Robust Statistical Modeling in Regression Settings via Trimming

Neyko M. Neykov

National Institute of Meteorology and Hydrology
Sofia, Bulgaria

Contents

1. Introduction to robust statistics

- Robust statistics is concerned with statistical procedure leading to inference that is stable with respect to departures of the data from model assumptions.
- Roughly speaking, the main purpose of any robust procedure is to give stable results in the presence or absence of outliers by best fitting the majority, the bulk of the data.
- Outliers 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 data sets with many observations and/or variables, and often they do not show up by simple visual inspection.
- Once found, the outliers should still be studded and interpreted, and not automatically be rejected (except in certain routine situation).

- Nowadays robust techniques have been developed in practically any field in statistical analysis. The milestones are books by Huber (1981)[7], Hampel et al. (1986)[5], Huber and Ronchetti (2009)[8], Maronna et al. (2019)[9].

- The book of Rousseeuw and Leroy (1987)[10] is mainly concerned with robust detection of regression and multivariate outliers based on trimming and is very practical.

- Atkinson and Riani (2000)[1], and Atkinson, Riani and Cerioli (2004)[2] and combine robustness with high BDP and various regression and multivariate data influential diagnostics and computer graphics. The so called Forward search algorithm is adapted 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)[6] gives robust methods in biostatistical modeling and statistical inference in general.

- Varmuza and Filzmoser (2008)[37] dissiminate the robust statistics in chemomemtrics.
- The book of Farcomeni and Greco (2015)[4] is about robust methods for data reduction techniques such as principal component, factor analysis, discriminant analysis, and clustering.
- Atkinson et al. (2025) [3] is open access book, presents robust statistical methods and procedures and their applications, with a focus on regression and provides numerous data analyses using MATLAB and R within regression framework.

2. The Least Squares Estimator and Related

- Consider the multiple linear regression model

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

where y_i are observed responses, $x_i^T = (x_{i1}, \ldots, x_{ip})$ are covariate vectors, $\beta_{p\times 1}$ is a vector of unknown parameters, ε_i are iid, $\mathrm{E}(\varepsilon_i)=0$, $\mathrm{var}(\varepsilon_i)=\sigma^2$. Denote by $r_i(\beta) = y_i - x_i^T \beta$ the regression residuals.

Definition 1. The LSE, Least Median of Squares (LMS), the Least Quantile of Squares (LQS) and the Least Trimmed Squares (LTS) are difined:

$$\min_\beta \sum_{i=1}^{n} r_i^2(\beta), \qquad \min_\beta \operatorname*{med}_i r_i^2(\beta), \qquad \min_\beta r^2_{\nu(k)}(\beta), \qquad \min_\beta \sum_{i=1}^{k} r^2_{\nu(i)}(\beta)$$

where $r^2_{\nu(1)}(\beta) \le r^2_{\nu(2)}(\beta) \le \ldots \le r^2_{\nu(n)}(\beta)$ are the ordered values of $r_i^2(\beta)$ at β; $\nu = (\nu(1), \ldots, \nu(n))$ is the permutation of the indices (depends on β); k is the trimming parameter such that $\lfloor \frac{n+p+1}{2} \rfloor \le k \le n$; see Rousseeuw (1984)[31]

- The objective function LMS, LQS and LTS are continuous, but non differentiable and possesses many local minima.
- The following representation of the LQS due to Krivulin (1992) is very useful as it clarifies its combinatorial nature and further reducing the problem

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

  where I_k is the set of all k–subsets of the set $\{1, \ldots, n\}$, whereas $I = \{i_1, \ldots, i_k\}$.
- The same holds about the LTS estimator

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

- Therefore all possible $\binom{n}{k}$ subsets of the data have to be fitted by the LSE.
- The LTS is given by those subsets with the smallest LSE fit criteria.
  Rousseeuw & van Driessen (2000) developed FAST-LTS algotithm to get an approximate estimate.

The Breakdown Point notion

- To aid the presentation we remind the replacement variant of the finite sample breakdown point (BDP) given in Hampel et al. (1986) [17], which is closely related to that introduced by Donoho and Huber (1983).
- Let $Z = \{z_i \in \mathbb{R}^p, \text{ for } i = 1, \ldots, n\}$ be a sample of size n.

  Definition 2. The BDP of an estimator T at Z is given by

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

  where $\tilde{Z}_m$ is a sample obtained from Z by replacing any m of the points in Z by arbitrary values from and $\|.\|$ is the Euclidean norm.
- In other words the BDP is the smallest fraction of contamination that can cause the estimator to take arbitrary large values.

3. Detection of multivariate outliers

- Let the a multivariate data set is given by $X_{n\times p} = (x_{ij})$, for $j = 1, \ldots, p$ and $i = 1, \ldots, n$, where n and p are the number of observations and variables.
- The classical estimates for the location and scatter are defined by

$$T(X) = \bar{x} = \sum_{i=1}^{n} x_i \quad \text{and} \quad C(X) = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

- The classical outlier detection procedure is based on the squared Mahalanobis distance defined by

$$MD_i^2 = (x_i - T(X))^T C(X)^{-1} (x_i - T(X)), \quad \text{for} \quad \mathrm{i} = 1, 2, \ldots, \mathrm{n}$$

  where $T(X)$ and $C(X)$ are the usual location and covariance estimates.
- Points whose MD_i^2 is large are flagged and/or deleted. The remaining data are processed in the usual way.

- However, that the sample mean and covariance in a multivariate data set are extremely sensitive to outliers.
- To repair this the Minimum Covariance Determinant (MCD) of Rousseeuw (1984) Rousseeuw and Leroy (1987) is recommended.

  Definition 3. The MCD is determined by the subset of size h whose covariance matrix has the smallest determinant. The MCD location estimate T is defined as the mean of that subset (the MCD scatter estimate C is defined as a multiple of its covariance matrix).

- If $h = \lfloor (n+p+1)/2 \rfloor$ then MCD get the maximal BDP, which is equal to 50%.
- The RD_i^2 distance is proposed instead of MD_i^2 which is defined by replacing $T(X)$ and $C(X)$ by their MCD analogs.

4. Detection of multiple linear regression outliers

- The hat elements, $h_{ii} = x_i^T (X^T X)^{-1} x_i$, the diagonal of the matrix $X(X^T X)^{-1} X^T$ should not be used as a tool for identification of leverage observations because is based on the classical estimates of the location and the covariance that possess zero breakdown point

$$h_{ii} = \frac{MD_i^2}{n-1} + \frac{1}{n}.$$

- Using the robust distance based on the one-step improvement MCD estimator RD_i for each case the data automatically falls into two categories: the regular observations with small RD_i and bad observations (outliers) with large RD_i.
- The distance-distance plot (D-D plot) plots the robust distances RD_i (based on MCD) versus the classical Mahalanobis distancesMD_i. The cutoff value $\sqrt{\chi^2_{p,0.975}}$ is indicated on both axes.
- Therefore the following procedure is recommended for regression outliers identification

(i) Perform LTS regression estimation

(ii) Compute the standardized residuals $r_i(\hat{b})/s$, where s is the final scale estimate and $r_i(\hat{b})$ are residuals based on LTS regression estimate $\hat{b}$.

(iii) Identify the ith observation as LTS regression outlier if $|r_i/s| > 2.5$.

(iv) Apply the MCD technique to the explanatory variables only. Compute the corresponding RD_i for each observation. Classify the data as *regular* and *bad* observations according to the magnitude....

(v) Plot the standardized residuals obtained in step (ii) versus the robust distance RD_i obtained in step (iv).

Example

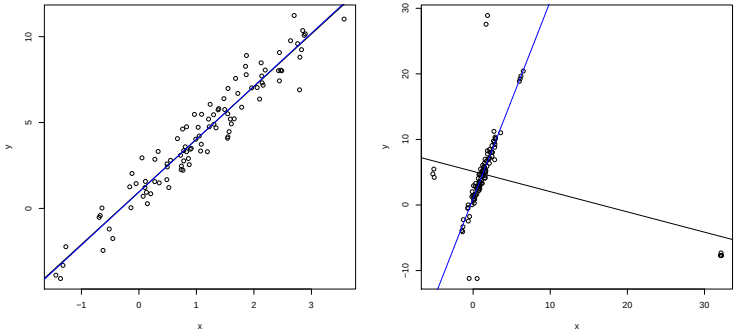

Figure 1: Plot left: data - $y_i = 1 + 3x_i + \varepsilon_i;\ x_i,\ \varepsilon_i \sim N(0,1)$, for $i = 1, \ldots, 112$, the LSE and LTS lines coincide; Plot right: contaminated data, LTS - the blue line; LSE - the black line.

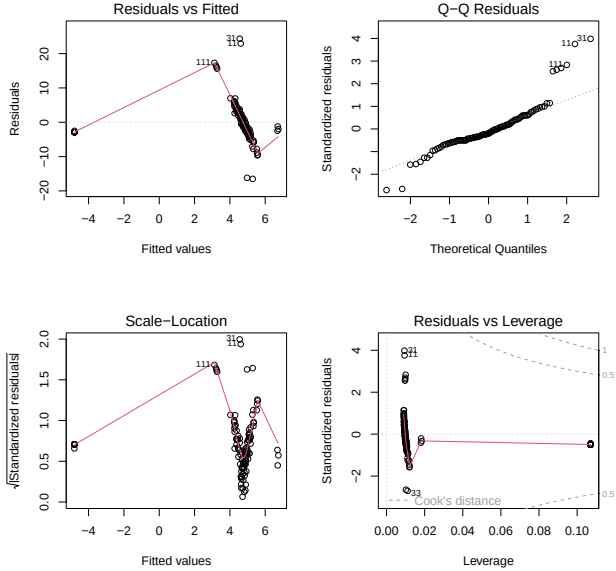

Figure 2: The classical LSE regression diagnostic plots

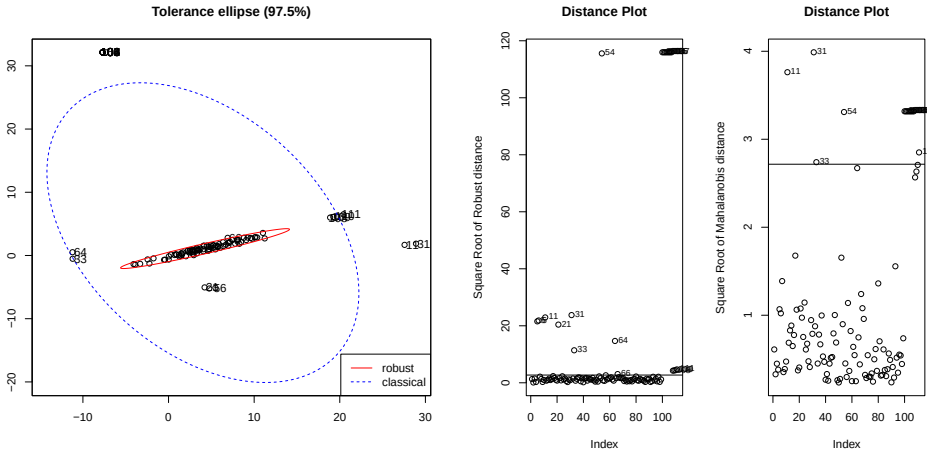

Figure 3: Plot left: Classical and MCD elippsoids; Plots right: Robust based MCD and Mahalanobis distances.

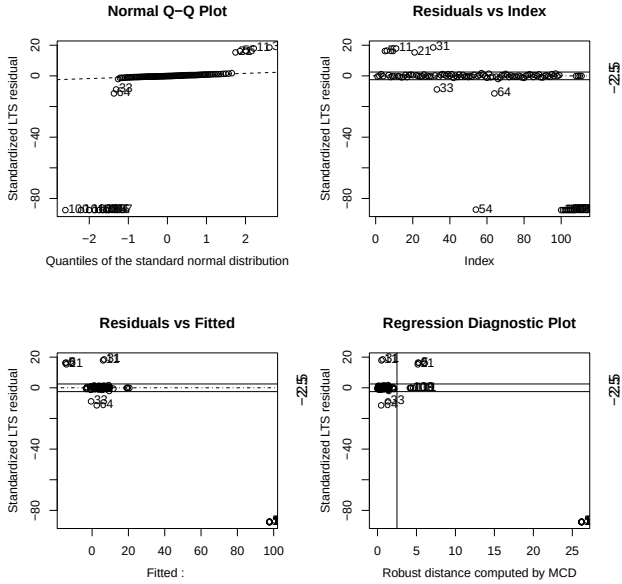

Figure 4: The LTS regression diagnostic plots in conjunction with the MCD based distances.

5. The trimmined likelihood estimator and related

- Let $z_i \in R^p$ for $i = 1, \ldots, n$ be i.i.d. observations with pdf $g(z, \theta)$, $\theta \subseteq \Theta^q$ is unknown parameter, and $l_i(\theta) = l(z_i, \theta) = -\log g(z_i, \theta)$.

- Neykov and Neytchev (1990) [25] proposed to replace in the Rousseeuw's estimators the squared residuals $r_i^2(\beta)$ by the negative log likelihoods $l_i(\theta)$ and thus the following two classes of estimators are defined:

Definition 4. The MedLE (MedLE(k)) and Trimmed Likelihood Estimator (TLE(k)) are defined as

$$\min_{\theta \in \Theta} l_{\nu(k)}(\theta) \qquad \text{and} \qquad \min_{\theta \in \Theta} \sum_{i=1}^{k} l_{\nu(i)}(\theta) \tag{3}$$

Remark 5. The MedLE and TLE are reduced to LMS and LTS in case of regression model with normal distribution of the errors and MVE and MVE covariance estimators in case multivariate normal distributions.

The Weighted Generalized Trimmed Estimator (wGTE)

Recall the definition of the weighted Generalized Trimmed Estimator (wGTE).
Let $f_i : \Theta \to \mathbb{R}^+$, $\Theta \subseteq \mathbb{R}^q$ and $F = \{f_i(\theta),\ \theta \in \Theta \text{ for } i = 1, \ldots, n\}$.

Definition 6. (Vandev & Neykov (1998) [36]) The wGTE is defined as

$$\min_{\theta \in \Theta^p} \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta) = \min_{\theta \in \Theta^p} \min_{I \in I_k} \sum_{i \in I} w_i f_i(\theta) = \min_{I \in I_k} \min_{\theta \in \Theta^p} \sum_{i \in I} w_i f_i(\theta) \tag{4}$$

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

- The wGTE trims those n-k observations which values would be highly unlikely to occur, had the fitted model been true.
- Therefore the optimization is infeasible for large n.

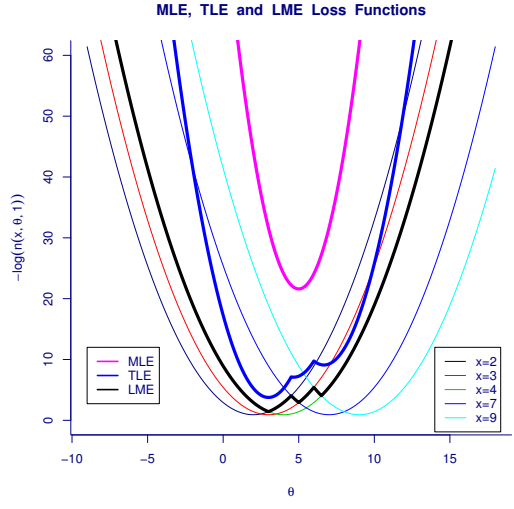

Figure 5: Negative log likelihoods of $N(\mu \in \{2,3,4,5,7,9\}, \sigma^2 = 1)$, MLE, TLE and median likelihood estimator **LME≡MedLE** loss functions

The BDP and d-fullness notion

Theorem 7. (Vandev & Neykov (1998) [36]). If F is d-full the BDP of the wGTE is not less than $(n-k)/n$ if $n \geq 3d$ and $(n+d)/2 \leq k \leq n-d$.

Definition 8. (Vandev (1993) [34]) Let $f_i : \Theta \to \mathbb{R}^+$, $\Theta \subseteq \mathbb{R}^q$. The set $F = \{f_1, \ldots, f_n\}$ is called d–full if for every $J \subset \{1, \ldots, n\}$ of cardinality d the function $g(\theta) = \max\limits_{j \in J} f_j(\theta)$ is subcompact.

Definition 9. (Vandev (1993) [34]) 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.

Remark 10. The asymptotic properties of the wGTE were studied by Čížek (2006a, 2008a, 2008b) [11], [12], [13], for the case of twice differentiable functions $f_i(\theta)$.

BDP of TLE and related within GLMs

Let $s : \Theta \to R^+$ and there exists $\alpha,\ \beta \in R$ with $\alpha \neq 0$ such that for all $\theta \in \Theta$

$$\alpha\, f_{\nu(k)}(\theta) \leq s(\theta) \leq \beta\, f_{\nu(k)}(\theta) \tag{5}$$

Definition 11. (Müller & Neykov (2003) [23]) The s estimator is defined as

$$\hat{\theta}_S := \arg\min_{\theta \in \Theta} s(\theta) \tag{6}$$

Theorem 12. (Müller & Neykov (2003) [23]). If F is d-full the BDP of the s is not less than $\frac{1}{n}\min(n-k, k-d)$

Remark 13. The lower bound of Theorem 12 is maximized if the trimming factor k satisfies $\left\lfloor \frac{n+d}{2} \right\rfloor \leq k \leq \left\lfloor \frac{n+d+1}{2} \right\rfloor$.

Remark 14. The wGTE estimators satisfies condition (5).

Remark 15. Theorem 12 is an extension of Theorem 1 of Vandev and Neykov (1998) [36] and provides the lower bound without additional assumptions on n and k.

Theorem 16. *Let* $\{l_i(y,\cdot);\ i=1,\ldots,n\}$ *is* d*-full and* $\left\lfloor \frac{n+d}{2} \right\rfloor \le k \le \left\lfloor \frac{n+d+1}{2} \right\rfloor$. *The BDP of the* $wTLE(k)$ *satisfies*

$$\epsilon^*(wTLE(k), y) \ge \frac{1}{n} \left\lfloor \frac{n-d+2}{2} \right\rfloor.$$

Theorem 17. *The BDP of the* $wTLE(k)$ *for the normal multiple linear, logistic linear* $(0 < s_i < t_i)$ *and linear Poisson* $(y_i > 0)$ *regression satisfies*

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

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

where $X := (x_1,\ldots,x_n)^\top \in R^{n \times p}$. $N(X)$ provides the maximum number of covariates lying in a subspace. $N(X) = p-1$ if any p obs. x_i^T are linearly independent.

FAST-TLE Algorithm

- The FAST-GTE algorithm consists of carrying out finitely many times a two-step procedure of a trial step followed by a refinement step;
- Trial step: $\widehat{\theta}^* := \underset{\theta\in\Theta^p}{\arg\min} \sum_{j=1}^{k^*} f_{i_j}(\theta)$, $F^* = \{f_{i_1}, \ldots, f_{i_{k^*}}\} \subset F$, $k^* \geq d$
- The refinement step is based on the so called concentration procedure:

  For $\widehat{\theta}^1 := \widehat{\theta}^*$ and $s = 1, 2, \ldots, S$ execute the steps:

  1. sort $f_i(\widehat{\theta}^s)$: $\quad f_{\nu(1)}(\widehat{\theta}^s) \leq \ldots \leq f_{\nu(k)}(\widehat{\theta}^s) \leq \ldots \leq f_{\nu(n)}(\widehat{\theta}^s)$
  2. get $F^k := \{f_{\nu(1)}, \ldots, f_{\nu(k)}\}$;
  3. compute $\widehat{\theta}^{s+1} := \underset{\theta\in\Theta^p}{\arg\min} \sum_{i=1}^{k} f_{\nu(i)}(\theta)$ based on F^k;
     - as a consequence $\quad \sum_{i=1}^{k} f_{\nu(i)}(\widehat{\theta}^{s+1}) \leq \sum_{i=1}^{k} f_{\nu(i)}(\widehat{\theta}^s)$;
  4. set $\widehat{\theta}^s := \widehat{\theta}^{s+1}$ and cycle steps 1-4 until convergence or $s = S$;
- The solution with the lowest lost function value of this procedure is stored.

6. Robust fitting of mixture models through trimming

- McLachlan and Peel (2000) [22]. 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 g distributions, conditional on the variables $x_i \in R^p$, in proportions $\pi_1, \ldots, \pi_g$, defined by

$$\varphi(y_i; x_i, \Psi) = \sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j), \tag{7}$$

  $\Psi = (\pi_1, \ldots, \pi_{g-1}, \theta_1, \ldots, \theta_g)^T$, $\theta_j \in \Theta^{p+1}$, $\pi_j \geq 0$, $j = 1, \ldots, g$, $\sum_{j=1}^{g} \pi_j = 1$.

- The log likelihood of Ψ is given by

$$\log L(\Psi) = \sum_{i=1}^{n} \varphi(y_i; x_i, \Psi) = \sum_{i=1}^{n} \log\Big\{ \sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j) \Big\} \tag{8}$$

- $\log L(\Psi)$ is not maximized directly.

- As a consequence of the EM algorithm, the expection of the trimmed complete-data negative log-likelihood estimator is minimized

$$\min_{\Psi} \min_{I \in I_k} \sum_{i \in I} \sum_{j=1}^{g} -\tau_j(y_i; x_i, \Psi^{(l)}) \{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}. \tag{9}$$

- Therefore Ψ have to be estimated for each $I \in I_k$ by

$$\min_{\Psi} \sum_{i \in I} -\tau_j(y_i; x_i, \Psi^{(l)}) \log \psi_j(y_i; x_i, \theta_j) \quad \text{for } j = 1, \ldots, g. \tag{10}$$

Proposition 18. *If the* $F_{\theta_j} = \{-\log \psi(x_1, \theta_j), \ldots, -\log \psi(x_n, \theta_j)\}$ *is* d_j*–full for any* $j = 1, \ldots, g$ *and* $d = \max(d_1, d_2, \ldots, d_g)$ *then the BDP of the mixture model distributions (9) satisfies* $\varepsilon_n^*(wTLE) \geq \frac{1}{n} \min(n - k, k - d)$.

Mixture of 3 regression lines with noise

- We consider mixture model with 3 simple regression lines:

$$\begin{aligned}
\text{Class 1}: \quad & y_i = 3 + 1.4x_i + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, 70 \\
\text{Class 2}: \quad & y_i = 3 - 1.1x_i + \varepsilon_i \quad \text{for} \quad i = 71, \ldots, 140 \\
\text{Class 3}: \quad & y_i = 0 + 0.1x_i + \varepsilon_i \quad \text{for} \quad i = 141, \ldots, 200 \\
\text{with} \quad & \varepsilon_i \sim N(0, 0.1) \\
\text{and} \quad & x \sim U(-3, -1) \cup U(1, 3).
\end{aligned}$$

- 50 outliers uniformly distributed in the area $[-4.5, 4.5] \times [-0.8, 2.8]$ were added to this data;
- The mixing proportions are $\pi_1 = \pi_2 = 0.35$ and $\pi_3 = 0.3$;
- A fit is successful if the 3 lines are correctly estimated.

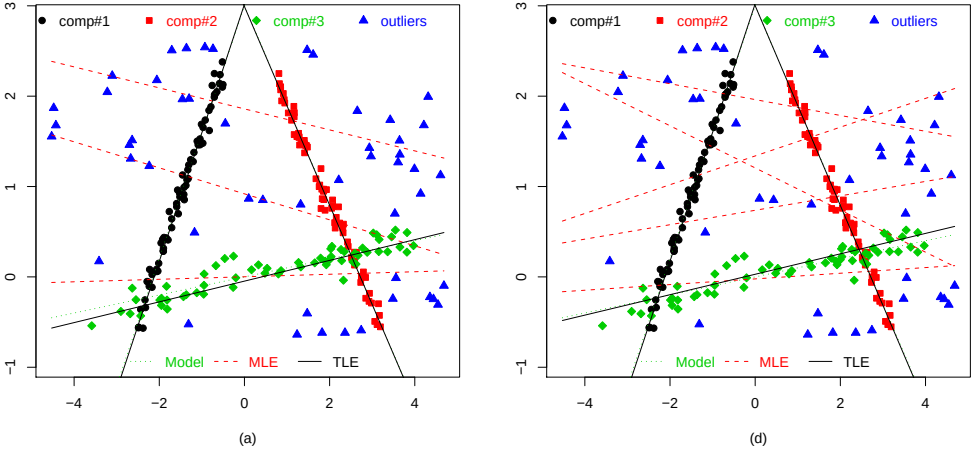

Figure 6: Mixture of three regressions: true model (dotted lines), fits based on the MLE (dashed lines) and FAST-TLE (solid lines) with (a) 20% trimming and 3 components, (b) 40% trimming and 5 components.

Mixture of multivariate normals with noise

- MacLachlan and Peel (2000) [22] considered a sample of 100 simulated points from a 3-component bivariate normal mixture model, to which 50 noise points were added from a uniform distribution over the range -10 to 10 on each variate;
- The parameter of the mixture model are:

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

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

- The mixing proportions are $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$;
- MacLachlan and Peel (2000) [22] modeled the data by a mixture of t–distributions.
- A fit is successful if all components are correctly estimated.

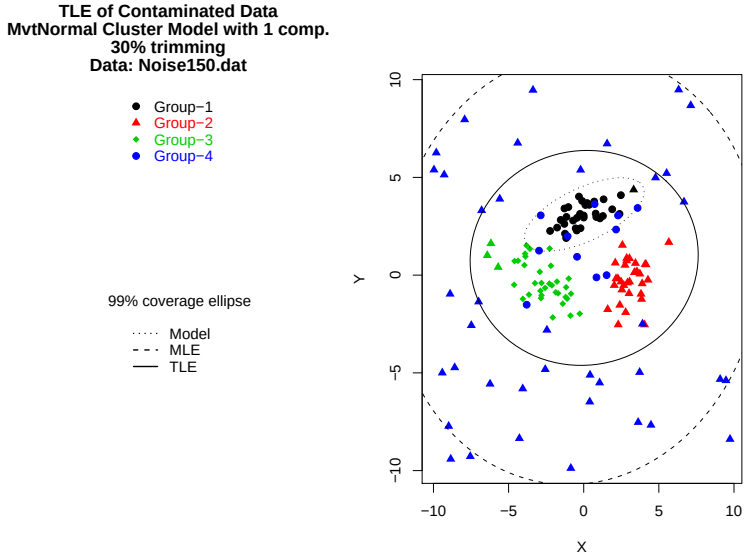

Figure 7: One component fit of mixture of 3 bivariate normals - 30% trimming.

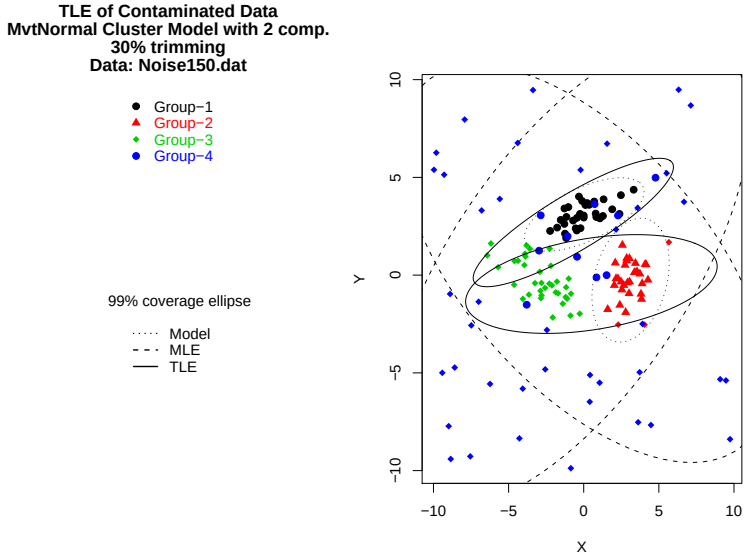

Figure 8: Two components fit of mixture of 3 bivariate normals - 30% trimming.

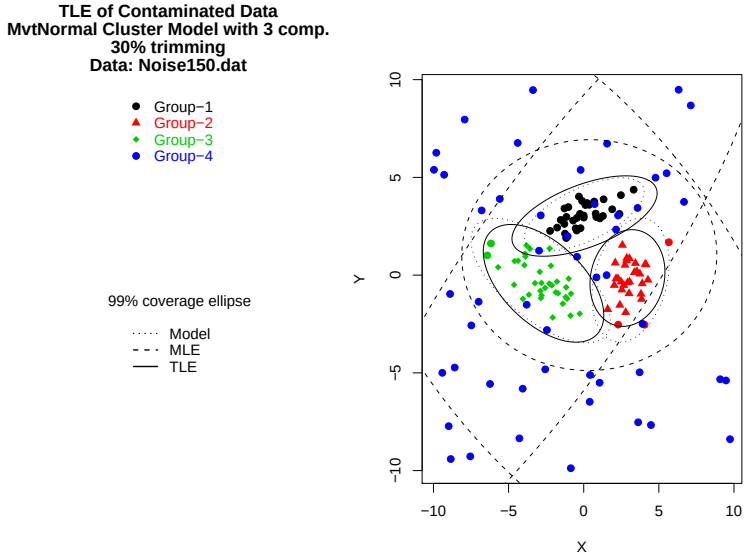

Figure 9: Three components fit of mixture of 3 bivariate normals - 30% trimming.

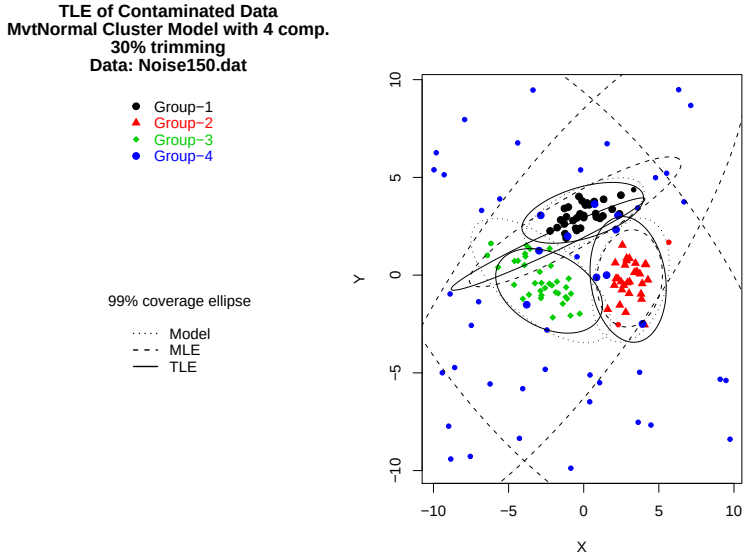

Figure 10: Four components fit of mixture of 3 bivariate normals - 30% trimming.

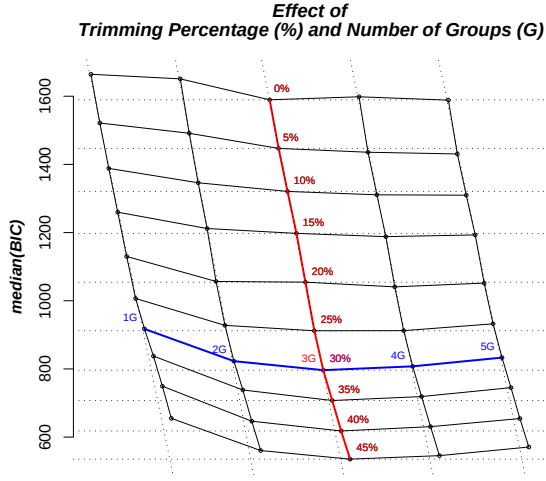

Figure 11: The TLE mixture components number and contamination level assessment.

Mixture of Poisson Regression Lines

- Mixture of two Poisson regression lines of size 40 each with expectations:

$$\begin{aligned} \text{Component 1}: \quad \log \lambda_1 &= 3 + 0.01x \quad \text{for} \quad i = 1, \ldots, 40 \\ \text{and} \qquad\qquad x &\sim U(20, 200) \\ \text{Component 2}: \quad \log \lambda_2 &= 3 - 0.01x \quad \text{for} \quad i = 41, \ldots, 80 \\ \text{and} \qquad\qquad x &\sim U(-20, -200) \end{aligned}$$

  to which 40 noise points were added

- The mixing proportions are $\pi_1 = \pi_2 = 0.5$
- To assess the quality of the fits we performed many MLE and TLE experiments starting respectively with 3, 4 and 5 mixture components over the same data set using the FlexMix program.
- A fit is successful if both components are correctly estimated.

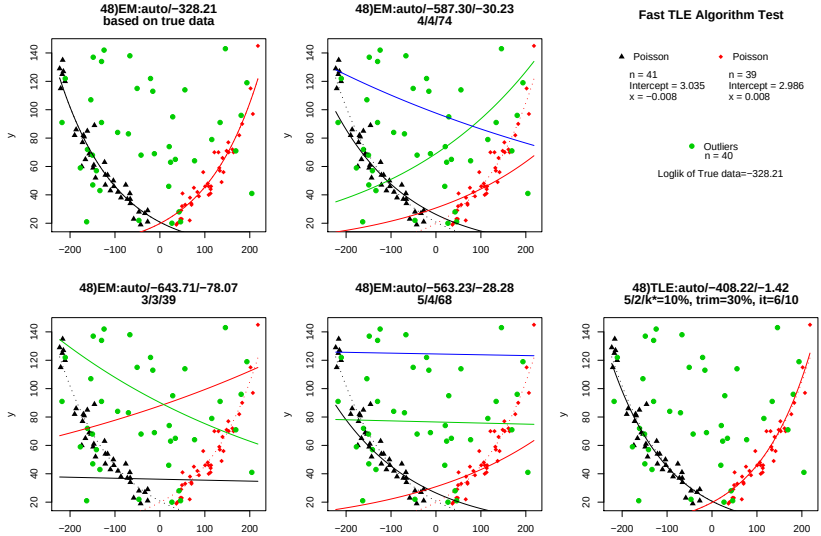

Figure 12: Mixture of 2 Poisson regressions lines plus 40 outliers.

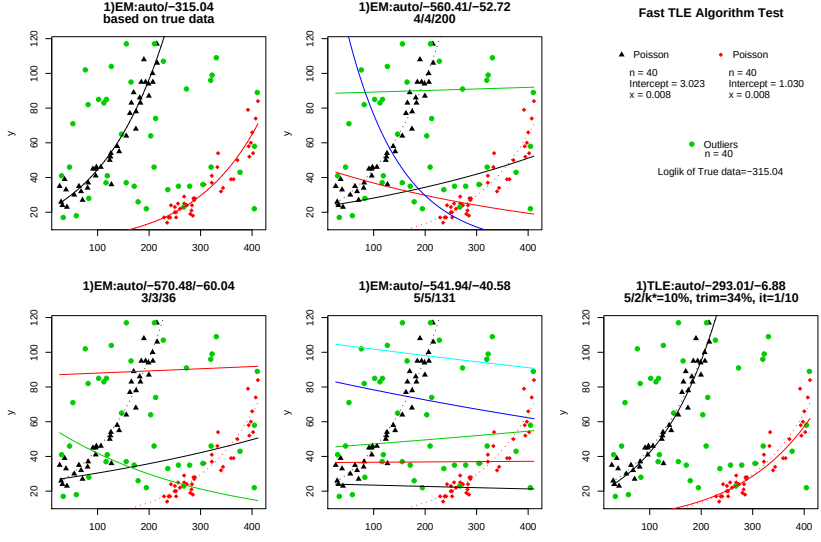

Figure 13: Mixture of 2 Poisson regressions lines plus 40 outliers.

7. Robust joint modeling through trimming

- A large class of parametric models assume that, observations y_i of the random variable Y for $i = 1, \ldots, n$, are independent and have probability distribution function $D_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ conditional on up to four distribution parameters, each of which can be a function of the explanatory variables.

- Let $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$ and $\mathbf{X_4}$ are $n \times J_k$ subsets of $\mathbf{X_{n \times p}}$ and $g_k(.)$ for $k = 1, 2, 3, 4$ be known monotonic link functions relating distribution parameters to predictors η_k as follow

$$\begin{aligned} g_1(\mu) = \eta_1 = \mathbf{X_1}\beta_\mathbf{1} \quad \text{(location)}; \qquad g_2(\sigma) = \eta_2 = \mathbf{X_2}\beta_\mathbf{2} \quad \text{(scale)} \\ g_3(\nu) = \eta_3 = \mathbf{X_3}\beta_\mathbf{3} \quad \text{(skewness)}; \qquad g_4(\tau) = \eta_4 = \mathbf{X_4}\beta_\mathbf{4} \quad \text{(kutosis)} \end{aligned} \tag{11}$$

- As the likelihood function based on $D_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ is a composite function of the link function $g_k(.)$ for $k = 1, 2, 3, 4$ and linear predictors η_k then 4(four) interlinked IRLS routines can be used to handle the computation in order to get MLE $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ of $\beta_1, \beta_2, \beta_3, \beta_4$, respectively, according to Green (1985)[16].

- Therefore, the index of fullness of the set of negative log or quasi likelihoods for the models (11) is equal to $\delta = \max[N(X_1)+1, N(X_2)+1, N(X_3)+1, N(X_3)+1]$, see **Theorem 17** for the definition of $N(X)$.

- For instance, from computational point of view, Green (1985)[16], this is equivalent to finding ML or quasi-likelihood estimates of β_1 and β_2 by solving iteratively the following two interlinked iterative weighted least squares problems:

$$\min_{\beta_1} (u_m - X_1\beta_1)^T W_m (u_m - X_1\beta_1) \tag{12}$$

$$\min_{\beta_2} (u_{d^*} - X_2\beta_2)^T W_{d^*} (u_{d^*} - X_2\beta_2), \tag{13}$$

u_m and u_{d^*} are the mean and dispersion adjusted dependent variable vectors

Theorem 19. *The BDP of this type of models equals* $\frac{1}{n}\min[n-k, k-\delta]$ *which is maximized if* k *satisfies* $\left\lfloor \frac{n+\delta}{2} \right\rfloor \leq k \leq \left\lfloor \frac{n+\delta+1}{2} \right\rfloor$.

Simulation examples

- This 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 $N(0, 1)$;

- 10% of data contamination is introduced 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. Similar results were obtained with 20% and 30% of contamination.

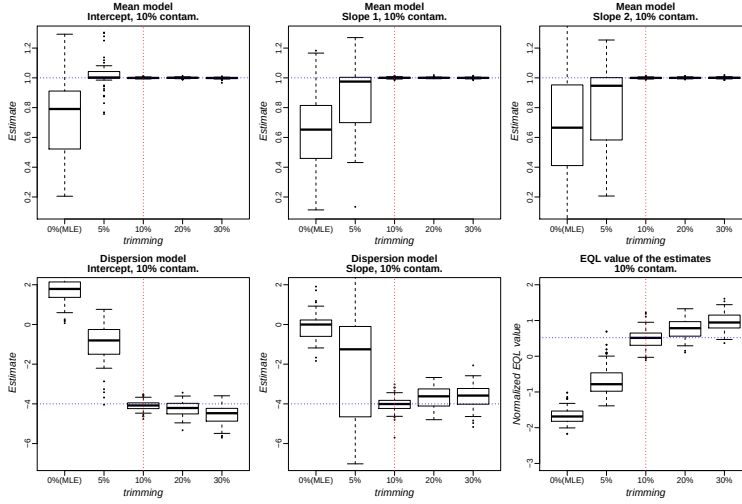

Figure 14: *Simulation experiment with 10% contamination*: boxplots of the estimates obtained from 1000 experiments for the joint normal mean and gamma dispersion GLMs. Lower right panel: boxplots for the EQL values, normalized by the sample size.

Student t distribution multiple linear regression

- The Student t density distribution is defined by

$$t_\nu(y;\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\sigma\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}.$$

- Denote by $r_i = y_i - x_i^T\beta$ the ith residuals of the multiple linear regression with Student t distribution. The score equations with respect to the first partial derivatives of β_j for $j = 1,\ldots,p$ are given by

$$\sum_{i=1}^{n}\frac{\partial \log t(y_i;x_i,\beta,\sigma,\nu)}{\partial\beta_j} = \sum_{i=1}^{n}\underbrace{\frac{\nu+1}{r_i^2+\nu\sigma^2}}_{w_i} r_i x_{ij} = \sum_{i=1}^{n} w_i r_i x_{ij} = 0$$

- It is seen that large residuals r_i induce small weights w_i, however, the estimate $\hat{\beta}$ remains vulnerable against the outliers (discordant) observations, the so called leverage points in the explanatory variables x_j for $j = 1,\ldots,p$.

- Consider t distribution heteroskedastic model with $\nu = 5$, $n = 100$, μ_i and σ_i defined by

$$\mu_i = x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = 1 + 3x_{i1} + 3x_{i2},$$
$$\log(\sigma_i^2) = z_i^T \lambda = \lambda_0 + \lambda_1 z_{i1} = 1 + 3z_{i1}$$

- The data variables, X and Z, were generated stochastically independent and identically $N(0, 1)$ on the real line.

- For each random run the responses were generated as $y_i \sim t(\mu_i, \log(\sigma_i^2), \nu)$, for $i = 1, \ldots, n$ and 24% of x_{i1}, x_{i2}, z_{i1} for $1 \leq i \leq n$ are replaced by outlers generated according to $N(3, 5)$. This means the true contamination level is 24% and $k = 100 - 24 = 76$ observations are not contaminated.

- The TLE behavior is studied with different trimming parameters $k = n * (1 - 0.05) = 95, k = n*(1-0.12) = 88, k = n*(1-0.18) = 82, k = n*(1-0.24) = 76$ and the results are compared by the MLE.

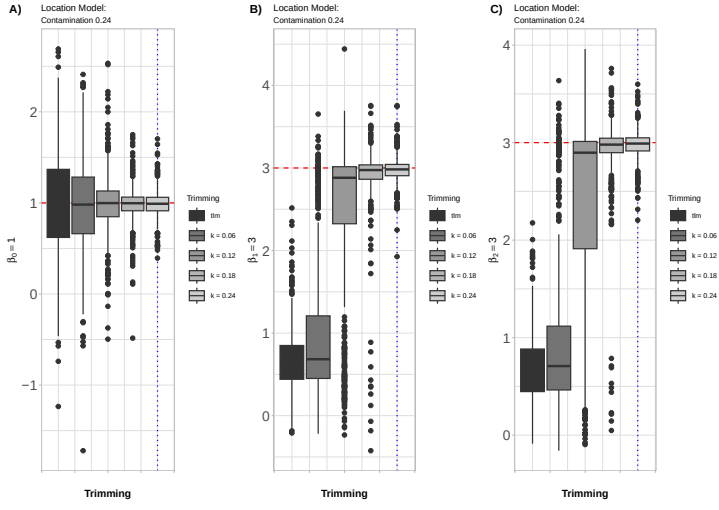

Figure 15: Simulation experiment with 24% covariates contamination of the location model. Student t distribution MLE of β_0, β_1 and β_2 are presented at left of each panel of plots whereas the TLE estimates for different trimming percentages $k = 0.05, 0.12, 0.18, 0.24$ are presented at the remaining panel of plots.

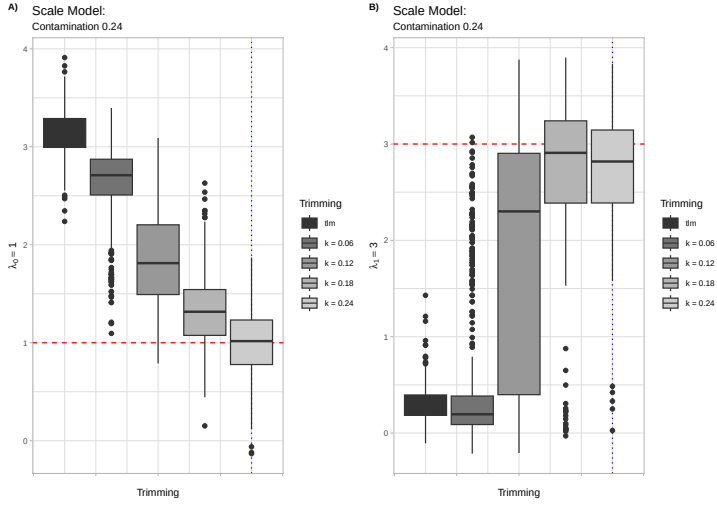

Figure 16: Simulation experiment with 24% covariate contamination of the scale model. Student t distribution MLE of λ_0, λ_1 and λ_2 are presented at left of each panel of plots whereas the TLE estimates for different trimming percentages $k = 0.05, 0.12, 0.18, 0.24$ are presented at the remaining panel of plots.

8. Trimmed Quantile Regression Estimator

Definition 20. (Koenker & Bassett (1978) [21], Koenker (2005) [19]) define the linear quantile regression (QR) estimator as any vector $\hat{\beta}_n(\tau)$ such that

$$\begin{aligned} \hat{\beta}_n(\tau) &:= \arg\min_{\beta \in R^p} \sum_{i=1}^{n} \rho_\tau(r_i(\beta)), \\ \rho_\tau(r_i(\beta)) &= \begin{cases} (\tau-1) r_i(\beta) & \text{if } r_i(\beta) < 0, \ 0 \le \tau \le 1 \\ \tau r_i(\beta) & \text{if } r_i(\beta) \ge 0, \end{cases} \end{aligned}$$

Definition 21. (Neykov et al. (2012) [29]) The Least Trimmed Quantile Regression (LTQR) estimator is defined as

$$\hat{\beta}_n^k(\tau) := \arg\min_{\beta} \min_{I \in I_k} \sum_{i \in I} \rho_\tau(r_i(\beta)), \tag{14}$$

where I_k is the set of all k–subsets of the set $\{1, \ldots, n\}$.

Theorem 22. (Neykov et al. (2012) [29]). *The BDP of the linear LTQR estimator equals* $\frac{1}{n}\min\{n-k, k-\mathcal{N}(X)-1\}$, *it is maximized for* k *such that* $\lfloor\{n+\mathcal{N}(X)+1\}/k \leq \lfloor\{n+\mathcal{N}(X)+2\}/2\rfloor$

Remark 23. The proof of this Theorem (in the Appendix of (Neykov et al. (2012) [29]) 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 $\tau \to 1$.

- The star cluster CYB OB1 dataset (Rousseeuw and Leroy (1987)[10]) consisting of 47 observations is considered. In the upper left corners of the plots of Figure 17 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 (2005)[19] and by the LTQR estimator using 4% and 9% trimming percentages.

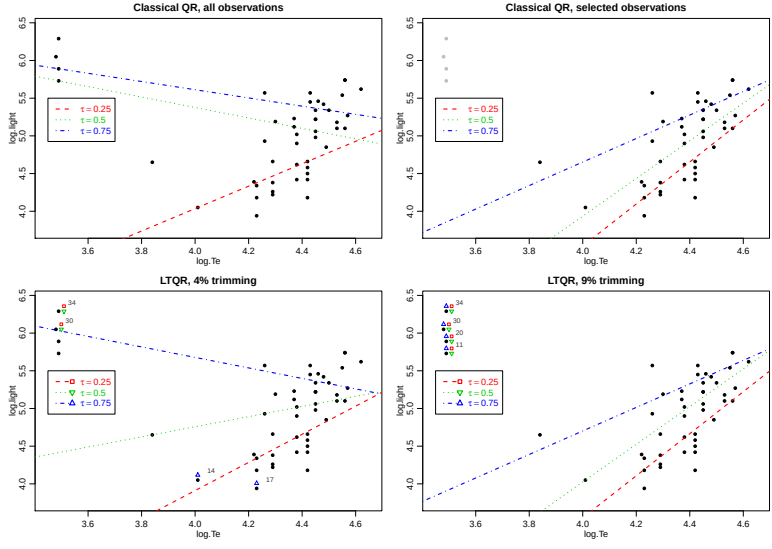

Figure 17: Star data: 0.25, 0.50, and 0.75 regression quantiles from Koenker and Bassett estimate based on whole data (upper left plot); selected cases (upper right plot); LTQR fits with 4% and 9% of trimming.

The END

Hawkins and Olive (2002) [18]:

"Because statistical analysis is generally just a small part of the effort and cost of any data gathering and analysis, one should not make too much of this computational load. We consider it clearly far better to use an analysis that takes 10 hours but finds all outliers than one that takes 10 seconds yet misses most outliers."

Thank you for your attention

References

[1] Atkinson, A. C. and Riani, M. 2000. *Robust diagnostic regression analysis.* Springer, NY.

[2] Atkinson, A. C. Riani, M. and Cerioli,A. 2004. *Exploring Multivariate Data with the Forward Search.* Springer, NY.

[3] Atkinson, A. C. Riani, M. , Corbellini, A., Perrotta, D. and Todorov, V. 2025. Robust Statistics Through the Monitoring Approach: Applications in Regression. Springer, NY.

[4] Farcomeni, A. and Greco,L. 2015. *Robust methods for data reduction.* CRC Press, New York.

[5] 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.

[6] Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.-P. 2009. *Robust methods in biostatistics*. Wiley, Chichester, U.K.

[7] Huber, P. 1981. *Robust statistics*. John Wiley & Sons, New York.

[8] Huber, P. and Ronchetti, E. 2009. *Robust statistics*. John Wiley & Sons, New York.

[9] Maronna,R. A., Martin, R. D. and Yohai, V. J., Salibian-Barrera, M. 2019. *Robust Statistics: Theory and Methods with R*, John Wiley and Sons, New York.

[10] Rousseeuw, P. J. and Leroy, A. M. 1987. *Robust regression and outlier detection*. Wiley, New York.

[11] P. Cizek (2008a) *Robust and efficient adaptive estimation of binary-choice regression models*. J. Amer.Statist.Assoc., **103(482)**, 687 – 696.

[12] P. Cizek (2008b). *General trimmed estimation: Robust approach to nonlinear and limited dependent variable models*. Econometric Theory, **24(6)**, 1500 – 1529.

[13] P. Cizek (2006a). *Least trimmed squares in nonlinear regression under dependence*. J. Statist. Plann. Inference, **136(11)**, 3967 – 3988.

[14] P. Cizek (2006b) Trimmed likelihood-based estimation in binary regression models. Austrian J. of Statist., **25(2-3)**, 223 – 232.

[15] Green P.J. (1984). *Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives*. J. Roy. Statist. Soc. Ser. B **46**, 149 – 192.

[16] Green, P.J. (1989). Generalized Linear Models and Some Extensions: Geometry and Algorithms. In: Decarli, A., Francis, B.J., Gilchrist, R., Seeber, G.U.H. (eds) Statistical Modelling. Lecture Notes in Statistics, vol 57. Springer, New York, pp. 26 – 36.

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

[18] 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.

[19] Koenker, R.W., 2005a. Quantile Regression. Cambridge University Press, Cambridge.

[20] Koenker, R.W., 2005b. Quantile Regression in R. http://cran.R-project.org/doc/packages/quantreg/quantreg.pdf

[21] Koenker, R.W. and Bassett, G. Jr., 1978. Regression quantiles. Econometrica 84, 33–50.

[22] McLachlan G.J. and Peel D. (2000). *Finite mixture models*. Wiley, NY.

[23] 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.

[24] Neykov N.M. and Müller C.H. (2003). *Breakdown point and computation of trimmed likelihood estimators in generalized linear models*. In: Developments in robust statistics, R. Dutter, P. Filzmoser, U. Gather and P. Rousseeuw, (eds.), Physica-Verlag, Heidelberg, 277 – 286.

[25] Neykov N.M. and Neytchev P.N. (1990). *A robust alternative of the maximum likelihood estimator*. In: Short communications of COMPSTAT'90, Dubrovnik, 99 – 100.

[26] Neykov, N.M., Filzmoser, P., Dimova, R. and Neytchev, P.N. (2004). *Mixture of Generalized Linear Models and the Trimmed Likelihood Methodology*. In: Proceedings in Computational Statistics, J. Antoch (ed.), Physica-Verlag, 1585 – 1592.

[27] Neykov, N. M., Filzmoser, P., Dimova, R. and Neytchev, P. N. (2007). Robust fitting of mixtures using the Trimmed Likelihood Estimator. Computational Statistics and Data Analysis, http://dx.doi.org/10.1016/j.csda.2006.12.024

[28] 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.

[29] Neykov, N.M., Čížek, P., Filzmoser,P. and Neytchev, P.N. (2012). *The least trimmed quantile regression.* Comput. Statist. Data Anal., **56**, 1757–1770.

[30] Neykov, N. M., Filzmoser, P. and Neytchev, P. N. (2014). *Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator.* Statistical Papers, **55**, 187–207.

[31] Rousseeuw, P. J. *Least median of squares regression.* J. Amer. Statist. Assoc. **79,** (1984) 851 – 857.

[32] Rousseeuw P.J. and Van Driessen K. (1999) *Computing LTS regression for large data sets.* Technical report, University of Antwerp, (submitted).

[33] Rousseeuw P.J. and Van Driessen K. (1999). *A fast algorithm for the MCD estimator.* Technometrics. **41**, 212 – 223.

[34] Vandev D.L. (1993). *A note on breakdown point of the least median squares and least trimmed squares*. Statistics and Probability Letters **16**, 117 – 119.

[35] Vandev D.L. and Neykov N.M. (1993). *Robust maximum likelihood in the Gaussian case*. In: New directions in data analysis and robustness, S. Morgenthaler, E. Ronchetti and W.A. Stahel, (eds.), (Birkhäuser Verlag, Basel, 259 – 264.

[36] Vandev D.L. and Neykov N.M. (1998). *About regression estimators with high breakdown point*. Statistics. **32**, 111 – 129.

[37] Varmuza, K. and Filzmoser, P. 2008. *Introduction to multivariate statistical analysis in chemometrics*. CRC press, New York.