecting the model with the smallest value of AIC.S.

When $\rho(t) = t^2$, this criterion reduces to Takeuchi's information criterion TIC (Takeuchi, 1976) for normal data.

3.4 Extensions

3.4.1 AIC for use with MM-estimators

A further step in robust estimation uses the S-scale estimator in an M-estimating equation. Let $\rho_1 : \mathbb{R} \to \mathbb{R}_+$ be a loss function such that $\rho_1(u) \leq \rho_0(u)$ for all $u \in \mathbb{R}$ and $\sup_u \rho_1(u) = \sup_u \rho_0(u)$. The MM-regression estimator $\widehat{\theta}_{mm}$ is defined as the global minimum of $f : \mathbb{R}^p \to \mathbb{R}_+$, with

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho_1 \left(\frac{y_i - \theta^t x_i}{\widehat{\sigma}_s} \right).$$

Thus,

$$\widehat{\theta}_{mm} = \operatorname*{argmin}_{\|\theta\| \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho_1 \left(\frac{y_i - \theta^t x_i}{\widehat{\sigma}_s} \right).$$

Since MM-estimators are M-estimators, the following form of a robust AIC version is obtained in a similar fashion as in Section 3.2.4,

$$\text{AIC}_\rho.\text{MM} = 2\sum_{i=1}^{n} \tilde{\rho}_d \left(\frac{Y_i - \widehat{\theta}_{mm}^t x_i}{\widehat{\sigma}_{mm}} \right) + 2 \text{ Trace}(J_{mm,n}^{-1} K_{mm,n}). \tag{3.13}$$

where $\widehat{\omega}_{mm} = (\widehat{\theta}_{mm},\widehat{\sigma}_{mm})$, $\widehat{\theta}_{mm}$ and $\widehat{\sigma}_{mm}$ are MM-estimators, with empirical information matrices gotten from the corresponding expressions for S-estimators (see Section 3.3.2) by replacing ρ_0 by $\tilde{\rho}_d$, $J_{mm,n} = -\sum_{i=1}^{n} \frac{\partial \psi(y_i|x_i,\widehat{\theta}_{mm})}{\partial \widehat{\omega}_{mm}}$ and $K_{mm,n} = \sum_{i=1}^{n} \psi(y_i|x_i,\widehat{\theta}_{mm})\psi^t(y_i|x_i,\widehat{\theta}_{mm})$, with ψ the derivative of $\tilde{\rho}_d$ with respect to $\widehat{\omega}_{mm}$. Again, the smallest such value points towards the preferred model.

Alternatively, in the same spirit as in Section 3.3.2 for robust estimators of scale, we propose robust AIC versions based on M- and MM-estimators as follows,

$$\text{AIC.M} = 2\,n\log(\widehat{\sigma}_m) + 2\ \text{Trace}(J_{m,n}^{-1}K_{m,n}), \tag{3.14}$$

$$\text{AIC.MM} = 2\,n\log(\widehat{\sigma}_{mm}) + 2\ \text{Trace}(J_{mm,n}^{-1}K_{mm,n}). \tag{3.15}$$

The model with the smallest AIC value indicates the preferred model. Our simulation studies show that these robust scale based-criteria (3.14) and (3.15) lead to a better performance as compared to the versions (3.9) and (3.13) with the scale estimator re-computed for each model.

3.4.2 Using uniform asymptotic results

Omelka and Salibián-Barrera (2010) obtain the uniform consistency and normality of the S- and MM-estimators over a contamination neighborhood $\mathcal{H}_{\epsilon_0}$. A difference with the (pointwise) asymptotic normality result is an increased variance, which will be reflected in the penalty term of the AIC when such asymptotic results are used. To make this more precise, let G_0 and F_0 be the cumulative distribution functions of X and u respectively. The cumulative distribution of (Y, X) under model (3.1) is then given by $H_0(y, x) = G_0(x)F_0(y - \theta_0^t x)$. In the presence of outliers, we make the assumption that the cumulative distribution function H of the data belongs to a contamination neighborhood of H_0 of size ϵ_0. More precisely,

$$H \in \mathcal{H}_{\epsilon_0} = \{(1-\epsilon)H_0 + \epsilon H^*; \epsilon \in [0, \epsilon_0]\},$$

where H^* is an arbitrary cumulative distribution function and $\epsilon_0 < 0.5$.

To define the penalty term, consider the functional form of the estimators. For each $\theta \in \mathbb{R}^p$ and $H \in \mathcal{H}_{\epsilon_0}$, define a functional $\sigma(.,\theta) : \mathcal{F} \subset \mathcal{H}_{\epsilon_0} \to \mathbb{R}_+$, and a scale function $\sigma(H, \theta)$ that satisfies

$$E_H\Big[\rho_0\left(\frac{Y - \theta^t X}{\sigma(H,\theta)}\right)\Big] = b,$$

where E_H is the expectation computed with respect to H. The associated functional S-estimators of location and scale satisfy

$\theta_s(H) = \arg\inf_{\theta\in\mathbb{R}^p} \sigma(H,\theta)$, and $\sigma_s(H) = \inf_{\theta\in\mathbb{R}^p} \sigma(H,\theta)$.
For MM-estimators

$$\theta_{mm}(H) = \underset{\|\theta\|\in\mathbb{R}^p}{\operatorname{argmin}} E_H \left[\rho_1\left(\frac{Y-\theta^t X}{\sigma_s(H)}\right)\right].$$

In practice $\rho_1 = \tilde{\rho}_d$ is often a re-scaled version of $\rho_0 = \rho_d$ (Tukey's bi-square family loss function).

Omelka and Salibián-Barrera (2010) shown that $\sqrt{n}(\widehat{\theta}_s - \theta_s(H)) \sim N_p(0,\Sigma_H)$, with $\Sigma_H = J_{us}^{-1} K_{us} J_{us}^{-1}$ and

$$\begin{aligned}
K_{us} &= E_H[\rho_0'^2(u_1(H))\frac{XX^t}{\sigma_s(H)^2}] + \frac{d_H}{b_H}\frac{d_H^t}{b_H}\mathrm{E}_H[(\rho_0(u_1(H))-b)^2] \\
&\quad - E_H[\rho_0'(u_1(H))(\rho_0(u_1(H))-b)X^t]\frac{d_H^t}{\sigma_s^2(H)} \\
&\quad - \frac{d_H}{b_H}\mathrm{E}_H[\rho_0'(u_1(H))(\rho_0(u_1(H))-b)\frac{X^t}{\sigma_s(H)}], \\
J_{us} &= \mathrm{E}_H\left[\rho_0''\left(\frac{Y-\theta_s(H)^t X}{\sigma_s(H)}\right)\frac{XX^t}{\sigma_s^2(H)}\right],
\end{aligned}$$

where $u_1(H) = (Y-\theta_s(H)^t X)/\sigma_s(H)$,

$$\begin{aligned}
d_H &= E_H\left[\rho_0''\left(\frac{Y-\theta_s(H)^t X}{\sigma_s(H)}\right)\frac{(Y-\theta_s(H)^t X)X^t}{\sigma_s(H)^2}\right] \\
b_H &= E_H\left[\rho_0'\left(\frac{Y-\theta_s(H)^t X}{\sigma_s(H)}\right)\frac{(Y-\theta_s(H)^t X)}{\sigma_s(H)}\right].
\end{aligned}$$

For the calculations of the penalty term in the robust AIC, we use the corresponding empirical information matrices, where $J_{us,n}$ is equal to $J_{s,n}$. Hence, the difference lies in the asymptotic variance component $K_{us,n}$, which results in a larger variance for uniform S-estimators by taking the contamination neighborhoods into account. This leads immediately to a robust AIC based on uniform asymptotic results for S-estimators,

$$\text{AIC.US} = 2\,n\log\widehat{\sigma}_s + 2\ \text{Trace}(J_{s,n}^{-1}K_{us,n}), \tag{3.16}$$

where, for example, $\rho_0 = \rho_d$ is Tukey's bi-square loss function. For MM-estimators we can use either the form with the ρ_d function, or the scale-based version, leading to the following definitions.

$$\text{AIC}_\rho.\text{UMM} = 2\sum_{i=1}^n \tilde{\rho}_d\left(\frac{Y_i - \widehat{\theta}_{mm}^t x_i}{\widehat{\sigma}_{mm}}\right) + 2\ \text{Trace}(J_{umm,n}^{-1}K_{umm,n}). \tag{3.17}$$

$$\text{AIC.UMM} = 2\,n \log(\widehat{\sigma}_{mm}) + 2\ \text{Trace}(J_{umm,n}^{-1} K_{umm,n}). \qquad (3.18)$$

where $\widehat{\sigma}_{mm}$ is the MM-estimator of scale and the matrices $J_{umm,n}$ and $K_{umm,n}$ are obtained in a similar fashion. Again, the smallest value of AIC towards the preferred model.

3.5 Numerical results

3.5.1 Simulation settings

The settings for the simulation study are as follows. For the number of variables p equal to either 6 or 10, the regression variables $X_1, \ldots, X_p$ are generated from a multivariate normal distribution with mean vector $\mu = (1, \ldots, p)$ and variance covariance matrices (i) a $(p \times p)$ identity matrix for independent Xs and (ii) for dependent Xs, we used for $p = 6$ the matrix Σ_1 reflecting independence between the group of the first three variables and the group of the last three variables, (iii) Σ_2 which is a situation where all six variables are correlated, and (iv) Σ_3 for the case $p = 10$ showing correlation within the group of the first six variables, within the group of the last four variables, and a constant correlation between the groups.

To shorten the display, define $I_r(a)$ as the square $r \times r$ matrix with the values 1 on the diagonal and the constant value a on all off-diagonal entries, and define $1_{r\times s}$ the matrix of dimension $r \times s$ consisting of values 1 everywhere. Then,

$$\Sigma_1 = \begin{pmatrix} I_3(0.6) & 0 \cdot 1_{3\times3} \\ 0 \cdot 1_{3\times3} & I_3(0.3) \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} I_3(0.6) & 0.4 \cdot 1_{3\times3} \\ 0.4 \cdot 1_{3\times3} & I_3(0.3) \end{pmatrix},$$

$$\Sigma_3 = \begin{pmatrix} I_6(0.6) & 0.3 \cdot 1_{6\times4} \\ 0.3 \cdot 1_{4\times6} & I_4(0.4) \end{pmatrix}.$$

For the case $p = 6$ we define the true model using the first three variables X_1, X_2, X_3. Hence, when using Σ_1, the set of important variables X_1, X_2, X_3 is not correlated with the unimportant variables X_4, X_5, X_6, this is in contrast to the situation when using Σ_2. For $p = 10$, the first six variables appear in the true model, while the remaining four variables are redundant. The chosen settings pose increased difficulty for variable selection.

For the mean structure, we have used the functions $m_1(x) = 1 + x_1 + x_2 + x_3$ for the setting with $p = 6$ and $m_2(x) = 1 + \sum_{j=1}^{6} x_j$ for the setting with $p = 10$, with $x = (x_1, \ldots, x_p)$. As error distribution we used $N(0, 0.7^2)$.

These values are kept fixed for all settings to reduce simulation variability. We took sample sizes equal to 50 and 100. Since the results were quite similar, we here only show the results for the sample size equal to 50. We have fitted all $2^p - 1$ possible models without interactions with these p variables.

We compare nine different AIC versions in this simulation study: classical AIC based on maximum likelihood estimation assuming a normal distribution (3.2), the scale based versions (3.12), (3.14)–(3.16), (3.18), as defined in Sections 3.3.2 and 3.4, and the versions using the ρ-function (3.9), (3.13), (3.17) of Sections 3.2.4 and 3.4.

To compute the robust M, S and MM-estimators, we used, respectively, the functions `rlm()`, `lmrob.S()` and `lmrob..M..fit()` from the `R` libraries `MASS` and `robustbase`. In order to investigate the robustness of the methods against outliers, we considered three situations: (i) vertical outliers (outliers in the response only), (ii) good leverage points (outliers in the response and the covariates), and (iii) bad leverage points (outliers in some of the covariates only). For case (i) we randomly generated different percentages of outliers (0%, 5%, 10%, 20%, 30% and 40%) from $N(50, 0.1^2)$ for each of the simulated cases. For case (ii) we considered the different percentages of outliers (0%, 5%, 10%, 20%, 30%) on the variables X_1, X_2 and X_4 are generated from a $N(100, 0.5^2)$ distribution, then generated Y to get good leverage points. For case (iii) different percentages of outliers (0%, 5%, 10%, 20%, 30%) on the variables X_1, X_2 and X_4 are generated from a $N(100, 0.5^2)$ distribution. For each of these settings we simulated 1000 samples.

3.5.2 Simulation results

A summary of the simulation results is provided by reporting the proportions of selected models that are

(C) Correct fit - The true model only.

(O) Overfit - Models containing all the variables in the true model plus some more that are actually redundant.

(U) Underfit - Models with only a strict subset of the variables in the true model.

(W) Wrong fit - All models that are not overfit (O), not a correct fit (C) nor underfit (U). These are the models where some of the relevant variables might be present (though not all of them) in addition to some of the redundant variables.

We first consider the vertical outliers case with outlying response values. Table 3.1 and Table 3.2 show detailed simulation results for one of the simulation settings with all AIC methods. As expected, the classical AIC works better than the robust AICs for the data without outliers. The classical AIC selects a large proportion of underfit or wrong fit models for the data with outliers, while a higher proportion of overfit and correct fit models are select by AIC.M with at most 20% contamination level. A higher proportion of overfit and correct fit models are select by AIC.S, AIC.US, AIC.MM and AIC.UMM. All of these methods work better for the cases with a high contamination level of outliers and break down at 50% of outliers in the data; this holds for both dependent and independent Xs. We present results for the AIC based on the ρ function in two versions: with known scale value (actually, estimated from the largest model and kept fixed for all models) and with unknown scale (re-estimated for each model).

Table 3.1: *Proportion of selected models from classical AIC, AIC with M-estimation (AIC$_\rho$.M), AIC with MM-estimation (AIC$_\rho$.MM), AIC with uniform MM-estimation (AIC$_\rho$.UMM) for both known (estimated in the largest model) and unknown scale σ. Data are generated with dependent Xs, mean structure m_1 for $p = 6$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$. S- and MM-estimators are computed with 50% breakdown point.*

ε			Based on loss function (ρ)				
			σ known		σ unknown		
%		AIC	M	MM	M	MM	UMM
0	C	0,480	0.453	0.446	0.020	0.003	0.004
	O	0,520	0.547	0.554	0.091	0.337	0.259
	U	0,000	0.000	0.000	0.179	0.001	0.008
	W	0,000	0.000	0.000	0.710	0.659	0.729
5	C	0,002	0.428	0.474	0.000	0.003	0.003
	O	0,001	0.572	0.526	0.000	0.329	0.255
	U	0,560	0.000	0.000	0.003	0.003	0.006
	W	0,437	0.000	0.000	0.997	0.665	0.736
10	C	0,005	0.489	0.511	0.000	0.003	0.004
	O	0,004	0.508	0.481	0.000	0.293	0.208
	U	0,454	0.002	0.001	0.001	0.002	0.009
	W	0,537	0.001	0.007	0.999	0.702	0.779
20	C	0,008	0.505	0.591	0.000	0.002	0.002
	O	0,004	0.344	0.312	0.006	0.298	0.228
	U	0,427	0.058	0.036	0.002	0.001	0.002
	W	0,561	0.093	0.061	0.992	0.699	0.768
30	C	0,012	0.008	0.430	0.027	0.000	0.000
	O	0,005	0.014	0.106	0.051	0.435	0.370
	U	0,409	0.285	0.271	0.139	0.000	0.000
	W	0,574	0.693	0.193	0.783	0.565	0.630
40	C	0,007	0.005	0.014	0.001	0.000	0.000
	O	0,008	0.013	0.000	0.000	0.828	0.826
	U	0,397	0.316	0.745	0.384	0.000	0.000
	W	0,588	0.666	0.241	0.615	0.172	0.174

Table 3.2: *Proportion of selected models from classical AIC, the robust scale versions: AIC.M, AIC.S, AIC.US, AIC.MM and AIC.UMM for unknown scale σ. Data are generated with dependent Xs, mean structure m_1 for $p = 6$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$. S- and MM-estimators are computed with 50% breakdown point.*

ε					Based on scale estimators		
					σ unknown		
%		AIC	AIC.M	AIC.S	AIC.US	AIC.MM	AIC.UMM
0	C	0,480	0.360	0.163	0.173	0.164	0.178
	O	0,520	0.552	0.808	0.795	0.816	0.799
	U	0,000	0.038	0.005	0.007	0.005	0.007
	W	0,000	0.050	0.024	0.025	0.015	0.016
5	C	0,002	0.366	0.214	0.217	0.216	0.219
	O	0,001	0.543	0.761	0.754	0.763	0.756
	U	0,560	0.041	0.004	0.005	0.005	0.006
	W	0,437	0.050	0.021	0.024	0.016	0.019
10	C	0,005	0.407	0.233	0.239	0.236	0.241
	O	0,004	0.522	0.741	0.734	0.740	0.734
	U	0,454	0.032	0.005	0.006	0.005	0.006
	W	0,537	0.039	0.021	0.021	0.019	0.019
20	C	0,008	0.406	0.417	0.417	0.420	0.419
	O	0,004	0.440	0.564	0.563	0.562	0.562
	U	0,427	0.069	0.005	0.006	0.005	0.006
	W	0,561	0.085	0.014	0.014	0.013	0.013
30	C	0,012	0.032	0.647	0.646	0.647	0.644
	O	0,005	0.029	0.343	0.344	0.343	0.346
	U	0,409	0.431	0.005	0.005	0.005	0.005
	W	0,574	0.508	0.005	0.005	0.005	0.005
40	C	0,007	0.018	0.906	0.906	0.906	0.906
	O	0,008	0.044	0.075	0.075	0.075	0.075
	U	0,397	0.453	0.014	0.014	0.014	0.014
	W	0,588	0.485	0.005	0.005	0.005	0.005

We observe from Table 3.1 and Table 3.2 that AIC based on the ρ function with a supposed to be known scale (estimated from the largest model) gives better results than when the scale is truly supposed to be unknown. In the latter case, all unknown parameters, including the scale, are treated as unknown and are estimated in the corresponding model, rather than in the largest model. A comparison of the scale versions of the AIC to those based on the ρ-function reveals that AIC.M, AIC.MM and AIC.UMM work better than $AIC_\rho.M$, $AIC_\rho.MM$ and $AIC_\rho.UMM$. For the rest of the paper we restrict to presenting the results using the scale-based versions of the AIC.

Figure 3.1 shows the results of the proportion of selected correct fit (C) and overfit (O) models by different model selection strategies. As expected, the classical AIC works slightly better than the robust AICs for the data without outliers. The classical AIC selects a small proportion of correct fit and overfit models, when the data contain outliers in the response variable. That means, the classical AIC method is ignoring some of the important variables in the model. AIC.M selects a large proportion of correct fit and overfit models until a 20% contamination level after which it gets influenced by the outliers and further shows a behaviour similar to the classical AIC. A higher proportion of correct fit models is selected by AIC.S for the data set with outliers. This method works fine also for the cases with a high contamination level of outliers and breaks down when there are 50% of outliers in the data. Figure 3.1 (a) and (b) presents a summary of the results for dependent Xs when using Σ_2 and for independent Xs respectively. It is observed that AIC.M selects a higher proportion of correct fit and overfit models for the independent case than for the dependent case. AIC.S selects a high proportion of correct fit and overfit models for both dependent and independent cases.

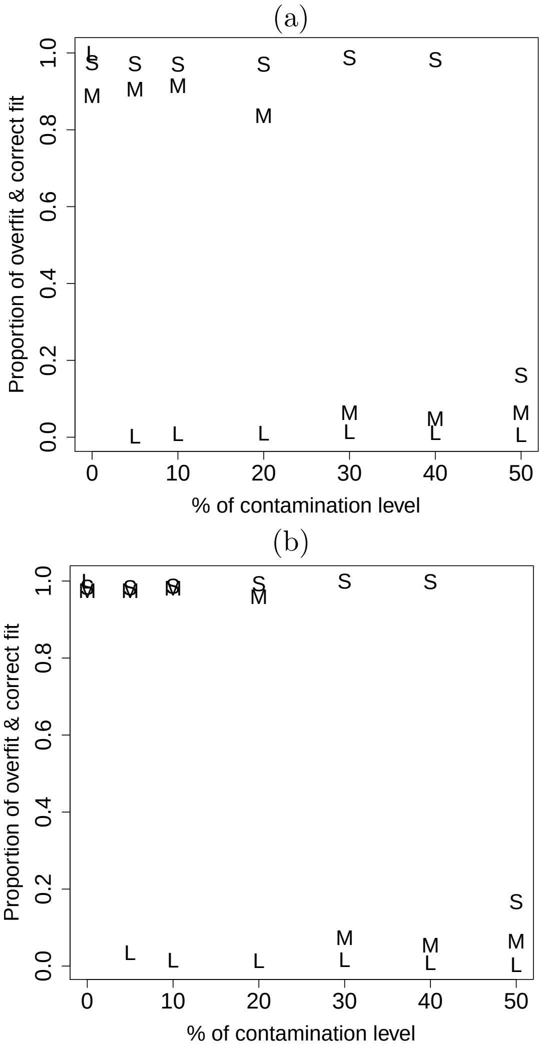


Figure 3.1: *Proportion of selected models from correct fit (C) and overfit (O) from classical AIC (L), AIC based on M-estimators (M) and AIC based on S-estimators (S) for data generated with mean structure m_1 for $p = 6$, error terms from $N(0, 0.7^2)$, sample size $n = 50$ and different percentages of outliers generated from $N(50, 0.1^2)$ for two different cases (a) dependent X_s with estimators with 50% breakdown point, (b) independent X_s with estimators with 50% breakdown point.*

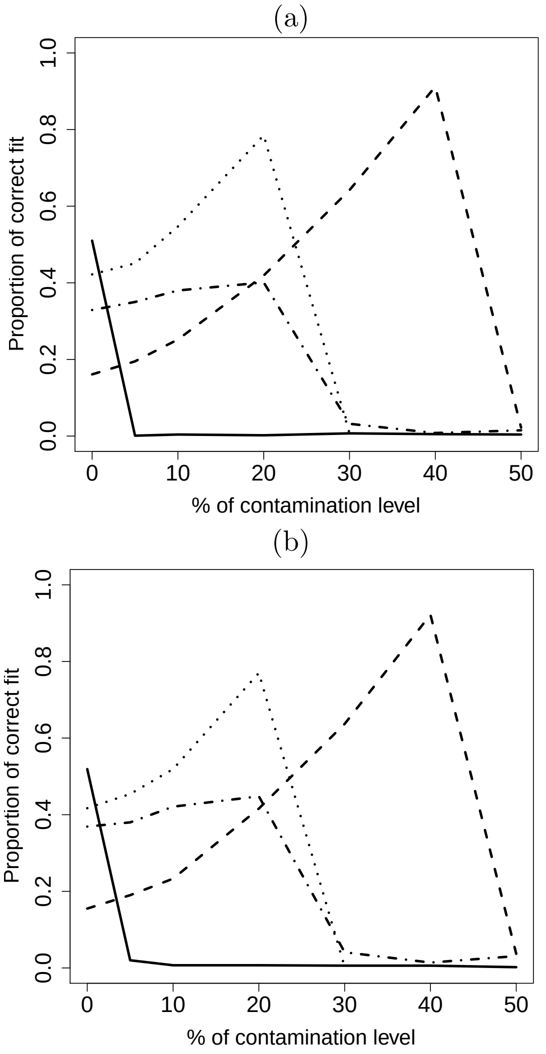


Figure 3.2: *Proportion of selected models from correct fit (C) from classical AIC (solid line), AIC based on M-estimators (dot-dashed line), AIC based on S-estimators with 50% breakdown point (dashed line) and AIC based on S-estimators with 30% breakdown point (dotted line) for data generated with mean structure m_1 for $p = 6$, $\varepsilon \sim N(0, 0.7^2)$, sample size $n = 50$ and different % outliers generated from $N(50, 0.1^2)$ for (a) dependent X_s as in Σ_2 and (b) independent X_s.*

The proportion of correct fit models from the classical AIC, and from AIC based on M- and S- estimators is given in Figure 3.2 (a) and (b) for dependent Xs when using Σ_2 and for independent Xs, respectively.

For small percentages of outliers (10%–20%), the AIC.S method (when tuned to a 50% breakdown point) is not doing well in selecting the correct model. Therefore, we re-compute AIC.S, now tuned to have a 30% breakdown point for the estimators. The corresponding results are plotted in Figure 3.2 (a) and (b). We observe that this significantly helps for the case of 20% outliers, resulting in a high proportion of correct models selected by AIC.S. When we consider the proportions of both overfit and correct fit models together, then AIC.S is performing well for any percentage of outliers with both considered breakdown points. We also computed AIC.US, AIC.MM and AIC.UMM in this simulation setting and observed that the results are similar to those of AIC.S.

A main message to be learned from this simulation study is that AIC based on M-estimators using expression (3.9) with the scale estimator computed in each model separately, rather than at the largest model, performs less well than the AIC.M based on a robust scale estimator. The AIC versions based on robust scale estimators are preferable. For best performance, the breakdown point of the estimators should be considered in relation with the proportion of outliers in the data to avoid underfitting.

More detailed simulation results are shown in Table 3.3 and Table 3.4. Again, as expected, the classical AIC works better than the robust AICs for the data without outliers. The classical AIC selects a large proportion of underfit or wrong fit models for the data with outliers, while a higher proportion of overfit and correct fit models are select by AIC.S, AIC.US, AIC.MM and AIC.UMM. All of these methods work better for the cases with a high contamination level of outliers and break at 50% of outliers in the data; this holds for both dependent and independent Xs. A further detailed investigation about this issue is reported at the end of this section. AIC based on M-estimators works fine for the data with small ($\leq$ 20%) contamination level. The S estimation based criteria AIC.S and AIC.US give similar results in most of the cases in Table 3.3 and Table 3.4.

Table 3.3: *Proportion of selected models from classical AIC (AIC), AIC with M-estimation (M), AIC with S-estimation (S), AIC with uniform S-estimation (US), AIC with MM-estimation (MM) and AIC with uniform MM-estimation (UMM), for data generated with dependent Xs, mean structure m_2 for $p = 10$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different % ε of outliers generated from $N(50, 0.1^2)$. S- and MM- estimators are computed with 50% breakdown point.*

ε		Dependent Xs					
%		AIC	AIC.M	AIC.S	AIC.US	AIC.MM	AIC.UMM
0	C	0.424	0.256	0.055	0.053	0.069	0.077
	O	0.576	0.516	0.600	0.592	0.795	0.780
	U	0.000	0.006	0.006	0.006	0.001	0.001
	W	0.000	0.222	0.339	0.349	0.135	0.142
5	C	0.001	0.259	0.079	0.069	0.106	0.099
	O	0.000	0.517	0.627	0.620	0.752	0.750
	U	0.278	0.007	0.004	0.004	0.002	0.002
	W	0.721	0.217	0.290	0.307	0.140	0.149
10	C	0.000	0.281	0.114	0.115	0.137	0.141
	O	0.000	0.505	0.633	0.620	0.739	0.730
	U	0.322	0.016	0.004	0.005	0.001	0.001
	W	0.678	0.198	0.249	0.260	0.123	0.128
20	C	0.000	0.273	0.236	0.235	0.263	0.250
	O	0.000	0.315	0.632	0.625	0.649	0.657
	U	0.310	0.048	0.000	0.001	0.002	0.003
	W	0.690	0.364	0.132	0.139	0.086	0.090
30	C	0.000	0.001	0.573	0.568	0.576	0.575
	O	0.000	0.003	0.364	0.364	0.363	0.363
	U	0.318	0.320	0.006	0.006	0.007	0.007
	W	0.682	0.676	0.057	0.062	0.054	0.055

Based on the results from Table 3.3 and Table 3.4, for dependent Xs the proportion of overfit models based on AIC.S and AIC.US is larger than for the case of independent Xs and based on AIC.MM and AIC.UMM is smaller than for the case of independent Xs.

Table 3.4: *Proportion of selected models from classical AIC, AIC with M-estimation, AIC with S-estimation, AIC with uniform S-estimation, AIC with MM-estimation and AIC with uniform MM-estimation, for data generated with independent X_s, mean structure m_2 for $p = 10$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$. S- and MM-estimators are computed with 50% breakdown point.*

ε		Independent Xs					
%		AIC	AIC.M	AIC.S	AIC.US	AIC.MM	AIC.UMM
0	C	0.404	0.288	0.029	0.030	0.069	0.074
	O	0.596	0.675	0.399	0.368	0.861	0.851
	U	0.000	0.000	0.016	0.006	0.001	0.001
	W	0.000	0.037	0.556	0.596	0.069	0.074
5	C	0.007	0.326	0.049	0.042	0.086	0.086
	O	0.009	0.635	0.427	0.394	0.839	0.826
	U	0.223	0.003	0.013	0.013	0.000	0.000
	W	0.761	0.036	0.511	0.551	0.075	0.088
10	C	0.001	0.341	0.078	0.079	0.150	0.149
	O	0.000	0.616	0.489	0.463	0.801	0.801
	U	0.264	0.004	0.007	0.003	0.002	0.002
	W	0.735	0.039	0.426	0.455	0.047	0.048
20	C	0.000	0.371	0.211	0.205	0.253	0.252
	O	0.000	0.455	0.591	0.574	0.713	0.714
	U	0.026	0.025	0.005	0.002	0.000	0.000
	W	0.974	0.149	0.193	0.219	0.034	0.034
30	C	0.000	0.002	0.588	0.576	0.604	0.597
	O	0.000	0.002	0.392	0.401	0.389	0.396
	U	0.283	0.271	0.000	0.000	0.000	0.000
	W 0.717	0.725	0.020	0.023	0.007	0.007	

Since Table 3.3 and Table 3.4 shows that the proportion of selected correct fit models is small for the cases with 5%, 10% and 20% contamination when estimators with 50% breakdown point are used, we recompute the AIC.S, AIC.US, AIC.MM and AIC.UMM for the cases with 0%, 5%, 10%, 20%, 30% contamination level, now with 30% breakdown point estimators.

These results are presented in Table 3.5 and Table 3.6. The AIC based on MM-estimators selects higher proportions of correct fit than the AIC based on S-estimators for the data without outliers. It clearly shows that for the case of 20% contamination, the proportion of selected correct fit models is now much larger for the methods AIC.S, AIC.US, AIC.MM and AIC.UMM.

Table 3.5: *Proportion of selected models from classical AIC, AIC with M-estimation, AIC with S-estimation, AIC with uniform S-estimation, AIC with MM-estimation and AIC with uniform MM-estimation, for data generated with dependent X_s, mean structure m_2 for $p = 10$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$. S- and MM-estimators are computed with a 30% breakdown point.*

ε		Dependent Xs					
%		AIC	AIC.M	AIC.S	AIC.US	AIC.MM	AIC.UMM
0	C	0.424	0.256	0.057	0.036	0.272	0.267
	O	0.576	0.516	0.601	0.257	0.707	0.710
	U	0.000	0.006	0.007	0.048	0.001	0.001
	W	0.000	0.222	0.335	0.659	0.020	0.022
5	C	0.001	0.259	0.342	0.340	0.338	0.348
	O	0.000	0.517	0.646	0.638	0.655	0.643
	U	0.278	0.007	0.001	0.000	0.000	0.000
	W	0.721	0.217	0.011	0.022	0.007	0.009
10	C	0.000	0.281	0.451	0.461	0.455	0.461
	O	0.000	0.505	0.546	0.533	0.538	0.532
	U	0.322	0.016	0.000	0.000	0.000	0.000
	W	0.678	0.198	0.003	0.006	0.007	0.007
20	C	0.000	0.273	0.820	0.818	0.817	0.818
	O	0.000	0.315	0.177	0.179	0.180	0.179
	U	0.310	0.048	0.001	0.001	0.001	0.001
	W	0.690	0.364	0.002	0.002	0.002	0.002
30	C	0.000	0.001	0.000	0.000	0.000	0.000
	O	0.000	0.003	0.001	0.003	0.000	0.000
	U	0.318	0.320	0.399	0.380	0.405	0.381
	W	0.682	0.676	0.600	0.617	0.595	0.619

Table 3.6: *Proportion of selected models from classical AIC, AIC with M-estimation, AIC with S-estimation, AIC with uniform S-estimation, AIC with MM-estimation and AIC with uniform MM-estimation, for data generated with independent X_s, mean structure m_2 for $p = 10$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$. S- and MM-estimators are computed with a 30% breakdown point.*

ε		Independent Xs					
%		AIC	AIC.M	AIC.S	AIC.US	AIC.MM	AIC.UMM
0	C	0.404	0.288	0.028	0.031	0.251	0.261
	O	0.596	0.675	0.398	0.369	0.740	0.730
	U	0.000	0.000	0.017	0.005	0.000	0.000
	W	0.000	0.037	0.557	0.595	0.009	0.009
5	C	0.007	0.326	0.310	0.317	0.318	0.329
	O	0.009	0.635	0.668	0.664	0.679	0.668
	U	0.223	0.003	0.000	0.000	0.000	0.000
	W	0.761	0.036	0.022	0.019	0.003	0.003
10	C	0.001	0.341	0.441	0.442	0.439	0.441
	O	0.000	0.616	0.558	0.558	0.561	0.559
	U	0.264	0.004	0.000	0.000	0.000	0.000
	W	0.735	0.039	0.001	0.000	0.000	0.000
20	C	0.000	0.371	0.804	0.804	0.803	0.804
	O	0.000	0.455	0.196	0.196	0.197	0.196
	U	0.026	0.025	0.000	0.000	0.000	0.000
	W	0.974	0.149	0.000	0.000	0.000	0.000
30	C	0.000	0.002	0.001	0.000	0.001	0.000
	O	0.000	0.002	0.004	0.005	0.002	0.003
	U	0.283	0.271	0.265	0.181	0.275	0.181
	W	0.717	0.725	0.730	0.814	0.722	0.816

Next, we present results of simulated data with outliers on the explanatory variables, in addition to outliers in the response variable (see the description of cases (ii) and (iii) above). The results are presented in Table 3.7. We have fitted all possible models with six explanatory variables in this setting.

Table 3.7: *Proportion of selected models from classical AIC, AIC with M-estimation, S-estimation and MM-estimation for data generated with dependent X_s, mean structure $m_1(x)$ for $p = 6$, error terms from a $N(0, 0.7^2)$, and for sample size $n = 50$. Considered different % ε of outliers generated for Y, X_1, X_2 and X_4 variables. S- and MM- estimators are computed with a 50% breakdown point.*

ε		Bad leverage points					Good leverage points			
%		AIC	M	S	MM	%	AIC	M	S	MM
0	C	0.510	0.329	0.161	0.168	0	0.510	0.329	0.161	0.168
	O	0.490	0.561	0.814	0.812		0.490	0.561	0.814	0.812
	U	0.000	0.041	0.003	0.004		0.000	0.041	0.003	0.004
	W	0.000	0.069	0.022	0.016		0.000	0.069	0.022	0.016
5	C	0.539	0.325	0.186	0.188	5	0.540	0.326	0.186	0.188
	O	0.460	0.626	0.771	0.794		0.460	0.626	0.771	0.794
	U	0.000	0.012	0.005	0.004		0.000	0.011	0.005	0.004
	W	0.001	0.037	0.038	0.014		0.000	0.037	0.038	0.014
10	C	0.520	0.345	0.153	0.154	10	0.534	0.358	0.186	0.192
	O	0.479	0.624	0.831	0.840		0.466	0.581	0.795	0.800
	U	0.000	0.009	0.001	0.001		0.000	0.016	0.001	0.000
	W	0.001	0.022	0.000	0.005		0.000	0.045	0.018	0.008
20	C	0.524	0.371	0.164	0.175	20	0.540	0.365	0.176	0.176
	O	0.475	0.586	0.818	0.814		0.460	0.560	0.804	0.809
	U	0.000	0.010	0.003	0.000		0.000	0.018	0.006	0.004
	W	0.001	0.033	0.000	0.011		0.000	0.057	0.014	0.011
30	C	0.545	0.329	0.172	0.174	30	0.536	0.323	0.173	0.174
	O	0.455	0.602	0.791	0.799		0.464	0.575	0.788	0.796
	U	0.000	0.015	0.003	0.004		0.000	0.031	0.008	0.004
	W	0.000	0.054	0.034	0.023		0.000	0.071	0.031	0.026

We simulated data with different percentages of outliers in the explanatory variables. We compute AIC values from these six different AIC methods. Based on these results in 3.7, we observe that the classical AIC method selects a large proportion of overfit and correct fit models for all cases. Therefore, based on this simulation results, it seems valid to use the classical AIC method for the cases with outliers only on the explanatory variables. Also, AIC based on S, uniform S, MM and uniform MM-estimation select a large proportion of overfit and correct fit models for all cases.

To investigate in more details the behaviour of the criteria when the percentage of outliers increases, we checked the scale estimators of the models for both M and S-estimation. In particular, we computed the average over 1000 simulation samples of scale estimates of each of the models in the simulation setting for data generated with dependent Xs, mean structure m_1 for $p = 6$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We considered three categories of selected models (i) correct fit, (ii) overfit and (iii) underfit & wrong fit. Table 3.8 presents the summary results of scale estimators based on the M- and S-estimation method.

Table 3.8: *Average of scale estimates from M-estimation (Scale.M) and S-estimation (Scale.S) over 1000 simulated samples of all models for data generated with dependent Xs, mean structure m_1 for $p = 6$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$.*

	ε	Correct fit		Overfit			Underfit & Wrong fit	
	%		Lower	Upper	Range	Lower	Upper	Range
Scale.M	0	0.667	0.636	0.658	0.022	0.930	2.504	1.573
	10	0.783	0.759	0.775	0.016	1.102	2.726	1.623
	20	1.125	1.160	1.432	0.271	1.543	3.689	2.146
	30	3.949	41.909	43.798	1.889	23.943	43.691	19.748
Scale.S	0	0.677	0.653	0.670	0.017	0.939	2.474	1.535
	10	0.808	0.801	0.806	0.006	1.135	2.861	1.726
	20	0.994	1.002	1.022	0.020	1.403	3.442	2.040
	30	1.316	1.348	1.426	0.077	1.825	4.422	2.597

Table 3.8 shows, first, that the scale estimates increase with the percentage of outliers. Second, the cases with 0% and 10% outliers result on average in a larger scale estimate of correctly fitted models than for models that are overfit; indeed the table shows that the average value for correct

fits is larger than the upper value over the simulation results of the overfit models. In such cases, the model selection criteria AIC based on M- and S- estimation will often select an overfit model, since a small scale estimate is preferable since we are minimizing the AIC values. On the other hand, for data with a larger contamination level of outlying cases (20% and 30% outliers) the averaged scale estimates are the smallest for the correctly fitted models. Therefore AIC based on M- and S-estimation will tend to select the correctly fitted model more often. The observed ranges of the scale estimates over the simulation study for the overfit models are smaller than for the underfit and wrongly fitted models, and both ranges increase with the percentage of outliers. The effect of the outliers on the penalty part of the various AIC is seen to be non-influential. As a comparison, we redid the simulation exercise (results are in Table 3.9) with using for the penalty the number of parameters in the model (which is not influenced by the number of outliers) and came to the same conclusion. It is the behaviour of the robust scale estimators in misspecified models that explains the obtained results for model selection.

Additionally, we have considered the proposed AIC based on robust scale M- and S-estimators with different penalty terms and compared them with classical AIC and generalized information criterion based on S-estimators(GIC.S). We presented the results in Table 3.9 from the simulation study, Data are generated with dependent Xs, mean structure m_1 for $p = 6$, error terms from a $N(0, 1)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(100, 0.5^2)$. We denote the robust scale versions of AIC based on M-estimator as AIC.M in (3.14) and based on S-estimator as AIC.S in (3.12). Here we used the penalty term is full matrices with dimension $(p+1) \times (p+1)$ in (3.14) and (3.12) and denoted AIC.M1 and AIC.S1 respectively. We used the penalty, the number of parameters in the model in the robust scale versions of AICs based on M- and S-estimators as in (3.14) and (3.12), denote AIC.M2 and AIC.S2 respectively. Table 3.9 shows that AIC.M1 and AIC.M2 are not much difference than AIC.M.

Table 3.9: *Proportion of selected models. Data are generated with dependent Xs, mean structure m_1 for $p = 6$, $n = 50$, error terms from a $N(0,1)$. Different percentages ε of outliers from $N(100, 0.5^2)$. S-estimators are computed with 50% breakdown point.*

		0%	10%	20%	30%	40%
AIC	C	0.520	0.004	0.004	0.005	0.008
	O	0.480	0.004	0.009	0.009	0.008
	U	0.000	0.407	0.398	0.395	0.396
	W	0.000	0.585	0.589	0.591	0.588
AIC.M	C	0.254	0.277	0.260	0.013	0.014
	O	0.384	0.370	0.283	0.012	0.042
	U	0.117	0.147	0.219	0.454	0.452
	W	0.245	0.206	0.238	0.521	0.492
AIC.M1	C	0.242	0.136	0.172	0.013	0.014
	O	0.379	0.584	0.462	0.015	0.043
	U	0.118	0.046	0.093	0.434	0.446
	W	0.261	0.234	0.273	0.538	0.497
AIC.M2	C	0.139	0.144	0.163	0.014	0.011
	O	0.501	0.529	0.460	0.018	0.062
	U	0.063	0.060	0.093	0.404	0.440
	W	0.297	0.267	0.284	0.564	0.487
AIC.S	C	0.141	0.194	0.352	0.574	0.654
	O	0.695	0.638	0.474	0.298	0.046
	U	0.023	0.038	0.065	0.074	0.256
	W	0.141	0.130	0.109	0.054	0.044
AIC.S1	C	0.138	0.192	0.349	0.572	0.660
	O	0.699	0.637	0.475	0.300	0.046
	U	0.021	0.039	0.064	0.073	0.250
	W	0.142	0.132	0.112	0.055	0.044
AIC.S2	C	0.254	0.342	0.484	0.652	0.577
	O	0.576	0.489	0.338	0.173	0.027
	U	0.046	0.063	0.088	0.121	0.361
	W	0.124	0.106	0.090	0.054	0.035
GIC.S	C	0.135	0.197	0.351	0.569	0.658
	O	0.718	0.643	0.494	0.316	0.047
	U	0.017	0.032	0.053	0.062	0.251
	W	0.130	0.128	0.102	0.053	0.044

The scale version of AIC based on S-estimator with different penalty terms gives similar results in Table 3.9. But the number of parameters in the model is not influenced by the number of outliers in the data. We presented R-code for all AICs with different penalty terms in section 5.2.

3.5.3 Employment data in East-central Europe

We used the data set coded ZA3132 from the website http://zacat.gesis.org /webview/index.jsp, named “The evaluation of programs to assist young unemployed in post communist East-Central Europe 1996-1998”. We used a subset of this dataset consisting of the response variable, the current monthly earnings (USA $) during 1996-1998, and 16 explanatory variables (see below for the details). Cases with missing values were removed from the resulting dataset, leading to the subset of 114 observations that we used here.

The explanatory variables are as follows: X_1 age; X_2 gender; X_3 marital status (1-single,2-married/cohabiting, 3-other); X_4 highest level of education (1-less than elementary school, 2-elementary school, 3-vocational school, 4-professional or technical school, 5-high school/lycee/ gymnasium/grammar school, 6-college, 7-university); X_5 age completed full-time education; X_6 the subject or field specialized in (0-nothing in particular, 1-construction & related, 2-vehicle & machinery repairs, 3-engineering, 4-catering & hospitality, 5-personal & consumer services, 6-health & related, 7-teaching, 8-professional services, 9-other academic subjects); X_7 number of proper jobs since leaving school; X_8 number of holidays away from home during the last 12 months; X_9 amount of time for family; X_{10} amount of time for friends; and X_{11} amount of time for yourself (1-not enough, 2-about right, 3-too much); X_{12} education matches work experience (1-yes(totally), 2-yes(partly), 3-not at all, 4-no work experience); X_{13} use of motor car; X_{14} use of satellite or cable TV; X_{15} use of personal computer; X_{16} use of mobile telephone.

The variable X_4, highest level of education might be an endogenous variable in which case a traditional linear regression model is no longer valid. The endogeneity problem can be solved by introducing instrumental variables (see, e.g. Johnston and DiNardo, 1997). We took X_5, the age at completion of the full-time education, as an instrumental variable since the variables X_4 and X_5 are highly correlated ($\text{corr}(X_4, X_5) = 0.6$) while the correlation between X_5 and the response variable is small ($\text{corr}(Y, X_5) = 0.06$). We used a two-stage least squares method to fit the regression models for these data. In stage (1), we fit a regression model of X_4 on the instrumental variable X_5 to obtain the fitted values $\widehat{X}_4$. Because the variable X_4 is an ordered categorical variable, we used a proportional odds logistic regression model in stage (1). Hereby we used the function `polr()` in `R`. In stage (2) we regress Y on all other Xs and $\widehat{X}_4$. We use the model selection procedure in stage (2).

Using standardized residuals plots, it turns out that 8 response values (7%) are outside the range $(-2, 2)$ and can be considered as vertical outliers. We used the chi-square plot to detect multivariate outliers as in Garrett (1989). In such a plot the ordered robust Mahalanobis distance of the data is plot against the quantiles of the chi-squared distribution. This method applied to the continuous covariates X_1 and X_8 showed that 6 observations (5%) can be considered as leverage points. We therefore set the S and MM estimation methods to use a 30% breakdown point.

Table 3.10: *Employment data in East-central Europe. The selected explanatory variables from the classical AIC, AIC with M-estimation, S-estimation and MM-estimation, tuned for a 30% breakdown point.*

Criteria	Selected variables		
	Best model	Second best model	Third best model
AIC	X_2,X_3,X_{15},X_{16}	X_3,X_{15},X_{16}	X_2,X_{14},X_{15},X_{16}
AIC.M	X_3,X_4,X_8,X_{10},X_{11}	X_3,X_4,X_6,X_{10},X_{12}, X_{15},X_{16}	X_1,X_4,X_{12},X_{14}, X_{15}
AIC.S	X_3,X_4,X_{11},X_{12},X_{14}, X_{15},X_{16}	X_3,X_4,X_{11},X_{12}, X_{14},X_{16}	X_3,X_4,X_{12},X_{14}, X_{15},X_{16}
AIC.MM	X_3,X_4,X_{11},X_{12},X_{14}, X_{15},X_{16}	X_3,X_4,X_{11},X_{12}, X_{14},X_{16}	X_3,X_4,X_{12},X_{14}, X_{15},X_{16}

We have fitted all 2^{15} models with a combination of any of these explanatory variables and computed several AIC values for each model. The best three selected models based on each AIC method are given in Table 3.10.

The classical AIC method selects a model with four explanatory variables, while AIC based on M-estimation selects a model with five variables. For classical AIC method, the number of selected variables is lower than for the other criteria. This is in line with the simulation results where we observed that classical AIC has the tendency to select underfit models in the presence of outliers.

The proposed methods based on S- and MM-estimators select the same best model with seven variables. Variables X_3 marital status, X_4 highest level of education and X_{11} amount of time for yourself, coincide with the selected variables from the M-estimation method. In addition, the S and MM-based criteria select X_{12} education matches work experience; X_{14} use of satellite or cable TV; X_{15} use of personal computer; and X_{16} use of mobile telephone to explain the current monthly earnings.

Table 3.11: *Employment data in East-central Europe with outliers removed. The selected explanatory variables from the classical AIC, AIC with M-estimation, S-estimation and MM-estimation, tuned for a 30% breakdown point.*

Criteria		Selected variables	
	Best model	Second best model	Third best model
AIC	X_2,X_3,X_7,X_8, X_{15},X_{16}	X_2,X_3,X_7,X_8,X_{10}, X_{15},X_{16}	X_2,X_3,X_7,X_{10}, X_{15},X_{16}
AIC.M	X_1,X_3,X_4,X_8,X_{10} ,X_{11},X_{12},X_{15}	X_3,X_7,X_9,X_{10},X_{11}	X_3,X_4,X_8,X_{10},X_{11}, X_{12},X_{14}, X_{15}
AIC.S	X_3,X_4,X_{10},X_{11}, X_{12},X_{15}	X_3,X_4,X_{10},X_{11},X_{12}, X_{15},X_{16}	X_3,X_6,X_7,X_9,X_{12}, X_{15},X_{16}
AIC.MM	X_3,X_4,X_{10},X_{11}, X_{12},X_{15}	X_3,X_4,X_{10},X_{11},X_{12}, X_{15},X_{16}	X_3,X_6,X_7,X_9,X_{12}, X_{15},X_{16}

We have refitted all 2^{15} models with a combination of any of these explanatory variables for the cleaned data (outliers removed) and computed several AIC values for each model. The best three selected models based on each AIC method are given in Table 3.11.

The classical AIC method now also selects models with more variables than when the outliers were still present (Table 3.10), indeed the three best models contain six or seven explanatory variables. Also AIC based on M-estimation selects a model with eight variables as the best one.

3.5.4 Hofstedt's highway data

We have used Hofstedt's highway data that are available from the R library `alr3` as `data(highway)` (see also Weisberg, 2005). In this dataset there are 39 observations on several highway related measurements. The response variable is the accident rate per million vehicle miles in the year 1973 and there are 11 potential explanatory variables:

X_1 Average daily traffic count(1000's); X_2 Truck volume as a percentage of the total volume; X_3 Number of lanes of traffic; X_4 Number of access point per mile; X_5 Number of signalized interchanges/mile; X_6 Number of freeway-type interchanges/mile; X_7 The speed limit in 1973; X_8 The length of the segments in miles; X_9 The lane width in feet; X_{10} Width in feet of outer shoulder on the roadway; X_{11} An indicator of the type of roadway or the source of funding for the road.

We have fitted all 2^{11} possible models with a combination of any of these covariates and computed several AIC values for each model. We have checked the outliers in this data set using studentized deleted residuals criteria and found that the absolute value of the standardized residuals is larger than the Bonferroni critical value of the t distribution, $t(1-\alpha/2n; n-p-1) = t(1-0.1/78; 39-11-1) \simeq 3.3$ for 4 observations. These observations (10% of the data) are considered as vertical outliers. We used the chi-square plot to detect multivariate outliers as in Garrett (1989). This method detects 14 observations as outliers in Xs in this data.

Table 3.12: *Highway data. The selected explanatory variables from best three models based on classical AIC, AIC with M-estimation, S-estimation and MM-estimation using estimators with 50% breakdown point.*

Criteria	Selected variables		
	Best model	Second best model	Third best model
AIC	X_4,X_5,X_7,X_8	X_2,X_4,X_7,X_8	X_4,X_7,X_8
AIC.M	X_5,X_7,X_{11}	X_5,X_6,X_7	X_3,X_5,X_7,X_{11}
AIC.S	X_2,X_3,X_5,X_6,X_7, X_9,X_{10}	X_1,X_2,X_4,X_5,X_9, X_{10},X_{11}	X_4,X_5,X_8,X_{11}
AIC.MM	X_3,X_4,X_8,X_9,X_{10}	X_4,X_5,X_8,X_{11}	X_1,X_3,X_4,X_5,X_8,X_{11}

The classical AIC selects a model with four explanatory variables, see Table 3.12, and thus omits seven potential explanatory variables. The robust model selection strategies as given in this chapter select models with more variables. AIC based on M-estimation selects a model with three variables. For this example, two of the selected variables coincide with those of AIC, the other one is different. All of the best five models based on AIC and AIC.M contain only few variables (3, 4 or 5 variables based on AIC and 4 or 5 variables based on AIC.M).

Table 3.13: *Highway data. The selected explanatory variables from highest ranked models based on AIC.S, AIC.US, AIC.MM and AIC.UMM using estimators with 30% breakdown point.*

Variables in the selected models											Number of	Rank of AIC			
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	variables	S	US	MM	UMM
0	1	1	0	1	1	1	1	0	0	1	7	1	1	1101	1059
1	1	0	0	1	0	1	0	1	1	1	7	2	2	684	636
0	1	1	1	0	0	0	1	1	0	0	5	3	3	216	195
0	0	1	1	0	0	0	1	1	1	1	6	4	4	1	1
0	0	0	1	1	0	0	1	0	0	1	4	5	5	2	2
0	0	1	1	0	0	0	1	1	1	0	5	6	7	3	4
1	0	1	1	1	0	0	1	0	0	1	6	7	9	4	6
0	0	1	1	1	0	0	1	0	0	1	5	8	6	5	3

Table 3.13 presents the five best models as ranked by their AIC values, using AIC with S, uniform S, MM and uniform MM-estimators. The corresponding ranks given by AIC and AIC.M are large for these same models, indicating low preference. AIC.S and AIC.US select the best model with seven variables, this is for the situation where the breakdown point of the estimators is tuned to 30% to accommodate the about 10% of outliers in the data. The model selected by AIC.MM and AIC.UMM corresponds to the 4th ranked model by AIC.S and contains six variables.

3.6 Discussion

In this chapter the definition of the AIC is extended to be used with S- and MM-estimators.

It turns out that the classical (non-robust) AIC works well for data sets with only few outlying observations, and with data where the outliers are only in the explanatory variables. The use of AIC based on M-estimation is not encouraged for data sets with high contamination levels of outliers, based on our simulation results (this holds for all considered variants of the criterion), The versions of AIC that use robust scale estimators arising from S- and MM-estimators perform well in the comparison. For these methods, the breakdown point of the estimation method should be tuned in accordance with the percentage of outliers in the data. These methods are particularly useful when there are outliers in the response variable.

In line with the known properties of the non-robust AIC, these versions of AIC based on S- and MM-estimators, have the tendency to slightly overfit, which ensures that no important variables are lost when this method is used as a screening step to indicate potential important variables in a full analysis of the data. In our simulation studies, the average number of redundant variables in overfitted models was between 2 and 3. The proposed AIC method based on S- and MM-estimators gave good results both for independent and dependent explanatory variables, for both sample sizes considered as well as for the different numbers of true and redundant variables in the simulated models. While this is a limited simulation study only, we expect the conclusions to hold to similar modelling situations as well.

While our study has focussed on the AIC as a variable selection tool, it might be of interest to extend other robust variable selection methods that currently mainly deal with M-estimators, to more advanced robust estimation methods, such as S- or MM-estimators.

Chapter 4

Robust estimation and a conditional Akaike information criterion for linear mixed models

We study model selection on both the fixed and random effects in the setting of linear mixed models that are estimated using outlier robust S-estimators. The derived marginal and conditional information criteria are in the style of Akaike's information criterion but avoid the use of a fully specified likelihood by a suitable S-estimation approach that minimizes a scale function. We derive the appropriate penalty terms and provide an implementation using R. The setting of semiparametric additive models fit with penalized regression splines, attractive because of its link with mixed models, is worked out as a specific application. Simulated data and real data examples illustrate the effectiveness of the proposed criteria.

4.1 Introduction

We consider mixed linear models of the form $Y = X\beta + Zu + \varepsilon$, where u and ε are independent random variables, not necessarily normally distributed. Outlying values may be present in either u or ε. Variable selection in mixed linear models by means of the Akaike information criterion (AIC, Akaike, 1973) which is defined as minus twice the value of the maximized log-likelihood of the model plus twice the number of estimated parameters in the model, may be done using the marginal log-likelihood of Y. Vaida and Blanchard (2005) have shown that in linear mixed models the resulting marginal AIC is not appropriate for variable selection when the random effects are of interest. They proposed the conditional Akaike information which uses the conditional likelihood of the response Y given the random effects u. The penalty term in the conditional AIC is related to the effective degrees of freedom of a linear mixed model (Hodges and Sargent, 2001). Liang et al. (2008) have proposed a corrected conditional AIC that accounts for the estimation of the variance components. Greven and Kneib (2010) study the theoretical properties of both the marginal and the conditional corrected AIC for the selection of variables in linear mixed models, and they provide a computationally feasible penalty term. All of the mentioned papers use maximum likelihood or restricted maximum likelihood for estimation.

In this chapter we derive a marginal and conditional AIC for linear mixed models that no longer requires likelihood based estimation methods. In particular, we work with robust S-estimators that can accommodate the presence of outliers in (i) the response values, (ii) the random effects. We derive an expression for the penalty term that explicitly takes the estimation of the variance components into account and that can be computed in a straightforward way.

4.2 S-estimation in linear mixed models

We model the vector of observations for the ith subject, $i = 1, \ldots, n$, as

$$Y_i = X_i\beta + \sum_{j=1}^{r} Z_{ij}u_{ij} + \varepsilon_i, \tag{4.1}$$

where Y_i has length m_i, X_i is a $m_i \times p$ design matrix of fixed effects, Z_{ij} is a $m_i \times q_j$ design matrix for the random effects. The p-vector β is fixed, while the q_j-vectors u_{ij} are random with mean zero and variance matrix G_j. The random error ε_i has mean zero, and its variance matrix is denoted by R_i. The total number of observations is equal to $N = \sum_{i=1}^{n} m_i$, resulting in vectors Y and ε of length N, a $N \times p$ design matrix $X = (X_1, \ldots, X_n)^t$ for the fixed effects, a $m_i \times q$ design matrix $Z_i = (Z_{i1}, \ldots, Z_{ir})$ for the random effects, $u_i = (u_{i1}^t, \ldots, u_{ir}^t)^t$ is a $q \times 1$ vector. We denote $Z = \text{diag}(Z_1, \ldots, Z_n)$, Z is a $N \times nq$, $u = (u_1^t, \ldots, u_n^t)^t$ is a $nq \times 1$ vector, $G_i = \text{diag}(G_1, \ldots, G_r)$, $G = \text{diag}(G_1, \ldots, G_n)$, and let $q = \sum_{j=1}^{r} q_j$. We assume that the set of random effects $\{u_{ij}; i = 1, \ldots, n, j = 1, \ldots, r\}$ and the set of error terms $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent, that $\text{Var}(u_{ij}) = G_j = \sigma_j^2 I_{q_j}$ and that $\text{Var}(\varepsilon) = R = \sigma_0^2 I_N$, with I_N the identity matrix with N rows. We define $R_i = \sigma_0^2 I_{m_i}$ and $V = \text{Var}(Y) = R + ZGZ^t$. In the balanced case where all $m_i = m$, we define the $m \times m$ matrices $\text{Var}(Y_i) = V_0$, and $\text{Var}(\varepsilon_i) = R_0 = \sigma_0^2 I_m$, for $i = 1, \ldots, n$, $j = 1, \ldots, r$.

The most frequent assumption in linear mixed models is that both the errors ε and the random effects u have Gaussian distributions. Outliers, extreme observations that are unlike most of the other observations in the sample, may occur for any of the observed random effects as well as for the observed error terms. Consequently, in such case the distributions of the errors and/or random effects may be non-Gaussian. Welsh and Richardson (1997) present several approaches and give an overview on how to robustly estimate parameters in linear mixed models. In this chapter we use the high-breakdown S-estimators of Copt and Victoria-Feser (2006) for both the parameters of the mean as well as for the variance components. For the purpose of developing a conditional AIC, we need in addition the predictions of the random effects.

Copt and Victoria-Feser (2006) work with the marginal likelihood in the linear mixed model where all $m_i = m$, and define the S-estimator for the vector β and the variance components $\sigma^2 = (\sigma_0^2, \ldots, \sigma_r^2)$ as the values for β and σ^2 that minimize $\det(V_0)$ subject to the constraint

$$\frac{1}{n}\sum_{i=1}^{n} \rho[\{(Y_i - X_i\beta)^t V_0^{-1}(Y_i - X_i\beta)\}^{1/2}] = b_1. \tag{4.2}$$

An appropriate choice of the function ρ and of the value of b_1 will lead to robust estimators with a high breakdown point.

The loss function ρ_0 is a function that is even, continuously differentiable, non-decreasing on $[0, \infty)$, satisfies that $\rho_0(0) = 0$ and is bounded for above by 1, that is, $\sup_{\varepsilon \in \mathbb{R}} \rho_0(u) = 1$. We define $b_1 = E_{F_0}[\rho_0(\varepsilon)]$ to ensure consistency of the scale estimator under the central model F_0 and assume that $\epsilon_0 < b_1 < 1 - \epsilon_0$, here F_0 is the cumulative distribution function of ε. The notation E_{F_0} means that the expectation is computed with respect to F_0. When $\rho(x) = x^2$, the estimation method boils down to maximum likelihood estimation. A translated Tukey biweight function ρ is proposed by Rocke (1996) and is used in this chapter. A translated Tukey biweight function can control the probability of an estimator giving a null weight to extreme observation and it is called asymptotic rejection probability(ARP). The translated Tukey biweight ρ function is given by,

$$\rho(d; c.M) = \begin{cases} \frac{d^2}{2}, & 0 \le d \le M \\ \rho_{M \le d \le M+c}(d; c, M), & M \le d \le M + c \\ \frac{M^2}{2} + \frac{c(5c+16M)}{30}, & d > M + c, \end{cases}$$

with $M + c < \infty$ and

$$\begin{aligned} \rho_{M \le d \le M+c}(d; c, M) = {} & \frac{M^2}{2} - \frac{M^2(M^4 - 5M^2c^2 + 15c^4)}{30c^4} \\ & + d^2\left(0.5 + \frac{M^4}{2c^4} - \frac{M^2}{c^2}\right) + d^3\left(\frac{4M}{3c^2} - \frac{4M^3}{3c^4}\right) \\ & + d^4\left(\frac{3M^2}{2c^4} - \frac{1}{2c^2}\right) - \frac{4Md^5}{5c^4} + \frac{d^6}{6c^4}. \end{aligned}$$

This translated Tukey biweight ρ function leads to the weight function,

$$u(d;c.M) = \begin{cases} 1, & 0 \leq d \leq M \\ \left(1-\left(\frac{d-M}{c}\right)^2\right)^2, & M \leq d \leq M+c \\ 0, & d > M+c, \end{cases}$$

where the constants c and M can be chosen to achieve the desired breakdown point and ARP.

We consider the conditional model for $Y|u$. In a first setting we assume that the random effects have a normal distribution $u_j \sim N(0, G_j)$. The conditional S-estimator (predictor) for the vectors β, u and the variance σ_0^2 are those parameter values that minimize det(R_0)=$\mid R_0 \mid$ subject to the constraint

$$\frac{1}{n}\sum_{i=1}^{n} \rho[\{(Y_i - X_i\beta - Z_iu)^t R_0^{-1}(Y_i - X_i\beta - Z_iu)\}^{1/2}] = b_1. \tag{4.3}$$

By following the idea of Henderson (unpublished paper, 1973) we provide an iterative system that gives in addition to estimators of (β, σ^2) the predictions of the random effects. In a likelihood setting the Henderson approach starts by phrasing the joint likelihood of (Y, u) as the product of the likelihood of $Y|u$ and the likelihood of u. In our context this leads to the following joint Lagrangian function, the maximization of which leads to estimators and predictors simultaneously,

$$\mathrm{L}_{\text{joint}}(\beta, u, \sigma^2) = \log|R_0| + \frac{\tau_1}{n}\sum_{i=1}^{n}\{\rho(d_i) - b_1\} + \log|G| + u^t G^{-1} u, \tag{4.4}$$

where $d_i = d_i(\beta, u, R_0) = \{(Y_i - X_i\beta - Z_iu)^t R_0^{-1}(Y_i - X_i\beta - Z_iu)\}^{1/2}$ and τ_1 is a Lagrange multiplier. The estimators of β and σ^2 that result from this procedure are identical to those obtained by Copt and Victoria-Feser (2006) by using a marginal Lagrangian (4.4) and by omitting the part related to the marginal density of u. The derivation of the estimators is given in Appendix 4.6.

Result 4.1. The S-estimators $\widehat{\beta}$ and $\widehat{\sigma}^2$ of the fixed effect parameters β and of the variance components σ^2 and the S-predictions $\widehat{u}$ of the random effects u in the linear mixed model (4.1) obtained by maximizing the joint Lagrangian (4.4), assuming normality of the random effects, are equivalently obtained by iteratively solving the following set of equations,

$$\begin{aligned}
\widehat{\beta} &= (X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\widehat{W}\widehat{V}^{-1}Y && (4.5)\\
\widehat{u} &= \frac{\widehat{\tau}_1}{2n}\widehat{G}Z^t\widehat{W}\widehat{V}^{-1}(Y-X\widehat{\beta}) && (4.6)\\
\widehat{\sigma}_0^2 &= (\widehat{d}^t\widehat{W}\widehat{d})^{-1}(Y-X\widehat{\beta}-Z\widehat{u})^t\widehat{W}(Y-X\widehat{\beta}-Z\widehat{u}) && (4.7)\\
\widehat{\sigma}_j^2 &= \widehat{u}_j^t\widehat{u}_j/q_j, && (4.8)
\end{aligned}$$

where $\widehat{W}=\mathrm{diag}_{i=1,\ldots,n}(\psi(\widehat{d}_i)/\widehat{d}_iI_m)$, $\widehat{d}_i=d_i(\widehat{\beta},\widehat{u},\widehat{R}_0)$, $\widehat{d}=(\widehat{d}_1,\ldots,\widehat{d}_n)^t$ and the vector $\widehat{u}$ decomposes in components $\widehat{u}_j$ with length q_j, $j=1,\ldots,r$, $\widehat{\tau}_1=2mn(\widehat{d}^t\widehat{W}\widehat{d})^{-1}$,

$$\widehat{V}=\widehat{R}+Z(\frac{\widehat{\tau}_1}{2n}\widehat{G})Z^t\widehat{W}. \qquad (4.9)$$

When $\rho(t)=t^2$ the S-scale estimator $\widehat{\sigma}_0$ reduces to the sample standard deviation. In this case we have that $\widehat{W}=2\,I_n$ and that $\widehat{\tau}_1=n$. Hence, as expected, $\widehat{V}=\widehat{R}+Z\widehat{G}Z^t$ and (4.5) and (4.6) correspond to the maximum likelihood fixed and random effects formulae where $\widehat{\beta}_{ML}=(X^t\widehat{V}^{-1}X)^tX^t\widehat{V}^{-1}Y$ and $\widehat{u}_{ML}=\widehat{G}Z^t\widehat{V}^{-1}(Y-X\widehat{\beta}_{ML})$.

To accommodate possible outliers on the random effects we consider robust S-prediction of the random effects simultaneous with S-estimation of the fixed effects and variance components. For this purpose we define a new joint Lagrangian function that is to be optimized over β, u and σ^2,

$$\mathrm{L}_{\mathrm{joint},2}(\beta,u,\sigma^2)=\log|R_0|+\frac{\tau_1}{n}\sum_{i=1}^{n}\{\rho(d_i)-b_1\}+\log|G|+\frac{\tau_2}{r}\sum_{j=1}^{r}\{\rho_2(d_{2j})-b_2\}. \qquad (4.10)$$

Here $d_{2,j}=(u_j^tG_j^{-1}u_j)^{1/2}$, ρ and ρ_2 are both Tukey's bi-square family loss functions, which might be taken to be different functions, b_1 and b_2 are constants associated with the breakdown point of the estimator. Generally b_1 and b_2 are is defined by $b=E(\rho(\sqrt{U}))$, where U is a Chi-squared distribution with p degrees of freedom, p is number of parameters in the model.. Here τ_1 and τ_2 are Lagrange multipliers.

Result 4.2. The S-estimators $\widetilde{\beta}$ and $\widetilde{\sigma}^2$ of the fixed effect parameters β and of the variance components σ^2 and the S-predictions $\widetilde{u}$ of the random effects u in the linear mixed model (4.1) obtained by maximizing the joint Lagrangian (4.10), without assuming normality, are equivalently obtained by iteratively solving the following set of equations,

$$\begin{aligned}
\widetilde{\beta} &= (X^t\widetilde{W}\widetilde{V}^{-1}X)^{-1}X^t\widetilde{W}\widetilde{V}^{-1}Y & (4.11)\\
\widetilde{u} &= \frac{r\widetilde{\tau}_1}{n\widetilde{\tau}_2}\left(\widetilde{G}^{-1/2}\widetilde{W}_2\widetilde{G}^{-1/2}\right)^{-1} Z^t\widetilde{W}\widetilde{V}^{-1}(Y - X\widetilde{\beta}) & (4.12)\\
\widetilde{\sigma}_0^2 &= (\widetilde{d}^t\widetilde{W}\widetilde{d})^{-1}(Y - X\widetilde{\beta} - Z\widetilde{u})^t\widetilde{W}(Y - X\widetilde{\beta} - Z\widetilde{u}) & (4.13)\\
\widetilde{\sigma}_j^2 &= \frac{\widetilde{\tau}_2}{2rq_j}\widetilde{u}_j^t\widetilde{W}_{2j}\widetilde{u}_j & (4.14)
\end{aligned}$$

where $\widetilde{W} = \text{diag}_{i=1,\ldots,n}(\psi(\widetilde{d}_i)/\widetilde{d}_i I_m)$, $\widetilde{d}_i = d_i(\widetilde{\beta}, \widetilde{u}, \widetilde{R}_0)$, $\widetilde{d} = (\widetilde{d}_{11}, \ldots, \widetilde{d}_{1n})^t$, $\widetilde{\tau}_1 = 2nm(\widetilde{d}^t\widetilde{W}\widetilde{d})^{-1}$,

$$\widetilde{V} = \widetilde{R} + Z\left(\frac{r\widetilde{\tau}_1}{n\widetilde{\tau}_2}(\widetilde{G}^{-1/2}\widetilde{W}_2\widetilde{G}^{-1/2})^{-1}\right)Z^t\widetilde{W}, \qquad (4.15)$$

$\widetilde{d}_{2j} = \widetilde{G}_j^{-1/2}\widetilde{u}_j$, $\widetilde{d}_2 = (\widetilde{d}_{21}, \ldots, \widetilde{d}_{2r})^t$, $\widetilde{W}_2 = \text{diag}_{j=1,\ldots,r}(\psi_2(\widetilde{d}_{2j})/\widetilde{d}_{2j} I_{q_j})$ and $\widetilde{\tau}_2 = 2rq\left(\widetilde{d}_2^t\widetilde{W}_2\widetilde{d}_2\right)^{-1}$. Here ψ and ψ_2 are the first derivatives of, respectively, ρ and ρ_2.

When $\rho_2(x) = x^2$, the estimators presented in Result 4.2 coincide with those of Result 4.1.

4.3 AIC for S-estimation in linear mixed models

When both the error terms and the random effects are Gaussian,

$$\begin{aligned}
\log f(Y|\,\widehat{\beta}, \widehat{u}, \widehat{R}) &= -\frac{1}{2}N\log(2\pi) - \frac{1}{2}\log(|\widehat{R}|)\\
&\quad -\frac{1}{2}(Y - X\widehat{\beta} - Z\widehat{u})^t\widehat{R}^{-1}(Y - X\widehat{\beta} - Z\widehat{u}), (4.16)
\end{aligned}$$

with maximum likelihood or restricted maximum likelihood estimators $\widehat{\beta}$, $\widehat{u}$, $\widehat{\sigma}_\varepsilon^2$.

A marginal AIC follows from an immediate application of the original AIC (Akaike, 1973), it counts the number of estimated parameters to be used in the penalty part of the criterion and it uses the marginal likelihood of Y, with maximum likelihood estimators inserted for the unknown parameters,

$$\text{mAIC} = -2\,\log f_Y(Y;\widehat{\beta},\widehat{V}) + 2(p+r+1). \tag{4.17}$$

Vaida and Blanchard (2005) obtain for variable selection when the random effects are of interest a conditional AIC, defined as

$$\text{cAIC} = -2\,\log f_{Y|u}(Y|\,\widehat{\beta},\widehat{u},\widehat{R}) + 2(\text{Tr}(H)+1), \tag{4.18}$$

where $f_{Y|u}$ is the conditional likelihood for $Y|u$ and $H = C(C^tR^{-1}C + B)^{-1}C^tR^{-1}$, where $C = (X,Z)$ and $B = \text{diag}(0_p, G^{-1})$, where 0_p is a vector of zeros of length p. The added value of 1 in the penalty term reflects the estimation of the error variance σ_0^2.

The boundedness of the function ρ for S-estimation has as a consequence that the transformation $\exp(-\rho)$ does not lead to a density function since its integral will be infinite. Hence a substitution of the model's density f by $\exp(-\rho)$ in expressions for the AIC is not valid when working with S-estimators, in contrast to the case of M-estimation where the unbounded ρ functions lead to valid density functions. Motivated by the m-variate normal likelihood with mean function μ and variance matrix Σ, a cAIC expression for M-estimation (Ronchetti, 1997) would replace the sum of the Mahalanobis distances by $\sum_{i=1}^n \rho(y_i;\mu,\Sigma)$. For S-estimation this, however, is the constant number nb. Indeed, the marginal multivariate S-estimator of (β,u,V) is defined by the minimization of $|V_0|$ subject to the constraint (4.2), while the conditional multivariate S-estimator of (β,u,R) is defined by the minimization of $|R_0|$ subject to the constraint (4.3). S-estimation requires a different approach towards defining the AIC. Following Tharmaratnam and Claeskens (2011a), we come to the definition of a marginal and conditional AIC for use with S-estimation as

$$\text{mAIC.S1} = 2\log|\ \widehat{V}\ | + 2\,(p+q+1), \tag{4.19}$$

$$\text{cAIC.S1} = 2\log \mid \widehat{R} \mid +2\,\text{Tr}(\widehat{H}_S + 1), \tag{4.20}$$

where, from application of Result 4.1, the matrix $\widehat{H}_S = (I_N - \widehat{R}\widehat{V}^{-1}\widehat{P})$, with $\widehat{P} = I_N - X(X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\widehat{W}\widehat{V}^{-1}$.

When robustness in both ε and u is considered we use instead the matrices $\widetilde{R}, \widetilde{V}, \widetilde{W}$ (see Result 4.2), with the corresponding matrices $\widetilde{H}_S$ and $\widetilde{P}$, leading to

$$\text{mAIC.S2} = 2\log \mid \widetilde{V} \mid +2\,(p + r + 1), \tag{4.21}$$

$$\text{cAIC.S2} = 2\log \mid \widetilde{R} \mid +2\,\text{Tr}(\widetilde{H}_S + 1), \tag{4.22}$$

Liang et al. (2008) obtain that $\Phi_0 = \text{Tr}(\partial\widehat{Y}/(\partial Y))$ is a better penalty term than $\text{Tr}(H) + 1$, since it takes the effect of the estimation of the variance components into account. This is further studied and explicitly computed by Greven and Kneib (2010, Thm. 3). A large part of the difficulty in arriving at computable expressions is that the estimators $(\widehat{\beta}, \widehat{u}, \widehat{\sigma}^2)$ are also dependent on Y. We can write the corrected conditional AIC from Greven and Kneib (2010) as,

$$\text{ccAIC} = -2\,\log f_{Y|u}(Y|\,\widehat{\beta}, \widehat{u}, \widehat{R}) + 2\,\Phi_0. \tag{4.23}$$

For the case of S-estimation we explicitly obtain the generalized degrees of freedom for both situations with one or two levels of robustness. In these calculations we consider σ_ε^2 to be unknown, and hence we do not need any additional adjustments in the penalty Φ_0 to account for the estimation of the error variance.

Theorem 4.1. *The generalized degrees of freedom $\Phi_S = Tr(\partial\widehat{Y}/(\partial Y))$ when the estimators are obtained via the joint Lagrangian* (4.4) *are computed as:*

$$\phi_{S1} = Tr\left(I_N - \widehat{R}\widehat{V}^{-1}\widehat{P} - B\right) \tag{4.24}$$

where

$$\begin{aligned} B &= \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial Y}Y \\ &= \left(\frac{\partial\widehat{R}\widehat{V}^{-1}\widehat{P}Y}{\partial Y_1}, \frac{\partial\widehat{R}\widehat{V}^{-1}\widehat{P}Y}{\partial Y_2}, \ldots, \frac{\partial\widehat{R}\widehat{V}^{-1}\widehat{P}Y}{\partial Y_N}\right) \end{aligned}$$

here $B_k = \partial \widehat{R}\widehat{V}^{-1}\widehat{P}Y/\partial Y_k; k = 1, 2, \ldots, N$ *and* B_k *is the* k*th column of the matrix* B.

$$\begin{aligned} B_k &= \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial\sigma_0^2} Y \frac{\partial\widehat{\sigma}_0^2}{\partial Y_k} + \sum_{j=1}^{r} \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial\sigma_j^2} Y \frac{\partial\widehat{\sigma}_j^2}{\partial Y_k} \\ &= D_1 D_{2k} + \sum_{j=1}^{r} D_{3j} D_{4jk}. \end{aligned}$$

Here,

$$\begin{aligned} D_1 &= \left[I_N - \widehat{R}\widehat{V}^{-1}\left((I_N - \widehat{P})D_{v\sigma_0} - D_{w\sigma_0}\right)\right]\widehat{V}^{-1}\widehat{P}, \\ D_{2k} &= -H_{\sigma_0}^{-1} H_{\sigma_0 Y_k}, \\ D_{3j} &= -\tau_1/(2n)\widehat{R}\widehat{V}^{-1}\widehat{P}ZZ^t\widehat{W}\widehat{V}^{-1}\widehat{P}, \\ D_{4jk} &= -H_{\sigma_j}^{-1} H_{\sigma_j Y_k}, \end{aligned}$$

with $\widehat{V}$, $\widehat{W}$ *and* $\widehat{\tau}_1$ *as in Result 4.1,* $\widehat{A}_j = \widehat{\tau}_1/(2n)Z_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}Y$, $D_{v\sigma_0} = \partial\widehat{V}/\partial\sigma_0^2$, $D_{w\sigma_0} = \partial\widehat{W}/\partial\sigma_0^2$, $D_{\tau_1\sigma_0} = \partial\widehat{\tau}_1/\partial\sigma_0^2$, $D_{vY_k} = \partial\widehat{V}/\partial Y_k$, $D_{wY_k} = \partial\widehat{W}/\partial Y_k$, $D_{\tau_1 Y_k} = \partial\widehat{\tau}_1/\partial Y_k$. $D_{(V^{-1}P_k)Y_k} = \partial(\widehat{V}^{-1}\widehat{P}_k)/\partial Y_k$, P_k *is a* k*th column of matrix* P.

Further,

$H_{\sigma_0} = -N/\sigma_0^4 - N^{-1}Y^t\widehat{P}^t\widehat{V}^{-1}\Big[\widehat{W}\widehat{V}^{-1}\widehat{P}YD_{\tau_1\sigma_0} - 2\,\tau_1\widehat{W}Y\widehat{V}^{-1}$
$\{(I_N - \widehat{P})D_{v\sigma_0} - D_{w\sigma_0}\}\widehat{V}^{-1}\widehat{P} + \tau_1 D_{w\sigma_0}\widehat{V}^{-1}\widehat{P}Y\Big]$,
$H_{\sigma_0 Y_k} \quad = \quad -n^{-1}Y_k^t\widehat{P}_k^t\widehat{V}^{-1} \quad \Big[2\,\tau_1\widehat{W}\widehat{V}^{-1}\widehat{P}_k + D_{\tau_1 Y_k}\widehat{W}\widehat{V}^{-1}\widehat{P}_kY_k + 2\,\tau_1\widehat{W}$
$D_{(V^{-1}P_k)Y_k}Y_k + \tau_1 D_{wY_k}\widehat{V}^{-1}\widehat{P}_kY_k\Big]$,
$H_{\sigma_j} = -q_j/\sigma_j^4 - 2\,\widehat{A}_j^t\{\tau_1/(2n)Z_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}Z_j\}\widehat{A}_j$,
$H_{\sigma_j Y_k} \quad = \quad -n^{-1}\widehat{A_j}^t\Big[\tau_1 Z_j^t\widehat{W}\widehat{V}^{-1}\widehat{P_k} + \tau_1 Z_j^t D_{wY_k}\widehat{V}^{-1}\widehat{P}_kY_k + \tau_1 Z_j^t\widehat{W}$
$D_{(V^{-1}P_k)Y_k}Y_k + D_{\tau_1 Y_k}Z_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}_kY_k\Big]$.

Theorem 4.2. *The generalized degrees of freedom Φ_{S2} when the estimators are obtained via the joint Lagrangian* (4.10) *are computed as:*

$$\phi_{S2} = Tr\left(I_N - \widetilde{R}\widetilde{V}^{-1}\widetilde{P} - \widetilde{B}\right) \tag{4.25}$$

where

$$\widetilde{B} = \frac{\partial(\widetilde{R}\widetilde{V}^{-1}\widetilde{P})}{\partial Y}Y = \left(\frac{\partial\widetilde{R}\widetilde{V}^{-1}\widetilde{P}Y}{\partial Y_1}, \frac{\partial\widetilde{R}\widetilde{V}^{-1}\widetilde{P}Y}{\partial Y_2}, \dots, \frac{\partial\widetilde{R}\widetilde{V}^{-1}\widetilde{P}Y}{\partial Y_N}\right)$$

here

$$\begin{aligned}\widetilde{B}_k &= \frac{\partial(\widetilde{R}\widetilde{V}^{-1}\widetilde{P})}{\partial\sigma_0^2}Y\frac{\partial\widetilde{\sigma}_0^2}{\partial Y_k} + \sum_{j=1}^{r}\frac{\partial(\widetilde{R}\widetilde{V}^{-1}\widetilde{P})}{\partial\sigma_j^2}Y\frac{\partial\widetilde{\sigma}_j^2}{\partial Y_k}\\ &= \widetilde{D}_1\widetilde{D}_{2k} + \sum_{j=1}^{r}\widetilde{D}_{3j}\widetilde{D}_{4jk}.\end{aligned}$$

The quantities $\widetilde{D}_1$ and $\widetilde{D}_{2k}$ are the same as in Theorem 4.1 though use the estimators $\widetilde{V}$, $\widetilde{W}_2$, $\widetilde{d}_2$ and $\widetilde{\tau}_2$ as from Result 4.2, $\widetilde{D}_{3j} = -\widetilde{R}\widetilde{V}^{-1}\widetilde{P}\widetilde{D}_{v\sigma_j}\widetilde{V}^{-1}\widetilde{P}$ and $\widetilde{D}_{4jk} = -\widetilde{H}_{\sigma_j}^{-1}\widetilde{H}_{\sigma_jY_k}$.
Here $\widetilde{D}_{v\sigma_j} = \partial\widetilde{V}/\partial\sigma_j^2$, $\widetilde{D}_{\tau_2\sigma_j} = \partial\widetilde{\tau}_2/\partial\sigma_j^2$, $\widetilde{D}_{d_{2j}\sigma_j} = \partial\widetilde{d}_{2j}/\partial\sigma_j^2$, $\widetilde{D}_{W_2\sigma_j} = \partial\widetilde{W}_2/\partial\sigma_j^2$, $\widetilde{D}_{\tau_2Y_k} = \partial\widetilde{\tau}_2/\partial Y_k$, $\widetilde{D}_{d_{2j}Y_k} = \partial\widetilde{d}_{2j}/\partial Y_k$, $\widetilde{D}_{W_{2j}Y_k} = \partial\widetilde{W}_{2j}/\partial Y_k$.
Further,
$\widetilde{H}_{\sigma_j} = -q_j/\widetilde{\sigma}_j^4 + 1/(2r\widetilde{\sigma}_j^2)\widetilde{d}_{2j}^t \Big[\left(\widetilde{\tau}_2/\widetilde{\sigma}_j^2\right)\widetilde{W}_{2j}\widetilde{d}_{2j} - \widetilde{W}_{2j}\widetilde{d}_{2j}\widetilde{D}_{\tau_2\sigma_j} - 2\widetilde{\tau}_2\widetilde{W}_{2j}\widetilde{D}_{d_{2j}\sigma_j} - \widetilde{\tau}_2\widetilde{D}_{W_{2j}\sigma_j}\widetilde{d}_{2j}\Big]$,
$\widetilde{H}_{\sigma_jY_k} = -1/(2r\widetilde{\sigma}_j^2)\widetilde{d}_{2j}^t\Big[\widetilde{W}_{2j}\widetilde{d}_{2j}\widetilde{D}_{\tau_2Y_k} + 2\widetilde{\tau}_2\widetilde{W}_{2j}\widetilde{D}_{d_{2j}Y_k} + \widetilde{\tau}_2\widetilde{D}_{W_{2j}Y_k}\widetilde{d}_{2j}\Big]$.

The generalized degrees of freedom from Theorems 4.1 and 4.2 lead to corrected versions of the conditional AIC,

$$\text{ccAIC.S1} = 2\log\mid\widehat{R}\mid + 2\,\Phi_{S1}, \tag{4.26}$$

$$\text{ccAIC.S2} = 2\log\mid\widetilde{R}\mid + 2\,\Phi_{S2}. \tag{4.27}$$

4.4 Numerical results

4.4.1 Algorithm

An iterative procedure is required to compute the S-estimators in Results 4.1 and 4.2, as is the case for other S-estimation schemes, e.g. in linear regression models. The algorithm to obtain the estimators from Result 4.1 is described in the following steps,

Step 1: Let $\widehat{\beta}^{(0)}$, $\widehat{u}^{(0)}$, $(\widehat{\sigma}_0^2)^{(0)}$ and $(\widehat{\sigma}_j^2)^{(0)}$ be the initial values, for which we use maximum likelihood estimators.

Step 2: Set $k = 0$. Iterate the following steps until convergence:

(i) Compute the $\widehat{d}_i^{(1)}$, weights $\widehat{W}^{(1)}$ and $\widehat{\tau}^{(1)}$ as in Result 4.1.

(ii) Compute $\widehat{\beta}^{(1)}$ and $\widehat{u}^{(1)}$ by substituting $(\widehat{\sigma}_0^2)^{(0)}$, $(\widehat{\sigma}_j^2)^{(0)}$, $\widehat{d}_i^{(1)}$, $\widehat{W}^{(1)}$ and $\widehat{\tau}^{(1)}$ in the equations (4.5) and (4.6).

(iii) Compute $(\widehat{\sigma}_0^2)^{(1)}$, $(\widehat{\sigma}_j^2)^{(1)}$ by substituting $\widehat{\beta}^{(1)}$, $\widehat{u}^{(1)}$, $\widehat{d}_i^{(1)}$, $\widehat{W}^{(1)}$ and $\widehat{\tau}^{(1)}$ in the equations (4.7) and (4.8).

(iv) If either $k = maxit$ (the maximum number of iterations) or $\|\widehat{\beta}^{(k)} - \widehat{\beta}^{(k+1)}\| < \epsilon \, \|\widehat{\beta}^{(k)}\|$ where $\epsilon > 0$ is a fixed small constant (the tolerance level) , then set $\widehat{\beta}^{F} = \widehat{\beta}^{(k)}$ and stop.

Step 3: Compute the final estimators $(\widehat{\sigma}_0^2)^{(F)}$, $(\widehat{\sigma}_j^2)^{(F)}$ by substituting $\widehat{\beta}^{(F)}$, $\widehat{u}^{(F)}$, $\widehat{d}_i^{(F)}$, $\widehat{W}^{(F)}$ and $\widehat{\tau}^{(F)}$ in the equations (4.7) and (4.8).

We used a similar algorithm for obtaining the estimates from Result 4.2. We have coded the above algorithm in R. In our experience the above algorithm converges without problems in the majority of the cases. The algorithm with $\epsilon = 10^{-6}$ and $maxit = 500$ converges generally in less than 50 iterations. For all of our simulation experiments, we have never encountered a situation where the algorithm diverged.

4.4.2 Simulation results – S-estimators

Case 1. We consider a model $Y = m_1(x) + \varepsilon$, with $x = (x_1, \ldots, x_6)$, the true mean function $m_1(x) = 1 + d\, sin(\pi x_1) + x1 + x_2 + x_3$, where $d = 15$.

The covariates are generated from a multivariate normal distribution with mean vector $\mu = (1, 2, \cdots, 6)$ and the covariance matrix Σ,

$$\Sigma = \begin{pmatrix} I_3(0.6) & 0 \cdot 1_{3\times 3} \\ 0 \cdot 1_{3\times 3} & I_3(0.3) \end{pmatrix},$$

while the errors ε come from a $N(0,1)$ distribution. The sample size $n = 100$. To investigate the robustness of the estimation method against outliers, we generated, using different percentages of outliers (0%, 10%, 20%, 30% and 40%), for each of the simulated cases outlying error terms from a normal distribution with mean 100 and standard deviation 0.5. We fit a cubic thin plate regression splines model

$$Y_i = \beta_0 + \sum_{j=1}^{6} \beta_j x_{ji} + \sum_{k=1}^{K} u_k \mid x_{1i} - \kappa_k \mid^3 + \varepsilon_i,$$

using ML estimation and the S1 and S2-estimation methods, see Results 4.1 and 4.2.

We use a mixed model formulation. where the u_k are random variables with mean zero and variance σ_u^2. We placed the knots according to the quantiles of the data. For sample size $n = 100$ there were 24 knots. For the non-robust estimation methods we have used the R library SemiPar, function `spm`, for the robust estimation methods, we used our own implementation of the algorithm in Section 4.4.1.

We compute the median squared prediction error (MSPE) to check the fit of the estimated models. Denoting $\widehat{m}_r(x)$ the estimated value of $m(x)$ for simulation run r, $(r = 1, 2, \ldots, 1000)$, the MSPE for the rth simulation run is defined by,

$$\text{MSPE}_r = \text{median}\{[m(x_i) - \widehat{m}_r(x_i)]^2, i = 1, \ldots, n\}.$$

To visualize the variability of the obtained estimates, we construct boxplots on the log scale of the MSPE values, see Figure 4.1.

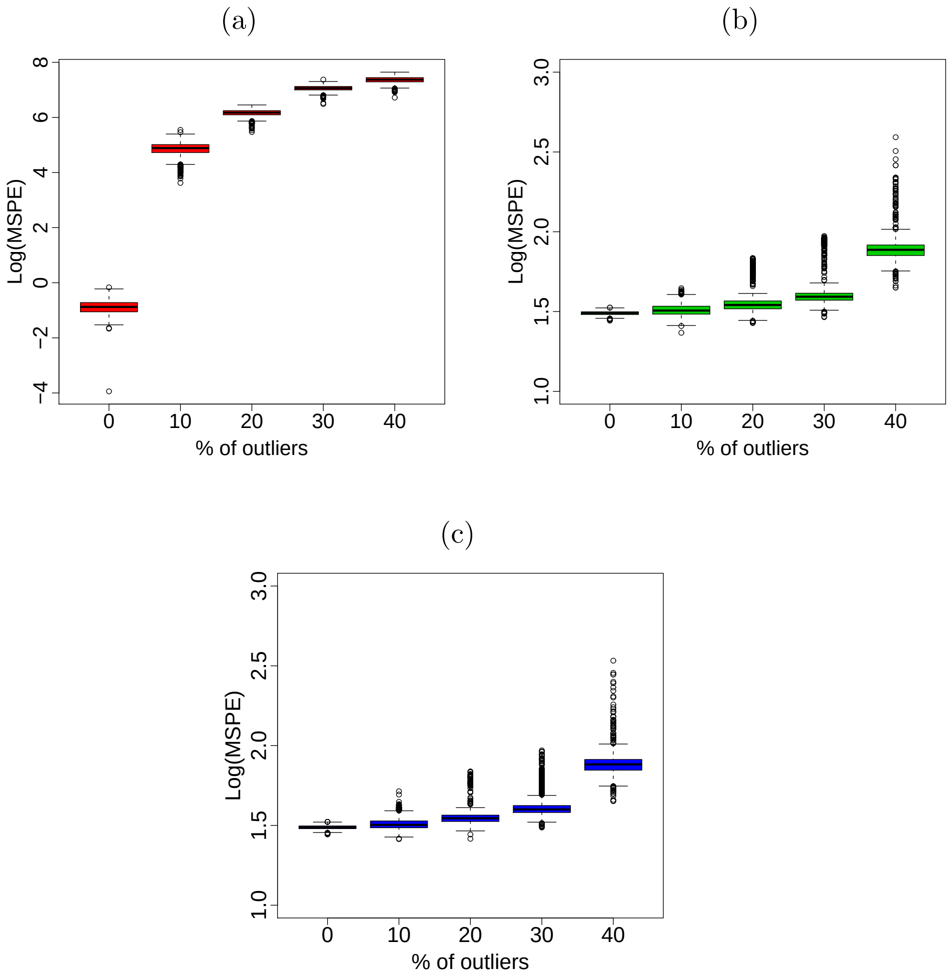


Figure 4.1: *Box plots of log scale of the median squared prediction error using (a) ML-estimation, (b) S1-estimation and (c) S2- estimation for samples with mean structure $m_1(x)$, error distribution $N(0, 1^2)$ and different percentages of outliers $N(100, 0.5^2)$, for sample size $n = 100$.*

It is observed that the MSPEs of the S1-estimators and S2-estimators remain stable as the proportion of contamination increases, though they become more variable for 40% of outliers. The ML-estimator's MSPEs increase in the presence of outliers, even with only 10% of outliers. Both S-estimation methods perform about equally well.

4.4.3 Simulation results – Variable selection

We use different versions of the AIC for linear mixed models with the following notation:
mAIC – Marginal AIC based on the ML-estimator ((4.17), Vaida and Blanchard (2005))
cAIC – Conditional AIC based on the ML-estimator ((4.18),(Vaida and Blanchard, 2005))
ccAIC – Corrected conditional AIC based on the ML-estimator using Φ_0 (Greven and Kneib, 2010)
mAIC.S1 – Marginal AIC based on the S1-estimator (Section 4.3, (4.19))
cAIC.S1 – Conditional AIC based on the S1-estimator (Section 4.3, (4.20))
ccAIC.S1 – Corrected conditional AIC based on the S1-estimator (Section 4.3, (4.26)).

Data are generated according to three settings. For case 1, see Section 4.4.2. Case 2 is taken from Greven and Kneib (2010), where $m_2(x) = 1 + x + 2d(0.3 - x)^2$. The covariate values x are generated from a uniform distribution on the interval $[0, 1]$. In the model, d is a constant and increasing values of d correspond to the increased non-linearity. We generate 11 different models corresponding to $d = (0, 5, 10, \ldots, 50)$. The model is linear in x when $d = 0$. In the case of no outliers, the error terms ε have a standard normal distribution.

Case 3: $m_3(x) = 1 + 2d_1\,cos(\pi x_1) + d_2\,sin((0.5 - x_2)^2) + x_3$, with $d_1 = 15, d_2 = 25$. The covariates $x_1, \ldots, x_6$ are generated from a multivariate normal distribution which is the same as in case 1. The full model that is fit to the data is

$$Y_i = \beta_0 + \sum_{j=1}^{6} \beta_j x_{ji} + \sum_{k=1}^{K} u_{1k} \mid x_{1i} - \kappa_k \mid^3 + \sum_{k=1}^{K} u_{2k} \mid x_{2i} - \kappa_k \mid^3 + \varepsilon_i,$$

that is, cubic thin plate splines are used to model smooth functions of x_1, x_2, while $x_3, \ldots, x_6$ enter the model in a linear way.

We fit model with all possible combinations of the six covariates, resulting in $(2^6 - 1)$ different models.

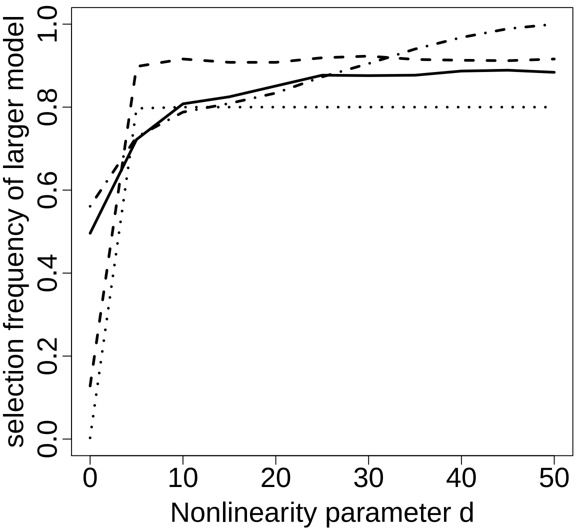
We first discuss the results from case 2. For each value of the constant d, for each simulated data set, we use the AIC, ccAIC, mAIC.S1 and ccAIC.S1 to decide on either the linear model (with $d = 0$) or the more complex model (with the given value of d). To assess the performance of the marginal and the conditional AIC for distinguishing between linear and non-linear models, we compute the frequency of selecting the nonlinear model for each d value.

We use 1000 simulated data sets for both cases with (a) no outliers and (b) 20% outliers on the error terms, generated from a $N(100, 0.5^2)$ distribution for the sample size $n = 100$.

From Figure 4.2 we observe that the corrected conditional AIC selects a larger proportion of nonlinear models than the marginal AIC (which is the true model when $d \neq 0$).

This holds for both maximum likelihood estimators and S1-estimators. In these penalized spline models, the random effects correspond to the spline coefficients. The conditional AIC is better suited to decide on the inclusion of random effects (i.e. nonlinear effects in this setup) than the marginal AIC. The results do not change much for different values of d.

(a)



(b)

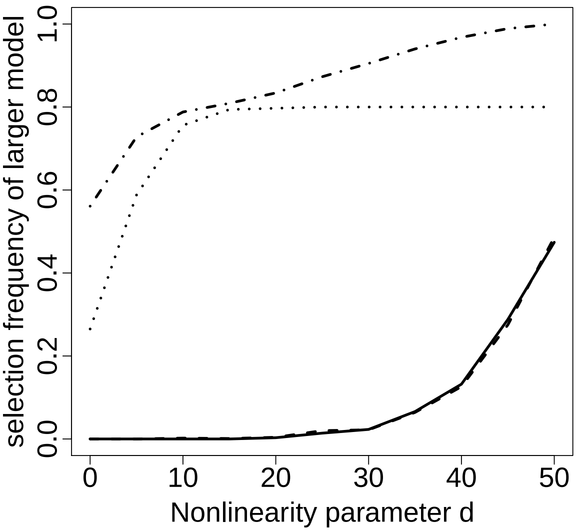


Figure 4.2: *Proportion of selected larger models from marginal AIC (solid line), ccAIC (dashed line), mAIC.S1 (dotted line) and ccAIC.S1 (dot-dashed line) with mean function $m_2(x)$. (a) no outliers in the data, (b) 20% of outliers in the error variables ε.*

For cases 1 and 3 there are six covariates used for fitting the models, some of them are redundant. A summary of the simulation results for these cases is provided by reporting the proportions of selected models that are
(C) Correct fit – The true model only.
(O) Over fit – Models containing all the variables in the true model plus some more that are actually redundant.
(U) Under fit – Models with only a strict subset of the variables in the true model.
(W) Wrong fit – All models that are not overfit (O), not a correct fit (C) nor underfit (U). These are the models where some of the relevant variables might be present (though not all of them) in addition to some of the redundant variables.

For case 1 we add outliers on the response variable that are generated from a $N(100, 0.5^2)$ distribution, using three situations of 10%, 20% and 30% of outliers. We fit a collection of models to these data, where, for case 1, the covariate x_1 is always included in the model, while we choose amongst the other covariates $x_2, \ldots, x_5$. This results in $2^5 - 1$ models. The simulation results are shown in Table 4.1 and as expected, the AIC based on maximum likelihood estimators works better than the AIC based on S-estimators for the data without outliers. However, the ML-based AIC selects a large proportion of underfit or wrong fit models for the data with outliers. A higher proportion of overfit and correct fit models are selected by the AIC based on S1-estimators. Because the situation of this example requires selection amongst the parametric components of the models, and the nonparametric (random) part of the models is included in all of the models, we observe a comparable behavior of the marginal and conditional AIC for S1-estimators. The AICs based on S-estimators are preferred to the ML-versions for the cases with a high contamination level of outliers, the methods break down with 50% of outliers in the data.

Table 4.1: *Proportion of selected models from mAIC, cAIC, ccAIC, mAIC.S1, cAIC.S1 and ccAIC.S1 for data generated with dependent xs, mean structure m_1 for $p = 6$, error terms from a $N(0, 1)$ distribution, and for sample size $n = 100$. We consider different % ε of outliers generated from $N(100, 0.5^2)$. S1- estimators are computed with 50% breakdown point.*

%		mAIC	cAIC	ccAIC	mAIC.S1	cAIC.S1	ccAIC.S1
0	C	0.535	0.509	0.509	0.450	0.476	0.487
	O	0.465	0.491	0.491	0.300	0.398	0.396
	U	0.000	0.000	0.000	0.083	0.076	0.074
	W	0.000	0.000	0.000	0.167	0.050	0.043
10	C	0.019	0.020	0.021	0.374	0.372	0.380
	O	0.013	0.011	0.010	0.451	0.460	0.472
	U	0.544	0.604	0.612	0.065	0.023	0.030
	W	0.424	0.365	0.357	0.110	0.145	0.118
20	C	0.020	0.017	0.019	0.327	0.364	0.371
	O	0.018	0.014	0.023	0.429	0.432	0.443
	U	0.582	0.632	0.602	0.073	0.077	0.083
	W	0.380	0.337	0.356	0.171	0.127	0.103
30	C	0.014	0.013	0.015	0.283	0.274	0.291
	O	0.018	0.014	0.017	0.491	0.492	0.480
	U	0.564	0.663	0.670	0.084	0.092	0.097
	W	0.404	0.310	0.298	0.142	0.142	0.132

For case 3 we conduct selection amongst the parametric and nonparametric (random) components of the model. This results in fitting $2^6 - 1$ different models to the data. Again, outliers on the response variable are generated from a $N(100, 0.5^2)$ distribution in different percentages (10%, 20% and 30%). Based on the results from Table 4.2 we clearly observe that the performance of the two marginal AICs (mAIC and mAIC.S1) is inferior to that of the conditional AICs, which is to be expected since in this setting we select both the parametric and the nonparametric components in the model.

Table 4.2: *Proportion of selected models from mAIC, cAIC, ccAIC, mAIC.S1, cAIC.S1 and ccAIC.S1 for data generated with dependent xs, mean structure m_3 for $p = 6$, error terms from a $N(0,1)$ distribution, and for sample size $n = 100$. We consider different % of outliers on ε, generated from $N(100, 0.5^2)$. S1- estimators are computed with 50% breakdown point.*

%		mAIC	cAIC	ccAIC	mAIC.S1	cAIC.S1	ccAIC.S1
0	C	0.383	0.494	0.442	0.270	0.432	0.499
	O	0.307	0.471	0.483	0.210	0.361	0.371
	U	0.231	0.000	0.000	0.364	0.059	0.062
	W	0.079	0.035	0.075	0.156	0.148	0.068
10	C	0.009	0.010	0.011	0.257	0.422	0.474
	O	0.003	0.006	0.008	0.200	0.343	0.352
	U	0.654	0.638	0.685	0.346	0.056	0.059
	W	0.334	0.346	0.296	0.198	0.179	0.115
20	C	0.006	0.008	0.010	0.236	0.409	0.432
	O	0.002	0.004	0.016	0.216	0.337	0.428
	U	0.672	0.683	0.698	0.318	0.052	0.107
	W	0.320	0.305	0.276	0.230	0.202	0.033
30	C	0.002	0.006	0.005	0.283	0.399	0.422
	O	0.001	0.004	0.007	0.491	0.376	0.395
	U	0.710	0.693	0.706	0.084	0.079	0.106
	W	0.287	0.297	0.282	0.142	0.146	0.077

Table 4.2 shows that the conditional S1-methods have a good performance also in the case that no outliers are present in the data, and these methods are preferred in the case of outliers. Higher proportions of correct and overfit models are obtained when the corrected versions of the conditional AIC are used.

4.5 Discussion

The need for robust model selection methods in linear mixed models has lead us to develop the generalized degrees of freedom for S-estimation methods. In multilevel models, extreme or outlying observations might occur at any level. The proposed estimation method and the subsequent generalized degrees of freedom that we have used in a conditional AIC, could presumably be developed along similar lines for other models, such as generalized linear mixed models.

4.6 Appendix. Computation of S-estimators for linear mixed models

4.6.1 Proof of Result 4.1

Setting the partial derivatives of L_{joint} in (4.4) with respect to β, u and the vector σ^2 to zero, and solving for these values, yields estimators $\widehat{\beta}, \widehat{u}, \widehat{\sigma}^2$. We arrive at

$$\begin{aligned}
\widehat{\beta} &= (X^t\widehat{W}\widehat{R}^{-1}X)^{-1}X^t\widehat{W}\widehat{R}^{-1}(\mathbf{y} - Z\widehat{\mathbf{u}}) && (4.28)\\
\widehat{u} &= (Z^t\widehat{W}\widehat{R}^{-1}Z + \frac{2n}{\widehat{\tau}_1}\widehat{G}^{-1})^{-1}Z^t\widehat{W}\widehat{R}^{-1}(Y - X\widehat{\boldsymbol{\beta}}). && (4.29)
\end{aligned}$$

Substituting (4.29) in equation (4.28) yields (4.5), while substituting (4.5) in (4.29) yields (4.6).
We let $\widehat{V}^{-1} = (\widehat{R}^{-1} - \widehat{R}^{-1}Z(Z^t\widehat{W}\widehat{R}^{-1}Z + \frac{2n}{\widehat{\tau}}\widehat{G}^{-1})^{-1}Z^t\widehat{W}\widehat{R}^{-1})$ from which it follows that $\widehat{V} = \widehat{R} + Z(\frac{\widehat{\tau}}{2n}\widehat{G})Z^t\widehat{W}$.

Equating the partial derivative of L_{joint} with respect to σ_0^2 to zero yields first, by solving for τ_1, that

$$m = \frac{\widehat{\tau}_1}{2n}\sum_{i=1}^{n} W(\widehat{d}_i)(y_i - X_i\widehat{\beta} - Z_i\widehat{u})^t\widehat{R}_i^{-1}(y_i - X_i\widehat{\beta} - Z_i\widehat{u}),$$

from which follows that $\widehat{\tau}_1 = 2mn(\widehat{d}^t\widehat{W}\widehat{d})^{-1}$. Second, solving for σ_0^2 yields that

$$\widehat{\sigma}_0^2 = \frac{\widehat{\tau}_1}{2mn}(Y - X\widehat{\beta} - Z\widehat{u})^t\widehat{W}(Y - X\widehat{\beta} - Z\widehat{u}),$$

from which (4.7) follows. The partial derivatives of L_{joint} with respect to σ_j^2 $(j = 1, \ldots, q)$, which occur only in the matrix G give that $\widehat{\sigma}_j^2 = \widehat{u}_j^t \widehat{u}_j / q_j$.

4.6.2 Proof of Result 4.2

The estimators for β, σ_0^2 and τ_1 are obtained similarly as in Result 4.1 though now starting from the joint Lagrangian (4.10). The expressions for the predictors $\widetilde{u}$ and for the variance component estimators are different. After substituting the estimator $\widetilde{\beta}$ in the next equation,

$$\widetilde{u} = \left(Z^t \widetilde{W} \widetilde{R}^{-1} Z + \frac{n\widetilde{\tau}_2}{q\widetilde{\tau}_1} \widetilde{G}^{-1/2} \widetilde{W}_2 \widetilde{G}^{-1/2} \right)^{-1} Z^t \widetilde{W} \widetilde{R}^{-1} (Y - X\widetilde{\beta}),$$

the estimator $\widetilde{u}$ of (4.12) results. From

$$\frac{\partial \mathrm{L}_{\text{joint2}}(\widetilde{\beta}, \widetilde{u}, \sigma^2)}{\partial \sigma_j^2} |_{\sigma_j^2 = \widetilde{\sigma}_j^2} = \frac{\partial}{\partial \sigma_j^2} \{ \log \mid G \mid + \frac{\widetilde{\tau}_2}{r} \sum_{k=1}^{r} \rho_2(\widetilde{d}_{2k}) \} = 0, \qquad (4.30)$$

for all $j = 1, \ldots, r$, (4.14) follows. Since (4.30) implies that also the sum over $j = 1, \ldots, r$ of these partial derivatives is equal to zero, $\widetilde{\tau}_2 = 2rq(\sum_{j=1}^{r} \widetilde{u}_j^t \widetilde{W}_{2j} \widetilde{G}_j^{-1} \widetilde{u}_j)^{-1} = 2rq(\widetilde{d}_2^t \widetilde{W}_2 \widetilde{d}_2)^{-1}$.

4.7 Appendix. Generalized degrees of freedom for the S-estimators

4.7.1 Proof of Theorem 4.1

We start from model (4.1) and assume that all variance components are unknown. The generalized degrees of freedom is defined by $\Phi_{S1} = \operatorname{Tr}\left(\partial\widehat{Y}/(\partial Y)\right)$. From (4.5) and (4.6) it follows that

$$\widehat{Y} \quad = \quad X\widehat{\beta} + Z\widehat{u} = X\widehat{\beta} + Z\left(\frac{\widehat{\tau}_1}{2n} \widehat{G} Z^t \widehat{W} \widehat{V}^{-1} (Y - X\widehat{\beta}) \right).$$

The expression of V from Result 4.1, see (4.9), leads to rewriting
$Z(\frac{\widehat{\tau}_1}{2n}\widehat{G})Z^t\widehat{W} = \widehat{V} - \widehat{R}$, from which it follows that

$\widehat{Y} = X\widehat{\beta} + (I_N - \widehat{R}\widehat{V}^{-1})(Y - X\widehat{\beta}) = Y - \widehat{R}\widehat{V}^{-1}\widehat{P}Y$
where $\widehat{P} = I_N - X(X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\widehat{W}\widehat{V}^{-1}$. Thus

$$\phi_{S1} = \mathrm{Tr}\left(I_N - \widehat{R}\widehat{V}^{-1}\widehat{P} - \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial Y}Y\right). \tag{4.31}$$

With $\widehat{R} = \widehat{\sigma}_0^2 I_N$ and $\widehat{G}_j = \widehat{\sigma}_j^2 I_{q_j}$, $j = 1, \ldots, r$, Y is a vector of length N, Y_k is the kth element of the vector Y. We define the $N \times N$ matrix B,

$$B = \left(\frac{\partial\widehat{R}\widehat{V}^{-1}\widehat{P}Y}{\partial Y_1}, \frac{\partial\widehat{R}\widehat{V}^{-1}\widehat{P}Y}{\partial Y_2}, \ldots, \frac{\partial\widehat{R}\widehat{V}^{-1}\widehat{P}Y}{\partial Y_N}\right) \tag{4.32}$$

here $B_k = \partial\widehat{R}\widehat{V}^{-1}\widehat{P}Y/\partial Y_k; k = 1, 2, \ldots, N$ and B_k is the kth column of the matrix B. B_k is a $N \times 1$ column matrix. We can re-write B_k using the chain rule as follows,

$$B_k \quad = \quad \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial\sigma_0^2}Y\frac{\partial\widehat{\sigma}_0^2}{\partial Y_k} + \sum_{j=1}^{r}\frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial\sigma_j^2}Y\frac{\partial\widehat{\sigma}_j^2}{\partial Y_k}. \tag{4.33}$$

A further application of the chain rule leads to

$$\frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial\sigma_0^2} = \widehat{V}^{-1}\widehat{P} - \widehat{R}\widehat{V}^{-1}\Big\{\frac{\partial\widehat{V}}{\partial\sigma_0^2} - X(X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t$$

$$(\widehat{W}\widehat{V}^{-1}\frac{\partial\widehat{V}}{\partial\sigma_0^2} - \frac{\partial\widehat{W}}{\partial\sigma_0^2})\Big\}\widehat{V}^{-1}\widehat{P}. \tag{4.34}$$

Starting from (4.9),

$$\frac{\partial\widehat{V}}{\partial\sigma_0^2} = I_N + Z(\frac{1}{2n}\widehat{G})Z^t\widehat{W}\frac{\partial\widehat{\tau}_1}{\partial\sigma_0^2} + Z(\frac{\widehat{\tau}_1}{2n}\widehat{G})Z^t\frac{\partial\widehat{W}}{\partial\sigma_0^2}, \tag{4.35}$$

where

$$\begin{aligned}
\frac{\partial\widehat{\tau}_1}{\partial\sigma_0^2} \quad &= \quad -2mn(\widehat{d}^t\widehat{W}\widehat{d})^{-1}\big(2\widehat{d}^t\widehat{W}\frac{\partial\widehat{d}}{\partial\sigma_0^2} + \widehat{d}^t\frac{\partial\widehat{W}}{\partial\sigma_0^2}\widehat{d}\big)(\widehat{d}^t\widehat{W}\widehat{d})^{-1} \\
\frac{\partial\widehat{W}}{\partial\sigma_0^2} \quad &= \quad \mathrm{diag}_{i=1,\ldots,n}\left[\left(\frac{\widehat{d}_i\psi'(\widehat{d}_i) - \psi(\widehat{d}_i)}{\widehat{d}_i^2}\right)\frac{\partial\widehat{d}_i}{\partial\sigma_0^2}I_m\right] \\
\frac{\partial\widehat{d}_i}{\partial\sigma_0^2} \quad &= \quad \frac{1}{2\widehat{d}_i}(Y_i - X_i\widehat{\beta} - Z_i\widehat{u})^t\widehat{R}_i^{-1}\widehat{R}_i^{-1}(Y_i - X_i\widehat{\beta} - Z_i\widehat{u}).
\end{aligned}$$

Since from (4.9) it follows that $\frac{\partial \widehat{V}}{\partial \sigma_j^2} = \frac{\widehat{\tau}_1}{2n} Z_j Z_j^t \widehat{W}$ and since $\frac{\partial \widehat{P}}{\partial \sigma_j^2} = \frac{\widehat{\tau}_1}{2n}(I_N - \widehat{P}) Z_j Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P}$,

$$\frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial \sigma_j^2} = -\frac{\widehat{\tau}_1}{2n} \widehat{R}\widehat{V}^{-1}\widehat{P} Z_j Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P}. \qquad (4.36)$$

Define for $j = 0, \ldots, r$

$$\frac{\partial \mathrm{L}_{\text{joint}}(\widehat{\beta}, \widehat{u}, \sigma^2)}{\partial \sigma_j^2}\Big|_{\sigma_j^2 = \widehat{\sigma}_j^2} = h(\widehat{\sigma}_j^2(Y), Y) = 0. \qquad (4.37)$$

Using the estimators from Result 4.1,

$$\begin{aligned} h(\widehat{\sigma}_0^2(Y), Y) &= \frac{m}{\widehat{\sigma}_0^2} - \frac{\widehat{\tau}_1}{n}(Y - X\widehat{\beta} - Z\widehat{u})^t \widehat{R}^{-1} \widehat{W} \widehat{R}^{-1} (Y - X\widehat{\beta} - Z\widehat{u}) \\ &= \frac{m}{\widehat{\sigma}_0^2} - \frac{\widehat{\tau}_1}{n} Y^t \widehat{P}^t \widehat{V}^{-1} \widehat{W} \widehat{V}^{-1} \widehat{P} Y. \qquad (4.38) \end{aligned}$$

In this expression $\widehat{\tau}_1$, $\widehat{P}$ and $\widehat{W}$ are a function of Y and $\widehat{\sigma}_0^2$. Take the full differentiation of $h(\widehat{\sigma}_0^2(Y), Y)$ with respect to Y_k,

$$\frac{dh(\widehat{\sigma}_0^2(Y), Y)}{dY_k} = \frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial \sigma_0^2} \frac{d\widehat{\sigma}_0^2}{dY_k} + \frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial Y_k} = 0,$$

to find that

$$\frac{d\widehat{\sigma}_0^2}{dY_k} = -\left[\frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial \sigma_0^2}\right]^{-1} \frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial Y_k}. \qquad (4.39)$$

From Result 4.1,

$$\begin{aligned} \frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial Y_k} &= -\frac{1}{n} Y_k^t \widehat{P}_k^t \widehat{V}^{-1} \Big[2\widehat{\tau}_1 \widehat{W} \widehat{V}^{-1} \widehat{P}_k + \frac{\partial \widehat{\tau}_1}{\partial Y_k} \widehat{W} \widehat{V}^{-1} \widehat{P}_k Y_k \\ &\quad + 2\widehat{\tau}_1 \widehat{W} \frac{\partial(\widehat{V}^{-1}\widehat{P}_k)}{\partial Y_k} Y_k + \widehat{\tau}_1 \frac{\partial \widehat{W}}{\partial Y_k} \widehat{V}^{-1} \widehat{P}_k Y_k \Big], \end{aligned}$$

with

$$\begin{aligned} \frac{\partial \widehat{\tau}_1}{\partial Y_k} &= -2mn(\widehat{d}^t \widehat{W} \widehat{d})^{-1} \left[2\widehat{d}^t \widehat{W} \frac{\partial \widehat{d}}{\partial Y_k} + \widehat{d}^t \frac{\partial \widehat{W}}{\partial Y_k} \widehat{d} \right] (\widehat{d}^t \widehat{W} \widehat{d})^{-1} \\ \frac{\partial \widehat{d}_i}{\partial Y_k} &= \frac{(Y_i - X_i \widehat{\beta} - Z_i \widehat{u})^t \widehat{R}_i^{-1}}{\widehat{d}_i} \\ \frac{\partial \widehat{W}}{\partial Y_k} &= \operatorname{diag}_{i=1,\ldots,n} \left[\left(\frac{\widehat{d}_i \psi'(\widehat{d}_i) - \psi(\widehat{d}_i)}{\widehat{d}_i^2} \right) \frac{\partial \widehat{d}_i}{\partial Y_k} I_m \right]. \end{aligned}$$

Here P_k is the kth column of the matrix P. Further it follows from (4.9) and from matrix differentiation rules that

$$\frac{\partial(\widehat{V}^{-1}\widehat{P}_k)}{\partial Y_k} = -\widehat{V}^{-1}\frac{\partial \widehat{V}}{\partial Y_k}\widehat{V}^{-1}\widehat{P}_k + \widehat{V}^{-1}\frac{\partial \widehat{P}_k}{\partial Y_k},$$

where

$$\begin{aligned}
\frac{\partial \widehat{V}}{\partial Y_k} &= \frac{1}{2n}Z\frac{\partial \widehat{\tau}_1}{\partial Y_k}\widehat{G}Z^t\widehat{W} + Z\frac{\widehat{\tau}_1}{2n}\widehat{G}Z^t\frac{\partial \widehat{W}}{\partial Y_k} \\
\frac{\partial \widehat{P}_k}{\partial Y_k} &= X(X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\left(\widehat{W}\widehat{V}^{-1}\frac{\partial \widehat{V}}{\partial Y_k} - \frac{\partial \widehat{W}}{\partial Y_k}\right)\widehat{V}^{-1}\widehat{P}_k.
\end{aligned}$$

With the calculations done so far, we immediately obtain that

$$\begin{aligned}
\frac{\partial h(\widehat{\sigma}_0^2(Y),Y)}{\partial \sigma_0^2} &= -\frac{n}{\widehat{\sigma}_0^4} - \frac{1}{n}Y^t\widehat{P}^t\widehat{V}^{-1}\left[\widehat{W}\widehat{V}^{-1}\widehat{P}Y\frac{\partial \widehat{\tau}_1}{\partial \sigma_0^2}\right. \\
&\quad \left. +2\widehat{\tau}_1\widehat{W}\frac{\partial(\widehat{V}^{-1}\widehat{P})}{\partial \sigma_0^2}Y + \widehat{\tau}_1\frac{\partial \widehat{W}}{\partial \sigma_0^2}\widehat{V}^{-1}\widehat{P}Y\right].
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial(\widehat{V}^{-1}\widehat{P})}{\partial \sigma_0^2} &= -\widehat{V}^{-1}\Big\{\frac{\partial \widehat{V}}{\partial \sigma_0^2} - X(X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\big(\widehat{W}\widehat{V}^{-1}\frac{\partial \widehat{V}}{\partial \sigma_0^2} - \frac{\partial \widehat{W}}{\partial \sigma_0^2}\big)\Big\} \\
&\quad \widehat{V}^{-1}\widehat{P}.
\end{aligned}$$

We consider next the functions $h(\widehat{\sigma}_j^2(Y),Y)$ for $j = 1,\ldots,r$. Using the expressions from Result 4.1,

$$\begin{aligned}
h(\widehat{\sigma}_j^2(Y),Y) &= \frac{q_j}{\widehat{\sigma}_j^2} - \widehat{u}_j^t\widehat{G}_j^{-1}\widehat{G}_j^{-1}\widehat{u}_j \\
&= \frac{q_j}{\widehat{\sigma}_j^2} - \frac{\widehat{\tau}_1}{2n}Y^t\widehat{P}^t\widehat{V}^{-1}\widehat{W}Z_j\frac{\widehat{\tau}_1}{2n}Z_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}Y \\
&= \frac{q_j}{\widehat{\sigma}_j^2} - \widehat{A}_j^t\widehat{A}_j,
\end{aligned}$$

where $\widehat{A}_j = \frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P} Y$. By the full differentiation of h, this further leads to

$$\frac{d\widehat{\sigma}_j^2}{dY_k} = - \left[\frac{\partial h(\widehat{\sigma}_j^2(Y), Y)}{\partial \sigma_j^2} \right]^{-1} \frac{\partial h(\widehat{\sigma}_j^2(Y), Y)}{\partial Y_k}, \tag{4.40}$$

where via similar calculations we arrive at

$$\begin{aligned} \frac{\partial h(\widehat{\sigma}_j^2(Y), Y)}{\partial Y_k} &= -2\widehat{A}_j^t \frac{\partial \widehat{A}_j}{\partial Y_k} \\ \frac{\partial h(\widehat{\sigma}_j^2(Y), Y)}{\partial \sigma_j^2} &= -\frac{q_j}{\widehat{\sigma}_j^4} - 2\widehat{A}_j^t \frac{\partial \widehat{A}_j}{\partial \sigma_j^2} \\ &= -\frac{q_j}{\widehat{\sigma}_j^4} - 2\widehat{A}_j^t \left(\frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P} Z_j \right) \widehat{A}_j, \end{aligned}$$

where it holds that

$$\begin{aligned} \frac{\partial \widehat{A}_j}{\partial Y_k} &= \frac{\widehat{\tau}_1}{2n} Z_j^t \{ \widehat{W} \widehat{V}^{-1} \widehat{P}_k + \frac{\partial \widehat{W}}{\partial Y_k} \widehat{V}^{-1} \widehat{P}_k Y_k + \widehat{W} \frac{\partial (\widehat{V}^{-1} \widehat{P}_k)}{\partial Y_k} Y_k \} \\ &\quad + \frac{1}{2n} \frac{\partial \widehat{\tau}_1}{\partial Y_k} Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P}_k Y_k \\ \frac{\partial \widehat{A}_j}{\partial \sigma_j^2} &= \frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \frac{\partial (\widehat{V}^{-1} \widehat{P})}{\partial \sigma_j^2} Y = \frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \frac{\widehat{\tau}_1}{2n} \widehat{V}^{-1} \widehat{P} Z_j Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P} Y. \end{aligned}$$

This proves Theorem 4.1.

4.7.2 Proof of Theorem 4.2

The proof goes along the same lines as that of Theorem 4.1, with this difference that we use the estimators from Result 4.2, and in particular the expressions for (4.31) and (4.33) with these estimators, in addition to (4.37) using now L_{joint2}, leads to considering

$$\begin{aligned}
\frac{\partial h(\widetilde{\sigma}_0^2(Y),Y)}{\partial Y_k} &= -\frac{1}{n}Y_k^t\widetilde{P}_k^t\widetilde{V}^{-1}\left[\widetilde{W}\widetilde{V}^{-1}\widetilde{P}_k\big(2\widetilde{\tau}_1+\frac{\partial\widetilde{\tau}_1}{\partial Y_k}Y_k\big)\right.\\
&\quad\left.+2\widetilde{\tau}_1\widetilde{W}\frac{\partial(\widetilde{V}^{-1}\widetilde{P}_k)}{\partial Y_k}Y_k+\widetilde{\tau}_1\frac{\partial\widetilde{W}}{\partial Y_k}\widetilde{V}^{-1}\widetilde{P}_kY_k\right]\\
\frac{\partial h(\widetilde{\sigma}_0^2(Y),Y)}{\partial\sigma_0^2} &= -\frac{n}{\widetilde{\sigma}_0^4}-\frac{1}{n}Y^t\widetilde{P}^t\widetilde{V}^{-1}\left[\widetilde{W}\widetilde{V}^{-1}\widetilde{P}Y\frac{\partial\widetilde{\tau}_1}{\partial\sigma_0^2}\right.\\
&\quad\left.+2\widetilde{\tau}_1\widetilde{W}Y\frac{\partial(\widetilde{V}^{-1}\widetilde{P})}{\partial\sigma_0^2}+\widetilde{\tau}_1\frac{\partial\widetilde{W}}{\partial\sigma_0^2}\widetilde{V}^{-1}\widetilde{P}Y\right],
\end{aligned}$$

from which we in a similar way arrive at the estimator

$$\frac{d\widetilde{\sigma}_0^2}{dY_k}=-\left[\frac{\partial h(\widetilde{\sigma}_0^2(Y),Y)}{\partial\sigma_0^2}\right]^{-1}\frac{\partial h(\widetilde{\sigma}_0^2(Y),Y)}{\partial Y_k}. \tag{4.41}$$

The quantities $\widetilde{R}$, $\widetilde{\tau}_1$ and $\widetilde{W}$ do not depend on $\widetilde{\sigma}_j^2; j=1,\ldots,r$. From (4.15),

$$\begin{aligned}
\frac{\partial\widetilde{V}}{\partial\sigma_j^2} &= Z\frac{r\widetilde{\tau}_1}{n\widetilde{\tau}_2}(\widetilde{G}_j^{-1/2}\widetilde{W}_{2j}\widetilde{G}_j^{-1/2})^{-1}\left[\widetilde{G}_j^{-1/2}(\widetilde{W}_{2j}\widetilde{G}_j^{-1}+\frac{\partial\widetilde{W}_{2j}}{\partial\sigma_j^2})\widetilde{G}_j^{-1/2}\right.\\
&\quad\left.\times(\widetilde{G}_j^{-1/2}\widetilde{W}_{2j}\widetilde{G}_j^{-1/2})^{-1}-\frac{1}{\widetilde{\tau}_2}\frac{\partial\widetilde{\tau}_2}{\partial\sigma_j^2}\right]Z^t\widetilde{W}
\end{aligned}$$

With $\partial\widetilde{d}_{2j}/\partial\sigma_j^2=-\frac{1}{2}\widetilde{G}_j^{-1}\widetilde{d}_{2j}$, and δ_{jk} the Kronecker delta such that $\delta_{jk}=1$ if and only if $j=k$, and $\delta_{jk}=0$ otherwise,

$$\begin{aligned}
\frac{\partial\widetilde{W}_{2j}}{\partial\sigma_j^2} &= \operatorname{diag}_{k=1,\ldots,r}\left[\left(\delta_{jk}\frac{\widetilde{d}_{2k}\psi'(\widetilde{d}_{2k})-\psi(\widetilde{d}_{2k})}{\widetilde{d}_{2k}^2}\right)\frac{\partial\widetilde{d}_{2k}}{\partial\sigma_j^2}I_{q_k}\right]\\
\frac{\partial\widetilde{\tau}_2}{\partial\sigma_j^2} &= -2qr(\widetilde{d}_2^t\widetilde{W}_2\widetilde{d}_2)^{-1}\left[2\widetilde{d}_2^t\widetilde{W}_2\frac{\partial\widetilde{d}_2}{\partial\sigma_j^2}+\widetilde{d}_2^t\frac{\partial\widetilde{W}_2}{\partial\sigma_j^2}\widetilde{d}_2\right](\widetilde{d}_2^t\widetilde{W}_2\widetilde{d}_2)^{-1}
\end{aligned}$$

All this taken together gives us $\partial(\widetilde{R}\widetilde{V}^{-1}\widetilde{P})/(\partial\sigma_j^2)$. Defining

$$\begin{aligned}\frac{\partial \mathrm{L}_{\text{joint2}}(\widetilde{\beta},\widetilde{u},\sigma^2)}{\partial\sigma_j^2}|_{\sigma_j^2=\widetilde{\sigma}_j^2} &= 0\\ &= h_2(\widetilde{\sigma}_j^2(Y),Y) = q_j/\sigma_j^2 - \frac{\widetilde{\tau}_2}{r\widetilde{\sigma}_j^2}\widetilde{d}_{2j}^t\widetilde{W}_{2j}\widetilde{d}_{2j},\end{aligned}$$

it follows that

$$\begin{aligned}\frac{\partial h_2(\widetilde{\sigma}_j^2(Y),Y)}{\partial Y_k} &= -\frac{1}{2r\widetilde{\sigma}_j^2}\widetilde{d}_{2j}^t\left[\widetilde{W}_{2j}\widetilde{d}_{2j}\frac{\partial\widetilde{\tau}_2}{\partial Y_k} + 2\widetilde{\tau}_2\widetilde{W}_{2j}\frac{\partial\widetilde{d}_{2j}}{\partial Y_k}\right.\\ &\quad\left.+\widetilde{\tau}_2\frac{\partial\widetilde{W}_{2j}}{\partial Y_k}\widetilde{d}_{2j}\right] \qquad (4.42)\end{aligned}$$

$$\begin{aligned}\frac{\partial h_2(\widetilde{\sigma}_j^2(Y),Y)}{\partial\sigma_j^2} &= -\frac{q_j}{\widetilde{\sigma}_j^4} + \frac{1}{2r\widetilde{\sigma}_j^2}\widetilde{d}_{2j}^t\left[\frac{\widetilde{\tau}_2}{\widetilde{\sigma}_j^2}\widetilde{W}_{2j}\widetilde{d}_{2j} - \widetilde{W}_{2j}\widetilde{d}_{2j}\frac{\partial\widetilde{\tau}_2}{\partial\sigma_j^2}\right.\\ &\quad\left.-2\widetilde{\tau}_2\widetilde{W}_{2j}\frac{\partial\widetilde{d}_{2j}}{\partial\sigma_j^2} - \widetilde{\tau}_2\frac{\partial\widetilde{W}_{2j}}{\partial\sigma_j^2}\widetilde{d}_{2j}\right], \qquad (4.43)\end{aligned}$$

where

$$\begin{aligned}\frac{\partial\widetilde{\tau}_2}{\partial Y_k} &= -2qr(\widetilde{d}_{2j}^t\widetilde{W}_{2j}\widetilde{d}_{2j})^{-1}\left[2\widetilde{d}_{2j}^t\widetilde{W}_{2j}\frac{\partial\widetilde{d}_{2j}}{\partial Y_k} + \widetilde{d}_{2j}^t\frac{\partial\widetilde{W}_{2j}}{\partial Y_k}\widetilde{d}_{2j}\right]\\ &\quad\times(\widetilde{d}_{2j}^t\widetilde{W}_{2j}\widetilde{d}_{2j})^{-1}\\ \frac{\partial\widetilde{d}_{2j}}{\partial Y_k} &= \frac{r\widetilde{\tau}_1}{n\widetilde{\tau}_2}\widetilde{G}_j^{-1/2}\left(\widetilde{G}_j^{-1/2}\widetilde{W}_{2j}\widetilde{G}_j^{-1/2}\right)^{-1}Z_j^t\widetilde{W}\widetilde{V}^{-1}\widetilde{P}_kY_k\\ \frac{\partial\widetilde{W}_{2j}}{\partial Y_k} &= \operatorname{diag}_{j=1,\ldots,r}\left[\left(\frac{\widetilde{d}_{2j}\psi_2'(\widetilde{d}_{2j}) - \psi_2(\widetilde{d}_{2j})}{\widetilde{d}_{2j}^2}\right)\frac{\partial\widetilde{d}_{2j}}{\partial Y_k}\right]\end{aligned}$$

This leads to

$$\frac{d\widetilde{\sigma}_j^2}{dY_k} = -\left[\frac{\partial h_2(\widetilde{\sigma}_j^2(Y),Y)}{\partial\sigma_j^2}\right]^{-1}\frac{\partial h_2(\widetilde{\sigma}_j^2(Y),Y)}{\partial Y_k}, \qquad (4.44)$$

from which the stated results follows.

Chapter 5

Implementations in the software package R

This chapter discusses the functions implemented in the R software to fit and study the proposed methods in the dissertation. Once we propose a new statistical method, we want it to be publicly available since then the proposed methods can more easily be used in practice by practitioners in their corresponding fields of application.

We present the implemented functions for each of our developed methods as given in the previous chapters. In the first part of this chapter, we discuss the functions used in S-estimation for penalized regression splines. This is followed by robust versions of AIC functions for regression models and finally we give the functions used for the S-estimation method and for model selection based on S-estimators for linear mixed models.

5.1 R functions for: "S-Estimation for penalized regression splines"

We used as a loss function the Tukey bi-square function to compute the S-estimator in this dissertation and we define the ρ function, the first derivative of the ρ function (ψ function) and the first derivative of the ψ

function here.

```
# Define rho function
Rho=function(x, cc){
U=x/cc
U1=3 * U^2 - 3 * U^4 + U^6
U1[abs(U) > 1] = 1
return(U1)}
# Define psi function
Psi=function(x, cc){
U = x/cc
U1 = 6/cc * U * (1 - U^2)^2
U1[abs(U) > 1] = 0
return(U1)}
```

To decide on the convergence of the proposed estimates of Chapters 2 and 4, we used the norm function, which is given here,

```
# Define norm function
norm = function(a) sqrt(sum(a^2))
```

To compute the non-robust penalized least squares estimators we have used the following function in Chapter 2.

```
pen.ls = function(y, X, D, lambda){
beta.ls = as.vector(solve(t(X)%*%X+lambda*D)%*%t(X)%*%y)
Sbeta.ls = mad( y - X %*% beta.ls)
return(list(beta=beta.ls,Sbeta=Sbeta.ls))}
# Define generalized cross validation function for
# LS-penalized regression
gcv = function(y, X, D, lambda){
# y is the response vector
# X is the big design matrix
# D is the penalty matrix
# lambda is the value of the penalty constant to be evaluated
tmp = solve( t(X) %*% X + lambda * D ) %*% t(X)
beta = as.vector( tmp %*% y )
n = length(y)
r = as.vector(y - X %*% beta)
H = X %*% tmp
return( n * sum( r^2 ) / (n - sum(diag(H)))^2 )}
# GCV search for penalized LS-estimators
pen.ls.gcv = function(y, X, D, lambdas){
ll = length(lambdas)
# GCVs for the LS estimator
gcvs = rep(0, ll)
for(i in 1:ll){
gcvs[i] = gcv(y, X, D, lambdas[i])}
# find the best lambda
lam = max( lambdas[ gcvs == min(gcvs) ] )
beta.ls = as.vector(solve(t(X)%*%X+lam*D)%*%t(X)%*%y)
Sbeta.ls = mad( y - X %*% beta.ls)
# get the LS estimated mean
yhat.ls = as.vector(X%*%solve(t(X)%*%X+lam*D)%*%t(X)%*%y)
return(list(beta=beta.ls, Sbeta=Sbeta.ls, yhat = yhat.ls,
lam=lam, gcv=min(gcvs)))}
```

Define the function for penalized M-estimators for fixed lambda - (Proposed in Lee and Oh (2007))

```
pen.m= function(y,X,N,D,lambda,num.knots,p,epsilon=1e-6){
# store the values in matrix
results = matrix(ncol=n+1,nrow=N)
# start with penalized LS
tmp = pen.ls(y, X, D, lambda)
beta1 = as.vector( tmp$beta )
mhat1 = as.vector( X %*% tmp$beta )
sigma1 = tmp$Sbeta
results[1,] = c(mhat1, sigma1)
mhat = mhat1
mbetaresults = matrix(ncol=num.knots+2+p-1, nrow=N)
mbetaresults[1,] = c(beta1)
mbeta = beta1
for (j in 2:N){
res = as.vector(y-X%*%mbeta)
# sigma = 1.4826*median(abs(res))
sigma = mad(res)
cval = 1.345*sigma
psi1 = ifelse( abs(res)<=cval, 2*res, 2*cval*sign(res) )
z = mhat + (psi1/2)
mbeta = solve( t(X)%*%X + D*lambda ) %*% t(X) %*% z
mhat = as.vector(X%*%solve(t(X)%*%X+D*lambda)%*%t(X)%*%z)
results[j,] = c(mhat,sigma)
mbetaresults[j,] = c(mbeta)
ifelse(((norm(mbetaresults[j,]-mbetaresults[j-1,])/
norm(mbetaresults[j-1,]))<epsilon),break,next)}
return(list(outmbeta=as.vector(mbetaresults[j,]),
sigma=as.vector(results[j,n+1]), iterations=j))}
```

Define robust cross validation function for M-penalized regression. This function is proposed in Cantoni and Ronchetti (2001a).

```
mrcv = function(mm, y, X, D, lambda,n){
# mm has the fit returned by pen.m()
# y is the response vector
# X is the big design matrix
# D is the penalty matrix
# lambda is the value of the penalty constant to be evaluated
# n = length(y)
res = as.vector(y - X %*% mm$outmbeta )
sigma = mad(res)
cval = 1.345*sigma
psi1 = ifelse( abs(res)<=cval, 2*res, 2*cval*sign(res) )
psi1dash = ifelse( abs(res)<=cval, 2,0 )
Epsi1dash = sum(psi1dash)/n
II= diag(c(rep(1,ncol(X))))
SS =X %*% solve(II+ lambda * (sigma/Epsi1dash)* D)%*%t(X)
return( mrcv=1/n * (sigma^2/Epsi1dash^2)* sum(psi1^2/
(1- diag(SS))^2 ))}
```

CV search for penalized M-estimators

```
pen.m.cv=function(y,X,NN,D,lambdas,num.knots,p,epsilon){
ll = length(lambdas)
# MCVs for the M estimator
best.cv = +Inf
mrcvs = rep(0,ll)
for(i in 1:ll){
mm= pen.m(y,X,NN,D,lambdas[i],num.knots,p,epsilon=1e-6)
mrcvs[i] = mrcv(mm, y, X, D, lambdas[i],n)
if( mrcvs[i] <= best.cv ){
best.mm = mm
best.cv = mrcvs[i]}}
# find the best lambda
#rlam = best.gcv
rlam= lambdas[mrcvs==best.cv]
yhat.m = as.vector( X %*% best.mm$outmbeta )
return(list(yhat = yhat.m, lam=rlam, gcv=min(mrcvs),
sigma.m=best.mm$sigma, iterations=best.mm$iterations))}
```

Define the function for penalized M-estimators with smoothing parameter selection using genaralized cross validation. This function is proposed in Lee and Oh (2007) and we used this function to compare our proposed estimation method of Penalized S-estimators.

```
pen.m.gcv = function(y,X,N,D,lambdas,num.knots,p,epsilon){
# store the values in matrix
results = matrix(ncol=n+1,nrow=N)
```

```
ll = length(lambdas)
# start with penalized LS
tmp = pen.ls.gcv(y, X, D, lambdas)
beta1 = as.vector( tmp$beta )
mhat1 = as.vector( X %*% tmp$beta )
sigma1 = tmp$Sbeta
results[1,] = c(mhat1, sigma1)
mhat = mhat1
mbetaresults = matrix(ncol=num.knots+2+p-1, nrow=N)
mbetaresults[1,] = c(beta1)
mbeta = beta1
for (j in 2:N){
res = as.vector(y-X%*%mbeta)
# sigma = 1.4826*median(abs(res))
sigma = mad(res)
cval = 1.345*sigma
psi1 = ifelse(abs(res)<=cval,2*res,2*cval*sign(res))
z = mhat + (psi1/2)
# GCV
gcvs = rep(0, ll)
for(i in 1:ll){
    gcvs[i] = gcv(z, X, D, lambdas[i])    }
# find the best lambda
lambda = max( lambdas[ gcvs == min(gcvs) ] )
mbeta = solve( t(X)%*%X + D*lambda ) %*% t(X) %*% z
mhat = as.vector(X%*%solve(t(X)%*%X+D*lambda)%*%t(X)%*%z)
results[j,] = c(mhat,sigma)
mbetaresults[j,] = c(mbeta)
ifelse(((norm(mbetaresults[j,]-mbetaresults[j-1,])
/norm(mbetaresults[j-1,]))<epsilon),break,next)}
return(list(outmbeta=as.vector(mbetaresults[j,]),
yhat = mhat, lam = lambda, gcv = min(gcvs),
leeoutmatrix=as.vector(results[j,1:n]),
sigma=as.vector(results[j,n+1]),iterations=j))}
```

Define the scale function

```
s.scale = function(r,cc=1.54764,b=.5,max.it=1000,ep){
s1 = mad(r)
if(abs(s1)<1e-10) return(s1)
s0 = s1 + 1
it = 0
while( ( abs(s0-s1) > ep ) && (it < max.it) ) {
it = it + 1
s0 = s1
s1 = s0*mean(Rho(r/s0,cc=cc))/b}
return(s1) }
```

We define here the function for the proposed estimation method of penalized S-estimators as in (2.14) of Chapter 2.

```
pen.s = function(y,X,N,D,lambda,num.knots,p,beta1,
Sbeta1,cc=1.54764,b=.5,epsilon=1e-6){
# store the values in matrix
betahats = matrix(ncol=num.knots+2+p-1+1,nrow=N)
betahats[1,] = c(beta1,Sbeta1)
beta = beta1
for (i in 2:N){
# update Sbeta conditional on beta
r = as.vector(y-X%*%beta)
Sbeta = s.scale(r, cc=cc, b=b, N, ep=1e-4)
rs = r / Sbeta
Wbeta = Psi(rs, cc) / rs
taubeta = n*(Sbeta)^2 / sum( r^2 * Wbeta )
# update beta conditional on Sbeta from above
beta = solve( t(X * Wbeta) %*% X + (D*lambda/taubeta) )
       %*% t(X * Wbeta) %*% y
betahats[i,] = c(beta,Sbeta)
ifelse(((norm(betahats[i,]-betahats[i-1,])/
norm(betahats[i-1,]))<epsilon),break,next)}
return(list(outmatrix=betahats[1:i,1:(num.knots+2+p-1+1)],
estimates=betahats[i,1:(num.knots+2+p-1)],
scale=betahats[i,num.knots+2+p-1+1],
iterations=i,weights=Wbeta)) }
```

The next function is used to obtain the penalized S-estimates with different initial candidate values, this is done to come as close as possible to the global minimum of the criterion function, see Chapter 2.

```
initial.S= function(y, X, D,lambda, num.knots, p,
           NN, cc, b, NNN){
# To get best beta w.r.t objective function
uubeta = matrix(0, ncol=(NNN+2),nrow=num.knots+2+p-1)
# We need to use the pen.s.gcv function instead of pen.s()
uuiteration = rep(0,(NNN+2))
# We need to use the pen.s.gcv function instead of pen.s()
uuscale = rep(0,(NNN+2))
# We need to use the pen.s.gcv function instead of pen.s()
uuweights = matrix(0, ncol=(NNN+2),nrow=n)
objval = rep(0,(NNN+2))        # To get min of objval
# Initial candidates from Resampling
for (ii in 1:NNN){
indices = sample(n,num.knots+2+p-1+1)
Xs = X[indices,]
ys = y[indices]
init = pen.ls(ys, Xs, D, lambda)
uu1 = pen.s(y, X, 20, D, lambda, num.knots, p,
      init$beta, init$Sbeta, cc=cc, b=b)
uubeta[,ii]= as.vector(uu1$estimates)
uuscale[ii] = uu1$scale
uuweights[,ii] = as.vector(uu1$weights)
uuiteration[ii] = uu1$iterations
objval[ii] = ((n*(uu1$scale^2))+(lambda*
    as.numeric(t(uu1$estimates)%*%D%*%uu1$estimates)))}
# Initial candidates from M-estimator
initM = pen.m(y, X, N=NN, D, lambda, num.knots, p)
uuM = pen.s(y, X, 20, D, lambda, num.knots, p,
      initM$outmbeta, initM$sigma, cc=cc, b=b)
uubeta[,(NNN+1)]= as.vector(uuM$estimates)
uuscale[(NNN+1)] = uuM$scale
uuweights[,(NNN+1)] = as.vector(uuM$weights)
uuiteration[(NNN+1)] = uuM$iterations
objval[(NNN+1)] = ((n*(uuM$scale^2))+(lambda*
    as.numeric(t(uuM$estimates)%*%D%*%uuM$estimates)))
# Initial candidates from LS-estimator
initLS = pen.ls(y, X, D, lambda)
uuLS = pen.s(y, X, 20, D, lambda, num.knots, p,
       initLS$beta, initLS$Sbeta, cc=cc, b=b)
uubeta[,(NNN+2)]= as.vector(uuLS$estimates)
uuscale[(NNN+2)] = uuLS$scale
uuweights[,(NNN+2)] = as.vector(uuLS$weights)
uuiteration[(NNN+2)] = uuLS$iterations
objval[(NNN+2)] = ((n*(uuLS$scale^2))+(lambda*
    as.numeric(t(uuLS$estimates)%*%D%*%uuLS$estimates)))
# find the best estimators with respect to objective function
bestbeta = as.vector( uubeta[ ,objval == min(objval)])
weights = s.vector(uuweights[ ,objval == min(objval)])
scale = uuscale[objval == min(objval)]
iterations = uuiteration[objval == min(objval)]
return(list(estimates=bestbeta, scale=scale, weights=weights,
      iterations=iterations)) }
```

We define the function for robust generalized cross validation for S-penalized regression splines as given in (2.18). We have used this function to select the smoothing parameter for the penalized S-regression spline estimation method for all simulation studies and for the real data examples in Chapter 2.

```
rgcv = function(uu, y, X, D, lambda) {
# uu has the fit returned by pen.s()
# y is the response vector
# X is the big design matrix
# D is the penalty matrix
# lambda is the value of the penalty constant to be evaluated
n = length(y)
nw = sum( uu$weights > 0 )
r = as.vector(y - X %*% uu$estimates )
aa = n * uu$scale^2 / sum( r^2 * uu$weights )
H = (X * uu$weights) %*% solve( t(X * uu$weights) %*% X +
     lambda/aa * D )  %*% t(X * uu$weights)
return(rgcv=nw * sum( r^2 * uu$weights )
       /(nw - sum(diag(H)))^2)}
```

The smoothing parameter selection using a generalized cross validation (GCV) search for penalized S-estimators function is given here.

```
pen.s.gcv = function(y, X, D, lambdas, num.knots,
        p, NN, cc, b,NNN){
ll = length(lambdas)
# GCVs for the S estimator
best.gcv = +Inf
rgcvs = rep(0,ll)
for(i in 1:ll){
uu= initial.S(y, X, D, lambdas[i], num.knots,
    p, NN, cc, b, NNN)
rgcvs[i] = rgcv(uu, y, X, D, lambdas[i])
if( rgcvs[i] <= best.gcv )    {
best.uu = uu
best.gcv = rgcvs[i]    }}
# find the best lambda
rlam= lambdas[rgcvs==best.gcv]
yhat.s = as.vector( X %*% best.uu$estimates )
return(list(yhat = yhat.s, lam=rlam, gcv=min(rgcvs),
iter.s=best.uu$iterations))}
```

5.2 R functions for: "A comparison of robust versions of the AIC based on M, S and MM-estimators"

We define the function for classical AIC selection (or more precisely, for TIC selection) for normal regression models as, AIC $= 2n\log(\sqrt{SSE/n}) + 2\mathrm{Tr}(J^{-1}K) + n\log(2\pi) + n$. This differs from the AIC for normal models that is used in most statistical software packages in the fact that we do not just count the number of parameters but use instead $\mathrm{Tr}(J^{-1}K)$.

```
AIC.scale.L= function(y, X,n,beta.L,scale.L){
U=UU=UU3=UU4=matrix(ncol=1,nrow=n)
for(i in 1:n){
U[i,]=((y[i,]-X[i,] %*% beta.L)^2/scale.L^4)
UU3[i,]= (((y[i,]-X[i,] %*% beta.L)/scale.L)^3)
UU4[i,]= (((y[i,]-X[i,] %*% beta.L)/scale.L)^4)
UU[i,]= t(X[i,]) %*% (X[i,])}
SCJ = 2/scale.L^2
SCK=(2+(sum(UU4)/n)-3)/scale.L^2
SCK1=(sum(UU3)/n)/scale.L^2
SCK.c=SCK1 *as.matrix(c(colMeans(X)),ncol(X),1)
SC.c=matrix(0,ncol(X),1)
SC.r=matrix(0,1,ncol(X))
J.beta= (t(X) %*% X)/(n*scale.L^2)
J=rbind(cbind(J.beta,SC.c),cbind(SC.r,SCJ))
inv.J= solve(J)
K.beta=(t(X)%*% X)/(n*scale.L^2)
K=rbind(cbind(K.beta,SCK.c),cbind(t(SCK.c),SCK))
AIC.CL =2* n*log(sqrt(SSE/(n)))+ 2 *sum(diag(inv.J %*% K))
        + n * log(2*pi)+n
return(AIC.CL) }
```

We define the AIC function for M-estimation where J and K are the full matrices, considering β and σ^2 as parameters (AIC.M1). We used this function in Table 3.9 in Chapter 3.

```
AIC.scale.M= function(y, X,n,beta.M,scale.M,cval=1.345){
U=U1=U2=UU=UU1=UUU=matrix(ncol=1,nrow=n)
UU2=matrix(ncol=ncol(X),nrow=n)
for(i in 1:n){
U[i,]=dPsiM((y[i,]-X[i,] %*% beta.M)/scale.M ,cval)
U1[i,]=(dPsiM((y[i,]-X[i,] %*% beta.M)/scale.M ,cval))*
       (((y[i,]-X[i,] %*% beta.M)/scale.M)^2)
U2[i,]=(PsiM((y[i,]-X[i,] %*% beta.M)/scale.M ,cval))*
       ((y[i,]-X[i,] %*% beta.M)/scale.M)
UU[i,]= (PsiM((y[i,]-X[i,] %*% beta.M)/scale.M ,cval))^2
UU1[i,]= ((PsiM((y[i,]-X[i,] %*% beta.M)/scale.M ,cval))^2)*
         (((y[i,]-X[i,] %*% beta.M)/scale.M)^2)
UU2[i,]=X[i,]*(PsiM((y[i,]-X[i,] %*% beta.M)/scale.M,cval))^2*
         ((y[i,]-X[i,] %*% beta.M)/scale.M)
UUU[i,]=RhoM((y[i,]-X[i,] %*% beta.M)/scale.M ,cval)}
SC.c=matrix(0,ncol(X),1)
SC.r=matrix(0,1,ncol(X))
J.betaM=(t(X)%*%diag(as.vector(U))%*% X)*(1/(n*scale.M^2))
SCJM = (sum(U1)-2* sum(U2)-n) * (1/(n*scale.M^2))
JM=rbind(cbind(J.betaM,SC.c),cbind(SC.r,SCJM))
inv.JM= solve(JM)
K.betaM=(t(X)%*%diag(as.vector(UU))%*%X)*(1/(n*scale.M^2))
SCKM=(sum(UU1)-2* sum(U2)+n) * (1/(n*scale.M^2))
SCKM.c=as.matrix(c(colMeans(UU2)),ncol(X),1)/(n*scale.M^2)
KM=rbind(cbind(K.betaM,SCKM.c),cbind(t(SCKM.c),SCKM))
AIC.M = 2* n*log(scale.M)+ 2*(sum(diag(inv.JM %*%KM)))
return(AIC.M)}
```

We define the AIC function for M-estimation where J and K are matrices considering only the vector β as parameters (AIC.M in (3.14)). We used this function in Chapter 3 for all simulation studies and real data examples.

```
AIC.M= function(y, X, beta.m,scale.m, cval=1.345){
U=UU=UUU=matrix(ncol=1,nrow=n)
for(i in 1:n){
U[i,]=dPsiM((y[i,]-X[i,] %*% beta.m)/scale.m ,cval)
UU[i,]=(PsiM((y[i,]-X[i,] %*% beta.m)/scale.m ,cval))^2 }
J= (t(X) %*% diag(as.vector(U))%*% X*(1/(scale.m^2)))/n
inv.J= solve(J)
K= (t(X) %*% diag(as.vector(UU))%*% X*(1/(scale.m^2)))/n
AIC = 2*n*(log(scale.m)) + 2* sum(diag(inv.J %*%(K)))
return(AIC) }
```

We define the AIC function for S-estimation where J and K are full matrices, considering β and σ^2 (AIC.S1). We used this function in Table 3.9 in Chapter 3.

```
AIC.scale.S= function(y, X,n,beta.S,scale.S,cc,b) {
U=U1=U2=UU=UU1=matrix(ncol=1,nrow=n)
UU2=matrix(ncol=ncol(X),nrow=n)
for(i in 1:n){
U[i,]=dPsi((y[i,]-X[i,] %*% beta.S)/scale.S ,cc)
U1[i,]=(dPsi((y[i,]-X[i,] %*% beta.S)/scale.S ,cc))*
       (((y[i,]-X[i,] %*% beta.S)/scale.S)^2)
U2[i,]=(Psi((y[i,]-X[i,] %*% beta.S)/scale.S ,cc))*
       ((y[i,]-X[i,] %*% beta.S)/scale.S)
UU[i,]=(Psi((y[i,]-X[i,] %*% beta.S)/scale.S ,cc))^2
UU1[i,]=((Psi((y[i,]-X[i,] %*% beta.S)/scale.S ,cc))^2)*
        (((y[i,]-X[i,] %*% beta.S)/scale.S)^2)
UU2[i,]=X[i,]*(((Psi((y[i,]-X[i,]%*%beta.S)/scale.S,cc))^2)*
        ((y[i,]-X[i,] %*% beta.S)/scale.S)) }
J.betaS= (t(X)%*%diag(as.vector(U))%*%X)*(1/(n*scale.S^2))
SCJS = (sum(U1)-2* sum(U2)-n) * (1/(n*scale.S^2))
SC.c=matrix(0,ncol(X),1)
SC.r=matrix(0,1,ncol(X))
JS=rbind(cbind(J.betaS,SC.c),cbind(SC.r,SCJS))
inv.JS= solve(JS)
K.betaS= (t(X)%*%diag(as.vector(UU))%*%X)*(1/(n*scale.S^2))
SCKS=(sum(UU1)-2* sum(U2)+n) * (1/(n*scale.S^2))
SCKS1=as.matrix(c(colMeans(UU2)),ncol(X),1)*(1/(n*scale.S^2))
KS=rbind(cbind(K.betaS,SCKS1),cbind(t(SCKS1),SCKS))
AIC.S = 2* n*log(scale.S) + 2 *sum(diag(inv.JS %*%KS))
return(AIC.S) }
```

We define the AIC function for S-estimation where J and K are matrices considering only the vector β as parameters (AIC.S in (3.12)). We used this function in Chapter 3 for all simulation studies and real data examples. We used this function for MM-estimators (AIC.MM in (3.15)) by using instead of S-estimators the regression MM-estimators and scale MM-estimators.

```
AIC.S= function(y, X, beta.s,scale.s, cc=1.54764){
U=UU=matrix(ncol=1,nrow=n)
for(i in 1:n){
U[i,]=dPsi((y[i,]-X[i,] %*% beta.s)/scale.s ,cc)
UU[i,]=(Psi((y[i,]-X[i,] %*% beta.s)/scale.s ,cc))^2 }
J= (t(X)%*%diag(as.vector(U))%*% X*(1/(scale.s^2)))/n
inv.J= solve(J)
K= (t(X)%*%diag(as.vector(UU))%*% X*(1/(scale.s^2)))/n
AIC =2*n*(log(scale.s))+ 2* sum(diag(inv.J %*%(K)))
return(AIC) }
```

We define the AIC function for S-estimation with J and K the matrices that only consider β, for the uniform asymptotic results (AIC.US in (3.16)). We used this function in Chapter 3 for all simulation studies and real data examples. We can use this function for MM-estimators AIC.UMM in (3.18) too.

```
AIC.scale.S.unif= function(y, X, beta.s,scale.s, cc, b){
U=UU=UU2=UUU=UUUb=UUUb2=UUUU=BH1=DH1=matrix(ncol=1,nrow=n)
for(i in 1:n){
U[i,]=dPsi((y[i,]-X[i,]%*%beta.s)/scale.s,cc)
UU[i,]=Psi((y[i,]-X[i,]%*%beta.s)/scale.s,cc)
UU2[i,]=(Psi((y[i,]-X[i,]%*%beta.s)/scale.s,cc))^2
UUU[i,]=Rho(((y[i,]-X[i,]%*%beta.s)/scale.s),cc)
UUUb[i,]=Rho(((y[i,]-X[i,]%*%beta.s)/scale.s),cc)-b
UUUb2[i,]=(Rho(((y[i,]-X[i,]%*%beta.s)/scale.s),cc)-b)^2
UUUU[i,] = UU[i,]*UUUb[i,]
BH1[i,]=(UU[i,] * ((y[i,]-X[i,] %*% beta.s)/scale.s))
DH1[i,]=U[i,] * ((y[i,]-X[i,] %*% beta.s)/scale.s) }
Vsi= (t(X)%*%X)/(scale.s^2)
Jsi= (t(X)%*%diag(as.vector(U))%*%X*(1/(scale.s^2)))/n
inv.Jsi= solve(Jsi)
dh=sum(DH1)/n * as.matrix(X/scale.s)
bh=sum(BH1)/n
dbh=dh/bh
E1=(t(X)%*%diag(as.vector(UU2))%*%X*(1/(scale.s^2)))/n
E2=(t(dbh)%*%diag(as.vector(UUUb2))%*%dbh*(1/(scale.s^2)))/n
E3=((t(X)/scale.s)%*%diag(as.vector(UUUU))%*%dbh)/n
E4=(t(dbh)%*%diag(as.vector(UUUU))%*%(X/scale.s))/n
Ksi= (E1+E2-E3-E3)
AIC.CS = 2*n*(log(scale.s))+2*sum(diag(inv.Jsi %*%Ksi))
return(AIC.CS) }
```

AIC based on log(scale) and on the number of parameters in the model (AIC.M2 and AIC.S2). We used this function in Table 3.9 in Chapter 3.

```
AIC.scale= function(y, X,n,scale){
AIC= 2* n*log(scale) + 2 *(ncol(X)+1)
return(AIC) }
```

Generalized information criteria (GICS) based on S-estimator. We used this function in Table 3.9 in Chapter 3.

```
GIC.scale.S= function(y, X,n,beta.S,scale.S,cc,b){
U=U2=UU=UU3=UUU=matrix(ncol=1,nrow=n)
U1=U3=UU1=UU2=matrix(ncol=ncol(X),nrow=n)
for(i in 1:n){
U[i,]=dPsi((y[i,]-X[i,] %*% beta.S)/scale.S ,cc)
U1[i,]=X[i,]*(dPsi((y[i,]-X[i,]%*%beta.S)/scale.S,cc))*
       (((y[i,]-X[i,]%*%beta.S)/scale.S))
U2[i,]=(Psi((y[i,]-X[i,]%*%beta.S)/scale.S,cc))*
       ((y[i,]-X[i,]%*%beta.S)/scale.S)
U3[i,]=(Psi((y[i,]-X[i,]%*%beta.S)/scale.S,cc))%*% X[i,]
UU1[i,]=X[i,]*((Psi((y[i,]-X[i,]%*%beta.S)/scale.S,cc))^2)*
         (((y[i,]-X[i,]%*%beta.S)/scale.S)^2)
UU2[i,]=X[i,]*(((Psi((y[i,]-X[i,]%*%beta.S)/scale.S,cc)))*
         (((y[i,]-X[i,]%*%beta.S)/scale.S)^2-1))
UU3[i,]= (Psi((y[i,]-X[i,]%*%beta.S)/scale.S,cc))*
         ((y[i,]-X[i,]%*%beta.S)/scale.S)*
         (((y[i,]-X[i,]%*%beta.S)/scale.S)^2-1)
UUU[i,]=((y[i,]-X[i,]%*%beta.S)/scale.S)^2}
b11=(t(X)%*%diag(as.vector(U))%*%X)*(1/(n*scale.S^2))
b12=as.matrix(c(colMeans(U1)),ncol(X),1)*(1/(n*scale.S^2))
b21=-t(as.matrix(c(colMeans(U3)),ncol(X),1))*(1/(n*scale.S^2))
b22=-sum(U2)/(n*scale.S^2)
B=rbind(cbind(b11,b12),cbind(b21,b22)); inv.B=solve(B)
a11=(t(X)%*%diag(as.vector(U2))%*% X)*(1/(n*scale.S^2))
a12=as.matrix(c(colMeans(UU2)),ncol(X),1)*(1/(n*scale.S^2))
a21=t(as.matrix(c(colMeans(UU1)),ncol(X),1)*(1/(n*scale.S^2)))
a22=sum(UU3)/(n*scale.S^2)
A=rbind(cbind(a11,a12),cbind(a21,a22))
GIC.S = 2* n*log(scale.S) + 2 *sum(diag(inv.B %*%A))
return(GIC.S) }
```

We used the Huber loss function to compute the M-estimator in this dissertation and define the ρ function, the first derivative of the ρ function (ψ function) and the first derivative of the ψ function here.

```
# Define Rho Huber function
RhoM= function(x, cval){
Rho1 = ifelse(abs(x)<=cval,(x^2),(2*cval*abs(x)-cval^2))
return(Rho1) }

# Define Psi Huber function
PsiM= function(x, cval) {
Psi1 = ifelse(abs(x)<=cval,2*x,2*cval*sign(x))
return(Psi1) }

# Define dPsi Huber function
dPsiM= function(x, cval){
dPsi1 = ifelse(abs(x)<=cval,2,0)
return(dPsi1)}
```

All subsets search, make an indicator matrix containing all possible combinations of variables. We used this function in all simulation studies and real data example in Chapter 3.

```
combinations = function(n){
comb = NULL
if (n<25) {
for( i in 1:n) comb = rbind(cbind(1,comb),cbind(0,comb))
return(comb) }
else {error("this value will probably block your computer,
 try on your own risk")} }
```

5.3 R-functions for: "Robust model selection for additive penalized regression splines models"

A translated Tukey biweight ρ function and the corresponding weight function of the translated Tukey biweight ρ function (this is taken from Copt and Victoria-Feser, 2006).

```
biwt.rho<- function(x,c)
{
  hulp = x^2/2-x^4/(2*c^2)+x^6/(6*c^4);
  rho = hulp*(abs(x)<c)+c^2/6*(abs(x)>=c)
  rho
}

biwt.wt <- function(e,k){
    ifelse (abs(e) <= k,(1 - (e/k)^2)^2,0)
}

biwt.psi <- function(e,k){
    ifelse (abs(e) <= k, e*(1 - (e/k)^2)^2,0)
}
```

We defined the S-estimators for additive penalized spline smoothing for the case with outliers only on individual level in Result 4.1. We used this function for all simulation studies in Chapter 4. S-estimators for additive penalized spline smoothing for the case with outliers only on the individual level

```
AdditivePenS1 = function(y,X,Z,NN,N,n,m,p,q,beta1,u1,
                R1,G1,V1, c1, epsilon){
#store the values in matrix
betahats = matrix(ncol=(p+(n*q)),nrow=NN)
betahats[1,] = c(beta1,u1)
beta=beta1; u=u1; R=R1; G=G1; V=V1
for (j in 2:NN){
# Compute mahalanobis distance and weight function
d2=rep(0,N); d1=rep(0,N)
for(i in 1:N){
mu = X[i,]%*% beta + Z[i,]%*% u
d2[i] = mahalanobis(Y[i],mu,R[i,i])}
d1 = sqrt(d2)
h = floor((N+(p+(n*q))+1)/2)
quantile = h/(N+1)
d=(d1*sqrt(qchisq(quantile,(p+(n*q)))))/(sort(d1)[h])
# Compute the weights
wld =rep(1,N)
wld = as.vector(biwt.wt(d1,c1))
W=diag(wld)
tau1=2*n*m*(1/as.vector(t(d1)%*%W%*%d1))
# Compute the fixed effect and random effect parameters
beta=as.vector(ginv(t(X)%*%W%*%ginv(V)%*% X)%*%t(X)
     %*%W%*%ginv(V)%*%Y)
u=as.vector((tau1/(2*n))*G%*%t(Z)%*%W%*%ginv(V)
  %*%(Y-X%*%beta))
# Compute the variance components
G=(as.vector(t(u)%*% u)/(n*q))*diag(rep(1,(n*q)))
R=(as.vector((1/as.vector(t(d1)%*%W%*%d1))%*%
  (t(Y-X%*%beta-Z%*%u)%*%W%*%(Y-X%*%beta-Z%*%u))))
  *diag(rep(1,N))
V= R+(Z%*%(n/as.vector(t(d1)%*%W%*%d1)*G)%*%t(Z))%*%W
betahats[j,] <- c(beta,u)
ifelse(((norm(betahats[j,]-betahats[j-1,]))<epsilon),
        break,next)}
yhat= X%*% beta+Z%*%u
return(list(beta=betahats[j-1,1:p],
       u=betahats[j-1,(p+1):(p+q)],
       R=R, G=G,V=V, iterations=j-1,yhat=yhat))}
```

We defined the S-estimators for additive penalized spline smoothing for the case with outliers both on individual level and cluster level in Result 4.2. We used this function for all simulation studies in Chapter 4.

```
AdditivePenS2 = function(Y, X,Z,NN,N,n,m,p,q,r,beta1,
                        u1,R1,G1,V1, c1, epsilon){
#store the values in matrix
betahats = matrix(ncol=(p+(n*q)),nrow=NN)
betahats[1,] = c(beta1,u1)
beta=beta1; u=u1; R=R1; G=G1; V=V1
for (j in 2:NN){
# Compute mahalanobis distance and weight function
d2=rep(0,N); d1=rep(0,N)
for(i in 1:N){
mu = X[i,]%*% beta + Z[i,]%*% u
d2[i] = mahalanobis(Y[i],mu,R[i,i])}
d1 = sqrt(d2)
h = floor((N+(p+(q*n))+1)/2)
quantile <- h/(N+1)
d=(d1*sqrt(qchisq(quantile,(p+(n*q)))))/(sort(d1)[h])
# Compute the weights
wld =rep(1,N)
wld = as.vector(biwt.wt(d1,c1))
W=diag(wld)
tau1=2*n*m*(1/as.vector(t(d1)%*%W%*%d1))
# Compute distance22 and weight2 function
d22=rep(0,(n*q))
for(kk in 1:(n*q)){
d22[kk] = (sqrt(1/G[kk,kk]) *u[kk])}
# Compute the weights
wld2 =rep(1,(n*q))
wld2 = as.vector((Psi(d22,c1)/d22))
W2=diag(wld2)
gw2.inv=c(rep(1,(n*q)))
for(k in 1:(n*q) ){
gw2.inv[k] = G[kk,kk]/(wld2[j])}
GW2.INV = diag(gw2.inv)
tau2 = 2*q*r*(1/as.vector(t(d22)%*%W2%*%d22))
# Compute the fixed effect and random effect parameters
beta=as.vector(ginv(t(X)%*%W%*%ginv(V)%*%X)%*%t(X)
     %*%W%*%ginv(V)%*%Y)
u=as.vector(((q*tau1)/(n*tau2))*((GW2.INV)%*%t(Z)%*%W
  %*%ginv(V)%*%(Y-X%*% beta)))
# Compute the variance components
G=(as.vector(t(u)%*%W2%*% u)*(tau2/(2*(n*q)^2)))
  *diag(rep(1,(n*q)))
R=(as.vector((1/as.vector(t(d1)%*%W%*% d1)) %*%
  (t(Y-X%*%beta-Z%*%u)%*%W%*%(Y-X%*%beta-Z%*%u))))
  *diag(rep(1,N))
V=R+((q*tau1)/(n*tau2))*(Z%*%(GW2.INV)%*%t(Z)%*%W)
betahats[j,] = c(beta,u)
ifelse(((norm(betahats[j,]-betahats[(j-1),]))<epsilon)
        ,break,next)
}
yhat = (X%*% beta) + (Z %*% u)
return(list(beta=betahats[(j-1),1:p],
       u=betahats[(j-1),(p+1):(p+q)],
       R=R, G=G,V=V, iterations=j-1,yhat=yhat))
}
```

We defined the function for the conditional AIC for linear mixed models based on S-estimators for the case with outliers only on individual level in Theorem 4.1. We used this function for all simulation studies in Chapter 4. We used the function for the Conditional AIC for the linear mixed models based on maximum likelihood estimators for only outliers on individual level from Greven and Kneib (2010).

```
ccAIC.S1=function(Y,X,Z,beta,u,R,G,V,N,n,m,p,q,c1){
#Compute the weights
d2 = rep(0,N)
d = rep(0,N)
for(i in 1:N){
mu = X[i,]%*%beta+Z[i,]%*%u
d2[i] = mahalanobis(Y[i],mu,R[i,i])
}
d1 = sqrt(d2)
h = floor((N+(p+(n*q))+1)/2)
quantile = h/(N+1)
d = (d1*sqrt(1/qchisq(quantile,(p+(n*q)))))/(sort(d1)[h])
wld =rep(1,N)
wld = as.vector(biwt.wt(d,c1))
W=diag(wld)
tau1 = 2*n*m*(1/as.vector(t(d1)%*%W%*%d1))

#Conditional AIC with \hat\rho penalty term
H.S = (X%*%solve(t(X)%*%W%*%solve(V)%*%X)%*%t(X)%*%W
      %*%solve(V))+(Z %*%((n*as.vector(solve(t(d1)%*% W
      %*%d1)))*G)%*%t(Z)%*%W%*%solve(V))-(Z%*%
      ((n*as.vector(solve(t(d1)%*%W%*%d1)))*G)%*%t(Z)
       %*%W%*%solve(V)%*%X%*%solve(t(X)%*%W%*%solve(V)
       %*%X)%*%t(X)%*%W%*%solve(V))
C.AIC.S= -2*determinant(R)$modulus[1]- 2*sum(diag(H.S))

# Conditional AIC with \phi_s1 penalty term
solveV=solve(V)  #### inverse of V matrix
IN=diag(1,N,N)    #### Identity matrix with N x N
# define P matrix
P =IN-X%*%solve(t(X)%*%W%*%solveV%*%X)%*%t(X)
      %*%W%*%solveV
```

```
dd.sig0=matrix(0,N,1)
ddi.sig0=dWi.sig0=c(rep(0,N))
dV.sig0=dW.sig0=matrix(0,N,N)

for(i in 1:N){
ddi.sig0[i]=(1/(2*d1[i]))*t(Y[i]-X[i,]%*%beta+Z[i,]%*%u)
              *(1/R[i,i])*(1/R[i,i])*
              (Y[i]-X[i,]%*%beta+Z[i,]%*%u)
}
dd.sig0 = as.matrix(ddi.sig0)
for(i in 1:N){
dWi.sig0[i]=((d[i]*Psi(d[i],cc=c1)-dPsi(d[i],cc=c1))/
             d[i]^2)*ddi.sig0[i]
}
dW.sig0=diag(dWi.sig0)

dtau1.sig0=-2*m*n*(1/as.vector(t(d)%*%W%*%d))*
            (as.vector(2*(t(d)%*%W%*%dd.sig0)+
            (t(d)%*%dW.sig0%*%d)))*
            (1/as.vector(t(d)%*%W%*%d))

# Equation (4.35)
dV.sig0=IN+(Z%*%G%*%t(Z)%*%W)*((1/(2*n))*dtau1.sig0)+
        (tau1/(2*n))*(Z%*% G%*%t(Z)%*%dW.sig0)

dRVP.sig0=dRVP.sigj=matrix(0,N,N)
dsig0.Yk=dhsig0.sig0=dhsig0.Yk=matrix(0,1,1)
dsigj.Yk=dhsigj.sigj=dhsigj.Yk=matrix(0,1,1)
dRVP.sigj =array(0,dim=c(N,N,(n*q)))
dVP.sig0=matrix(0,N,N)

# Equation (4.34)
dRVP.sig0=solveV%*%P-R%*%solveV%*%(dV.sig0-
            (X%*%solve(t(X)%*% W%*%solveV %*% X)%*%t(X)%*%
            (W%*%solveV%*%dV.sig0-dW.sig0)))%*%solveV%*%P

dVP.sig0=-solveV%*%(dV.sig0-(X%*%solve(t(X)%*% W%*%solveV
            %*%X)%*%t(X)%*%(W%*%solveV%*%dV.sig0
            -dW.sig0)))%*%solveV%*%P

dd.Yk=matrix(0,N,1)
ddi.Yk=dWi.Yk=c(rep(0,N))
dW.Yk=matrix(0,N,N)
for(i in 1:N){
ddi.Yk[i]=(1/(d1[i]))*t(Y[i]-X[i,]%*%beta+Z[i,]%*%u)
             *(1/R[i,i])
}
dd.Yk = as.matrix(ddi.Yk)
for(i in 1:N){
dWi.Yk[i]=((d[i]*dPsi(d[i],cc=c1)-Psi(d[i],cc=c1))/
             d[i]^2)*ddi.Yk[i]
}
dW.Yk=diag(dWi.Yk)

# Equation (4.32) and Equation (4.33)
B=matrix(0,N,N)
sum.dRVP.sigj.Y.dsigj.Yk=matrix(0,N,1)
for(k in 1:N){
dhsig0.sig0= -(n/R[k,k]^4)-(((1/n)*t(Y))%*%t(P)%*%solveV
                 %*%(((W %*%solveV%*%P%*%Y)*dtau1.sig0)+
                 ((2*tau1)*(W%*%dVP.sig0%*%Y))+
                 ((tau1*dW.sig0)%*%solveV%*%P%*%Y)))

dtau1.Yk=-2*m*n*(1/as.vector(t(d)%*%W%*%d))*
           (as.vector(2*(t(d)%*%W%*%dd.Yk)+
           (t(d)%*%dW.Yk%*%d)))*(1/as.vector(t(d)%*%W%*%d))
```

```
dV.Yk=matrix(0,N,N)
dV.Yk=(((1/(2*n))*dtau1.Yk*Z)%*%G%*%t(Z)%*%W)+
      ((Z*(tau1/(2*n)))%*%G%*%t(Z)%*%dW.Yk)

dVPk.Yk=dPk.Yk=matrix(0,N,1)
dPk.Yk=(X%*%solve(t(X)%*%W%*%solveV%*%X)%*%t(X)%*%
       (W%*%solveV%*%dV.Yk-dW.Yk))%*%solveV%*%P[,k]

dVPk.Yk=-solveV%*%dV.Yk%*%solveV%*%P[,k]
          +solveV%*%dPk.Yk

dhsig0.Yk=-(((1/n)*Y[k])*t(P[,k]))%*%solveV%*%(((2*tau1*W)
          %*%solveV%*%P[,k])+((dtau1.Yk*W)%*%solveV%*%
          (P[,k]*Y[k]))+((2*tau1*W)%*%(dVPk.Yk*Y[k]))+
          ((tau1*dW.Yk)%*%solveV%*%(P[,k]*Y[k])))

#Equation (4.39)
dsig0.Yk= -( 1/as.vector(dhsig0.sig0)) * dhsig0.Yk
for(j in 1:(n*q)){
#Equation (4.36)
dRVP.sigj[,,j]=(-(tau1/(2*n))*R)%*%(solveV%*%P%*%Z[,j]%*%
                t(Z[,j])%*%W%*%solveV%*%P)

dhsigj.sigj=-((n*q)/(G[j,j])^4)-(2*(tau1/(2*n))^3*t(t(Z[,j])
             %*%W%*%solveV%*%P%*%Y))%*%Z[,j]%*%W%*%solveV
             %*%P%*%Z[,j]%*%t(Z[,j])%*%W%*%solveV%*%P%*%Y

dhsigj.Yk=-(2*(tau1/(2*n))*t(t(Z[,j])%*%W%*%solveV%*%P%*%Y))
           %*%(((tau1/(2*n))*t(Z[,j])%*%((W%*%solveV%*%P[,k])
           +(dW.Yk%*%solveV%*%(P[,k]*Y[k]))+(W%*%
           (dVPk.Yk*Y[k]))))+(((dtau1.Yk/(2*n))*t(Z[,j]))%*%
            W%*%solveV%*%(P[,k]*Y[k])))
#Equation (4.40)
dsigj.Yk[j] = -( 1/as.vector(dhsigj.sigj)) * dhsigj.Yk

sum.dRVP.sigj.Y.dsigj.Yk=sum.dRVP.sigj.Y.dsigj.Yk+
                         (dRVP.sigj[,,j]%*%Y %*%dsigj.Yk[j])
}
B[,k]=dRVP.sig0%*%Y%*%dsig0.Yk+sum.dRVP.sigj.Y.dsigj.Yk
}
# Equation (4.24) and (4.31)
phiS1= sum(diag(IN - (R %*% solveV %*% P)- B))
# Equation (4.26)
CC.AIC.S1= -2* determinant(R)$modulus[1]- 2*phiS1
return(list(CAICS=C.AIC.S,CCAICS1=CC.AIC.S1))
}
```

We can define the function for the conditional AIC for linear mixed models based on S-estimators for the case with outliers both on individual level and cluster level in Theorem 4.2 similarly and use the S-estimators from Result 4.2.

Chapter 6

Future research

Further research could be done on some specific issues related to the results in this dissertation. I present some possible extensions in this Chapter.

We used and proposed the model selection method AIC for data sets where the sample size is strictly larger than the number of independent variables. In practice, often we need to deal with data sets that contain more independent variables than the sample size. It would be interesting to extend our current work to derive robust model selection methods based on S-estimators for data with more independent variables than the sample size. One idea could be to add an l_1 penalty to the optimization criterion in the spirit of the lasso estimation method (Tibshirani, 1996). This would form an extension on the robust lasso method based on least absolute deviation (LAD) estimators (Wang et al., 2007).

Another interesting direction for further research is an extension of the robust model selection methods in this dissertation to the context of generalized linear mixed models. There exist several non robust model selection methods for generalized linear mixed models in the literature, for example, Cai et al. (2006), Chen et al. (2003) and Lavergne et al. (2008). Robust estimation based on M-estimators in generalized linear mixed models is proposed in Yau and Kuk (2002). Some of the ideas in the given references could turn out useful for the application of S-estimation methods and of model selection using these S-estimation in this setting.

Another possible extension of the robust model selection of mixed models is in the context of survival models and of frailty models. Liang and Zou (2008) and (Ibrahim and Chen, 2005) proposed some model selection methods for survival models. Hjort and Claeskens (2006) is concerned with variable selection methods for the proportional hazards regression model based on a focussed information criterion. Xu et al. (2009) proposed a semiparametric model selection method with application to proportional hazards mixed models using profile likelihood. All of these model selection methods are non robust for outliers in the data. Ideas in those papers could be used to propose a robust version of AIC for survival models. The frailty model, on the other hand, can be represented in a mixed model form, which suggests using a robust conditional Akaike information criterion for linear mixed models. Ha et al. (2007) study an Akaike information criterion (AIC) for selecting a frailty structure from a set of non-nested frailty models. They propose two new AIC criteria, one based on a conditional likelihood and the other on an extended restricted likelihood (ERL) as given by Lee and Nelder (1996). The frailty models and several numerical techniques are discussed in detail in the book (Duchateau and Janssen, 2008).

List of Figures

List of Tables

Bibliography

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Statistics and Probability Letters*, 56:289–300.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

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

Besse, P., Cardot, H., and Ferraty, F. (1997). Simultaneous nonparametric regression of unbalanced longitudial data. *Computational Statistics & Data Analysis*, 24:255–270.

Brumback, B., Ruppert, D., and Wand, M. (1999). Comment on "variable selection and function estimation in additive nonparametric regression using a data-based prior". *Journal of the American Statistical Association*, 94:794–797.

Cai, B., Dunson, D. B., and Gladen, T. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics*, 62:446–457.

Cantoni, E. and Ronchetti, E. (2001a). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, 11(2):141–146.

Cantoni, E. and Ronchetti, E. (2001b). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030.

Chen, M.-H., Ibrahim, J. G., Shao, Q.-M., and Weiss, R. E. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111(1-2):57 – 76.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge, UK.

Claeskens, G., Krivobokova, T., and Opsomer, J. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96:529–544.

Copt, S. and Victoria-Feser, M. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association*, 101(473):292–300.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. *The Annals of Statistics*, 11(2):530–551.

Craven, P. and Whaba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403.

Davies, L. (1990). The asymptotics of S-estimators in the linear regression model. *The Annals of Statistics*, 18(4):1651–1675.

Donoho, D. and Huber, P. (1983). The notion of breakdown-point. In Bickel, P., Doksum, K., and Hodges, J. J., editors, *A Festschrift for Erich L. Lehmann*, pages 157–184, Belmont, CA. Wadsworth.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model.* Springer New York, New york, USA.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121. With comments and a rejoinder by the authors.

Garrett, R. (1989). The chi-square plot: a tools for multivariate outlier recognition. *Journal of Geochemical Exploration*, 32(1/3):319–341.

Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional akaike information criteria in linear mixed models. *Biometrika*, 97(4):773–789.

Ha, I. D., Lee, Y., and MacKenzie, G. (2007). Model selection for multi-component frailty models. *Statistics in Medicine*, 26:4790–4807.

Hall, P. and Jones, M. C. (1990). Adaptive M-estimation in nonparametric regression. *The Annals of Statistics*, 18:1712–1728.

Hall, P. and Opsomer, J. (2005). Theory for penalized spline regression. *Biometrika*, 92:105–118.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons, Inc., New York.

Härdle, W. and Gasser, T. (1984). Robust non-parametric function fitting. *Journal of the Royal Statistical Society, Series B*, 46:42–51.

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics.* Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester,West sussex, UK.

Hjort, N. and Claeskens, G. (2006). Focussed information criteria and model averaging for cox's hazard regression model. *Journal of the American Statistical Association*, 101:1449–1464.

Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika*, 88(2):367–379.

Hössjer, O. (1992). On the optimality of S-estimators. *Statistics and Probability Letters*, 14(5):413–419.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101.

Huber, P. (1979). Robust smoothing. In Wilkinson, G. and Launer, R., editors, *Robustness in Statistics*, New York. Academic Press.

Huber, P. (2004). *Robust Statistics*. John Wiley & Sons, Inc.

Ibrahim, J. G. and Chen, M.-H. (2005). *Bayesian Model Selection in Survival Analysis*. John Wiley & Sons, Ltd.

Johnston, J. and DiNardo, J. (1997). *Econometric Methods(Fourth edition)*. McGraw-Hill companies Inc.

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B*, 71(2):487–503.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 46:1071–1085.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83(4):875–890.

Lavergne, C., Martinez, M.-J., and Trottier, C. (2008). Empirical model selection in generalized linear mixed effects models. *Computational Statistics*, 23:99–109.

Lee, T. C. M. and Oh, H.-S. (2007). Robust penalized regression spline fitting with application to additive mixed modeling. *Computational Statistics*, 22(1):159–171.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, 58:619–678.

Liang, H., Wu, H., and ZOU, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, 95(3):773–778.

Liang, H. and Zou, G. (2008). Improved aic selection strategy for survival analysis. *Computational Statistics & Data Analysis*, 52(5):2538–2548.

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

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics Theory and Methods*. John Wiley & Sons, Ltd.

Maronna, R. A. and Yohai, V. J. (1991). The breakdown point of simultaneous general M-estimates of regression and scale. *Journal of the American Statistical Association*, 86(415):699–703.

Müller, S. and Welsh, A. (2005). Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, 100:1297–1310.

Oh, H.-S., Nychka, D. W., Brown, T., and Charbonneau, P. (2004). Period analysis of variable stars by robust smoothing. *Applied Statistics*, 53(1):15–30.

Oh, H.-S., Nychka, D. W., and Lee, T. C. M. (2007). The role of pseudo data for robust smoothing with application to wavelet regression. *Biometrika*, 94(4):893–904.

Omelka, M. and Salibián-Barrera, M. (2010). Uniform asymptotics for S- and MM-regression estimators. *The Annals of the Institute of Statistical Mathematics*, 62:897–927.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1:505–527. With discussion.

Qian, G. and Künsch, H. R. (1998). On model selection via stochastic complexity in robust linear regression. *Journal of Statistical Planning and Inference*, 75:91–116.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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. (1985). Robust model selection in regression. *Statistics and Probability Letters*, 3:21–23.

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

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

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

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In J. Franke, W. H. and Martin, R. D., editors, *Robust and nonlinear time series analysis (Heidelberg, 1983)*, volume 26 of *Lecture Notes in Statist.*, pages 256–272. Springer, New York.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge, UK.

Salibián-Barrera, M. and Van Aelst, S. (2008). Robust model selection using fast and robust bootstrap. *Computational Statistics & Data Analysis*, 52:5121–5135.

Salibian-Barrera, M. and Yohai, V. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15:414–427.

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

Sommer, S. and Staudte, R. G. (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics*, 37:323–336.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18. In Japanese.

Tharmaratnam, K. and Claeskens, G. (2011a). A comparison of robust versions of the AIC based on M, S and MM-estimators. *Statistics*, in press.

Tharmaratnam, K. and Claeskens, G. (2011b). Robust model selection in additive penalized regression splines models. In *The conference proceedings of the 26th International Workshop on Statistical Modelling*, Valncia.

Tharmaratnam, K., Claeskens, G., Croux, C., and Salibian-Barrera, M. (2010). S-estimation for panalized regression splines. *Journal of Computational and Graphical Statistics*, 19(3):609–625.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92:351–370.

Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25(3):347–355.

Weisberg, S. (2005). *Applied Linear Regression.* John Wiley & Sons Inc., Hoboken, New Jersey., 3rd edition.

Welsh, A. H. and Richardson, A. M. (1997). Approaches to the robust estimation of mixed models. In *Robust inference*, volume 15 of *Handbook of Statistics.*, pages 343–384. North-Holland, Amsterdam.

Xu, R., Vaida, F., and Harrington, D. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica*, 19:819–842.

Yau, K. K. W. and Kuk, A. Y. C. (2002). Robust estimation in generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:101–117.

Doctoral dissertations from the Faculty of Business and Economics

A list of doctoral dissertations from the Faculty of Business and Economics can be found at the following website:
`http://www.kuleuven.be/doctoraatsverdediging/archief.htm`.