Robust joint modeling of mean and dispersion through trimming

N.M. Neykov[*,a], P. Filzmoser[b], P.N. Neytchev[a]

[a] *National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66 Tsarigradsko chaussee, 1784 Sofia, Bulgaria*

[b] *Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria*

Abstract

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.

Key words: Extended quasi-likelihood, Extended trimmed quasi-likelihood, Generalized linear models, Joint modeling of mean and dispersion, Breakdown point, Outlier detection

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

$$g(\mu_i) = x_i^T \beta, \qquad h(\phi_i) = z_i^T \lambda \qquad \text{and} \qquad \operatorname{var}(y_i) = \phi_i V(\mu_i), \tag{1}$$

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 $\operatorname{var}(y_i) = \phi\mu_i$ and binomial distribution with $\operatorname{var}(y_i) = \phi\mu_i(1 - \mu_i/m_i)$ 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

$$Q^+(\beta, \lambda) \quad = \quad \sum_{i=1}^{n} q^+(y_i; \mu_i(\beta), \phi_i(\lambda)) = \sum_{i=1}^{n} q^+(y_i; \mu_i, \phi_i) = \sum_{i=1}^{n} -\frac{1}{2}\left\{\log\left[2\pi\phi_i V(y_i)\right] + \frac{d_i}{\phi_i}\right\}, \tag{2}$$

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

*Corresponding author. Tel.: +3592 4624597, Fax: +3592 9744409

Email addresses: Neyko.Neykov@meteo.bg (N.M. Neykov), p.filzmoser@tuwien.ac.at (P. Filzmoser), Plamen.Neytchev@meteo.bg (P.N. Neytchev)

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 (2) 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,

$$E(y_i) = \mu_i \qquad g(\mu_i) = \eta_i = x_i^T \beta \qquad \text{var}(y_i) = \phi_i V(\mu_i) \tag{3}$$

$$E(d_i) = \phi_i \qquad h(\phi_i) = \xi_i = z_i^T \lambda \qquad \text{var}(d_i) = 2\phi_i^2. \tag{4}$$

Setting the initial value for ϕ_i to a constant, the mean model (3) produces the deviances d_i as responses for the dispersion model (4) with dispersion parameter 2 and log-link function h, which in turn produces the prior weights $1/\phi_i$ for the mean model (3). 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 (4), where ρ_{ii} is the ith diagonal element of the projection matrix of the mean model (3) (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:

$$\min_{\beta} \left(u_m - X\beta\right)^T W_m \left(u_m - X\beta\right) \tag{5}$$

$$\min_{\lambda} \left(u_{d^*} - Z\lambda\right)^T W_{d^*} \left(u_{d^*} - Z\lambda\right), \tag{6}$$

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 $W_m^{1/2} X (X^T W_m X) X^T W_m^{1/2}$, 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 (5) 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) 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), Smyth (2009b) and Dunn (2009).

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 paper 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.

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

$$\hat{\theta}^{\mathrm{k}}_{\mathrm{wGTE}} := \arg\min_{\theta\in\Theta} \left\{ S(\theta) = \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta) \right\}, \tag{7}$$

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 (7). The combinatorial nature of the optimization problem is emphasized by the representation

$$\min_{\theta\in\Theta^p} S(\theta) = \min_{\theta\in\Theta^p} \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta) = \min_{\theta\in\Theta^p} \min_{I\in I_k} \sum_{i\in I} w_i f_i(\theta) = \min_{I\in I_k} \min_{\theta\in\Theta^p} \sum_{i\in I} w_i f_i(\theta), \tag{8}$$

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 (7) 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 (8). Then we obtain a particular case of a wGTE which we will call the maximum Extended Trimmed Quasi-Likelihood (ETQL) estimator.

Definition 2. The maximum ETQL estimator $(\hat{\beta}, \hat{\lambda})$ of (β, λ) is defined as

$$\max_{\beta,\lambda} Q^+_{\text{trim}}(\beta, \lambda) \quad = \quad \max_{\beta,\lambda} \max_{I\in I_k} \sum_{i\in I} q^+(y_i; \mu_i, \phi_i) = \max_{I\in I_k} \max_{\beta,\lambda} \sum_{i\in I} q^+(y_i; \mu_i, \phi_i). \tag{9}$$

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 (3) and (4) 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 (9). 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 (3) and (4) 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 (4) 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 (3) and (4) 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 (3) and (4) 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, the finite sample BDP of the maximum ETQL estimator equals $\frac{1}{n}\min\{n-k, k-\max[\mathcal{N}(X), \mathcal{N}(Z)]-1\}$. This BDP is maximized for $\lfloor\{n+\max[\mathcal{N}(X), \mathcal{N}(Z)]+1\}/2\rfloor \leq k \leq \lfloor\{n+\max[\mathcal{N}(X), \mathcal{N}(Z)]+2\}/2\rfloor$ and equals to $\frac{1}{n}\lfloor\{n-\max[\mathcal{N}(X), \mathcal{N}(Z)]-1\}/2\rfloor$.

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.

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 (8), we assume that the set F is d–full and

$k \geq d$. 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 $F^{old} = \{f_{i_1}, \ldots, f_{i_l}\} \subset F = \{f_1, \ldots, f_n\}$ 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 $F^{new} = \{f_{\nu(1)}, \ldots, f_{\nu(k)}\} \subset F$ 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 1. On the basis of steps 3 and 4 $S(\widehat{\theta}^{new}) \leq S(\widehat{\theta}^{old})$.
Proof. From the definition of $\widehat{\theta}^{old}$ and $\widehat{\theta}^{new}$ it follows that

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

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 (7)"*. There is no guarantee that the achieved solution will be the global minimizer of (7) 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.

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/bloodcpk.html`, 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 1 (left column) shows the (condensed) output of the statistical analysis, using the function *fitjoint* of the R package *JointModeling*.

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 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

```
### 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 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.

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 2 for $s = 2$ and Figure 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.

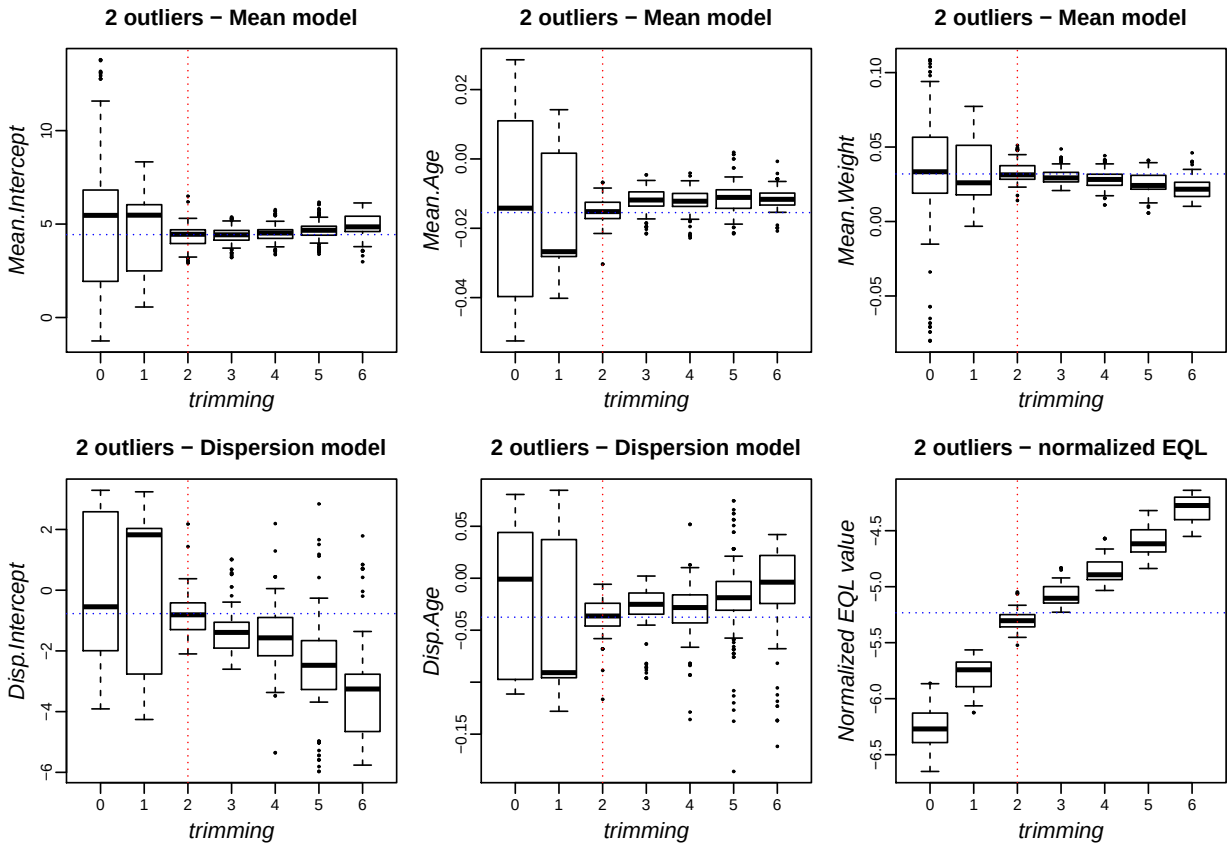

Figure 2: Boxplots of the joint modeling parameter estimates (intercept, Age and Weight in the mean model; intercept and Age in the dispersion model) when placing $s = 2$ outliers at any positions of the response variable, and varying the number of trimmed observations. Lower right panel: boxplots for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

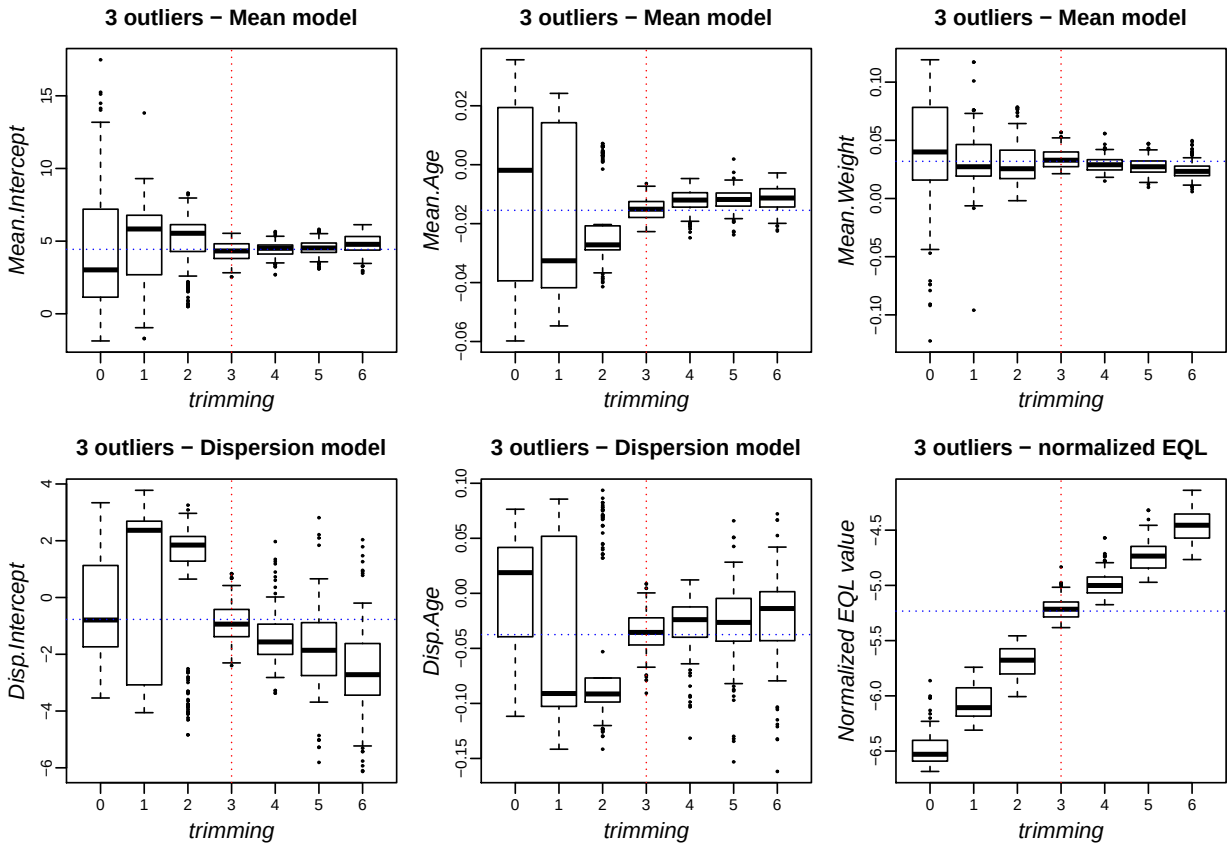

Figure 3: Boxplots for the joint modeling parameter estimates (intercept, Age and Weight for the mean model; intercept and Age for the dispersion model) when placing $s = 3$ outliers at any positions of the response variable, and varying the number of trimmed observations. Lower right panel: boxplots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

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`).

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 $y_{40} := y_{40} + 14$, 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 (4) 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.

5.2. Results and discussion of the 1st simulation experiment

The plots in Figures 4–6 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 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 5 are based on the experiments with contaminated data. Figure 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 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.

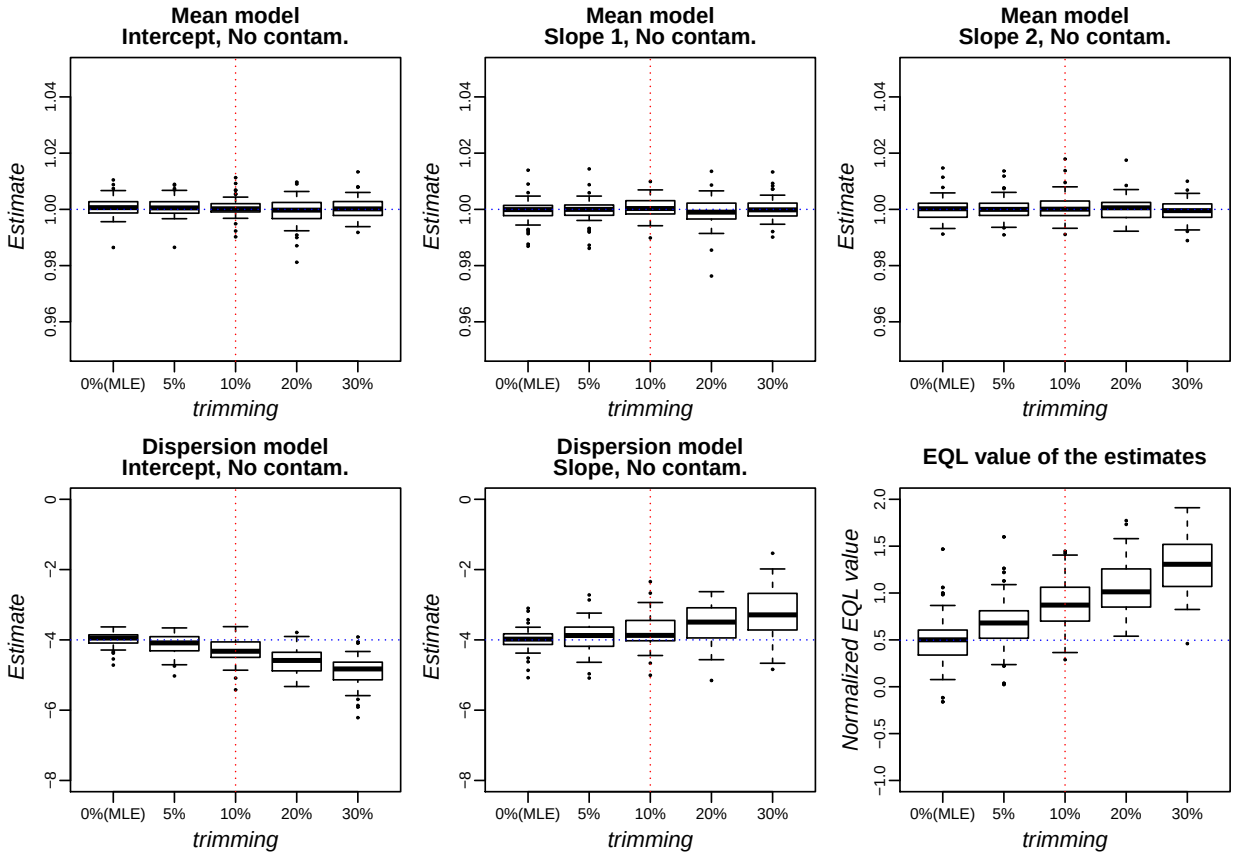

Figure 4: *1st simulation experiment without contamination*: boxplots of the estimates obtained from 1000 experiments for the joint normal mean and gamma dispersion GLMs parameters. Lower right panel: boxplots for the EQL values, normalized by the sample size.

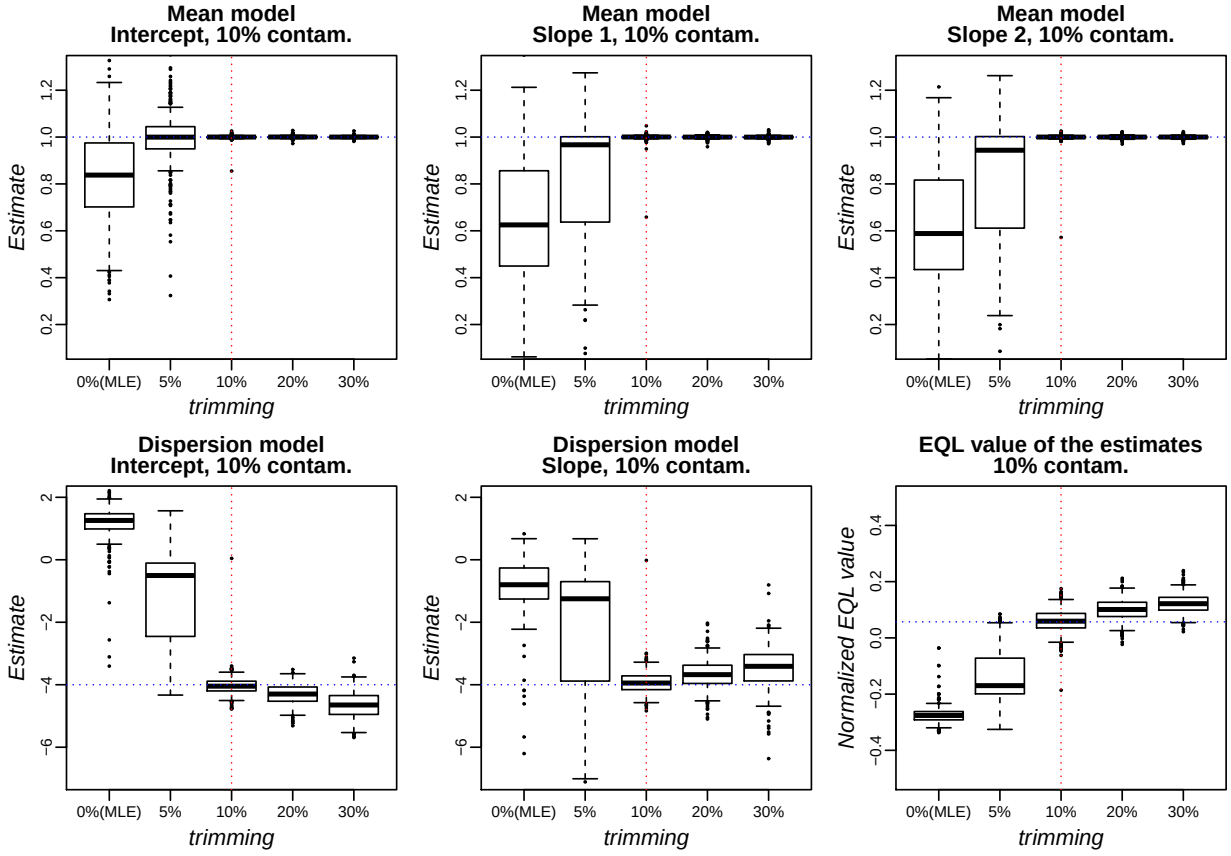

Figure 5: *1st simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the joint normal mean and gamma dispersion GLMs. Lower right panel: boxplots for the EQL values, normalized by the sample size.

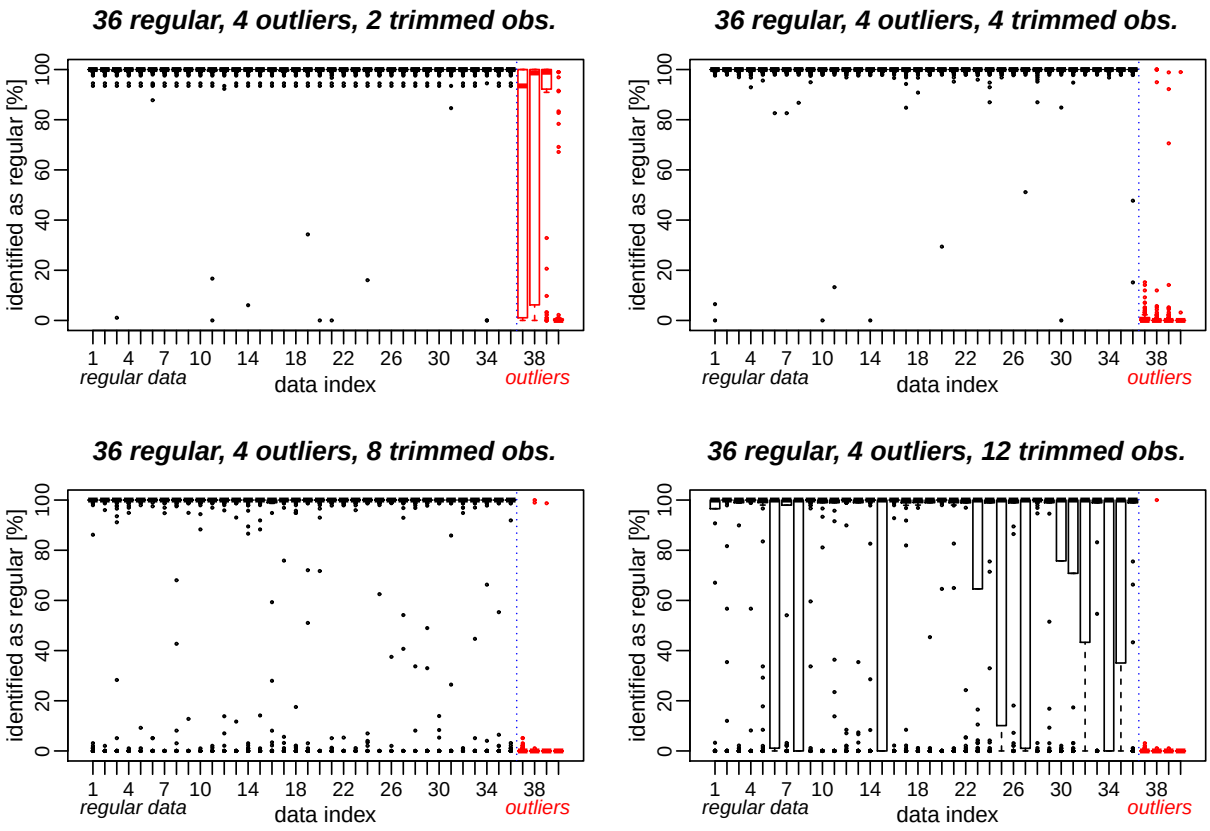

Figure 6: *1st simulation experiment with 10% contamination*: relative frequency distribution that an observation is identified as regular within the computational procedure of the algorithm within 1000 experiments.

It is interesting to look at the effect of model misspecification on the EQL and ETQL estimators. Figure 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 7 and 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.

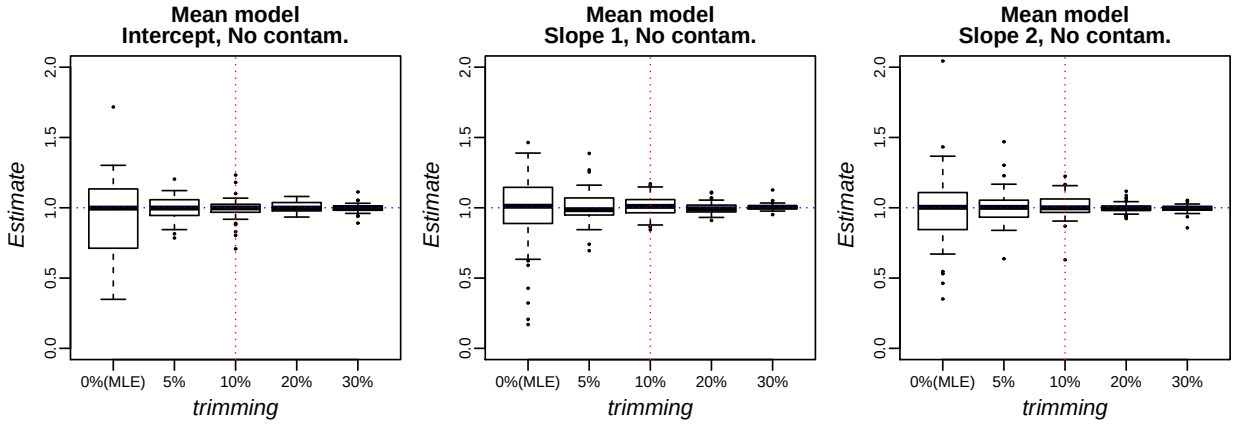

Figure 7: *1st simulation experiment without contamination*: boxplots for the estimates obtained from 1000 experiments for the normal mean model; dispersion parameter treated as constant.

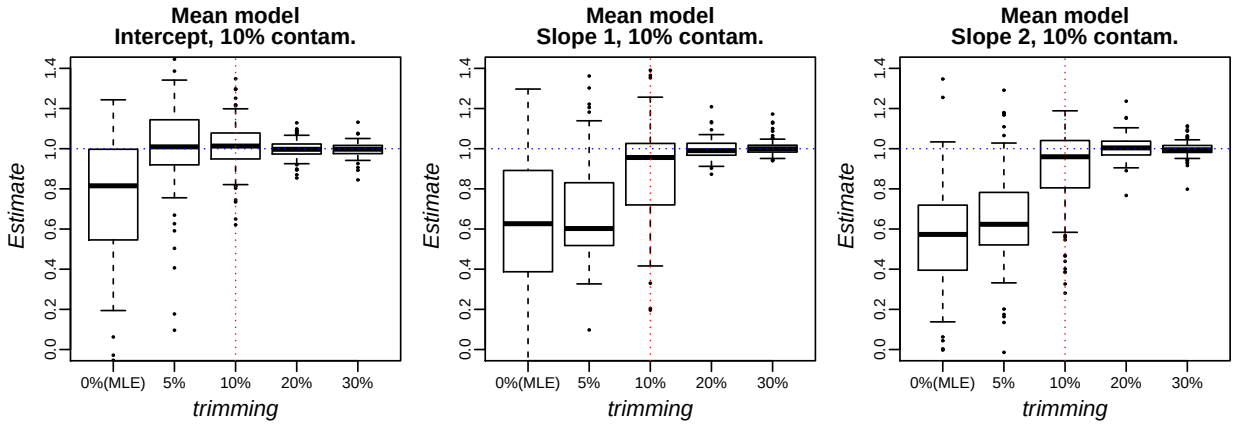

Figure 8: *1st simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the normal mean model; dispersion parameter treated as constant.

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 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 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 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.

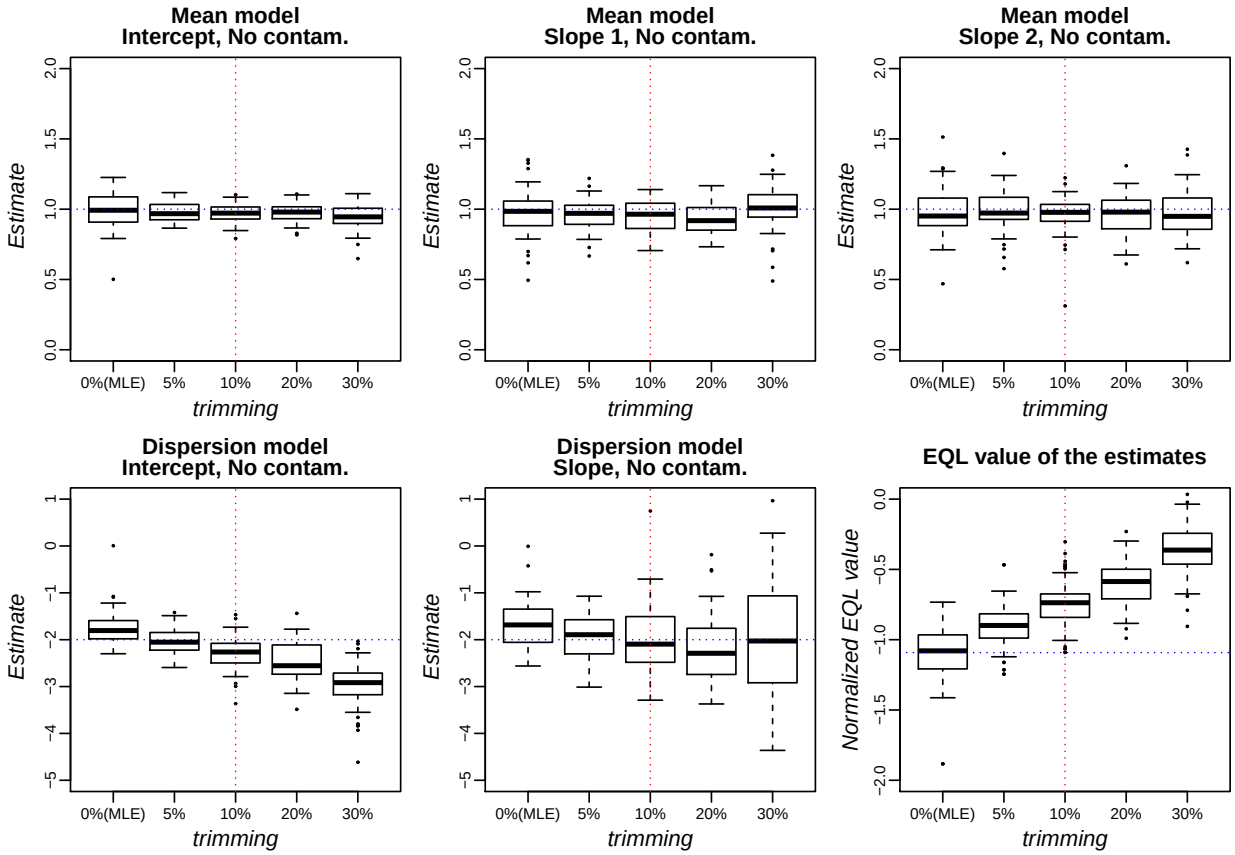

Figure 9: *2nd simulation experiment without contamination*: boxplots of the estimates obtained from 1000 experiments for the gamma mean and dispersion GLMs. Lower right panel: boxplots for the EQL values, normalized by the sample size.

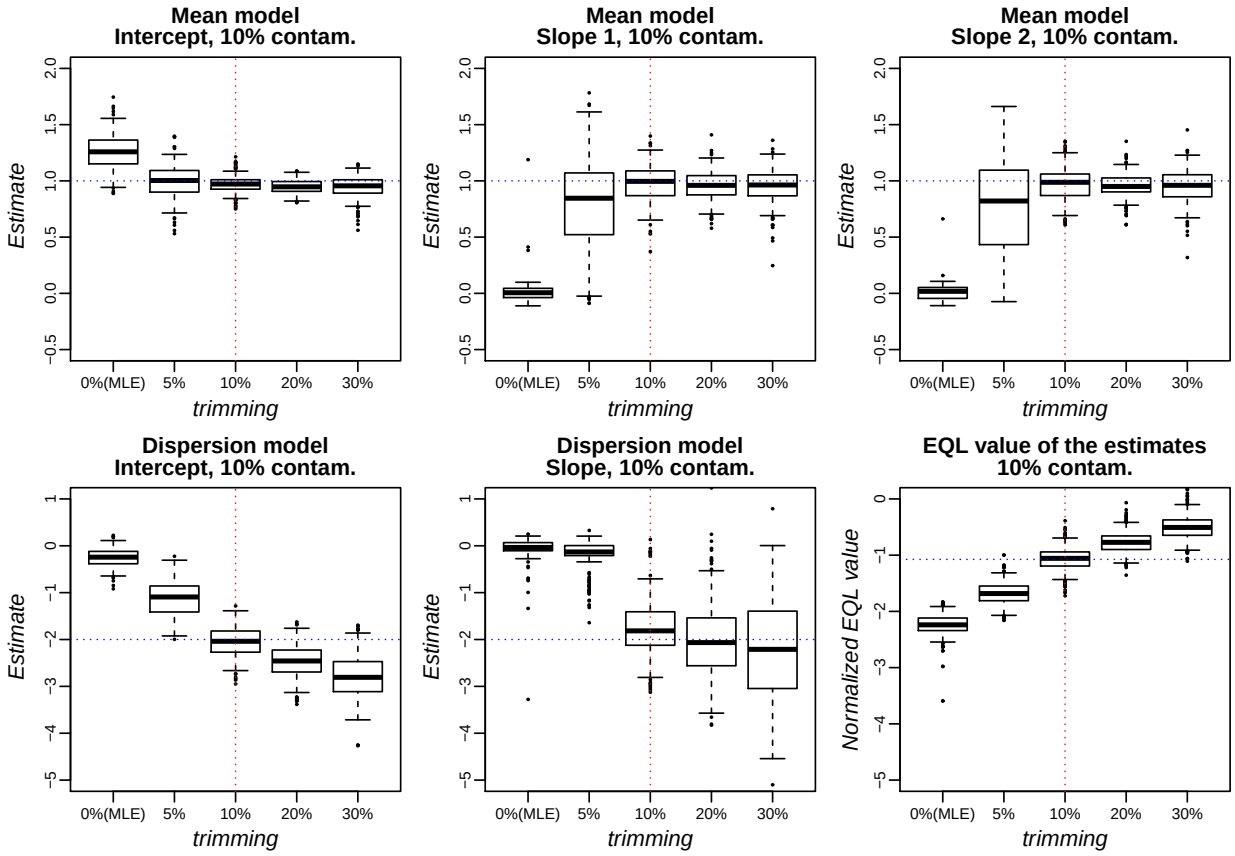

Figure 10: *2nd simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the gamma mean and dispersion GLMs. Lower right panel: boxplots for the EQL values, normalized by the sample size.

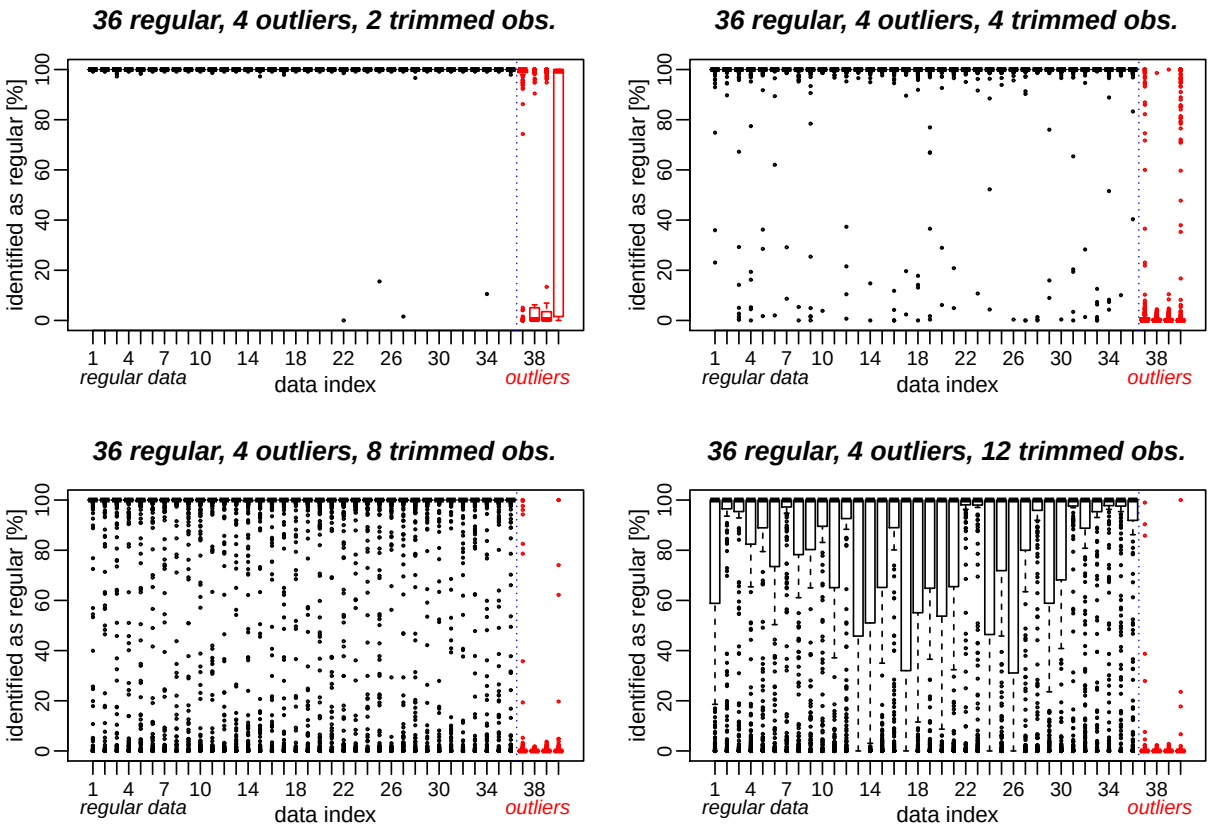

Figure 11: the relative frequency distribution that an observation is identified as regular within the computational procedure of the algorithm within 1000 experiments for the gamma mean and dispersion GLMs.

Similar as before, we want to show the effect of model misspecification. Figures 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 12. In the case of 10% contamination, from the plots of Figures 13 we see the EQL fails completely. ETQL, on the other hand, gives a very precise answer for different trimming percentages.

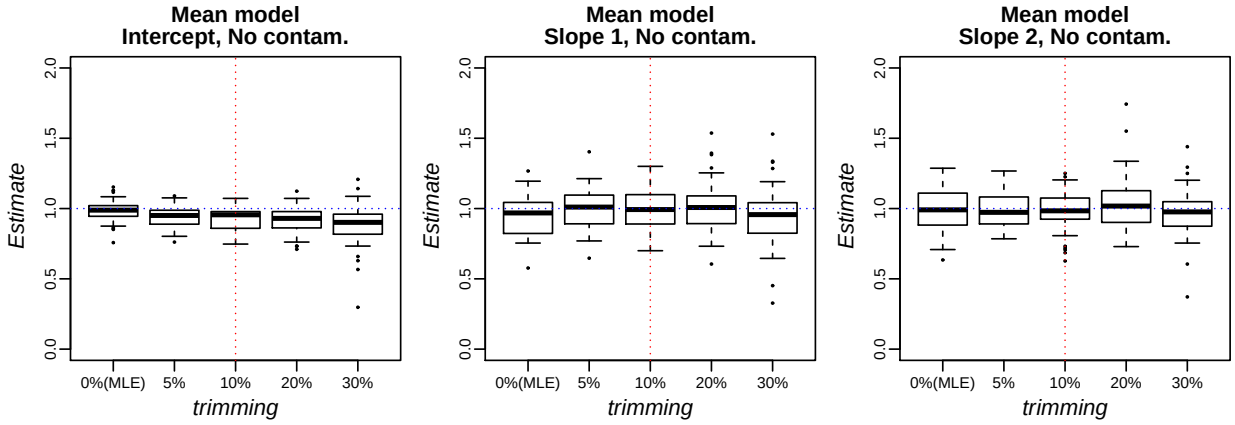

Figure 12: *2nd simulation experiment without contamination*: boxplots of the estimates obtained from 1000 experiments for the gamma mean GLMs; dispersion parameter treated as constant.

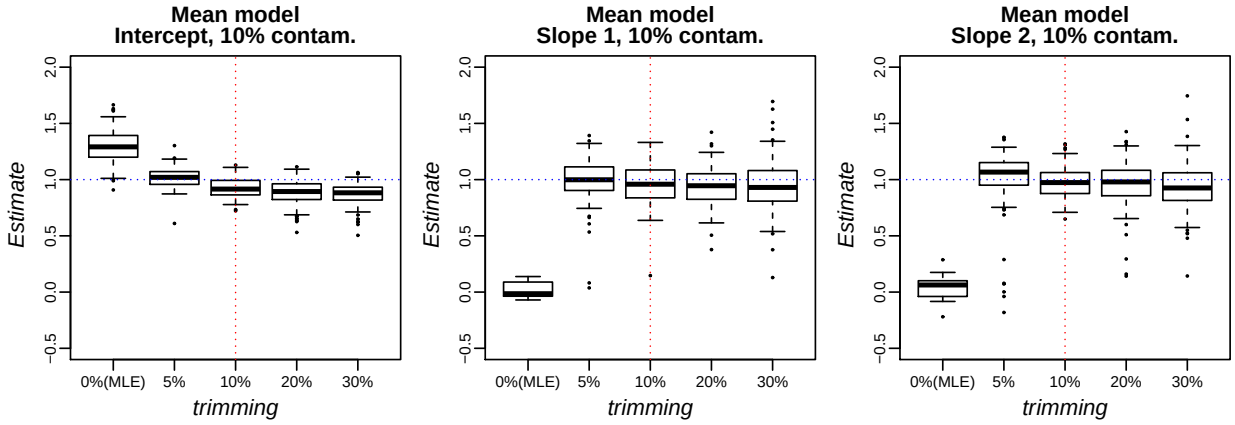

Figure 13: *2nd simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the gamma mean GLMs; dispersion parameter treated as constant.

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 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 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.

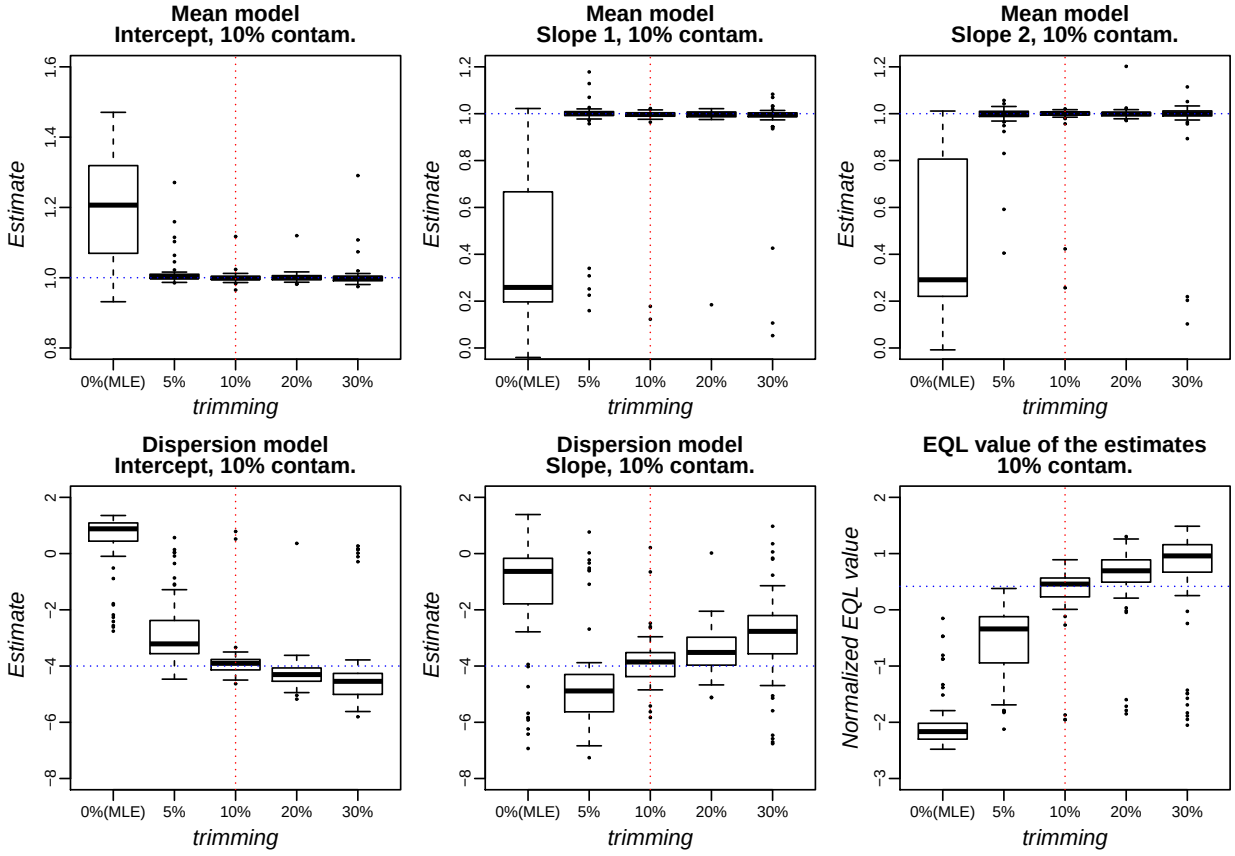

Figure 14: *3rd simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the Tweedie distribution with mean and dispersion model and power equal to 1.

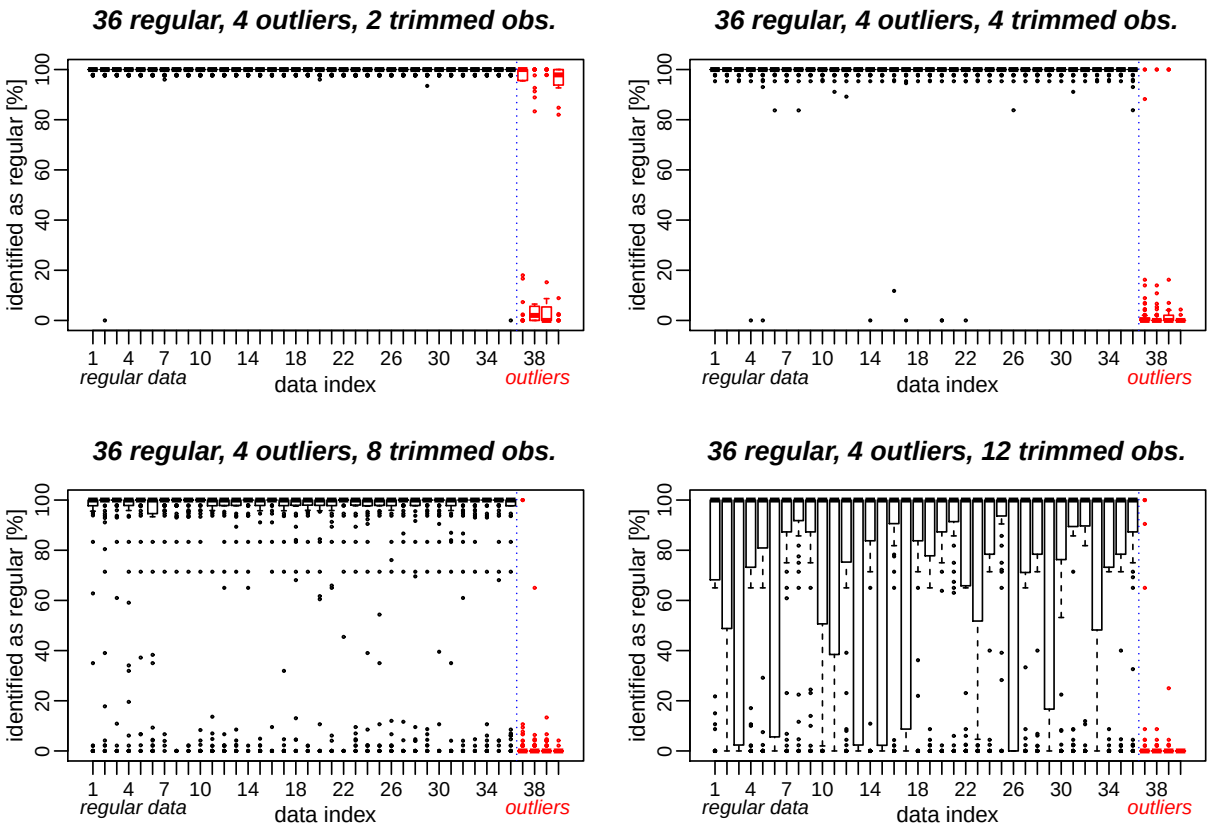

Figure 15: the relative frequency distribution that an observation is identified as regular within the computational procedure of the algorithm within 1000 for the Tweedie distribution with mean and dispersion model and power equal to 1.

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. 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 (2), leading to the ETQL loss function (9). All standard regression diagnostic techniques developed within the context of GLMs, McCullagh and Nelder (1989), can be used for model assessment and for the detection of structure in the remaining data.

R code of our method is available at `http://www.statistik.tuwien.ac.at/public/filz/programs.html`.

Acknowledgment

The authors are very grateful to Drs Ralitza Gueorguieva and Christophe Croux for their valuable comments on an earlier draft of the paper. We would like to thank the associate editor and two anonymous referees for their constructive comments on our work that greatly improved the presentation of these results. The authors are thankful to the Vienna University of Technology and the ESF (COST Action IC0702) for supporting the stay of N. Neykov and P. Neytchev in Vienna.

References

Cantoni, E. and Ronchetti, E., 2001. Robust inference for generalized linear models. J. Amer. Statist. Ass. 96: 1022-1030.

Cheng, T.-C., 2011. Robust diagnostics for the heteroscedastic regression model. Computational Statistics & Data Analysis 55: 1845-1866.

Čížek, P., 2008. General trimmed estimation: Robust approach to nonlinear and limited dependent variable models. Econometric Theory 24: 1500-1529.

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 70: 779-802.

Efron, B. 1986. Double exponential families and their use in generalized linear regression. J. Amer. Statist. Ass. 81: 709-721.

Dimova, R. and Neykov, N.M., 2004. Generalized d-fullness technique for breakdown point study of the trimmed likelihood estimator with applications. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhäuser, Basel, 83-91.

Dunn, P., 2009. Tweedie exponential family models.http://cran.R-project.org/doc/ packages/tweedie.pdf

Gallegos, M.T. and Ritter, G., 2005. A robust method for cluster analysis. Ann. Statist. 33: 347-380.

Gallegos, M.T. and Ritter, G., 2010. Using combinatorial optimization in model-based trimmed clustering with cardinality constraints. Computational Statistics & Data Analysis 54: 637-654.

Garcia-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A., 2008. A general trimming approach to robust cluster analysis. Ann. Statist. 36: 1324-1345.

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.

Hawkins, D.M. and Khan, D.M., 2009. A procedure for robust fitting in nonlinear regression. Computational Statistics & Data Analysis 53: 4500-4507.
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. 97: 136-159.
Herwindiati, D.E., Djauhari, M.A. and Mashuri, M., 2009. Robust multivariate outlier labeling. Communications in Statistics: Simulation and Computation. 36: 1287-1294.
Jørgensen, B., 1997. The theory of dispersion models. London: Chapman and Hall.
Lee, Y. and Nelder, J.A. 1998. Generalized linear models for the analysis of quality-improvement experiments. Can. J. Statist. 26: 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. 49: 413-419.
Lee, Y., Nelder, J.A. and Pawitan, Y., 2006. Generalized Linear Models with Random Effects: Unified analysis via h-likelihood. London: Chapman and Hall/CRC.
Markatou, M., Basu, A. and Lindsay, B., 1997. Weighted likelihood estimating equations: The discrete case with applications to logistic regression. J. Statist. Plann. Inference. 57: 215-232.
McCullagh, P. and Nelder, J.A., 1989. Generalized linear models. London: Chapman and Hall.
Müller, C.H. and Neykov, N.M., 2003. Breakdown points of the trimmed likelihood and related estimators in generalized linear models. J. Statist. Plann. Inference 116: 503-519.
Maronna,R.A., Martin, R.D. and Yohai, V.J., 2006. Robust Statistics: Theory and Methods, John Wiley and Sons, New York.
Nelder, J.A. and Pregibon, D., 1987. An extended quasi-likelihood function. Biometrika 74: 221-232.
Neykov, N.M. and Neytchev, P., 1990. A Robust Alternative of the Maximum Likelihood Estimators. COMPSTAT'90 - Short Communications, Dubrovnik, Yugoslavia, 99-100.
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, 277-286.
R Development Core Team, 2006. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
Ribatet,M. and Iooss,B., 2009. Joint modeling of mean and dispersion package. http://cran.R-project.org/ doc/ packages/ JointModeling.pdf
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 54: 163-190.
Rousseeuw, P.J. and Van Driessen, K., 1999b. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41: 212-223.
Smyth, G.K., 1989. Generalized linear models with varying dispersion. J. R. Statist. Soc. B 51: 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 10: 696-709.
Vandev, D.L. and Neykov, N.M., 1993. Robust maximum likelihood in the Gaussian case. In: Ronchetti, E., Stahel, W.A. (Eds.), New directions in data analysis and robustness. Birkhäuser Verlag, Basel, pp 259-264.
Vandev, D.L. and Neykov, N.M., 1998. About regression estimators with high breakdown point, Statistics 32: 111-129.
Visek, J.A., 1996. On high breakdown point estimation. Computational Statistics 11: 137-146.