Robustified Maximum Likelihood

D.Vandev (SU), N.Neykov (BAS)

to be presented at Skiathos, 14 Meeting of ESI

Abstract

It is well known that the Maximum Likelihood Estimator (MLE) can be very sensitive to some deviations from the assumptions, in particular to unexpected outliers in the data. To overcome this problem many robust alternatives of the MLE have been developed in the last decades.

Neykov and Neytchev (1990), following the definitions of the Least Median of Squares (LMS) and Least Trimmed Squares (LTS) estimators of Rousseeuw (1984) introduced an extension of the maximum likelihood principle in the case of estimating the parameters of any unimodal distribution with regular density.

We are going to present here the recent results in this field.

Content of the talk

- Introduction - brief history

Content of the talk

- Introduction - brief history
- Weighted Trimmed Likelihood

Content of the talk

- Introduction - brief history
- Weighted Trimmed Likelihood
- Definition of breakdown point

Content of the talk

- Introduction - brief history
- Weighted Trimmed Likelihood
- Definition of breakdown point
- Definition of d-fullness

Content of the talk

- Introduction - brief history
- Weighted Trimmed Likelihood
- Definition of breakdown point
- Definition of d-fullness
- Multivariate normal distribution

Content of the talk

- Introduction - brief history
- Weighted Trimmed Likelihood
- Definition of breakdown point
- Definition of d-fullness
- Multivariate normal distribution
- Linear regression

Content of the talk

- Introduction - brief history
- Weighted Trimmed Likelihood
- Definition of breakdown point
- Definition of d-fullness
- Multivariate normal distribution
- Linear regression
- Nonstandard Linear regression setups

Content of the talk

- Introduction - brief history
- Weighted Trimmed Likelihood
- Definition of breakdown point
- Definition of d-fullness
- Multivariate normal distribution
- Linear regression
- Nonstandard Linear regression setups
- Algorithms

Introduction

Consider the multiple regression 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 $\sigma^2 > 0$.

Introduction

Consider the multiple regression 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

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

$$\text{LTS}(k)(r_1, \ldots, r_n) = \operatorname*{argmin}_\theta \sum_{i=1}^{k} r^2_{\nu(i,\theta)}. \tag{3}$$

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 using "smallest residuals" only.

Neykov and Neytchev (1990) proposed to replace in these estimators the squared residuals with $-\log\psi(x_i,\theta)$ of the individual observations. 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.

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 using "smallest residuals" only.

Neykov and Neytchev (1990) proposed to replace in these estimators the squared residuals with $-\log\psi(x_i,\theta)$ of the individual observations. 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.

$$\mathsf{LME}(k) \quad = \quad \underset{\theta}{\operatorname{argmin}}\{-\log\psi(x_{\nu(k,\theta)},\theta)\}, \tag{4}$$

$$\mathsf{LTE}(k) \quad = \quad \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{k}\{-\log\psi(x_{\nu(i,\theta)},\theta)\}. \tag{5}$$

Thus the idea was to maximize the likelihood over only k observations with "largest likelihood", not over all observations.

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 using "smallest residuals" only.

Neykov and Neytchev (1990) proposed to replace in these estimators the squared residuals with $-\log\psi(x_i,\theta)$ of the individual observations. 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.

$$\mathsf{LME}(k) \quad = \quad \underset{\theta}{\operatorname{argmin}}\{-\log\psi(x_{\nu(k,\theta)},\theta)\}, \tag{4}$$

$$\mathsf{LTE}(k) \quad = \quad \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{k}\{-\log\psi(x_{\nu(i,\theta)},\theta)\}. \tag{5}$$

Thus the idea was to maximize the likelihood over only k observations with "largest likelihood", not over all observations.

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

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 $\psi(e, \theta)$ with unknown vector parameter θ. Let the weights w_i for $i = 1, \ldots, n$ be fixed nonnegative numbers.

$$\text{WTL(k)}(x_1, \ldots, x_n) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{k} \{-w_i \log \psi(x_{\nu(i,\theta)}, \theta)\} \tag{6}$$

where $\psi(x_{\nu(i,\theta)}, \theta) \geq \psi(x_{\nu(i+1,\theta)}, \theta)$ are the ordered density values and 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 $i = 1, \ldots, n$, whereas the maximum likelihood estimator is derived if $k = n$.

Definition of breakdown point

The most important characteristic of any robust estimator is its breakdown point (BP). 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.

Definition 1. The breakdown point of an estimator $T(\Omega)$ is given by

$$\varepsilon_n^*(T) = \frac{1}{n} \max\{m : \sup_{\tilde{\Omega}_m} \|T(\tilde{\Omega}_m)\| < \infty\}, \tag{7}$$

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

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

Definition of d-fullness

In order to study the breakdown properties of general estimators of the type (4) and (5) 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 (6). First the definition

Definition of d-fullness

In order to study the breakdown properties of general estimators of the type (4) and (5) 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 (6). First the definition

Definition 2. 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 Lebesgue 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

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

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 $(n-k)/n$.*

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

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

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 needed to make unique guess for the estimated parameter.

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.

The breakdown point of the linear regression estimators

Consider the class of regression estimators defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{k} w_i \rho(|r|_{\nu(i,\theta)}), \tag{8}$$

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 (8) 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 shall remind that the observations $x_i \in R^p$ for $i = 1, \ldots, n$ are in general position if the convex hull any $p+1$ of them has nonzero measure.

The class of estimators (8) is quite general - it contains also:

- the Least Squares Estimators (LSE) if $\rho(|r|_{(i)}) = r^2_{(i)}$ and the Least Absolute Value Estimator (LAV) if $\rho(|r|_{(i)}) = |r|_{(i)}$ and $w_i \equiv 1$ for $i = 1, 2, \ldots, n$;

The class of estimators (8) is quite general - it contains also:

- the Least Squares Estimators (LSE) if $\rho(|r|_{(i)}) = r^2_{(i)}$ and the Least Absolute Value Estimator (LAV) if $\rho(|r|_{(i)}) = |r|_{(i)}$ and $w_i \equiv 1$ for $i = 1, 2, \ldots, n$;

- the Chebishev minmax estimator if $\rho(|r|_{(n)}) = |r|_{(n)}$, $w_n = 1$ and $w_i = 0$ for $i = 1, 2, \ldots, n-1$;

The class of estimators (8) is quite general - it contains also:

- the Least Squares Estimators (LSE) if $\rho(|r|_{(i)}) = r^2_{(i)}$ and the Least Absolute Value Estimator (LAV) if $\rho(|r|_{(i)}) = |r|_{(i)}$ and $w_i \equiv 1$ for $i = 1, 2, \ldots, n$;

- the Chebishev minmax estimator if $\rho(|r|_{(n)}) = |r|_{(n)}$, $w_n = 1$ and $w_i = 0$ for $i = 1, 2, \ldots, n-1$;

- the LMS and LTS estimators of Rousseeuw (1984);

The class of estimators (8) is quite general - it contains also:

- the Least Squares Estimators (LSE) if $\rho(|r|_{(i)}) = r^2_{(i)}$ and the Least Absolute Value Estimator (LAV) if $\rho(|r|_{(i)}) = |r|_{(i)}$ and $w_i \equiv 1$ for $i = 1, 2, \ldots, n$;

- the Chebishev minmax estimator if $\rho(|r|_{(n)}) = |r|_{(n)}$, $w_n = 1$ and $w_i = 0$ for $i = 1, 2, \ldots, n-1$;

- the LMS and LTS estimators of Rousseeuw (1984);

- the h-trimmed weighted L_q estimators of Müller (1995) if $\rho(|r|_{(i)}) = |r|^q_{(i)}$;

The class of estimators (8) is quite general - it contains also:

- the Least Squares Estimators (LSE) if $\rho(|r|_{(i)}) = r^2_{(i)}$ and the Least Absolute Value Estimator (LAV) if $\rho(|r|_{(i)}) = |r|_{(i)}$ and $w_i \equiv 1$ for $i = 1, 2, \ldots, n$;

- the Chebishev minmax estimator if $\rho(|r|_{(n)}) = |r|_{(n)}$, $w_n = 1$ and $w_i = 0$ for $i = 1, 2, \ldots, n-1$;

- the LMS and LTS estimators of Rousseeuw (1984);

- the h-trimmed weighted L_q estimators of Müller (1995) if $\rho(|r|_{(i)}) = |r|^q_{(i)}$;

- the rank-based linear regression estimators proposed by Hössjer (1994), where the weights w_i are generated by a function of the residuals ranks.

In conclusion, if $k = (n + p + 1)/2$ we easily find the highest breakdown point that is derived by Rousseeuw and Leroy (1987) and Hössjer (1994) respectively about the LQS and LTS, and the rank-based regression estimators. The usefulness of d–fullness is evident:

- The breakdown point can be exemplified by the range of values of k. This allows the statistician to choose the tuning parameter k according to the expected percent of outliers in data. The corresponding estimator will possess breakdown point less than the highest possible but it will be more efficient at the same time

In conclusion, if $k = (n + p + 1)/2$ we easily find the highest breakdown point that is derived by Rousseeuw and Leroy (1987) and Hössjer (1994) respectively about the LQS and LTS, and the rank-based regression estimators. The usefulness of d–fullness is evident:

- The breakdown point can be exemplified by the range of values of k. This allows the statistician to choose the tuning parameter k according to the expected percent of outliers in data. The corresponding estimator will possess breakdown point less than the highest possible but it will be more efficient at the same time

- The number of observations n in the linear regression setup must be at least $3(p+1)$

In conclusion, if $k = (n + p + 1)/2$ we easily find the highest breakdown point that is derived by Rousseeuw and Leroy (1987) and Hössjer (1994) respectively about the LQS and LTS, and the rank-based regression estimators. The usefulness of d–fullness is evident:

- The breakdown point can be exemplified by the range of values of k. This allows the statistician to choose the tuning parameter k according to the expected percent of outliers in data. The corresponding estimator will possess breakdown point less than the highest possible but it will be more efficient at the same time

- The number of observations n in the linear regression setup must be at least $3(p+1)$

- It is sufficient only that the observations $x_i \in R^p$ to be in general position not the pair (y_i, x_i^T) of observations for $i = 1, 2, \ldots, n$ as it is usually assumed.

Linear regression with exponential q–th order distributions

Let the errors ε_i of the regression model (1) are i.i.d. with q–th power exponential distribution, i.e. the density function of ε_i is given by

$$\phi(\varepsilon, \beta, \sigma) = \frac{q(1/2)^{(1+1/q)}}{\sigma\Gamma(1/2)} \exp\{-\frac{1}{2}|\frac{\varepsilon}{\sigma}|^q\},$$

The Gaussian for $q = 2$, the Laplace for $q = 1$, the double exponential for $0 < q < 2$, the leptokurtic for $1 < q < 2$, the platikurtic for $q > 2$, the rectangular for $q \to \infty$ distributions are obtained as particular cases.

In order to obtain the breakdown properties of the $WTL(k)$ regression estimators of β and σ (when estimating simultaneously), we need to study the d–fullness of the set $f_i(\beta, \sigma), \ldots, f_n(\beta, \sigma)$, where

$$f_1(\beta, \sigma) = -\log(\phi(\varepsilon_i, \beta, \sigma)) = \frac{1}{2}|r_i/\sigma|^q + \log(\sigma) + C_1$$

and

$$C_1 = \log(\Gamma(1/2)) - \log(q(1/2)^{(1+1/q)}).$$

Theorem. *The breakdown point of the $WTL(k)$ estimators of β and scale σ in the linear regression model with q–th order exponential distribution for any $q \in [1, \infty)$ is equal to $(n-k)/n$ if $n \geq 3(p+1)$, $x_i \in R^p,\ i = 1, \ldots, n$ are in general position, the weights $w_i \geq 0$ for $i = 1, 2, \ldots, n,\ \ w_k > 0$, and the index k is within the bounds $(n+p+1)/2 \leq k \leq n-p-1$.*

If σ is known or treated as a nuisance parameter in $f_i(\beta, \sigma)$ the $LME(k)$ ($LTE(k)$) estimators of β are equivalent to the LQS (LTS) regression estimators of Rousseeuw (1984).

The grouped binary regression

A high breakdown point estimator based on LQS regression estimator of Rousseeuw (1984) for binary regression data was considered by Christmann (1994). The type of the data under consideration has the form (y_i, x_i^T) for $i = 1, \ldots, m$ where y_i is assumed to be binomially distributed, $b(y_i \mid n_i, \pi_i)$, where the group size is n_i, the probability of success is π_i, x_i is a $p \times 1$–dimensional vector of covariates (explanatory variables) and the total number of observations is $n = n_1 + n_2 + \cdots + n_m$.

We assume that π_i follows the linear logistic regression model

$$\pi_i = \exp(x_i^T \beta)/(1 + \exp(x_i^T \beta),$$

where β is a $p \times 1$–dimensional vector of unknown parameters, but probit or other link functions could be used.

Hereafter we shall also assume that $0 < y_i < n_i$ for each i. We study the d–fullness

of the set $\{f_1(\beta), \ldots, f_n(\beta)\}$, where

$$f_i(\beta) = f_i(y_i, x_i^T, \beta) = -\log(b(y_i \mid n_i, \pi_i)) =$$

$$= -y_i(x_i^T\beta) + n_i \log(1 + exp(x_i^T\beta)) - \log \binom{n_i}{y_i}.$$

Theorem. *The breakdown point of the $WTL(k)$ estimators in the grouped binary logistic linear regression model defined above is equal to $(m-k)/m$ if the data $x_i \in R^p$ for $i = 1, \ldots, m$ are in general position, the weights $w_i \geq 0$ for $i = 1, 2, \ldots, m$, $w_k > 0$ for $k = \max\{i : w_i > 0\}$, $m \geq 3(p+1)$ and the index k is within the bounds $(m + p + 1)/2 \leq k \leq m - p - 1$.*

Remark 2. Note that the meaning of breakdown point here is different from the one given by Definition 1 (7) as we consider the triple (n_i, y_i, x_i^T) as one observation.

The above results also hold in the case of a homoscedastic linear regression model with replications.

Algorithms

The problem of finding the minima in (6) is not easy. The objective function has multiple local minima. Thus one has to combine methods of nonlinear and discrete opimisation.

For calculating LME(k) (like least median of residuals and Minimum Volume Ellipsoid) estimators of covariance matrix

- resampling techinque was proposed by (Rousseeuw and Leroy, 1987).A subset of k ($k \geq d$) observations is drawn at random and the LS estimate is calculated. This procedure is repeated many times, and the fit with the lowest TL objective function (6) is retained.;
- iterating improvement - (Rousseeuw and Zomeren, 1989);
- An iterative approximate algorithm for finding the LTE and LME estimates was considered by Neykov and Neytchev (1990) based on the resampling techinque. They calculated ML estimates on subsets which may be used with any kind of error

- simulated annealing - (Todorov, 1992); Combining Branch&Bound with iterating improvement
- A Stochastic Optimisation Algorithm was proposed by Vandev (1995)in [4].

References

[1] Neykov, N., P. Neytchev, A robust alternative of the maximum likelihood estimators, Short communications of COMPSTAT'90, Dubrovnik, Yugoslavia, 99-100, 1990

[2] Vandev, D. A note on breakdown point of the least median squares and least trimmed squares. Statist. & Prob. Letters, 16, 1993, 117-119.

[3] Vandev, D. L.,N. M. Neykov, Robust maximum likelihood in the Gaussian case, In: New Directions in Statistical Data Analysis and Robustness, (Eds. S. Morgenthaler, E. Ronchetti, and W. A. Stahel), Basel, Birkhauser Verlag, 1993, 257-264.

[4] Vandev, D., Computation of the Trimmed L_1 -median, In: Multidimensional Anal-

ysis in Behavioral Sciences. Philosophic to technical, Ed. I.Parchev, Prof.Marin Drinov Publ.House, Sofia, 1995, 152-157.

[5] Maya Z. Marincheva and Dimitar L. Vandev, On High Breakdown Point Estimators of Location and Scale, In: Statistical Data Analysis, Proceedings of SDA-95, SDA-96,Sofia,1995,51–57.

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

[7] Atanasov, D. V. and Neykov, N. M. (2001). On the finite sample BP of the WTL estimators and the d–fullness of a set of continous functions. CDAM conference, Minsk, Belarus, (submitted)