Robust Methods in Industrial Statistics

Dimitar Vandev

University of Sofia, Faculty of Mathematics and Informatics,
Sofia 1164, J.Bourchier 5,
E-mail: vandev@fmi.uni-sofia.bg

Abstract

The talk aims to make a short overview of the most popular methods of the robust statistics and to outline there place in the applications. Roughly speaking, robustness means stability of statistical inference under the variations of the accepted distribution models.

The basic reasons of research in this field are of a general mathematical character. 'Optimality' and 'stability' are the mutually complementary characteristics of any statistical procedure. It is a wellknown fact that the behavior of many optimal decisions is rather sensible to 'small deviations' from prior assumptions. In statistics, the remarkable example of such unstable optimal procedure is given by the least squares method: its performance may become extremely poor under small deviations from normality.

The field of industrial statistics refers to the problems of data processing in industry (electromechanical and energetic), economics and finances (financial mathematics), defense (detection of air targets), medicine (cardiology, pharmacokinetics), food technology (wine industry e.g.). We are going to indicate the place of robust methods in these problems.

A large part of this talk is based on the recent book of Shevlyakov and Vilchevski (2002).

1 General remarks

The field of mathematical statistics called robust statistics appeared due to the pioneer works of Tukey (1960), Huber (1964), and Hampel (1968); it has been intensively developed since the sixties and is rather definitely formed by present. The term 'robust' (strong, sturdy) as applied to statistical procedures was proposed by Box (1953).

1.1 The forms of data representation.

We begin with the customary forms of data representation:

(i) as a sample $\{x_1, \ldots, x_n\}$ of real numbers $x_i \in R$ being the easiest form to handle;

(ii) as a sample$\{x_1, \ldots, x_n\}$ of realvalued vectors $x_i \in R^m$;

(iii) as a realization $x(t), t \in [0, T]$ of a realvalued continuous process (function);

(iv) as a sample of 'nonnumerical nature' data representing qualitative variables;

(v) as semantic type data (statements, texts, pictures, etc.).

1.2 Types of statistical characteristics

The experience of treating various statistical problems shows that nearly all of them are solved with the use of only a few qualitatively different types of data statistical characteristics.

These characteristics may be classified as follows:

- the measures of location (central tendency, mean values);
- the measures of spread (dispersion, scale, scatter);

- the measures of interdependence (association, correlation);
- the characteristics of extreme values;
- the characteristics of data distributions or the measures of shape.

1.3 The main aims of Industrial Statistics

These aims may be formulated as follows:

(A1) compact representation of data,

(A2) estimation of model parameters describing mass phenomena,

(A3) prediction and optimization.

A human mind cannot efficiently work with large volumes of information, since there exist natural psychological bounds of perception. Thus it is necessary to provide a compact data output of information: only in this case we may expect a satisfactory final decision. Note that data processing often begins and ends with this first item (A1).

The next step (A2) is to suggest an explanatory underlying model for the observed data and phenomena. It may be a regression model, or a distribution model, or any other, desirably a simple one: an essentially multiparametric model is usually a bad model. Parametric models refer to the first to be considered and examined.

Finally, all previous aims are only the steps to the last (A3): here we have to state that this aim remains a main challenge to statistics and to science as a whole.

2 Huber minimax approach

The convincing arguments for robust statistics are given in Tukey (1960); Huber (1981); Hampel et al. (1986). Here we only recall that the classical examples of robust and nonrobust estimators of location are given by the sample median and sample mean, respectively.

As it was said above, robust statistics deal with the consequences of possible deviations from the assumed statistical model and suggests the methods protecting statistical procedures against such deviations. Thus the statistical models used in robust statistics are chosen so that to account possible violations of the assumptions about the underlying distribution. For description of these violations, the concrete forms of neighborhoods of the underlying model are formed with the use of an appropriately chosen metric, for example, the Kolmogorov, Prokhorov, or Lévy Hampel et al. (1986); Huber (1981)). Hence the initial model (basic or ideal) is enlarged up to the socalled supermodel that describes both the ideal model and the deviations from it.

Defining a robust procedure, it is useful to answer three main questions:

- Robustness of what? Here we one defines the type of a statistical procedure (point or interval estimation, hypotheses testing, etc.);
- Robustness against what? Here one specifies the supermodel;
- Robustness in what sense? Here the criterion of quality of a statistical procedure and some related requirements towards its behavior are considered.

The wide spectrum of the problems observed in robust statistics can be explained by the fact that there exists a variety of answers to each of the above questions.

According Bickel (1976) the main supermodels in robust statistics are of two types: local and global.

A local type suggests setting an ideal (basic) model, and then the related supermodel is defined as a neighborhood of this ideal model. A global supermodel represents some class $\mathcal{F}$ of distributions with given properties that also comprises an ideal model.

For example, Hodges and Lehmann (1963) consider the supermodel in the form of all absolutely continuous symmetric distributions.

Birnbaum and Laska (1967) propose the supermodel as a finite collection of distribution functions: $\mathcal{F} = \{F_1, F_2, \ldots, F_k\}$.

Andrews et al. (1972) examine estimators in the supermodels containing distributions with heavier tails than the normal. In particular, they use the Tukey supermodel based on the quantile function, the inverse to the distribution function. This supermodel comprises rather accurate approximations to the normal, Laplace, logistic, Cauchy, and Student distributions.

Various supermodels are used to study deviations from normality: the family of powerexponential distributions with the normal, Laplace, and uniform distributions as particular cases; the family of the Student tdistributions with the normal and Cauchy distributions; also the influence of nonnormality can be studied with the use of the measures of asymmetry and kurtosis, the positive values of the latter indicate gross errors and heavy tails.

For describing gross errors and outliers, the most popular is the Tukey (1960) supermodel based on the Gaussean low:

$$\mathcal{F} = \left\{ F : F(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(\frac{x-\theta}{k}), \quad 0 < \varepsilon < 1, 1 < k) \right\}. \qquad (1)$$

Huber (1964) considered more general model

$$\mathcal{F} = \left\{ F : F(x) = (1-\varepsilon)F_0(x) + \varepsilon H(x) \right\}, \qquad (2)$$

where F_0 is some given distribution (the ideal model) and $H(x)$ is an arbitrary continuous distribution.

2.1 M-estimators of location

The first general approach to robust estimation is based on the minimax principle (Huber, 1964; Huber, 1972; Huber, 1981). The minimax approach aims at the least favorable situation for which it suggests the best solution. Thus, in some sense, this approach provides a guaranteed result, perhaps too pessimistic. However, being applied to the problem of estimation of the location parameter, it yields a robust modification of the principle of maximum likelihood.

Let $x_1, \ldots, x_n$ be a random sample from a distribution F with density $f(x-\theta)$ in a convex class $\mathcal{F}$, where θ is the location parameter. Assume that F is a symmetric unimodal distribution, hence θ is the center of symmetry to be estimated. Then the M-estimator $\widehat{\theta}_n$ of the location parameter is defined as some solution of the following minimization problem

$$\widehat{\theta}_n = \mathbf{argmin}_{\theta} \sum_{i=1}^{n} \rho(x_i - \theta), \tag{3}$$

where $\rho(u)$ is an even non-negative function called the contrast function; $\rho(x_i-\theta)$ is the measure of discrepancy between the observation x_i and the center θ.

- Choosing $\rho(u) = u^2$, we have the least squares (LS) method with the sample mean $\overline{x}_n$ as an estimator;
- for $\rho(u) = |u|$, we have the least absolute values (LAV) method with the sample median as the estimator;
- most important, for a given density $f(x)$, the choice $\rho(u) = -\log f(u)$ yields the maximum likelihood estimator (MLE).

It is convenient to formulate the properties of M-estimators in terms of the derivative of the contrast function $\psi(u) = \rho'(u)$ called the score function. In this case, the M-estimator $\widehat{\theta}_n$ is defined as a solution of the following implicit equation

$$\sum_{i=1}^{n} \psi(x_i - \theta) = 0. \tag{4}$$

Under rather general regularity conditions imposed on the class of score functions Ψ and on the related class of densities $\mathcal{F}$, M - estimators are consistent, asymptotically normal with the asymptotic variance

$$V_M(\psi, f) \stackrel{def}{=} \mathbf{D}\,(n^{1/2}\widehat{\theta}_n) = \frac{\mathbf{E}_F \psi^2}{(\mathbf{E}_F \psi')^2} = \frac{\int \psi^2 dF}{(\int \psi' dF)^2}. \tag{5}$$

2.2 Huber minimax property

The following regularity conditions defining the classes $\mathcal{F}$ are sufficient (for details Hampel et al. (1986), pp.125 - 127)):
$\mathcal{F}1$: f is twice continuously differentiable and satisfies $f(x) > 0$ for all $x \in R$,
$\mathcal{F}2$: the Fisher information for location satisfies $0 < I(f) < \infty$.

Let f^* be the least favorable density in $\mathcal{F}$:

$$f^* = \underset{f \in \mathcal{F}}{\mathbf{argmin}}\, I(f), \quad I(f) = \int \left[\frac{f'(x)}{f(x)}\right]^2 f(x)dx. \tag{6}$$

Then the optimal contrast function and score function are calculated by maximum likelihood method for the least favorable density:

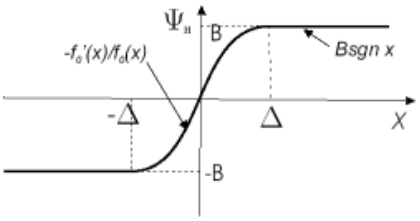
$$\rho^* = -\log f^*, \quad \psi^* = f^{*\prime}/f^*$$

Then the Huber minimax property is

$$V_M(\psi^*, f) \le V_M(\psi^*, f^*) = \sup_{f \in \mathcal{F}} \inf_{\psi \in \Psi} V_M(\psi, f). \tag{7}$$

Thus the main problem is to solve (6) for different classes $\mathcal{F}$.

For the mixture class of Huber (2) if $h(x)$ satisfy conditions ($\mathcal{F}1$) and ($\mathcal{F}2$) along with the additional logconvexity condition we have:



$$\psi^*(x) = \begin{cases} -f_0'(x)/f_0(x), & |x| \le \Delta \\ B\,\mathbf{sgn}(x), & \Delta < |x|. \end{cases}$$

$$f^*(x) = \begin{cases} (1-\varepsilon)f_0(x), & |x| \le \Delta \\ A\exp(-Bx), & \Delta < |x|. \end{cases}$$

Figure 1: Score Function

2.3 L-estimators of location

L-estimators were proposed by Daniel (1920) and forgotten for 30 years. The linear combinations of order statistics (L-estimators) are defined as

$$\widehat{\theta}_n = \sum_{i=1}^{n} C_i x_{(i)}, \tag{8}$$

where $x_{(i)}$ is the i-th order statistic. The normalization condition in 8 provides equivariancy of L-estimators under translation. The trimmed mean:

$$\overline{x}_{tr}(k) = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)} \tag{9}$$

and the Winsorized mean:

$$\overline{x}_W(k) = \frac{1}{n}(kx_{(k)} + \sum_{i=k+1}^{n-k} x_{(i)} + kx_{(n-k+1)}) \tag{10}$$

are typical representatives of this class.

The L-estimators may be easily represented in the form:

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} h(\frac{i}{n+1}) x_{(i)}, \tag{11}$$

where the function h is a function of bounded variation on $[0,1]$, $h(t) = h(1-t)$ and $\int_0^1 h(t)dt = 1$. These conditions on h along with the regularity conditions ($\mathcal{F}1$) and ($\mathcal{F}2$) on the distribution provide consistency and asymptotic normality of L-estimators (8) with asymptotic variance

$$V_L(h,f) \stackrel{def}{=} \mathbf{D}\,(n^{1/2}\widehat{\theta}_n) = \int_0^1 K^2(t)dt, \tag{12}$$

where

$$K(t) = (h(t)F^{-1}(t) - \theta), \quad \theta = \int_0^1 h(t)F^{-1}(t)dt$$

2.4 R-estimators of location

R-estimators proposed in Hodges and Lehmann (1963) are based on rank tests. There are several methods of their construction. Let $y_1, \ldots, y_n$ and $z_1, \ldots, z_n$ be independent samples from the distributions $F(x)$ and $F(x-\theta)$ respectively. For testing the hypothesis $\theta = 0$ against the alternative $\theta > 0$ the following statistic is used:

$$W_n(y_1, \ldots, y_n, z_1, \ldots, z_n) = \sum_{i=1}^{n} J(\frac{s_i}{2n+1}) \tag{13}$$

where s_i is the rank of $y_i, i = 1, \ldots n$, in the united sample of size $2n$.

Let $J(t)$, $0 \leq t \leq 1$, satisfy the following conditions:

1. $J(t)$ is increasing;
2. $J(t) + J(1-t) = 0$ for all $t \in [0, 1]$;
3. the functions J' and $f(F^{-1})$ are of bound variation on $[0, 1]$,
4. $\int_0^1 J'(t)f(F^{-1}(t))dt \neq 0$.

Under these conditions, Hájek and Šidák (1967) proved that the test with the critical region $W_n > c$ has certain optimal in power properties. The R-estimator $\widehat{\theta}_n$ based on this test is defined as a solution of the equation:

$$W_n(x_1 - \theta, \ldots, x_n - \theta, -(x_1 - \theta), \ldots, -(x_n - \theta)) = 0 \tag{14}$$

Under the above conditions, $\widehat{\theta}_n$ is consistent and asymptotically normal with asymptotic variance

$$V_R(J, F) \stackrel{def}{=} \mathbf{D}\,(n^{1/2}\widehat{\theta}_n) = \frac{\int_0^1 J^2(t)dt}{(\int J'(F(x))f^2(x)dx)^2}. \tag{15}$$

For any fixed function $F(x)$, it is possible to find the function $J(t)$ minimizing asymptotic variance $V_R(J, F)$. The test based on such function $J(t)$ also has optimal properties for this F. In particular, the logistic distribution $F(x) = (1 + e^{-x})^{-1}$ produces the well known Wilcoxon test. The corresponding estimator of location is the Hodges-Lehmann median:

$$\widehat{\theta}_n = \mathbf{med}\left\{\frac{x_{(i)} + x_{(k)}}{2},\ 1 \le i < k \le n\right\}. \tag{16}$$

2.5 Applications of measures of location

Probably the most important place is the statistical quality control. Stromberg et al. (1998) developed Control Charts for the Median and Iinterquartile Range.

Römisch et al. (2001) tested all kinds of estimators for Determination of the Geographical Origin of Wines from East European Countries.

3 Hampel approach

The main advantage of robust methods is their lower sensitivity to possible variations of data statistical characteristics. Thus it is necessary to have specific mathematical tools allowing to analyze the sensitivity of estimators to outliers, roundingoff errors, etc. On the other hand, such tools make it possible to solve the inverse problem: to design estimators with the required sensitivity. Now we introduce the abovementioned apparatus, namely the sensitivity curves and the influence functions.

3.1 The sensitivity curve

Let $\{T_n\}$ be a sequence of statistics. Let $T_n(X)$ denote the statistic from $\{T_n\}$ on the sample $X = (x_1, \ldots, x_n)$, and let $T_{n+1}(x, X)$ denote the same statistic on the

sample $(x_1, \ldots, x_n, x)$. Then the function

$$SC_n(x; T_n, X) = (n+1)[T_{n+1}(x, X) - T_n(X)] \tag{17}$$

is called the sensitivity curve for this statistic Tukey (1977). In particular,

$\overline{x}_n$	$SC_n(x; \overline{x}_n, X) = x - \overline{x}_n$
$\mathbf{med}(X)$ $(n = 2k+1)$	$SC_n(x; \mathbf{med}(X), X) = \begin{cases} (n+1)(x_{(k)} - x_{(k+1)})/2, & x \le x_{(k)} \\ (n+1)(x - x_{(k+1)})/2, & x_{(k)} \le x \le x_{(k+2)} \\ (n+1)(x_{(k+2)} - x_{(k+1)})/2, & x_{(k+2)} \le x \end{cases}$
$\overline{x}_{tr}(1)$	$SC_n(x; \overline{x}_{tr}(1), X) = \begin{cases} x_{(1)}, & x \le x_{(1)} \\ x, & x_{(1)} \le x \le x_{(n)} \\ x_{(n)}, & x_{(n)} \le x \end{cases}$

Table 1: Example sensitivity curves

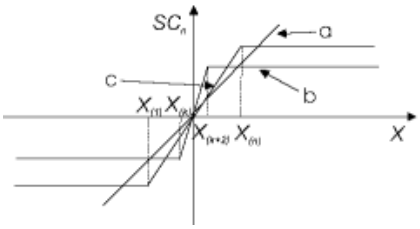


Figure 2: Sensitivity curves

We can see that the sensitivity curve (a.) of the sample mean is unbounded, hence only one extreme observation can completely destroy the estimator. In addition, the maximal error of the trimmed mean (curve c.) is of order $(x_{(n)} - x_{(1)})/n$, while this of median – of $(x_{(k+2)} - x_{(k)})$.

3.2 The influence function

Let F be a fixed distribution and $T(F)$ be a functional defined on some set $\mathcal{F}$ of distributions satisfying conditions ($\mathcal{F}1$) and ($\mathcal{F}2$). Let the estimator T_n be constructed in the form $T_n = T(F_n)$. Then we define the influence function as:

$$IF(x, T, F) = \lim_{t \to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t}. \tag{18}$$

$\overline{x}_n$	$IF(x,T,F) = x - T(F) = x - \int x dF(x)$
$\mathbf{med}(X)$	$T(F) = F^{-1}(1/2), \quad IF(x,T,F) = \mathbf{sgn}(x)/(2f(0))$
$\overline{x}_{tr}(k)$ $\alpha = k/n$	$IF(x,T,F) = \begin{cases} F^{-1}(\alpha)/(1-2\alpha), x \le F^{-1}(\alpha) \\ x/(1-2\alpha), \quad F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha) \\ F^{-1}(\alpha)/(1-2\alpha), \quad F^{-1}(1-\alpha) \le x \end{cases}$

Table 2: Example influence functions

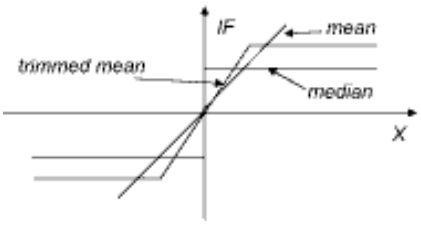


Figure 3: Influence functions

Comparing Fig.2 and Fig.3, we see that the forms of influence and sensitivity curves are similar. In fact $SC_n(x;T,F) \to IF(x;T,F)$ as $n \to \infty$.

The influence function for the M-estimator with the score function ψ is of the form Hampel et al. (1986)

$$IF(x;\psi,F) = \frac{\psi}{\int \psi dF(x)}$$

Fernholz (1983) showed that T_n is asymptotically normal with asymptotic variance

$$V(T,F) = \int IF^2(x;T,F)dF(x). \tag{19}$$

4 Measures of robustness

4.1 Local measures of sensitivity

From the influence function, the following robustness measures can be defined (Hampel (1968); Hampel (1974).

Gross-error sensitivity

$$\gamma^*(T,F) = \sup_x |IF(x;T,F)|$$

is an upper bound to the asymptotic bias of the estimator and measures the worst influence of an infinitesimal contamination. The estimators T having finite $\gamma^*(T,F)$ are called B-robust.

Local-shift sensitivity

$$\lambda^*(T,F) = \sup_{x \neq y} \frac{|IF(y;T,F) - IF(x;T,F)|}{|y-x|}$$

accounts the effects of rounding-off and grouping of the observations.

Rejection point

$$\rho^*(T,F) = \inf_{r>0}\{r : IF(x;T,F) = 0, \forall |x| > r\}$$

defines the observations to be rejected completely.

Change-of-variance function

$$CVF(x;T,F) = \lim_{t \to 0} \frac{V(T,(1-t)F + t\delta_x) - V(T,F)}{t}$$

was introduced by Hampel et al. (1986) by analogy with the influence function IF. Here $V(T,F)$ is the asymptotic variance.

Change-of-variance sensitivity

$$k^*(T,F) = \sup_x \frac{CVF(x;F,T)}{V(T,F)}$$

The estimator $T_n = T(F_n)$ of the functional $T(F)$ is called V-robust if $k^*(T,F) < \infty$.

All the above-introduced measures of robustness based on the influence function and its derivatives are of a local character being evaluated at the model distribution F. Hence it is desirable to have a measure of the global robustness of

the estimator over the chosen class of distributions, in other words, in the chosen supermodel $\mathcal{F}$.

4.2 Break-down point

Since the general definition of a supermodel is based on the concept of a distance in the space of all distributions, the same concept is involved into the construction for a measure of the global robustness. Let d be such a distance. Then the break-down point of the estimator $T_n = T(F_n)$ for the functional $T(F)$ at $\mathcal{F}$ is defined by

$$\varepsilon^*(T, \mathcal{F}) = \sup_{\varepsilon<1}\{\varepsilon : \sup_{F:d(F,F_0)<\varepsilon} |T(F) - T(F_0)| < \infty\}.$$

The breakdown point characterizes the maximal deviation (in the sense of a metric chosen) from the ideal model F_0 that provides the boundedness of the estimator bias.

Breakdown point as applied to the Huber supermodel

$$\varepsilon^*(T, \mathcal{F}) = \sup_{\varepsilon<1}\{\varepsilon : \sup_{F:F=(1-\varepsilon)F_0+\varepsilon H} |T(F) - T(F_0)| < \infty\}. \tag{20}$$

This notion defines the largest fraction of gross errors that still keeps the bias bounded. Here is the replacement variant of the finite sample breakdown point given by Hampel et al. (1986).

Let $\Omega = \{\omega_i \in R^p$, for $i = 1, \ldots, n\}$ be a sample of size n. The breakdown point of an estimator $T(\Omega) \in R^q$ is given by

$$\varepsilon_n^*(T) = \frac{1}{n}\max\{m : \sup_{\tilde{\Omega}_m} \|T(\tilde{\Omega}_m)\| < \infty\}, \tag{21}$$

where $\tilde{\Omega}_m$ is any sample obtained from Ω by replacing any m of the points in Ω by arbitrary values.In other words there should exist a compact set such that the estimator T remains in it even if we replace any m elements of the sample Ω by arbitrary ones. The largest m/n for which this property holds is the breakdown

point.

5 Multidimensional Statistics

All the definitions and methods can be easily extended to multivariate and multiparametric case when one estimates location.

5.1 LTS and LMS

The multiple regression is probably most used statistical procedure in the industrial statistics. Consider the model

$$y_i = x_i^T \beta + \varepsilon_i.$$

where y_i is an observed response, x_i is a $p \times 1$–dimensional vector of explanatory variables and β is a $p \times 1$ vector of unknown parameters. Classically ε_i, $i = 1, \ldots, n$ are assumed to be i.i.d. $N(0, \sigma^2)$, for some $\sigma^2 > 0$.

The LMS (Least Median of Squares) and LTS (Least Trimmed Squares) estimators were proposed by Rousseeuw (1984) as robust alternatives of the LSE

$$\text{LMS}(r_1, \ldots, r_n) = \operatorname*{argmin}_\theta \operatorname{med}\{r_i^2, i = 1, ..., n\}, \tag{22}$$

$$\text{LTS}(k)(r_1, \ldots, r_n) = \operatorname*{argmin}_\theta \sum_{i=1}^{k} r^2_{\nu(i,\theta)}. \tag{23}$$

Here $\nu(i, \theta)$ is a permutation of the indices, such that $r^2_{\nu(i,\theta)} \leq r^2_{\nu(i+1,\theta)}$. Thus the idea was to minimize the sum of squares using "smallest residuals" only.

5.2 Covariance

The estimation of unknown covariance matrix of observed data or estimated parameters turned out to be not so easy. In fact, only two methods are used in

practice and very little is known about their properties. These are

- MVE - minimum volume ellipsoid;
- MCD - minimum covariance determinant.

In both cases one have to choose fixed percentage (e.g. 90%) of observed data having corresponding optimal property. Then the estimator is build using only these data.

5.3 Applications

We should mention here the works of Mili et al. (1991)and Mili et al. (1994) on Power Systems. The abstract of the second paper:

The exact fit points of the Least Median of Squares (LMS) and the Least Trimmed Squares (LTS) estimators in electric power systems are investigated. The expression of the maximum possible exact fit point is derived, and the corresponding quantile index of the ordered squared residual is determined. It is found that these values hinge on the surplus of the network, defined as one less than the smallest number of measurements whose deletion from the data set decreases the rank of the Jacobian matrix. Based on the surplus concept, a system decomposition scheme is developed; it significantly increases the number of outliers that can be handled by the LMS and the LTS estimators. In addition, it dramatically reduces the computing time of these estimators, opening the door to their application in a real-time environment, even for large-scale systems.

6 Robustified Maximum likelihood

Neykov and Neytchev (1990) proposed to replace in these estimators (LMS and LTS) the squared residuals with - log likelihood´s of the individual observations and thus to obtain robustified likelihood.

Let the observations $x_1, x_2, \ldots, x_n$ be generated by an arbitrary probability density function $\psi(x, \theta)$ with unknown vector parameter θ.

$$\text{LME}(k) = \underset{\theta}{\operatorname{argmin}} \{-\log \psi(x_{\nu(k,\theta)}, \theta)\}, \tag{24}$$

$$\text{LTE}(k) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{k} \{-\log \psi(x_{\nu(i,\theta)}, \theta)\}. \tag{25}$$

Thus the idea was to maximize the likelihood over the best k observations (with "largest likelihood").

Both estimators may be easily combined into one. However it took some time (5 years) to understand.

6.1 Weighted Trimmed Likelihood (WTL)

WTL estimators were introduced independently by Hadi and Luceño (1997) and Vandev and Neykov (1998).

Let the observations $x_1, x_2, \ldots, x_n$ be generated by an arbitrary probability density function $f(x, \theta)$ with unknown vector parameter θ. Let the weights w_i for $i = 1, \ldots, n$ be fixed nonnegative numbers.

$$\text{WTL(k)}(x_1, \ldots, x_n) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{k} \{-w_i \log f(x_{\nu(i,\theta)}, \theta)\} \tag{26}$$

where $f(x_{\nu(i,\theta)}, \theta) \geq f(x_{\nu(i+1,\theta)}, \theta)$ are the ordered density values. ν is a permutation of the indices $1, \ldots, n$, which may depend on θ.

- The LME(k) estimator is obtained if $w_i = \delta_{i,k}$,
- the LTE(k) is obtained if $w_i = 1$ for $i = 1, \ldots, n$,
- the maximum likelihood estimator – if $k = n$.

6.2 Definition of d-fullness

In order to study the breakdown properties of general estimators of the type (24) and (25) Vandev (1993) developed a d–fullness technique. He proved that their breakdown point is not less than $(n-k)/n$ if k is within the range of values $(n+d)/2 \leq k \leq (n-d)$ for some constant d which depends upon the density considered.

Now we present a simple generalization of this result for the case of WTL estimators (26). First the definition

Definition *A finite set F of n functions is called d–full, if for each subset of cardinality d of F, the supremum of this subset is a subcompact function.*

We remind that a real valued function $g(\theta)$ defined on a topological space Θ is called subcompact, if its Lebesque sets $\{\theta : g(\theta) \leq C\}$ are compact (or empty) for any constant C.

Let the finite set $F = \{f_i(\theta) \geq 0, i = 1, \ldots, n, \text{ for } \theta \in \Theta\}$ be d–full and Θ is a topological space. Consider the estimator of θ

$$R(k) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{k} w_i f_{\nu(i,\theta)}(\theta).$$

Here $f_{\nu(i,\theta)}(\theta) \leq f_{\nu(i+1,\theta)}(\theta)$ are the ordered values of f_i at θ. The weights $w_i \geq 0, w_k = 1$. From a statistical point of view $R(k)$ can be considered as a set of estimates if the functions $f_i(\theta)$ are appropriately chosen, e.g. depend on the observations.

Theorem. *Under these conditions if $n \geq 3d$ and $(n+d)/2 \leq k \leq n-d$, then the breakdown point of the estimator $R(k)$ is not less than $(n-k)/n$.*

Thus if one knows the value of d for the set $\{f_i(\theta)\}$, one easily make conclusions about the conditions on k to have appropriate BP.

The value d may be interpreted as
number of observations necessary to make unique guess for the estimated parameter.

6.3 Multivariate normal distribution

Vandev and Neykov (1993) determined the value of d for the set of log-density functions for the multivariate normal case. When estimating only the mean $d = 1$. When one need to estimate the covariance matrix $d = p + 1$. Let $x_i \in R^p,\ i = 1, \ldots, n$ have density

$$\phi(x, \mu, S) = (2\pi)^{-p/2}(\det(S))^{-1/2} \exp(-(x-\mu)'S^{-1}(x-\mu)/2).$$

Theorem.*If $n \geq d$ and $(n+d)/2 \leq k \leq n-d$, then the breakdown point of the $WTL(k)$ of μ and S is equal to $(n-k)/n$.*

Later Marincheva and Vandev (1995) considered a general elliptic family. Atanasov and Neykov (2001) calculated the fullness parameters for the Lognormal, Poisson, Gamma, Geometric and Logarithmic series distributions and thus determined the BPs of the WTL estimators for the corresponding Generalized Linear Models.

7 Linear regression

7.1 Theory

Consider the class of regression estimators defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{k} w_i \rho(|r|_{\nu(i,\theta)}), \tag{27}$$

where ρ is strictly increasing continuous function such that $\rho(0) = 0$.

This class of estimators is regression, scale and affine equivariant following the reasoning of Rousseeuw and Leroy (1987).

Theorem.*The breakdown point of the regression estimators (27) is equal to $(n-k)/n$ if the index k is within the bounds $(n+p+1)/2 \leq k \leq n-p-1$, $n \geq 3(p+1)$ and the data points $x_i \in R^p$ for $i = 1, \ldots, n$ are in general position.*

We should remind that the observations $x_i \in R^p$ for $i = 1, \ldots, n$ are in general position if any p of them are linearly independent.

This class of estimators (27) contains also:

- Least Squares Estimators (LSE) if $\rho(|r|_{(i)}) = r^2_{(i)}$
- Least Absolute Value Estimator (LAV) if $\rho(|r|_{(i)}) = |r|_{(i)}$ and $w_i \equiv 1$ for $i = 1, 2, \ldots, n$;
- Chebishev minmax estimator if $\rho(|r|_{(n)}) = |r|_{(n)}$, $w_n = 1$ and $w_i = 0$ for $i = 1, 2, \ldots, n-1$;
- LMS and LTS estimators of Rousseeuw (1984);
- h-trimmed weighted L_q estimators of Müller (1995) if $\rho(|r|_{(i)}) = |r|^q_{(i)}$,
- rank-based linear regression estimators proposed by Hössjer (1994), where the weights w_i are generated by a function of the residual's ranks.

Rousseeuw and Hubert (1999) introduced a notion of depth in the regression setting. It provides the "rank" of any line (plane), rather than ranks of observations or residuals. In simple regression they can compute the depth of any line by a fast algorithm. For any bivariate dataset Z_n of size n there exists a line with depth at least $n = 3$. The largest depth in Z_n can be used as a measure of linearity versus convexity.

In both simple and multiple regression they introduce the deepest regression method, which generalizes the univariate median and is equivariant for monotone transformations of the response. Throughout, the errors may be skewed and heteroscedastic.

They also consider depth-based regression quantiles. They estimate the quantiles of y given x, as do the KoenkerBassett regression quantiles, but with the advantage of being robust to leverage outliers. They explore the analogies between depth in regression and in location, where Tukey's halfspace depth is a special case of this general definition.

7.2 Algorithms

Peña and Yohai (1999) propose a procedure for computing a fast approximation to regression estimates based on the minimization of a robust scale. The procedure can be applied with a large number of independent variables where the usual algorithms require an unfeasible or extremely costly computer time. Also, it can be incorporated in any high-breakdown estimation method and may improve it with just little additional computer time. The good performance of the procedure allows identification of multiple outliers, avoiding masking effects.

References

ANDREWS, D. F., BICKEL, P. J., ET AL. (1972). *Robust Estimates of Location*. Princeton Univ. Press, Princeton.

BICKEL, P. J. (1976). Another look at robustness: a review of reviews and some new developments. *Scand. J. Statist. Theory and Appl.*, 3(4):pp. 145 – 168.

BIRNBAUM, A. and LASKA, E. (1967). Optimal robustness: a general method with application to linear estimators of location. *J. Amer. Statist. Assoc.*, 62:pp. 1230 – 1240.

BOX, G. E. P. (1953). Nonnormality and test on variances. *Biometrika*, 40:pp. 318 – 335.

DANIEL, C. (1920). Observations weighted according to order. *Amer. J. Math.*, 42:pp. 222 – 236.

FERNHOLZ, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Springer, New York.

HADI, A. and LUCEÑO, A. (1997). Maximum Trimmed Likelihood Estimators: A Unified Approach, Examples and Algorithms. *Comput. Statist. and Data Analysis*, 25:pp. 251 – 272.

HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

HAMPEL, F. R., RONCHETTI, E., ET AL. (1986). *Robust Statistics. The Approach Based on Influence Functions*. Wiley, New York.

HAMPEL, F. R. (1968). *Contributions to the Theory of Robust Estimation*. Ph.D. thesis, University of California, Berkeley.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:pp. 383 – 393.

HODGES, Z. L. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.*, 34:pp. 598 – 611.

HÖSSJER, O. (1994). Rank-based estimates in the linear model with high breakdown point. *J. Amer. Statist. Assoc.*, 89(425):pp. 149–158.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:pp. 73 – 101.

HUBER, P. J. (1981). *Robust Statistics*. John Wiley, New York.

MARINCHEVA, M. Z. and VANDEV, D. (1995). On high breakdown point estimators of location and scale. In: *Statistical Data Analysis, Proceedings of SDA-95, SDA-96*, pp. 51–57. Sofia.

MILI, L., CHENIAE, M., ET AL. (1994). Robust State Estimation of Electric Power System. *IEEE Transactions on circuit and systems-I: Fundamental Theory and Applications*, 41(5):pp. 349 – 358.

MILI, L., PHANIRAJ, V., ET AL. (1991). Least Median of Squares Estimation in Power System. *IEEE Transactions on Power Systems*, 6(2):pp. 511 – 523.

MÜLLER, C. H. (1995). Breakdown points for designed experiments. *J. Statist. Plann. Inference*, 45:pp. 413 – 427.

NEYKOV, N. and NEYTCHEV, P. (1990). A robust alternative of the maximum likelihood estimators. In: *Short communications of COMPSTAT'90*, pp. 99 – 100. Dubrovnik, Yugoslavia.

PEÑA, D. and YOHAI, V. (1999). A fast procedure for outlier diagnostics in large regression problems. *J. Amer. Statist. Assoc.*, 94(446):pp. 434–445. ISSN 0162-1459.

RÖMISCH, U., VANDEV, D., ET AL. (2001). Determination of the geographical origin of wines from east european countries by methods of multivariate data analysis. In: *Seminar, Region Oesterreich- Schweiz (ROeS) of the International Biometric Society*. 24 - 27 September 2001, Mayrhofen.

ROUSSEEUW, P. J. and HUBERT, M. (1999). Regression depth. *J. Amer. Statist. Assoc.*, 94(446):pp. 388–402.

ROUSSEEUW, P. J. and LEROY, A. (1987). *Robust Regression and Outlier Detection*. John Wiley and Sons, New York.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388):pp. 871–880.

SHEVLYAKOV, G. L. and VILCHEVSKI, N. O. (2002). *Robustness in data analysis: criteria and methods*, volume 6 of *Modern Probability and Statistics'*. VSP, Utrecht. V. Yu. Korolev, V. M. Zolotarev, Editors-in-Chief.

STROMBERG, A. J., GRIFFITH, W., ET AL. (1998). Control Charts for the Median and Iinterquartile Range. Department in Statistics, University of Kentucky, USA.

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In: OLKIN, I. (ed.), *Contributions to Probability and Statistics*, pp. 448 – 485. Stanford Univ. Press, Stanford.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

VANDEV, D. L. and NEYKOV, N. M. (1993). Robust maximum likelihood in the gaussian case. In: MORGENTHALER, S., RONCHETTI, E., ET AL. (eds.), *New Directions in Statistical Data Analysis and Robustness*, pp. 257 – 264. Birkhauser Verlag, Basel.

VANDEV, D. L. and NEYKOV, N. M. (1998). About regression estimators with high breakdown point. *Statistics,* 32:pp. 111 – 129.

VANDEV, D. (1993). A note on breakdown point of the least median squares and least trimmed squares. *Statist. Probab. Lett.*, 16:pp. 117 – 119.