Nationa Institute of Meteorology and Hydrology
Bulgarian Academy of Sciences

Doctoral disertation

Robust Statistical Modelling Through Trimming

Neyko M. Neykov

Contents

General Introduction - Motivation and Outlook of Robust Statistics

0.1 Introduction

Robust statistics is concerned with statistical procedures leading to inference that is stable with respect to departures of the data from model assumptions. Data inadequacies often occur in the form of contamination, anomalous values or outliers, which are in disagreement with the data generating mechanism. Outliers do occur in real data, for example as gross errors. They are data which are far way from the *majority*, *the bulk* of the data. They can be due to: gross errors - copying or punching error, in particular wrong decimal point, interchange of two values with different meaning, equipment failures etc. Outliers are sample values which cause surprise in relation to the majority of the sample. They are usually $influential$ observations, that is, their deletion often causes major changes in estimates, confidence regions, tests, and so on. As the values and frequency of outliers strongly fluctuate from sample to sample, outliers can make conclusions of a statistical analysis unreliable.

Outliers are more likely to occur in datasets with many observations and/or variables, and often they do not show up by simple visual inspection. Thus in multiple regression (and other complex designs), outliers may be vary hard to find. In the computer age, we therefore need reliable routine methods for finding all outliers. Once found, the outliers should still be studded and interpreted, and not automatically be rejected (except in certain routine

situation). We also should have reasonably efficient methods for dealing with "doubtful outliers" (borderline or, rather, bortherzone cases).

The main purpose of any robust procedure is to give resistant (stable) results in the presence or absence of outliers by best fitting the majority, the bulk of the data. This means that robust estimation finds a fit, which is similar to the fit we would have found without the outliers. Robust statistics deals with deviations (such as gross errors) from ideal parametric models (such as normality) and with statistical procedures that are still reliable and reasonably efficient under smaller and larger deviations from the parametric model used. Robust procedures are based on the use of parametric models and must be resistant with respect to outliers and efficient with respect to the sample variation of the majority.

A great deal of work has been done in developing the theory and methodology of robustness. Nowadays robust techniques have been developed in practically any field in statistical analysis. The milestones are books by Huber (1981), Hampel et al. (1986), Staudte and Sheather (1990), Maronna et al. (2006), Huber and Ronchetti (2009).

The book of Rousseeuw and Leroy (1987) is mainly concerned with robust detection of regression and multivariate outliers based on trimming and is very practical. Marazzi (1993) documents a set of FORTRAN routines for robust statistics with interface to S-PLUS. Atkinson and Riani (2000), and Atkinson, Riani and Cerioli (2004) combine robustness with high BDP and various regression and multivariate data influential diagnostics and computer graphics. Atkinson and Riani (2001) adapted the so called Forward search algorithm in order to compute the parameter estimate of the generalized linear linear regression models within the exponential family of distributions.

The book of Heritier et al. (2009) gives robust methods in biostatistical modeling and statistical inference in general. Varmuza and Filzmoser (2008) dissiminate the robust statistics in chemomemtrics whereas the book of Farcomeni and Greco (2015) is about robust methods for data reduction techniques such as principal component and factor analysis, discriminata analysis, and clustering. Review papers about robustness with high BDP can be found in Hubert et al. (2005) and Hubert et al. (2008).

A global measure of robustness of a statistical estimator is the finite sample breakdown point (BDP). It measures the smallest fraction of contamination that can cause the estimator to take arbitrary large values. We now recall the replacement variant of the finite sample BDP given in Hampel et al. (1986), which is closely related to that introduced by Donoho and Huber (1983). Let $\Omega = \{\omega_i \in R^p, \text{ for } i = 1, \ldots, n\}$ be a sample of size n.

Definition 0.1 *The breakdown point of an estimator T at Ω is given by*

$$\varepsilon_n^*(T) = \max\{\frac{m}{n} : \sup_{\tilde{\Omega}_m} \|T(\Omega) - T(\tilde{\Omega}_m)\| < \infty\},$$

where $\tilde{\Omega}_m$ is any sample obtained from Ω by replacing any m of the points in Ω by arbitrary values and $\|.\|$ is the Euclidean norm. .

Thus, there is a compact set such that the estimator T remains in it even if we replace any m elements of the sample Ω by arbitrary ones. The largest m/n for which this property holds is the breakdown point.

One way to construct a positive BDP estimator is to employ a standard estimator and to trim some unusual, discordant, unlikely observations from the corresponding objective function. For example in linear regression, this is the case of the Least Median of Squares (LMS) and Least Trimmed Squares (LTS) introduced by Rousseeuw (1984), the Least Trimmed Absolute Deviations (LTAD) by Bassett (1991), and the Maximum Trimmed Likelihood estimator (TLE) by Neykov and Neytchev (1990).

This thesis is dedicated to robust estimators based on trimming and their applications in fitting statistical models to data.

0.2 Regression estimators based on trimming

In order to aid the presentation we remind some basic definitions.

Consider the classical linear regression model

$$y_i = x_i^T\theta + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, n,$$

where $y_i \in R^1$, $x_i \in R^p$, $\theta \in R^p$, ε_i are i.i.d., $\mathrm{E}(\varepsilon_i) = 0$, and $\mathrm{var}(\varepsilon_i)=\sigma^2 > 0$. Denote the residuals by

$$r_i(\theta) := y_i - x_i^T\theta \quad \text{for} \quad i = 1, \ldots, n.$$

Definition 0.2 *The Least Squares Estimator (LSE) is defined as*

$$\hat{\theta}_{LSE} := \arg\min_{\theta} \sum_{i=1}^{n} r_i^2(\theta).$$

Definition 0.3 *The Least Absolute Deviations (LAD) is defined as*

$$\hat{\theta}_{LAD} := \arg\min_{\theta} \sum_{i=1}^{n} |r_i(\theta)|,$$

Definition 0.4 *The Least Median of Squares (LMS), Least Quantile of Squares (LQS) and Least Trimmed Squares (LTS) are defined by Rousseeuw (1984):*

$$\begin{aligned}
\hat{\theta}_{MED} &:= \arg\min_{\theta} \operatorname{med}_i r_i^2(\theta), \\
\hat{\theta}_{LQS} &:= \arg\min_{\theta} r^2_{\nu(k)}(\theta), \\
\hat{\theta}_{LTS} &:= \arg\min_{\theta} \sum_{i=1}^{k} r^2_{\nu(i)}(\theta),
\end{aligned}$$

where $r^2_{\nu(1)}(\theta) \le r^2_{\nu(2)}(\theta) \le \ldots \le r^2_{\nu(n)}(\theta)$ are the ordered values of $r_i^2(\theta)$ at θ, $\nu = (\nu(1), \ldots, \nu(n))$ is the permutation of the indices (depends on θ), k is the trimming parameter such that $\lfloor \frac{n+p+1}{2} \rfloor \le k \le n$ if the observations are in general position, i.e., any p of them are linearly independent.

Definition 0.5 *The Least Trimmed Absolute Deviations (LTAD) is defined in Rousseeuw and Leroy (1987):*

$$\hat{\theta}_{LTAD} := \arg\min_{\theta} \sum_{i=1}^{k} f_{\nu(i)}(\theta),$$

where $f_{\nu(1)}(\theta) \le f_{\nu(2)}(\theta) \le \ldots \le f_{\nu(n)}(\theta)$ are the ordered values of $f_i(\theta) = |r_i(\theta)|$ at θ, $\nu = (\nu(1), \ldots, \nu(n))$ is the permutation of the indices (depends on θ), k is the trimming parameter such that $\lfloor \frac{n+p+1}{2} \rfloor \le k \le n$ if the observations are in general position, i.e., any p of them are linearly independent.

From the above definitions it follows that the minima are achieved over a subsample of size k. On the other hand the objective function LMS, LQS and LTS are continuous, but non differentiable and possesses many local minima. Therefore one need non-smooth and/or combinatorial optimization in general to get the corresponding minima. The following representation of the LQS due to Krivulin (1992) is very useful as it clarifies its combinatorial nature over the observations

$$\min_{\theta} r^2_{\nu(k)}(\theta) = \min_{\theta} \min_{I \in I_k} \max_{i \in I} r_i^2(\theta),$$

where I_k is the set of all k–subsets of the set $\{1, \ldots, n\}$, whereas $I = \{i_1, \ldots, i_k\}$. This representation seems to be more useful than the original because it is based on the well-known *min* and *max* functions. Moreover, it allows of further reducing the problem. One can change the order of the operations of taking minimum and get

$$\min_{\theta} r^2_{\nu(k)}(\theta) = \min_{\theta} \min_{I \in I_k} \max_{i \in I} r_i^2(\theta) = \min_{I \in I_k} \min_{\theta} \max_{i \in I} r_i^2(\theta).$$

The same holds about the LTS estimators

$$\min_{\theta} \sum_{i=1}^{k} r^2_{\nu(i)}(\theta) = \min_{\theta} \min_{I \in I_k} \sum_{i \in I} r_i^2(\theta) = \min_{I \in I_k} \min_{\theta} \sum_{i \in I} r_i^2(\theta).$$

Therefore all possible $\binom{n}{k}$ subsets of the data have to be fitted by the LSE. The LQS and LTS estimators trim at most $n - k$ of the observations that do not follow the assumed mode. Computing the LQS and LTS is infeasible for large data sets. To get an approximate estimate a FAST-LTS was developed by Rousseeuw & van Driessen (2000).

Similar algorithm holds about LQS, however, the LTS is $\sqrt{n}$ consistent and asymptotically normal which are desired properties of any statistical estimator.

The same hold also about the LQAD estimator $f_{\nu(k)}(\theta)$ where $f_i(\theta) = |r_i(\theta)|$

$$\min_{\theta} f_{\nu(k)}(\theta) = \min_{\theta} \min_{I \in I_k} \max_{i \in I} f_i(\theta) = \min_{\theta} \min_{I \in I_k} \max_{i \in I} |r_i(\theta)| = \min_{I \in I_k} \min_{\theta} \max_{i \in I} |r_i(\theta)|,$$

which is the well known problem of fitting a linear function according to the so called l_∞ criterion or Chebishev norm as well as about the LTAD

$$\min_{\theta} \sum_{i=1}^{k} f_{\nu(k)}(\theta) = \min_{\theta} \min_{I \in I_k} \sum_{i \in I} f_i(\theta) = \min_{\theta} \min_{I \in I_k} \sum_{i \in I} |r_i(\theta)| = \min_{I \in I_k} \min_{\theta} \sum_{i \in I} |r_i(\theta)|.$$

In conclusion the optimization problems may be regarded as a "two-stage" problems of both combinatorial optimization over the observations and Least Squares optimization or Linear Programming if LQAD and LTAD-estimators are to be considered.

Due to the representations it follows that the LMS, LQS, LTS, LQAD and LTAD estimators are regression, scale and affine-equivariant estimators.

The BDP of the LMS, LTS and related trimmed rank based estimators were derived by Rousseeuw (1984), Rousseeuw and Leroy (1987), and Hössjer (1994), assuming that the observations are in general position whereas Müller (1997) and Mili and Coakley (1996) omitted this restriction and gave a general treatment with replicated regression data. We remind that the observations $x_i \in R^p$ for $i = 1, \ldots, n$ are in general position if any p of them are linearly independent. The BDP properties of the LTS and LTAD estimators were studied also by Vandev na Neykov (1998) based on the concept of d–fullness for data in general position.

The LTS estimator belongs to the class of affine equivariant estimators which achieves asymptotically the highest breakdown point 0.5. It is the right choice instead of the LMS estimator of Rousseeuw (1984) because of its asymptotic (normality and consistency) properties and computational aspects. Stromberg (1993) studied the properties of the nonlinear regression LTS estimator. Visek (2002) proved the consistency and asymptotic normality of the weighted LTS whereas Čížek (2002), and Gervini and Yohai (2002) did the same about the LTS adaptive choice of trimming.

Efficient FAST-LTS computational algorithm has been developed by Rousseeuw and Van Driessen (1999a). Agullo (2001) considered a branch and bound algorithm to get the LTAS estimate but it is appropriate for samples of small size. Hawkins and Olive (1999) and Hawkins and Olive (2002) studied the asymptotic properties of the LTAD and LTS estimators and discussed some computational algorithms as well.

The BDP properties of the LTS and LTAD estimators were studied also by Vandev na Neykov (1998) based on the concept of d–fullness for data in general position.

0.3 The trimmed likelihood and related estimators

Let $x_i \in R^p$ for $i = 1, \ldots, n$ be i.i.d. observations with pdf $\psi(x, \theta)$, $\theta \subseteq \Theta^q$ is unknown parameter, and $l_i(\theta) = l(x_i, \theta) = -\log \psi(z_i, \theta)$. Neykov and Neytchev (1990) proposed to replace in the Rousseeuw's estimators the squared residuals $r_i^2(\beta)$ by the negative log likelihoods $l_i(x_i, \theta)$ and thus the following two classes of estimators are defined:

Definition 0.6 *(Neykov and Neytchev, 1990) The minimum Median Likelihood Estimator (MedLE(k)) and minimum Trimmed Likelihood Estimator (TLE(k)) is defined as*

$$\hat{\theta}_{MedLE} := \arg\min_{\theta \in \Theta} l(x_{\nu(k)}, \theta) \qquad \text{and} \qquad \hat{\theta}_{TLE} := \arg\min_{\theta \in \Theta} \sum_{i=1}^{k} l(x_{\nu(i)}, \theta),$$

where $l(x_{\nu(1)}, \theta) \leq l(x_{\nu(2)}, \theta) \leq \ldots \leq l(x_{\nu(n)}, \theta)$ *are the ordered values of* $l(x_i, \theta)$ *for* $i = 1, \ldots, n$ *at* θ, $\nu = (\nu(1), \ldots, \nu(n))$ *is the corresponding permutation of the indices, which depends on* θ *and* k *is the trimming parameter.*

The basic idea behind the trimming in this estimator is in removal of those $n - k$ observations which values would be highly unlikely to occur, had the fitted model been true. The TLE coincides with the MLE if $k = n$. Due to the representation

$$\min_{\theta \in \Theta} \sum_{i=1}^{k} l(x_{\nu(i)}, \theta) = \min_{\theta \in \Theta} \min_{I \in I_k} \sum_{i \in I} l(x_i, \theta) = \min_{I \in I_k} \min_{\theta \in \Theta} \sum_{i \in I} l(x_i, \theta)$$

where I_k is the set of all k–subsets of the set $\{1, \ldots, n\}$, it follows that all possible $\binom{n}{k}$ partitions of the data have to be fitted by the MLE. Therefore, the TLE is given by the partition with that MLE fit for which the negative log likelihood is minimal.

Vandev (1993) put the LMS, LTS, LTAD, MedLE and TLE estimators into a general class of positive functions. Let $f : X \times \Theta \to \mathbb{R}^+$, where $\Theta \subseteq \mathbb{R}^q$ be an open set, and $F = \{f_i(\theta) = f(x_i, \theta), \text{ for } i = 1, \ldots, n\}$.

Definition 0.7 *(Vandev, 1993) The Generalized Median Estimator (GMedE(k)) and Generalized Trimmed Estimator (GTE(k)) are defined as*

$$\hat{\theta}^{\mathrm{k}}_{\mathrm{GMedE}} := \arg\min_{\theta \in \Theta^q} f_{\nu(k)}(\theta) \qquad \text{and} \qquad \hat{\theta}^{\mathrm{k}}_{\mathrm{GTE}} := \arg\min_{\theta \in \Theta} \sum_{i=1}^{k} f_{\nu(i)}(\theta)$$

where $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \ldots \leq f_{\nu(n)}(\theta)$ *are the ordered values of* f_i *at* θ, $\nu = (\nu(1), \ldots, \nu(n))$ *is the corresponding permutation of the indices, which depends on* θ, k *is the trimming parameter, the weights* $w_i \geq 0$ *for* $i = 1, \ldots, n$ *are associated with the functions* $f_i(\theta)$ *and are such that* $w_{\nu(k)} > 0$.

Vandev (1993) developed d–fullness technique in order to study their BDP properties.

Definition 0.8 *(Vandev, 1993) A set of functions* $F = \{f_i(\theta), \text{ for } i = 1, \ldots, n\}$ *is called* d*–full if for every subset* $J \subset \{1, \ldots, n\}$ *of cardinality* d *(*$|J| = d$*) the function* $g_{_J}(\theta) = \max_{j \in J} f_j(\theta)$, $\theta \in \Theta$, *is subcompact.*

Definition 0.9 *(Vandev, 1993; Vandev and Neykov, 1993) A function* $g : \Theta \to \mathbb{R}$, $\Theta \subseteq \mathbb{R}^q$ *is called subcompact if its Lebesgue set* $L_g(C) = \{\theta \in \Theta : g(\theta) \leq C\}$ *is a compact set for every real constant* C.

Later on Neykov (1995) introduced the Weighted Generalized Trimmed Estimators (wGTE(k)) and studied their BDP properties with the d–fullness concept.

Definition 0.10 *The weighted Generalized Trimmed Estimator (wGTE) is defined as*

$$\hat{\theta}_{WTE} := \arg\min_{\theta \in \Theta} \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta), \tag{1}$$

where $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \ldots \leq f_{\nu(n)}(\theta)$ *are the ordered values of* f_i *at* θ, $\nu = (\nu(1), \ldots, \nu(n))$ *is the corresponding permutation of the indices, which depends on* θ, k *is the trimming parameter, the weights* $w_i \geq 0$ *for* $i = 1, \ldots, n$ *are associated with the functions* $f_i(\theta)$ *and are such that* $w_{\nu(k)} > 0$.

A particula case of the wGTE(k), called WTLE, is obtained if the functions in (1) are replaced by the negative log-likeligoods. The main wGTE(k) results and findings with some additional application concerning characterization of the BDP of the WTLE within the framework of linear logistic regression and linear regression model with Laplace of order q of the errors under the assumption that the data are in general position were published in Vandev and Neykov (1998).

Finally, we note that Hadi and Luceño (1997) defined a version of the WTLE closely following Neykov and Neytchev (1990), and Vandev and Neykov (1993) and offered a computational algorithm in the univariate case. These authors "introduced" also the MedLE(k) but for some reason missed to cite that the MedLE(k) definition not only had already been given by Vandev (1993) and Vandev and Neykov (1993) but its BDP properties were characterized.

0.4 The thesis structute

This thesis is dedicated to robust estimators based on trimming and their applications in fitting statistical models to data. The estimators based on trimming have been developed as alternatives of the classical statistical estimators such as the Least Squares and Maximum Likelihood estimatos in order to reduce the outliers influence in data. The basic idea behind trimming is in the removal of those observations whose values would be highly unlikely to occur if the fitted model was true.

The following paper are summarized in the thesis as separate chapters:

Ch.1. Müller, Ch. and Neykov, N. M. (2003). Breakdown Points of the Trimmed Likelihood and Related Estimators in Generalized Linear Models. *J. Statist. Plann. and Inference*, **116**, 503-519.

Ch.2. Neykov, N. M. and Müller, Ch. (2003). Breakdown Point and Computation of Trimmed Likelihood Estimators in Generalized Linear Models. In: *Developments in Robust Statistics*, Dutter, R., Filzmoser, P., Gather, U., and Rousseeuw, P. (eds.), Physica-Verlag, Heidelberg, 277-286.

Ch.3. Dimova, R. and Neykov, N. M. (2004). Generalized d-fullness Technique for Breakdown Point Study of the Trimmed Likelihood Estimator with Applications. In: *Theory and Applications of Recent Robust Methods*, M. Hubert, G. Pison, A. Struyf and S. Van Aelst (eds.), Birkhauser, Basel, 83-92.

Ch.4. Neykov, N.M., Dimova, R. and Neytchev, P.N. (2005). Trimmed Likelihood Estimation of the Parameters of the Generalized Extreme Value Distribution: A Monte-Carlo Study. Pliska Stud. Math. Bulgar. 17, 187-200.

Ch.5. Neykov, N. M., Filzmoser, P., Dimova, R. and Neytchev, P. N. (2007). Robust fitting of mixtures using the Trimmed Likelihood Estimator. *Comput. Statist. Data Anal.*, **52**, 299-308. developed the so called Forward search algorithm to compute the LMS and

Ch.6. Neykov, N. M., Filzmoser, P. and Neytchev, P. N. (2012). Robust joint modeling of mean and dispersion through trimming. *Comput. Statist. Data Anal.* **56**, 34-48.

Ch.7. Neykov, N. M., Čížek, P., Filzmoser, P. and Neytchev, P.N. (2012). The least trimmed quantile regression. *Comput. Statist. Data Anal.* **56**, 1757-1770.

Ch.8. Neykov, N. M., Filzmoser, P. and Neytchev, P. N. (2014). Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. *Stat. Papers*, **55**, 187-207.

0.5 Short thesis survey

In chapter 2, a general class called S–estimators based on majorization - minorization of the kth ordered element of a d–full set of functions is introduced by Müller and Neykov (2003) and its BDP is characterized. As a consequence of this the main results Theorem 1 of Vandev and Neykov (1998) about the BDP of the wGTE estimator were extended and a lower bound without any additional sample size and trimming parameter assumption was found. Müller and Neykov (2003) consider the BDP behavior of the TLE estimator within the framework of the generalized linear models and gave a detail derivations for the linear logistic regression, Poisson linear regression model and linear regression model with Laplace error of q order omitting the requirement of Vandev and Neykov (1998) for general position of the data. The BDP of the linear regression S-estimator of Rousseeuw

and Yohai (1985) and Rousseeuw and Leroy (1987) is characterized using the d–fullness concept. Details can be found in Chapter 2.

We note that the results of Müller and Neykov (2003) about the linear logistic regression were further developed by Čížek (2008a) considering the adaptive maximum symmetrically trimmed likelihood estimator in order to overcome the problems with nonexistence of the TLE solution in case of no overlap between observations. Čížek (2008) prove the TLE consistency and asymptotic normality, and demonstrate its applicability in nonlinear regression, time series, and limited dependent variable models (these are the generalized linear models in econometric literature).

Computation of the TLE is unfeasible for large data sets because of its combinatorial nature. To get approximate TLE an algorithm called FAST-TLE was developed by Neykov and Müller (2003). The concentration steps of this algorithm are reduced to the FAST-LTS or FAST-MCD algorithms developed by Rousseeuw and Van Driessen (1999a) and Rousseeuw and Van Driessen (1999b) in case of normal linear regression parameter and multivariate Gaussian mean and covariance matrix estimation, respectively. Details can be found in Chapter 3.

The requirement for d–fullness of the set F is restrictive, more precisely, the condition *for every real constant* C in the definition of a subcompact function is not always satisfied. For instance the corresponding set F of negative log-likelihoods for the mixtures of univariate and multivariate normal are not d–full in the above sense. To overcome the problems Dimova and Neykov (2004) developed a generalized d–fullness techniques to study the BDP of a wider class of functions containing the class of subcompact functions. The main results are Proposion 3.1 and 3.2 which give the necessary conditions under which there exists a solution of the corresponding optimization problem and a lower bound for the BDP of the $wGTE(k)$ estimator for a set of functions F satisfying the conditions A1 and A2. These results are a generalization of Theorem 1 of Vandev and Neykov (1998). A generalization of the corresponding result of Vandev and Neykov (1998) and Müller and Neykov (2003) about linear logistic regression model with generalized link function is given by Proposion 3.3. Details are given in Chapter 4.

In Neykov et al. (2005) the applicability of the TLE is considered within the framework of the extreme value distributions. The index of fullness for the negative log-likelihoods of the Gumbel density is derived. The finite sample properties of the MLE and TLE are studied in a comparative simulation Monte Carlo study, The study shows that the MLE can be easily destroyed by one or several discordant observations. A strategy for trimming parameter choice is discussed. Details can be found in Chapter 5.

Another application of the TLE is proposed in Neykov et al. (2007) to estimate mixture of distributions in a robust way. The superiority of this approach in comparison with the MLE is illustrated by examples and simulation studies. The FAST-TLE algorithm is adapted to carry out the computation of the unknown parameters. The BDP of the TLE for the mixture components is characterized by the d–fullness concept. The relationship of the TLE with various other approaches that have incorporated robustness in fitting mixtures and clustering are also discussed in this context such as the classification trimmed likelihood estimator. An adaptive way for selection of the TLE trimming parameter k based on the robust version of the Bayesian Information Criteria is proposed as well. The proposed robust estimation technique is demonstrated on 3 data sets with contamination - mixture of 3 simple regression lines, mixture of two Poisson regression and mixture of 3 components of two-variate Gaussian distributions. A strategy for trimming parameter choice based on trimmed version of the Bayesian Information Criteria is offered. Details are given in Chapter 6.

It is worth to mention that a group of researchers from Spain has published a series of papers based on the classification trimmed likelihood estimator. Details can be found in Garcia-Escudero et al. (2008), Garcia-Escudero et al. (2010a), Garcia-Escudero et al. (2010b), Garcia-Escudero et al. (2011), Garcia-Escudero et al. (2013), Garcia-Escudero et al. (2014), Garcia-Escudero et al. (2015) and Garcia-Escudero et al. (2016). In order to disseminate robustness usage within the framework of clusterwise linear regression and multivariate data clustering these authors developed fast and reliable software in R following the ideas of the FAST-LTS of Rousseeuw and Van Driessen (1999a)and FAST-TLE of Neykov and Müller (2003) algorithms. Details can be found in Fritz et al. (2013a),

Fritz et al. (2013b) and Ruwet et al. (2013).

The MLE and the Extended Quasi-Likelihood (EQL) estimators have commonly been used to estimate the unknown parameters within the joint modeling of mean and dispersion framework. Particular cases of this framework are the classical linear regression model with heteroskedastic errors and generalized linear models with distributions from the linear exponential family with non-constant dispersion. In order to overcome the sensitivity of these estimators to outliers in the data Neykov et al. (2012a) introduced the maximum Extended Trimmed Quasi-Likelihood (ETQL) estimator to estimate the unknown parameters in a robust way. The BDP of this class of estimators is characterized by Theorem 6.1. The superiority of the proposed estimator in comparison with the classical MLE and EQL estimators is illustrated by examples and an extended Monte Carlo simulation study. Details are given in Chapter 7.

The linear quantile regression estimator is very popular and widely used during the last 40 years. It is well known that this estimator can be very sensitive to leverage observations in data (the discordant observations in the explanatory variables). In order to reduce the influence of the leverage observations in data, the least trimmed quantile regression estimator is proposed by Neykov et al. (2012b) in order to estimate the unknown parameters in a robust way. The BDP of the proposed estimator is characterized by Theorem 7.1 whereas its consistency is proved by Theorem 7.2. The performance of this approach in comparison with the classical one is illustrated by an example and an extended simulation study. Details can be found in Chapter 8.

The Penalized Maximum Likelihood Estimator (PMLE) has been widely used for explanatory variable selection in high-dimensional data when the sample size n is comparable or less than p the dimension of the explanatory variables. The penalized least squares estimator and MLE are non-robust against outliers in the data just as the classical estimators. To overcome this problem, the penalized M-estimator has been employed (Fan and Li, 2001; Fan and Lv, 2010). However, within regression models, M-estimators are not robust against outlying observations in the explanatory variables, the so called leverage points, and therefore penalized M-estimators are not robust in such settings as well. Only some

redescending M-estimators are robust in linear regression settings with fixed designs, e.g., Mizera and Müller,(1999) and Bühlmann and van der Geer (2011)). Neykov et al. (2014) proposed the Penalized Maximum Trimmed Likelihood Estimator (PMTLE) to estimate the unknown parameters in a robust way. The computation of the PMTLE takes advantage of the same technology as used for PMLE but here the estimation is based on subsamples only. The BDP properties of the PMTLE are discussed using the notion of d–fullness in several settings. The performance of the proposed estimator is evaluated in a simulation study for the classical multiple linear and Poisson linear regression models. Details are given in Chapter 9.

The newly developed methodology based on trimming has been widely used in real everyday practice. We assess this by the increasing citation number of our papers given in the Appendix.

Chapter 1

Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models

Summary. Lower bounds for breakdown points of trimmed likelihood (TL) estimators in a general setup are expressed by the fullness parameter of Vandev (1993) and results of Vandev and Neykov (1998) are extended. A special application of the general result are the breakdown points of TL estimators and related estimators as the S estimators in generalized linear models. For the generalized linear models, a connection between the fullness parameter and the quantity $\mathcal{N}(X)$ of Müller (1995) is derived for the case that the explanatory variables may be not in general position which happens in particular in designed experiments. These results are in particular applied to logistic regression and log-linear models where also upper bounds for the breakdown points are derived.

1.1 Introduction

Assume that the distribution of an observation Y_n has the density $f_n(y_n, \theta)$ and that the observations $Y_1, \ldots, Y_N$ are independent. Let $y := (y_1, \ldots, y_N)^\top$ be the vector of all realized observations, $l_n(y, \theta) := -\log f_n(y_n, \theta)$ the log-likelihood, and $l(y, \theta) := (l_1(y, \theta), \ldots, l_N(y, \theta))^\top$. Maximum likelihood (ML) estimators are maximizing the likelihood, i.e. minimzing

$$\sum_{n=1}^{N} l_n(y, \theta)$$

. Trimming the least likely observations, i.e. the observations with the largest $l_n(y, \theta)$, leads to trimmed likelihoods. Maximizing the trimmed likelihood provides the trimmed likelihood estimators $TL_h(y)$ given by

$$TL_h(y) := \arg\min_{\theta} \sum_{n=1}^{h} l_{(n)}(y, \theta),$$

where $N - h$ observations are trimmed and $l_{(1)}(y, \theta) \leq \ldots \leq l_{(N)}(y, \theta)$. These estimators can be also extended to weighted trimmed likelihood estimators WTL_h defined by

$$WTL_h(y) := \arg\min_{\theta} \sum_{n=1}^{h} w_n \, l_{(n)}(y, \theta),$$

where the weights satisfy $w_n \geq 0$ for $n = 1, \ldots, h$ and $w_h > 0$. See e.g. Hadi and Luccño (1997) and Vandev and Neykov (1998).

The weighted trimmed estimators will be used if some outliers are expected. Outliers are observations which differ from the majority of the observations and in particular do not possess the density $f_n(y_n, \theta)$. If the number of ouliers is known, then this number of observations should be trimmed. Since usually the number of ouliers is unknown we could choose the trimming parameter h such that the protection against outliers is as good as possible. A well known measure of protection against outliers is the breakdown point which here will be studied.

In the case of normal distribution with known variance, the trimmed likelihood estimators coincide with the least trimmed squares (LTS) estimators of Rousseeuw (1984,

1985) and Rousseeuw and Leroy (1987). Breakdown points of LTS estimators for linear regression were derived in Rousseeuw (1984, 1985), Rousseeuw and Leroy (1987), Vandev (1993), Vandev and Neykov (1993), Coakley and Mili (1993), Hössjer (1994), Müller (1995, 1997), Mili and Coakley (1996) and Hössjer (1994) showed also consistency and asymptotic normality. Trimmed likelihood estimators for normal distribution with unknown variance were regarded in Bednarski and Clarke (1993) who derived their asymptotic properties like Fisher consistency, asymptotic normality and compact differentiability.

Up to now, not much is known about trimmed likelihood estimators for distributions different from the normal distribution. There are approaches on robust and in particular high breakdown point estimators for logistic regression and other nonlinear models given by Stefanski, Carroll, and Ruppert (1986), Copas (1988), Künsch, Stefanski and Carroll (1989), Stromberg and Ruppert (1992), Carroll and Pederson (1993), Wang and Carroll (1993, 1995), Christmann (1994), Sakata and White (1995), Hubert (1997), Christmann and Rousseeuw (1999). But these approaches do not concern trimmed likelihood estimators.

Only Vandev and Neykov (1998) derived breakdown points of trimmed likelihood estimators for logistic regression and exponential linear models with unknown dispersion. Their approach bases on the concept of d-fullness developed by Vandev (1993). However, they could only derive breakdown points under the restriction that the explanatory variables $x_1, \ldots, x_N$ of the logistic regression and the exponential linear model are in general position. This restriction was also used in the approaches of Rousseeuw (1984, 1985) and Rousseeuw and Leroy (1987) concerning LTS estimators. Müller (1995, 1997) and Mili and Coakley (1996) dropped this restriction and showed that then the breakdown point of LTS estimators is determined by $\mathcal{N}(X)$ defined as

$$\mathcal{N}(X) := \max_{0 \neq \beta \in I\!R^p} \operatorname{card} \left\{ n \in \{1, \ldots, N\};\ x_n^\top \beta = 0 \right\},$$

where $X := (x_1, \ldots, x_N)^\top \in I\!R^{N \times p}$. Hence $\mathcal{N}(X)$ provides the maximum number of explanatory variables lying in a subspace. If the explanatory variables are in general position then $\mathcal{N}(X) = p - 1$ which is the minimum value for $\mathcal{N}(X)$. In other cases $\mathcal{N}(X)$

is much higher. These other cases appear mainly when the explanatory variables are not random but fixed and this happens in particular if they are given by an experimenter in a designed experiment.

In this chapter we are showing that the quantity $\mathcal{N}(X)$ determines the breakdown point not only of LTS estimators in linear models but also of any trimmed likelihood estimator and related estimators as the S estimators in generalized linear models. In particular, we will show how the fullness parameter of Vandev (1993) is connected with $\mathcal{N}(X)$. This leads to a general approach about lower bounds for breakdown points in generalized linear models with and without dispersion parameters. Although the approach is a generalization and combination of that in Müller (1995, 1997), Mili and Coakley (1996) and Vandev and Neykov (1998) it is much simpler and the proofs are shorter. In particular, restrictions of the sample size and the trimming factor h which are used in Vandev and Neykov (1998) can be dropped.

In Section 1.2, the most general result concerning a lower bound for breakdown points of trimmed likelihood estimators in general models is presented. The first application of the general result is given in Section 1.3 for generalized linear models without dispersion parameter. Here it is shown how the fullness parameter d of Vandev (1993) is connected with the quantity $\mathcal{N}(X)$. From these results, lower bounds for the breakdown points in linear models, in logistic regression models and in log-linear models appear as simple examples. Since also upper bounds for the breakdown points are derived for the logistic regression and the log-linear models by special considerations, the logistic regression model and the log-linear model are treated separately in Section 1.4 and Section 1.5, respectively. The second application of the general result of Section 1.2 concerns generalized linear models with dispersion parameter and is presented in Section 1.6. Here we also derive breakdown points of S estimators by completing a proof of Rousseeuw and Yohai (1984) and Rousseeuw and Leroy (1987).

1.2 Breakdown points of trimmed likelihood estimators in general models

Let Θ be an topological space. For example, $\Theta = [0,1]$ for binomial experiments, $\Theta = [0,\infty)$ for variance estimation, $\Theta = I\!R^p$ for regression experiments, or $\Theta = I\!R \times I\!R^+$. In such general setting breakdown points of an estimator for $\theta \in \Theta$ are defined as follows, where $\text{int}(\Theta)$ denotes the interior of Θ. Compare e.g. Hampel et al. (1986), p. 97.

Definition 1.1 *The breakdown point of an estimator $\hat{\theta} : \mathcal{Y}^N \to \Theta$ at $y \in \mathcal{Y}^N$ is defined as*

$$\epsilon^*(\hat{\theta}, y) := \frac{1}{N} \min \Big\{ M;\ \textit{there exists no compact set } \Theta_0 \subset \textit{int}(\Theta) \textit{ with } \{\hat{\theta}(\overline{y}); \overline{y} \in \mathcal{Y}_M(y)\} \subset \Theta_0 \Big\},$$

where $\mathcal{Y}_M(y) := \big\{ \overline{y} \in \mathcal{Y}^N;\ \textit{card}\{n;\ y_n \neq \overline{y}_n\} \leq M \big\}$ is the set of contaminated samples corrupted by at most M observations.

In the case $\Theta = I\!R^p$, we have that $N \cdot \epsilon^*(\hat{\theta}, y)$ is the smallest number M of contaminated observations so that $\{\hat{\theta}(\overline{y});\ \overline{y} \in \mathcal{Y}_M(y)\}$ is unbounded.

In some situations the breakdown point satisfies $\epsilon^*(\hat{\theta}, y) = 0$, which can only happen if $\hat{\theta}(y)$ is not uniquely defined and given by values not lying in a compact subset of $\text{int}(\Theta)$. Then the estimator is not identifiable. Typically the values of a nonidentifiable estimator are lying in a whole subspace of Θ, and this happens in particular in complex models, for example, in models where the observations depend on several explanatory variables. Maximum likelihood estimators are not identifiable if the maximum of $\prod_{n=1}^N f_n(y_n, \theta)$, or equivalently the minimum of $\sum_{n=1}^N l_n(y, \theta)$, is attained at several θ. Then, setting $\gamma(\theta) = \sum_{n=1}^N l_n(y, \theta)$, a breakdown point equal to zero means that the set $\{\theta \in \Theta;\ \gamma(\theta) \leq C\}$ is not contained in a compact subset of $\text{int}(\Theta)$ for all $C \geq \min_\theta \gamma(\theta)$. Since we want to extend these considerations to trimmed likelihood estimators, we make the following definition.

Definition 1.2 *A function $\gamma : \Theta \to I\!R$ is called sub-compact if the set $\{\theta \in \Theta;\ \gamma(\theta) \leq C\}$ is contained in a compact set $\Theta_C \subset \textit{int}(\Theta)$ for all $C \in I\!R$.*

This definition of sub-compactness is similar to the definition of Vandev and Neykov (1998) but not the same since Vandev and Neykov demanded that $\{\theta;\ \gamma(\theta) \leq C\}$ itself is a compact set. But this is for our purposes too restrictive. With Definition 1.2 we have the following conclusion.

Lemma 1.1 *The breakdown point of a maximum likelihood estimator $\hat{\theta}$ at y satisfies $\epsilon^*(\hat{\theta}, y) > 0$ if γ given by $\gamma(\theta) = \sum_{n=1}^{N} l_n(y, \theta)$ is sub-compact.*

Since $\max_{n=1,\ldots,N} l_n(y,\theta) \leq \sum_{n=1}^{N} l_n(y,\theta) \leq N \max_{n=1,\ldots,N} l_n(y,\theta)$ Lemma 1.1 holds also if γ is given by $\gamma(\theta) = \max_{n=1,\ldots,N} l_n(y,\theta)$. To study the breakdown behavior of maximum likelihood estimators at subsamples, we use the definition of d-fullness of Vandev and Neykov (1998).

Definition 1.3 *A finite set $\Gamma = \{\gamma_n : \Theta \to I\!R;\ n = 1, \ldots, N\}$ of functions is called d-full if for every $\{n_1, \ldots, n_d\} \subset \{1, \ldots, N\}$ the function γ given by $\gamma(\theta) := \max\{\gamma_{n_k}(\theta);\ k = 1, \ldots, d\}$ is sub-compact.*

The d-fullness of the log-likelihood functions $\{l_n(y,\cdot);\ n = 1, \ldots, N\}$ provides positive breakdown points of the maximum likelihood estimators at any subsample with d observations. Moreover, as Vandev and Neykov (1998) showed, d-fullness is also related to the breakdown points of TL and WTL estimators. Here we extend this result and show that the proof of this extension is even much shorter and simpler than that of Vandev and Neykov. For this proof, we use in particular the fact that the definition of d-fullnes is based on the maximum of $\gamma_{n_k}(\theta)$ instead of the sum. The extension concerns any estimator S of the form

$$S(y) := \arg\min_{\theta \in \Theta} s(y, \theta)$$

with $s : \mathcal{Y}^N \times \Theta \to I\!R$, where $s(y,\theta)$ can be estimated by $l_{(h)}(y,\theta)$ such that there exists $\alpha,\ \beta \in I\!R$ with $\alpha \neq 0$ and $h \in \{1, \ldots, N\}$ such that

$$\alpha\, l_{(h)}(\tilde{y}, \theta) \leq s(\tilde{y}, \theta) \leq \beta\, l_{(h)}(\tilde{y}, \theta) \tag{1.1}$$

for all $\tilde{y} \in \mathcal{Y}^N$ and $\theta \in \Theta$. It is obvious that $s(y, \theta)$ of the TL and WTL estimators satisfies condition (1.1). But there are also other estimators which fall under condition (1.1). One of these estimators, the S-estimator, is treated in Section 1.6.

Theorem 1.1 *If the estimator S satisfies condition (1.1) and $\{l_n(y, \cdot);\ n = 1, \ldots, N\}$ is d-full, then*

$$\epsilon^*(S, y) \geq \frac{1}{N} \min\{N - h + 1, h - d + 1\}.$$

The proof of Theorem 1.1 bases on the following lemma.

Lemma 1.2 *If $\{l_n(y, \cdot);\ n = 1, \ldots, N\}$ is d-full, $M \leq N - h$, and $M \leq h - d$, then $l_{(d)}(y, \theta) \leq l_{(h)}(\overline{y}, \theta) \leq l_{(N)}(y, \theta)$ for all $\overline{y} \in \mathcal{Y}_M(y)$ and $\theta \in \Theta$.*

Proof of Lemma 1.2. Regard $n_1, \ldots, n_h$ with $l_{(k)}(\overline{y}, \theta) = l_{n_k}(\overline{y}, \theta)$ for $k = 1, \ldots, h$. Since $h \geq M + d$ we have $1 \leq k(1) < \ldots < k(d) \leq h$ with $l_{n_{k(i)}}(\overline{y}, \theta) = l_{n_{k(i)}}(y, \theta)$. Then we obtain

$$l_{(h)}(\overline{y}, \theta) = l_{n_h}(\overline{y}, \theta) \geq l_{n_{k(d)}}(\overline{y}, \theta) \geq l_{n_{k(i)}}(\overline{y}, \theta) = l_{n_{k(i)}}(y, \theta)$$

for all $i = 1, \ldots, d$. This implies $l_{(h)}(\overline{y}, \theta) \geq l_{(d)}(y, \theta)$. The other inequality follows similarly. □

Proof of Theorem 1.1. Let $M = \min\{N-h, h-d\}$. Lemma 1.2 together with assumption (1.1) provide that

$$\alpha\, l_{(d)}(y, \theta) \leq s(\overline{y}, \theta) \leq \beta\, l_{(N)}(y, \theta)$$

for all $\overline{y} \in \mathcal{Y}_M(y)$ and $\theta \in \Theta$. This means

$$\alpha\, l_{(d)}(y, S(\overline{y})) \leq s(\overline{y}, S(\overline{y})) = \min_\theta s(\overline{y}, \theta) \leq \beta \min_\theta l_{(N)}(y, \theta)$$

for all $\overline{y} \in \mathcal{Y}_M(y)$. Setting $C_0 := \frac{\beta}{\alpha} \min_\theta l_{(N)}(y, \theta)$ we have $\{S(\overline{y});\ \overline{y} \in \mathcal{Y}_M(y)\} \subset \{\theta \in \Theta;\ l_{(d)}(y, \theta) \leq C_0\}$ so that we have only to show that γ given by

$$\begin{aligned} \gamma(\theta) &:= l_{(d)}(y, \theta) = \max\{l_{(1)}(y, \theta), \ldots, l_{(d)}(y, \theta)\} \\ &= \max\{l_{n_1(\theta)}(y, \theta), \ldots, l_{n_d(\theta)}(y, \theta)\} \end{aligned}$$

is sub-compact. Assume that this is not the case. Then there exists $C \in I\!R$ such that $\{\theta;\ \gamma(\theta) \leq C\}$ is not contained in a compact set. Hence, there exists a sequence $(\theta_m)_{m \in I\!N} \in \{\theta;\ \gamma(\theta) \leq C\}$ such that every subsequence of $(\theta_m)_{m \in I\!N}$ is not converging. Because of $\{n_1(\theta_m), \ldots, n_d(\theta_m)\} \subset \{1, \ldots, N\}$ we have a subsequence $(\theta_{m(k)})_{k \in I\!N}$ and $n_1, \ldots, n_d$ such that $\{n_1(\theta_{m(k)}), \ldots, n_d(\theta_{m(k)})\} = \{n_1, \ldots, n_d\}$ for all $k \in I\!N$. This implies $\gamma(\theta_{m(k)}) = \max\{l_{n_1}(y, \theta_{m(k)}), \ldots, l_{n_d}(y, \theta_{m(k)})\} \leq C$ for all $k \in I\!N$. However, $\max\{l_{n_1}(y, \cdot), \ldots, l_{n_d}(y, \cdot)\}$ is sub-compact since $\{l_1(y, \cdot), \ldots, l_N(y, \cdot)\}$ is d-full. This provides that $(\theta_{m(k)})_{k \in I\!N}$ contains a convergent subsequence which is a contradiction. Hence γ is sub-compact. $\square$

Note that Theorem 1 of Vandev and Neykov (1998) provides a lower bound of the breakdown point of weighted trimmed likelihood estimators which is $(N - h + 1)/N$. However this lower bound is derived under the additional assumptions of $N \geq 3\,d$ and $(N + d)/2 \leq h \leq N - d$. Since $(N + d)/2 \leq h$ implies $h - d \geq (N - d)/2 \geq N - h$ the lower bound of Vandev and Neykov is not better than that of Theorem 1.1. Hence Theorem 1.1 is not only an extension of Theorem 1 of Vandev and Neykov to other estimators but also provides the lower bound without additional assumptions on N and h.

Note also that the lower bound of Theorem 1.1 is maximized if the trimming factor h satisfies $\lfloor \frac{N+d}{2} \rfloor \leq h \leq \lfloor \frac{N+d+1}{2} \rfloor$ where $\lfloor z \rfloor := \max\{n \in I\!N;\ n \leq z\}$. A simple consequence of this fact is the following result concerning trimmed likelihood estimators.

Theorem 1.2 *Assume that* $\{l_n(y, \cdot);\ n = 1, \ldots, N\}$ *is* d*-full and* $\lfloor \frac{N+d}{2} \rfloor \leq h \leq \lfloor \frac{N+d+1}{2} \rfloor$. *Then the breakdown point of any weighted trimmed likelihood estimator* WTL_h *satisfies*

$$\epsilon^*(WTL_h, y) \geq \frac{1}{N} \left\lfloor \frac{N - d + 2}{2} \right\rfloor .$$

In the next sections, we derive the fullness parameter d and thus the lower bound for the breakdown point for special models.

1.3 Application on generalize linear models without dispersion parameter

Assume that the distribution of the observations Y_n have densities $f(y_n, x_n, \beta)$ given by a linear exponential family, that is

$$f(y_n, x_n, \beta) = \exp\{T(y_n)^\top g(x_n^\top \beta) + c(x_n^\top \beta) + b(y_n)\},$$

where $T : \mathcal{Y} \to I\!R^r$, $g : I\!R \to I\!R^r$, $c : I\!R \to I\!R$, and $b : \mathcal{Y} \to I\!R$ are known functions with $\mathcal{Y} \subset I\!R^q$, $x_n \in \mathcal{X} \subset I\!R^p$, $n = 1, \ldots, N$, are known explanatory variables and $\beta \in I\!R^p$ is unknown. Then the log-likelihood functions are given by

$$l_n(y, X, \beta) = -T(y_n)^\top g(x_n^\top \beta) - c(x_n^\top \beta) - b(y_n),$$

where $X = (x_1, \ldots, x_n)^\top$. For estimating β we can use again trimmed or weighted trimmed likelihood estimators to protect ourselves against the influence of outliers not coming from the model. The breakdown point of these estimators is determined according to Theorem 1.2 by the fullness parameter of $\{l_1(y, X, \cdot), \ldots, l_N(y, X, \cdot)\}$. We will now show that, under fixed X, this fullness parameter depends on the quantity $\mathcal{N}(X)$ of Müller (1995) defined in the introduction. Intuitively it is clear that we only can expect identifiability of β with d observations and thus d-fullness if d explanatory variables always span the whole $I\!R^p$. This is just satisfied by $d \geq \mathcal{N}(X) + 1$ by definition of $\mathcal{N}(X)$. We even have $d = \mathcal{N}(X) + 1$ for a lot of generalized linear models, however sometimes with some restrictions on the sample space. The formal proof of the relation between the fullness parameter and the quantity $\mathcal{N}(X)$ is based on the following lemma.

Lemma 1.3 *Let $X \in I\!R^{N \times p}$ and $I \subset \{1, \ldots, N\}$ with cardinality $\mathcal{N}(X) + 1$. Then the set $\{\beta \in I\!R^p;\ \max_{i \in I} |x_i^\top \beta| \leq D\}$ is bounded for all $D \in I\!R$.*

Proof of Lemma 1.3. We have the following inclusion

$$\begin{aligned}
&\{\beta \in I\!R^p;\ \max_{i\in I} |x_i^\top \beta| \le D\} \\
&\quad \subset \quad \left\{\beta \in I\!R^p;\ \frac{1}{\mathcal{N}(X)+1} \sum_{i\in I} (x_i^\top \beta)^2 \le D^2\right\} \\
&\quad = \quad \left\{\beta \in I\!R^p;\ \frac{1}{\mathcal{N}(X)+1}\, \beta^\top \sum_{i\in I} x_i\, x_i^\top \beta \ \le D^2\right\}.
\end{aligned}$$

Because I is of cardinality $\mathcal{N}(X)+1$ the definition of $\mathcal{N}(X)$ implies that the matrix $\sum_{i\in I} x_i\, x_i^\top$ is of full rank. Hence the set

$$\left\{\beta \in I\!R^p;\ \frac{1}{\mathcal{N}(X)+1}\, \beta^\top \sum_{i\in I} x_i\, x_i^\top \beta \ \le D^2\right\}$$

is bounded. □

Theorem 1.3 *If the function γ_z given by $\gamma_z(\theta) = -T(z)^\top g(\theta) - c(\theta) - b(z)$ is sub-compact for all $z \in \mathcal{Y}$ then the family $\{l_n(y, X, \cdot);\ n = 1, \ldots, N\}$ is $\mathcal{N}(X)+1$-full for all $y \in \mathcal{Y}^N$ and all $X \in \mathcal{X}^N$.*

Proof of Theorem 1.3. Regard any $C \in I\!R$ and any $I \subset \{1, \ldots, N\}$ with cardinality $\mathcal{N}(X)+1$. Because of the sub-compactness of γ_z there exists D_i, $i \in I$, such that

$$\begin{aligned}
&\left\{\beta \in I\!R^p;\ \max_{i\in I} l_i(y, X, \beta) \le C\right\} \\
&\quad = \quad \bigcap_{i\in I} \{\beta \in I\!R^p;\ l_i(y, X, \beta) \le C\} = \bigcap_{i\in I} \left\{\beta \in I\!R^p;\ \gamma_{y_i}(x_i^\top \beta) \le C\right\} \\
&\quad \subset \quad \bigcap_{i\in I} \left\{\beta \in I\!R^p;\ |x_i^\top \beta| \le D_i\right\} \subset \left\{\beta \in I\!R^p;\ \max_{i\in I} |x_i^\top \beta| \le \max_{i\in I} D_i\right\}.
\end{aligned}$$

The last set is contained in compact set because of Lemma 1.3. □

Example 1.1 (Linear models)

In a linear model where the errors have normal distribution with known variance the log-likelihood function is

$$\begin{aligned} & l_n(y, X, \beta) \\ &= \frac{1}{2}\frac{(y_n - x_n^\top \beta)^2}{\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \\ &= -y_n \frac{1}{\sigma^2} x_n^\top \beta + \frac{(x_n^\top \beta)^2}{2\sigma^2} + \frac{(y_n)^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2). \end{aligned}$$

Since $\gamma_z(\theta) = -z\,\frac{1}{\sigma^2}\,\theta + \frac{\theta^2}{2\sigma^2} + \frac{z^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$ is sub-compact the condition of Theorem 1.3 is satisfied so that the set $\{l_n(y, X, \cdot);\ n = 1, \ldots, N\}$ is $\mathcal{N}(X)+1$-full. Hence Theorem 1.2 provides that any weighted trimmed likelihood estimator with $\left\lfloor \frac{N+\mathcal{N}(X)+1}{2} \right\rfloor \leq h \leq \left\lfloor \frac{N+\mathcal{N}(X)+2}{2} \right\rfloor$ has a breakdown point not less than $\frac{1}{N}\left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$. This result was already obtained by Müller (1995, 1997) since the trimmed likelihood estimators coincide with the least trimmed squares estimators in this case. In Müller (1995, 1997) it was also shown that $\frac{1}{N}\left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$ is an upper bound for regression equivariant estimators as well. Since also weighted trimmed likelihood estimators are regression equivariant we even have that the breakdown point of the WTL estimators is exactly $\frac{1}{N}\left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$.

Note that Vandev and Neykov (1998) also treated this linear model and the least trimmed squares estimators. But they made the assumption that $x_1, \ldots, x_N$ are in general position, that is $\mathcal{N}(X) = p - 1$. Under this assumption, they showed only that the set $\{l_n(y, X, \beta);\ n = 1, \ldots, N\}$ is $p + 1$-full although it is p-full.

Theorem 1.3 together with Theorem 1.1 provide only a lower bound for the breakdown points. Since regression equivariance makes only sense for linear models but not for other generalized linear models an upper bound cannot be derived by regression equivariance as it was shown for linear models by Müller (1995, 1997). In other generalized linear models it also can happen that even the maximum likelihood estimators never breaks down. However, as soon as the ML estimator has a breakdown point less than or equal to $\frac{1}{N}$ it is obvious that the following upper bound for the breakdown point holds.

Lemma 1.4 *If the breakdown point of the maximum likelihood estimator satisfies*

$$\epsilon^*(ML, y, X) \leq \frac{1}{N}$$

for all $y \in \mathcal{Y}^N$ *and* $X \in \mathcal{X}^N$*, then we have for any weighted trimmed likelihood estimator* WTL_h

$$\epsilon^*(WTL_h, y, X) \leq \frac{1}{N}(N - h + 1).$$

Proof of Lemma 1.4 . From the definition of the weighted trimmed likelihood estimator it follows that it is a maximum likelihood estimator over a subsample of h observations out of N. So the smallest fraction of the data which has to be contaminated in order to render the estimator completely meaningless is $(N - h + 1)/N$. Therefore its breakdown point is not greater than $(N - h + 1)/N$. □

The assumption $\epsilon^*(ML, y, X) \leq \frac{1}{N}$ of Lemma 1.4 must be shown for each generalized linear model separately. Under this assumption we see that, as for linear models, the efficiency and the breakdown point are contrary properties. For large h, the efficiency is high while the breakdown point is small.

The upper bound given by Lemma 1.4 is only useful if h is large enough. In particular for $h \geq \left\lfloor \frac{N+\mathcal{N}(X)+1}{2} \right\rfloor$ we obtain $\epsilon^*(WTL_h, y, X) \leq \frac{1}{N} \left\lfloor \frac{N-\mathcal{N}(X)+2}{2} \right\rfloor$. This is very similar to the upper bound for regression equivariant estimators in linear models. However, for $h < \left\lfloor \frac{N+\mathcal{N}(X)+1}{2} \right\rfloor$ no reasonable upper bound is provided by Lemma 1.4. As for linear models, the breakdown point should be small if h is too small. However, this cannot be shown in full generality so that special considerations for each generalized linear model are necessary. Therefore we treat the logistic regression model and the log-linear model as examples in the following two sections .

1.4 Logistic regression

Let t_n the total number of observations and $s_n \in \{0, \ldots, t_n\}$ the number of successes under condition x_n. In a logistic regression model, it is assumed that the number of successes S_n has a binomial distribution with parameters t_n and π_n where $\pi_n = \exp(x_n^\top \beta)/(1+\exp(x_n^\top \beta))$ is the probability of success explained by the explanatory variable x_n. Setting $y = (s, t)$

with $t = (t_1, \ldots, t_N)^\top$ and $s = (s_1, \ldots, s_N)^\top$, the log-likelihood function is

$$\begin{aligned} l_n(y, X, \beta) &= l_n(s, t, X, \beta) \\ &= -s_n\, x_n^\top \beta + t_n\, \log(1 + \exp(x_n^\top \beta)) - \log(\tbinom{t_n}{s_n}) \\ &= \gamma_{s_n, t_n}(x_n^\top \beta). \end{aligned} \tag{1.2}$$

The function $\gamma_{u,v}$ given by $\gamma_{u,v}(\theta) = -u\theta + v\, \log(1 + \exp(\theta)) - \log(\binom{v}{u})$ is sub-compact as soon as $0 < u < v$ so that according to Theorem 1.3 the set $\{l_n(y, X, \cdot);\ n = 1, \ldots, N\}$ is $\mathcal{N}(X)$+1-full for all $y = (s, t)$ satisfying $0 < s_n < t_n$ for $n = 1, \ldots, N$. Hence, Theorem 1.1 provides $\frac{1}{N}\, \min\{N - h + 1, h - \mathcal{N}(X)\}$ as a lower bound for the breakdown point of any weighted trimmed likelihood estimator WTL_h. This lower bound is also an upper bound as the following theorem shows. For that let be $\mathcal{Y}^*$ the set of all $y = (s, t)$ with $0 < s_n < t_n$ for $n = 1, \ldots, N$. Here we exclude the case $s_n = 0$ or $s = t_n$ to avoid problems described in Christmann and Rousseeuw (1999) concerning missing overlap. A combination of our results and those of Christmann and Rousseeuw would provide a result for $0 \leq s_n \leq t_n$ but this is beyond the scope of this chapter.

Theorem 1.4 *The breakdown point of any weighted trimmed likelihood estimator WTL_h for logistic regression satisfies*

$$\min_{y \in \mathcal{Y}^*} \epsilon^*(WTL_h, y, X) = \frac{1}{N}\, \min\{N - h + 1, h - \mathcal{N}(X)\}.$$

Proof of Theorem 1.4. After the remarks above we have only to show that $\frac{1}{N}\, \min\{N - h + 1, h - \mathcal{N}(X)\}$ is an upper bound. If $h > \left\lfloor \frac{N + \mathcal{N}(X) + 1}{2} \right\rfloor$ the upper bound follows from Lemma 1.4 if we can show $\epsilon^*(ML, y, X) \leq \frac{1}{N}$ for all y for the maximum likelihood estimator. Hence regard any $y = (s, t)$. If $ML(y, X)$ is not contained in a compact subset of $\mathbb{R}^p$ then $\epsilon^*(ML, y, X) = 0$. Otherwise, since the second derivative of $\sum_{n=1}^N l_n(y, X, \beta)$ with respect to β is a positive semidefinite matrix, it is sufficient and necessary for $\hat{\beta} = ML(y, X)$ that $X^\top s = X^\top e(t, \hat{\beta})$ holds where

$$e(t, \hat{\beta}) := \left(t_1\, \frac{\exp(x_1^\top \hat{\beta})}{1 + \exp(x_1^\top \hat{\beta})}\ ,\ \ldots\ ,\ t_N\, \frac{\exp(x_N^\top \hat{\beta})}{1 + \exp(x_N^\top \hat{\beta})} \right)^\top .$$

Regard the sequence $y^k = (s^k, t^k) \in \mathcal{Y}_1((s,t))$ with $s_1^k = 1$ and $t_1^k = k$ for all $k \in I\!N$. Assume that $\hat{\beta}^k = ML(y^k, X)$ is bounded. Then

$$\begin{aligned} X^\top e(t^k, \hat{\beta}^k) = x_1\, t_1^k \frac{\exp(x_1^\top \hat{\beta}^k)}{1+\exp(x_1^\top \hat{\beta}^k)} &+ x_2\, t_2 \frac{\exp(x_2^\top \hat{\beta}^k)}{1+\exp(x_2^\top \hat{\beta}^k)} + \ldots \\ &+ x_N\, t_N \frac{\exp(x_N^\top \hat{\beta}^k)}{1+\exp(x_N^\top \hat{\beta}^k)} \end{aligned}$$

is unbounded while $X^\top s^k$ is bounded which is a contradiction to $X^\top s^k = x^\top e(t^k, \hat{\beta}^k)$. Hence $ML(y^k, X)$ is not bounded so that $\epsilon^*(ML, y, X) \leq \frac{1}{N}$.

Now regard the case $h \leq \left\lfloor \frac{N+\mathcal{N}(X)+1}{2} \right\rfloor$. W.l.o.g. we can assume that there exists β_0 such that $x_n^\top \beta_0 = 0$ for $n = 1, \ldots, \mathcal{N}(X)$. Then by definition of $\mathcal{N}(X)$ we have $x_n^\top \beta_0 \neq 0$ for $n = \mathcal{N}(X)+1, \ldots, N$. At least $M_0 = \left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$ of these n satisfy $x_n^\top \beta_0 < 0$ otherwise we can regard $-\beta_0$. W.l.o.g. let $x_n^\top \beta_0 < 0$ for $n = N - M_0 + 1, \ldots, N$. Setting $M = h - \mathcal{N}(X)$ we have $M \leq M_0$. Now regard the following special sample $y = (s,t)$ with $s_n = 1$ for $n = 1, \ldots, N$, $t_n = 2$ for $n = 1, \ldots, \mathcal{N}(X)$ and $t_n = u > 2$ for $n = \mathcal{N}(X)+1, \ldots, N$. As corrupted sample we use $\overline{y} = (\overline{s}, \overline{t})$ with $\overline{s}_n = 0$, $\overline{t}_n = t_n$ for $n = N - M + 1, \ldots, N$ and $\overline{s}_n = s_n$, $\overline{t}_n = t_n$ for $n = 1, \ldots, N - M$. Then $y \in \mathcal{Y}^*$ and $\overline{y} \in \mathcal{Y}_M(y)$. Moreover, we have

$$\min_\beta l_n(\overline{s}, \overline{t}, X, \beta) = \min_\beta l_n(s, t, X, \beta) = l_n(s, t, X, k\beta_0) = \log(2)$$

for $n = 1, \ldots, \mathcal{N}(X)$ and all $k \in I\!R$, and

$$\min_\beta l_n(s, t, X, \beta) \geq \min_\mu \left(-\mu + u \log(1 + \exp(\mu)) - \log(u)\right) > \log(2)$$

for $n = \mathcal{N}(X) + 1, \ldots, N$. This implies

$$\min_\beta \sum_{n=1}^{h} w_n\, l_{(n)}(\overline{s}, \overline{t}, X, \beta) \geq \sum_{n=M+1}^{h} w_n\, \log(2).$$

Since we have with the property $x_n^\top \beta_0 < 0$ for $n = N - M + 1, \ldots, N$ for k large enough

$$\begin{aligned} \sum_{n=1}^{h} l_{(n)}(\overline{s}, \overline{t}, X, k\beta_0) &= \sum_{n=M+1}^{h} w_n\, l_{n-M}(s, t, X, k\beta_0) \\ &\quad + \sum_{n=1}^{M} w_n\, t_{N-n+1}\, \log\left(1 + \exp(x_{N-n+1}^\top\, k\, \beta_0)\right) \\ &\quad \stackrel{k\to\infty}{\longrightarrow} \sum_{n=M+1}^{h} w_n\, \log(2), \end{aligned}$$

the estimator $WTL_h(\overline{y}, X)$ is not contained in a bounded subset of $\mathbb{R}^p$ so that $\epsilon^*(WTL_h, y, X) \leq \frac{1}{N}M = \frac{1}{N}(h - \mathcal{N}(X)).\square$

The proof of Theorem 1.4 shows that for deriving the upper bound for $h > \left\lfloor \frac{N+\mathcal{N}(X)+1}{2} \right\rfloor$ it is necessary to assume that also the total numbers t_n can be contaminated by outliers. Without this assumption even the maximum likelihood estimator need not to break down by one or more corrupted observations. The assumption of possibly contaminated total numbers makes sense as soon as the total numbers are given by random which is often the case (see the example below). If the total numbers are given by random then the log-likelihood function (1.2) should have additional terms. But since these terms are additive they do not influence the determination of the maximum likelihood estimator if these terms are independent of β. However they can influence the trimmed likelihood estimators by changing the order of the $l_n(y, X, \beta)$ and would make the second step of the proof of Theorem 1.4 (for $h \leq \left\lfloor \frac{N+\mathcal{N}(X)+1}{2} \right\rfloor$) more complicated though it also would work. Thus here we regarded for simplicity the simple trimmed likelihood estimators based on the log-likelihood functions given by (1.2).

Theorem 1.4 in particular shows that the maximum breakdown point for logistic regression is attained for h satisfying $\left\lfloor \frac{N+\mathcal{N}(X)+1}{2} \right\rfloor \leq h \leq \left\lfloor \frac{N+\mathcal{N}(X)+2}{2} \right\rfloor$ and equals $\frac{1}{N}\left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$. Hence we have the same maximum breakdown point value and the same optimal trimming proportion h as for linear models.

Note, that for the special case that $x_1, \ldots, x_N$ are in general position, that is $\mathcal{N}(X) = p - 1$, a lower bound for the breakdown point similar to that in Theorem 1.4 was already obtained by Vandev and Neykov (1998) under the additional restriction of $N \geq 3(p + 1)$. Thereby, they showed again that the set $\{l_n(y, X, \beta);\ n = 1, \ldots, N\}$ is only $p + 1$-full although it is p-full.

Example 1.2 (Toxicological experiment with fish eggs)

This example involves data which resulted from a toxicological experiment conducted at the University of Waterloo, Canada, and are presented in O'Hara Hines and Carter (1993, p.13). Six different concentrations of the toxicant potassium cyanate (KSCN) were applied

to 48 vials of trout fish eggs. Each vial contained between 61 and 179 eggs. The eggs in half the vials were allowed to water harden for several hours before the toxicant was applied (this is a process in which the surface of a fish eggs becomes toughened after a few hours in water). For the remaining vials, the toxicant was applied immediately after fertilization. After 19 days of the start of the experiment the number of dead eggs in each vial was counted.

Treating the number of dead eggs in each vial as the response, a logistic regression model was fitted to the data with covariates for water hardening (0 if the toxicant was applied before water hardening and 1 after), and for a linear and quadratic term in log-concentration of toxicant. The quadratic term in log-concentration is used to describe a sharp increase in mortality caused by the two highest concentrations. Thus the logistic regression model is

$$\text{logit}\left(\frac{p}{1-p}\right) = \beta_1 + \beta_2 * WH + \beta_3 * \log_{10}(Ct) + \beta_4 * \log_{10}(Ct)^2$$

The maximum likelihood estimator for $(\beta_1, \beta_2, \beta_3, \beta_4)^\top$ based on all observations is $ML(y, X) = (10.28, 0.03, -11.4, 2.50)^\top$.

O'Hara Hines and Carter (1993) pinpoint the observations 38, 39 and 26 as possible outliers. They also reported that Pregibon's influence diagnostics indicated that the observations 38 and 39 were pinpointed as potential outliers. The MLE without the observations 38 and 39 is $(15.40, 0.27, -15.53, 3.26)^\top$ and without the observations $26, 38$ and 39 is $(14.04, 0.32, -14.64, 3.11)^\top$.

Markatou et al. (1997) analyzed the same data. The observations 38 and 39 are identified as potential outliers, whilst their methods gave a weight nearly 1 to observations 26 by means of the negative exponential RAF (Residual Adjustment Function) downweight function. When the Hellinger RAF was used for the construction of the weights, observations 13, 32, 40, 43 and 44 received a weight of 0. They reported that examination of those observations revealed that observations 32 and 40 had a 0 response, while observations 43 and 44 had the lowest mortality at concentration levels 720 and 1440, respectively, at the same water-hardening level. The MLE without the observations 13, 32, 40, 43 and 44 is

Table 1.1:

	WH	Concentration (Ct)	No Eggs	No Dead		WH	Concentration	No Eggs	No Dead
1	1	90	111	8	25	0	90	130	7
2	1	90	97	10	26	0	90	179	25
3	1	90	108	10	27	0	90	126	5
4	1	90	122	9	28	0	90	129	3
5	1	180	68	4	29	0	180	114	12
6	1	180	109	6	30	0	180	149	4
7	1	180	109	11	31	0	180	121	4
8	1	180	118	6	32	0	180	105	0
9	1	360	98	6	33	0	360	102	4
10	1	360	110	5	34	0	360	145	21
11	1	360	129	9	35	0	360	61	1
12	1	360	103	17	36	0	360	118	3
13	1	720	83	2	37	0	720	99	29
14	1	720	87	3	38	0	720	109	53
15	1	720	118	16	39	0	720	99	40
16	1	720	100	9	40	0	720	70	0
17	1	1440	140	60	41	0	1440	100	14
18	1	1440	114	47	42	0	1440	127	10
19	1	1440	103	49	43	0	1440	132	8
20	1	1440	110	20	44	0	1440	113	3
21	1	2880	143	79	45	0	2880	145	113
22	1	2880	131	85	46	0	2880	103	84
23	1	2880	111	78	47	0	2880	143	105
24	1	2880	111	74	48	0	2880	102	78

$(6.49, -0.23, -8.42, 1.97)^\top$.

For satisfying the assumption $0 < s_n < t_n$ of Theorem 1.4, we dropped the observations 32 and 40 for our calculations so that only 46 observations are available. Since 24 observations satisfy WH=1, we have $\mathcal{N}(X) = 24$. Hence, according to Theorem 1.4, the maximum breakdown point is 11/46 and is attained by any weighted trimmed likelihood estimator with $h = 35$ or $h = 36$. Since an exact algorithm for calculating the trimmed likelihood estimator with weights $w_n = 1$ and $h = 36$ had run too long, we used a genetic algorithm and obtained $TL_{36}(y, X) = (7.36, -0.12, -9.29, 2.16)^\top$. The trimmed observations were 13, 14, 20, 21, 38, 39, 41, 42, 43, 44. Hence there is some coincidence with the results of Markatou et al. (1997) with respect to the estimate and the trimmed observations.

1.5 Log-linear models

The response variables Y_n of a log-linear model have a Poisson distribution with parameter $\lambda_n = \exp(x_n^\top \beta)$ so that the log-likelihood function is

$$l_n(y, X, \beta) = -y_n \, x_n^\top \beta + \exp(x_n^\top \beta) + \log(y_n!).$$

The function γ_z given by $\gamma_z(\theta) = -z\theta + \exp(\theta) + \log(z!)$ is sub-compact as soon as $z > 0$ so that according to Theorem 1.3 the set $\{l_n(y, X, \cdot);\ n = 1, \ldots, N\}$ is $\mathcal{N}(X)+1$-full for all y satisfying $y_n > 0$ for all $n = 1, \ldots, N$. Hence, Theorem 1.1 provides $\frac{1}{N} \min\{N - h + 1, h - \mathcal{N}(X)\}$ as a lower bound for the breakdown point of any weighted trimmed likelihood estimator WTL_h. This lower bound is also an upper bound as the following theorem shows. For that let be $\mathcal{Y}^*$ the set of all y with $y_n > 0$ for $n = 1, \ldots, N$.

Theorem 1.5 *The breakdown point of any weighted trimmed likelihood estimator WTL_h for a log-linear model satisfies*

$$\min_{y \in \mathcal{Y}^*} \epsilon^*(WTL_h, y, X) = \frac{1}{N} \min\{N - h + 1, h - \mathcal{N}(X)\}.$$

Proof of Theorem 1.5. The proof is similar to that for the logistic regression model. The first step is to show $\epsilon^*(ML, y, X) \leq \frac{1}{N}$ for using Lemma 1.4. Again a sufficient and

necessary condition for $\hat{\beta} = ML(y, x)$ lying in a bounded subset of $\mathbb{R}^p$ is $X^\top y = X^\top e(\hat{\beta})$ where here $e(\hat{\beta}) := (\exp(x_1^\top \hat{\beta}), \ldots, \exp(x_1^\top \hat{\beta}))^\top$. Then, as soon as $y^k \in \mathcal{Y}_1(y)$ is unbounded also the corresponding $\hat{\beta}^k = ML(y^k, X)$ cannot be bounded.

The second step is to construct a sample $y \in \mathcal{Y}^*$ and a corrupted sample $\overline{y} \in \mathcal{Y}_M(y)$ with $M = h - \mathcal{N}(X)$ such that $WTL_h(\overline{y}, X)$ is not contained in a bounded subset of $\mathbb{R}^p$. For that, let $x_1, \ldots, x_n$ and β_0 as in the proof of Theorem 1.4. Set $y_n = 1$ for $n = 1, \ldots, \mathcal{N}(X)$, $y_n = z > e^2 - \frac{1}{2}$ for $n = \mathcal{N}(X) + 1, \ldots, N$, $\overline{y}_n = y_n$ for $n = 1, \ldots, N - M$, and $\overline{y}_n = 0$ for $n = N - M + 1, \ldots, N$. Then we have

$$\min_\beta l_n(\overline{y}, X, \beta) = \min_\beta l_n(y, X, \beta) = l_n(y, X, k\beta_0) = 1$$

for $n = 1, \ldots, \mathcal{N}(X)$ and all $k \in \mathbb{R}$, and

$$\begin{aligned}
\min_\beta l_n(y, X, \beta) &\geq \min_\mu \left(-z\mu + \exp(\mu) + \log(z!)\right) \\
&= -z\,\log(z) + z + \log(z!) \\
&\geq -z\,\log(z) + z + \left(z + \frac{1}{2}\right)\log\left(z + \frac{1}{2}\right) - z - \frac{1}{2}\log\left(\frac{1}{2}\right) \\
&\geq \frac{1}{2}\log\left(z + \frac{1}{2}\right) > 1
\end{aligned}$$

for $n = \mathcal{N}(X) + 1, \ldots, N$. The rest follows as in the proof of Theorem 1.4. □

Again the maximum breakdown point for log-linear models is $1/N\,\lfloor N - \mathcal{N}(X) + 1/2 \rfloor$ and coincides with the maximum breakdown point for linear models. This maximum breakdown point is also attained by the same trimming proportion h.

1.6 Application on exponential linear models with dispersion parameter

Assume that the observations Y_n are distributed with q-th power exponential distribution, i.e., the density function is given by

$$f(y_n, x_n, \beta, \sigma) = \frac{q\,(1/2)^{(1+1/q)}}{\sigma\,\Gamma(1/2)}\,\exp\left(-\frac{1}{2}\left|\frac{y_n - x_n^\top \beta}{\sigma}\right|^q\right\},$$

where Γ is here the gamma function. Special cases of this distribution are the normal ($q = 2$), the Laplace ($q = 1$), the double exponential ($0 < q < 2$), the leptokurtic ($1 < q < 2$), the platikurtic ($q > 2$) and the rectangular distribution ($q \to \infty$). The fullness parameter of $\{l_n(y, X, \cdot);\ n = 1, \ldots, N\}$, where

$$l_n(y, X, \beta, \sigma) = \frac{1}{2} \left| \frac{y_n - x_n^\top \beta}{\sigma} \right|^q + \log(\sigma) - \log\left(\frac{q\,(1/2)^{(1+1/q)}}{\Gamma(1/2)} \right) \tag{1.3}$$

with $\beta \in I\!R^p$ and $\sigma \in I\!R^+$, was derived in Vandev and Neykov (1998) under the assumption that $x_1, \ldots, x_N$ are in general position, i.e. $\mathcal{N}(X) = p - 1$. They showed in Lemma 3 that the fullness parameter is $p + 1$. Here we show that the fullness parameter is even p and that the fullness parameter can be also determined in the case where $x_1, \ldots, x_N$ are not in general position.

Lemma 1.5 *If the log-likelihood function is given by (1.3) with $q > 0$ then the set $\{l_n(y, X, \cdot);\ n = 1, \ldots, N\}$ is $\mathcal{N}(X)+1$-full.*

Proof of Lemma 1.5. We have to show that γ given by

$$\gamma(\beta, \sigma) := \max_{i \in I} \frac{1}{2} \left| \frac{y_i - x_i^\top \beta}{\sigma} \right|^q + \log(\sigma) - K$$

with $K \in I\!R$ is sub-compact for all $I \subset \{1, \ldots, N\}$ with cardinality $\mathcal{N}(X) + 1$. We will do it with the same trick as in Vandev and Neykov (1998) but with a shorter proof. Take any $C \in I\!R$ and set $\tilde{\beta}(\sigma) := \arg\min\{\gamma(\beta, \sigma);\ \beta \in I\!R^p\}$ and $\tilde{\sigma}(\beta) := \arg\min\{\gamma(\beta, \sigma);\ \sigma \in I\!R^+\}$. Then $\tilde{\beta}(\sigma)$ is independent of σ such that $\tilde{\beta}(\sigma) =: \tilde{\beta}$. Setting

$$\gamma_1(\sigma) := \gamma(\tilde{\beta}(\sigma), \sigma) = \max_{i \in I} \frac{1}{2} \left| \frac{y_i - x_i^\top \tilde{\beta}}{\sigma} \right|^q + \log(\sigma) - K$$

we see that γ_1 is a sub-compact function. Hence, there exists a compact set $\Theta_1 \subsetneqq I\!R^+$ such that $\{\sigma;\ \gamma_1(\sigma) \le C\} \subset \Theta_1$. Moreover, we have that with $\eta(\beta) := \max_{i \in I} \left| y_i - x_i^\top \beta \right|$

$$\tilde{\sigma}(\beta) = \eta(\beta)\ \left(\frac{q}{2} \right)^{1/q}$$

so that

$$\gamma_2(\beta) := \gamma(\beta, \tilde{\sigma}(\beta)) = \frac{1}{q} + \log\left(\eta(\beta)\right) + \frac{1}{q}\ \log\left(\frac{q}{2} \right) - K.$$

Example 1.1 implies that η is sub-compact. Since the logarithm is monoton also γ_2 is sub-compact so that $\{\beta;\ \gamma_2(\beta) \leq C\} \subset \Theta_2$ for some compact set $\Theta_2 \subsetneqq I\!R^p$. Then we have

$$\begin{aligned}&\{(\beta,\sigma) \in I\!R^p \times I\!R^+;\ \gamma(\beta,\sigma) \leq C\}\\ &\quad\subset\ \{(\beta,\sigma) \in I\!R^p \times I\!R^+;\ \gamma_1(\sigma) \leq C \text{ and } \gamma_2(\beta) \leq C\} \subset \Theta_2 \times \Theta_1. \square\end{aligned}$$

Theorem 1.1 and Lemma 1.5 immediately imply that any weighted trimmed likelihood estimator WTL_h for (β,σ) has a breakdown point not less than $\frac{1}{N}\min\{N-h+1, h-\mathcal{N}(X)\}$ and that the lower bound of the breakdown point attains its maximum value of $\frac{1}{N}\left\lfloor\frac{N-\mathcal{N}(X)+1}{2}\right\rfloor$ if $\left\lfloor\frac{N+\mathcal{N}(X)+1}{2}\right\rfloor \leq h \leq \left\lfloor\frac{N+\mathcal{N}(X)+2}{2}\right\rfloor$. This maximum lower bound is also the upper bound for the breakdown point since the estimator for β is regression equivariant so that the upper bound follows from Müller (1995, 1997).

However, also other robust estimators can be used for distributions with unknown dispersion parameter. Estimators with good breakdown properties are the S-estimators. An S-estimator S_c is defined by (see Rousseeuw and Yohai 1984, Rousseeuw and Leroy 1987, p. 135)

$$S_c(y,X) := \arg\min_{\beta} s_c(y,X,\beta),$$

where $s_c(y,X,\beta)$ is given as solution of

$$b_c(y,X,\beta,s_c(y,X,\beta)) := \frac{1}{N}\sum_{n=1}^{N}\rho_c\left(\frac{|y_n - x_n^\top\beta|}{s_c(y,X,\beta)}\right) = K.$$

Usually K is chosen as the expectation $E_{\beta,\sigma}(b_c(Y,X,\beta,\sigma))$ to get consistency under the model distribution. If ρ_c is strictly increasing on $[0,c]$ and constant on $[c,\infty)$ then S-estimators have high breakdown points. This was shown in Rousseeuw and Yohai (1984) for $x_1,\ldots,x_N$ in general position and in Mili and Coakley (1996) for general $x_1,\ldots,x_N$. However they showed only the inequality

$$\alpha\, l_{(h)}(y,X,\beta) \leq s_c(y,X,\beta) \leq \beta\, l_{(h)}(y,X,\beta) \tag{1.4}$$

for all $y,\ X,\ \beta$, and c satisfying $\rho_c(c) = K\,N/(N-h+1)$, where $l_n(y,X,\beta) = |y_n - x_n^\top\beta|$. A detailed proof of the inequality (1.4) for $h = \left\lfloor\frac{N+1}{2}\right\rfloor$ was given in the book of Rousseeuw and

Leroy (1987), p. 136-139. Also in this book, they conclude from (1.4) without additional arguments that the breakdown points of $\arg\min_\beta l_{(h)}(y, X, \beta)$ and $S_c(y, X)$ coincide. But the proof of Theorem 1.1 shows that indeed additional arguments are necessary and that they base on the concept of d-fullness. Hence only now a complete proof of the breakdown points of the S-estimators is possible.

Theorem 1.6 *Any S-estimator S_c with $\rho_c(c) = K\,N/(N-h+1)$ has a breakdown point satisfying $\epsilon^*(S_c, y, X) \geq \frac{1}{N}\min\{N-h+1, h-\mathcal{N}(X)\}$. If*
$\lfloor (N+\mathcal{N}(X)+1)/2 \rfloor \leq h \leq \lfloor (N+\mathcal{N}(X)+2)/2 \rfloor$ *then*
$\epsilon^*(S_c, y, X) = \frac{1}{N}\left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$.

Proof of Theorem 1.6. It follows from Example 1.1 that the set $\{l_n(y, X, \cdot);\ n = 1, \ldots, N\}$ is $\mathcal{N}(X)+1$-full. Hence, Theorem 1.1 and inequality (1.4) provide the lower bounds. That $\frac{1}{N}\left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$ is also an upper bound follows from the result of Müller (1995, 1997) concerning regression equivariant estimators. □

Chapter 2

Breakdown Point and Computation of the Trimmed Likelihood Estimators in Generalized Linear Models

Summary. A review of the studies concerning the finite sample breakdown point (BP) of the trimmed likelihood (TL) and related estimators based on the d–fullness technique of Vandev (1993), and Vandev and Neykov (1998) is made. In particular, the BP of these estimators in the frame of the generalized linear models (GLMs) depends on the trimming proportion and the quantity $\mathcal{N}(X)$ introduced by Müller (1995). A faster iterative algorithm based on resampling techniques for derivation of the TLE is developed. Examples of real and artificial data in the context of grouped logistic and log-linear regression models are used to illustrate the properties of the TLE.

2.1 Introduction

The Weighted Trimmed Likelihood (WTL) estimators are defined by Hadi and Luceño (1997), and Vandev and Neykov (1998) as

$$\mathrm{WTL}_k(y_1,\ldots,y_n) := \arg\min_{\theta\in\Theta^p} \sum_{i=1}^{k} w_i f(y_{\nu(i)},\theta), \tag{2.1}$$

where $f(y_{\nu(i)},\theta) \le f(y_{\nu(i+1)},\theta)$, $f(y_i,\theta) = -\log\ \varphi(y_i,\theta)$, $y_i \in \mathcal{Y} \subset R^q$ for $i = 1,\ldots,n$ are iid observations with probability density $\varphi(y,\theta)$, which depends on an unknown parameter $\theta \in \Theta^p \subset R^p$, $\nu = (\nu(1),\ldots,\nu(n))$ is the corresponding permutation of the indices, which depends on θ, k is the trimming parameter and $w_i \ge 0$ are known weights such that $w_k > 0$.

The WTL estimators reduce to TLE if $w_i = 1$ for $i = 1,\ldots,k$. In the case of normal regression and appropriate choice of the weights, the WTLE reduce to the LMS and LTS estimators of Rousseeuw (1984) and Rousseeuw and Leroy (1987). Similarly, the WTLE coincide with the MVE and MCD estimators of the multivariate location and scatter considered by Rousseeuw and Leroy (1987) in the multivariate normal case, see Vandev and Neykov (1993). The Fisher consistency, asymptotic normality and compact differentiability of the TLE for normal distributions with unknown variance are derived by Bednarski and Clarke (1993).

The BP (i.e. the smallest fraction of contamination that can cause the estimator to take arbitrary large values) properties of the WTLE were studied by Vandev and Neykov (1998) using the d–fullness technique developed by Vandev (1993). It was proved that the BP of the WTLE is not less than $(n-k)/n$ if the set $F = \{f(y_i,\theta), i = 1,\ldots,n\}$ is d-full, $n \ge 3d$ and $(n+d)/2 \le k \le n-d$. We remind that, according to Vandev (1993), a finite set F of n functions is called d-full if for each subset of cardinality d of F, the supremum of this subset is a subcompact function. A real valued function $g(\theta)$ is called subcompact if the sets $L_{g(\theta)}(C) = \{\theta : g(\theta) \le C\}$ are compact for any constant C.

Vandev and Neykov (1993), and Vandev and Marincheva (1996) determined the value of d for the multivariate normal and general elliptical family of distributions, respectively. Vandev and Neykov (1998) did the same about some linear and logistic regression models

under the restriction that the observations are in general position. Similarly, the fullness parameters for the Lognormal, Poisson, Gamma, Geometric and Logarithmic series distributions were derived by Atanasov (1998), and the BPs of the WTLE of the corresponding GLMs were characterized (see, Atanasov and Neykov (2001)).

There are approaches on robust and in particular high BP estimators for logistic regression and other nonlinear models given by Copas (1988), Carroll and Pederson (1993), Christmann (1994), Christmann and Rousseeuw (2001), Hubert (1997), Künsch et al. (1989), Markatou et al. (1997), Stromberg (1992), to name a few, but these approaches do not concern TLE.

The BP of the LMS, LTS and related regression estimators were derived by Rousseeuw (1984), Rousseeuw and Leroy (1987), and Hössjer (1994), assuming that the observations are in general position. Müller (1995), and Mili and Coakley (1996) omitted this restriction and showed that then the BP of these estimators is determined by

$$\mathcal{N}(X) := \max_{0 \neq \beta \in R^p} \operatorname{card} \left\{ i \in \{1, \ldots, n\};\ x_i^\top \beta = 0 \right\},$$

where $X := (x_i^\top)$ is the data matrix of the explanatory variables $x_i \in R^p$. If x_i are in general position then $\mathcal{N}(X) = p - 1$ whereas in other cases, e.g., ANOVA models or designed experiments $\mathcal{N}(X)$ is much higher.

Müller and Neykov (2003) relaxed the compactness condition in the above definition assuming only that the set $L_{g(\theta)}(C)$ is contained in a compact set. However, the meaning of the term subcompact function is retained since if the function $g(\theta)$ is continuous or has at most countable many discontinuities then $L_{g(\theta)}(C)$ is a compact set. The following theorem characterizes the BPs of any estimator S defined by $S(y) := \arg\min_{\theta \in \Theta} s(y, \theta)$, where $s(y, \theta)$ can be estimated by $f(y_{\nu(k)}, \theta)$ and satisfies the conditions $\alpha\, f(y_{\nu(k)}, \theta) \leq s(y, \theta) \leq \beta\, f(y_{\nu(k)}, \theta)$ for some constants $\alpha \neq 0$ and β, and therefore of the WTLE in particular.

Theorem 2.1 *If* $\{f(y_i, \theta);\ i = 1, \ldots, n\}$ *is d-full, then the BP of the estimator* S *is not less than* $\frac{1}{n} \min\{n - k + 1, k - d + 1\}$.

This theorem is an extension of Theorem 1 of Vandev and Neykov (1998) and provides the lower bound of the BP without additional assumptions on k and n (see, Müller and Neykov (2003)).

Thus, if one wants to study the BP of the WTL and related S estimators for a particular distribution, one has to find the fullness parameter d for the corresponding set of log likelihoods and then the BP can be exemplified by the range of values of k by Theorem 2.1.

An application of this technique is made, by Müller and Neykov (2003), for the general linear exponential families of distributions (known dispersion parameter) of y_i depending on unknown vector parameter $\beta \in R^p$ and known $x_i \in R^p$ for $i = 1, \ldots, n$. The log likelihoods of these families are $f(y_i, x_i, \beta) = -T(y_i)^\top g(x_i^\top \beta) - c(x_i^\top \beta) - h(y_i)$ for suitably defined vectors and functions. The following theorem holds.

Theorem 2.2 *The set* $\{f(y_i, x_i, \beta);\ i = 1, \ldots, n\}$ *is* $\mathcal{N}(X)+1$*–full if the function* $\gamma_z(\theta) = -T(z)^\top g(\theta) - c(\theta) - h(z)$ *is subcompact in* θ *for all* $z \in \mathcal{Y}$ *and arbitrary* $x_i \in R^p$ *.*

For the particular cases of normal, logistic and log-linear regression models Müller and Neykov (2003) show that the corresponding $\gamma_z(\theta)$ are subcompact. Therefore, according to Theorem 2.1 and some additional arguments it is shown that the BP of the WTL estimators is $\frac{1}{n}\min\{n - k + 1, k - \mathcal{N}(X)\}$. If k satisfies $\lfloor (n + \mathcal{N}(X) + 1)/2 \rfloor \leq k \leq \lfloor (n + \mathcal{N}(X) + 2)/2 \rfloor$ this BP is maximized and equal to $\frac{1}{n}\lfloor (n - \mathcal{N}(X) + 1)/2 \rfloor$, where $\lfloor r \rfloor := \max\{n \in N; n \leq r\}$. As a consequence, the results of Mili and Coakley (1996), Müller (1997), Vandev and Neykov (1998), and Atanasov (1998) for these models are derived.

In this way, a unifying theory for the BP of the WTL and related estimators is developed.

2.2 The FAST-TLE Algorithm

From the definition of the WTLE it follows that its minima are achieved over a subsample of size k. The objective function (2.1) is continuous, but non differentiable and possesses many local minima. Therefore one need nonsmooth and/or combinatorial optimization in general. In the univariate case Hadi and Luceño (1997) developed several algorithms for TL estimation.

Neykov and Neytchev (1990) considered an iterative approximate algorithm for finding the TLE which is based on the resampling technique proposed by Rousseeuw and Leroy (1987). Many subsets of k different observations out of n are drawn at random and the MLE is calculated for any one. The estimate with the lowest TL objective function (2.1) is retained. There is no guarantee that the achieved estimate will be the global minimizer of (2.1) but one can hope that it would be a close approximation to it.

In this chapter we offer a more efficient TLE algorithm called the FAST-TLE as it reduces to the FAST-LTS algorithm developed by Rousseeuw and Van Driessen (1999a) in the normal linear regression case. The corner stone of this algorithm is an analog of the so called *C–step* procedure proposed by these authors. We shall follow closely the terminology and exposition of their paper in order to present the algorithm in a more readable form to those who are acquainted with it.

So as to make sure that there always exists a solution to the optimization problem (2.1), we assume that the set F is d–full and $k \geq d$ Neykov (1995). Then the idea behind the FAST-TLE algorithm can be described as follows.

Given $H^{old} = \{y_{j_1}, \ldots, y_{j_k}\} \subset \{y_1, \ldots, y_n\}$ then:

- take $\widehat{\theta}^{old}$ to be either arbitrary or compute $\widehat{\theta}^{old} := MLE$ based on H^{old};
- define $Q^{old} := \sum_{i=1}^{k} f(y_{j_i}, \widehat{\theta}^{old})$;
- sort $f(y_i, \widehat{\theta}^{old})$ for $i = 1, \ldots, n$ in ascending order, $f(y_{\nu(i)}, \widehat{\theta}^{old}) \leq f(y_{\nu(i+1)}, \widehat{\theta}^{old})$, and get the permutation $\nu = (\nu(1), \ldots, \nu(n))$;

- put $H^{new} := \{y_{\nu(1)}, \ldots, y_{\nu(k)}\}$;
- compute $\widehat{\theta}^{new} := MLE$ based on H^{new};
- define $Q^{new} := \sum_{i=1}^{k} f(y_{\nu(i)}, \widehat{\theta}^{new})$.

Proposition 2.1 *On the basis of the above statements* $Q^{new} \leq Q^{old}$.

Proof of Proposition 2.1. Since $\widehat{\theta}^{new}$ is the MLE based on H^{new} then

$$Q^{new} = \sum_{i=1}^{k} f(y_{\nu(i)}, \widehat{\theta}^{new}) \leq \sum_{i=1}^{k} f(y_{\nu(i)}, \widehat{\theta}^{old}) \leq \sum_{i=1}^{k} f(y_{j_i}, \widehat{\theta}^{old}) = Q^{old}. \square$$

We call this step in our algorithm *C–step* just like Rousseeuw and Van Driessen (1999a) where C is reserved for 'concentration' since H^{new} is more concentrated (has a lower sum of negative log likelihoods) than H^{old}.

Clearly, repeating *C–step* yields an iterative process. When $Q^{new} = Q^{old}$ the process terminates; otherwise we need more *C–steps*. In this way a nonnegative monotonically decreasing sequence $Q^1 \geq Q^2 \geq Q^3 \geq \ldots$ is defined, which by a classical theorem in analysis is always convergent. Moreover, the convergence is guaranteed after a finite number of steps since there are only finitely many k–subsets out of $n!/(k!(n-k)!)$ in all. Finally, we note that this is only a necessary condition for a global minimum of the TL objective function. This gives us a hint as to how to implement an algorithm. Actually, we will be using the suggestion made by Rousseeuw and Van Driessen (1999a): *"Take many initial choices of* H^{old} *and apply C–steps to each until convergence, and keep the solution with lowest value of"* (2.1).

However, this would not be of much use unless we can tell: how to generate different sets H^{old} to start the algorithm; the necessary number of H^{old} sets; how to avoid duplication of work since several H^{old} may yield the same solution; is it possible to reduce the number of *C–steps*?

Unfortunately, at this stage we cannot provide reasonable answer to all these issue alike Rousseeuw and Van Driessen (1999a) as the structure of the data in GLMs beyond the

linear regression case is usually more complicated. However, it is worth to discuss some of these aspects based on the experience concerning the grouped binary linear logistic and Poisson regression cases.

First, we consider the possibilities for the sample sizes of H^{old}. Since the parameter of fullness of the GLMs is given explicitly by $\mathcal{N}(X)$ then any k within the bounds $\mathcal{N}(X)+1 \leq k \leq n$ can be chosen to draw a random k–subset in order to compute $\widehat{\theta}^{old}$. A recommendable choice of k is $\lfloor (n+\mathcal{N}(X)+1)/2 \rfloor$ as the BP of the TLE is maximized. However, following the same reasoning as Rousseeuw and Van Driessen (1999a) and because $\widehat{\theta}^{old}$ can be arbitrary, one should draw subsamples with a smaller $k^* := \mathcal{N}(X)+1$ size as the chance to get at least one outlier free subsample is larger. In practice, for the case of initial choices of H^{old}, we draw finitely many random subsamples of size k^*, calculate ML estimate $\widehat{\theta}^{old}$ for any one, and keep those 10 different subsamples of size k whose TL values evaluated at $\widehat{\theta}^{old}$ are lowest. In this way the resampling process would guarantee better initial choice of H^{old} sets. The recommendable choice of k^* and k could be used as defaults in a software implementation. If the expected percentage of outliers in data is low then a larger value of k can be chosen by the user in order to increase the efficiency of the TL estimator.

Second, as the regression models we consider belong to the linear exponential families an iteratively reweighted least squares algorithm discussed by Green (1994) for obtaining the MLE can be used. Therefore any modern Gauss-Newton nonlinear regression program can be used to carry out the computations as the iteratively reweighted Gauss-Newton, Fisher scoring and Newton-Raphson algorithms are identical to these families, see Jennrich and Moore (1975). In all the applications of the MLE handled by such a program called NLR, see Neytchev et al. (1994), convergence to $\widehat{\theta}^{old}$ discussed in the previous paragraph is reached in about 6 iterations starting from an arbitrary value $\theta^o := (0, \ldots, 0)$.

Third, each *C–step* calculates MLE based on k observations, and the negative log likelihoods for all n observations. In practice, we need 4 or 5 *C–steps* at most to reach convergence starting from $\widehat{\theta}^{old}$ at the first *C–step*, which leads to a faster convergence at the remaining *C–steps*.

A combination of the above elements yields the basis of our algorithm.

If the data set is large one can apply partitioning and nestings in a similar way as in FAST-LTS of Rousseeuw and Van Driessen (1999a), i.e., the entire data is partitioned in a representative way to several data subsets with smaller size. Applying the above algorithm to any subset the best 10 estimates $\widehat{\theta}^{old}_{sub}$ can be calculated. The process continues by making *C–steps* over the merged set, which is composed by pooling the subsets. In this way the best 10 estimates $\widehat{\theta}^{old}_{merged}$ can be obtained, and at last to find the best solution $\widehat{\theta}_{full}$.

When the data set is small all possible subsets with the default size k can be considered for calculation of the TLE skipping the *C–steps* procedure.

The above algorithm can be implemented easily using the environment of the software packages such as GLIM, S-PLUS, SAS, etc.

2.3 Applications

We illustrate our theory and algorithm by three examples. As a first one we analyzed a subset of the data set 28 of Hand et al. (1994) concerning the vaccination successes in three different areas (1=Staffordshire, 2=Cardiff, 3=Sheffield) by using two types of needles (1=fixed, 2=detachable). In the original data set an additional factor, the vaccine batch, was given. This factor was dropped since it had no significant influence and reduces the model's low degree of freedom once more. So a subset of the data with design matrix $X = (x_i^\top)$ with $x_i \in R^4$ is given in Table 2.1. The logistic regression model $\text{logit}(p/(1-p)) = x_i^\top \beta$ is used. As $\mathcal{N}(X) = 6$ the maximum BP attained by any TLE with $k = 8$ is $\frac{2}{9}$. We obtained $TL_8(y, X) = (2.05, -0.92, -0.12, -0.21)^\top$ and $ML(y, X) = (2.01, -0.92, -0.17, -0.15)$ for β. The mean absolute difference between these estimates is less than 0.04. It seems that there are no large influential outliers in the sample. To study the behavior of the estimators in the presence of one outlier we replaced s_1 and t_1 by $s_1 = 0$ and $t_1 = u$, respectively, where u attains several large values. For a study with two outliers we additionally replaced s_9 and t_9 by $s_9 = 0$ and $t_9 = u$. Table 2.2 provides the mean absolute difference between the estimators at the original and the contaminated samples.

These results show clearly that the TLE is stable in the presence of one outlier and

Table 2.1: Subset of data set 28 of Hand et al. (1994).

t_i	s_i	Area	Needle	x_{1i}	x_{2i}	x_{3i}	x_{4i}
228	223	1	1	1	-1	-1	-1
221	210	1	2	1	-1	-1	1
230	218	1	1	1	-1	-1	-1
221	181	2	1	1	1	-1	-1
213	158	2	2	1	1	-1	1
200	160	2	1	1	1	-1	-1
223	198	3	1	1	0	2	-1
228	189	3	2	1	0	2	1
216	177	3	1	1	0	2	-1

Table 2.2: Mean absolute differences.

Estimator	u	10	20	50	100	200	500	1000
MLE	1 outlier	0.15	0.23	0.39	0.54	0.71	0.94	1.11
TLE_8	1 outlier	0.07	0.07	0.07	0.07	0.07	0.07	0.07
MLE	2 outliers	0.15	0.24	0.41	0.57	0.77	1.06	1.28
TLE_8	2 outliers	0.07	0.11	0.19	0.28	0.40	0.55	0.67

breaks down (explodes) in the presence of two outliers. However, the explosion of it and the MLE is not linear in u, it is more logarithmical.

The second example is about a toxicological experiment conducted at the University of Waterloo, Canada, and discussed in O'Hara Hines and Carter (1993) with $n = 48$ observations. A logistic regression model is fitted to the data with covariates for water hardening (WH), and for a linear and quadratic term in log concentration (C) of toxicant

$$\text{logit}(p/(1-p)) = \beta_1 + \beta_2 * WH + \beta_3 * \log_{10}(C) + \beta_4 * \log_{10}(C^2), \tag{2.2}$$

where $\beta_1, \beta_2, \beta_3$, and β_4 are unknown parameters.

Based on all observations the MLE is $(10.28, 0.03, -11.4, 2.50)^\top$. O'Hara Hines and Carter (1993) pinpoint the observations 38, 39 and 26 as possible outliers. They also reported that Pregibon's influence diagnostics indicated the observations 38 and 39 as potential outliers. The MLE without the cases 38 and 39 is $(15.40, 0.27, -15.53, 3.26)^\top$ whereas without the cases 26, 38 and 39 is $(14.04, 0.32, -14.64, 3.11)^\top$.

Markatou et al. (1997) analyzed the same data. They identified the observations 38 and 39 as potential outliers, whilst their methods gave a weight nearly 1 to observations 26 by means of the negative exponential RAF (Residual Adjustment Function) downweight function. When the Hellinger RAF was used for the construction of the weights, observations 13, 32, 40, 43 and 44 received a weight of 0. They reported that examination of those observations revealed that observations 32 and 40 had a 0 response, while observations 43 and 44 had the lowest mortality at concentration levels 720 and 1440, respectively, at the same water-hardening level. The MLE without the observations 13, 32, 40, 43 and 44 is $(6.49, -0.23, -8.42, 1.97)^\top$.

We dropped the observations 32 and 40 in TLE analysis as subcomactness can not be proved because of zero response according to Müller and Neykov (2003), and Vandev and Neykov (1998). Since 24 observations satisfy WH=1, we have $\mathcal{N}(X) = 24$. Hence, the maximum breakdown point is 11/46 and is attained by any TL estimator with $k = 35$ or $k = 36$. Using the TLE algorithm we obtained $TL_{36} = (7.36, -0.12, -9.29, 2.16)^\top$. The trimmed observations are 13, 14, 20, 21, 38, 39, 41, 42, 43, 44. The Pearson residuals

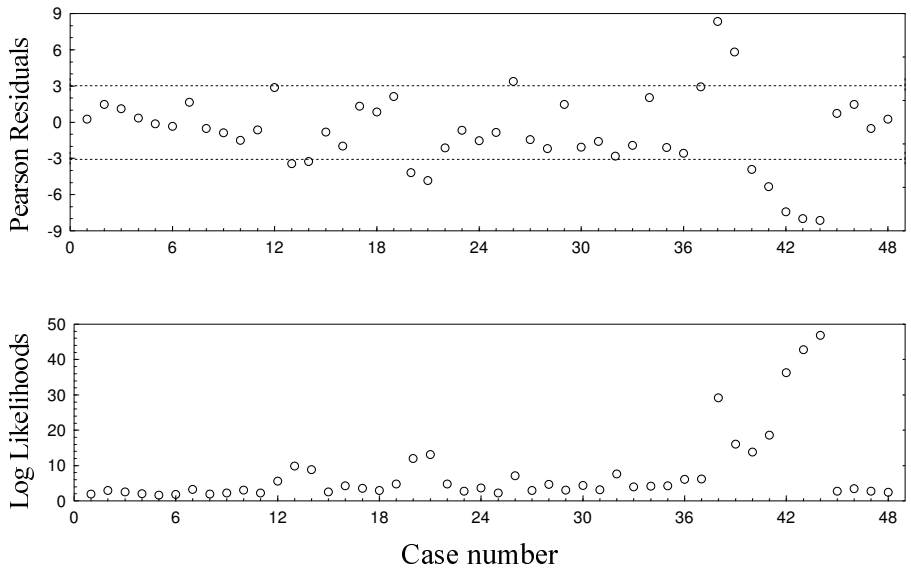

Figure 2.1: O'Hara Hines and Carter (1993) data: Index plot associated with Pearson residuals and log likelihood values based on TLE.

diagnostic calculated by the TL_{36} estimate indicate these observations as potential outliers (see Figure 1). As a bench-mark, the value of 3 is considered. Hence there is some coincidence with the results of Markatou et al. (1997) with respect to the estimate and the trimmed observations.

Next example is about the data set 340 of Hand et al. (1994), given in Table 2.3, concerning the amount of newspaper and TV publicity t_i following $i = 17$ murder-suicides through deliberate crashing of private aircraft and the number y_i of fatal crashes during the week immediately following.

Since fatal crashes are rare events a log-linear model can be assumed where the amount t_i of publicity is the explanatory variable. For simplicity we assume a linear influence

Table 2.3: Data set 340 of Hand et al. (1994).

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
t_i	376	347	322	104	103	98	96	85	82	63	44	40	5	5	0	0	0
y_i	8	5	8	4	6	4	8	6	4	2	7	4	3	2	4	3	2

of t_i, i.e., $x_i = (1, t_i)^\top$. Then the maximum BP is 7/17 and is attained by any TLE with $k = 10$ or $k = 11$ as $\mathcal{N}(X) = 3$. We obtained $TL_{11}(y, X) = (1.086, 0.002)^\top$ and $ML(y, X) = (1.310, 0.002)^\top$ for $\beta = (\beta_1, \beta_2)^\top$. Both estimators provides a very small estimate of the slope of the regression line but they differ with respect to the estimated intercept. This difference is caused by the fact that the TLE trims the highest numbers of crashes at i=1, 3, 5, 7, 8, 11. A scatter plot of this two-dimensional data set is given in Figure 2, along with the MLE (squares) and TLE (triangles) fits.

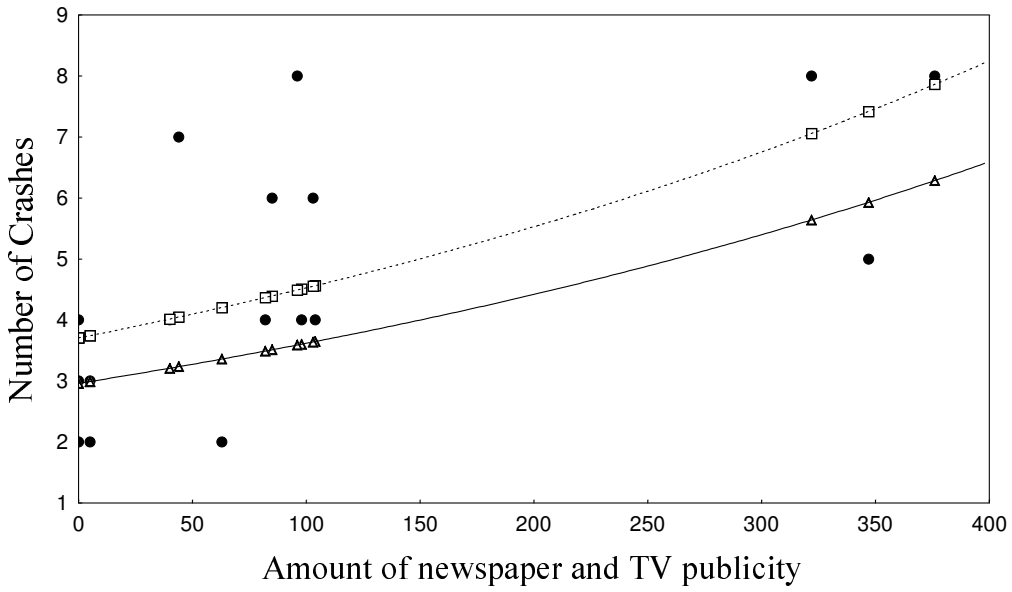

Figure 2.2: Scatterplot of 340 data set of Hand et al. (1994) with MLE (dashed) and TLE (solid) curves.

Chapter 3

Generalized d-fullness Technique for Breakdown Point Study of the Trimmed Likelihood Estimator with Application

Summary. The d–fullness technique of Vandev Vandev (1993) for the finite sample breakdown point study of the Weighted Trimmed Likelihood Estimator is extended. The proposed generalized d–fullness technique is illustrated over the generalized logistic regression model.

3.1 Introduction

The classical Maximum Likelihood Estimator (MLE) can be very sensitive to outliers in the data. In fact, even a single outlier can ruin completely the MLE. To overcome this problem many robust alternatives of the MLE have been developed (see, Atkinson and Riani (2000), Bednarski and Clarke (1993), Beran (1982), Christmann (1994), Field and Smith (1994), Hampel et al. (1986), Huber (1981), Hubert (1997), Marazzi and

Yohai (2004), Markatou et al. (1997), Neykov and Neytchev (1990), Shane and Simonoff (2001), Vandev and Neykov (1993), Windham (1995), and Choi et al. (2000)).

Hadi and Luceño (1997), and Vandev and Neykov (1998) introduced a robust parametric modification of the MLE called the Weighted Trimmed Likelihood (WTL) estimator. The basic idea behind the trimming in the proposed estimator is in the removal of those observations whose values would be highly unlikely to occur if the fitted model was true. These authors showed that under appropriate choices of the trimming parameter and the weights, the WTL estimator reduces to the MLE, to the LMS and LTS estimators of Rousseeuw (1984) in the case of normal regression, and to the MVE and MCD estimators of the multivariate location and scatter introduced by Rousseeuw (1986) in the multivariate normal case.

The $\sqrt{n}$ consistency of the WTL estimator is derived by Čížek (2002).

Algorithms for TL estimation in the univariate case were developed by Hadi and Luceño (1997), whereas Neykov and Müller (2003) proposed a FAST-TLE algorithm in the framework of the generalized linear models.

The breakdown point (BP) properties of the WTL estimator were studied by Vandev and Neykov (1998), and Müller and Neykov (2003) using the d–fullness technique of Vandev (1993). According to Vandev and Neykov (1998), a set $F = \{f_1, \ldots, f_n\}$ of arbitrary functions $f_i : \Theta \to \mathbb{R}^+$, $\Theta \subseteq \mathbb{R}^q$, is called d–full if for every subset $J \subset \{1, \ldots, n\}$ of cardinality d ($|J| = d$) the function $g_J(\theta) = \max_{j \in J} f_j(\theta)$, $\theta \in \Theta$, is subcompact. A function $g : \Theta \to \mathbb{R}$, $\Theta \subseteq \mathbb{R}^q$ is called subcompact if its Lebesgue set $L_g(C) = \{\theta \in \Theta : g(\theta) \leq C\}$ is contained in a compact set for every real constant C, as defined by Müller and Neykov (2003).

The requirement for d–fullness of the set F is restrictive, more precisely, the condition *for every real constant* C in the definition of a subcompact function is not always satisfied. For instance the corresponding set F of log likelihoods for the mixtures of univariate/multivariate normal or binomial distributions, just to name but a few, are not d–full in the above sense.

In this chapter a generalized d–fullness technique is proposed to study the BP of the

WTL estimator for a wider class of functions containing the class of subcompact functions. Section 2 defines the concept of breakdown point and generalized d-fullness. This technique is illustrated over the the generalized logistic regression model in section 3. The lemmas and propositions proofs are given in the Appendix.

3.2 Generalized d-fullness Technique

To aid the presentation we remind the replacement variant of the finite sample BP given in Hampel et al. (1986), which is closely related to that introduced by Donoho and Huber (1983). Let $X = \{x_i \in \mathcal{X} \subseteq \mathbb{R}^p, \text{ for } i = 1, \ldots, n\}$ be a sample of size n.

Definition 3.1 *The BP of an estimator T at X is given by*

$$\varepsilon_n^*(T) = \frac{1}{n} \max\{m : \sup_{\tilde{X}_m} \|T(\tilde{X}_m)\| < \infty\},$$

where $\tilde{X}_m$ is a sample obtained from X by replacing any m of the points in X by arbitrary values from $\mathcal{X}$, and $\|.\|$ is the Euclidean norm.

We now recall the definition of the Weighted Trimmed estimator given in Vandev and Neykov (1998). Let $f : X \times \Theta \to \mathbb{R}^+$, where $\Theta \subseteq \mathbb{R}^q$ be an open set, and $F = \{f_i(\theta) = f(x_i, \theta), \text{ for } i = 1, \ldots, n\}$.

Definition 3.2 *The Weighted Trimmed estimator is defined as*

$$\mathrm{W}_k := \arg\min_{\theta \in \Theta} \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta), \tag{3.1}$$

where $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \ldots \leq f_{\nu(n)}(\theta)$ are the ordered values of f_i at θ, $\nu = (\nu(1), \ldots, \nu(n))$ is the corresponding permutation of the indices, which depends on θ, k is the trimming parameter, the weights $w_i \geq 0$ for $i = 1, \ldots, n$ are associated with the functions $f_i(\theta)$ and are such that $w_{\nu(k)} > 0$.

The W_k estimator is too general to be of practical use. However, many high breakdown statistical estimators can be derived from it. For instance, let $f_i(\theta) = g(|r_i(\theta)|)$ for $i =$

$1, \ldots, n$, where g be a continuos monotonic function such that $g(0) = 0$ and $r_i(\theta) = y_i - x_i^T\theta$ be the ith linear regression residual, generated by the observation $(y_i, x_i^T) \in \mathbb{R}^{p+1}$ and the unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Then the LMS, LTS and LQS regression estimators of Rousseeuw (1984), the h–trimmed weighted L_q estimator of Müller (1995), and the Hössjer (1994) rank-based estimator can be obtained as special cases of the W_k estimator. If $x_i \in \mathcal{X}$ for $i = 1, \ldots, n$ be i.i.d. observations with probability density function $\phi(x, \theta)$, which depends on an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^q$ and $f(x_i, \theta) = -\log \phi(x_i, \theta)$, then the W_k estimator coincides with the WTL_k estimator proposed by Hadi and Luceño (1997), and Vandev and Neykov (1998). In particular, when $\phi(x, \theta)$ is the multivariate normal dencity function, the W_k reduces to the MVE and MCD estimators of the multivariate location and scatter considered by Rousseeuw (1986) (see, Vandev and Neykov (1993)). Vandev and Neykov (1998) prove that the BP of the W_k is not less than $(n-k)/n$ if the set F is d-full, $n \geq 3d$ and $(n+d)/2 \leq k \leq n-d$. Müller and Neykov (2003) extend their result finding the lower bound of the BP without additional assumptions on k and n. Moreover, for the generalized linear models it was shown that the fullness parameter d is related with the quantity $\mathcal{N}(X)$ of Müller (1995), where X is the matrix of the explanatory variables which may not be in general position, as is the case with the designed experiment, or are generated by qualitative factors.

We need the following notations in the presentation of the generalized d–fullness technique.

Let g be a function such that $g : \Theta \to \mathbb{R}$, $\partial\Theta$ be the set of the boundary points of Θ, and $\Theta_\infty = \{\{\theta_k\}_{k=1}^\infty : \theta_k \in \Theta, \|\theta_k\| \to \infty\}$ be the set of all sequences whose norm tends to infinity. Then $\underline{g}$ is defined as

$$\underline{g} = \begin{cases} \inf\limits_{\theta^* \in \partial\Theta} \liminf\limits_{\theta_k \to \theta^*} g(\theta_k), & \text{if } \Theta \text{ is bounded, or} \\ \\ \inf\limits_{\theta^* \in \partial\Theta} \liminf\limits_{\substack{\theta_k \to \theta^* \\ \{\theta_k\} \in \Theta_\infty}} g(\theta_k), & \text{if } \Theta \text{ is unbounded.} \end{cases} \tag{3.2}$$

Let us introduce the following conditions:

A1. $F = \{f_i(\theta) \geq 0,\ i = 1, \ldots, n,\ \text{for } \theta \in \Theta\}$ be a set of continuous functions;

A2. There exists $\theta_0 \in \Theta$, such that for every subset $J \subset \{1, \ldots, n\}$ of cardinality d, $c^{**} g_J(\theta_0) < C$, where $g_J(\theta) = \max_{j \in J} f_j(\theta)$ and $C = \inf_J \underline{g_J}$, $c^{**} = \frac{c^*}{w^*}$, $c^* = \sum_{i=1}^{k} w_{\nu(i)}$, and $w^* = \min\{w_j > 0,\ j = 1, \ldots, n\}$.

Remark 3.1 *The d-full sets class of functions is a special case of the class of sets of functions satisfying A1 and A2 conditions, because if* $\underline{g} = \infty$*, then* g *is a subcompact function. This follows from Lemma 3.1 given in the Appendix.*

The following proposition gives the necessary conditions under which there exists a solution of the optimization problem 3.1.

Proposition 3.1 W_k *is non-empty compact set if* $k \geq d$*, A1 and A2 hold.*

The next proposition gives a lower bound for the BP of W_k for a set of functions F satisfying A1 and A2 conditions. It is a generalization of the corresponding result of Vandev and Neykov (1998) who required d–fullness of F.

Proposition 3.2 *The BP of* W_k *is not less than* $\frac{1}{n}\min(n-k, k-d)$ *if A1 and A2 hold.*

Remark 3.2 *The above propositions hold for the* WTL_k *estimator. Thus, if one wishes to study the BP of the* WTL_k *estimators for a particular distribution, one has to establish the validity of the conditions A1 and A2 for the corresponding set of functions* $f_i(\theta) = -\log \phi(x_i, \theta)$ *for* $i = 1, \ldots, n$*. Then the BP can be exemplified by the range of values of* k *by Proposition* 3.2*.*

3.3 Application on a generalized logistic regression model

As an illustration of the above propositions we consider the grouped binary linear regression model with generalized logistic link. The type of the data under consideration is of the

form (y_i, x_i^T) for $i = 1, \ldots, N$. It is assumed that, y_i is binomially distributed, $b(y_i \mid n_i, \pi_i)$, where the group size is n_i, the probability of success is π_i, and x_i is a p–dimensional vector of covariates (explanatory variables). The total number of observations is $n = n_1 + n_2 + \cdots + n_N$. We will assume that $0 < y_i < n_i$ for each i, and π_i follows the Prentice (1976) generalized logistic distribution

$$\pi_i = (1 + \exp(-\eta_i))^{-a},$$

where $a > 0$, $\eta_i = x_i^T \beta$ is the linear predictor and β is a p–dimensional vector of unknown parameters.

The particular case, when a=1, is considered by Müller and Neykov (2003) who proved that the BP of the WTL_k estimator is equal to $\frac{\min(N-k+1, k-\mathcal{N}(X))}{N}$, where $\mathcal{N}(X) = \max_{0 \neq \beta \in \mathbb{R}^p} \mathrm{card}\{i \in \{1, \ldots, N\}; x_i^T \beta = 0\}$.

We will show that the set $F = \{f(y_i, \eta_i, a), i = 1, \ldots, N\}$, where $f(y_i, \eta_i, a) = -\log \binom{n_i}{y_i} + y_i a \log(1 + e^{-\eta_i}) - (n_i - y_i) \log(1 - (1 + e^{-\eta_i})^{-a})$, satisfies Conditions A1 and A2.

It is obvious that $\lim_{a \to 0} f(y_i, \eta_i, a) = +\infty$, $\lim_{a \to +\infty} f(y_i, \eta_i, a) = +\infty$, and $\lim_{\eta_i \to \pm\infty} f(y_i, \eta_i, a) = +\infty$. Therefore $f(y_i, \eta_i, a)$ is a subcompact function because $\underline{f} = +\infty$.

Proposition 3.3 *The set* $\{f(y_i, x_i, \beta, a), i = 1, \ldots, N\}$ *is* $\mathcal{N}(X) + 1$*-full.*

As a consequence of this proposition, the following corollary is obtained.

Corollary 3.1 *The set* WTL_k *for the grouped binary linear regression model with generalized logistic link is a non empty compact set if* $k \geq \mathcal{N}(X) + 1$.

Applying Theorem 2 of Müller and Neykov (2003) we get the following

Corollary 3.2 *If* $\lfloor (N + \mathcal{N}(X) + 1)/2 \rfloor \leq k \leq \lfloor (N + \mathcal{N}(X) + 2)/2 \rfloor$*, then the BP of the* WTL_k *estimator for the grouped binary linear regression model with generalized logistic link is*

$$\varepsilon_N^*(WTL_k) \geq \frac{1}{N} \left\lfloor \frac{N - \mathcal{N}(X) + 1}{2} \right\rfloor.$$

We remind that $\lfloor z \rfloor := \max\{n : n \leq z\}$.

3.4 Appendix

The proofs in this section are based on the following

Lemma 3.1 *Let $g : \Theta \to \mathbb{R}$ be continuous function, $\Theta \subseteq \mathbb{R}^q$ be an open set. If there exists $\theta_0 \in \Theta$ and a real constant $a \geq 1$, such that $ag(\theta_0) < C$, where $C \leq \underline{g}$, then the set $S = \{\theta : g(\theta) < C\}$ is bounded and non empty.*

Proof of Lemma 3.1. We shall note that the set S is non empty since $\theta_0 \in S$. Let us assume that S is unbounded, $\{\theta_j\}_{j=1}^{\infty}$ be a sequence from S such that $\|\theta_j\| \underset{j\to\infty}{\longrightarrow} \infty$ and $r_j = \max\{\|\theta_1\|, \ldots, \|\theta_j\|\}$. Then we have that $\inf\limits_{\|\theta_j\| \geq r_j} g(\theta) \leq g(\theta_j) < C$. Taking a limit we get $\underline{g} < C$, which is a contradiction with $C \leq \underline{g}$. □

We will use the representation $f_{(k)}(\theta) = \min\limits_{I \in I_k} \max\limits_{i \in I} f_i(\theta)$ which holds at any fixed θ and I_k is the set of all subsets of $\{1, \ldots, n\}$ consisting of k elements in the propositions proof (see, Krivulin (1992)).

Proof of Proposition 3.1: The following inclusions hold

$$
\begin{aligned}
\mathrm{W}_k &= \left\{\theta : \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta) \le \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\vartheta) \ \forall\ \vartheta \in \Theta\right\} \\
&= \left\{\theta : \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta) \le \inf_{\vartheta\in\Theta} \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\vartheta)\right\} \\
&\subseteq \left\{\theta : \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta) \le \inf_{\vartheta\in\Theta} \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(k)}(\vartheta)\right\} \\
&\subseteq \left\{\theta : \min_{I\in I_k} \sum_{i\in I} w_i f_i(\theta) \le c^* \inf_{\vartheta\in\Theta} f_{\nu(k)}(\vartheta)\right\} \\
&\subseteq \bigcup_{I\in I_k} \left\{\theta : \sum_{i\in I} w_i f_i(\theta) \le c^* \inf_{\vartheta\in\Theta} f_{\nu(k)}(\vartheta)\right\} \\
&= \bigcup_{I\in I_k} \left\{\theta : \sum_{i\in I} w_i f_i(\theta) \le c^* \inf_{\vartheta\in\Theta} \min_{I\in I_k} \max_{i\in I} f_i(\vartheta)\right\} \\
&\subseteq \bigcup_{I\in I_k} \left\{\theta : \sum_{i\in I} w_i f_i(\theta) \le c^* \inf_{\vartheta\in\Theta} \max_{i\in I} f_i(\vartheta)\right\} \\
&\subseteq \bigcup_{I\in I_k} \left\{\theta : w^* \max_{i\in I} f_i(\theta) \le c^* \inf_{\vartheta\in\Theta} \max_{i\in I} f_i(\vartheta)\right\} \\
&\subseteq \bigcup_{I\in I_k} \left\{\theta : \max_{i\in I} f_i(\theta) \le c^{**} \max_{i\in I} f_i(\theta_0)\right\} \\
&\subseteq \bigcup_{I\in I_k} \bigcup_{J\subset I} \left\{\theta : \max_{j\in J} f_j(\theta) \le c^{**} \max_{j\in J} f_j(\theta_0)\right\} \\
&= \bigcup_{I\in I_k} \bigcup_{J\subset I} \{\theta : g_{_J}(\theta) \le c^{**} g_{_J}(\theta_0)\} \\
&\subset \bigcup_{I\in I_k} \bigcup_{J\subset I} \{\theta : g_{_J}(\theta) < C\}
\end{aligned}
$$

The latter set is a non-empty bounded set according to Lemma 3.1, since A2 condition is satisfied. Therefore, W_k is a compact set, since the functions from F are continuous. □

Proof of Proposition 3.2: Let $\tilde{F} = \{\tilde{f}_i(\theta) = f(\tilde{x}_i, \theta) \text{ for } i = 1, \ldots, n\}$ be obtained from F upon replacement of $m = \min\{n-k, k-d\}$ observations from the sample X with arbitrary ones from $\mathcal{X}$, the weights w_i correspond to the functions that belong to F and $\tilde{\mathrm{W}}_k$

is the corresponding analog of W_k, defined over $\tilde{F}$. The number of the original functions in $\tilde{F}$ is $n-m \geq k$. Hence there exists $I^* \in I_k$ such that $\tilde{f}_i(\theta) \equiv f_i(\theta)$ for $i \in I^*$. The following inequalities hold $\tilde{f}_{\nu(k)}(\theta) = \min\limits_{I \in I_k} \max\limits_{i \in I} \tilde{f}_i(\theta) \leq \max\limits_{i \in I^*} f_i(\theta)$ and $f_{\nu(d)}(\theta) \leq \tilde{f}_{\nu(k)}(\theta)$ (see, Müller and Neykov (2003)). Then we have the inclusions

$$
\begin{aligned}
\tilde{\mathrm{W}}_k &= \left\{ \theta : \sum_{i=1}^{k} w_{\nu(i)} \tilde{f}_{\nu(i)}(\theta) \leq \inf_{\vartheta \in \Theta} \sum_{i=1}^{k} w_{\nu(i)} \tilde{f}_{\nu(i)}(\vartheta) \right\} \\
&\subseteq \left\{ \theta : w^* \tilde{f}_{\nu(k)}(\theta) \leq c^* \inf_{\vartheta \in \Theta} \tilde{f}_{\nu(k)}(\vartheta) \right\} \\
&\subseteq \left\{ \theta : w^* \tilde{f}_{\nu(k)}(\theta) \leq c^* \inf_{\vartheta \in \Theta} \max_{i \in I^*} f_i(\vartheta) \right\} \\
&\subseteq \left\{ \theta : w^* \tilde{f}_{\nu(k)}(\theta) \leq c^* \max_{i \in I^*} f_i(\theta_0) \right\} \\
&\subseteq \bigcup_{\substack{J \subset I^* \\ |J|=d}} \left\{ \theta : w^* \tilde{f}_{\nu(k)}(\theta) \leq c^* \max_{j \in J} f_j(\theta_0) \right\} \\
&\subseteq \bigcup_{\substack{J \subset I^* \\ |J|=d}} \left\{ \theta : w^* f_{\nu(d)}(\theta) \leq c^* \max_{j \in J} f_j(\theta_0) \right\} \\
&\subset \bigcup_{\substack{J \subset I^* \\ |J|=d}} \left\{ \theta : f_{\nu(d)}(\theta) < C \right\}
\end{aligned}
$$

According to Lemma 3.1 the final set is a non-empty bounded set. Therefore, $\tilde{\mathrm{W}}_k$ is a compact set, since the functions from $\tilde{F}$ are continuous. □

Proof of Proposition 3.3. Let $C \in \mathbb{R}$ is arbitrary. Since $f(y_i, \eta_i, a)$ is a subcompact function of η_i and a, there exist constants B_i and A_i for $i \in I \subset \{1, \ldots, N\}$, $\mathrm{card}(I) = \mathcal{N}(X) + 1$, such that the set

$$
\begin{aligned}
& \{\beta \in \mathbb{R}^p, a > 0 : \max_{i \in I} f(y_i, x_i, \beta, a) \leq C\} \\
= & \bigcap_{i \in I} \{\beta \in \mathbb{R}^p, a > 0 : f(y_i, x_i, \beta, a) \leq C\} \\
= & \bigcap_{i \in I} \{\beta \in \mathbb{R}^p, a > 0 : f(y_i, \eta_i = x_i^T \beta, a) \leq C\} \\
\subset & \bigcap_{i \in I} \{\{\beta \in \mathbb{R}^p : |x_i^T \beta| \leq B_i\} \times \{a : 0 < a \leq A_i\}\}
\end{aligned}
$$

is contained in a bounded set. (The set $\{\beta \in \mathbb{R}^p | x_i^T \beta| \leq B_i\}$ is bounded for all B_i according to Lemma 3 of Müller and Neykov (2003).) $\square$

Chapter 4

TLE of the Parameters of the GEV Distributions: A Monte-Carlo Study

Summary. The applicability of the Trimmed Likelihood Estimator (TLE) proposed by Neykov and Neytchev (1990) to the extreme value distributions is considered. The effectiveness of the TLE in comparison with the classical MLE in the presence of outliers in various scenarios is illustrated by an extended simulation study. The FAST-TLE algorithm developed by Neykov and Müller (2003) is used to get the parameter estimate. The computations are carried out in the R environment using the package *ismev* originally developed by Coles (2001) and ported in R by Stephenson (2002).

4.1 Introduction

The extreme value distributions theory has been intensively developed. The book of Coles (2001) provides a useful theoretical background. The Maximum Likelihood is the standard technique for statistical inference in extremes. It is well known that the MLE can be very sensitive to outliers in the data. Indeed, the simulation study of Barão and Tawn (1999) shows that in the presence of outliers, the parameter estimates are significantly influenced and thus the return period. Relatively little attention to robustness has been paid in the

context of extreme values. To overcome this problem Dupuis and Field (1998), Dupuis and Morgenthaler (2002), and Dupuis and Tawn (2001) estimate robustly the parameters of various extreme value distributions using the so called B-optimal robust M-estimators of **?**. It is concluded that these estimators are more efficient than MLE under some model assumptions violation. Unfortunately, these estimators do not possess a high Breakdown Point (BP) and hence are not appropriate for the modeling purposes, as with the increasing the number p of the explanatory variables their BP decreases to zero as $1/p$. (Roughly speaking, the BP is the smallest fraction of contamination that can cause the estimator to take an arbitrarily large value.) In practice, one needs robust estimators that possess a high BP resistant against high percentage of surrogate (aberrant, anomalous) observations in data. For instance, such observations arise when data are collected by different ways.

Several parametric robust alternatives of the ML estimator possessing high BP have been developed, e.g., Choi et al. (2000), Markatou et al. (1997), Neykov and Neytchev (1990), and Windham (1995). To our knowledge, none of these high BP estimators has been used for the purposes of the extreme value modeling. Thus, the main goal of the paper is to develop a robust parametric approach for extreme values statistical modeling based on the TLE proposed by Neykov and Neytchev (1990). The TLE is looking for that sub-sample of k observations out of n the original data size with the optimal likelihood. The trimming number of observations can be chosen by the user in appropriate bounds to get a high BP and optimal efficiency. Details about the properties of the TLE can be found in Vandev and Neykov (1993), Vandev and Neykov (1998), Neykov and Müller (2003), Čížek (2002), **?**, and Dimova and Neykov (2004). Because the TLE accommodates the classical MLE, the extreme value methodology, which is based mainly on the MLE, can be adapted and further developed.

In this paper we consider an application of the TLE to the Generalized Extreme Value (GEV) distribution, however, the generalized Pareto distribution or the Poison point approaches for modeling of extreme values can be used instead. A simulation study is performed to illustrate the effectiveness of the TLE in comparison with the MLE.

4.2 Basic definitions and notions

In the following, the GEV distribution is introduced. It arises as the limiting distribution of the maxima of a series of independent and identically distributed (i.i.d.) observations. The distribution function of the GEV is given by

$$G(x;\mu,\sigma,\xi) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} & \text{if} \quad \xi \neq 0, \\ \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} & \text{if} \quad \xi = 0. \end{cases} \tag{4.1}$$

where $\{x : 1+\xi(x-\mu)/\sigma > 0\}$, $\sigma > 0$, and μ, σ, ξ are location, scale and shape parameters, respectively, see Coles (2001).

The Fréchet and Weibull distributions are obtained for $\xi < 0$ and $\xi > 0$, respectively. The case of $\xi = 0$ is interpreted as the limit of the GEV as $\xi \to 0$, widely known as the Gumbel distribution. The MLE is completely regular if $\xi > -0.5$, it exists but is not completely regular if $-1 < \xi < -0.5$ and it does not exist if $\xi < -1$, according to Smith Smith (1985).

We now recall the definition of the Trimmed Likelihood Estimator. Let $x_1, \ldots, x_n$ be i.i.d. observations with density function $f(x,\theta)$, depending on unknown parameter θ and $l(x_i,\theta) = -\log f(x_i,\theta)$.

Definition 1. The Trimmed Likelihood Estimator (TLE) is defined in Neykov and Neytchev (1990) as

$$\hat{\theta} := \arg\min_{\theta\in\Theta} \sum_{i=1}^{k} l(x_{\nu(i)},\theta), \tag{4.2}$$

where $l(x_{\nu(1)},\theta) \leq l(x_{\nu(2)},\theta) \leq \ldots \leq l(x_{\nu(n)},\theta)$ are the ordered values of $l(x_i,\theta)$ for $i = 1,\ldots,n$ at θ, $\nu = (\nu(1),\ldots,\nu(n))$ is the corresponding permutation of the indexes, which depends on θ and k is the trimming parameter.

The basic idea behind the trimming in this estimator is in removal of those $n-k$ observations which values would be highly unlikely to occur, had the fitted model been true. The TLE coincides with the MLE if $k = n$. Due to the representation

$$\min_{\theta\in\Theta} \sum_{i=1}^{k} l(x_{\nu(i)},\theta) = \min_{\theta\in\Theta} \min_{I\in I_k} \sum_{i\in I} l(x_i,\theta) = \min_{I\in I_k} \min_{\theta\in\Theta} \sum_{i\in I} l(x_i,\theta)$$

where I_k is the set of all k–subsets of the set $\{1,\ldots,n\}$, it follows that all possible $\binom{n}{k}$ partitions of the data have to be fitted by the MLE. Therefore, the TLE is given by the partition with that MLE fit for which the negative log likelihood is minimal.

General conditions for the existence of a solution of (5.2) can be found in Dimova and Neykov (2004), whereas the consistency is proved in Čížek (2002). The BP of the TLE is studied by Vandev and Neykov (1998), **?**, and Müller and Neykov (2003) using the d–fullness technique proposed by Vandev (1993). According to Vandev (1993), the set $F = \{l(x_i,\theta), i = 1,\ldots,n\}$ is called d-full if for any subset of cardinality d of F, the supremum of this subset is a subcompact function. A real valued function $g(\theta)$ is called subcompact if the sets $L_{g(\theta)}(C) = \{\theta : g(\theta) \leq C\}$ are compact for any constant C. The BP of the TL is not less than $\frac{1}{n}\min\{n-k, k-d\}$ if the corresponding set of negative loglikelihoods is d–full (see, Müller and Neykov (2003)). It is easy to show that in case of Gumbel distribution $d = 2$. When the location parameter is a monotone function of a linear predictor, $\mu = h(z_i^T\beta)$, where $\beta \in R^p$ is unknown parameter and $Z := (z_i^\top)$ is the data matrix of rank p of the explanatory variables $z_i \in R^p$, then $d = p+1$. Determination of the d–fullness parameter for the Fréchet and Weibull distributions is not considered because of the complexity of parameters' domain.

Increasing k, the estimator will possess a BP point less than the highest possible, but it will be more efficient at the same time.

Computation of the TLE is infeasible for large data sets because of its combinatorial nature. To get approximate TLE an algorithm called FAST-TLE is developed in Neykov and Müller (2003). It reduces to the FAST-LTS or FAST-MCD algorithms in the normal regression or multivariate Gaussian cases. The basic idea behind the FAST-TLE algorithm consists of carrying out finitely many times a two-step procedure: a trial step followed by a refinement step. In the trial step a subsample of size k^* is selected randomly from the data sample and then the model is fitted to the subsample to get a trial MLE. The refinement step is based on the so called concentration procedure. The cases with the k smallest negative log likelihoods from the trial fit are found. Fitting the model to these k cases gives an improved fit. Repeating the improvement step yields an iterative process.

Convergence is always guaranteed after a finite number of steps since there are only finitely many k–subsets out of $\binom{n}{k}$ in all. The estimate with the lowest TL objective function is retained. There is no guarantee that this value will be the global minimizer but one can hope that it would be a close approximation to it. The trial subsample size k^* should be greater than or equal to $p+1$ which is needed for the existence of the MLE but the chance to get at least one outlier free subsample is larger if $k^* = p+1$. Any k within the interval [p+1, n] can be chosen in the refinement step. A recommendable choice of k is $\lfloor (n+p+1)/2 \rfloor$ because then the BP of the TLE is maximized, where $\lfloor r \rfloor := \max\{n \in N; n \leq r\}$. We note that, if the data set is small, all possible subsets of size k can be considered.

4.3 Simulation design

We compare the performance of the MLE and the TLE through a simulation study for a range of different situations of GEV generated data sets. The regular data follow the model

$$y_i \sim \text{GEV}(\mu_i = 1 + x_i, \sigma = 1, \xi = 0.0), \text{ where } x \sim N(0, 7).$$

The outliers follow the model

$$y_i \sim \begin{cases} \text{U}(y_{max} + \mu_i, y_{max} + (y_{max} - y_{min}) + \mu_i) & \text{if} \quad x_i \geq \bar{x}, \\ \text{U}(y_{min} - (y_{max} - y_{min}) - \mu_i, y_{min} - \mu_i) & \text{if} \quad x_i < \bar{x}. \end{cases}$$

The regular observations and outliers union comprises the contaminated sample of size $n = 100$. Thus, samples with levels of contamination 0%, 10%, 20%, 30% and 40% are considered. The trimming percentage $\frac{n-k}{n}100\%$ is held fixed at 0%, 5%, ..., 45%. The MLE and TLE are computed over the regular and contaminated data. These estimators are compared using the mean, median, root mean square error and various quantiles criteria over 400 independent replications of the simulation experiment at any contamination level.

All the computations were carried out in the R environment using the *ismev* package

originally developed by Coles (2001) in S-Plus and ported in R by Stephenson Stephenson (2002).

4.4 Simulation results

On all plots in Figures 1-4, the data sets that constitute the regular observations are represented by bullets, while the outliers, if any, are represented by triangles. The ML and TL fits are based on the contaminated samples. Exceptions are the upper left plots where the ML fits are based only on the regular observations. The dashed lines describe the generated model, whereas the straight lines describe the ML and TL fits in all of these plots. Due to space limitations, only some selected TL fits are presented. In all other plots, the empty triangles or tiny circles describe the trimmed observations (either regular or outliers). The two numbers in the plots' title represent the (trimmed)log-likelihood value of the current estimate and the (trimmed)log-likelihood value evaluated over the regular data at this estimate.

It is a well known fact that the MLE can easily be influenced by a single bad observation whereas the TLE is resistant up to $\frac{n-k}{n}100\%$ percentage of outliers. To explore the behavior of the estimators the simulation experiment was replicated more than 400 times. As a consequence, a series of estimates were obtained and their distribution was studied.

The plot panels on Fig. 1-4 represent some of these experiments. Generally, the plots indicate that the MLE becomes completely useless if the percentage of observations that do not follow the model is large, while the TLE gives better fits. However, the quality of the TLE fits depends on the trimming percentage $\frac{n-k}{n}100\%$. As it could be expected, the TL estimates are more stable for those values of k that satisfy the inequality $\frac{n-k}{n}100\% \geq \alpha$, where α is the contamination level. The series of box-plots in the "Intercept", "Slope", "Scale" and "Shape" panels on Figure 5 give a more detailed characterization of the distribution of the GEV parameters' estimates conditional on the different trimming percentages. Any of these box-plots series exhibits some specific properties that could serve as a guide to get an idea about the optimal trimming percentage. One can see that the box-plots

variation in any panel becomes more stable by increasing the trimming percentage. A large percentage of trimming exhibits great influence on the scale and shape estimates. This is because the relationship between the trimming percentage and the scale estimate is inversely proportional while it is generally nonlinear for the shape parameter estimate. For instance, even a single outlier can drive the MLE scale estimate to infinity. However, increasing the trimming percentage leads to underestimating of the scale. So, trimming with large percentages outside the location-scale distributions framework should be done with great care. It can be seen that there is a common interval of trimming percentages where the parameters estimates become more stable in all these panels ("Intercept", "Slope", "Scale" and "Shape"). Therefore, an optimal choice of the trimming percentage could be the minimal value of that interval.

Usually, the percentage of outliers in real data is unknown. Therefore, one can proceed by a TLE, based on a decreasing range of values for k, starting with $k = n$. However, the TL estimation procedure must be repeated several times at any particular value of k. When the parameters estimate stabilization occurs then following the previous recommendations on the trimming choice, not only the unknown GEV parameters but also the outliers percentage in the data can be estimated robustly.

4.5 Summary and conclusions

The simulation study demonstrates that the TLE is a useful alternative to the MLE in the framework of extreme value modeling. The extreme values data can be analyzed with the TLE methodology just as with the classical MLE, however, over sub-samples. Therefore the computation can be carried out by a standard MLE procedure for fitting extreme value distributions to data closely following the FAST-TLE algorithm of Müller and Neykov (2003). Such procedures are widely available in software packages such as S-PLUS, R, SAS. The TLE will lead to greater computational effort, but having in mind the growing power of modern-day processors and memory, one can afford it.

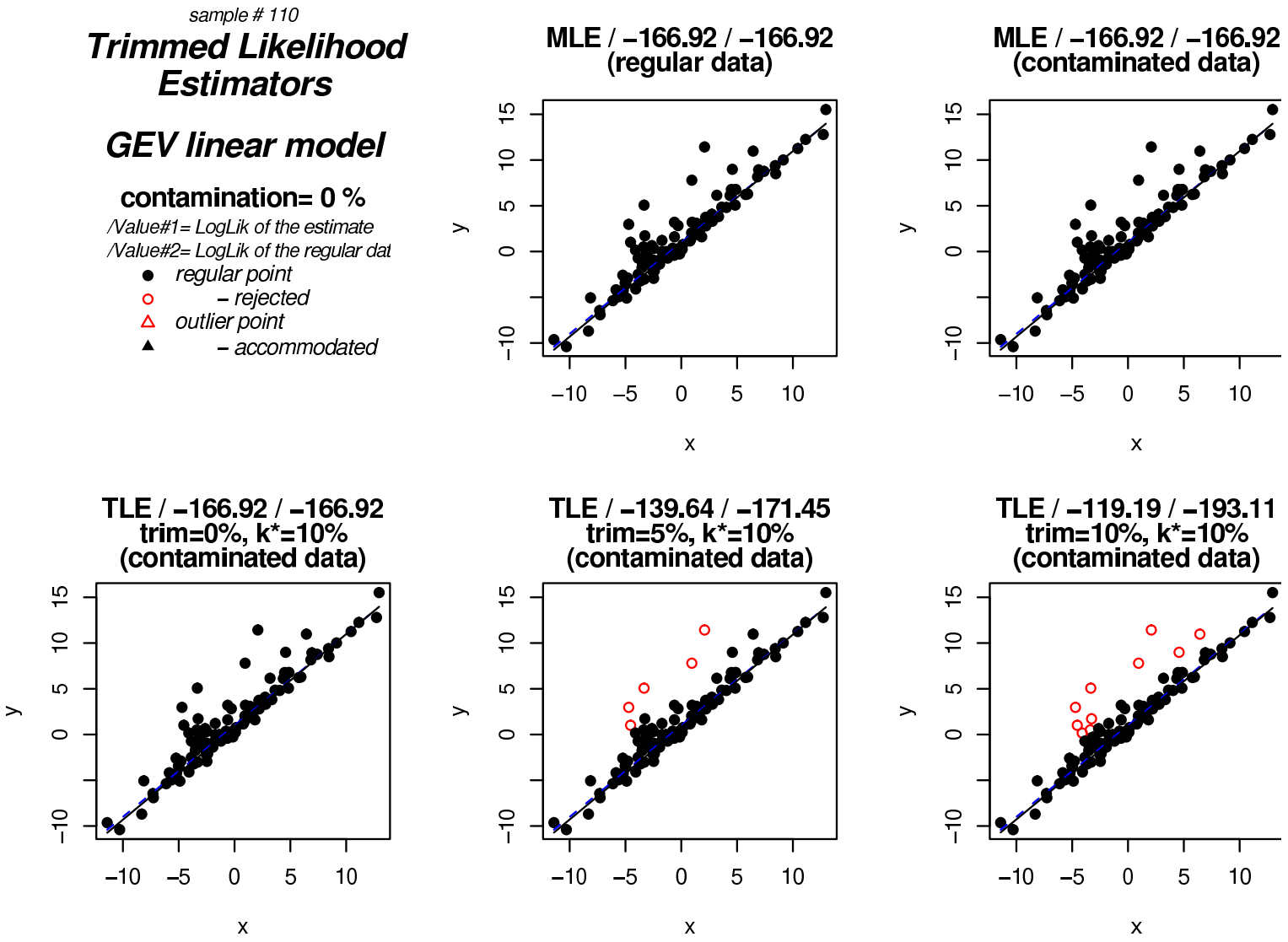

Figure 4.1: Experiments with zero percentage of contamination.

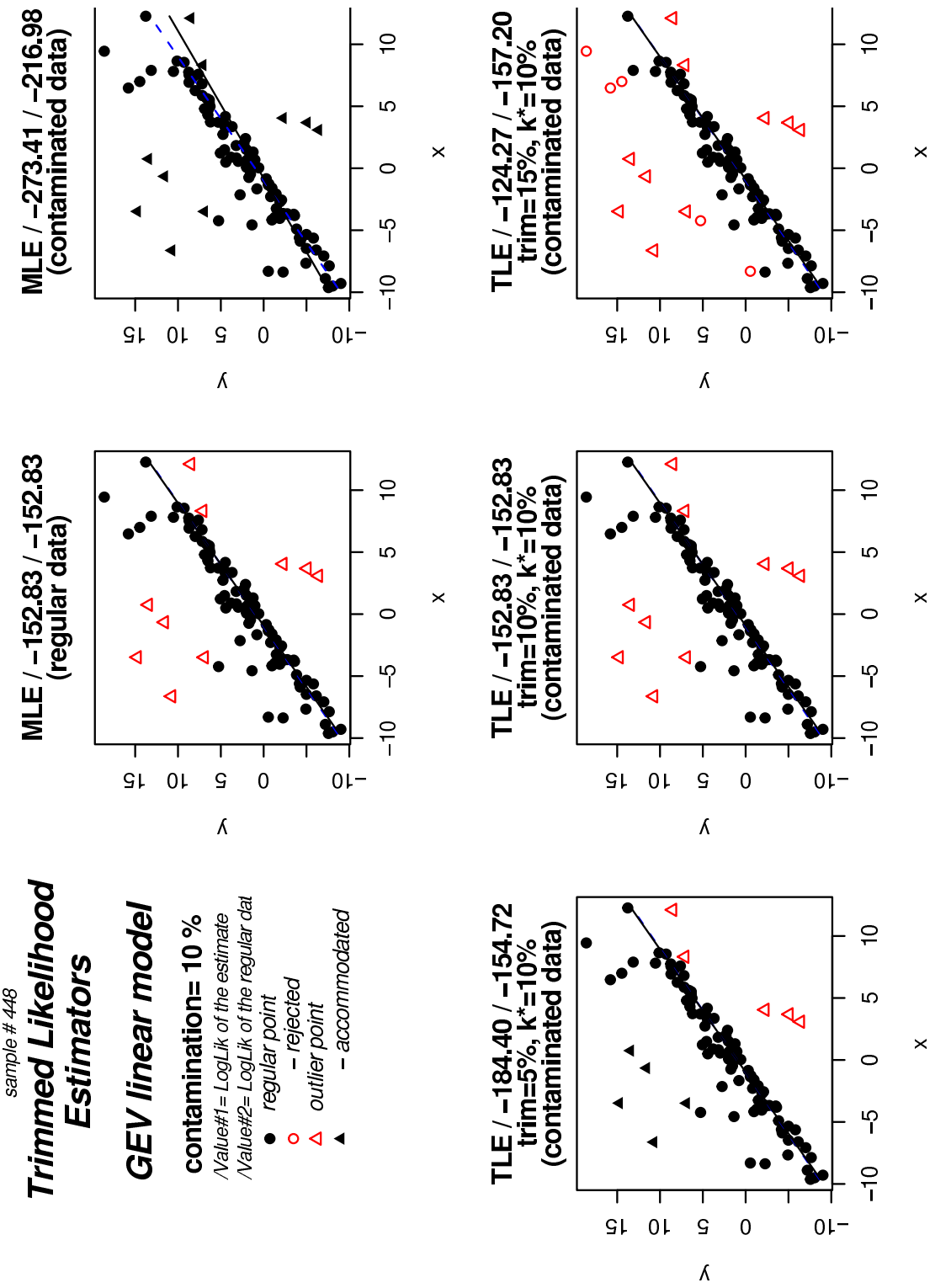

Figure 4.2: Experiments with 10% of contamination.

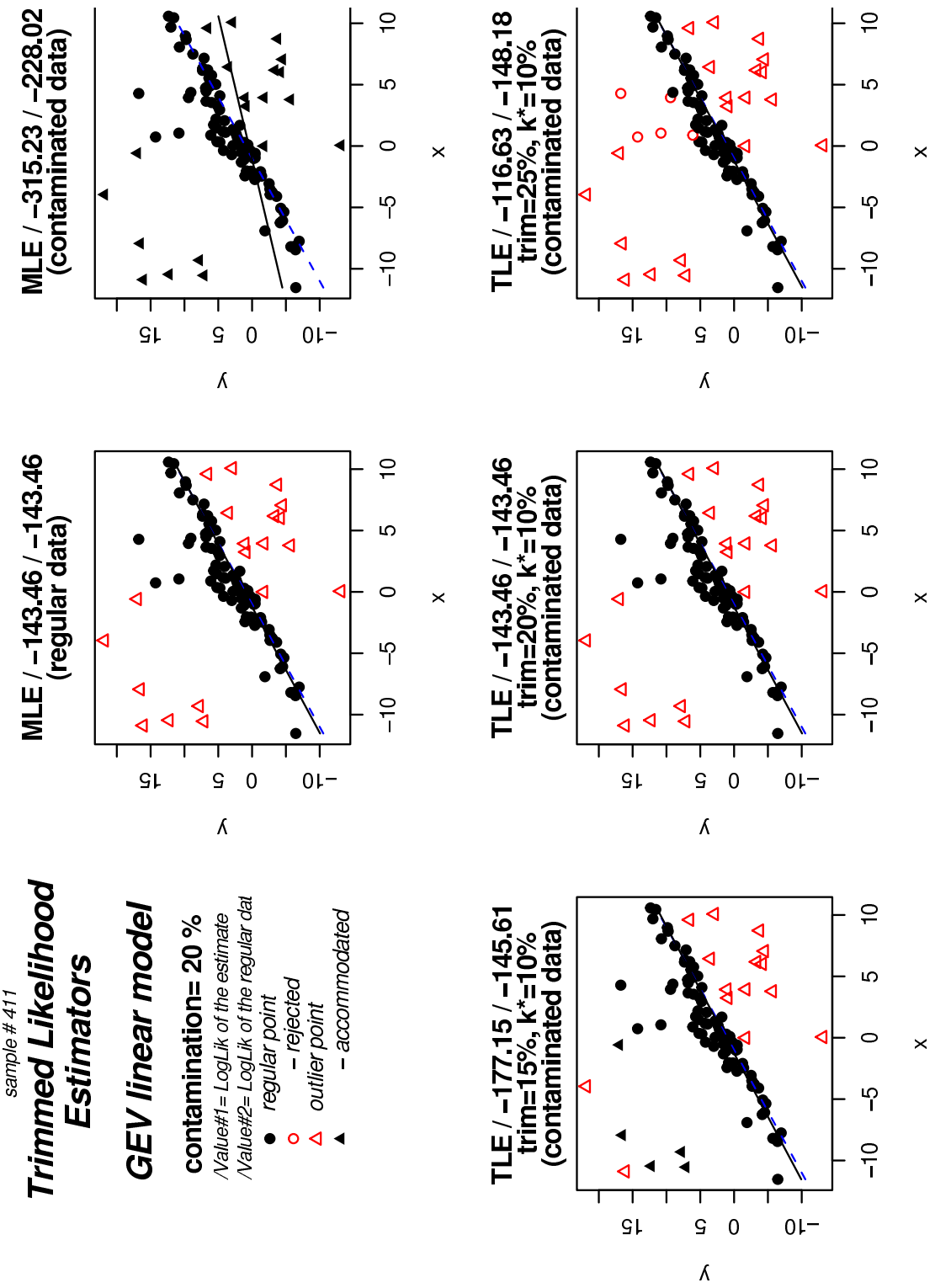

Figure 4.3: Experiments with 20% of contamination.

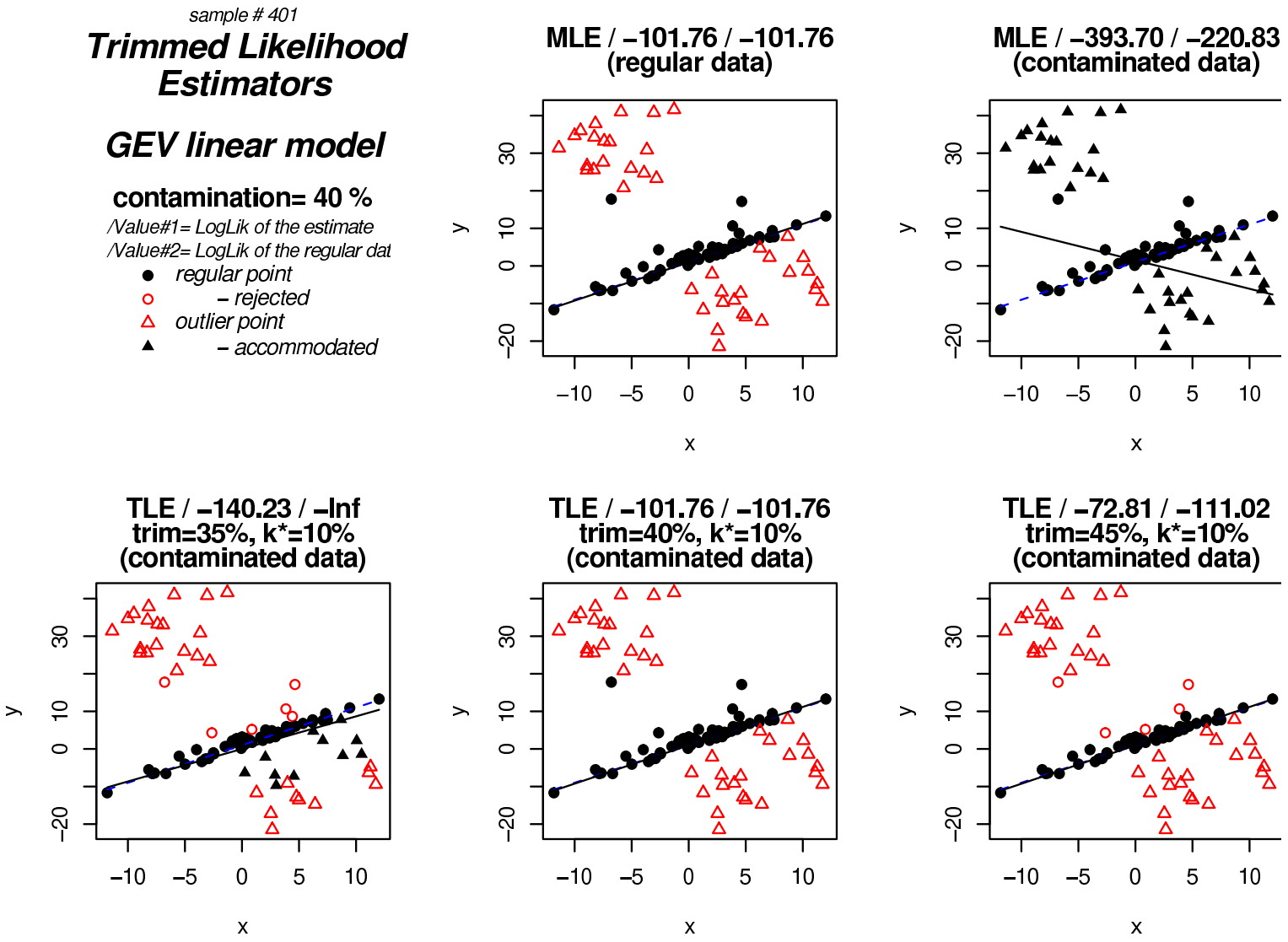

Figure 4.4: Experiments with 40% of contamination.

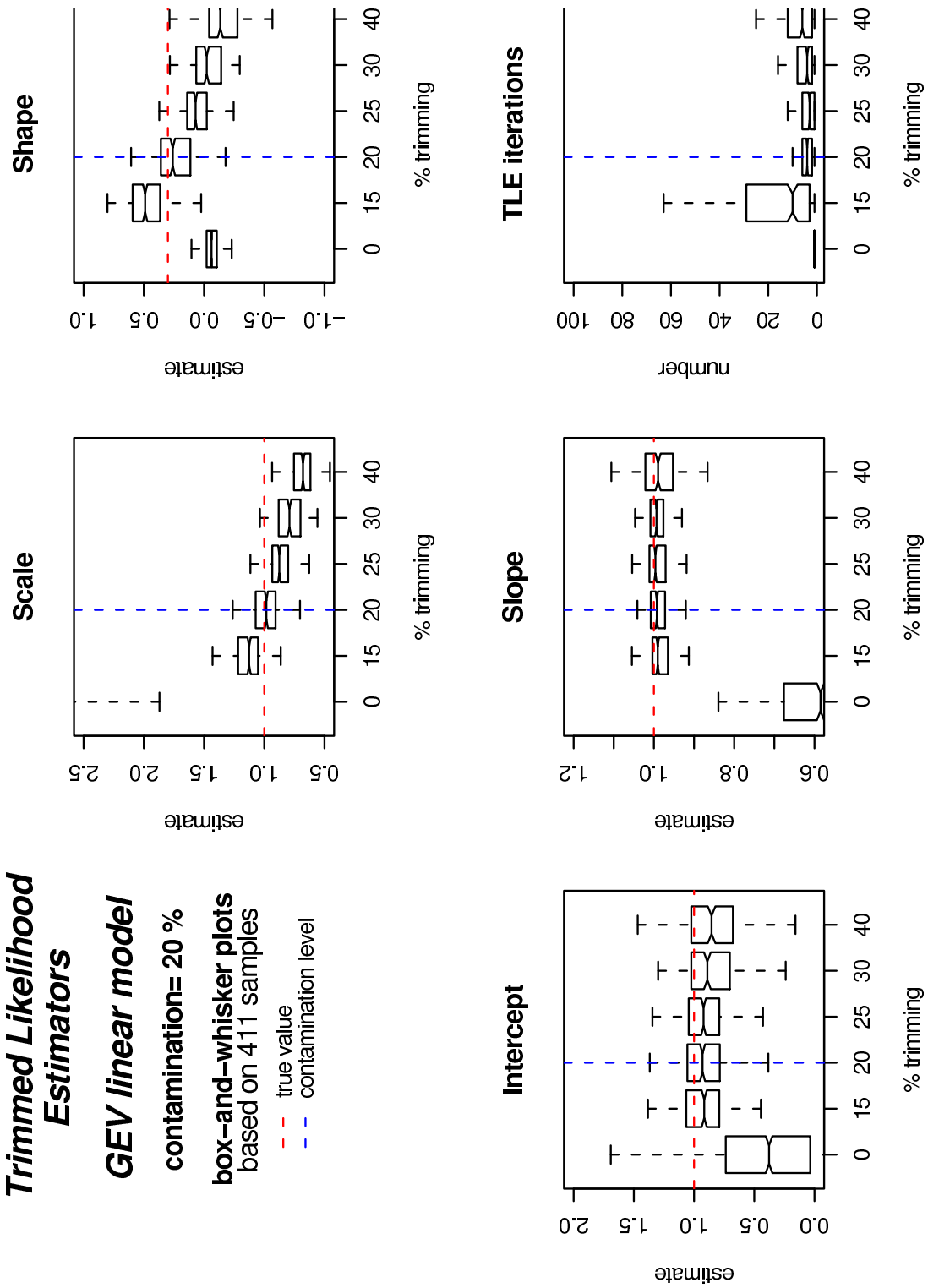

Figure 4.5: Distribution of the GEV estimates of location (intercept and slope), scale and shape parameters based on 411 experiments.

Chapter 5

Robust fitting of mixtures using the Trimmed Likelihood Estimator

Summary. The Maximum Likelihood Estimator (MLE) has commonly been used to estimate the unknown parameters in a finite mixture of distributions. However, the MLE can be very sensitive to outliers in the data. In order to overcome this the Trimmed Likelihood Estimator (TLE) is proposed to estimate mixtures in a robust way. The superiority of this approach in comparison with the MLE is illustrated by examples and simulation studies. Moreover, as a prominent measure of robustness, the Breakdown Point (BDP) of the TLE for the mixture component parameters is characterized. The relationship of the TLE with various other approaches that have incorporated robustness in fitting mixtures and clustering are also discussed in this context.

5.1 Introduction

Finite mixtures of distributions have been widely used to model a wide range of heterogeneous data. In most applications the mixture model parameters are estimated by the MLE via the expectation-maximization (EM) algorithm, see e.g.McLachlan and Peel (2000). It is well known, however, that the MLE can be very sensitive to outliers in the data. In fact,

even a single outlier can completely ruin the MLE which in mixture settings means that at least one of the component parameters estimate can be arbitrarily large. To overcome this, robust parametric alternatives of the MLE have been developed, e.g., Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987). A direct fitting of mixture models to data by these robust estimators is of limited use. The reason is that these robust estimators are designed to fit a parametric model to the majority of the data whereas the remaining data which do not follow the model are considered as outliers. In practice, however, the data could be quite heterogeneous without having a homogeneous part consisting of at least 50% of the data. Fortunately, since the EM algorithm is capable to transfer a complex mixture MLE problem into relatively simple single component MLE problems, some of the ideas of robust estimation have been adapted to mixture models. Details can be found in Campbell (1984), Kharin (1996), Davé and Krishnapuram (1997), Medasani and Krishnapuram (1998), McLachlan and Peel (2000), Hennig (2003), just to name a few. In this way robustness has been adapted to meet the problem with outliers in mixtures of the location-scale family of distributions. Generally speaking, robust fitting of mixtures of distributions outside this family has not been developed yet. Exceptions are Markatou (2000) and Neykov et al. (2004) who discussed fitting mixtures of Poisson regressions based on the weighted MLE and Trimmed Likelihood Estimator (TLE) via simulations.

Thus, after many years of parallel development of fitting mixtures, cluster analysis, outlier detection and robust techniques, the need for a synthesis of some of these methods beyond the location scale family of distributions has become apparent. Such a synthesis can be a flexible and powerful tool for an effective analysis of heterogeneous data. So, the aim of this chapter is to make a step toward the achievement of this goal by offering a unified approach based on the TLE methodology. Because the TLE accommodates the classical MLE, the finite mixture methodology based on the MLE can be adapted and further developed. In this chapter the superiority of this approach in comparison with the MLE is illustrated.

The paper is organized as follows. In Section 2, the basic properties of the weighted TLE are presented. In Section 3 we briefly discuss the EM algorithm and explain how robustness

can be incorporated. Moreover, the TLE software implementation and adjustments to the framework of mixtures with existing software are presented. Comparisons of the MLE and TLE by examples and simulations are presented in Section 4. Finally, in Section 5 the conclusions are given.

5.2 The Trimmed Likelihood methodology

Definition 5.1 *The Weighted Trimmed Likelihood Estimator (WTLE) is defined in Hadi and Luceño (1997) and in Vandev and Neykov (1998) as*

$$\hat{\theta}_{WTLE} := \arg\min_{\theta\in\Theta^p} \sum_{i=1}^{k} w_{\nu(i)} f(y_{\nu(i)};\theta), \tag{5.1}$$

where $f(y_{\nu(1)};\theta) \le f(y_{\nu(2)};\theta) \le \ldots \le f(y_{\nu(n)};\theta)$ *for a fixed* θ, $f(y_i;\theta) = -\log\varphi(y_i;\theta)$, $y_i \in I\!R^q$ *for* $i = 1,\ldots,n$ *are i.i.d. observations with probability density* $\varphi(y;\theta)$, *which depends on an unknown parameter* $\theta \in \Theta^p \subset I\!R^p$, $\nu = (\nu(1),\ldots,\nu(n))$ *is the corresponding permutation of the indices, which depends on* θ, k *is the trimming parameter and the weights* $w_i \ge 0$ *for* $i = 1,\ldots,n$ *are nondecreasing functions of* $f(y_i,\theta)$ *such that at least* $w_{\nu(k)} > 0$.

The basic idea behind trimming in (5.1) is the removal of those $n-k$ observations whose values would be highly unlikely to occur if the fitted model was true. The combinatorial nature of the WTLE is emphasized by the representation

$$\min_{\theta\in\Theta^p} \sum_{i=1}^{k} w_{\nu(i)} f(y_{\nu(i)};\theta) = \min_{\theta\in\Theta^p} \min_{I\in I_k} \sum_{i\in I} w_i f(y_i;\theta) = \min_{I\in I_k} \min_{\theta\in\Theta^p} \sum_{i\in I} w_i f(y_i;\theta),$$

where I_k is the set of all k–subsets of the set $\{1,\ldots,n\}$. Therefore, it follows that all possible $\binom{n}{k}$ partitions of the data have to be fitted by the MLE, and the WTLE is given by the partition with the minimal negative log likelihood.

The WTLE accommodates: (i) the MLE if $k = n$; (ii) the TLE if $w_{\nu(i)} = 1$ for $i = 1,\ldots,k$ and $w_{\nu(i)} = 0$ otherwise; (iii) the Median Likelihood Estimator (MedLE) if $w_{\nu(k)} = 1$ and $w_{\nu(i)} = 0$ for $i \ne k$; If $\varphi(y;\theta)$ is the multivariate normal density function then

the MedLE and TLE coincide with the MVE and MCD estimators Rousseeuw and Leroy (1987). If $\varphi(y;\theta)$ is the normal regression error density, the MedLE and TLE coincide with the LMS and LTS estimators Rousseeuw and Leroy (1987). Details can be found in Vandev and Neykov (1993) and Vandev and Neykov (1998). General conditions for the existence of a solution of (5.1) can be found in Dimova and Neykov (2004) whereas the asymptotic properties are investigated in Čížek (2004). The Breakdown Point (BDP) (i.e. the smallest fraction of contamination that can cause the estimator to take arbitrary large values) of the WTLE is not less than $\frac{1}{n}\min\{n-k, k-d\}$ for some constant d which depends on the density considered, see Müller and Neykov (2003). The choice of d in mixture settings will be discussed in Section 3.

The FAST-TLE algorithm. Computing the WTLE is infeasible for large data sets because of its combinatorial nature. To get an approximative TLE solution an algorithm called FAST-TLE was developed in Neykov and Müller (2003). It reduces to the FAST-LTS and FAST-MCD algorithms considered in Rousseeuw and Van Driessen (1999a,b) in the normal regression or multivariate Gaussian cases, respectively. The basic idea behind the FAST-TLE algorithm consists of carrying out finitely many times a two-step procedure: a trial step followed by a refinement step. In the trial step a subsample of size k^* is selected randomly from the data sample and then the model is fitted to that subsample to get a trial ML estimate. The refinement step is based on the so called concentration procedure: (a) The cases with the k smallest negative log likelihoods based on the current estimate are found, starting with the trial MLE as initial estimator. (Instead of the trial MLE any arbitrarily plausible value can be used.); (b) Fitting the model to these k cases gives an improved fit. Repeating (a) and (b) yields an iterative process. The convergence is always guaranteed after a finite number of steps since there are only finitely many k–subsets out of $\binom{n}{k}$. At the end of this procedure the solution with the lowest trimmed likelihood value is stored. There is no guarantee that this value will be the global minimizer of (5.1) but one can hope that it would be a close approximation to it. The trial subsample size k^* should be greater than or equal to d which is necessary for the existence of the MLE but the chance to get at least one outlier free subsample is larger if $k^* = d$. Any k within

the interval [d, n] can be chosen in the refinement step. A recommendable choice of k is $\lfloor (n+d+1)/2 \rfloor$ because then the BDP of the TLE is maximized according to Müller and Neykov (2003). The algorithm could be accelerated by applying the partitioning and nesting techniques as in Rousseeuw and Van Driessen (1999a) and Rousseeuw and Van Driessen (1999b). We note that if the data set is small all possible subsets with size k can be considered.

5.3 Finite mixtures and robustness

To make the robust approaches in mixture and cluster settings more understandable we will briefly sketch the MLE within these frameworks based on the EM algorithm. For more details see McLachlan and Peel (2000).

The MLE and EM algorithm. Let (y_i, x_i^T) for $i = 1, \ldots, n$ be a sample of i.i.d. observations such that y_i is coming from a mixture of distributions $\psi_1(y_i; x_i, \theta_1), \ldots, \psi_g(y_i; x_i, \theta_g)$ conditional on $x_i \in I\!R^p$, in proportions $\pi_1, \ldots, \pi_g$ defined by

$$\varphi(y_i; x_i, \Psi) = \sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j), \tag{5.2}$$

where $\Psi = (\pi_1, \ldots, \pi_{g-1}, \theta_1, \ldots, \theta_g)^T$ is the unknown parameter vector. The proportions satisfy the conditions $\pi_j > 0$ for $j = 1, \ldots, g$, and $\sum_{j=1}^{g} \pi_j = 1$. The MLE of Ψ is defined as a maximum of the log likelihood

$$\log L(\Psi) = \sum_{i=1}^{n} \log\Big\{ \sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j) \Big\}. \tag{5.3}$$

Under certain assumptions on $\psi_j(y_i; x_i, \theta_j)$ for $j = 1, \ldots, g$ the MLE of Ψ exists and belongs to a compact set. However, the resulting MLE is not reasonable if these assumptions are violated. Usually (5.3) is not maximized directly. The EM algorithm is a standard technique to obtain the MLE of Ψ. It is assumed that each observation (y_i, x_i^T) is associated with an unobserved state $z_i = (z_{i1}, z_{i2}, \ldots, z_{ig})^T$ for $i = 1, \ldots, n$, where z_{ij} is one or zero, depending on whether y_i does or does not belong to the jth

component. Treating (y_i, x_i^T, z_i^T) as a complete observation, its likelihood is given by $P(y_i, x_i, z_i) = P(y_i, x_i|z_i)P(z_i) = \prod_{j=1}^{g} \psi_j(y_i; x_i, \theta_j)^{z_{ij}} \pi_j^{z_{ij}}$. Therefore the complete-data log-likelihood is defined by

$$\log L_c(\Psi) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}. \tag{5.4}$$

Considering the z_{ij} as missing the EM algorithm proceeds iteratively in two steps, called the E-step and M-step for expectation and maximization respectively. The E-step on the $(l+1)$th iteration involves the calculation of the conditional expectation of the complete-data log-likelihood, given the observed data (y_i, x_i^T) and using the current estimate $\Psi^{(l)}$ of Ψ,

$$Q(\Psi; \Psi^{(l)}) = \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; x_i, \Psi^{(l)}) \{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}, \tag{5.5}$$

where $\tau_j(y_i; x_i, \Psi^{(l)}) = \pi_j^{(l)} \psi_j(y_i; x_i, \theta_j^{(l)}) / \sum_{h=1}^{g} \pi_h^{(l)} \psi_h(y_i; x_i, \theta_h^{(l)})$ is the current estimated posterior probability that y_i belongs to the jth mixture component. The function $Q(\Psi; \Psi^{(l)})$ minorizes $\log L(\Psi)$, i.e., $Q(\Psi; \Psi^{(l)}) \leq \log L(\Psi)$ and $Q(\Psi^{(l)}; \Psi^{(l)}) = \log L(\Psi^{(l)})$. The M-step in the $(l+1)$th iteration maximizes $Q(\Psi; \Psi^{(l)})$ with respect to Ψ. This yields a new parameter estimate $\Psi^{(l+1)}$. These two steps are alternately repeated until convergence occurs.

The maximization problem can be simplified as (5.5) can be seen to consist of two parts. The first depends only on the parameters $\pi_1, \ldots, \pi_{g-1}$ whereas the second part depends only on $\theta_1, \ldots, \theta_g$. Consequently, the prior probabilities π_j are updated by

$$\pi_j^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_j(y_i; x_i, \Psi^{(l)}) \tag{5.6}$$

and the expression for θ_j is maximized,

$$\max_{\theta_1, \ldots, \theta_g} \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_j(y_i; x_i, \Psi^{(l)}) \log \psi_j(y_i; x_i, \theta_j), \tag{5.7}$$

considering the posterior probabilities $\tau_j(y_i; x_i, \Psi^{(l)})$ as the prior weights. Under the assumption that θ_j (for $j = 1, \ldots, g$) are distinct a priori, expression (5.7) is maximized for each component separately,

$$\max_{\theta_j} \sum_{i=1}^{n} \tau_j(y_i; x_i, \Psi^{(l)}) \log \psi_j(y_i; x_i, \theta_j), \quad \text{for } j = 1, \ldots, g. \tag{5.8}$$

In case that θ_j are non-distinct, many techniques exist to reformulate (5.7) by single summations, see McLachlan and Peel (2000).

The classification EM algorithm. This approach consists of assigning the observation (y_i, x_i^T) to the hth component if $\tau_h(y_i; x_i, \Psi^{(l)}) \geq \tau_j(y_i; x_i, \Psi^{(l)})$ for $j = 1, \ldots, g$. In case of equal estimated posterior probabilities an observation is assigned arbitrarily to one of the components. Hence instead of (5.8) the following expression is maximized

$$\max_{\theta_j} \sum_{i=1}^{n_j} \log \psi_j(y_i; x_i, \theta_j) \qquad \text{for } j = 1, \ldots, g, \tag{5.9}$$

where n_j is the jth cluster sample size and $n_1 + n_2 + \ldots + n_g = n$. This is a k–means-type algorithm which converges in a finite number of iterations. The resulting estimates are neither MLE nor consistent, see McLachlan and Peel (2000). However, they could be used as starting values in the EM algorithm.

The expressions (5.8) and (5.9) are standard MLE problems. In this way the EM algorithm decomposes complex MLE problems into more simple ones that can be solved by widely available software packages.

The Breakdown Point of the WTLE in mixture settings. As a consequence of the EM algorithm, the BDP of the WTLE in mixture settings can be characterized via the BDP of the trimmed version of (5.5), the trimmed conditional expectation of the complete-data negative log-likelihood estimator

$$\min_{\Psi} \min_{I \in I_k} \sum_{i \in I} \sum_{j=1}^{g} -\tau_j(y_i; x_i, \Psi^{(l)}) \{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}. \tag{5.10}$$

Here only the BDP of the WTLE for the parameters θ_j for $j = 1, \ldots, g$ will be treated because the BDP for $\pi_1, \ldots, \pi_g$ needs special consideration. Therefore the fullness index d of the set $F_\theta = \{\sum_{j=1}^{g} - \log \psi_j(y_i; x_i, \theta_j) \text{ for } i = 1, \ldots, n\}$ has to be characterized using the d–fullness technique of Vandev and Neykov (1993), and Müller and Neykov (2003). It can be proved easily that the fullness index of F_θ is equal to $d = \sum_{j=1}^{g} d_j$ under the assumption that θ_j are distinct a priori and the sets $F_{\theta_j} = \{- \log \psi_j(y_i; x_i, \theta_j) \text{ for } i = 1, \ldots, n\}$ are d_j–full for $j = 1, \ldots, g$. Derivation of the fullness index of any of the

sets F_{θ_j} is a routine task. Consequently, there always exists a solution of the optimization problem (5.10) if k^* and k are within the interval $[\sum_{j=1}^{g} d_j, n]$. If k satisfies $\left\lfloor (n + \sum_{j=1}^{g} d_j)/2 \right\rfloor \leq k \leq \left\lfloor (n + \sum_{j=1}^{g} d_j + 1)/2 \right\rfloor$ the BDP of the WTLE is maximized and equal to $\frac{1}{n} \left\lfloor (n - \sum_{j=1}^{g} d_j)/2 \right\rfloor$. Generally, the fullness index of F_θ is less than the above in case of non-distinct parameters. The fullness indices d_j are equal if $\psi_j(y_i; x_i, \theta_j)$ for $j = 1, \ldots, g$ belong to the same distribution family, e.g., $d_j = p + 1$ in the p-variate normal case Vandev and Neykov (1993). Therefore the BDP of the WTLE in mixtures of p-variate heteroscedastic normals is equal to $\frac{1}{n} \lfloor (n - g(p + 1))/2 \rfloor$. The index of fullness of a mixture of p-variate homoscedastic normals is $g + p$ and thus the BDP of the WTLE in this setting is equal to $\frac{1}{n} \lfloor (n - g - p)/2 \rfloor$. The WTLE reduces to the weighted MCD estimator in both cases if $g = 1$ whereas the BDPs coincide with the BDP of the MCD estimator which is equal to $\frac{1}{n} \lfloor (n - p - 1)/2 \rfloor$. The same holds for mixtures of multiple normal and Poisson regressions with intercept and rank p of the covariates matrix. If the data are not in general position (which is often the case with mixtures of GLMs) this number should be much larger, at least $g(\mathtt{N}(X) + 2)$, see Müller and Neykov (2003) for the definition of $\mathtt{N}(X)$.

Robust fitting of mixtures. If one is able to perform all k–subsets MLE fits of n cases for the mixture model (5.2) then the WTLE could be found. As this is infeasible for large n the FAST-TLE algorithm can be used to get an approximation. The FAST-TLE algorithm is a general approach for robust estimation and thus any MLE procedure for fitting mixtures can be used. However, the usage of the EM algorithm has a number of conceptual advantages. For instance, fitting mixtures of p-variate normals by the FAST-TLE using the classification EM algorithm reduces to the cluster analysis estimation techniques described by Garcia-Escudero et al. (2003), Gallegos and Ritter (2005), and Hardin and Rocke (2004) under the restriction that the covariance matrices are spherical, homoscedastic and heteroscedastic, respectively. FAST-TLE fitting mixture of normal regressions using the classification EM algorithm would coincide with carrying out cluster-wise regression by the FAST-LTS algorithm of Rousseeuw and Van Driessen (1999a).

Generally, other techniques for robust fitting of mixtures or clustering can be derived by

replacing the g standard MLE problems in (5.8) or (5.9) by appropriate g robust estimation problems. This idea was adapted by Campbell (1984) in robustly fitting mixtures of normals involving the M-estimators Huber (1981) of location and scale. The usage of M-estimators for the cluster-wise multiple linear regression case is discussed by Hennig (2003).

Software adjustments of the FAST-TLE to mixtures. Since the trial and refinement steps are standard MLE procedures, the FAST-TLE algorithm can be easily implemented using widely available software. We illustrate this in the framework of mixtures of linear regression models, multivariate heteroscedastic normals, and Poisson regressions using the program FlexMix of Leisch (2004). FlexMix was developed in R (`http:/www.R-project.org`) as a computational engine for fitting arbitrary finite mixture models, in particular, mixtures of GLMs and model-based cluster analysis by using the EM algorithm.

In the mixture setting with g components, the trial sample size k^* must be at least $g(p+1)$ to overcome the degenerated case of unbounded likelihood. Thus we recommend a larger trial sample size to increase the chance to allocate at least $p+1$ cases to each mixture component. If this is not the case, any program would fail to get an estimate that could serve as a trial estimate. If this happens a new random subsample of k^* observations has to be drawn and supplied to the software estimation procedure. This trial and error process continues until a trial estimate is derived. The refinement subsample size k has to be $\lfloor (n+g(p+1))/2 \rfloor$ to ensure the highest BDP of the TLE. If the expected percentage α of outliers in the data is a priory known, a recommendable choice of k is $\lfloor n(1-\alpha) \rfloor$ to increase the efficiency of the TLE.

Most of the software procedures for fitting mixtures, in particular the FlexMix program, maximize the expression (5.8) or (5.9) according to the user specified weight option. For instance, if the hard weighting option is specified then the classification EM algorithm is performed by FlexMix. We recommend this option within the trial step only. Hence depending on the weight option various algorithms can be designed.

As a final remark we note that in the refinement steps the negative log likelihoods $-\log \varphi(y_i; x_i, \Psi)$ defined by (5.2) are evaluated at the current estimate $\hat{\Psi}$ and then sorted

in ascending order to get the indices of those k cases with the smallest negative log-likelihoods, starting with the trial estimate Ψ^* of Ψ at the first iteration of the refinement step. In practice, we need 4 or 5 refinement steps at most to reach convergence.

5.4 Examples

In the examples below we compare MLE and FAST-TLE approaches using the program FlexMix as a computational MLE and FAST-TLE procedure. Sometimes FlexMix returns less components than initially specified. This is because FlexMix allows a removal of components containing less observations than a user specified percentage to overcome numerical instabilities. Since the true number of mixture components is unknown in practice, FlexMix is always run with various numbers of components. The Bayesian Information Criterion (BIC) based on the MLE and FAST-TLE can then be used to determine the number of mixture components. In this way we can assess the quality of the fits as in our examples the number of components and their parameters are known. A fit is considered as successful if *all* components are correctly estimated even if some non-sense fits occur additionally. Correct estimation means that at least 95% of the observations that are assigned to a particular component are indeed generated from this model.

Mixture of three regression lines with noise

In this example we consider a mixture of three simple normal linear regressions with additional noise. The regression lines were generated according to the models $y_{1i} = 3+1.4x_i+\epsilon_i$ (70 data points), $y_{2i} = 3-1.1x_i+\epsilon_i$ (70 data points), and $y_{3i} = 0.1x_i+\epsilon_i$ (60 data points), where x_i is uniformly distributed in the intervals [-3,-1] and [1,3], respectively, and ϵ_i is a standard normal distribution with $\sigma = 0.1$. To these data we added 50 outliers uniformly distributed in the area $[-4.5, 4.5] \times [-0.8, 2.8]$. The points that follow the models are marked by rhombs, squares and bullets whereas the outliers are marked by triangles. The plots in Figure 5.1 are typical results of the MLE and FAST-TLE fits. The dotted, dashed and solid lines correspond to the true models, MLE and FAST-TLE fits, respec-

Figure 5.1: Mixture of three regressions: true model (dotted lines), fits based on the MLE (dashed lines) and FAST-TLE (solid lines) with (a) 20% trimming and 3 components, (b) 40% trimming and 4 components.

tively. Starting with an increasing percentage of trimming from 20 to 45 and number of components from 2 to 5 the FAST-TLE algorithm converged to the correct two components mixture model in almost all trials whereas the MLE failed.

Mixture of three bivariate normal models with noise

By this example the behavior of the FAST-TLE is studied for the simulated data set discussed in McLachlan and Peel (2000). This data consists of 100 observations generated from a 3-component bivariate normal mixture model with equal mixing proportions and component parameters, respectively as

$$\mu_1 = (0 \quad 3)^T, \quad \mu_2 = (3 \quad 0)^T, \quad \mu_3 = (-3 \quad 0)^T,$$

$$\Sigma_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & .5 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} .1 & 0 \\ 0 & .1 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & .5 \end{pmatrix}.$$

Fifty outliers, generated from a uniform distribution over the range -10 to 10 on each variate are added to the original data. Thus a sample of 150 observations is obtained. McLachlan and Peel (2000) model this data by a mixture of t–distributions and reduce the influence of the outliers.

The original observations, the outliers, the 3 components MLE and FAST-TLE fits with 15%, 25%, 35% and 45% trimming are presented in Figure 5.2 (a)–(d). The original observations, i.e., data that follow the models are marked by rhombs, squares and bullets whereas the outliers are marked by triangles. The dotted contours of the ellipses on the plots correspond to the true models whereas the solid and dashed contours of the 99% confidence ellipses correspond to the FAST-TLE and MLE fits, respectively. For the robust fits we can see that a lower or higher trimming percentage than the true contamination

Figure 5.2: Data set of McLachlan and Peel (2000) with mixtures of 3 normals with noise: true model (dotted lines) and fits of a three component normal mixture based on the MLE (dashed lines) and FAST-TLE (solid lines) with (a) 15%, (b) 25%, (c) 35%, and (d) 45% of trimming.

level still allows the correct estimation of the ellipsoid centers while the covariances are overestimated or underestimated due to the too high or low trimming percentage. The fits are excellent if the specified trimming percentage is close to the true percentage of contamination. The classical MLE fits are poor when using 3 or even more components.

Generally, in real applications the number of mixture components is unknown and the BIC is widely used for model assessment. The trimmed analog of BIC is defined by $\text{TBIC} = -2\log(TL_k(\tilde{\Psi})) + m\log(k)$, where $TL_k(\tilde{\Psi})$ is the maximized trimmed likelihood, k is the trimming parameter, and m is the number of parameters in the mixture model. Obviously, TBIC reduces to BIC if $k = n$. To get an impression of the empirical distribution of these quantities for this example a limited Monte Carlo simulation study was conducted for a range of different situations. We fit the simulated three bivariate mixtures of normals with 1 to 5 components and vary the trimming percentage from 0% to 45% in steps of 5%. The experiment was independently replicated 500 times for any combination. The resulting TBIC median values (rounded) are presented in Table 5.1. The smallest values for each column are marked in *italics*. One can see that these values stabilize in a model with 3 components which is the correct model. A two-phase regression fit of the 3rd row values against the trimming percentages detects a change point between 25% and 30% trimming which could be interpreted as a data contamination estimate. We note that the true contamination level in this data set is slightly higher, however, a part of the noise observations conforms the mixture model. From this and other similar studies we could conclude that the TBIC might be used to assess robustly the number of mixture

Table 5.1: Simulation experiment for the data of McLachlan and Peel (2000): resulting TBIC median values (rounded) based on different numbers of components (rows) and different trimming percentages (columns).

	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%
1	1672	1510	1382	1253	1119	1003	915	837	749	650
2	1654	1494	1338	1202	1054	920	822	734	643	559
3	1585	1436	1313	1190	*1047*	*902*	*795*	*709*	*620*	*538*
4	1595	*1429*	*1304*	*1178*	1040	908	807	720	631	549
5	*1594*	1430	1309	1184	1051	922	822	736	647	566

components and the percentage of contamination in the data.

Mixture of two Poisson regression models with noise

In this example we consider two Poisson regression models with equal mixing proportions and with additional noise. For each Poisson regression model 100 data points are generated according to the Poisson distribution with means $\log \lambda_1 = 3 - 0.008x$ and $\log \lambda_2 = 3 + 0.008x$, where x is uniformly distributed in the intervals [-225,-25] and [25,225], respectively. For the noise we generated 50 points from a uniform distribution over the range of each variate. The plots in Figure 5.3 are typical results of the MLE and FAST-TLE fits for a simulated data set. The points that follow the models are marked by squares and rhombs whereas the outliers are marked by triangles. The dotted, dashed and solid lines correspond to the true models, MLE and TLE fits, respectively. Starting with an increasing number of components from 2 to 5 the FAST-TLE algorithm converged to the correct two components mixture model in most of the trials whereas the MLE failed, see Figure 5.3.

In order to get more insight we generated 100 independent data samples according to the above model. Each data set was fitted by a mixture model with 2, 3, 4 and 5 components and with 20% trimming. Similar to the previous examples, the estimated number of components as returned by FlexMix can be smaller than initially specified. For

Figure 5.3: Mixture of two Poisson regression components: true model (dotted lines), fits based on the MLE (dashed lines) and FAST-TLE (solid lines) with (a) 20% trimming and 2 components, and (b) 40% trimming and 4 components.

each considered model we count how often a model with a certain number of components is returned among all simulated data sets. The results for the MLE and FAST-TLE are reported in Table 5.2. The number of specified components is presented by the rows in the table, and the number of returned components by the columns. Additionally to the frequencies we provide the number of successful fits (number below in *italics*), i.e., both Poisson regression components of the mixture model were correctly estimated. For the MLE method we see that the chance for successful fits increases only for a larger required number components. Overall, the method has severe problems in estimating the models since only 37 out of 400 fits were successful. For FAST-TLE the increase of the initial number of components has almost no effect, since a model with 2 components is optimal in more than 90% of the fits. Moreover, these models are almost always successful fits. In total, 392 out of the 400 experiments were successful.

5.5 Summary and conclusions

The TLE methodology can be used for robustly fitting mixture models. We have demonstrated by examples and simulations that in presence of outliers the TLE gives very reliable estimates comparable to the mixture model generating parameters. Applying the FAST-TLE algorithm to mixtures boils down to carrying out the classical MLE on subsamples. Procedures for mixture models based on the MLE are widely available and thus the method is easy to implement. Software in R is available from the authors upon request. The TBIC is a useful indicator for determining the number of components and contamination level in the data. If the trimming percentage is chosen too large, some of the observations that follow the model will be trimmed and incorrectly identified as outliers. Therefore an addi-

Table 5.2: Simulation results for the mixture of two Poisson regressions. Models with 2, 3, 4, and 5 components were fitted for 100 simulated data sets. Out of 400 fits, 37 were successful for MLE and 392 were correctly estimated by FAST-TLE.

	MLE returned components					FAST-TLE returned components				
started	2	3	4	5	Total	2	3	4	5	Total
2	100				100	100				100
	1				*1*	*98*				*98*
3		100			100	93	7			100
		2			*2*	*92*	*7*			*99*
4		94	6		100	96	4	0		100
		4	*4*		*8*	*94*	*4*			*98*
5		19	15	66	100	94	6		0	100
		3	*7*	*16*	*26*	*91*	*6*			*97*
Total	100	213	21	66	400	383	7	0	0	400
	1	*9*	*11*	*16*	*37*	*375*	*17*			*392*

tional inspection of the FAST-TLE posterior weights can be helpful in distinguishing these observations from real outliers. The TLE will lead to greater computational effort, but having in mind the growing power of modern-day processors and memory, one can afford this.

Chapter 6

Robust joint modeling of mean and dispersion through trimming

Summary. The Maximum Likelihood Estimator (MLE) and Extended Quasi-Likelihood (EQL) estimator have commonly been used to estimate the unknown parameters within the joint modeling of mean and dispersion framework. However, these estimators can be very sensitive to outliers in the data. In order to overcome this disadvantage, the usage of the maximum Trimmed Likelihood Estimator (TLE) and the maximum Extended Trimmed Quasi-Likelihood (ETQL) estimator is recommended to estimate the unknown parameters in a robust way. The superiority of these approaches in comparison with the MLE and EQL estimator is illustrated by an example and a simulation study. As a prominent measure of robustness, the finite sample Breakdown Point (BDP) of these estimators is characterized in this setting.

6.1 Introduction

Let y_i be independently observed responses with means μ_i and known variance function $V(\mu_i)$, for $i = 1, \ldots, n$. Nelder and Pregibon (1987) consider a general quasi-likelihood

model

$$g(\mu_i) = x_i^T\beta, \qquad h(\phi_i) = z_i^T\lambda \qquad \text{and} \qquad \text{var}(y_i) = \phi_i V(\mu_i), \tag{6.1}$$

where ϕ_i is the dispersion parameter, g and h are known monotonic differentiable link functions, x_i and z_i are the covariate vectors of dimensions p and q affecting the means and dispersions, and β and λ are vectors of unknown regression parameters, respectively. The linear exponential family of distributions with a known constant $\phi_i = \phi$ is a special case of this general setting. The widely used over-dispersed Poisson distribution with $\text{var}(y_i) = \phi\mu_i$ and binomial distribution with $\text{var}(y_i) = \phi\mu_i(1 - \mu_i/m_i)$ and trial number m_i are also accommodated by this model.

For joint inferences on β and λ, Nelder and Pregibon (1987) suggest to maximize an extended quasi-likelihood (EQL) (strictly extended quasi-log-likelihood) function

$$Q^+(\beta,\lambda) = \sum_{i=1}^{n} q^+(y_i;\mu_i(\beta),\phi_i(\lambda)) = \sum_{i=1}^{n} q^+(y_i;\mu_i,\phi_i) \tag{6.2}$$

$$= \sum_{i=1}^{n} -\frac{1}{2}\left\{\log\left[2\pi\phi_i V(y_i)\right] + \frac{d_i}{\phi_i}\right\}, \tag{6.3}$$

in which $d_i \equiv d(y_i;\mu_i) = -2\int_{y_i}^{\mu_i} \frac{y_i-u}{V(u)}du$ denotes the individual deviance function corresponding to $V(\mu_i)$.

The EQL is an approximate log-likelihood which is exact in the normal, inverse Gaussian and gamma cases (Smyth, 1989). Therefore the Maximum Likelihood Estimation (MLE) can be employed as an estimation criterion for these distributions. Actually, the EQL is not a proper density, but a distribution can be derived by suitably normalizing it. Nelder and Pregibon (1987) proposed using the unnormalized EQL due to convenience in implementation. The EQL does not require full distributional assumption, only specification of the form of the first two moments. In many cases this provides a greater flexibility within the statistical modeling framework eliminating the necessity of specifying the full distribution for the data. However, if an exponential dispersion family with a variance function $V(\mu)$ exists then the EQL is the log-likelihood function based on a saddlepoint approximation to that family (McCullagh and Nelder, 1989; Jørgensen, 1997). A related

approach to the EQL proposed by Efron (1986) is based on the double-exponential family of distributions. Lee and Nelder (2000) notice that the unnormalized EQL and Efron's (1986) unnormalized double-exponential family are equivalent up to some constant terms and therefore both approaches lead to identical inferences.

Equation (6.3) shows that the quasi-likelihood estimator $\hat{\beta}$ of β can be found by minimizing the deviance function $\sum_{i=1}^{n} d_i$ instead of maximizing directly $Q^+(\beta, \lambda)$. Then the quasi-likelihood estimator $\hat{\lambda}$ of λ can be obtained by using $\hat{\mu}_i = \mu_i(\hat{\beta})$. The parameters ϕ_i and μ_i are orthogonal in the sense of Cox and Reid (1987) as $E\left(\partial^2 Q^+ / \partial\mu_i \partial\phi_i\right) = 0$ and this implies orthogonality between β and λ. Therefore the optimization of the $p+q$ dimensional problem reduces to two separate optimization problems of dimensions p and q. As a consequence, the unknown parameters β and λ can be estimated by alternating between two GLMs, a standard and a gamma,

$$E(y_i) = \mu_i \qquad g(\mu_i) = \eta_i = x_i^T \beta \qquad \text{var}(y_i) = \phi_i V(\mu_i) \tag{6.4}$$

$$E(d_i) = \phi_i \qquad h(\phi_i) = \xi_i = z_i^T \lambda \qquad \text{var}(d_i) = 2\phi_i^2. \tag{6.5}$$

Setting the initial value for ϕ_i to a constant, the mean model (6.4) produces the deviances d_i as responses for the dispersion model (6.5) with dispersion parameter 2 and log-link function h, which in turn produces the prior weights $1/\phi_i$ for the mean model (6.4). This alternation process continues until convergence is reached, see Smyth (1989) for a comprehensive exposition. McCullagh and Nelder (1989) referred to this procedure as "joint modeling of mean and dispersion".

In order to reduce the bias in estimating the dispersion parameters, when the number of mean parameters is relatively large compared to sample size, Lee and Nelder (1998) recommend using adjusted deviances $d_i^* = d_i/(1-\rho_{ii})$ as responses instead of d_i and prior weights $1-\rho_{ii}$ in the dispersion model (6.5), where ρ_{ii} is the ith diagonal element of the projection matrix of the mean model (6.4) (Smyth and Verbyla, 1999; Lee et al., 2006). The proposed modification is called restricted maximum likelihood (REML) adjustment algorithm. It provides the MLE and REML estimators for β and ϕ, respectively, in case of normal models with non-homogeneous errors. Details about estimation adjustments can

be found in McCullagh and Nelder (1989), Smyth (1989), Smyth and Verbyla (1999), Lee and Nelder (2000), and Lee et al. (2006).

From a computational point of view, (Green, 1984), this is equivalent to finding ML or quasi-likelihood estimates of β and λ by solving iteratively the following two interlinked weighted least squares problems:

$$\min_{\beta} (u_m - X\beta)^T W_m (u_m - X\beta) \tag{6.6}$$

$$\min_{\lambda} (u_{d^*} - Z\lambda)^T W_{d^*} (u_{d^*} - Z\lambda), \tag{6.7}$$

where X and Z are the $n \times p$ and $n \times q$ matrices of covariates, u_m and u_{d^*} are the mean and dispersion adjusted dependent variable vectors with elements $u_{m,i} = x_i^T\beta + \frac{\partial \eta_i}{\partial \mu_i}(y_i - \mu_i)$ and $u_{d^*,i} = z_i^T\lambda + \frac{\partial \xi_i}{\partial \phi_i}(y_i - \phi_i)$, and $W_m = \text{diag}\left((\phi_i(\partial\eta_i/\partial\mu_i)^2 V(\mu_i))^{-1}\right)$ and $W_{d^*} = \text{diag}\left((2(1-\rho_i)\phi_i^2(\partial\xi_i/\partial\phi_i)^2)^{-1}\right)$ are the working weight matrices, ρ_{ii} is the ith diagonal element of the matrix $W_m^{1/2}X(X^TW_mX)X^TW_m^{1/2}$, and all these elements are evaluated at the current estimates of β and λ. More precisely, holding λ fixed at the current estimate $\hat{\lambda}$ at each iteration, W_m and u_m are updated and (6.6) is solved again for β until convergence. Similarly, holding β fixed at the current estimate $\hat{\beta}$, at each iteration, W_d^* and v_d^* are updated and (6.7) is solved again for λ until convergence. Cycling between these two Iteratively Reweighted Least Squares (IRLS) algorithms until convergence results in the EQL estimates of β and λ. Thus a standard linear regression routine can be adapted to calculate $\hat{\beta}$ and $\hat{\lambda}$ via an IRLS algorithm.

The use of EQL provides a greater flexibility of the GLMs modeling, and the availability of software such as the R packages *dglm*, *JointModeling*, *statmod* and *tweedie*, facilitate and enlarge its applicability. Information on these R packages is given in Smyth (2009a), Ribatet and Iooss (2009), and Smyth (2009b).

Unfortunately, the MLE and EQL estimator can be highly sensitive to a small proportion of observations that departs from the model, (Hampel at al., 1986). The non-robustness of the MLE and quasi-likelihood estimators against outliers within the single GLM has been studied extensively in the literature, e.g., Markatou et al. (1997), Cantoni and Ronchetti (2001), Müller and Neykov (2003), Maronna et al. (2006) and the references

therein.

In this chapter we consider robust estimation for joint modeling of the mean and dispersion trough trimming in order to reduce the influence of outliers. The paper is organized as follows. In Section 2 we recall the weighted Generalized Trimmed Estimator (wGTE), we define the maximum Extended Trimmed Quasi Likelihood (ETQL) estimator and discuss its breakdown property. In Section 3 an approximate computational procedure for the wGTE optimization is proposed. Section 4 compares the behavior of classical and robust estimation on a simple data example. In Section 5 a simulation study is performed to illustrate the effectiveness of the proposed estimator in comparison with the EQL. Finally, conclusions are given in Section 7.

6.2 Maximum extended trimmed quasi-likelihood estimator

The definition of the weighted Generalized Trimmed Estimator (wGTE) given by Vandev and Neykov (1998) is as follows. Let $f_i : \Theta \to \mathbb{R}^+$, where $\Theta \subseteq \mathbb{R}^q$ be an open set and $F = \{f_i(\theta) \text{ for } i = 1, \ldots, n\}$ be d–full. According to Vandev and Neykov (1993), the set F is called d–full if for any subset of cardinality d of F, the supremum of this subset is a subcompact function. A real valued function $\varphi(\theta)$ is called subcompact if the sets $L_{\varphi(\theta)}(C) = \{\theta : \varphi(\theta) \leq C\}$ are contained in a compact set for any constant C.

Definition 6.1 *The wGTE, $\hat{\theta}^{\mathrm{k}}_{\mathrm{wGTE}}$, of θ is defined as the solution of the optimization problem*

$$\hat{\theta}^{\mathrm{k}}_{\mathrm{wGTE}} := \arg\min_{\theta\in\Theta} \left\{ S(\theta) = \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta) \right\}, \tag{6.8}$$

where $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \ldots \leq f_{\nu(n)}(\theta)$ are the ordered values of f_i at θ and $\nu = (\nu(1), \ldots, \nu(n))$ is the corresponding permutation of the indices, which depends on θ, $k \leq n$. The weights $w_i = w(f_i(\theta)) \geq 0$ for $i = 1, \ldots, n$ are such that $w_{\nu(k)} > 0$, and $w(.)$ is a non-negative decreasing function.

The trimming parameter k determines the robustness properties of the wGTE as those $n-k$ functions $f_i(\theta)$ with the largest values are excluded from the objective function (6.8). The combinatorial nature of the optimization problem is emphasized by the representation

$$\min_{\theta\in\Theta^p} S(\theta) = \min_{\theta\in\Theta^p} \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta) \quad = \quad \min_{\theta\in\Theta^p} \min_{I\in I_k} \sum_{i\in I} w_i f_i(\theta) \tag{6.9}$$

$$= \quad \min_{I\in I_k} \min_{\theta\in\Theta^p} \sum_{i\in I} w_i f_i(\theta), \tag{6.10}$$

where I_k is the set of all k–subsets of the set $\{1,\ldots,n\}$. Therefore, it follows that all possible $\binom{n}{k}$ partitions of the set $\{f_1,\ldots,f_n\}$ have to be considered and $\hat{\theta}^{\mathrm{k}}_{\mathrm{wGTE}}$ is defined by the partition with the minimal value of $S(\theta)$. An exact computation of the wGTE is not feasible for large data sets and therefore an approximation is proposed below.

The wGTE accommodates many statistical estimators. For instance, it reduces to the Least Trimmed Squares (LTS) estimator of Rousseeuw (1984) if the set F is comprised of the squared linear regression residuals and the weights are defined by $w_{\nu(i)} = w(f_{\nu(i)}(\hat{\theta}) \leq f_{\nu(k)}(\hat{\theta})) = 1$, for $i \leq k$, and otherwise 0. Similarly, the maximum Trimmed Likelihood Estimator (TLE) of Neykov and Neytchev (1990) is derived if F is comprised of the negative log-likelihoods. The finite sample breakdown point (BDP) of the wGTE which is a global measure of robustness of a statistical estimator is characterized by Theorem 1 of Vandev and Neykov (1998). Roughly speaking, the BDP is the smallest fraction of contamination that can cause the estimator to take arbitrary large values. The BDP of the wGTE is not less than $\frac{1}{n}\min\{n-k, k-d\}$ if F is d–full. This BDP is maximized for $\lfloor\{n+d+1\}/2\rfloor \leq k \leq \lfloor\{n+d+2\}/2\rfloor$ when it approximately equals $1/2$ for large n, where the notation $\lfloor a\rfloor$ stands for the largest integer less than or equal to a. Therefore selecting the value of k properly one can control the level of robustness of the wGTE. We note that the d-fullness index ensures the existence of a solution and provides positive BDP of the optimization problem (6.8) at any subset of d functions. See Müller and Neykov (2003), and Dimova and Neykov (2004) for a general treatment. Further, the asymptotic properties of the wGTE were studied by Čížek (2008) for the case of twice differentiable functions f.

Let $\theta = (\beta, \lambda)$ and replace $f_i(\theta) := f_i(\beta, \lambda) = -q^+(y_i; \mu_i(\beta), \phi_i(\lambda))$ in (6.9). Then we obtain a particular case of a wGTE which we will call the maximum Extended Trimmed Quasi-Likelihood (ETQL) estimator.

Definition 6.2 *The maximum ETQL estimator* $(\hat{\beta}, \hat{\lambda})$ *of* (β, λ) *is defined as*

$$\max_{\beta,\lambda} Q^+_{\text{trim}}(\beta, \lambda) = \max_{\beta,\lambda} \max_{I \in I_k} \sum_{i \in I} q^+(y_i; \mu_i, \phi_i) \tag{6.11}$$

$$= \max_{I \in I_k} \max_{\beta,\lambda} \sum_{i \in I} q^+(y_i; \mu_i, \phi_i). \tag{6.12}$$

The maximum ETQL estimate is thus the EQL estimate calculated from some k-subset of the n cases. Therefore for all k-subsets the two interlinked GLMs given by (6.4) and (6.5) have to be solved simultaneously and the ETQL estimates $(\hat{\beta}, \hat{\lambda})$ of (β, λ) is defined by that k-subset with the maximal value of (6.11). This means that those $n - k$ observations with the largest deviance residuals are excluded from the loss function. Consequently, the finite sample BDP of the maximum ETQL estimator can be derived as the lower finite sample BDP of these two interconnected GLMs. Thus we have to determine the fullness indices of the negative log-likelihoods sets of both GLMs and then the finite sample BDP can be exemplified by the range of values of k (Vandev and Neykov, 1998; Müller and Neykov, 2003). Because the negative log-likelihoods of the two GLMs (6.4) and (6.5) are proportional to their corresponding unit deviance functions it is more convenient to determine the fullness indices of these latest quantities. For fixed λ, the unit deviances $d(y_i, \mu_i)$ for $i = 1, \ldots, n$ are convex functions in both arguments (Jørgensen, 1997, p. 24-25, 49-50) and thus subcompact functions in μ_i. Similarly, for fixed β, we can conclude that the dispersion gamma GLMs (6.5) unit deviances are subcompact functions in ϕ_i as well. A direct prove follows easily. Indeed, denote by $d_\gamma(d_i, \phi_i) = 2(d_i/\phi_i + \log(\phi_i/d_i) - 1)$ the gamma dispersion GLMs unit deviance. Its limit behavior with respect to the boundary points is $\lim_{\phi_i \to \infty} d_\gamma(d_i, \phi_i) = \lim_{\phi_i \to 0} d_\gamma(d_i, \phi_i) = +\infty$. Hence $d_\gamma(d_i, \phi_i)$ is subcompact function in ϕ_i for $i = 1, \ldots, n$ according to Lemma 4.1 of Dimova and Neykov (2004). Therefore the sets of unit deviances of (6.4) and (6.5) are $\mathcal{N}(X) + 1$ and $\mathcal{N}(Z) + 1$ full, respectively, according to Theorem 3 of Müller and Neykov (2003), where $\mathcal{N}(X) = \max_{0 \neq \beta \in \mathbb{R}^p} \text{card}\{i \in$

$\{1,\ldots,m\}; x_i^T\beta = 0\}$ provides the maximum number of covariates, explanatory variables, $x_i \in R^p$ lying in a subspace, the meaning of $\mathcal{N}(Z)$ is the same. If the observations x_i^T, respectively z_i^T, are linearly independent then $\mathcal{N}(X) = p-1$, $\mathcal{N}(Z) = q-1$, and these are the minimal values for $\mathcal{N}(X)$ and $\mathcal{N}(Z)$. If the covariates are qualitative variables such as factors with several levels, then $\mathcal{N}(X)$ and $\mathcal{N}(Z)$ are much larger. Thus the quantity $\max(\mathcal{N}(X),\mathcal{N}(Z))+1$ determines the minimal number of observations that ensure the existence of solutions of the interlinked GLMs (6.4) and (6.5) with positive BDPs. Hence, the finite sample BDPs of the mean and dispersion GLMs estimators equal to $\min\{n-k, k-\mathcal{N}(X)-1\}/n$ and $\min\{n-k, k-\mathcal{N}(Z)-1\}/n$ according to Müller and Neykov (2003). Therefore we have the following

Theorem 6.1 *The finite sample BDP of the maximum ETQL estimator equals*

$$\frac{1}{n}\min\{n-k, k-\max[\mathcal{N}(X),\mathcal{N}(Z)]-1\}$$

and attains its maximum at

$$\lfloor\{n+\max[\mathcal{N}(X),\mathcal{N}(Z)]+1\}/2\rfloor \le k \le \lfloor\{n+\max[\mathcal{N}(X),\mathcal{N}(Z)]+2\}/2\rfloor$$

which equals to $\frac{1}{n}\lfloor\{n-\max[\mathcal{N}(X),\mathcal{N}(Z)]-1\}/2\rfloor$*, where* $\lfloor a\rfloor$ *stands for the largest integer less than or equal to* a*.*

Note that Nelder and Pregibon (1987) warn that aliasing of the parameters could occur when $Z = X$ is used in modeling both mean and dispersion. This problem might occur also with the ETQL estimator because it is the EQL estimate calculated from some k-subset of the n cases.

6.3 Computational procedure for the wGTE

We propose a computational algorithm to determine an approximate solution of the wGTE. In order to ensure the existence of a solution to the optimization problem (6.9), we assume that the set F is d–full and $k \ge d$. Then the algorithm consists of carrying out finitely many times a two-step procedure of a trial step followed by a refinement step:

Trial step:

1. Let $F^{old} = \{f_{i_1}(\theta), \ldots, f_{i_l}(\theta)\} \subset F = \{f_1(\theta), \ldots, f_n(\theta)\}$ where $l \geq d$;
2. Let $\widehat{\theta}^{old}$ be arbitrary or the minimizer of $\sum_{j=1}^{l} f_{i_j}(\theta)$;

Refinement step:

3. Let $F^{new} = \{f_{\nu(1)}(\theta), \ldots, f_{\nu(k)}(\theta)\} \subset F$ where $f_{\nu(1)}(\widehat{\theta}^{old}) \leq \ldots \leq f_{\nu(n)}(\widehat{\theta}^{old})$ be the sorted values $f_i(\widehat{\theta}^{old})$ for $i = 1, \ldots, n$;
4. Let $\widehat{\theta}^{new}$ be the minimizer of $S(\theta) = \sum_{i=1}^{k} f_{\nu(i)}(\theta)$ where $f_{\nu(i)} \in F^{new}$ for $i = 1, \ldots, k$;
5. Let $\widehat{\theta}^{old} := \widehat{\theta}^{new}$;
6. Cycle steps 3 to 5, until convergence or a finite number of cycles is reached.

Proposition 6.1 *On the basis of steps 3 and 4* $S(\widehat{\theta}^{new}) \leq S(\widehat{\theta}^{old})$.

Proof of Proposition 6.1. From the definition of $\widehat{\theta}^{old}$ and $\widehat{\theta}^{new}$ it follows that

$$S(\widehat{\theta}^{new}) = \sum_{i=1}^{k} f_{\nu(i)}(\widehat{\theta}^{new}) \leq \sum_{i=1}^{k} f_{\nu(i)}(\widehat{\theta}^{old}) = S(\widehat{\theta}^{old}).\square$$

Clearly, the convergence is guaranteed after a finite number of steps since there are only finitely many k–subsets out of $\binom{n}{k}$ in all. We note that this is only a necessary condition for a global minimum of the wGTE objective function. Actually, we will be using the suggestion made by Rousseeuw and Van Driessen (1999b) *Take many initial choices of* F^{old} *and apply the refinement step to each until convergence, and keep the solution with lowest value of* $S(\theta)$ *of (6.8).* There is no guarantee that the achieved solution will be the global minimizer of (6.8) but according to our experiments the approximation is sufficiently good.

An important issue is the choice of the sets F^{old} for starting the algorithm. When the data set is small, all possible subsets with the default size k can be considered. If the cardinality of F is large, one can randomly partition F in a representative way into several

non-overlapping subsets $F_1, \ldots, F_m$ of size $n^* \approx n/m$. The trimming parameter k^* for any of these subsets can be chosen within the interval $[d, n^*]$. A recommended choice for k^* is within the interval $[d, \lfloor (n^* + d + 1)/2 \rfloor]$ to guarantee a positive BDP of the estimators. However, following the same reasoning as in Rousseeuw and Van Driessen (1999b), and because $\widehat{\theta}^{old}$ can be arbitrary, one could draw subsamples with a smaller size $k^{**} := d$ as the chance to get at least one outlier free subsample is larger. In case of data replications, k^{**} must be much larger than d (Müller and Neykov, 2003). Thus within the trial steps the initial estimate must be based on k^{**} whereas within the refinement step the trimming parameter must be $k^* = \lfloor (n^*+d+1)/2 \rfloor$ in order to maximize their BDP. As a consequence of the computational procedure of the GTE applied to each of the subsets $F_1, \ldots, F_m$, the optimal subsets $F^{new}_{opt(1)}, \ldots, F^{new}_{opt(m)}$ each of cardinality k^* are obtained. We remind that the trial and refinement steps are performed finitely many times in order to obtain these optimal subsets. Pooling the sets into $F^{old}_{pooled} = F^{new}_{opt(1)} \cup \ldots \cup F^{new}_{opt(m)}$ with cardinality mk^* we can compute a reliable initial estimate $\widehat{\theta}^{old}_{pooled}$ for the refinement step over F with an optimal trimming parameter $k = \lfloor (n + d + 1)/2 \rfloor$. In this way an approximate GTE and a subset F^{final}_{opt} with cardinality k are obtained.

One can recycle this procedure g times. As a consequence, g pooled sets $F^{final}_{opt(1)}, \ldots, F^{final}_{opt(g)}$, each of cardinality $k = \lfloor (n + d + 1)/2 \rfloor$ would be obtained. Obviously, one must expect a large overlap between these g sets. Pooling these sets into a merged set F^{old}_{merged} with cardinality $k_{trim} > k$ we can get a reliable initial $\widehat{\theta}^{old}_{merged}$ estimate for the last refinement step over F. In this way an approximate GTE, $\widehat{\theta}^{k_{trim}}_{GTE}$ of θ and the corresponding subset $F^{final} \subset F$ with cardinality k_{trim} can be obtained. This final approximate GTE would possess a BDP less than the highest, however, it would be more efficient as $k_{trim} \geq k$. On the other hand, the number of times each observation entered the optimal subsamples $F^{final}_{opt(i)}$ for $i = 1, \ldots, g$ could serve as a self control in designing the subset F^{old}_{merged}. Clearly, a preference would be given to those observations with a relatively higher frequency of inclusion.

Finally, the remaining $n - k$, respectively $n - k_{trim}$, observations that are dropped out of F could be treated as outlying and need additional consideration. Special attention

should be given to those cases with the lowest percentage of inclusion.

We note that particular cases of the "refinement step" procedure have been developed for the computational needs of various high BDP estimators: (i) the concentration steps considered by Visek (1996), Rousseeuw and van Driessen(1999a), and Hawkins and Olive (2002) within the linear LTS regression estimator, and Hawkins and Khan (2009) within the nonlinear LTS regression estimator; (ii) the concentration steps proposed by Rousseeuw and van Driessen(1999b), and Herwindiati et al. (2007) within the multivariate location and scale Minimum Covariance Determinant and Minimum Vector Variance Estimators framework; (iii) the concentration steps discussed by Neykov and Müller (2003), Gallegos and Ritter (2005), Neykov et al. (2007), Garcia-Escudero et al. (2008), Cuesta-Albertos et al. (2008), and Gallegos and Ritter (2010) within the trimmed likelihood and classification trimmed likelihood estimators framework. In all these considerations the corresponding set F of functions is comprised of regression residuals, various multivariate distances and negative log likelihoods.

6.4 Example

As an illustrative example we consider a data set of Zuliani et al. (1983) that has also been used by Smyth and Verbyla (1999). The data are available at `http://www.statsci.org/data/general/blood` and they contain the age, weight (kg) and blood CPK (creatine phosphokinase) concentrations of 18 cross country skiers. The skiers are participants in a 24 hour cross-country relay. The blood CPK concentration was recorded 12 hours into the relay. The CPK is an enzyme contained in muscle cells which is necessary for the storage and release of energy. Leakage of the enzyme CPK into the blood is a symptom of muscle stress.

Examining the relationship of the log-CPK concentrations and the age of each skier, Smyth and Verbyla (1999) detect a decreasing linear trend, and a decreasing variability with increasing age. Instead of stabilizing the variance via transformation they fit the blood CPK concentrations to the age directly by a double generalized linear gamma model with a log-link. Smyth and Verbyla (1999) only used the age variable and not the information

of the weight of the skiers. As a result, the age variable is significant in both the mean and the dispersion model. Here we additionally use the weight variable in the mean model. Figure 6.1 (left column) shows the (condensed) output of the statistical analysis, using the function *fitjoint* of the R package *JointModeling*.

```
### Analysis of original data:



> ori.Gamma <- fitjoint("glm",'CPK~Age+Weight','d~Age',
    family.mean = Gamma(link = "log"),data = bloodcpk)

# EQL: -97.51943

> summary(ori.Gamma$mod.mean)

# Mean Coefficients (output condensed)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.435063   1.021730   4.341 0.000582
Age         -0.015482   0.005809  -2.665 0.017642
Weight       0.031938   0.012960   2.464 0.026289

    Null deviance: 34.536  on 17  degrees of freedom
Residual deviance: 15.000  on 15  degrees of freedom

> summary(ori.Gamma$mod.disp)

# Dispersion Coefficients (output condensed)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.77186    0.98183  -0.786    0.443
Age         -0.03749    0.02510  -1.494    0.155

    Null deviance: 55.664  on 17  degrees of freedom
Residual deviance: 53.050  on 16  degrees of freedom

> anova(ori.Gamma$mod.mean,test="Chisq")

# Mean Analysis of Deviance Table (output condensed)
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    17     34.536
Age     1   14.011        16     20.525 0.0001818
Weight  1    5.525        15     15.000 0.0187495

> anova(ori.Gamma$mod.disp,test="Chisq")

# Dispersion Analysis of Deviance Table (output condensed)
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    17     55.664
Age   1   2.6143        16     53.050     0.199
```

```
### Analysis of modified data:

> bloodcpk$CPK[15] <- bloodcpk$CPK[15]+3000
  # original value 420
> mod.Gamma <- fitjoint("glm",'CPK~Age+Weight','d~Age',
    family.mean = Gamma(link = "log"),data = bloodcpk)

# EQL: -107.2937

> summary(mod.Gamma$mod.mean)

# Mean Coefficients (output condensed)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.63500    1.71681   0.952   0.3560
Age          0.00468    0.01448   0.323   0.7510
Weight       0.06334    0.02257   2.807   0.0133

    Null deviance: 27.190  on 17  degrees of freedom
Residual deviance: 15.000  on 15  degrees of freedom

> summary(mod.Gamma$mod.disp)

# Dispersion Coefficients (output condensed)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.56902    1.11748  -3.194  0.00565
Age          0.06446    0.02723   2.367  0.03088

    Null deviance: 52.365  on 17  degrees of freedom
Residual deviance: 42.783  on 16  degrees of freedom

> anova(mod.Gamma$mod.mean,test="Chisq")

# Mean Analysis of Deviance Table (output condensed)
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    17     27.191
Age     1   0.0932        16     27.097   0.80651
Weight  1  12.0973        15     15.000   0.00526

> anova(mod.Gamma$mod.disp,test="Chisq")

# Dispersion Analysis of Deviance Table (output condensed)
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    17     52.365
Age   1   9.5827        16     42.783     0.025
```

Figure 6.1: Analysis of the blood CPK concentrations using the function *fitjoint* of the R package *JointModeling*. Left: output for the original data; right: output for the modified data.

The output shows that the parameter estimates of the mean model are significant according to the Wald (t-) test statistics and the LR tests (deviance table). However, for the dispersion model the parameter estimates are not significant, see the probability tails

of the Wald and LR tests. This means that there is no heterogeneity model as the age of the skiers is not an influential dispersion covariate. Almost the same results (not presented here) are obtained using the function *dglm* from R package *dglm*.

To get an impression about the influence of outliers on the parameter estimation and on the inference, we added the value 3000 to case number 15 of the response variable CPK. The original data value is 420, and the range of the CPK values is from 200 to 1340. Thus, in this case the modified value would be easily identifiable, and this experiment is only used for illustrative purposes. In general, however, outliers or influential observations might not be extreme along one coordinate (multivariate outliers), and then it is not straightforward to identify them.

The output of the analysis of the modified data is shown in Figure 6.1 (right column). The parameter estimates, as well as the inference, have changed drastically. For instance, the parameter estimate for the covariate age is no longer significant in the mean model but it is significant in the dispersion model according to the Wald tests and the LR test. Similar results are obtained by simply deleting observation 15 from the EQL analysis.

Using this data example, we want to study the effect of the outliers in more detail. Particularly, we are interested in estimating the number of observations to be trimmed, and the effect of trimming on the estimates. We added a value 3000 to the response variable CPK of s randomly selected cases for $s = 2, 3$ by using all possible combinations for s (18 for $s = 2$, and 153 for $s = 3$). Then we compute the maximum ETQL estimates of the new data by trimming t observations (for $t = 0, 1, \ldots, 6$). The resulting estimates are shown in Figure 6.2 for $s = 2$ and Figure 6.3 for $s = 3$. Each boxplot represents the results of the estimated parameter, depending on the number t of trimmed observations. The horizontal lines in the plots show the EQL estimates for the original data, while the vertical lines indicate the "correct" number of trimmed observations (i.e.$t = s$). The plots show that the parameter estimates are very close to the EQL estimates of the original data (horizontal line) in the case $t = s$. If the trimming percentage is too low ($t < s$), the variability of the parameter estimates increases considerably. The parameter estimates remain quite stable if the trimming percentage is chosen higher ($t > s$).

Clearly, the EQL/ETQL value (normalized by the sample size k) has to increase with increasing trimming percentage, but also there a certain break can be seen when t is chosen at least as large as s.

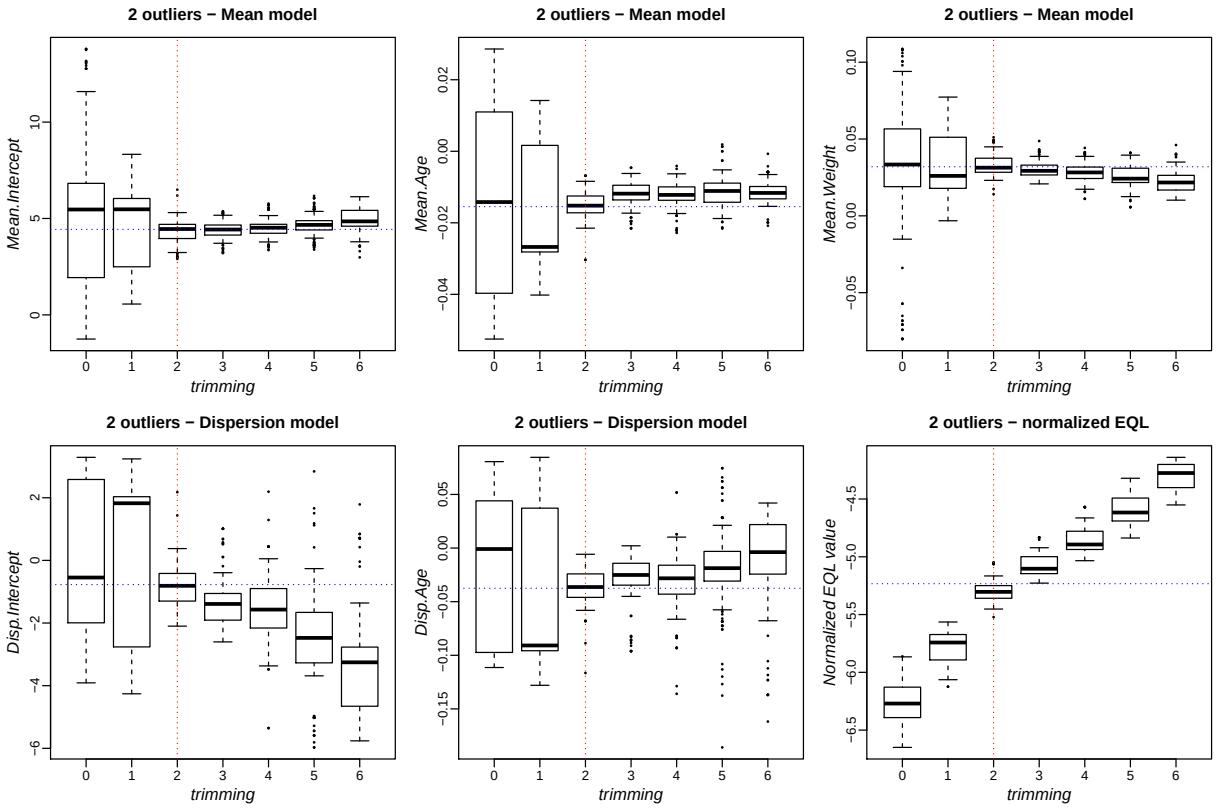

Figure 6.2: Boxplots of the joint modeling parameter estimates (intercept, Age and Weight in the mean model; intercept and Age in the dispersion model) when placing $s = 2$ outliers at any positions of the response variable, and varying the number of trimmed observations. Lower right panel: boxplots for the EQL value, normalized by the sample size.

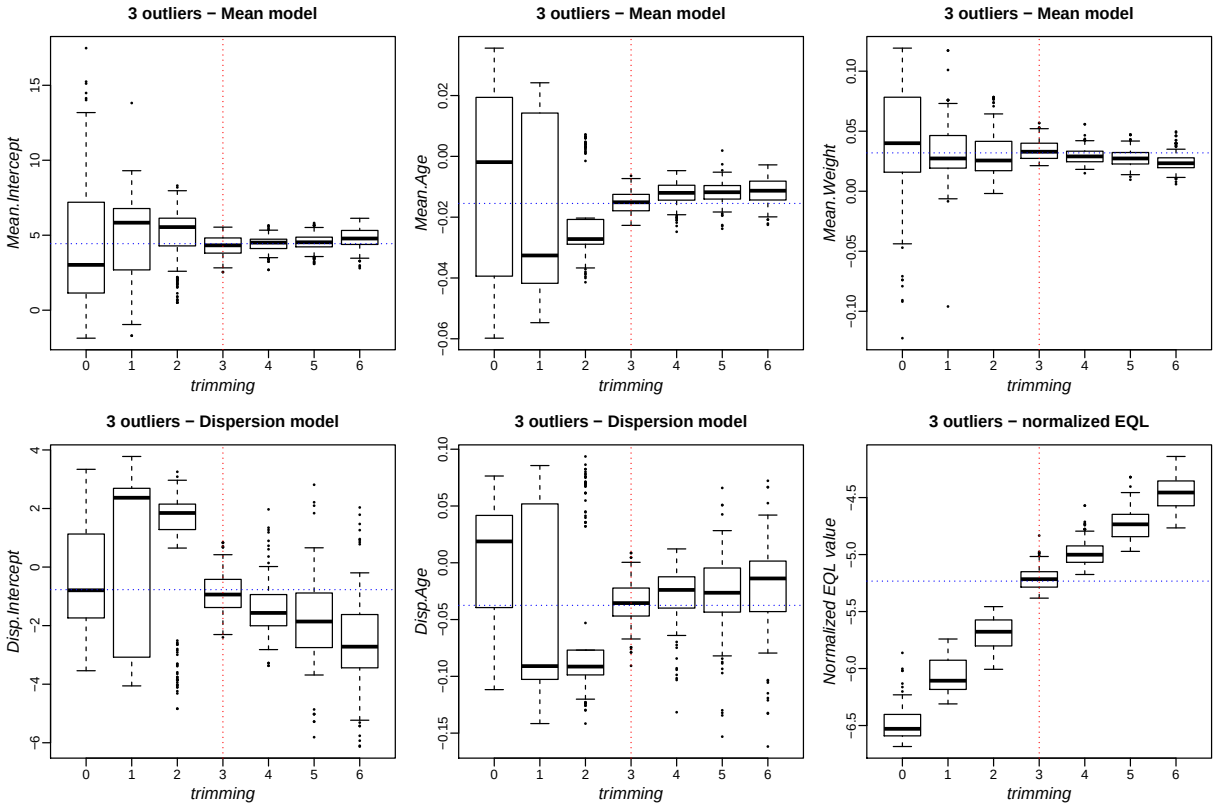

Figure 6.3: Boxplots for the joint modeling parameter estimates (intercept, Age and Weight for the mean model; intercept and Age for the dispersion model) when placing $s = 3$ outliers at any positions of the response variable, and varying the number of trimmed observations. Lower right panel: boxplots for the EQL value, normalized by the sample size.

6.5 Simulation experiments

We compare the performance of the EQL and the maximum ETQL estimator through a simulation study in order to explore their behavior in situations of correct and incorrect dispersion model specification. The estimators are applied to outlier-free and contaminated data with different percentages of trimming.

Since the trial and refinement steps are standard EQL procedures, the wGTE algorithm can be easily implemented using widely available software. We illustrate this in the joint mean and dispersion modeling framework using the packages *dglm* of Smyth (2009) and *JointModeling* of Ribatet and Iooss (2009) which were developed in R (http://www.R-project.org).

6.5.1 Simulation design

The *1st experiment* concerns the classical heteroscedastic normal linear regression model. The regression model was generated according to

$$\begin{aligned} y_i &= 1 + x_{i1} + x_{i2} + \sqrt{\phi_i}\epsilon_i \quad \text{for} \quad i = 1, \ldots, 40 \\ \log(\phi_i) &= -4 - 4x_{i3}, \end{aligned}$$

where x_{i1}, x_{i2} and x_{i3} are uniformly distributed in the intervals [0,1] and ϵ_i is simulated from a standard normal distribution. Data contamination is introduced by modifying four generated values as follows: $x_{37,3} := x_{37,3} - 5$, $x_{38,2} := x_{38,2} - 5$, $x_{39,1} := x_{39,1} + 5$, and $y_{40} := y_{40} - 10$. In this way three of the outliers are leverage points whereas the last one is an outlier in the response variable. Both packages gave almost the same results.

In the *2nd experiment* a gamma mean GLMs is used. The data sample of size 40 is generated according to mean and dispersion models

$$\begin{aligned} \log(\mu_i) &= 1 + x_{i1} + x_{i2} \quad \text{for} \quad i = 1, \ldots, 40 \\ \log(\phi_i) &= -2 - 2x_{i3}, \end{aligned}$$

where the covariates x_{i1}, x_{i2} and x_{i3} are uniformly distributed in the intervals $[-1, 1]$.

Therefore the observations y_i are $Gamma(\phi_i\mu_i, \phi_i^{-1})$ distributed with scale and shape parameters $\phi_i\mu_i$ and ϕ_i^{-1}, respectively. Data contamination is introduced by replacing four generated values as follows: $x_{37,1} := x_{37,1} \pm 14$, $x_{38,2} := x_{38,2} \pm 20$, $x_{39,3} := x_{39,3} \pm 20$ and $y_{40} := y_{40} \pm 14$, where $\pm$ means that the sign plus or minus is randomly selected. As before, three outliers are leverage points whereas the last one is of type response outlier. We note that digamma dispersion GLMs is used instead of gamma dispersion GLMs (6.5) in case of gamma mean GLMs, see Smyth (1989), and Lee et al (2005). Thus the packages *dglm* of Smyth (2009) was used to handle the computations.

In the *3rd experiment* data are generated according to the Tweedie family of distributions with variance function of the form $var(y_i) = \phi_i\mu_i^\theta$ with power parameter $\theta = 1$, mean μ_i and dispersion ϕ_i defined by

$$\begin{aligned} \log(\mu_i) &= 1 + x_{i1} + x_{i2} \quad \text{for} \quad i = 1, \ldots, 40 \\ \log(\phi_i) &= -4 - 4x_{i3}. \end{aligned}$$

The covariates x_{i1}, x_{i2} and x_{i3} are uniformly distributed in the intervals [0,1]. Data contamination is introduced by modifying four generated values as follows: $x_{37,3} := x_{37,3} - 5$, $x_{38,2} := x_{38,2} \pm 5$, $x_{39,1} := x_{39,1} \pm 5$, and $y_{40} := y_{40} + 10$. Similar as before, three outliers are leverage points, the last one is a response outlier. The tweedie distribution from the *tweedie* R package developed by Dunn (2009) was used for data generation whereas the package *dglm* of Smyth (2009) was used to handle the computations. The Tweedie family of distributions belongs to the exponential dispersion model which accommodates the widely used GLMs. Gaussian, Poisson, gamma and inverse-Gaussian families are special cases. Details can be found in Jørgensen (1997).

The simulation experiments were replicated 1000 times. As a consequence, a series of estimates were obtained and their distributions are visualized in boxplots. The series of boxplots for the intercept and slope parameters for both the mean and dispersion panels provide a more detailed characterization of the estimates.

6.5.2 Results and discussion of the 1st simulation experiment

The plots in Figures 6.4–6.6 present the results from the *1st experiment* based on outlier-free (non-contaminated) and contaminated data, and for correctly specified normal mean and gamma dispersion GLMs, respectively. The results of the experiments with non-contaminated data are given in the plot panels of Figure 6.4. From the upper plots one can see both EQL and ETQL estimators perform well in fitting the mean model. The lower panel plots shows that the EQL estimators perform well with respect to the dispersion parameter estimates. However, the variation of the ETQL estimates is larger and bias is observed as the percentage of trimming increases. An obvious reason for this effect is the reduction of sample size due to the special kind of trimming based on the concentration procedure. The results given in the plots panels of Figure 6.5 are based on the experiments with contaminated data. Figure 6.5 shows that the EQL estimator becomes completely useless if part of the data (here 10% contamination) does not follow the model, while the ETQL estimator fits well provided the trimming percentage $\frac{n-k}{n}100\%$ is larger than the percentage of the contamination. The ETQL estimates show the same effect of increased variability for the dispersion model estimation with an increased percentage of trimming. The plots of Figures 6.6 give an impression about the distributions of the trimmed observations when applying the ETQL estimator with different trimming levels within the 1000 experiments. Each boxplot summarizes for a specific observation the outcomes of the 1000 experiments, which are the relative frequencies that the observation is identified as regular, non-outlying, within the computational procedure of the algorithm. Due to the data generation, the last four observations are outliers, and they are correctly identified in the majority of simulation runs and in the majority of the individual steps of the computation, as long as the chosen trimming percentage is not too small. The best stability of this outlier identification is reached for 10% trimming, which corresponds to the actual outlier generation. We note that a similar simulation experiment was considered by Cheng (2011) in order to study the small sample behavior of the restricted (residual) maximum trimmed likelihood estimator.

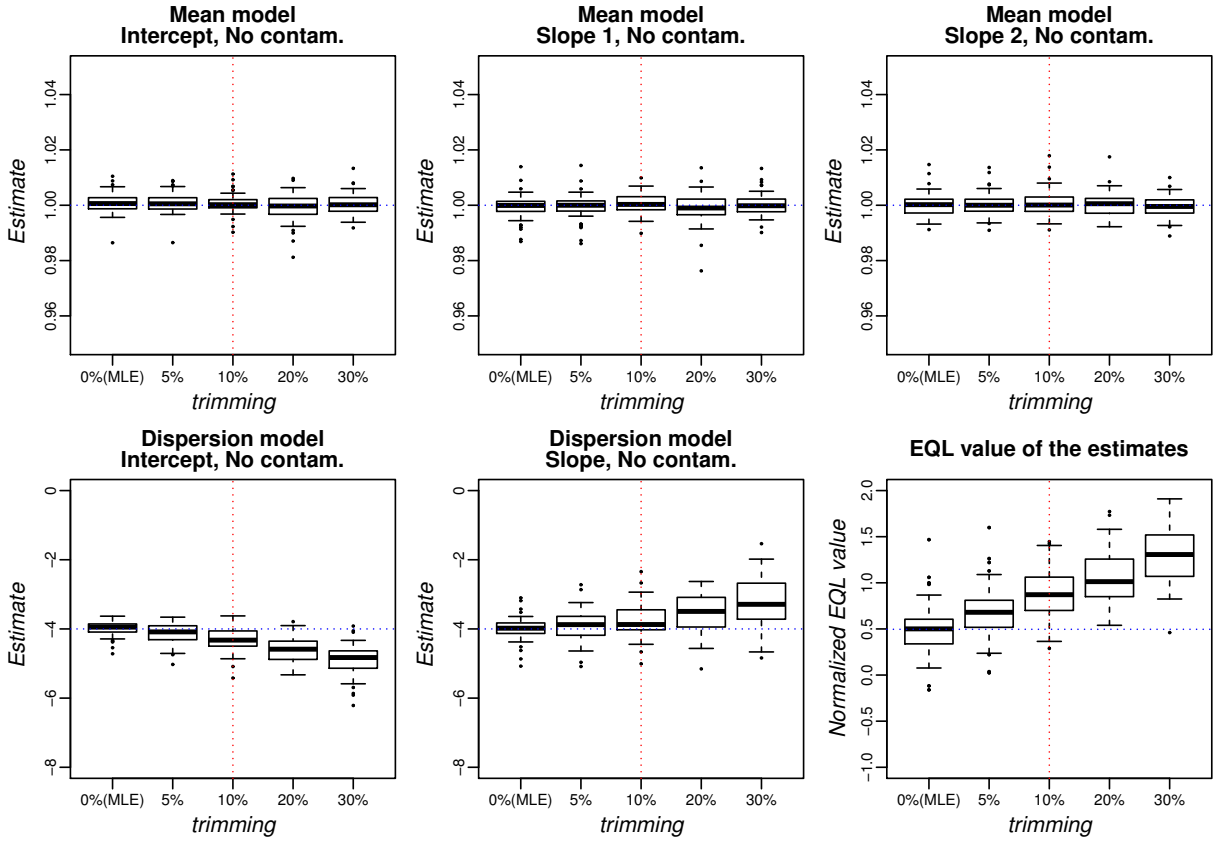

Figure 6.4: *1st simulation experiment without contamination*: boxplots of the estimates obtained from 1000 experiments for the joint normal mean and gamma dispersion GLMs parameters. Lower right panel: boxplots for the EQL values, normalized by the sample size.

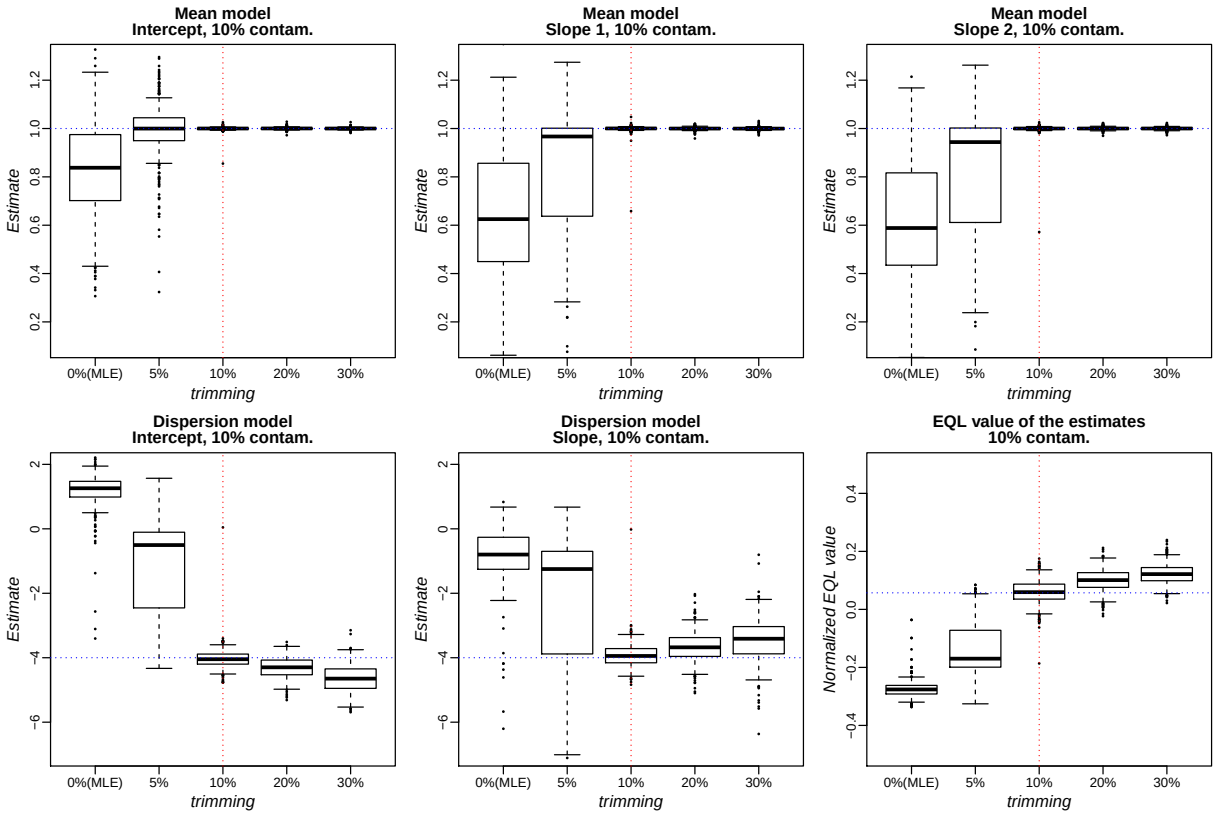

Figure 6.5: *1st simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the joint normal mean and gamma dispersion GLMs. Lower right panel: boxplots for the EQL values, normalized by the sample size.

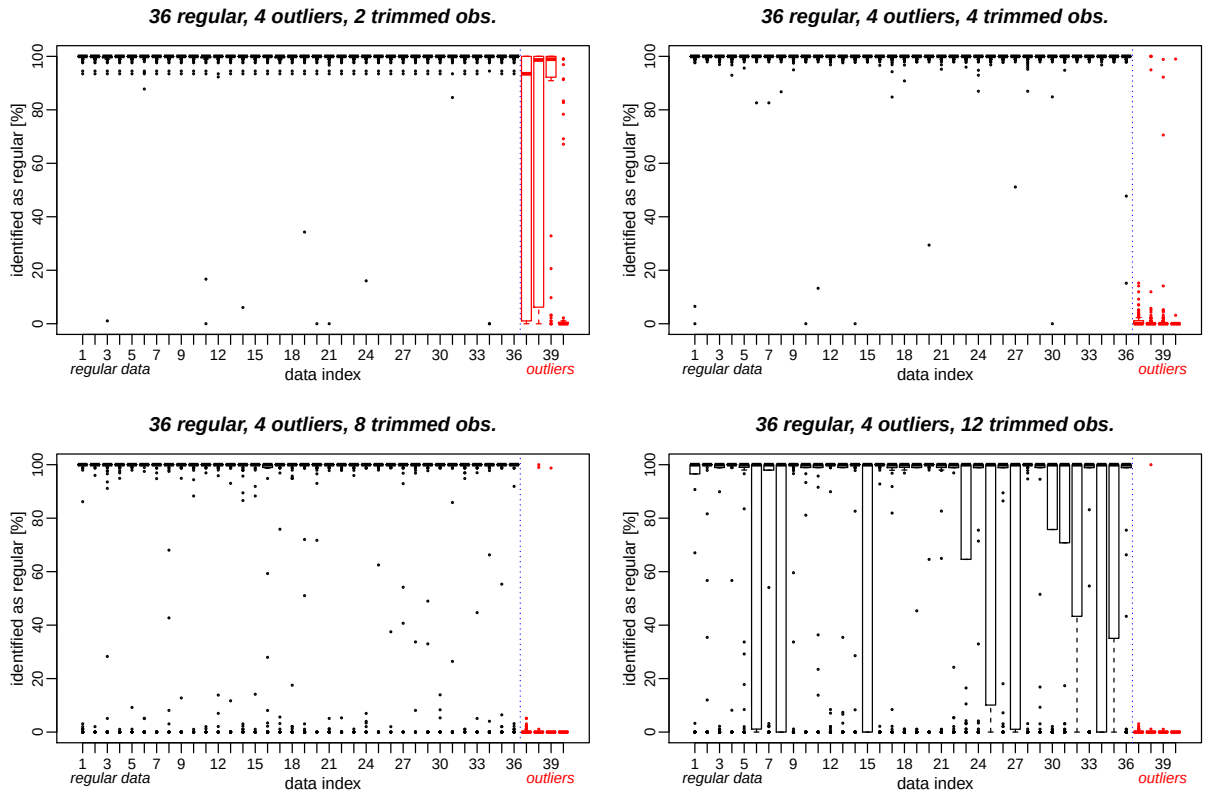

Figure 6.6: *1st simulation experiment with 10% contamination*: relative frequency distribution that an observation is identified as regular within the computational procedure of the algorithm within 1000 experiments.

It is interesting to look at the effect of model misspecification on the EQL and ETQL estimators. Figure 6.5 shows that if the mean model is wrong (because the trimming percentage is zero or too small), then dispersion estimation is affected also. As soon as the appropriate amount of trimming is used, the dispersion parameters are also reasonable. On the other hand, one can check how the estimation of the mean parameters varies if the dispersion is treated as constant. Since the data are heteroscedastic this might have an effect on the mean estimation. In our context, this case reduces the EQL and ETQL estimation problems to ordinary LSE and LTS estimation for the linear regression model. Using the design of the *1st experiment*, the plots presented in Figures 6.7 and 6.8 show the resulting estimates for non-contaminated and contaminated data by varying the trimming percentage among the 1000 simulation experiments. Similar to the previous results we can see that the EQL estimator is useless if a part of the data (here 10% contamination) does not follow the model. For the ETQL estimator the trimming percentage needs to be sufficiently high in order to achieve stable results. For both the uncontaminated and the contaminated data, the results improve if the percentage of trimming is increased. Obviously, this corresponds to trimming data points that generate heteroscedasticity.

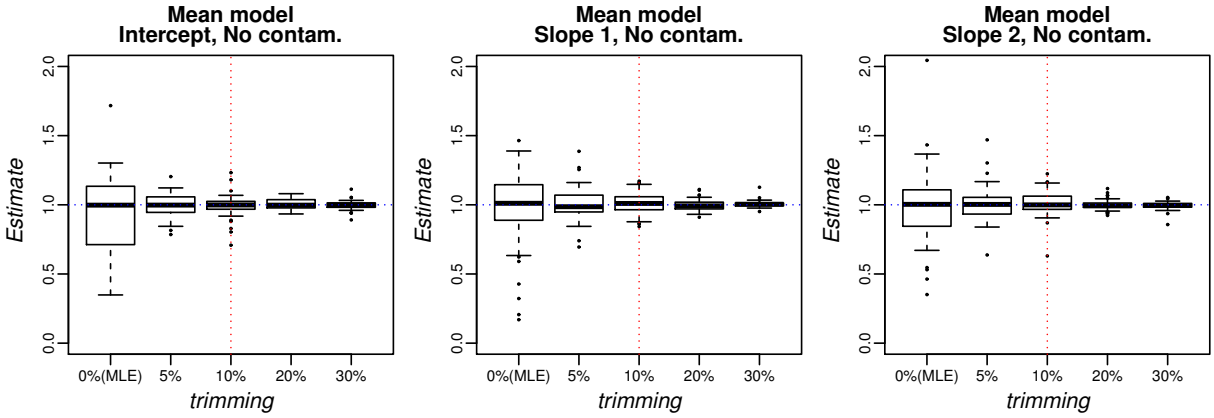

Figure 6.7: *1st simulation experiment without contamination*: boxplots for the estimates obtained from 1000 experiments for the normal mean model; dispersion parameter treated as constant.

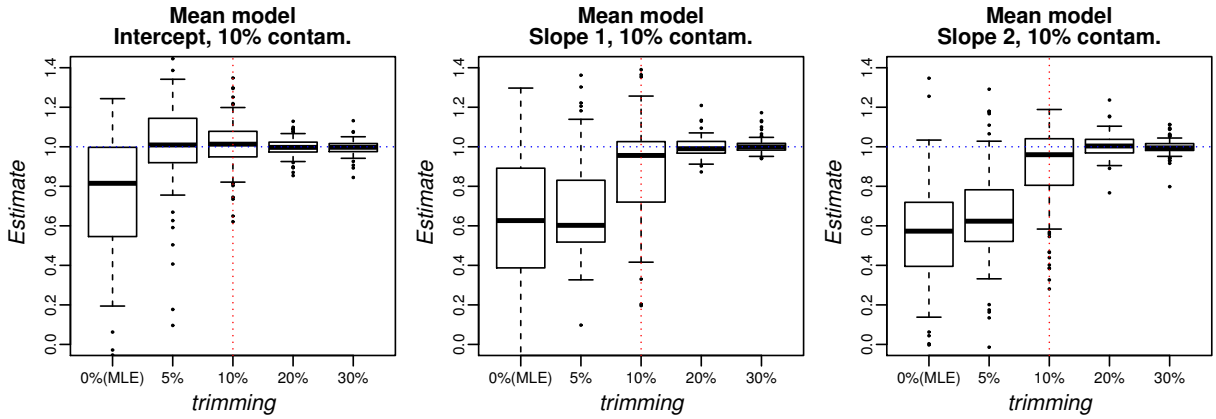

Figure 6.8: *1st simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the normal mean model; dispersion parameter treated as constant.

6.5.3 Results and discussion of the 2nd simulation experiment

In order to provide a direct comparison with the *1st simulation experiment*, we show the same sequence of plots for this experiment. Figure 6.9 presents the results for the uncontaminated data, where the EQL estimator is supposed to perform the best. Using the ETQL estimator, the mean model parameters estimates are still comparable to the EQL estimates for a moderate trimming percentage. However, the dispersion model is much more sensitive to trimming. In case of 10% contamination, the results change drastically, see Figure 6.10. The most precise and stable results are obtained for the ETQL with the correct trimming percentage of 10%. Using a higher percentage causes increasing instability especially for the dispersion parameters estimates. On the other hand, if trimming is too low or zero, the estimates are incorrect.

Figure 6.11 shows the relative frequencies of identifying observations as regular for the contaminated case. The trimming percentage used for the results in the upper left plot is smaller than the contamination level. Accordingly, not all four outliers are regularly identified. For the other plots the outliers were identified correctly in the vast majority of experiments because the trimming level was high enough.

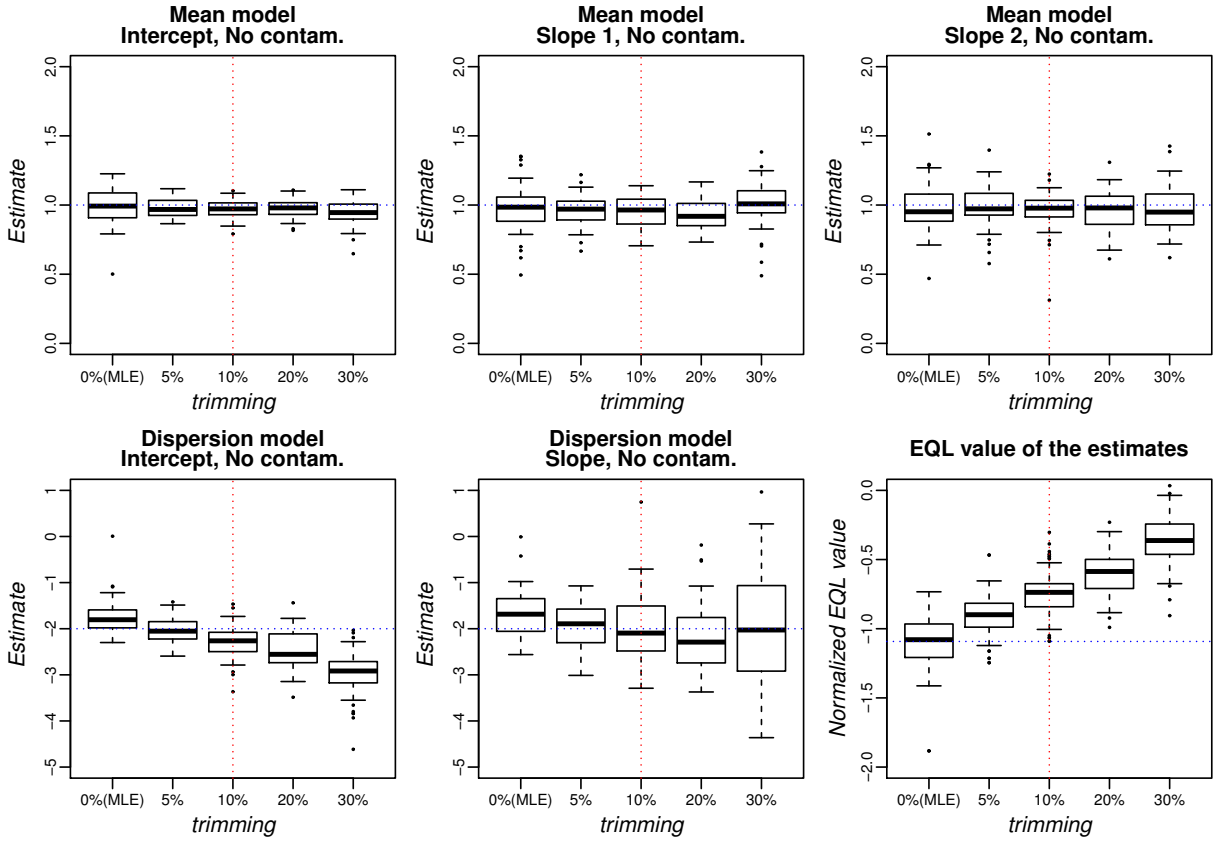

Figure 6.9: *2nd simulation experiment without contamination*: boxplots of the estimates obtained from 1000 experiments for the gamma mean and dispersion GLMs. Lower right panel: boxplots for the EQL values, normalized by the sample size.

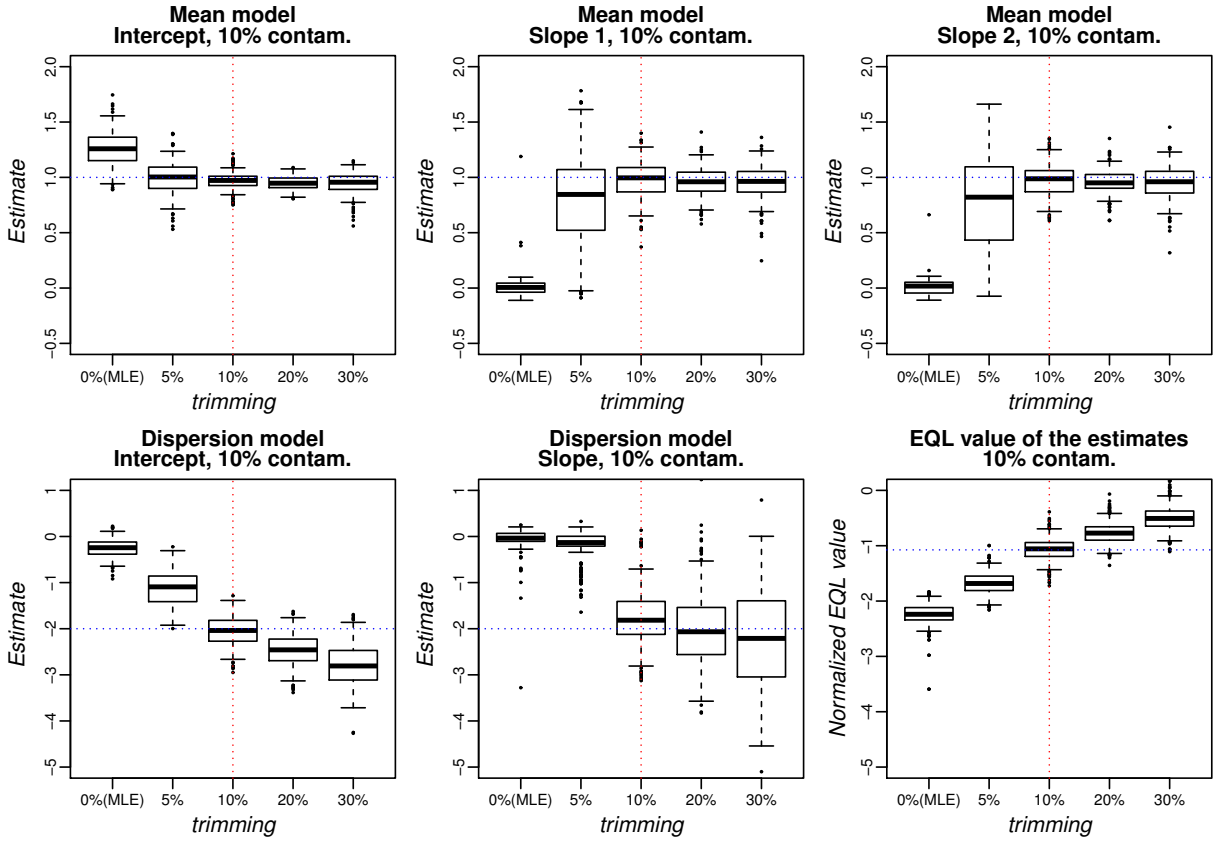

Figure 6.10: *2nd simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the gamma mean and dispersion GLMs. Lower right panel: boxplots for the EQL values, normalized by the sample size.

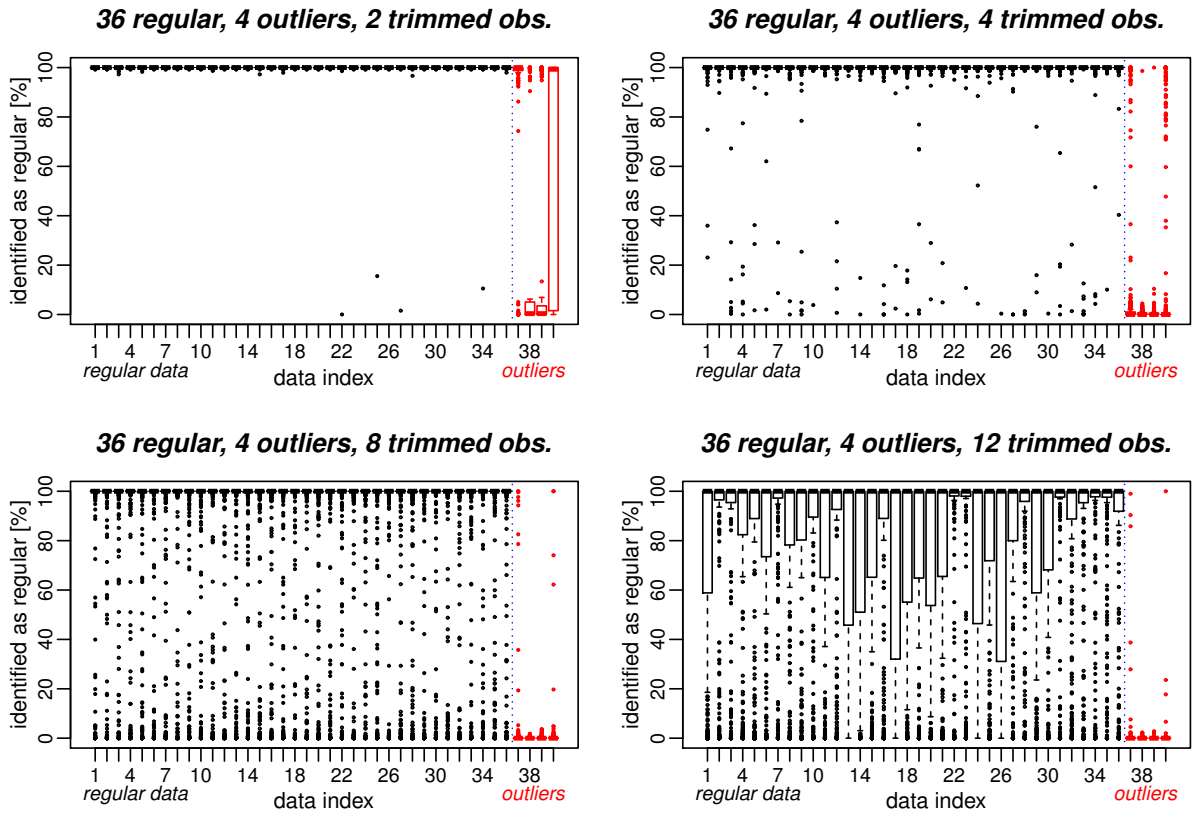

Figure 6.11: the relative frequency distribution that an observation is identified as regular within the computational procedure of the algorithm within 1000 experiments for the gamma mean and dispersion GLMs.

Similar as before, we want to show the effect of model misspecification. Figures 6.10 show that a precise estimation of the mean model parameters implies reasonable dispersion parameter estimates, and vice versa. When treating the dispersion parameter as unknown constant, the parameter estimates of the mean model are relatively stable in case of uncontaminated data for both the EQL and the ETQL estimator, in the latter case even for different trimming percentages, see Figures 6.12. In the case of 10% contamination, from the plots of Figures 6.13 we see the EQL fails completely. ETQL, on the other hand, gives a very precise answer for different trimming percentages.

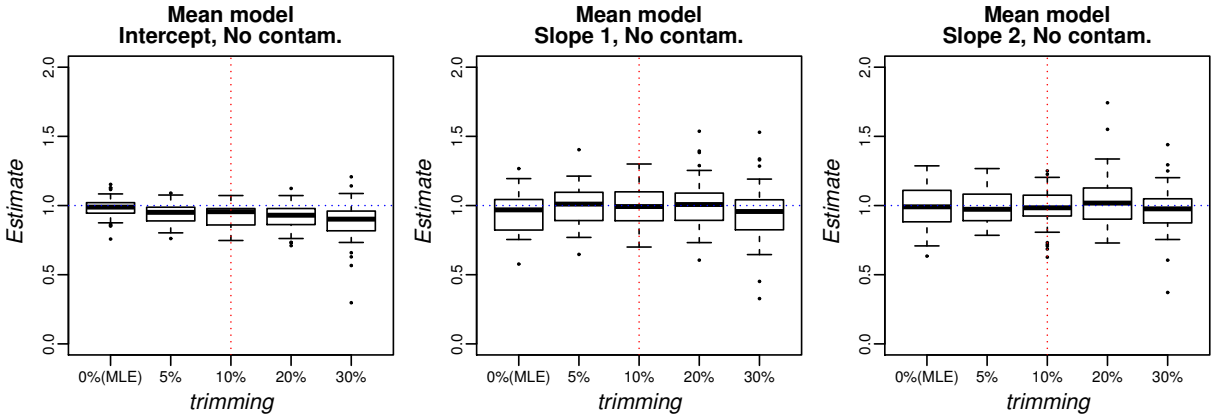

Figure 6.12: *2nd simulation experiment without contamination*: boxplots of the estimates obtained from 1000 experiments for the gamma mean GLMs; dispersion parameter treated as constant.

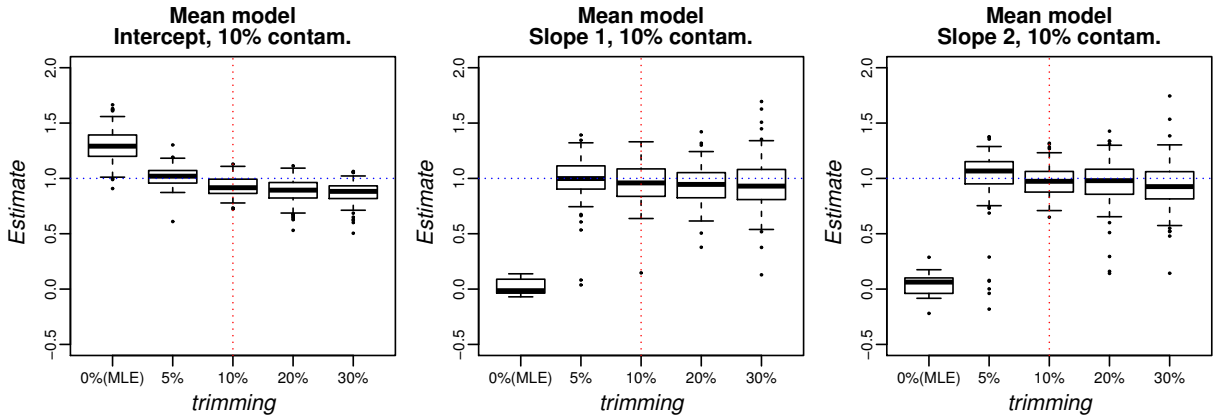

Figure 6.13: *2nd simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the gamma mean GLMs; dispersion parameter treated as constant.

6.5.4 Results and discussion of the 3rd simulation experiment

For this experiment only the results for the 10% contamination data are presented, because of the similarities with the *1st* and *2nd* simulation experiments for the uncontaminated data and model misspecification effect. From the plots of Figure 6.14 we see that the most precise and stable results are obtained for the ETQL with the correct trimming percentage of 10%. Using a higher percentage of trimming causes increasing variability for the dispersion parameter estimates due to the smaller sample size. On the other hand, if trimming is too low or zero, the estimates are incorrect. Figure 6.15 shows the relative frequencies of identifying observations as regular for the contaminated case. The trimming percentage used for the results in the upper left plot is smaller than the contamination level. Accordingly, not all four outliers are regularly identified. For the other plots the outliers were identified correctly in the vast majority of experiments because the trimming level was high enough.

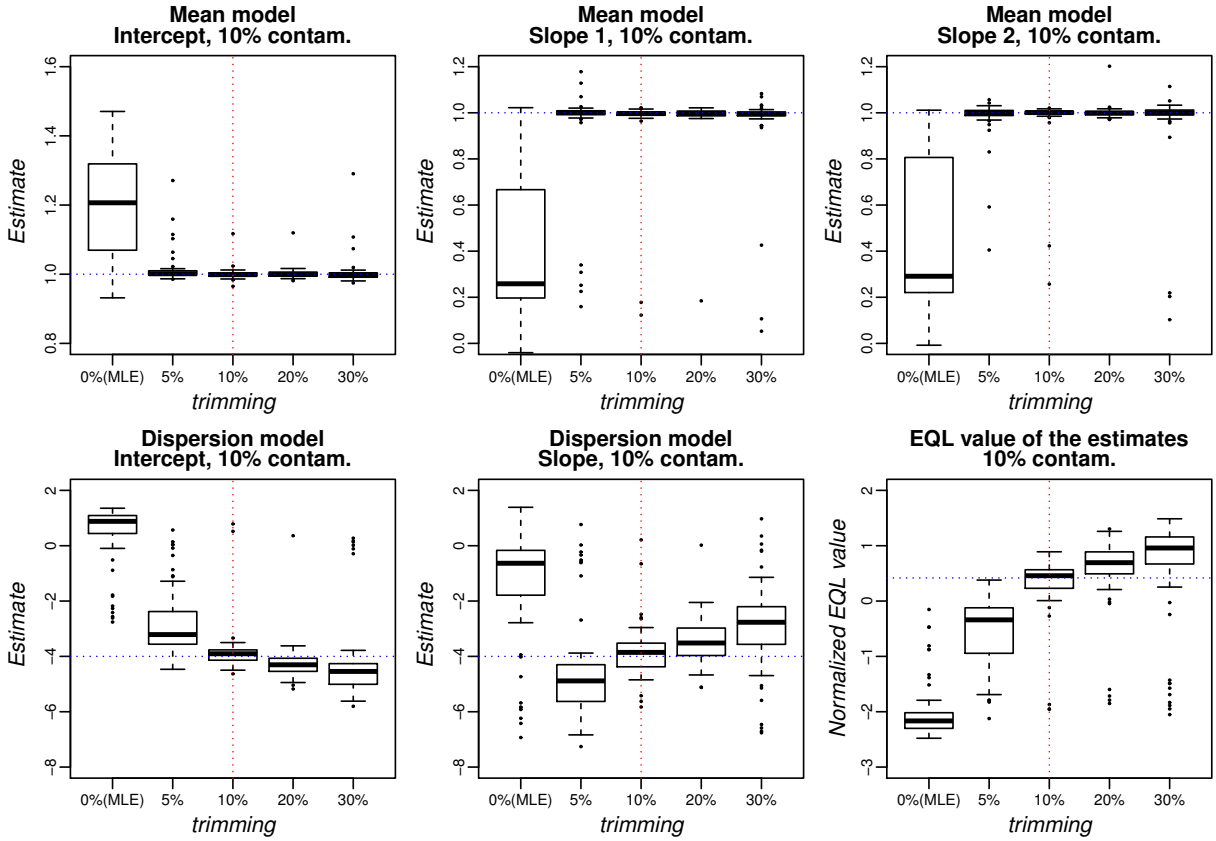

Figure 6.14: *3rd simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the Tweedie distribution with mean and dispersion model and power equal to 1.

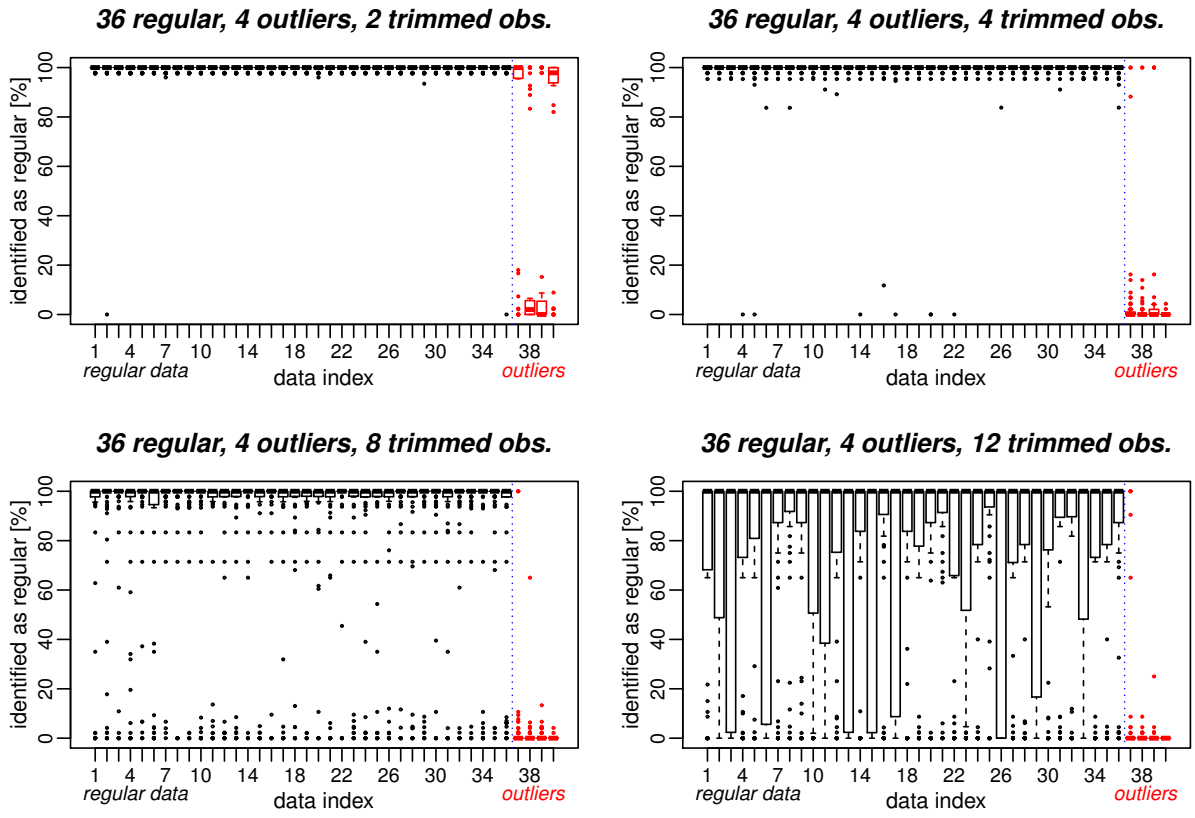

Figure 6.15: the relative frequency distribution that an observation is identified as regular within the computational procedure of the algorithm within 1000 for the Tweedie distribution with mean and dispersion model and power equal to 1.

Usually, the percentage of outliers in real data is unknown. A technique for the selection the trimming percentage $\frac{n-k}{n}100\%$ can thus be based on fitting the model across a range of different percentages of trimming, and by looking for stability of the parameter estimates. This suggests plotting the parameter estimates against the trimming percentage $\frac{n-k}{n}100\%$, where k varies within the interval $[(n + \max[\mathcal{N}(X), \mathcal{N}(Z)] + 1)/2, n]$, and selecting properly that value of k for which the parameters estimates become stable so as to guarantee simultaneously a positive BDP and a higher efficiency of the estimates. For instance, one can proceed by an ETQL estimator, based on a decreasing range of values for k, starting with $k = n$. In this way not only the unknown parameters but also the outlier percentage in the data can be estimated robustly.

6.6 Summary and conclusions

We introduced a robust version of the EQL framework for joint modeling of mean and dispersion based on the idea of trimming and characterized its breakdown point. The computation of the estimator takes advantage of the same technology as used for its classical counterpart, but here the estimation is based on subsamples only. Our algorithm consists of a trial and a refinement step, following the ideas of the fast-LTS and fast-MCD algorithms of Rousseeuw and Van Driessen (1999a, 1999b), and Neykov and Müller (2003).

An important choice for estimators based on trimming is the trimming percentage. In the simulation experiments an approach has been shown how this tuning parameter can be determined. As a by-product, data outliers are flagged. They contain important information for the analyst because of their deviations from the assumed underlying model. In more detail, the outliers are those $n-k$ observations with the largest deviance residuals, and they are excluded from the loss function (6.3), leading to the ETQL loss function (6.11).

After removing the identified outliers, all available diagnostic tools in the context of GLMs can be used (e.g., McCullagh and Nelder, 1989). This is important for checking the validity of the model and for the detection of structure in the (remaining) data.

Chapter 7

The Least Trimmed Quantile Regression

Summary. The linear quantile regression estimator is very popular and widely used. It is also well known that this estimator can be very sensitive to outliers in the explanatory variables. In order to overcome this disadvantage, the usage of the least trimmed quantile regression estimator is proposed to estimate the unknown parameters in a robust way. As a prominent measure of robustness, the breakdown point of this estimator is characterized and its consistency is proved. The performance of this approach in comparison with the classical one is illustrated by an example and simulation studies.

7.1 Introduction

Consider the multiple linear regression model

$$y_i = x_i^T \beta^0 + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, n, \tag{7.1}$$

where y_i is an observed response, $x_i^T = (x_{i1}, \ldots, x_{ip})$ is a vector of explanatory variables (covariates, carriers), and β^0 is the underlying value of a $p \times 1$ vector of unknown parameters β. Classically, ε_i, $i = 1, \ldots, n$, are assumed to be independent and identically distributed. Denote by $r_i(\beta) = y_i - x_i^T \beta$ the regression residuals.

Koenker and Bassett (1978) define the quantile regression (QR) estimator as any vector $\hat{\beta}_n(\tau)$ such that

$$\hat{\beta}_n(\tau) := \arg\min_{\beta \in R^p} \sum_{i=1}^{n} \rho_\tau(r_i(\beta)), \tag{7.2}$$

where

$$\begin{aligned} \rho_\tau(r(\beta)) &= |r(\beta)| \left[\tau 1_{\{r(\beta)\geq 0\}} + (1-\tau) 1_{\{r(\beta)<0\}}\right] \\ &= \begin{cases} (\tau-1) r(\beta) & \text{if } r(\beta) < 0, \\ \tau r(\beta) & \text{if } r(\beta) \geq 0, \end{cases} \end{aligned}$$

$0 < \tau < 1$, and $1_{\{A\}}$ is the usual indicator function of the set A, which equals 1 if A is true and 0 otherwise.

Different quantile regression estimates $\hat{\beta}_n(\tau)$ can be obtained for different values of τ. This offers the analyst a more complete statistical model than the mean regression, and nowadays, quantile regression has widespread applications. It can be derived as the maximum likelihood (ML) estimator of observations coming from an asymmetric Laplace (double exponential) distribution (e.g., Koenker and Machado, 1999). The quantile regression estimator is robust to skewed tails and departures from normality. In addition, under very general conditions, the asymptotic distribution of the vector of estimated coefficients is multivariate normal, which permits standard inferences to be carried out (Koenker and Bassett, 1978). The finite-sample distribution of quantile regression was also studied (e.g., Jurečková, 2010). Computational algorithms concerning quantile regression estimation are based on linear programming techniques as discussed in Koenker (2005a), or maximization-minorization techniques considered by Hunter and Lange (2000) and Chen (2004). The package *quantreg* developed in R by Koenker (2005b) (http://www.R-project.org) facilitates the wide use of quantile regression. For more details about quantile regression see Koenker (2005a).

Unfortunately, the quantile regression estimator, like other regression M-estimators, can be highly sensitive to outliers in the explanatory variables, see He et al. (1990). Therefore, many attempts based on the downweighting of distant observations appeared that led to a more robust form of quantile regression (e.g., Hubert and Rousseeuw, 1998, and Giloni et

al., 2006). Such procedures were shown to be robust in the regression with a small number of uniformly-distributed or fixed-design regressors (see Giloni et al., 2006, for the case of one and two regressors). The robustness of weighted quantile regression, however, diminishes in general with an increasing number of regressors, and even for a small number of covariates, the robustly weighted quantile regression can be substantially biased by outliers (e.g., Čížek, 2011). This led to the development of alternative estimators of the quantile regression model (7.1), which are generally based on saddle-point optimization problems (e.g., Rousseeuw and Hubert, 1999, and Adrover et al., 2004). The robustness of these procedures is independent of the complexity of the regression model and is proportional to $\min\{\tau, 1-\tau\}$, where τ refers to the quantile of interest (see Adrover et al., 2004, for an overview). The main disadvantages of these methods – the computational difficulties and non-standard asymptotic distributions – are generally related to their definitions based on nested optimization problems.

In this chapter, we consider an alternative approach to robust estimation in the framework of quantile regression, the least trimmed quantile regression (LTQR), which is based on trimming in order to reduce the influence of the outliers in the explanatory variables. The proposed method extends the robust location estimator of the median theoretically studied by Tableman (1994a,b) and the Least Trimmed Absolute deviation (LTA) estimator studied empirically by Hawkins and Olive (1999): we generalize them to the general quantile regression model (7.1), and additionally, prove the consistency of the proposed method and thus also of LTA. Contrary to existing highly robust methods of quantile regression discussed in the previous paragraphs, the LTA and proposed LTQR estimate the regression quantiles for the data that are not trimmed from the objective function, that is, the quantiles are determined for a subset of data. This allows us to achieve robustness properties independent of the quantile τ of interest. However, the superior robustness properties of the LTQR estimator impose also one constraint: although we can consistently estimate the regression coefficients of all variables in the model (7.1), the intercept will not be identified (i.e., the constant term will converge to another quantity than the τth quantile of errors ε_i at $x_i = 0$). If the intercept is of importance, it has to be identified

by some of the existing procedures.

The paper is organized as follows. Section 2 recalls the generalized trimmed estimator, which renders the proposed LTQR method, and discusses its computation. The LTQR estimator is defined and its breakdown property is discussed in Section 3, while the consistency of the proposed methods is discussed in Section 4. Section 5 demonstrates the different behavior of classical and robust estimation on a simple example. In Section 6, a simulation study is performed to illustrate the effectiveness of the proposed estimator in comparison with the classical quantile regression. Finally, the conclusions are given in Section 8 and proofs are provided in the Appendix.

7.2 The Generalized Trimmed Estimator

The LTQR estimator will be obtained as a special case of the Generalized Trimmed Estimator (GTE) given by Vandev and Neykov (1998). To introduce it, note that GTE can be defined for any regression model by means of an objective function $f_i : \Theta \to \mathbb{R}^+$, where $\Theta \subseteq \mathbb{R}^q$ is an open set. In particular, the GTE estimator $\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}}$ of θ is defined as the solution of the optimization problem

$$\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}} := \arg\min_{\theta\in\Theta} \left\{ S_{n,k}(\theta) = \min_{I\in I_k} \sum_{i\in I} w_i f_i(\theta) \right\}, \tag{7.3}$$

where I_k is the set of all k-subsets of the index set $\{1, \ldots, n\}$ and k is the trimming constant determining the number k of observations and their function values $f_i(\theta)$ kept in the objective function from the total number n of observations. Consequently, the trimming parameter k determines the robustness properties of the GTE as $n - k$ observations with the largest values of $f_i(\theta)$ are excluded from the loss function.

The robustness properties of the GTE can be described, for example, by the finite-sample breakdown point (BDP): it is a global measure of an estimator's robustness characterizing the minimum number of observations that, if arbitrarily modified, can cause the estimates to increase above any bound. The BDP of the GTE is characterized by Theorem 1 of Vandev and Neykov (1998) using the d-fullness technique. Dimova and

Neykov (2004) proved that the BDP of the GTE is not less than $\frac{1}{n}\min\{n-k, k-d\}$ if the set $F = \{f_i(\theta) : i = 1, \ldots, n\}$ is d-full ($d = p$ if any p observations are linearly independent, see Section 3 and the appendix for more details). The BDP is maximized for $k = \lfloor (n+d+1)/2 \rfloor$, when it approximately equals $1/2$ for large n. Further, the asymptotic properties of the GTE estimator (7.3) were studied by Čížek (2008) for the case of twice differentiable functions f.

The optimization problem (7.3) defining the GTE is of combinatorial nature due to the representation

$$\min_{\theta\in\Theta^p} S_{n,k}(\theta) = \min_{\theta\in\Theta^p} \min_{I\in I_k} \sum_{i\in I} w_i f_i(\theta) = \min_{I\in I_k} \min_{\theta\in\Theta^p} \sum_{i\in I} w_i f_i(\theta). \tag{7.4}$$

Therefore, it follows that all possible $\binom{n}{k}$ partitions of the set $\{f_1, \ldots, f_n\}$ have to be considered and $\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}}$ is defined by the partition with the minimal value of $S_{n,k}(\theta)$. Hence, an exact computation of the GTE is infeasible for large samples. To get an approximative GTE solution, an algorithm was developed in Neykov et al. (2012). It repeatedly (i) sets $s = 0$, selects a small subset $\{f_{i_1}, \ldots, f_{i_{k^*}}\}$ of k^* functions from F and forms $I_s = \{i_1, \ldots, i_{k^*}\}$, (ii) minimizes the objective function $\sum_{i\in I_s} f_i(\theta)$ with respect to θ, and uses the obtained estimate $\hat{\theta}_s$, (iii) sets $s = s + 1$, orders the functions of F in ascending order, $f_{\nu(1)}(\hat{\theta}_s) \leq f_{\nu(2)}(\hat{\theta}_s) \leq \ldots \leq f_{\nu(k)}(\hat{\theta}_s) \leq \ldots \leq f_{\nu(n)}(\hat{\theta}_s)$, where $\nu(.)$ is the permutation of the indices $\{1, 2, \ldots, n\}$, and forms $I_s = \{\nu(1), \ldots, \nu(k)\}$; the steps (ii) and (iii) are repeated as long as the newly obtained estimates $\hat{\theta}_s$ produce smaller values of the objective function $\sum_{i\in I_s} f_i(\theta)$.

To fully specify the algorithm, the size and choice of the initial subsets have to be specified (all possible subsets of size $k^* = k$ can be considered to obtain the precise instead of an approximative solution only in very small samples). First, the trial subsample size k^* should be greater than or equal to d, which is necessary for the existence of (7.3), but the chance to get at least one good subsample of data points is larger if $k^* = d$. Next, the initial subsets of observations are traditionally chosen as random subsamples of size k^*. As this requires a large number of initial subsets to be drawn to obtain a good approximation and because the QR estimator used from Section 7.3 on possesses some

robustness properties if there are no leverage points (cf. Giloni et al., 2006), we combine the random and a deterministic choice of initial subsamples. Specifically, we draw a number of initial subsamples of size k^* randomly, and additionally, n_{init} initial subsamples are taken as the ith observation and its (k^*-1)-nearest neighbors in the space of the explanatory variables, $i = 1, \ldots, n_{init}$. The algorithm could be further accelerated for large data sets by applying the partitioning and nesting techniques as in Rousseeuw and van Driessen (1999, 2006).

7.3 The Least Trimmed Quantile Regression Estimator

In this section, the Least Trimmed Quantile Regression (LTQR) estimator is introduced and the finite-sample BDP properties of the linear LTQR estimator are discussed. The LTQR estimator is a particular form of the GTE (7.3) that, contrary to many existing variants, employs a non-differentiable objective function $f_i(\beta) = \rho_\tau(r_i(\beta))$.

Definition 7.1 *The LTQR estimator is defined by*

$$\hat{\beta}_n^k(\tau) := \arg\min_{\beta} \left\{ Q_{n,k}(\beta) = \min_{I \in I_k} \sum_{i \in I} \rho_\tau(r_i(\beta)) \right\}, \tag{7.5}$$

where I_k is the set of all k–subsets of the set $\{1, \ldots, n\}$, $\rho_\tau(r_i(\beta))$ is defined by (7.2), and $0 < \tau < 1$.

From this definition, it can be seen that the maximum LTQR estimator is the classical QR estimator calculated for some k-subset of the n cases. Consequently, the LTQR estimator includes the quantile regression estimator (7.2) as a special case for $k = n$, and the LTA estimator for $\tau = 0.5$. As the linear LTQR estimator is a particular case of the GTE, its finite-sample BDP can be derived from the finite-sample BDP of the GTE.

Theorem 7.1 *Let $\mathcal{N}(X) = \max_{0 \neq \beta \in \mathbb{R}^p} \text{card}\{i \in \{1, \ldots, n\}; x_i^T \beta = 0\}$. Then the BDP of the linear LTQR estimator equals $\frac{1}{n} \min\{n - k, k - \mathcal{N}(X) - 1\}$. The BDP attains its*

maximum and equals to $\frac{1}{n}\lfloor\{n-\mathcal{N}(X)-1\}/2\rfloor$ *for* k *such that* $\lfloor\{n+\mathcal{N}(X)+1\}/2\rfloor \leq k \leq \lfloor\{n+\mathcal{N}(X)+2\}/2\rfloor$.

The quantity $\mathcal{N}(X)$, introduced by Müller (1995), provides the maximum number of explanatory variables $x_i \in R^p$ lying in a subspace. If any p observations x_i^T are linearly independent, then $\mathcal{N}(X) = p-1$, which is the minimal value for $\mathcal{N}(X)$. When the covariates are qualitative variables such as factors with several levels, $\mathcal{N}(X)$ can be much larger.

As $\mathcal{N}(X)$ is bounded and independent of n, the most robust choice of trimming $k = \lfloor\{n+\mathcal{N}(X)+1\}/2\rfloor$ guarantees a BDP which will be asymptotically equal to 1/2 and independent of τ. This is possible because LTQR estimates the quantiles only within the subset of observations that are not trimmed from the objective function, and as shown in the following section, it does not identify the intercept in model (7.1). On the other hand, the proof of Theorem 1 (in the Appendix) shows that the size of the compact set containing the LTQR estimates in the presence of contamination does depend on τ by means of $\min\{\tau, 1-\tau\}^{-1}$. Although the BDP can reach 1/2 for any τ, the maximum bias caused by contamination will be smallest for $\tau = 1/2$, it will increase as τ moves away from 1/2, and could be arbitrarily large if one requires $\tau \to 0$ or $\tau \to 1$.

7.4 Consistency of the LTQR estimator

Here it will be shown that the LTQR estimator (7.5) is a consistent estimator of the slope parameters in model (7.1). Moreover, the constant identified by the LTQR estimator will be found.

Let us now assume for the sake of simplicity that the distribution function F of the error term ε_i in (7.1) has an infinite support. Further, as the trimming constant k defining the LTQR estimator generally depends on the sample size n, we will write k_n to indicate this and assume $\lim_{n\to\infty} k_n/n = \lambda \in (0,1\rangle$ exists. In the location model, that is, in model (7.1) containing only the constant term, Tableman (1994a) then showed that the LTQR

estimator with $\tau = 0.5$ identifies the median on the shortest interval Δ such that $P(y_i \in \Delta) = \lambda$. To formalize this statement in the general case, let us first state assumptions on the data generating process.

Assumption D. The vectors (x_i, ε_i) form a sequence of independent and identically distributed random vectors with the finite $(1+\delta)$th moments for some $\delta > 0$.

Assumption F. Let the distribution function F be continuous, strictly increasing on its support, and having a differentiable density function f, which is supposed to be unimodal and bounded on its support.

Let us recall that, for an interval $\Delta(a, \lambda) = \langle F^{-1}(a), F^{-1}(a+\lambda)\rangle$, $a \in (0, 1-\lambda)$, and a fixed $\tau \in (0,1)$, Tableman (1994a, p. 390) proved in the location model that LTA applied to univariate data following the distribution function F converges to and thus consistently estimates

$$\mu^*(\tau) = F^{-1}(a^*(\tau) + \tau\lambda), \tag{7.6}$$

where $a^*(\tau) = \arg\min_{a\in(0,1)} \int_{\Delta(a,\lambda)} \rho_\tau(\varepsilon - F^{-1}(a+\tau\lambda))dF(\varepsilon)$.

Assuming that the intercept is the first element of the parameter vector β, we will now show in the regression case (7.1) that the LTQR estimator consistently estimates the parameter vector $\beta^*(\tau) = (\mu^*(\tau), 0, \ldots, 0)^T + \beta^0$, where the parameter $\beta^*(\tau)$ obviously equals to β^0 for all its elements, but the first one. The constant term obtained by the LTQR estimator thus corresponds to $\beta_1^0 + \mu^*(\tau)$, where in general $\mu^*(\tau) \neq 0$ ($\mu^*(\tau) = 0$ if F is symmetric and $\tau = 1/2$, for instance).

Theorem 7.2 *Let Assumptions D and F hold and let $\tau \in (0,1)$ be fixed. Assuming $\beta^0 \in B$ and $\beta^*(\tau) = (\mu^*(\tau), 0, \ldots, 0)^T + \beta^0$, where $\mu^*(\tau)$ is defined in (7.6) and B is a compact parametric space, the LTQR estimator defined for $k_n = [\lambda n]$ and $\lambda \in (0, 1\rangle$ consistently estimates $\beta^*(\tau)$, $\hat{\beta}_n^{k_n}(\tau) \to \beta^*(\tau)$ in probability as $n \to \infty$.*

The theorem shows that, under Assumption F, the LTQR estimator correctly identifies the coefficients of the regression variables, but provides a different estimate of the constant term. To obtain the intercept term representing the classical τth quantile, one can use

the residuals $r_i(\hat{\beta}_n^{k_n}(\tau))$ from the LTQR estimator fit, compute their empirical τth quantile q_τ, and add q_τ to the LTQR estimator intercept estimate. A possible caveat of such a procedure is its robustness: this newly estimated intercept has (asymptotically) a BDP bounded by $\min\{\tau, 1-\tau\}$, which is irrelevant for τ close to 0.5, but rather limiting for quantiles τ close to 0 or 1.

7.5 Examples

Since the trial and refinement steps of the GTE-LTQR algorithm are standard quantile regression procedures, the GTE algorithm can be easily implemented using widely available software. We illustrate this using the package *quantreg* of Koenker (2005), which was developed in R (R Development Core Team, 2011). In particular, we first compare the performance of classical linear quantile regression and the LTQR estimator through a real dataset and a simulation study. Later, some robustness properties of LTQR and existing robust methods are compared.

7.5.1 Star cluster CYB OB1 dataset

First, the well-known dataset on the star cluster CYB OB1 consisting of 47 observations is considered, which was already analyzed by Adrover et al. (2004) and Rousseeuw and Leroy (1987). In the upper left corners of the plots of Figure 7.1 one can see four points with high leverage that do not follow the trend of the data majority. The observations are plotted as tiny black bullets on all of the plots. Here we focus on estimating the regression quantiles τ of 0.25, 0.50, and 0.75 by both the classical QR estimator proposed by Koenker and Bassett (1978) and by the LTQR estimator using different trimming percentages. The upper plots show the results of the classical estimator for all data points (upper left) and for a reduced dataset where the four leverage points are deleted (upper right). It is evident that the leverage points have a strong influence on the classical estimator.

The remaining plots in Figure 7.1 show the results of the LTQR estimator on the original

data, with 4%, 9%, 11%, and 17% trimming. This corresponds to trimming 2, 4, 5, and 8 observations, respectively. The trimmed observations are marked by symbols: tiny squares for $\tau = 0.25$, upside-down triangles for $\tau = 0.50$, and normal triangles for $\tau = 0.75$. The corresponding LTQR regression lines are influenced by the leverage points in case the trimming percentage is too low (4%). For 9% trimming the four leverage points are identified as outliers and we obtain practically the same result as for the classical method applied to the reduced data. If the trimming percentage is chosen higher (11%, 17%), additional observations are identified as outliers, but the regression lines are very stable. It is interesting to see that not always the same additional observations are trimmed: this depends on the considered regression quantile τ. This phenomenon is corresponding to the definition of regression outliers, where observations that do not follow the assumed model can be treated as outliers. We can also see that even the LTQR fits for $\tau = 0.75$ with 11% and 17% of trimming are not influenced by the outliers like, for example, the maximum depth estimator in Adrover et al. (2004, Figure 2). The LTQR regression lines are similar to those of the robustified Koenker and Bassett (RobKB) method in Adrover et al. (2004, Figure 1), but look more plausible and stable because the LTQR median regression lines will intersect both with the $\tau = 0.25$ and $\tau = 0.75$ fitted lines for large values of the covariate, whereas the RobKB median regression line intersects with the $\tau = 0.75$ fitted line for small covariate values and with the $\tau = 0.25$ fitted line for large covariate values.

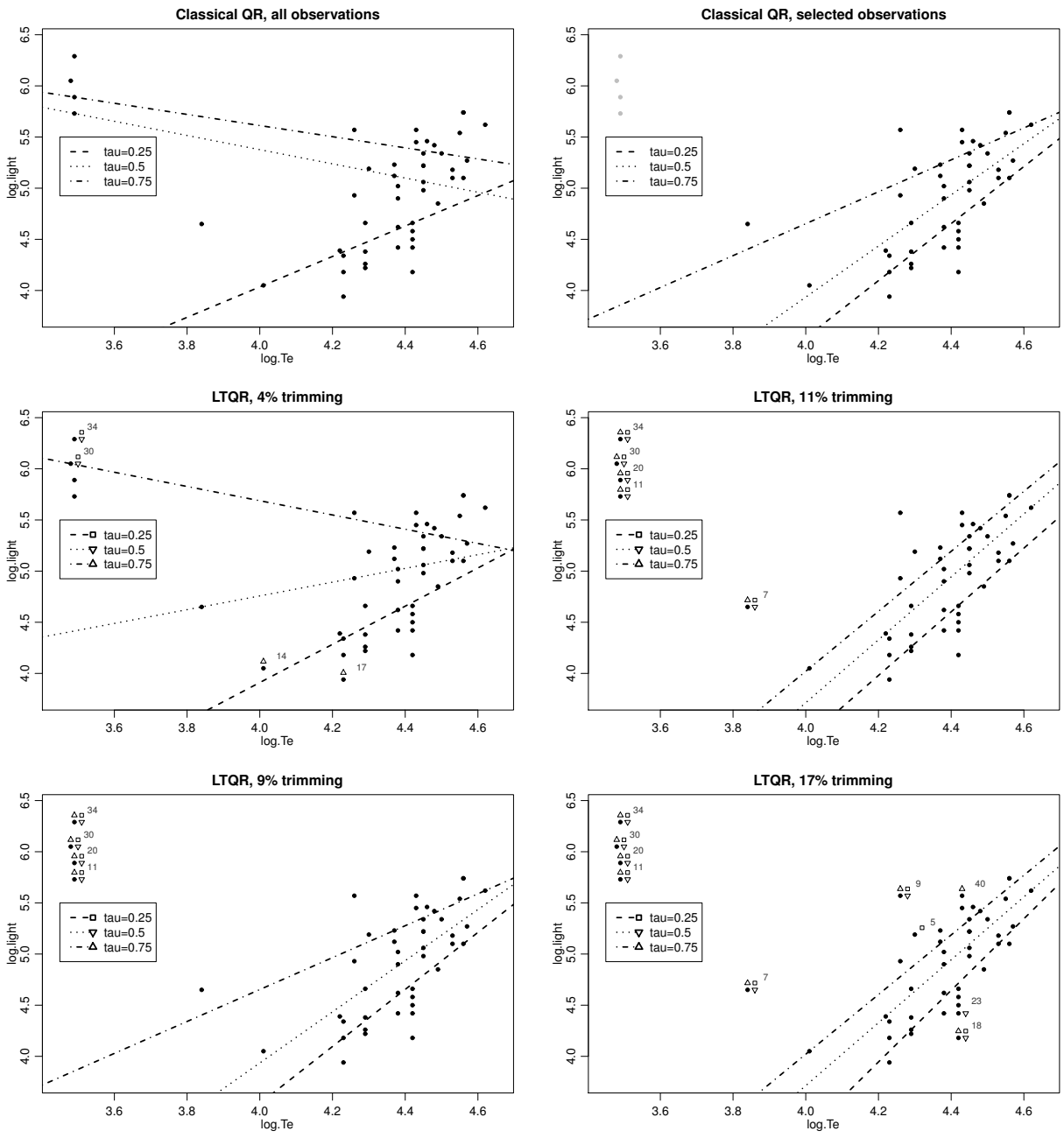

Figure 7.1: Star data: 0.25, 0.50, and 0.75 regression quantiles from Koenker and Bassett estimate, based on whole data (upper left) and on data without the four extreme points (upper right); LTQR fits with 4%, 9%, 11% and 17% of trimming (remaining plots).

7.5.2 Simulation experiments

We compare the performance of the QR and LTQR estimators through a simulation study within the classical heteroscedastic multiple linear regression model. The data were generated according to the model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sigma_i \varepsilon_i \quad \text{for} \quad i = 1, \ldots, 100, \\ \sigma_i &= \sqrt{\exp(0.11(x_{i1} + x_{i2}))}, \end{aligned}$$

where $\beta_0 = \beta_1 = \beta_2 = 0$ can be chosen without loss of generality because of the regression equivariance of LTQR, $\varepsilon_i \sim \mathrm{N}(z_\alpha, 1)$, and z_α is the α–quantile of N(0,1). (Note that this traditional form of heteroscedasticity implies a slight nonlinearity of the QR regression lines for $\tau \neq 0.5$.)

Two distribution types for the covariates are considered: in the *1st experiment*, the covariates x_{i1} and x_{i2} are uniformly distributed on the interval $[-10, 10]$, that is, $x_{ij} \sim \mathrm{U}[-10, 10]$ for $j = 1, 2$; in the *2nd experiment*, the covariates x_{i1} and x_{i2} are normally distributed, that is, $x_{ij} \sim \mathrm{N}(0, 1)$ for $j = 1, 2$. Data contamination is introduced by modifying the first $m = \lfloor 100\epsilon \rfloor$ observations for $\epsilon = 0.1, 0.2, 0.3$ as follows ($r = 2, 3, 4$): in the *1st experiment*, $x_{ij} \sim \mathrm{U}[-30, -20]$ for $j = 1, 2$ and $y_i \sim \mathrm{U}[-10r, -10r + 10]$; in the *2nd experiment*, $(x_{i1}, x_{i2}, y_i)^T \sim \mathrm{N}_3(\mu, \Sigma)$ where $\mu = (-10, -10, -10r)^T$ and $\Sigma = 3I_3$ for $i = 1, \ldots, m$. In this way all those generated outliers are bad leverage points of different magnitude. As the results are similar across different choices of r, we present the *1st experiment* with $r = 2$ and the *2nd experiment* with $r = 3$.

All simulation experiments were replicated 1000 times to explore the small sample behavior of the classical QR and LTQR estimators for the different quantile values $\tau = (0.5, 0.75, 0.90)$ and different trimming percentages over the clean and contaminated data. Subsequently, the simulated estimates were obtained and summarized in boxplots, see Figures 7.2–7.9. The plot panels for the upper rows of the figures show the results for the intercept term β_0, while the middle and bottom rows present the results of the slope parameters β_1 and β_2, respectively. The "correct" trimming percentages are indicated by

dotted vertical lines, and the true simulated parameters are indicated by horizontal dashed lines.

Figures 7.2 and 7.3 demonstrate the performance in the uncontaminated case for both uniformly and normally distributed regressors (QR corresponds to 0% trimming). One can see that, when increasing the trimming percentage, the intercept estimates are unbiased for $\tau = 0.5$ (the error distribution is symmetric), but the bias for the regression quantiles $\tau = 0.75$ and $\tau = 0.90$ increases with the amount of trimming. The reason for this effect was given in Section 4, where we noted that LTQR identifies the sum of the intercept and $\mu^*(\tau) = F^{-1}(a^*(\tau) + \tau\lambda)$ (see equation (7.6)), which depend both on the quantile τ and the amount of trimming $\lambda = \lim_{n\to\infty} k_n/n$. The estimates of the slopes are presented on the lower plot panels. Both QR and LTQR estimators are unbiased in agreement with the theory and perform well, although the variability of the estimates increases for larger amounts of trimming. This is caused by LTQR using less and less observations due to a higher amount of trimmed data points.

Figures 7.4–7.9 present the results for the *1st* and *2nd experiment* corresponding to an increasing proportion of outliers $\epsilon = 0.1, 0.2, 0.3$. When choosing the same trimming percentage as the contamination level, the resulting estimates are very precise – comparable to the uncontaminated case. Similarly, if the trimming is chosen higher than the contamination level (i.e., $1 - \lambda \geq \epsilon$), we observe essentially the same picture as for the uncontaminated case. On the other hand, the use of smaller trimming percentages (i.e., $1 - \lambda < \epsilon$) has an immediate effect on the quality of the estimates and this becomes more severe for high contamination levels. In such cases, both bias and variance of the estimates increase dramatically because the resulting procedure has not sufficient robustness.

Further, these boxplots on Figures 7.4–7.9 also reveal that the variation in any panel depends on the chosen trimming percentage. In general, the smallest variation is obtained by choosing the exact trimming percentage corresponding to the contamination level in the data. In practice, it is preferable to be conservative, and in case of doubts, choose a higher trimming proportion than necessary (and thus a bit higher variance of estimates) to prevent a substantial bias caused by the lack of robustness.

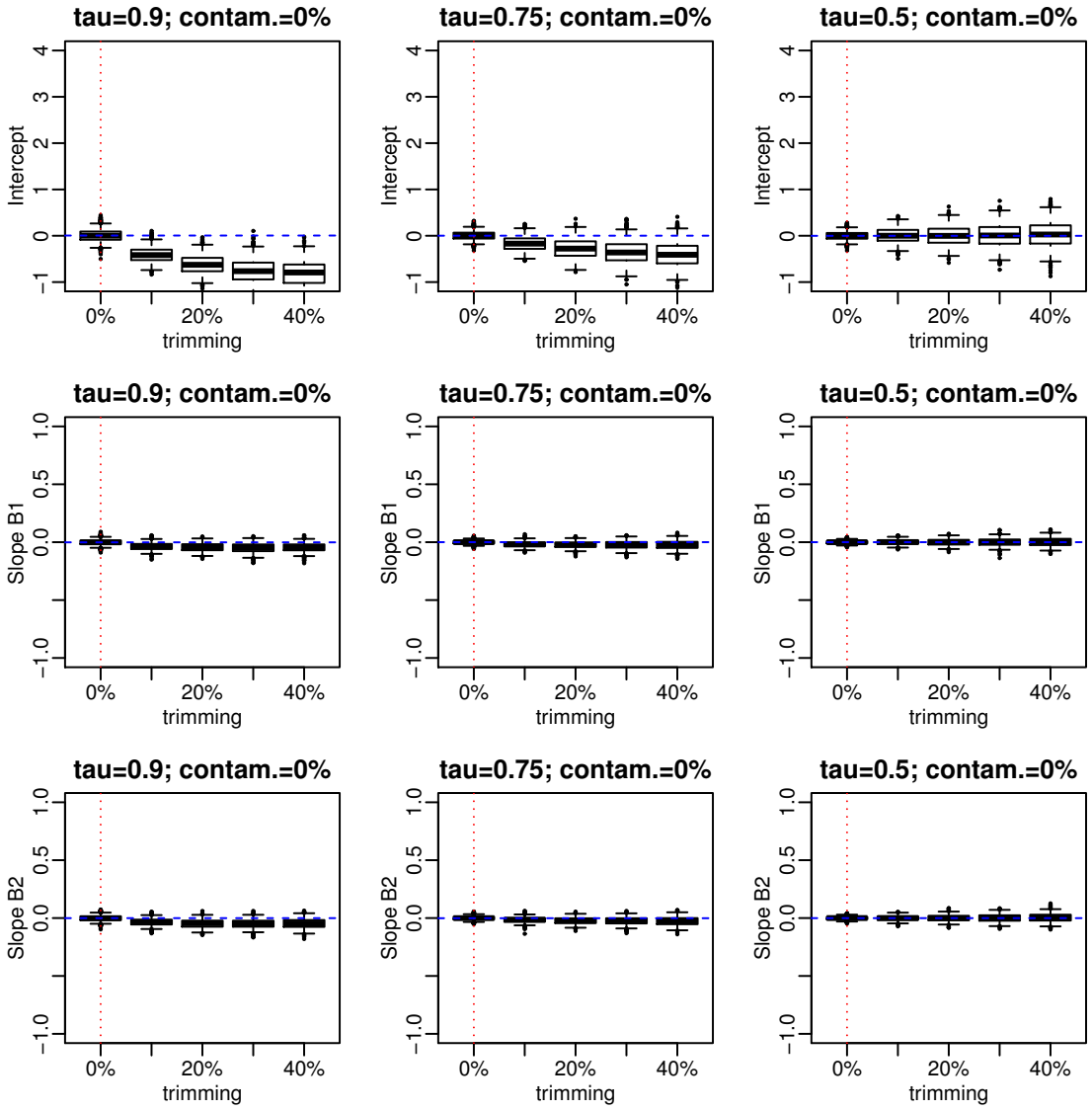

Figure 7.2: Boxplots of the estimates based on the originally generated data (0% contamination) and uniformly distributed covariates.

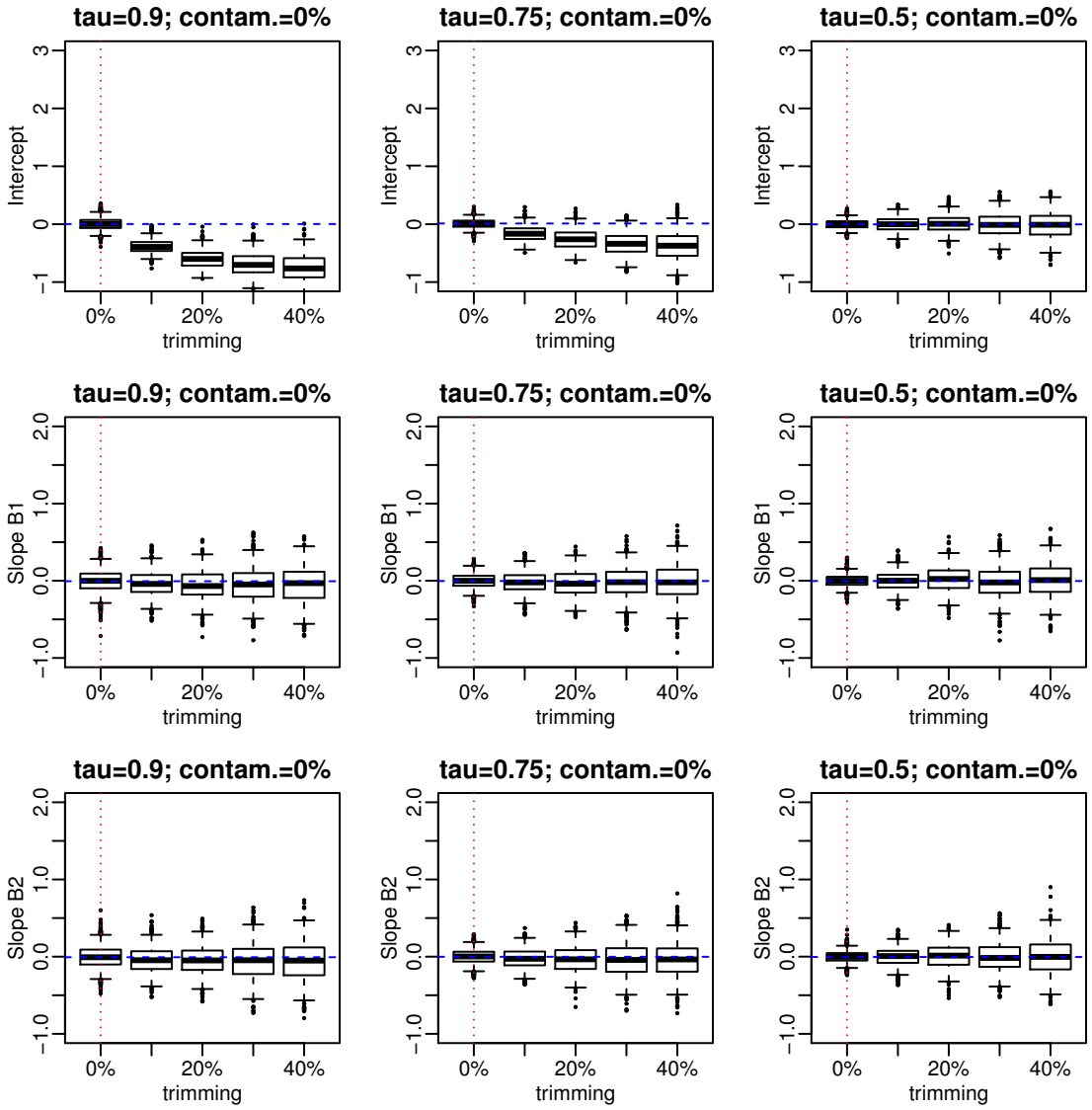

Figure 7.3: Boxplots of the estimates based on the originally generated data (0% contamination) and normally distributed covariates.

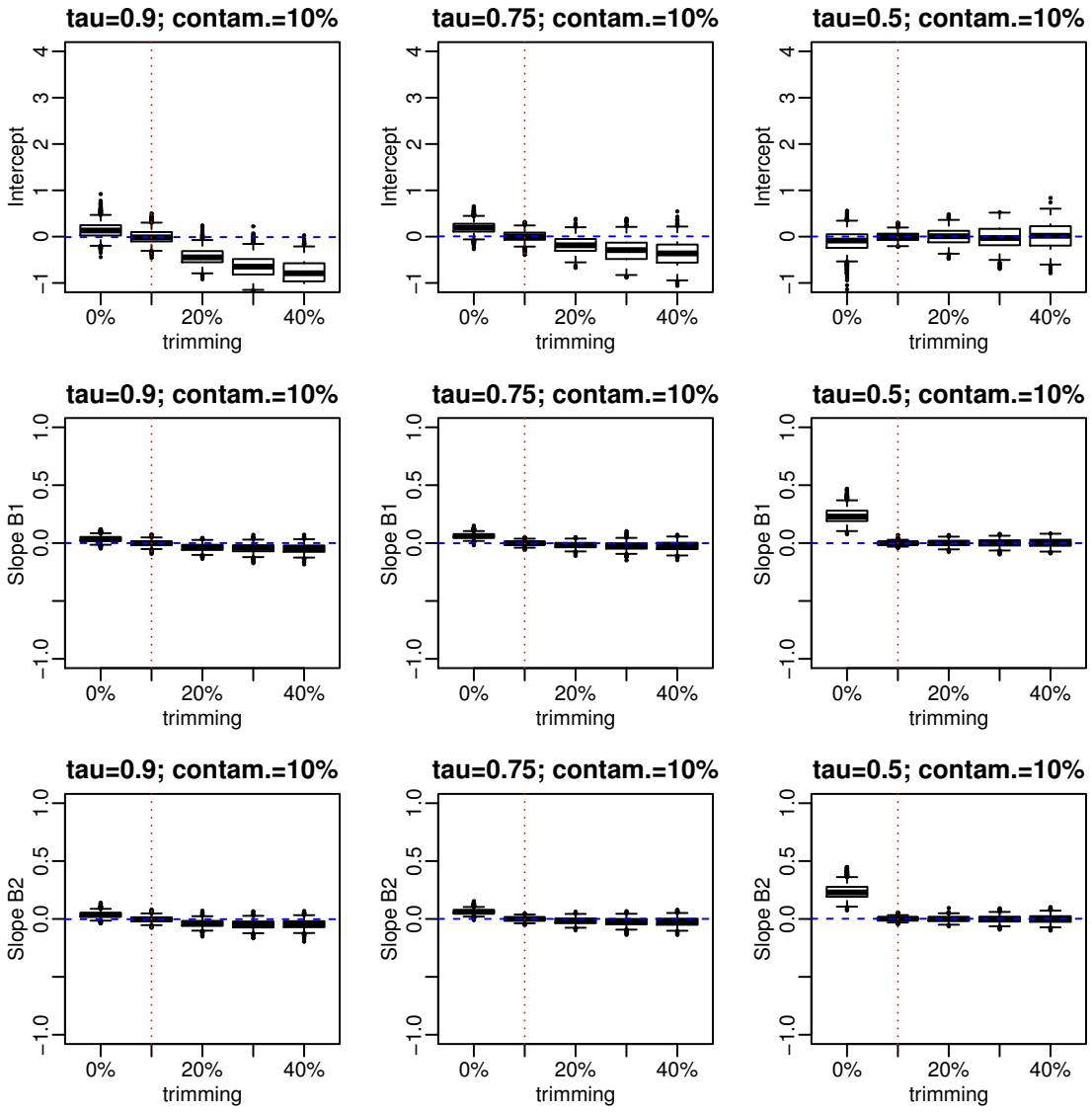

Figure 7.4: Boxplots of the estimates for the *1st experiment* with 10% contamination and uniformly distributed covariates.

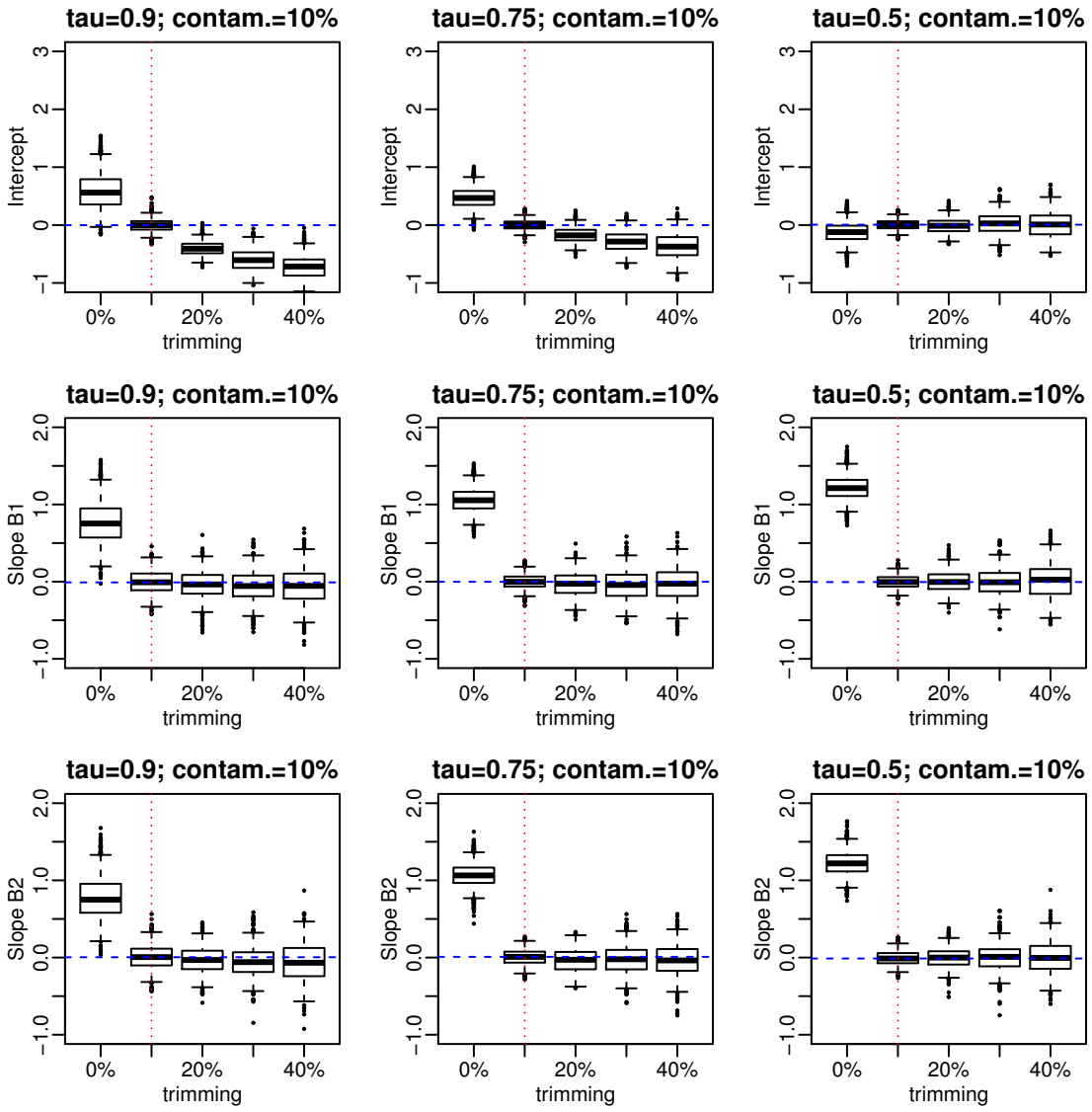

Figure 7.5: Boxplots of the estimates for the *2nd experiment* with 10% contamination and normally distributed covariates.

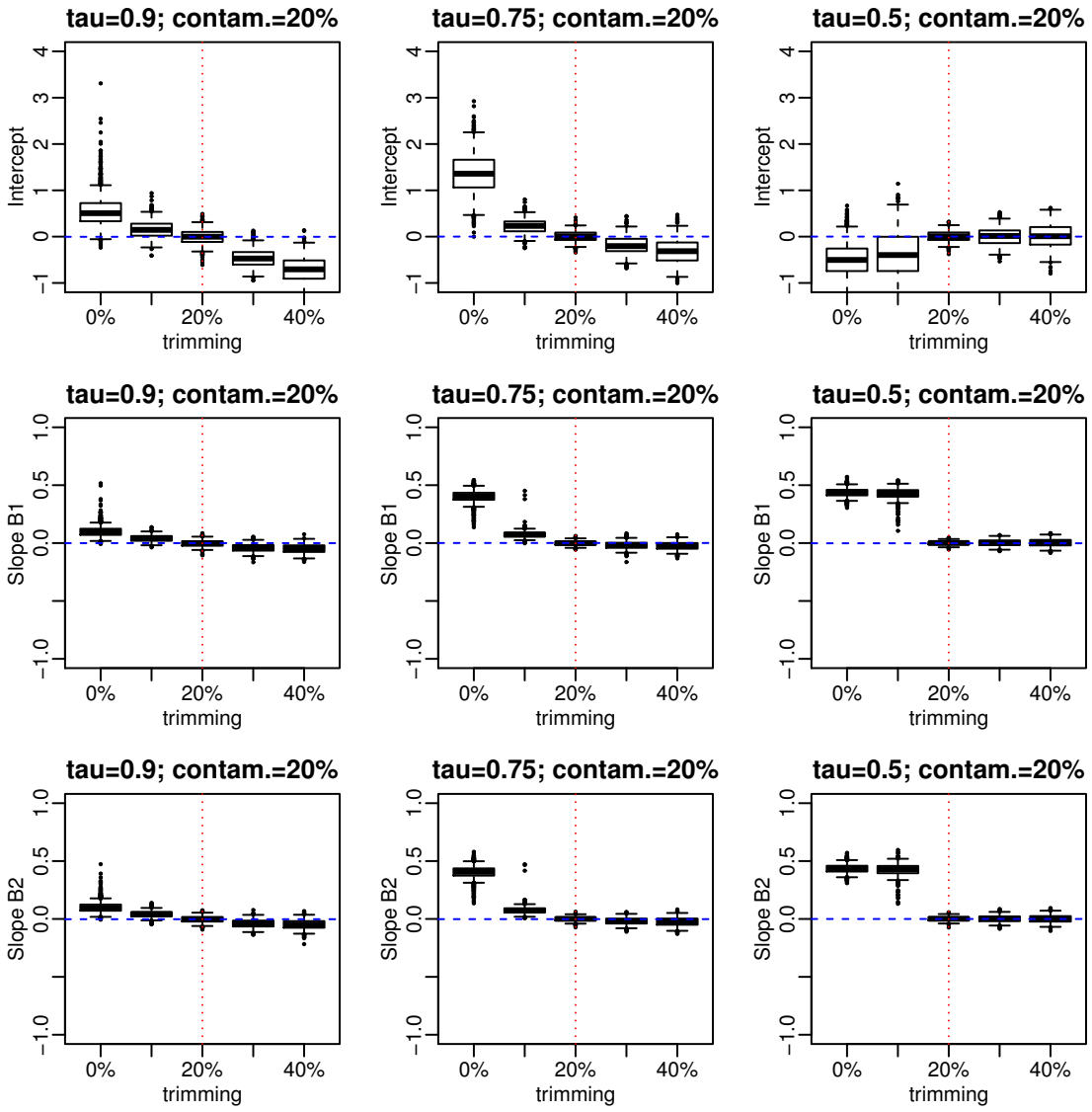

Figure 7.6: Boxplots of the estimates for the *1st experiment* with 20% contamination and uniformly distributed covariates.

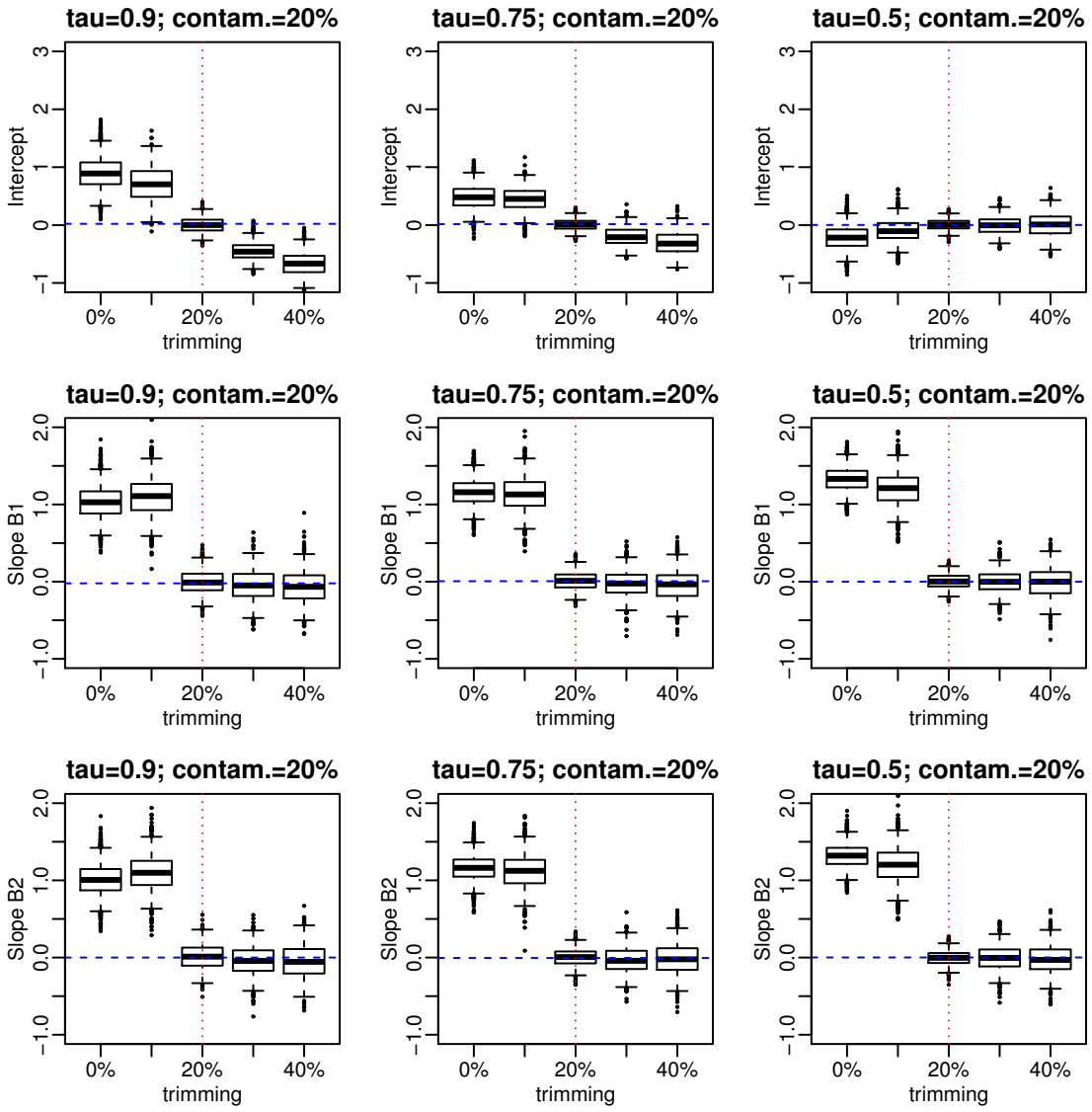

Figure 7.7: Boxplots of the estimates for the *2nd experiment* with 20% contamination and normally distributed covariates.

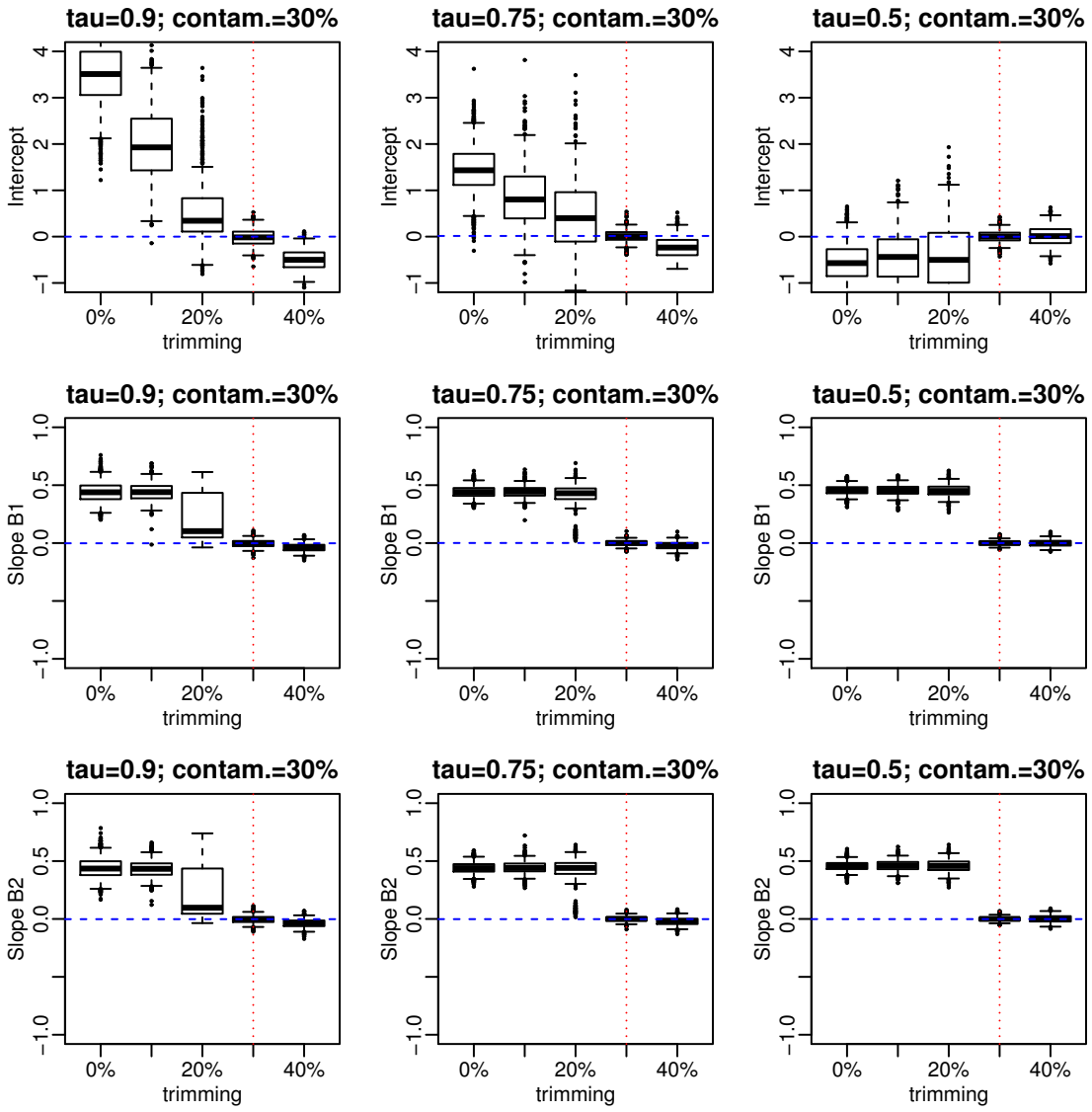

Figure 7.8: Boxplots of the estimates for the *1st experiment* with 30% contamination and uniformly distributed covariates.

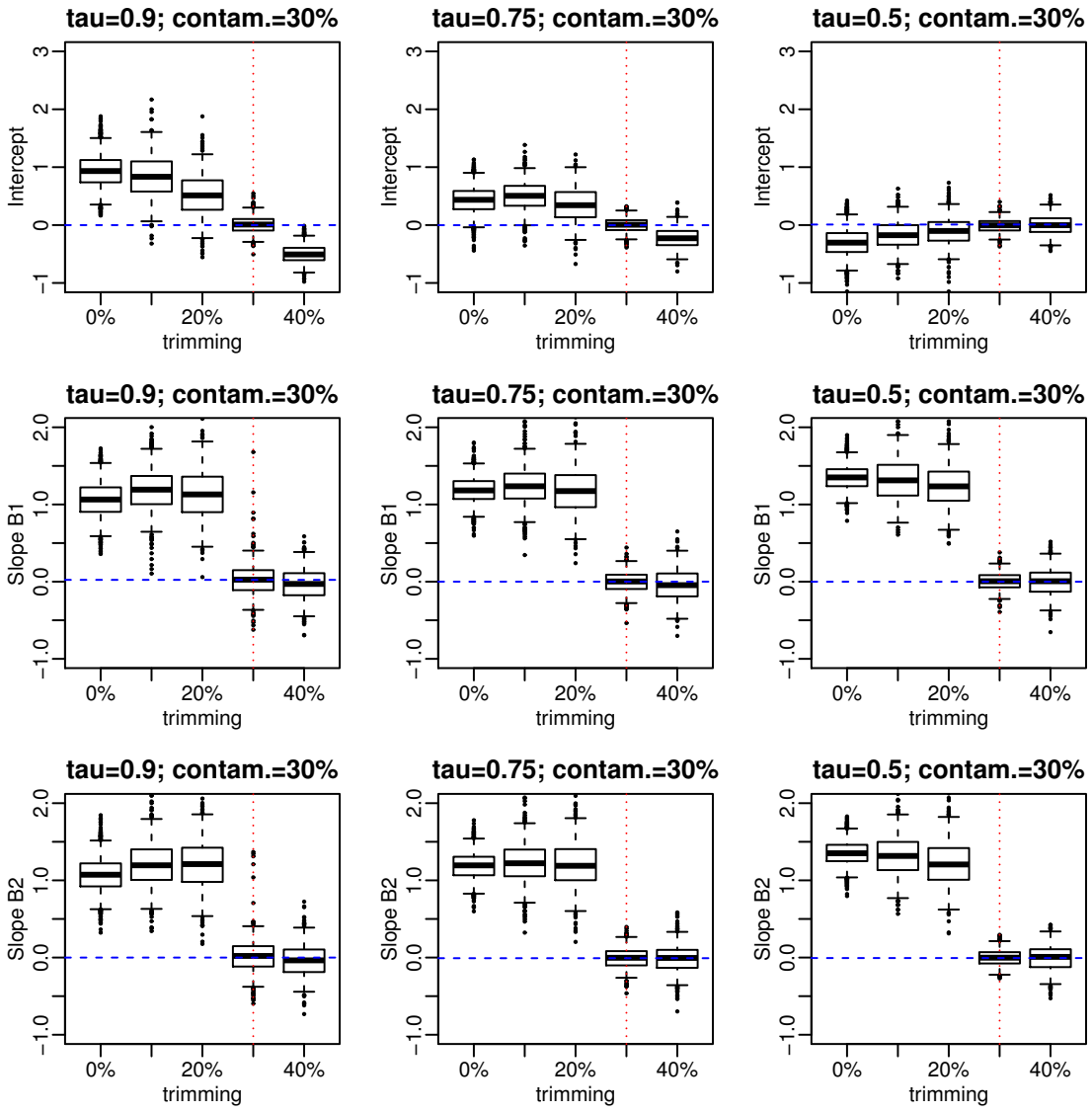

Figure 7.9: Boxplots of the estimates for the *2nd experiment* with 30% contamination and normally distributed covariates.

The percentage of outliers in real data is of course unknown or can be only roughly estimated. A technique for automatically selecting the trimming parameter $k_n = [\lambda n]$ or the trimming percentage $[(1-\lambda)n]100\%$ can be developed for LTQR in a straightforward way by mimicking the procedure of Čížek (2010) for the (reweighted) LTS regression estimator, which is in turn an adaptation of the method of Gervini and Yohai (2002). One only needs to take the asymmetric Laplace distribution instead of the normal one.

7.5.3 Comparison with other robust regression quantile estimators

In order to study the small sample behavior of the proposed estimator and to compare it with some robust regression quantile estimators considered by Rousseeuw and Hubert (1999) and Adrover et al. (2004), we estimated the maximum effect of the point contamination on the LTQR estimator in terms of the mean squared errors. The simulation experiment closely follows that of Adrover et al. (2004). Each experiment consists of $n = 50$ data points. The data generation is based on a multiple linear regression model $y_i = \beta_0^T \mathbf{x}_i + u_i$ for $i = 1, \ldots, n$, where $\mathbf{x}_i^T := (1, x_{i1}, \ldots, x_{i,p-1})$, the $p-1$ covariates follow independent unit normals, $u_i \sim N(z_\alpha, 1)$, z_α is the α–quantile of $N(0,1)$, and $\beta_0 = 0$, which can be chosen without loss of generality because of the regression equivariance of LTQR. The point contamination is introduced via replacement of the last $m = \lfloor \epsilon n \rfloor$ observations by the following outliers: $y_i := y_0 = 5b$ and $\mathbf{x}_i^T := \mathbf{x}_0^T = (1, 5\mathbf{e}_1^T) \in R^p$ for $i = n-m+1, \ldots, n$ for various values of ϵ (see Table 7.1), where $\mathbf{e}_1$ is the first element of the canonical basis of R^{p-1}. As in Adrover et al. (2004), the contamination slope b varies over a large grid from 0 to 10 with step 0.1 in order to search for the least favorable situations, and the experiment was performed for the number of regressors $p = 2$ and $p = 5$. Each simulation experiment was replicated 500 times. The trimming parameter k_n of the LTQR estimator is set equal to $\lfloor (1-\epsilon)n \rfloor$ and also to a slightly lower value $\lfloor (1 - \epsilon - 0.1 * (1-\tau))n \rfloor$. The maximum mean of the squared errors $\|\hat{\beta}_n^{k_n} - \beta_0\|^2$ (MaxMSE) is used as an error criterion and its values for different quantiles are presented in Table 7.1.

In the case of LTQR, MaxMSE is computed for the slope coefficients as well as for all coefficients including intercept: as LTQR does not consistently estimate the intercept, the MaxMSE computed for the whole coefficient vector necessarily reflects also its intercept bias, which is not directly related to the bias due to contamination. Additionally, we include results for the LTQR using larger percentages of trimming $1 - \lambda = \epsilon + 0.1 * (1 - \tau)$ than the contamination levels. This is done because in practice the exact contamination level is unknown (even though it can be estimated as discussed in the previous section) and thus a more realistic procedure is to use a larger trimming than expected contamination. These values for LTQR can be compared to the results for the classical Koenker–Bassett (K-B), the robustified Koenker–Bassett (RobKB) estimator of Adrover et al. (2004), and other alternatives such as the maximum depth estimator (MaxDep) of Rousseeuw and Hubert (1999). We implemented those estimators, but – because of the limited information concerning the estimation algorithms – we also report the original results of Adrover et al. (2004), who report the median of the squared errors. The FORTRAN software developed by Van Aelst et al. (2002) was used to handle the MaxDep computations.

As the LTQR estimator for $1 - \lambda = 0$, that is for $k_n = n$, reduces to the classical Koenker–Bassett estimator, its values for no contamination case $\epsilon = 0$ can be used as a reference for the mean squared errors of QR, see Table 7.1. For the positive levels of contamination, the LTQR performance is proportional to the level of contamination ϵ, but does not depend substantially on the estimated quantile. This is most visible if the contamination level $\epsilon \in (0.10, 0.12)$ is considered: the MaxMSE of LTQR increases only by 50% if $\tau = 0.50$ grows to 0.90. On the other hand, the most robust alternative RobKB (cf. Adrover et al., 2004) increases its bias $\epsilon \in (0.10, 0.12)$ by 50–100% if $\tau = 0.50$ changes to 0.75 and grows above any bound if $\tau = 0.90$ as it reaches its breakdown point 10% at this point. Thus, LTQR generally outperforms the other methods for $\tau > 0.5$ and higher contamination levels as a consequence of its breakdown point independent of the actual quantile (although BDP itself does not quantify the bias of an estimator). Additionally, LTQR performs better than competing methods at higher quantiles such as $\tau = 0.90$ irrespective of the contamination level. On the other hand, RobKB and MaxDep

perform better for quantiles closer to $\tau = 0.50$ and lower contamination levels. This can be expected especially in the case of RobKB as methods based on the pairwise differences or comparisons of observations typically outperform the corresponding methods minimizing plain sums of functions of individual observations (e.g., compare least trimmed squares and least trimmed differences estimator of Stromberg et al., 2000).

7.6 Summary and conclusions

A robust version of the linear quantile regression estimator is introduced which is based on the idea of trimming. The breakdown point and the consistency of the proposed estimator are characterized. The computation of the estimator is taking advantage of the same technology as used for its classical counterpart, but here the estimation is based on subsamples only. The used algorithm consisting of a trial and a refinement step (Neykov et al., 2012) follows the ideas of the FAST-LTS and FAST-MCD algorithms of Rousseeuw and van Driessen (1999, 2006) and Neykov and Müller (2003). The new estimator generally performs very well, which is confirmed by an example, by simulation studies, and by a comparison to other proposals.

An important choice for estimators based on trimming is the trimming percentage. In the numerical experiments, it has been shown that a trimming percentage lower than the contamination level can lead to very poor estimates, but any higher trimming percentage gives very reasonable results. Therefore, a general rule is to work with a conservative choice of the trimming percentage or to estimate the amount of trimming similarly to Čížek (2010) and Gervini and Yohai (2003).

Appendix: Proofs

The BDP will be derived using the d-fullness technique proposed by Vandev (1993). According to Vandev and Neykov (1998), the set $F = \{f_i(\theta); i = 1, \ldots, n\}$ is called d-full if the function $g(\theta) = \max_{j \in J} f_j(\theta)$, $\theta \in \Theta$, is subcompact for every subset $J \subset \{1, \ldots, n\}$

Table 7.1: The Monte Carlo maximum mean squared errors based on the LTQR estimator of the slope (LTQR 'no intercept'), LTQR estimator of the intercept and slope (LTQR with 'intercept'), maximum depth estimator (MaxDep), robustified Koenker-Bassett (RobKB) estimator, and Koenker-Bassett (K-B) estimator in samples of $n = 50$ observations. For LTQR, the trimming parameter equals $k_n = \lfloor \lambda n \rfloor$, where the fraction of trimmed observations $1 - \lambda$ equals exactly to the contamination level ϵ or is slightly larger $\epsilon + \delta$, where $\delta = 0.1 * (1 - \tau)$.

MaxMSE			LTQR [trimming $1-\lambda$]				MaxDep	RobKB	K-B	MaxDep	RobKB
			no intercept		intercept						
p	τ	ϵ	$[\epsilon]$	$[\epsilon+\delta]$	$[\epsilon]$	$[\epsilon+\delta]$			results of Adrover et al. (2004)		
2	0.50	0.00	0.04	0.05	0.08	0.09	0.10	0.09	0.07	0.10	0.10
		0.10	0.45	0.39	0.53	0.48	0.19	0.18		0.21	0.20
		0.20	1.03	0.94	1.20	1.11	0.84	0.81		0.79	0.97
	0.75	0.00	0.05	0.06	0.09	0.12		0.11	0.07	0.10	0.11
		0.06	0.28	0.25	0.34	0.31		0.16		0.24	0.21
		0.12	0.66	0.55	0.76	0.64		0.38		1.82	0.57
		0.20	1.07	0.94	1.24	1.10		1.43			
	0.90	0.00	0.07	0.08	0.15	0.15		0.16	0.11	0.14	0.14
		0.02	0.15	0.12	0.22	0.19		0.17		0.23	0.16
		0.04	0.25	0.20	0.31	0.28		0.26		0.77	0.26
		0.08	0.53	0.42	0.61	0.52		1.16			
		0.10	0.64	0.54	0.74	0.63		117.			
5	0.50	0.00	0.17	0.20	0.21	0.24	0.26	0.29	0.19	0.37	0.38
		0.10	0.78	0.77	0.87	0.87	0.65	0.61		0.74	0.66
		0.20	2.64	2.65	2.93	2.95	4.14	2.89		4.60	2.40
	0.75	0.00	0.20	0.23	0.25	0.29		0.38	0.21	0.36	0.36
		0.06	0.54	0.48	0.61	0.55		0.46		0.77	0.54
		0.12	1.22	1.10	1.33	1.21		0.89		2.87	1.70
		0.20	2.91	2.78	3.24	3.07		4.13			
	0.90	0.00	0.30	0.31	0.37	0.42		0.46	0.34	0.48	0.38
		0.02	0.40	0.34	0.47	0.45		0.52		0.71	0.44
		0.04	0.54	0.46	0.61	0.56		0.66		3.00	0.87
		0.08	0.98	0.81	1.08	0.93		2.86			
		0.10	1.24	1.04	1.35	1.17		119.			

of cardinality d. A function $g : \Theta \to \mathbb{R}$, $\Theta \subseteq \mathbb{R}^q$, is called subcompact if its Lebesgue set $L_g(C) = \{\theta \in \Theta : g(\theta) \leq C\}$ is contained in a compact set for every real constant C. The d-fullness index ensures the existence of a solution and provides positive BDP of the optimization problem (7.3) at any subset of functions with size $k \geq d$. **Proof of Theorem 7.1:** As the linear LTQR estimator is a particular case of the GTE, its finite-sample BDP equals to $\frac{1}{n}\min\{n-k, k-d\}$, provided the set of functions $F = \{\rho_\tau(r_i(\beta)); i = 1, \ldots, n\}$ is d-full, Dimova and Neykov (2004). Now we are ready to show that the set F is $d = \mathcal{N}(X) + 1$-full for any fixed $\tau \in (0, 1)$, where $\mathcal{N}(X)$ is defined as $\mathcal{N}(X) = \max_{0 \neq \beta \in \mathbb{R}^p} \text{card}\{i \in \{1, \ldots, n\}; x_i^T \beta = 0\}$. Let D be an arbitrary constant and τ be fixed. Then for any subset $J \subset \{1, \ldots, n\}$ of cardinality $\mathcal{N}(X) + 1$, the set

$$
\begin{aligned}
& \{\beta \in R^p : \max_{j\in J} \rho_\tau(x_j^T\beta - y_j) \le D\} \\
= & \left\{\beta \in R^p : \max_{j\in J} \left[\left|x_j^T\beta - y_j\right| \left(\tau 1_{\left\{x_j^T\beta - y_j \ge 0\right\}} + (1-\tau)\, 1_{\left\{x_j^T\beta - y_j < 0\right\}}\right)\right] \le D\right\} \\
\subseteq & \left\{\beta \in R^p : \min(\tau, 1-\tau) \max_{j\in J} \left|x_j^T\beta - y_j\right| \le D\right\} \\
= & \left\{\beta \in R^p : \max_{j\in J} \left|x_j^T\beta - y_j\right| \le \frac{D}{\min(\tau, 1-\tau)}\right\} \\
\subseteq & \left\{\beta \in R^p : \max_{j\in J} \left(\left|x_j^T\beta\right| - |y_j|\right) \le \frac{D}{\min(\tau, 1-\tau)}\right\} \\
\subseteq & \left\{\beta \in R^p : \max_{j\in J} \left|x_j^T\beta\right| \le \frac{D}{\min(\tau, 1-\tau)} + \max_{j\in J} |y_j|\right\} \\
\subseteq & \left\{\beta \in R^p : \frac{1}{N(X)+1}\beta^T \sum_{j\in J} x_j x_j^T \beta \le \left[\frac{D}{\min(\tau, (1-\tau))} + \max_{j\in J} |y_j|\right]^2\right\}
\end{aligned}
$$

is contained in a compact set. Indeed, the last set is bounded because J is of cardinality $\mathcal{N}(X)+1$ and the definition of $\mathcal{N}(X)$ implies that the matrix $\sum_{j\in J} x_j x_j^T$ has full rank. This last set is closed as the quadratic form $\beta^T \sum_{j\in J} x_j x_j^T \beta$ is a continuous function in β. Hence it is compact because it is closed and bounded.

Therefore, the finite sample BDP of the linear LTQR estimator is $\frac{1}{n}\min\{n-k, k-\mathcal{N}(X)-1\}$. This BDP is maximized for $\lfloor\{n+\mathcal{N}(X)+1\}/2\rfloor \le k \le \lfloor\{n+\mathcal{N}(X)+2\}/2\rfloor$ and equals to $\frac{1}{n}\lfloor\{n-\mathcal{N}(X)-1\}/2\rfloor$. □

Proof of Theorem 7.2: Let $Q_\lambda(\beta) = E[\rho_\tau(r_i(\beta)) \cdot 1_{\{\rho_\tau(r_i(\beta)) \le G_\beta^{-1}(\lambda)\}}]$ be the asymptotic form of $Q_{n,k_n}(\beta)$ defined in (7.5), where G_β and G_β^{-1} are the distribution and quantile functions of $\rho_\tau(r_i(\beta))$ (the uniform convergence of $Q_{n,k_n}(\beta)$ to $Q_\lambda(\beta)$ under Assumptions D and F is derived in Lemmas 2.1 and A.1 of Čížek, 2008).

Now, considering an interval $\Delta(a,\lambda) = \langle F^{-1}(a), F^{-1}(a+\lambda)\rangle$ for $a \in (0, 1-\lambda)$ and a fixed $\tau \in (0,1)$, Tableman (1994a, p. 390) proved in the location model that $Q_\lambda(\mu)$ applied to univariate data following the distribution function F has a unique minimum at $\mu^*(\tau) = F^{-1}(a^*(\tau)+\tau\lambda)$, where $a^*(\tau) = \arg\min_{a\in(0,1)} \int_{\Delta(a,\lambda)} \rho_\tau(\varepsilon - F^{-1}(a+\tau\lambda))dF(\varepsilon)$ (the proof is given for $\tau = 0.5$ by an argument, which directly applies also to general $\tau \in (0,1)$).

This result can be employed in the regression model (7.1). Conditionally on a given

value of covariates x, the distribution function of the response y equals $F_{y|x}(t) = F(t - x^T\beta^0), t \in R$, and the corresponding quantile function equals $F^{-1}_{y|x}(u) = F^{-1}(u) + x^T\beta^0$. Therefore, $Q_{\lambda|x}(\beta) = E[\rho_\tau(r_i(\beta)) \cdot 1_{\{\rho_\tau(r_i(\beta)) \leq G^{-1}_\beta(\lambda)\}}|x]$ is minimized at $\mu^*(\tau) + x^T\beta^0$ (conditionally on x). Consequently, $Q_\lambda(\beta)$ is minimized at $\beta^*(\tau) = (\mu^*(\tau), 0, \ldots, 0)^T + \beta^0$ unconditionally if the intercept is supposed to be the first element of the parameter vector β.

The limit $Q_\lambda(\beta)$ of the LTQR estimator objective function identifies the parameter vector $\beta^*(\tau)$. To prove the consistency of the LTQR estimator, we can apply now Theorem 3.1 of Čížek (2008): since we verified the identification condition for β^*, assume Assumptions D and F, and $\{\rho_\tau(r_i(\beta)) = \max\{\tau r_i(\beta), -(1-\tau)r_i(\beta)\}|\beta \in R^p\}$ forms a Vapnik-Chervonenkis class of functions (van der Vaart and Wellner, 1996, Lemmas 2.6.15 and 2.6.18), we only need to check Assumption D3 of Čížek (2008). This is however verified under Assumptions D and F by Lemma 2 of Čížek (2006). Thus, Theorem 3.1 of Čížek (2008) implies the claim of the theorem. □

Chapter 8

Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator

Summary. The Penalized Maximum Likelihood Estimator (PMLE) has been widely used for variable selection in high-dimensional data. Various penalty functions have been employed for this purpose, e.g., Lasso, weighted Lasso, or smoothly clipped absolute deviations (SCAD). However, the PMLE can be very sensitive to outliers in the data, especially to outliers in the covariates (leverage points). In order to overcome this disadvantage, the usage of the Penalized Maximum Trimmed Likelihood Estimator (PMTLE) is proposed to estimate the unknown parameters in a robust way. The computation of the PMTLE takes advantage of the same technology as used for PMLE but here the estimation is based on subsamples only. The breakdown point properties of the PMTLE are discussed using the notion of d–fullness. The performance of the proposed estimator is evaluated in a simulation study for the classical multiple linear and Poisson linear regression models.

8.1 Introduction

Let (y_i, x_i^T), for $i = 1, \ldots, n$, be identically and independently distributed observations, where y_i is the ith observation of the response variable Y and $x_i^T = (x_{11}, x_{12}, \ldots, x_{1p})$ is the ith row of the covariates matrix X. Assume that y_i depends on x_i through a linear predictor $\eta_i(\theta) = x_i^T\theta$ via the objective function $L(\eta_i(\theta), y_i)$. For instance, $L(\eta_i(\theta), y_i)$ might be a probabilistic model such as likelihood, quasi-likelihood or another discrepancy function related with the ith observation. Without loss of generality, we shall assume that $L(\eta_i(\theta), y_i)$ is the log-likelihood. The Maximum Likelihood Estimator (MLE) is defined as

$$\hat{\theta}_{n,MLE} := \arg\max_{\theta} \left\{ \ell_n(\theta) = \sum_{i=1}^{n} L(\eta_i(\theta), y_i) \right\}. \tag{8.1}$$

The Penalized MLE (PMLE) is defined as

$$\hat{\theta}_{n,PMLE} := \arg\max_{\theta} \left\{ \ell_n(\theta) - n\sum_{j=1}^{p} p_\lambda(|\theta_j|) \right\}. \tag{8.2}$$

Here, $p_\lambda(.)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$. Due to the penalty function, some of the components of θ are shrunk to zero automatically and thus variables selection is performed. A large value of λ tends to choose a simple model whereas a small value of λ inclines to a complex model. In real applications the parameter λ is not known. It may be chosen by cross-validation or using an information criterion like the Bayesian Information Criterion (BIC), see Bühlmann and van der Geer (2011).

Commonly used penalty functions are the L_1 penalty $p_\lambda(|\theta_j|) = \lambda\,|\theta_j|$ called LASSO (least absolute shrinkage and selection operator) by Tibshirani (1996), the L_q-norm penalty $p_\lambda(|\theta|) = \lambda\,|\theta_j|^q$ for $0 < q \leq 2$, (Frank and Friedman, 1993), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), which is a quadratic spline

$$p_\lambda(|\theta|) = \begin{cases} \lambda|\theta| & \text{if } |\theta| < \lambda, \\ \frac{(a^2-1)\lambda^2 - (|\theta|-a\lambda)^2}{2(a-1)} & \text{if } \lambda \leq |\theta| < a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\theta| \geq a\lambda, \end{cases} \tag{8.3}$$

where $a = 3.7$, or the minimum concavity penalty (MCP) $p'_\lambda(|\theta_j|) = (\lambda - |\theta|/a)_+$ considered by Zhang (2008). The SCAD and MCP are non-convex penalty functions which

possess the oracle property. This means that the important variables can be correctly selected with a high probability whereas the remaining variables will be dropped from the model. Antoniadis et al. (2011) gave a discussion about many other penalty functions and selection criteria for the regularization parameter λ for the generalized linear models (GLMs) framework.

The problem (8.2) is a convex optimization problem if the $\ell_n(\theta)$ is concave and the L_1 penalty is used. In general, for fixed parameter λ, the penalized likelihood with SCAD penalty function is non-convex and thus special algorithms have been developed to obtain a solution. For instance, Zou and Li (2008) propose an effective locally linear approximation algorithm (LLA) for optimization of (8.2) with the SCAD penalty function. The idea is to approximate (majorize) the SCAD function by a linear function at the mth iteration

$$p_\lambda(|\theta|) \approx p_\lambda(|\theta^{(m)}|) + p'_\lambda(|\theta^{(m)}|)(|\theta| - |\theta^{(m)}|). \tag{8.4}$$

As a consequence the penalized maximum likelihood (8.2) reduces to

$$\ell_n(\theta) - n \sum_{j=1}^{p} w_j^{(m)} |\theta_j| , \tag{8.5}$$

where $w_j^{(m)} = p'_\lambda(|\theta_j^{(m)}|)$. By the quadratic approximation of $\ell_n(\theta)$ at $\theta^{(m)}$ this optimization problem becomes weighted L_1 penalized least squares closely related with the adaptive LASSO estimation procedure (Zou, 2006) that produce sparse fits and performs variable selection automatically. The LLA algorithm is implemented as a function in the R (R Development Core Team, 2012) package SIS of Fan et al. (2009). Zou and Li (2008) discussed also other iterative approaches for solving the corresponding weighted L_1 penalized least squares problem efficiently by the least angle regression (LARS) algorithm (Efron et al., 2004).

For further consideration we use a well known result of Green (1984) concerning computational aspects of the MLE in fitting probabilistic regression models. Because the log-likelihood $\ell_n(\theta)$ is a composite function of the linear predictors $\eta_i(\theta)$, the Fisher scoring algorithm for maximization of $\ell_n(\theta)$ reduces to an iteratively re-weighted least squares

(IRLS) algorithm. Thus the optimization problem (8.5) at the $(m+1)$th iteration can be replaced by the following weighted least squares problem with weighted L_1 penalty

$$(z^{(m)} - X\theta)^T A^{(m)} (z^{(m)} - X\theta) + n \sum_{j=1}^{p} w_j^{(m)} |\theta_j| , \tag{8.6}$$

where $z = A^{-1}u + \eta$ is an adjusted dependent variable, $u = (\partial \ell_n(\theta)/\partial \eta)$, $\eta = X\theta$ and $A = (uu^T)$, and all these elements are evaluated at the current value $\theta^{(m)}$. The working weight matrix A is diagonal as the observations are independent.

Hence an efficient standard regression procedure with L_1 penalty, e.g., based on the LARS algorithm of Efron et al. (2004) or the coordinate descent algorithm (Friedman et al., 2007; Friedman et al., 2010), can be adapted to calculate $\hat{\theta}_{n,PMLE}$ via an IRLS algorithm. A discussion about the applicability and implementation of these two approaches for the penalized logistic regression model with the LLA majorant (surrogate) of the SCAD penalty function is presented by Breheny and Huang (2011). Computational algorithms within high-dimensional settings are discussed also in Bühlmann and van der Geer (2011), and also in the review paper of Fan and Lv (2010).

It is well known that the least squares estimator, the MLE and quasi-likelihood estimators can be highly sensitive to a small proportion of observations that departs from the model (Huber, 1981; Hampel et al., 1986; Maronna et al., 2006). Therefore the penalized least squares estimator and MLE are non-robust against outliers in the data too. To overcome this problem, the penalized M-estimator has been employed (Fan and Li, 2001; Fan and Lv, 2010). However, within regression models, M-estimators are not robust against outlying observations in the covariates, the so called leverage points, and therefore penalized M-estimators are not robust in such settings as well. We remind that only some redescending M-estimators are robust in linear regression settings with fixed designs (Mizera and Müller, 1999).

Several robust alternatives of the MLE that are robust simultaneously against outliers in the response and covariates have been developed, e.g., the weighted MLE of Markatou et al. (1997) and the maximum Trimmed Likelihood Estimator (TLE) of Neykov and Neytchev

(1990). To our knowledge, none of these estimators have been used in high dimensional data modeling. Thus, the goal of this chapter is to develop an alternative of the penalized MLE for variable selection based on the penalized maximum TLE (PMTLE) in order to reduce the influence of the outliers in the covariates. The TLE is looking for that subsample of $k > n/2$ observations out of n with the optimal likelihood. The trimming number of observations can be chosen by the user in appropriate bounds to get a high breakdown point (BDP) and optimal efficiency. Because the TLE accommodates the classical MLE, the variable selection methodology, which is based mainly on the penalized MLE, can be adapted and further developed. In this chapter the superiority of this approach in comparison with the penalized MLE is illustrated.

The paper is organized as follows. In Section 2 we define the Generalized Trimmed Estimator (GTE), consider its penalized version and characterize its finite sample BDP. The applicability of the PMTLE is considered for the iterative sure independence screening (ISIS) framework of Fan et al. (2009) in Section 3. In Section 4 a simulation study is performed to illustrate the effectiveness of the proposed estimator in comparison with the ISIS procedure for the classical multiple and Poisson linear regression models. Finally, conclusions are given in Section 5.

8.2 Penalized maximum trimmed likelihood estimator

For introducing the Penalized Maximum Trimmed Likelihood Estimator (PMTLE), we first need to review the definition and some properties of the Generalized Trimmed Estimator (GTE) introduced by Vandev and Neykov (1998). Let $f_i : \Theta^p \to \mathbb{R}^+$, where $\Theta^p \subseteq \mathbb{R}^p$ is an open set.

Definition 1. The GTE $\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}}$ of θ is defined as the solution of the optimization problem

$$\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}} := \arg\min_{\theta\in\Theta^p} \left\{ S_{n,k}(\theta) = \min_{I\in I_k} \sum_{i\in I} f_i(\theta) \right\}, \tag{8.7}$$

where I_k is the set of all k-subsets of the index set $\{1,\ldots,n\}$ and k is the trimming constant.

The trimming parameter k determines the robustness properties of the GTE as $n-k$ functions with the largest values of $f_i(\theta)$ are excluded from the loss function. The BDP of the GTE is not less than $\frac{1}{n}\min\{n-k, k-d\}$ if the set $F=\{f_i(\theta) : i=1,\ldots,n\}$ is d-full. F is called d–full if for any subset of cardinality d of F, the supremum of this subset is a subcompact function. A real valued function $\varphi(\theta)$ is called subcompact if the sets $L_{\varphi(\theta)}(C)=\{\theta:\varphi(\theta)\le C\}$ are contained in a compact set for every real constant C. Details can be found in Vandev and Neykov (1998), Müller and Neykov (2003), and Dimova and Neykov (2004). Thus, if one wants to study the BDP of the GTE, one has to find the fullness parameter d of F and then the BDP can be exemplified by the range of values of k. The BDP is maximized for $k=\lfloor(n+d+1)/2\rfloor$, when it approximately equals $1/2$ for large n. Therefore, by selecting the value of k properly one can control the level of robustness of the GTE. Further, the asymptotic properties of the GTE estimator (8.7) were studied by Čížek (2008) for the case of twice differentiable functions $f_i(\theta)$.

The optimization problem (8.7) defining the GTE is of combinatorial nature,

$$\min_{\theta\in\Theta^p} S_{n,k}(\theta)=\min_{\theta\in\Theta^p}\min_{I\in I_k}\sum_{i\in I} f_i(\theta)=\min_{I\in I_k}\min_{\theta\in\Theta^p}\sum_{i\in I} f_i(\theta). \tag{8.8}$$

Therefore, it follows that all possible $\binom{n}{k}$ partitions of the set $\{f_1,\ldots,f_n\}$ have to be considered and $\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}}$ is defined by the partition with the minimal value of $S_{n,k}(\theta)$. Hence, an exact computation of the GTE is not feasible for large samples. To get an approximative GTE solution, an algorithm was developed by Neykov et al. (2012a). It repeatedly (i) sets $s=0$, selects a small subset $\{f_{i_1},\ldots,f_{i_{k^*}}\}$ of k^* functions from F and forms $I_s=\{i_1,\ldots,i_{k^*}\}$, (ii) minimizes the objective function $\sum_{i\in I_s} f_i(\theta)$ with respect to θ, and uses the obtained estimate $\hat{\theta}_s$, (iii) sets $s=s+1$, orders the functions of F in ascending order, $f_{\nu(1)}(\hat{\theta}_s)\le f_{\nu(2)}(\hat{\theta}_s)\le\ldots\le f_{\nu(k)}(\hat{\theta}_s)\le\ldots\le f_{\nu(n)}(\hat{\theta}_s)$, where $\nu(.)$ is the permutation of the indices $\{1,2,\ldots,n\}$, and forms $I_s=\{\nu(1),\ldots,\nu(k)\}$; the steps (ii) and (iii) are repeated as long as the newly obtained estimates $\hat{\theta}_s$ produce smaller values of the objective function $\sum_{i\in I_s} f_i(\theta)$.

The trial subsample size k^* should be greater than or equal to d, which is necessary for the existence of (8.7). However, the chance to get at least one good subsample of data points is larger if $k^* = d$. Obviously, only for very small samples all possible subsets of size $k^* = k$ can be considered to obtain the precise instead of an approximative solution. The algorithm could be further accelerated for large data sets by applying the partitioning and nesting techniques as discussed by Neykov et al. (2012a).

Particular cases of the GTE are the least trimmed squares (LTS) estimator (Rousseeuw, 1984) if $f(\theta)$ in (8.7) is replaced by the squared regression residuals, the least median of squares (Rousseeuw, 1984), the maximum trimmed likelihood estimator (TLE) (Neykov and Neytchev, 1990) if $f(\theta) = -L(\eta_i(\theta); y_i)$, the least trimmed quantile regression (Neykov et al., 2012b), the extended trimmed quasi-likelihood estimator (Neykov et al., 2012a), to name a few.

For high dimensional statistical optimization problems, where p is large in comparison with the sample size n, we need to consider a penalized version of the GTE.

Definition 2. The penalized GTE is defined as

$$\begin{aligned}
\min_{\theta\in\Theta^p} S^P_{n,k}(\theta) &= \min_{\theta\in\Theta^p}\left\{\min_{I\in I_k}\sum_{i\in I} f_i(\theta) + k\sum_{j=1}^{p} p_\lambda(|\theta_j|)\right\} && (8.9)\\
&= \min_{I\in I_k}\left\{\min_{\theta\in\Theta^p}\left[\sum_{i\in I} f_i(\theta) + k\sum_{j=1}^{p} p_\lambda(|\theta_j|)\right]\right\} && (8.10)\\
&= \min_{I\in I_k}\left\{\min_{\theta\in\Theta^p}\sum_{i\in I}\left[f_i(\theta) + \sum_{j=1}^{p} p_\lambda(|\theta_j|)\right]\right\}. && (8.11)
\end{aligned}$$

One can see that the penalized GTE refers to a penalized optimization problem, however, defined over all k-subsets. Thus the aforementioned algorithm can be used to obtain an approximate solution. For fixed λ, the BDP of the penalized GTE can be characterized via the d-fullness index of the set of functions $F_\lambda = \{f_i(\theta) + \sum_{j=1}^{p} p_\lambda(|\theta_j|),\ i = 1, \ldots, n\}$. Let F be d-full. Due to the inclusion

$$\{\theta \in R^p : \max_{j\in J}(f_j(\theta) + \sum_{l=1}^{p} p_\lambda(|\theta_l|)) \le C\} \subseteq \{\theta \in R^p : \max_{j\in J} f_j(\theta) \le C\}$$

it follows that F_λ is d-full because F is d-full. We see that the set F_λ is even 1-full provided

the set $\{\theta \in R^p : \sum_{l=1}^{p} p_\lambda(|\theta_l|) \leq C\}$ is contained in a compact set. For the convex penalty functions such as L_1 this is obvious, whereas for the non-convex function such as SCAD the generalized d-fullness technique (Dimova and Neykov, 2004) can be employed. From a computational point of view, the LLA defined by (8.4) can be used to get an approximate solution of the penalized GTE with SCAD penalty function. As the LLA is a convex majorant of the SCAD function, this ensures d-fullness of the corresponding set of functions F_λ. Therefore we conclude that a solution of the penalized GTE always exists if the set of functions F_λ is d-full. We note that this solution may not be unique and thus additional conditions are required to achieve this.

From the penalized GTE definition it follows that when $k = n$, and for suitable choices of $f_i(\theta)$ and $p_\lambda(.)$, we can derive different penalized estimators such as the LASSO of Tibshirani (1996), the penalized L_1-likelihood of Tibshirani (1997), the penalized likelihood with the SCAD function of Fan and Li (2001), the LAD-LASSO of Wang et al. (2007), or the penalized M-estimator (Fan and Li, 2001). The lack of robustness with respect to outlying leverage points in the regression framework is the main weakness of these estimators. Exceptions are the high BDP penalized MCD estimator (Croux and Haesbroeck, 2010) and the penalized LTS estimator (Alfons et al., 2012) which are defined over subsamples. These estimators can also be derived from the penalized GTE by substituting $f_i(.)$ with the Mahalanobis distances and squared regression residuals, respectively.

Definition 3. The PMTLE is defined as a particular case of the penalized GTE when the function $f_i(\theta)$ in (8.9) is replaced by the negative log-likelihood of the ith observation.

The PMTLE can attain the highest BDP provided the set F_λ of penalized negative log-likelihoods is d–full. As the set F_λ inherits the index of fullness of F, it is sufficient to derive the index of fullness of the set F comprised by the negative log-likelihoods.

We remind that in the classical settings, when $p < n$, the d–fullness indices of various sets of functions have been characterized. For instance, Vandev and Neykov (1993) determined the index of fullness $d = p$ for the set of p-variate normal distributions. Müller and Neykov (2003) related the index of fullness of the negative log-likelihoods sets of the

linear logistic, Poisson and r-th power exponential distribution regression models with the quantity $\mathcal{N}(X)+1$, where $\mathcal{N}(X)=\max_{0\neq\theta\in\mathbb{R}^p}\text{card}\{i\in\{1,\ldots,n\};\ x_i^T\theta=0\}$ provides the maximum number of covariates $x_i\in R^p$ lying in a subspace, Müller (1995). If the observations x_i^T are linearly independent then $\mathcal{N}(X)=p-1$, and this is the minimal value for for $\mathcal{N}(X)$. If the covariates are qualitative variables such as factors with several levels, then $\mathcal{N}(X)$ is much larger. Neykov et al. (2012a) derived the index of fullness $d=\max(\mathcal{N}(X),\mathcal{N}(Z))+1$ of the set of extended quasi-log-likelihoods where X and Z are the mean and dispersion models covariates data matrices. Neykov et al. (2012b) characterized the index of fullness $d=\mathcal{N}(X)+1$ of the quantile linear regression residuals set. Hence the indices of fullness of the corresponding F_λ sets with convex penalty functions are available for direct use. As consequence of this, the BDP of the PMTLE for the above probabilistic models equals $\frac{1}{n}\min\{n-k,k-\mathcal{N}(X)-1\}$. If the parameter of trimming k satisfies the inequalities $\lfloor\{n+\mathcal{N}(X)+1\}/2\rfloor\leq k\leq\lfloor\{n+\mathcal{N}(X)+2\}/2\rfloor$ the BDP is maximized and equals $\frac{1}{n}\lfloor\{n-\mathcal{N}(X)-1\}/2\rfloor$. Obviously, the BDP of the PMTLE can be small in modeling experimental data with qualitative (categorical) covariates. Thus the PMTLE is more suitable for data with continuous covariates.

Now the question is how to proceed with the characterization of the BDP in high-dimensional data when $p>>n$. As Bühlmann and van der Geer (2011) pointed out: "The philosophy that will generally rescue us, is to 'believe' that in fact only a few coordinates of the θ are non-zero". Armed with this 'belief' we postpone the BDP discussion of the PMTLE to the next section.

In order to reduce the outlier's influence on the selection of the penalization parameter λ we recommend the usage of the penalized BIC based on trimming, defined by $\text{PTBIC}(\lambda)=-2\log(S^P_{n,k}(\hat{\theta}))+df(\lambda)\log(k)$ where $S^P_{n,k}(\hat{\theta})$ is the PMTLE and $df(\lambda)$ are the model degrees of freedom given by the non-zero estimated components of $\hat{\theta}$. Obviously, PTBIC reduces to BIC if $k=n$ and $\lambda=0$.

In the next section, the applicability of the PMTLE is investigated, and its BDP properties for the ultrahigh dimensional multiple linear regression and Poisson regression model are considered.

8.3 Robust SIS and ISIS based on trimming

The usage of penalization is limited in ultrahigh dimensional settings. According to Fan and Lv (2008), the ultrahigh dimensionality concerns with variable selection in the cases when p is much larger than n, i.e., $\log(p) = O(n^{\alpha})$ for some $0 < \alpha < 1$. In this section we focus on the so called sure independence screening (SIS) technique and its variations for variable selection developed by Fan et al. (2009). SIS is a preprocessing technique which aims at a drastic reduction of the number of covariates to a dimension less than the sample size by conventional marginal utility methods, with the hope to catch the most informative covariates, and then to use a penalization technique to select the carriers, see Fan and Lv (2010). Such a two-stage procedure is acceptable because the penalty based variable selection techniques work reasonably well with a moderate number of covariates. Fan and Lv (2008), Fan et al. (2009), and Fan and Lv (2010) have provided theoretical results that all important covariates can be selected by such a procedure with high probability. Unfortunately, the SIS techniques that rely on MLE, quasi-likelihood and robust M-estimators of Huber (1981), are not resistant against outliers in the covariates, and so their applicability is of limited use. This can be overcome by replacing these estimators by their high BDP counterparts based on trimming. The usage of the PMTLE for the classical multiple linear and Poisson regression models will be demonstrated in the following. In order to aid the presentation, we briefly review the SIS formulation, following closely Fan et al. (2009).

8.3.1 Variable ranking by marginal utility

Without loss of generality, we shall assume that $L(.)$ is the negative log-likelihood, although other loss functions such as the quasi-likelihood, the least squares can be used. Let us define the marginal utility of the jth covariate X_j, for $j = 1, \ldots, p$, by

$$L_0 = \min_{\theta_0} \frac{1}{n} \sum_{i=1}^{n} L(y_i, \theta_0), \tag{8.12}$$

$$L_j = \min_{\theta_0, \theta_j} \frac{1}{n} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{ij}\theta_j), \tag{8.13}$$

where L_j is the loss function of using $\theta_0 + x_{ij}\theta_j$ to predict y_i.

The idea behind SIS is to compute the marginal utilities $L_1, \ldots, L_p$, rank them in ascending order, $L_{\nu(1)} \leq \ldots \leq L_{\nu(q)} \leq \ldots \leq L_{\nu(p)}$, where $(\nu(1), \ldots, \nu(p))$ is the permutation of the indices $(1, \ldots, p)$, and select the q-vector of covariates $(X_{\nu(1)}, X_{\nu(2)}, \ldots, X_{\nu(q)})$ for further consideration. In this way the covariate X_j is selected by SIS according to the magnitude of its marginal utility. Computing the L_j is fast as the fitting model has two parameters only, and so even for ultrahigh dimensional data this is not an intensive computational problem. Fan and Lv (2008) recommend to take $q = \lfloor n/\log n \rfloor$ for multiple regression and $q = \lfloor n/(2\log n) \rfloor$ for Poisson regression. The parameter q is usually chosen large enough but $q < n$ to ensure the sure screening property. As q is specified in advance, only the q smallest marginal utilities have to be ordered, and an ordering of the remaining values is not required, hereby saving computation time. We note that the influence of θ_0 can be excluded by the marginal utility via the marginal likelihood ratio $LR_j = L_0 - L_j$ that assesses the increments of the log-likelihood and equals the deviance differences for GLMs. Obviously this will not change the ordering of L_j. For the multiple regression model this is equivalent to centering the dependent variable by its mean. On the other hand, the covariates have to be standardized to reduce the influence of their magnitude.

8.3.2 Penalized pseudo-likelihood

The subset of variables selected by SIS may still include many unimportant covariates. To improve performance, Fan et al. (2009), and Fan and Song (2010) recommend the usage of the penalized likelihood to further delete unimportant variables. By reordering the covariates if necessary, we may assume without loss of generality that $X_1, \ldots, X_q$ are the covariates recruited by SIS. Let $x_{i,q} = (x_{i1}, \ldots, x_{iq})^T$ and redefine $\theta = (\theta_1, \ldots, \theta_q)^T$. Minimization of the penalized log-likelihood

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,q}^T\theta) + \sum_{j=1}^{q} p_\lambda(|\theta_j|), \tag{8.14}$$

will yield a sparse regression parameter estimate θ, where the regularization parameter may be chosen by cross-validation. Let us denote the nonzero components of θ by $\widehat{\mathcal{M}}$.

Fan et al. (2009) refer to this two-stage procedure as SIS-Lasso or SIS-SCAD, depending on the choice of the penalty function. The screening stage solves only bivariate optimization, see (8.13), whereas the fitting part solves only the optimization problem (8.14) with moderate size q. This is an attractive feature in ultrahigh dimensional statistical learning.

8.3.3 Robust SIS-SCAD based on trimming

The two-stage SIS-SCAD estimation procedure are based on the MLE and penalized MLE which are not robust against outlying observations in the covariates in probabilistic regression models. A naive approach would be to replace the optimization problems (8.12), (8.13) and (8.14) by their counterparts based on trimming and to solve them separately to get the corresponding extremes keeping the trimming parameter k at the lowest possible levels to guarantee maximal BDP. This means that the GTE algorithm needs to be used in $p+2$ separate optimization problems.

However, the GTE combinatorial optimization principle dictates that the two-stage SIS-SCAD estimation procedure has to be applied to all k-subsets in order to get that k-subset with the optimal value of the objective function (8.14). In this way we formally define the two-stage Trimmed SIS-SCAD (TSIS-SCAD) estimation procedure as follows:

$$\min_{I\in I_k}\begin{cases} \text{SIS procedure} \\ \begin{cases} L_0^{trim} := \min\limits_{\theta_0} \frac{1}{k}\sum\limits_{i\in I} L(y_i,\theta_0) \\ L_j^{trim} := \min\limits_{\theta_0,\theta_j} \frac{1}{k}\sum\limits_{i\in I} L(y_i,\theta_0+x_{ij}\theta_j) \end{cases} \\ (X_1,\ldots,X_q) := (X_{\nu(1)},X_{\nu(2)},\ldots,X_{\nu(q)}) \\ \text{SIS}-\text{SCAD procedure} \\ S_{k,n}^{P,trim} := \min\limits_{\theta_0,\theta}\left(\frac{1}{k}\sum\limits_{i\in I} L(y_i,\theta_0+x_{i,q}^T\theta)+\sum_{j=1}^q p_\lambda(|\theta_j|)\right). \end{cases} \tag{8.15}$$

Therefore, for all k-subsets the linked optimization problems (8.15) have to be solved subsequently and the penalized TSIS-SCAD estimate is defined by that k-subset with the

minimal value of $S_{k,n}^{P,trim}$. Because this is not feasible for large data sets, an approximate estimate can be obtained by the use of the GTE algorithm. Obviously, the covariates have to be standardized using the means and standard deviations computed from each subset in order to reduce the influence of their magnitude as the GTE algorithm consists of optimization problems over data subsets.

An important choice for the algorithm is the trimming parameter k which controls the identifiability of the model parameters and determines the finite sample BDP of the estimator. Since here we will only consider simple linear and Poisson regression models, the d-fullness index of $F_j = \{L(y_i, \theta_0 + x_{ij}\theta_j) \text{ for } i = 1, \ldots, n\}$ is $\mathcal{N}(X_j) + 1$ for $j = 1, \ldots, p$, according to Müller and Neykov (2003). Thus, the finite sample BDP of the TLE utility estimator equals
$\frac{1}{n} \min\{n - k, k - \mathcal{N}(X_j) - 1\}$, whereas for the penalized TSIS-SCAD estimator the finite sample BDP is $\frac{1}{n} \min\{n - k, k - \mathcal{N}(X_{n\times q}) - 1\}$, see Müller and Neykov (2003). Therefore, the finite sample BDP of the two-stage TSIS-SCAD estimation procedure equals $\frac{1}{n} \min\{n - k, k - D - 1\}$ where $D = \max[\max_j \mathcal{N}(X_j), \mathcal{N}(X_{n\times q})]$. This BDP is maximized for $\lfloor \{n + D + 1\}/2 \rfloor \leq k \leq \lfloor \{n + D + 2\}/2 \rfloor$ and equals $\frac{1}{n} \lfloor \{n - D - 1\}/2 \rfloor$.

Instead of assigning a minimal value of the trimming parameter k to gain maximal BDP we prefer to take a subset of data of size $k = \lfloor \alpha n \rfloor$ for $\alpha \in (0.5, 1]$, provided all covariates are continuous. For instance, the choice $\alpha = 0.80$ ensures simultaneously a resistance against 20% outliers in the data and leads to a higher efficiency of the estimator.

8.3.4 Iterative feature selection

Independent variable screening as it is done in the SIS procedure may have poor performance if variables are marginally weakly correlated with the response variable but jointly related with the response, or if a variable is jointly uncorrelated with the response but its marginal correlation with the response is higher than for some other important variable. These problems are addressed by iterative SIS (ISIS) proposed by Fan and Lv (2008), Fan et al. (2009), and Fan and Song (2010) which incorporates the joint covariance information.

In the first step of ISIS, the two stage SIS-SCAD procedure is performed to select the subset $\widehat{\mathcal{M}}_1$ of covariates. Then Fan et al. (2009) propose to compute the following loss function in order to assess the importance of the covariate X_j which has not been included by the SIS-SCAD procedure:

$$L_j^{(2)} = \min_{\theta_0, \theta_{\widehat{\mathcal{M}}_1}, \theta_j} n^{-1} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1} + x_{ij}\theta_j), \tag{8.16}$$

for $j \in \widehat{\mathcal{M}}_1^c = \{1, \ldots, p\} \setminus \widehat{\mathcal{M}}_1$, where $x_{i,\widehat{\mathcal{M}}_1}$ is the sub-vector of x_i consisting of those elements in $\widehat{\mathcal{M}}_1$.

The optimization problem (8.16) is low-dimensional and thus easy to solve. The additional contribution of variable X_j given the existence of variables in $\widehat{\mathcal{M}}_1$ can be assessed by the marginal likelihood ratio test (difference by the two deviance functions for the GLM setting):

$$L_j^{LR} = \min_{\theta_0, \theta_{\widehat{\mathcal{M}}_1}} n^{-1} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1}) - L_j^{(2)}. \tag{8.17}$$

After ordering of L_j^{LR} in ascending order for $j \in \widehat{\mathcal{M}}_1^c$ we take the indices corresponding to the smallest m_2 elements and form the set $\widehat{\mathcal{A}}_2$.

The above pre-screening step is followed by the penalized likelihood for obtaining a sparse estimate

$$\theta_2 = \underset{\theta_0, \theta_{\widehat{\mathcal{M}}_1}, \theta_{\widehat{\mathcal{A}}_2}}{\arg\min} \left(n^{-1} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1} + x_{i,\widehat{\mathcal{A}}_2}^T \theta_{\widehat{\mathcal{A}}_2}) + \sum_{j \in \widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2} p_\lambda(|\theta_j|) \right). \tag{8.18}$$

As a result we obtain a new estimated set $\widehat{\mathcal{M}}_2$ of active indices consisting of those indices of θ_2 that are non-zero. Thus, this procedure allows to delete variables from the previously selected features with indices in $\widehat{\mathcal{M}}_1$. The process, which iteratively recruits and deletes features, can then be repeated until we obtain a set of indices $\widehat{\mathcal{M}}_l$ which either has reached the prescribed size q, or satisfies $\widehat{\mathcal{M}}_l = \widehat{\mathcal{M}}_{l-1}$. In this way a final estimated parameter vector θ_l is obtained.

In their R package *SIS*, Fan et al. (2009) chose $k_1 = \lfloor 2q/3 \rfloor$, and thereafter at the rth iteration, they take $m_r = q - |\widehat{\mathcal{M}}_{r-1}|$. This ensures that the iterated versions of SIS take at least two iterations to terminate.

8.3.5 Robust iterated variable selection based on trimming

Similar to robustifying the two-stage SIS-SCAD estimation procedure, we could replace the optimization problems (8.16) and (8.18) by their counterparts based on trimming and solve them for all k-subsets out of n cases in order to get that k-subset with the optimal objective value of (8.18). This way we formally define the two-stage Trimmed ISIS-SCAD (TISIS-SCAD) estimation procedure as

$$\min_{I\in I_k}\begin{cases} \text{ISIS procedure} \\ \begin{cases} L^{trim}_{0,\widehat{\mathcal{M}}_1} = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1}} k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0+x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1}) \\ L^{(2,trim)}_j = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1},\theta_j} k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0+x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1}+x_{ij}\theta_j) \end{cases} \\ \text{ISIS}-\text{SCAD procedure} \\ \tilde{S}^{P,trim}_{k,n} = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1},\theta_{\widehat{\mathcal{A}}_2}} \left(k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0+x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1}+x^T_{i,\widehat{\mathcal{A}}_2}\theta_{\widehat{\mathcal{A}}_2}) \right. \\ \qquad\qquad \left. + \sum\limits_{j\in\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2} p_\lambda(|\theta_j|)\right) \end{cases} \tag{8.19}$$

Therefore for all k-subsets the linked optimization problems (8.19) have to be solved subsequently and the penalized TISIS-SCAD estimate is defined by the k-subset with the minimal value of $\tilde{S}^{P,trim}_{k,n}$. This procedure would not be computationally feasible for larger data sets, and therefore an approximate estimate can be obtained by the use of the GTE algorithm. Again the variable standardization has to be done within the subsets.

Let $r = |\widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2|$ be the cardinality of $\widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2$ and $\widehat{\mathcal{M}}^*_1 = \widehat{\mathcal{M}}_1 + 1$. Similar to the previous section we can conclude that the corresponding utility sets are $(\mathcal{N}(X_{n\times r})+1)$ and $(\mathcal{N}(X_{n\times\widehat{\mathcal{M}}^*_1})+1)$ full, and these are the minimal numbers of observations that guarantee identifiability of θ (Müller and Neykov, 2003). Hence, the finite sample BDP of the TLE utility estimator defined by (8.19) equals $\frac{1}{n}\min\left\{n-k, k-\mathcal{N}(X_{n\times\widehat{\mathcal{M}}^*_1})-1\right\}$ whereas for the penalized maximum trimmed ISIS-SCAD estimator it is $\frac{1}{n}\min\{n-k, k-\mathcal{N}(X_{n\times r})-1\}$. Using the notation $\tilde{D} = \max[\mathcal{N}(X_{n\times\widehat{\mathcal{M}}^*_1}), \mathcal{N}(X_{n\times r})]$, the BDP of the two-stage TISIS-SCAD estimation procedure (8.19) equals $\frac{1}{n}\min\left\{n-k, k-\tilde{D}-1\right\}$. This BDP is maximized for $\left\lfloor\left\{n+\tilde{D}+1\right\}/2\right\rfloor \le k \le \left\lfloor\left\{n+\tilde{D}+2\right\}/2\right\rfloor$ and equals

$\frac{1}{n}\left\lfloor\left\{n-\tilde{D}-1\right\}/2\right\rfloor$.

As mentioned above, one can select $k = \lfloor \alpha n \rfloor$ with $\alpha = 0.80$, for instance.

8.4 Simulation study

In this section, we study the performance of SIS-SCAD, ISIS-SCAD and their trimmed counterparts on simulated data for the multiple and Poisson linear regression framework. Two different data configurations are presented and discussed.

8.4.1 Performance measures

According to the simulation designs described in the next sections we generate training data without and with contamination, and estimate the regression parameters θ with the different methods. In addition, n test set observations are generated according to the same scheme but without outliers. We denote the test set covariates by $\tilde{x}_i$ and the response by $\tilde{y}_i$, for $i = 1, \ldots, n$. The predictions $\tilde{\eta}_i = \tilde{x}_i^T\hat{\theta}$ for the linear regression model, and $\log(\tilde{\eta}_i) = \tilde{x}_i^T\hat{\theta}$ for Poisson regression are evaluated by the root mean squared error of prediction (RMSEP),

$$\text{RMSEP}(\hat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \tilde{\eta}_i)^2}.$$

The RMSEP is computed for each estimator and simulated test data set, and we report averages and medians over all simulations. Further, we compare also with the so called oracle estimator, where the true regression coefficients θ are used for the evaluation.

We evaluate the methods also according to their ability to select the correct variables, using the false positive rate (FPR) and the false negative rate (FNR). False positives refer to variables that are selected by the method, while their coefficients in the simulation design are zero. In contrast, a false negative is a coefficient estimated as zero, while it was

generated as non-zero. Formally, FPR and FNR can be defined as

$$FPR(\hat{\theta}) = \frac{|\{j \in \{1,\ldots,p\} : \hat{\theta}_j \neq 0 \wedge \theta_j = 0\}|}{\{j \in \{1,\ldots,p\} : \theta_j = 0\}} \tag{8.20}$$

$$FNR(\hat{\theta}) = \frac{|\{j \in \{1,\ldots,p\} : \hat{\theta}_j = 0 \wedge \theta_j \neq 0\}|}{\{j \in \{1,\ldots,p\} : \theta_j \neq 0\}} \tag{8.21}$$

These rates are computed for each simulated data set, and we will report average numbers over all simulations. The better the sparseness structure is identified by the method, the smaller these rates should be.

In order to compare the simulation results with those of Fan et al. (2009) for the Poisson regression model, we also report the median values of the evaluation measures $||\theta-\hat{\theta}||_1 = \sum_{i=0}^{p} |\theta_j-\hat{\theta}_j|$ and $||\theta-\hat{\theta}||_2 = (\sum_{i=0}^{p}(\theta_j-\hat{\theta}_j)^2)^{1/2}$, the AIC - Akaike's information criterion, and the BIC - Bayesian information criterion.

8.4.2 Simulation design - multiple linear regression

We use the 3^{rd} simulation design considered in Alfons et al. (2012) where the sparse LTS regression estimator with L_1 penalty (L1-penalized trimmed LTS, trimmed LASSO) was introduced. We compare their estimator with the SIS-SCAD and its trimmed version TSIS-SCAD, because SIS-SCAD exhibits better performance than SIS-LASSO according to the simulation study (without contamination) of Fan et. al. (2009). We note that Fan et al. (2009) denote SIS-SCAD and ISIS-SCAD as Van-SIS and Van-ISIS.

In this setting, we generate $n = 100$ observations from a p-dimensional normal distribution $N_p(0, \Sigma)$, with $p = 1000$. The covariance matrix $\Sigma = (\Sigma_{ij})_{1\leq i,j\leq p}$ is given by $\Sigma_{ij} = 0.5^{|i-j|}$, creating correlated predictor variables. The coefficient vector $\theta = (\theta_1,\ldots,\theta_p)^T$ has components $\theta_1 = \theta_7 = 1.5$, $\theta_2 = 0.5$, $\theta_4 = \theta_{11} = 1$, and $\theta_j = 0$ for $j \in \{1,\ldots,p\}\backslash\{1,2,4,7,11\}$.

The response variable is generated according to the multiple linear regression model $y_i = x_i^T\theta + \varepsilon_i$, where the error terms ε_i follow a normal distribution with $\mu = 0$ and $\sigma = 0.5$. We apply the same contamination scheme as Alfons et al. (2012), see also Khan et al. (2007), who proposed:

1. No contamination

2. Vertical outliers: 10% of the errors terms in the regression model follow a normal $N(20, \sigma^2)$, instead of a $N(0, \sigma^2)$.

3. Leverage points: Same as in 2., but the 10% contaminated observations contain high-leverage values, by drawing the predictor variables from independent $N(50, 1)$ distributions.

The results of the simulation experiment are given in Table 8.1. The first and second row of this table are taken from Table 3 of Alfons et al. (2012) in order to make a comparison. *L1-LTSraw* is the result of the L1-penalized trimmed LTS procedure, and *L1-LTS* is a reweighted version of the estimator (see Alfons et al., 2012). The means (*mean*) and medians (*med*), respectively, of the RMSEP, FPR and FNR over 500 simulation runs are reported for every method; ISIS-SCAD is denoted by *ISIS*, and its trimmed version by *TISIS-XX*, where *XX* shows the percentage of trimming - 10, 20, 25.

The results based on the means and medians are almost the same in our simulation experiments. Larger differences could refer to possible problems with the algorithm. We see that the performance of the ISIS-SCAD estimator is excellent for the scenario without contamination, and the RMSEP is close to the oracle estimator. However, ISIS-SCAD breaks down in the presence of vertical outliers or leverage points, whereas the robust methods L1-LTS and TISIS are stable. TISIS shows excellent performance: the RMSEP is close to the oracle estimator, and the false positive and false negative rates are very small. Moreover, the different trimming percentages result in about the same performance.

8.4.3 Simulation design - Poisson regression

The simulation configurations of this section are the same as in Fan et al. (2009). The following three settings of covariates $X_1, \ldots, X_p$ and regression coefficients $\theta_0, \theta_1, \ldots, \theta_p$, for $p = 1000$ and sample size $n = 200$ are generated:

1. $X_1, \ldots, X_p$ are independent and identically distributed $N(0,1)$ random variables; $\theta_0 = 5$, $\theta_1 = -0.5423$, $\theta_2 = -0.5314$, $\theta_3 = -0.5012$, $\theta_4 = -0.4850$, $\theta_5 = -0.4133$, $\theta_6 = -0.5234$, and $\theta_j = 0$ for $j > 6$;

2. $X_1, \ldots, X_p$ are jointly Gaussian, marginally $N(0,1)$, and with $cor(X_i, X_4) = 1/\sqrt{2}$ for all $i \neq 4$ and $cor(X_i, X_j) = 1/2$ if i and j are distinct elements of $\{1, \ldots, p\}\backslash\{4\}$; $\theta_0 = 5$, $\theta_1 = \theta_2 = \theta_2 = 0.6$, $\theta_4 = -0.9\sqrt{2}$; and $\theta_j = 0$ for $j > 4$;

3. $X_1, \ldots, X_p$ are jointly Gaussian, marginally $N(0,1)$, and with $cor(X_i, X_5) = 0$ for all $i \neq 5$, $corr(X_i, X_4) = 1/\sqrt{2}$ for all $i \notin \{4, 5\}$, and $cor(X_i, X_j) = 1/2$ if i and j are distinct elements of $\{1, \ldots, p\}\backslash\{4, 5\}$; $\theta_0 = 5$, $\theta_1 = \theta_2 = \theta_2 = 0.6$, $\theta_4 = -0.9\sqrt{2}$, $\theta_5 = 0.15$, and $\theta_j = 0$ for $j > 5$.

The first case with independent predictors is the simplest situation for variable selection. Here, the coefficients $\theta_1, \ldots, \theta_6$ were generated as $\left(\frac{\log n}{\sqrt{n}} + |Z|/8\right) U$ with $Z \sim N(0,1)$ and $U = 1$ with probability 0.5 and $U = -1$ with probability 0.5, independently of Z. The last two cases are more complicated because of serial correlations. Even more, although $\theta_4 \neq 0$, the choices of the other regression coefficients in Cases 2 and 3 ensure that $\text{cor}(X_4, Y) = 0$, which makes variable selection more difficult. The coefficient $\theta_0 = 5$ is used to control an appropriate signal-to-noise ratio.

The data (x_i^T, y_i) for $i = 1, \ldots, 200$ are independent copies of a pair where y_i is conditionally on x_i distributed as Poisson$(\mu(x))$, where $\log(\mu(x)) = \theta_0 + x_i^T\theta$.

We apply the following contamination scheme:

1. No contamination

2. Vertical outliers: 10% and 20% data contamination is introduced by changing respectively the first 20 and 40 observations to $y_i := y_i + \exp(7)$, for $i = 1, \ldots, 20$, respectively 40.

3. Leverage points: 10% and 20% data contamination is introduced by modifying respectively the first 20 and 40 rows of the covariates matrix according to $x_{ij} :=$

$$-3B_j \text{sign}(x_{ij}) \text{ for } i = 1, \ldots, 20, \text{ where } B_j = \max_{1 \leq i \leq n}(|x_{ij}|) \text{ for } j = 1, \ldots, p.$$

Following the suggestion of Fan et al. (2009), we perform the computation for ISIS-SCAD and TISIS-SCAD with $q = \left\lfloor \frac{n}{2 \log n} \right\rfloor = 18$ as a sensible choice based on asymptotic results. The final regularization parameter for the SCAD penalty was chosen via 10-fold cross-validation as recommended by Fan et al. (2009). However, the BIC is used to choose the SCAD regularization parameter at each intermediate stage of the ISIS procedures in the three cases.

The estimators were applied to the training data and evaluated on the test data with $n = 200$ observations, which were generated according to the same schemes without contamination. For the TISIS-SCAD procedure we report the result for different trimming percentages. In the tables below, we report several performance measures, all of which are based on 100 Monte Carlo repetitions. The tables contain the medians of these measures. The first two rows give the estimation errors $||\theta - \hat{\theta}||_1$ and $||\theta - \hat{\theta}||_2$, respectively, evaluated for the training data. In the 3^{rd} and 4^{th} row we report the FPR and FNR, respectively, for the training data. The fifth, sixth, seventh and eighth rows give Akaike's information criterion (Akaike, 1974), AIC, and the Bayesian information criterion (Schwartz, 1978), BIC, computed over the training and test (indicated by the additional "t") data. The last two rows give the RMSEP for the test data (RMSEP.t) and the true regression parameter (RMSEP.o). The symbols "*" in the tables refer to very big values greater than 250000. Two consecutive tables are used for one simulation setting, where the first table contains the results for the vertical outliers, and the second table is for the leverage points.

For the simulation experiments without contamination, our results for ISIS-SCAD closely follow those based on Van-ISIS presented at Tables 5-7 of Fan et al. (2009). In case of contamination (vertical outliers or leverage points) we see that the ISIS-SCAD estimator fails; all error measures are (much) worse, independent of the simulation scheme. An exception is the FPR, which means that in case of contamination the correct zero-coefficients are indeed set to zero. However, since FNR increases considerably, many non-zero coefficients are also set to zero. The robust version TISIS-SCAD shows excellent behavior for

all simulation schemes, and for uncontaminated and contaminated data. Generally, the results are close to the ISIS-SCAD estimator when no contamination is present. Remarkable are the results for FPR and FNR of TISIS-SCAD, which are not higher than 1% in all scenarios.

Table 8.1: Results for the simulation scheme in the multiple linear regression case, where $n = 100$ and $p = 1000$. The means and medians of RMSEP, FPR and FNR over 500 simulation runs are reported for every method: L1-LTSraw and L1-LTS refer to the raw and weighted penalized LTS regression estimator of Alfons et al. (2012), respectively, ISIS and TISIS (with the percentage of trimming) corresponds to the original and trimmed version of ISIS-SCAD, respectively, and Oracle uses the true regression parameters.

	No contamination			Vertical outliers			Leverage points		
Method	RMSEP	FPR	FNR	RMSEP	FPR	FNR	RMSEP	FPR	FNR
L1-LTSraw	0.79	0.02	0.00	0.74	0.02	0.00	0.72	0.02	0.00
L1-LTS	0.74	0.01	0.00	0.70	0.01	0.00	0.70	0.02	0.00
ISIS(mean)	0.53	0.00	0.00	4.89	0.01	0.75	2.17	0.01	0.33
ISIS(med)	0.52	0.00	0.00	4.88	0.01	0.79	2.13	0.01	0.40
TISIS-10(mean)	0.53	0.00	0.00	0.55	0.00	0.00	0.55	0.00	0.00
TISIS-20(mean)	0.53	0.00	0.00	0.56	0.00	0.01	0.57	0.00	0.03
TISIS-25(mean)	0.53	0.00	0.00	0.59	0.00	0.02	0.58	0.00	0.04
TISIS-10(med)	0.52	0.00	0.00	0.53	0.00	0.00	0.53	0.00	0.00
TISIS-20(med)	0.52	0.00	0.00	0.54	0.00	0.00	0.54	0.00	0.00
TISIS-25(med)	0.52	0.00	0.00	0.55	0.00	0.00	0.56	0.00	0.00
Oracle	0.50								

Table 8.2: Poisson regression, Case 1 of the simulation scheme with 0%, 10% and 20% of contamination by vertical outliers (VO), $n = 200$ and $p = 1000$.

	0% cont.	VO-10% contamination			VO-20% contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.12	3.58	0.13	0.17	4.83	0.14	0.18
$\|\|\theta - \hat{\theta}\|\|_2$	0.03	0.99	0.03	0.05	1.36	0.04	0.05
FPR	0.01	0.01	0.01	0.01	0.01	0.01	0.01
FNR	0	0	0	0	0.17	0	0
AIC	1544.82	*	1393.09	1175.19	*	1232.82	1022.87
AICt	1666.58	26502.49	1675.22	1749.27	*	1685.46	1786.56
BIC	1607.49	*	1453.76	1233.61	*	1291.25	1078.26
BICt	1729.24	26563.51	1737.89	1811.93	*	1748.13	1849.23
RMSPE.t	24.84	385.37	26.22	32.28	493.55	28.4	36.74
RMSEP.o	17.38						

Table 8.3: Poisson regression, Case 1 of the simulation scheme with 0%, 10% and 20% of contamination by leverage points (LP), $n = 200$ and $p = 1000$.

	0% cont.	LP-10% contamination			LP-20% contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.12	3.84	0.14	0.17	3.92	0.15	0.21
$\|\|\theta - \hat{\theta}\|\|_2$	0.03	1.41	0.04	0.05	1.42	0.04	0.06
FPR	0.01	0.01	0.01	0.01	0.01	0.01	0.01
FNR	0	0.83	0	0	1	0	0
AIC	1544.82	*	1381.66	1179.74	*	1228.92	1033.77
AICt	1666.58	*	1691.73	1727.89	*	1693.31	1770
BIC	1607.49	*	1442.22	1237.93	*	1286.54	1089.67
BICt	1729.24	*	1754.4	1790.56	*	1754.99	1832.67
RMSPE.t	24.84	493.32	27.61	31.63	511.07	29.26	43.89
RMSEP.o	17.38						

Table 8.4: Poisson regression, Case 2 of the simulation scheme with 0%, 10% and 20% of contamination by vertical outliers (VO), $n = 200$ and $p = 1000$.

	0% cont.	VO-10% contamination			VO-20% contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.26	5.54	0.28	0.32	6.54	0.29	0.33
$\|\|\theta - \hat{\theta}\|\|_2$	0.07	1.66	0.08	0.09	1.88	0.08	0.1
FPR	0.01	0.02	0.01	0.01	0.02	0.01	0.01
FNR	0	0.25	0	0	0.5	0	0
AIC	1535.93	*	1381.11	1174.58	*	1226.36	1024.64
AICt	1674.54	26274.8	1683.06	1703.42	38396.37	1686.06	1732.16
BIC	1598.6	*	1441.77	1233	*	1284.79	1080.53
BICt	1737.21	26334.17	1745.73	1766.09	38459.04	1748.73	1792.69
RMSPE.t	17.52	212.38	17.59	18.9	291.46	18.29	19.83
RMSPE.o	13.79						

Table 8.5: Poisson regression, Case 2 of the simulation scheme with 0%, 10% and 20% of contamination by leverage points (LP), $n = 200$ and $p = 1000$.

	0% cont.	LP-10% contamination			LP-20% contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.26	3.41	0.28	0.32	3.46	0.3	0.32
$\|\|\theta - \hat{\theta}\|\|_2$	0.07	1.66	0.08	0.09	1.66	0.09	0.09
FPR	0.01	0.01	0.01	0.01	0.01	0.01	0.01
FNR	0	0.75	0	0	0.75	0	0
AIC	1535.93	17359.63	1379.97	1174.5	16466.75	1226.01	1027.9
AICt	1674.54	18796.81	1680.45	1704.82	19419.59	1697.08	1716.28
BIC	1598.6	17389.32	1440.63	1232.54	16521.18	1284.44	1083.79
BICt	1737.21	18834.75	1743.11	1767.49	19468.78	1759.75	1778.95
RMSPE.t	17.52	157.28	17.66	18.41	159.36	18.22	19.34
RMSEP.o	13.79						

Table 8.6: Poisson regression, Case 3 of the simulation scheme with 0%, 10% and 20% of contamination by vertical outliers (VO), $n = 200$ and $p = 1000$.

	0% cont.	VO-10% contamination			VO-20% contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.26	5.63	0.27	0.31	6.6	0.29	0.33
$\|\|\theta - \hat{\theta}\|\|_2$	0.07	1.65	0.08	0.09	1.9	0.09	0.1
FPR	0.01	0.02	0.01	0.01	0.02	0.01	0.01
FNR	0	0.4	0	0	0.6	0	0
AIC	1539.62	*	1384.54	1177.33	*	1231.26	1031.1
AICt	1674.91	26836.25	1683.5	1705.71	47284.46	1689.55	1729.97
BIC	1602.29	*	1445.2	1235.75	*	1289.36	1086.99
BICt	1737.57	26898.92	1746.17	1768.38	47345.48	1752.22	1792.64
RMSPE.t	17.58	214.84	17.95	18.94	288.5	18.47	19.8
RMSEP.o	13.91						

Table 8.7: Poisson regression, Case 3 of the simulation scheme with 0%, 10% and 20% of contamination by leverage points (LP), $n = 200$ and $p = 1000$.

	0% cont.	LP-10% contamination			LP-20% contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.26	3.58	0.27	0.31	3.65	0.31	0.31
$\|\|\theta - \hat{\theta}\|\|_2$	0.07	1.67	0.07	0.09	1.67	0.09	0.09
FPR	0.01	0.01	0.01	0.01	0.01	0.01	0.01
FNR	0	0.8	0	0	1	0	0
AIC	1539.62	17816.74	1385.37	1181.93	16620.12	1229.87	1031
AICt	1674.91	19809.14	1677.79	1704.42	20470.47	1697.85	1724.25
BIC	1602.29	17849.23	1446.04	1240.27	16672.89	1288.3	1086.89
BICt	1737.57	19835.52	1740.46	1767.09	20526.54	1760.52	1786.92
RMSPE.t	17.58	164.91	17.86	18.84	166.76	19.06	19.96
RMSEP.o	13.91						

8.5 Summary and conclusions

We introduced a robust version of the penalized MLE based on the idea of trimming and characterized its BDP based on the notion of d-fullness. The finite sample properties of the proposed estimator were studied via an extended simulation study within high-dimensional multiple and Poisson linear regression settings. The new estimator generally performs very well, which is confirmed by the simulation experiments and by a comparison to other proposals. To handle the computations, the SIS/ISIS procedure of Fan et al. (2009) was used. However, any other procedure that implements penalization/regularization techniques can be employed instead. The computation of the estimator is taking advantage of the same technology as used for its classical counterpart, but here the estimation is based on subsamples only. The used algorithm consisting of a trial and a refinement step (Neykov et al., 2012a) follows the ideas of the FAST-LTS algorithm of Rousseeuw and van Driessen (1999), and Neykov and Müller (2003). An important choice for estimators based on trimming is the trimming percentage. In the numerical experiments, it has been shown that a trimming percentage lower than the contamination level can lead to very poor estimates, but any higher trimming percentage gives very reasonable results. Therefore, a general rule is to work with a conservative choice of the trimming percentage or to estimate the amount of trimming similarly to Čížek (2010), and Gervini and Yohai (2002).

Bibliography

Adrover, J., Maronna, R.A. and Yohai, V.J., 2004. Robust regression quantiles. J. Statist. Plann. Infer., **vol. 122**, pp. 187 – 202.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Auto. Control, **vol. 19**, pp. 716 – 723.

Alfons, A., Croux, C. and Gelper, S. 2013. Sparse least trimmed squares regression. Ann. Appl. Stat., **vol. 7**, pp. 226 – 248.

Antoniadis, A., Gijbels, I. and Nikolova, M. 2011. Penalized likelihood regression for generalized linear models with non-quadratic penalties. Ann. Inst. Stat. Math. **vol. 63**, 585 – 615.

Atanasov, D. V. 1998. About the finite sample breakdown point of the WTL estimators, MSc. Thesis, Faculty of Mathematics, Sofia Univ., (in Bulgarian).

Atanasov, D. V. and Neykov, N. M. 2001. On the finite sample breakdown point of the weighted trimmed likelihood estimators and the d–fullness of a set of continous functions. In: *Proceedings of the CDAM Confrence*, 10-14 September 2001, Minsk, Belarus, Aivazian, S., Yu. Kharin and H. Reider (eds.), **vol. 1,** pp. 52 – 57.

Atkinson, A. C. and Riani, M. 2000. *Robust diagnostic regression analysis.* Springer, NY.

Atkinson, A. C. Riani, M. and Cerioli,A. 2004. *Exploring Multivariate Data with the Forward Search.* Springer, NY.

Barão, M. I. and Tawn, J. A. 1999. Extremal analysis of short series with outliers: Sea-levels and athletic records. Appl. Statist. **vol. 48**, 469 – 487.

Bednarski, T. and Clarke, B. R. 1993. Trimmed likelihood estimation of location and scale of the normal distribution. Austral. J. Statist. **vol. 35,** pp. 141 – 153.

Beran, R. 1982. Robust estimation in models for independent nonindentically distributed data. Ann. Statist. **vol. 10**, pp. 415 – 428.

Breheny, P. and Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Stat., **vol. 5**, pp. 232 – 253.

Bühlmann, P. and van der Geer, S. 2011. *Statistics for High Dimensional Data: Methods Theory and Applications*. Springer. New York.

Campbell N. A. 1984. Mixture models and atypical values. Math. Geology **16**, pp. 465 – 477.

Cantoni, E. and Ronchetti, E. 2001. Robust inference for generalized linear models. J. Amer. Statist. Assoc. **vol. 96**, pp. 1022 – 1030.

Carroll, R. J. and Pederson, S. 1993. On robustness in the logistic regression model. J. R. Statist. Soc. B. **vol. 55,** pp. 693 – 706.

Chen, C. (2004). An adaptive algorithm for quantile regression. In: *Theory and Applications of Recent Robust Methods*. Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Birkhäuser, Basel, pp. 39 – 48.

Cheng, T.-C. 2011. Robust diagnostics for the heteroscedastic regression model. Comput. Statist. and Data Anal. **vol. 55**, pp. 1845 – 1866.

Christmann, A. 1994. Least median of weighted squares in logistic regression with large strata. Biometrika. **vol. 81,** pp. 413 – 417.

Christmann, A. and Rousseeuw, P. J. 2001. Measuring overlap in logistic regression. Comput. Statist. and Data Anal.. **vol. 28,** pp. 65 – 75.

Choi, E. P., Hall, P. and Presnell, B. 2000. Rendering parametric procedures more robust by empirically tilting the model. Biometrika **vol. 87**, pp. 453 – 465.

Čížek, P. 2002. *Robust estimation in nonlinear regression and limited dependent variable models.* http://econpapers.hhs.se/paper/wpawuwpem/0203003.htm

Čížek, P. 2004. General trimmed estimation: robust approach to nonlinear and limited dependent variable models. (Discussion Paper No. 130), Tilburg University, Center for Economic Research.

Čížek, P. 2006. Least trimmed squares in nonlinear regression under dependence. J. Statist. Plann. Infer., **vol. 136**, pp. 3967 – 3988.

Čížek, P., 2008. Robust and Efficient Adaptive Estimation of Binary-Choice Regression Models, J. Amer. Statist. Assoc., **vol. 103**, pp. 685 – 698.

Čížek, P. 2008. General trimmed estimation: Robust approach to nonlinear and limited dependent variable models. Econometric Theory, **vol. 24**, pp. 1500 – 1529.

Čížek, P. 2010. Reweighted least trimmed squares: an alternative to one-step estimators. CentER Discussion Paper 2010/91, Tilburg University, The Netherlands.

Čížek, P. 2011. Semiparametrically weighted robust estimation of regression models. Comput. Statist. and Data Anal., **vol. 55**, pp. 774 – 786.

Čížek, P. 2013. Reweighted least trimmed squares: an alternative to one-step estimators. TEST, **vol. 22**, pp. 514 – 533.

Coles, S. G. 2001. An introduction to statistical modeling of extreme values. Springer-Verlag, London.

Copas, J. B. 1988. Binary regression models for contaminated data (with discussion). J. R. Statist. Soc. B. **vol. 50,** pp. 225 – 265.

Cuesta-Albertos,J.A., Matrán, C. and Mayo-Iscar, A. 2008. Robust estimation in the normal mixture model based on robust clustering. J. R. Statist. Soc. B **vol. 70**, pp. 779 – 802.

Davé, R., Krishnapuram, R. 1997. Robust clustering methods: a unified view. IEEE Transactions on Fuzzy Systems **vol. 5**, pp. 270 – 293.

Demidenko, E. Z. 1989 *Optimization and regression.*(in Russian) Nauka, Moscow.

Dimova, R., Neykov, N. M. 2004. Generalized d-fullness technique for breakdown point study of the trimmed likelihood estimator with applications. In: *Theory and Applications of Recent Robust Methods*. Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Birkhäuser, Basel. pp. 83 – 91.

Donoho, D. L. and Huber, P. J. 1983. *The notion of breakdown point.* In: A festschrift for Eric Lehmann. P.J. Bickel, K. A. Doksum and J. L. Hodges (eds.). Belmont, CA: Wadsworth, 157 – 184.

Dupuis, D. J. and Field, C. A. 1998. Robust estimation of extremes. The Canadian J. of Statistics, **vol. 26**, 199 – 215.

Dupuis, D. J. and Morgenthaler, S. 2002. Robust weighted likelihood estimators with an application to bivariate extreme value problems. The Canadian J. of Statistics, **vol. 30**, 19 – 31.

Dupuis, D. J. and Tawn, J. A. 2001. Efects of misspecification in bivariate extreme value problems. Extremes, **vol. 4**, 315 – 330.

Dunn, P. 2009. Tweedie exponential family models.http://cran.R-project.org/doc/ packages/tweedie.pdf

Efron, B. 1986. Double exponential families and their use in generalized linear regression. J. Amer. Statist. Ass. **vol. 81**, pp. 709 – 721.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004. Least angle regression. Ann. Statist., **vol. 32**, pp. 407 – 499.

Fan, J. and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc., **vol. 96**, pp. 1348 – 1360.

Fan, J. and Lv, J. 2008. Sure independence screening for ultrahigh dimensional feature space (with discussion), J. R. Statist. Soc. B, **vol. 70**, pp. 849 – 911.

Fan, J. and Lv, J. 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica, **vol. 20**, pp. 101 – 148.

Fan, J., Samworth, R. and Wu, Y. 2009. Ultrahigh dimensional variable selection: beyond the linear model. J. Mach. Learn. Res., **vol. 10**, pp. 1829 – 1853.

Fan, J. and Song, R. 2010. Sure independence screening in generalized linear models with NP-dimensionality. Ann. Statist., **vol. 38**, pp. 3567 – 3604.

Farcomeni, A. and Greco,L. 2015. *Robust methods for data reduction.* CRC Press, New York.

Field, C. and Smith, B. (1994). Robust estimation - a weighted maximum likelihood approach. Int. Statist. Rev. **vol. 62**, pp. 405 – 424.

Frank, I.E. and Friedman, J.H. 1993. A statistical view of some chemometrics regression tools (with discussion). Technometrics, **vol. 35**, pp. 109 – 148.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. 2007. Pathwise coordinate optimization. Ann Appl Statist., **vol. 1**, pp. 302 – 332.

Friedman, J., Hastie, T. and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. J. Statist. Software, **vol. 33**, pp. 1 – 22. URL http://www.jstatsoft.org/v33/i01/.

Fritz, H., Garc?a-Escudero, L.A., and Mayo-Iscar, A. 2013. Robust constrained fuzzy clustering. Information Sciences, **vol. 245**, pp. 38 – 52.

Fritz, H., Garc?a-Escudero, L.A. and Mayo-Iscar, A. 2013. A fast algorithm for robust constrained clustering. Comput. Statist. and Data Anal. **vol. 61**, pp. 124 – 136.

Gallegos, M. T., Ritter, G. 2005. A robust methods for cluster analysis. Ann. Statist. **vol. 33**, pp. 347 – 380.

Gallegos, M. T. and Ritter, G. 2010. Using combinatorial optimization in model-based trimmed clustering with cardinality constraints. Comput. Statist. and Data Anal. **vol. 54**, pp. 637 – 654.

Garcia-Escudero, L. A., Gordaliza, A., Matran, C. 2003. Trimming tools in exploratory data analysis. Journal of Computational and Graphical Statistics, **vol. 12**, pp. 434 – 449.

Garcia-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. 2008. A general trimming approach to robust cluster analysis. Ann. Statist. **vol. 36**, pp. 1324 – 1345.

Garcia-Escudero, L. A., Gordaliza, A., Martin R. S., and Mayo-Iscar, A. 2010a Robust Clusterwise Linear Regression Through Trimming, submitted to Comput. Statist. and Data Analysis. **vol. 54**, pp. 3057 – 3069.

Garc?a-Escudero, L. A., Gordaliza, A., Matran, C. and Mayo-Iscar, A. 2010b. A review of robust clustering methods. Adv. Data Anal. Classif. **vol 4**, pp. 89 – 109.

Garc?a-Escudero, L. A., Gordaliza, A., Matr?n, C., Mayo-Iscar, A. 2011. Exploring the number of groups in robust model-based clustering. Statistics and Computing, **vol. 2**, pp. 585 – 599.

Garc?a-Escudero, L. A., Gordaliza, A. and Mayo-Iscar, A. 2013. Comments on: model-based clustering and classification with non-normal mixture distributions. Statistical Methods and Applications, **vol. 22**, 459 – 461.

Garc?a-Escudero, L. A., Gordaliza, A., and Mayo-Iscar, A. 2014. A constrained robust proposal for mixture modeling avoiding spurious solutions. Advances in Data Analysis and Classification, **vol. 8**, pp. 27 – 43.

Garcia-Escudero, L. A., Gordaliza, A., Martin R. S., and Mayo-Iscar, A. 2015. Avoiding Spurious Local Maximizers in Mixture Modeling. Stat. Computing. **vol. 25**, pp. 619 – 633.

Garc?a-Escudlero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., and Mayo-Iscar, A. 2016. The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. Comput. Statist. and Data Anal. **vol. 99**, pp. 131 – 147.

Gervini, D. and Yohai, V.J. 2002. A class of robust and fully efficient regression estimators. Ann. Statist. **vol. 30**, 583 – 616.

Giloni, A., Simonoff, J.S. and Sengupta, B. 2006. Robust weighted LAD regression. Comput. Statist. and Data Anal., **vol. 50**, pp. 3124 – 3140.

Green,P. J. 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, J. Roy. Statist. Soc. Ser. B **vol. 46,** pp. 149 – 192

Hadi, A. S. and Luceño, A. 1997. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Comput. Statist. and Data Anal., **vol. 25**, pp. 251 – 272.

Hand, D. J., Daly, F., Lunn, A.D., Mc Conway, K.J. and Ostrowski, E. 1994. *A Handbook of Small Data Sets* (Chapman & Hall, London

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P.J. and Stahel, W.A. 1986. Robust statistics. The approach based on influence functions. Wiley, New York.

Hardin, J., Rocke, D. M. 2004. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput. Statist. and Data Anal., **vol. 44**, pp. 625 – 638.

Hawkins, D. M. and Khan, D. M. 2009. A procedure for robust fitting in nonlinear regression. Comput. Statist. and Data Anal. **vol. 53**, pp. 4500 – 4507.

Hawkins, D. M. and Olive, D. J. 1999. Applications and algorithms for least trimmed sum of absolute deviations regression. Comput. Statist. and Data Anal.. 32, pp. 119 – 134.

Hawkins, D. M. and Olive, D. J. 2002. Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussions). J. Amer. Statist. Assoc. **vol. 97**, pp. 136 – 159.

He, X., Jurečkova, J., Koenker, R. and Portnoy, S., 1990. Tail behavior of regression estimators and their breakdown points. Econometrics, **vol. 58**, pp. 1195 – 1214.

Hennig, C. 2003. Clusters, outliers, and regression: fixed point clusters. J. of Multivariate Analysis, **vol. 86**, pp. 183 – 212.

Herwindiati, D. E., Djauhari, M. A. and Mashuri, M. 2009. Robust multivariate outlier labeling. Communications in Statistics: Simulation and Computation. **vol. 36**, pp. 1287 – 1294.

Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.-P. 2009. *Robust methods in biostatistics*. Wiley, Chichester, U.K.

Hössjer, O. 1994. Rank-based estimates in the linear model with high breakdown point. J. Amer. Statist. Assoc. **vol. 89,** pp. 149 – 158.

Huber, P. 1981. *Robust statistics*. John Wiley & Sons, New York.

Hubert, M. 1997. The breakdown value of the L_1 estimator in contingency tables. Statistics and Probability Letters. **vol. 33,** pp. 419 – 425.

Huber, P. and Ronchetti, E. 1981. *Robust statistics*. John Wiley & Sons, New York.

Hubert, M. and Rousseeuw, P. J. 1998. The catline for deep regression. J. Multivariate Analysis, **vol. 66**, pp. 270 – 296.

Hubert, M., Rousseeuw, P. J. and Van Aelst, S. 2005. Multivariate Outlier Detection and Robustness. In: Handbook of Statistics, **vol. 23**: *Data Mining and Computation in Statistics*, C.R. Rao, E. Wegman, and J.L. Solka (eds.), Amsterdam: Elsevier North-Holland, pp. 263 – 302.

Hubert, M., Rousseeuw, P. J. and Van Aelst, S. 2008. High-breakdown robust multivariate methods. Statistical Science, **vol. 23**, pp. 92 – 119.

Hunter, D. and Lange, K. (2000) Quantile regression via an MM. J. of Computational and Graphical Statistics, **vol. 9**, pp. 60 – 77.

Jennrich, R. I. and Moore, R. H. 1975. Maximum likelihood estimation by means of non-linear least squares. Proc. of the Statistical Computing Section of the Amer. Statist. Assoc. pp. 57 – 65.

Jørgensen, B., 1997. *The theory of dispersion models*. London: Chapman & Hall.

Jurečková, J. 2010. Finite-sample distribution of regression quantiles. Statistics and Probability Letters, **vol. 80**, pp. 1940 – 1946.

Khan, J. A., Van Aelst, S. and Zamar, R. H. 2007. Robust linear model selection based on least angle regression. J. Amer. Statist. Assoc., **vol. 102**, 1289 – 1299.

Kharin, Yu. S. 1996. *Robustness in Statistical Pattern Recognition*. Kluwer Academic Publishers, Dordrecht, London.

Koenker, R. W. 2005a. *Quantile Regression*. Cambridge University Press, Cambridge.

Koenker, R. W. 2005b. Quantile Regression in R. http://cran.R-project.org/doc/packages/quantreg/quantreg.pdf

Koenker, R. W. and Bassett, G. Jr. 1978. Regression quantiles. Econometrica, **vol. 84**, pp. 33 – 50.

Koenker, R. and Machado, J. 1999. Goodness of fit and related inference processes for quantile regression. J. Amer. Statist. Assoc. **vol. 94**, pp. 1296 – 1309.

Krivulin, N. 1992. An analysis of the least median of squares regression problem. In: Proceedings in Comput. Statist., Y. Dodge and J. Whittaker (eds.), Heidelberg, Physica-Verlag, pp. 471 – 476.

Künsch, H. R., Stefanski, L. A. and Carroll, R. J. 1989. Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. J. Amer. Statist. Assoc. **vol. 84,** pp. 460 – 466.

Lee, Y. and Nelder, J.A. 1998. Generalized linear models for the analysis of quality-improvement experiments. Can. J. Statist. **vol. 26**, pp. 95 – 105

Lee, Y. and Nelder, J.A. 2000. The relationship between double-exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex ration data. Appl. Statist. **vol. 49**, pp. 413 – 419.

Lee, Y., Nelder, J.A. and Pawitan, Y. 2006. Generalized Linear Models with Random Effects: Unified analysis via h-likelihood. London: Chapman & Hall/CRC.

Leisch, F. 2004. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. J. of Statist. Soft. 11, http://www.jstatsoft.org/

Marazzi, A. and Yohai, V. 2004. Adaptively truncated maximum likelihood regression with asymmetric errors. J. Statist. Plann. Inference, **vol. 122**, pp. 271 – 291.

Markatou, M., Basu, A. and Lindsay, B. 1997. Weighted likelihood estimating equations: The discrete case with applications to logistic regression. J. Statist. Plann. Inference. **vol. 57,** pp. 215 – 232.

Markatou, M. 2000. Mixture models, robustness, and the weighted likelihood methodology. Biometrics, **vol. 56**, pp. 483 – 486.

Maronna,R. A., Martin, R. D. and Yohai, V. J. 2006. *Robust Statistics: Theory and Methods*, John Wiley and Sons, New York.

Medasani, S., Krishnapuram, R. 1998. Robust mixture decomposition via maximization of trimmed log-likelihood with application to image database organization. In: Proceedings of the North American Fuzzy Information Society Workshop, Pensacola, August 1998, pp. 237 – 241.

McCullagh, P. and Nelder, J.A. 1989. *Generalized linear models*. London: Chapman & Hall.

McLachlan, G. J., Peel, D. 2000. *Finite mixture models*. Wiley, New York.

Mili, L. and Coakley, C. W. 1996. Robust estimation in structured linear regression. Ann. Statist. **vol. 15,** 2593 – 2607.

Mizera, I. and Müller, C. H. 1999. Breakdown points and variation exponents of robust M-estimators in linear models. Ann. Statist., **vol. 27**, 1164 – 1177.

Müller, Ch. H. 1995. Breakdown points for designed experiments. J. Statist. Plann. Inference. **vol. 45,** pp. 413 – 427.

Müller, Ch. H. 1997. *Robust Planning and Analysis of Experiments* (Shpringer, New York

Müller, C. H., Neykov, N. M. 2003. Breakdown points of the trimmed likelihood and related estimators in generalized linear models. J. Statist. Plann. Inference, **vol. 116**, pp. 503 – 519.

Nelder, J.A. and Pregibon, D. 1987. An extended quasi-likelihood function. Biometrika **vol. 74**, pp. 221 – 232.

Neykov, N. M. and Neytchev, P. N. 1990. A robust alternative of the maximum likelihood estimator. Short communications of COMPSTAT, Dubrovnik , pp. 99 – 100.

Neykov, N. M. (1995). *Robust methods with high breakdown point in the multivariate statistical analysis* Ph.D. Thesis, Faculty of Mathematics, Sofia University, (in Bulgarian).

Neykov, N. M. and Müller, C. H. 2003. Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (Eds.), Developments in robust statistics. Physica-Verlag, Heidelberg, pp. 277 – 286.

Neykov, N. M., Filzmoser, P., Dimova, R., and Neytchev, P. N. 2004. Mixture of generalized linear models and the Trimmed Likelihood methodology. In: Antoch (Ed.), Proceedings in Computational Statistics. Physica-Verlag, pp. 1585 – 1592.

Neykov, N. M., Dimova, R. and Neytchev, P. N. 2005. Trimmed Likelihood Estimation of the Parameters of the Generalized Extreme Value Distribution: A Monte-Carlo Study. Pliska Stud. Math. Bulgar. **vol. 17**, 187 – 200.

Neykov, N. M., Filzmoser, P., Dimova, R. and Neytchev, P. N. 2007. Robust fitting of mixtures using the Trimmed Likelihood Estimator. Comput. Statist. and Data Anal., **vol. 52**, pp. 299 – 308.

Neykov, N. M., Filzmoser, P. and Neytchev, P. N. 2012a. Robust joint modeling of mean and dispersion through trimming. Comput. Statist. and Data Anal. 56, 34 – 48.

Neykov, N. M., Cizek, P., Filzmoser, P. and Neytchev, P. N. 2012b. The least trimmed quantile regression. Comput. Statist. and Data Anal., **vol. 56**, 1757 – 1770.

Neykov, N. M., Filzmoser, P. and Neytchev, P. N. 2014. Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. Stat. Papers, **vol. 55**, 187 – 207.

Neytchev, P. N. Neykov,N. M. and Todorov,V. K. 1994. User's manual of REGRESS PC program system for fitting models to data. TR of NIMH, Sofia

O'Hara Hines, R. J. and Carter, E. M. 1993. Improved added variable and partial residual plots for the detection of influential observations in generalized linear models. Appl. Statist. **vol. 42,** pp. 3 – 20.

Prentice, R. L. 1976. A generalization of the probit and logit methods for dose response curves. Biometrics **vol. 32**, pp. 761 – 768.

Ribatet, M. and Iooss, B. 2009. Joint modeling of mean and dispersion package. http://cran.R-project.org/ doc/ packages/ JointModeling.pdf

Ritter, G. 2010. *Robust Cluster Analysis and Variable Selection*. Chapman & Hall / CRC Press.

Rousseeuw, P. J. (1984). Least median of squares regression. J. Amer. Statist. Assoc. **vol. 79,** pp. 851 – 857.

Rousseeuw, P. J. 1986. Multivariate Estimation with High Breakdown Point. In: Mathematical Statistics and Applications **Vol I** B, W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (eds.), Dordrecht: Reidel Publishing Company , pp. 283 – 297.

Rousseeuw, P. J. and Hubert, M. 1999. Regression depth. J. Amer. Statist. Assoc., **vol. 94**, pp. 388 – 402.

Rousseeuw, P. J. and Leroy, A. M. 1987. *Robust regression and outlier detection*. Wiley, New York.

Rousseeuw, P. J. and Van Driessen, K. 1999a. Computing least trimmed of squares regression for large data sets. Estadistica, **vol. 54**, pp. 163 – 190.

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

Ruwet, C., Garc?a-Escudero, L. A., Gordaliza, A., and Mayo-Iscar, A. 2013. On the breakdown behavior of the TCLUST clustering procedure. TEST, **vol. 22**, pp. 466 – 487.

Shane, K. V. and Simonoff, J. S. 2001. A Robust approach to categorical data analysis. J. Computational and Graphical Statistics, **vol. 10**, 135 – 157.

Schwartz, G. 1978. Estimating the dimension of a model. Ann. Statist., **vol. 6**, 461 – 464.

Smyth, G. K. 1989. Generalized linear models with varying dispersion. J. R. Statist. Soc. B **vol. 51**, pp. 47 – 60.

Smyth, G. K. 2009a. Double generalized linear models. http://cran.R-project.org/ doc/ packages/ dglm.pdf

Smyth, G. K. 2009b. Statistical Modeling. http://cran.R-project.org/ doc/ packages/ statmod.pdf

Smyth, G. K. and Verbyla, A. P. 1999. Adjusted likelihood methods for modelling dispersion in generalized linear models. Environmetrics **vol. 10**, pp. 696 – 709.

Smith, R. L. 1985. Maximum Likelihood estimation in a class of non regular cases. Biometrika, **vol. 72**, 67 – 90.

Stephenson, A. G. 2002. EVD: Extreme Value Distributions. R-News, 2, 31 – 32, URL http://CRAN.R-project.org/doc/Rnews/

Stromberg, A. J. and Ruppert, D. 1992. Breakdown in nonlinear regression. J. Amer. Statist. Assoc. **vol. 87,** pp. 991 – 997.

Stromberg, A. J., Hössjer, O. and Hawkins, D. M. 2000. The Least Trimmed Differences Regression Estimator and Alternatives. J. Amer. Statist. Assoc., **vol. 95**, pp. 853 – 864.

Tableman, M. 1994a. The asymptotics of the least trimmed absolute deviations (LTAD) estimator. Statistics and Probability Letters, **vol. 19**, pp. 387 – 398.

Tableman, M. 1994b. The influence functions for the least trimmed squares and the least trimmed absolute deviations estimator. Statistics and Probability Letters, **vol. 19**, pp. 329 – 337.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. J. Roy. Statist. Soc. Ser. B, **vol. 58**, 267 – 288.

Tibshirani, R. 1997. The lasso method for variable selection in the Cox model. Statist. Med., **vol. 16**, 385 – 95.

Van Aelst, S., Rousseeuw, P. J., Hubert, M. and Struyf, A. 2002. The deepest regression method. J. Multivariate Analysis, **vol. 81**, pp. 138 – 166.

Van der Vaart, A. W. and Wellner, J. A. 1996. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer, New York.

Vandev, D. L. 1993. A note on breakdown point of the least median squares and least trimmed squares. Statistics and Probability Letters **vol. 16,** pp. 117 – 119.

Vandev, D. L. and Marincheva, M. 1996. The BP of the WLT estimators in the general elliptic family of distribution. In: *Proc. of Statistical Dada Analysis*, Vandev, D.L. (ed.) Varna, pp. 25 – 31.

Vandev, D. L. and Neykov, N. M. 1993. Robust maximum likelihood in the Gaussian case, In: *New Directions in Data Analysis and Robustness*, Morgenthaler, S., Ronchetti, E. and Stahel, W. A. (eds.), Birkhäuser Verlag, Basel, pp. 259 – 264.

Vandev, D. L. and Neykov, N. M. 1998. About regression estimators with high breakdown point. Statistics, **vol. 32,** pp. 111 – 129.

Wang, H., Li, G. and Jiang, G. 2007. Robust regression shrinkage and consistent variable selection through the LAD-lasso. J. of Business & Economic Statist. **vol. 25**, 347 – 355.

Varmuza, K. and Filzmoser, P. 2008. *Introduction to multivariate statistical analysis in chemometrics*. CRC press, New York.

Windham, M. P. 1995. Robustifying model fitting. J. Roy. Statist. Soc. Ser. B **vol. 57**, 599 – 609.

Zhang, C. H. 2008. Discussion of One-step sparse estimates in nonconcave penalized likelihood models by H. Zou and R. Li. Ann. Statist., **vol. 36**, 1553 – 1560.

Zou, H. 2006. The Adaptive lasso and its oracle properties. J. Amer. Statist. Assoc., **vol. 101**, 1418 – 1429.

Zou, H. and Li, R. 2008. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Ann. Statist., **vol. 36**, 1509 – 1533.