PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **17** (2005), 323–335

PLISKA
STUDIA MATHEMATICA
BULGARICA

STOCHASTIC OPTIMIZATION IN ROBUST STATISTIC

D. Vandev [1]

The paper studies a stochastic optimization algorithm for computing of robust estimators of location proposed by Vandev (1992). A random approximation of the exact solution was proposed which is much cheaper in time and easy to program.

Two examples are presented. Besides standard estimators of location like trimmed mean also robust regressions (LMS and LTS) introduced by Rousseeuw and Leroy are considered. MATLAB programs are included.

1. Introduction

Many authors considered robust estimators of the covariance matrix and the location in the multidimensional case. When a high level of contamination is expected it is appropriate to use estimators with high breakdown point. Such estimators are the minimum volume ellipsoid (MVE) and the minimum covariance determinant (MCD), introduced by Rousseeuw and Leroy [9]. On the other hand in the robust regression literature very popular is the Least Median of the squares (LME) estimator which also has high breakdown point. Recently Neykov and Neytchev [6] proposed a robust alternative of the maximum likelihood estimators. Namely let $f(\theta, x)$ be the likelihood functions of the individual observation x. We denote by X the finite set of all observations. Here θ is the vector of unknown parameters. Let $A(\theta) = \{-log(f(\theta, x)), x \in X\}$ be the (increasingly) ordered set

[1] Partialy supported by Pro-ENBIS GTC1 -2001-43031

2000 *Mathematics Subject Classification:* 62J05, 62G35

Key words: robust estimators of location, least median of squares, stochastic approximation algorithm, Monte – Carlo study

of the values of f at a fixed point θ. Denote by $M(k,\theta)$ the k-smallest and by $S(k,\theta)$ the sum of the k smallest numbers of the set $A(\theta)$. The minimizers of these two random functions are to be considered as estimators in statistical sense.

Vandev [13] has shown that MVE and MCD estimators may be extracted from this robustified version in the gaussean case. The same is true for LME in regression. It was also shown that in general all robustified maximum likelihood estimators have high break down point.

Computationally both (trimmed and least median) problems are not easy to solve in a conventional way because the functions involved have many local minima. Thus the minimization turns out to be a serious combinatorial problem. Up to now mainly the resampling technique is used for the purpose, see Rousseeuw & Leroy [9].

In this paper an algorithm is presented for approximate calculating of LME(k) and LTE(k). Hawkins [2] used a feasible set algorithm for exact calculation of the minima. Our proposition is based on the well known Robins-Monro [8] procedure for stochastic optimization, which was already successfully used by Martin and Masreliez [5] in the robust estimation. We will call the algorithm RM algorithm.

2. Robust estimators in statistics

For modeling gross errors and outliers in the sample, the most popular is the Tukey supermodel [12] based on the Gaussean low:

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

Huber [3] considered more general model

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

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

2.1. 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 breakdown 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

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

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 [1].

2.2. LMS and LTS

The multiple regression is probably most used statistical procedure in the 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 [10] as robust alternatives of the LSE

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

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

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

Theorem 1. *The breakdown point of the regression estimators (4 and 5) is equal to* $(n-k)/n$ *if the index* k *is within the bounds* $(n+p+1)/2 \le k \le n-p-1$, $n \ge 3(p+1)$ *and the data points* $x_i \in R^p$ *for* $i = 1, \ldots, n$ *are in general position.*

This theorem was first proved by Rousseeuw [10] and then easily by Vandev [13] with different technique.

2.3. Robustified Maximum Likelihood

Neykov and Neytchev [6] 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.

$$\text{(6)} \qquad \mathrm{LME}(k) = \underset{\theta}{\operatorname{argmin}} \{-\log \psi(x_{\nu(k,\theta)}, \theta)\},$$

$$\text{(7)} \qquad \mathrm{LTE}(k) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{k} \{-\log \psi(x_{\nu(i,\theta)}, \theta)\}.$$

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

2.4. Traditional Algorithms

The main resource of information is the excellent WEB page [16] created in Antwerpen. Another useful source of information is the dissertation of Werner [15], where all problems of detection of multidimesional outliers are studied thoroughly. Pena [7] proposed 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.

Mahalanobis distances

Given the sample, they are defined as

$$d_i = (x_i - \mu)' C^{-1} (x_i - \mu),$$

where μ and C are suitable (robust) estimators of the mean and the covariance matrix. When no outliers are present and the data are normally distributed these numbers follow χ^2 distribution with p degrees of freedom.

Thus all traditional algorithms are based on some (robust) estimators of the mean and the covariance as a first step and choosing the suitable cutoff value for marking outliers with higher Mahalanobis distance.

First we will consider two algorithms for detecting a small number of outliers (they do not need special robust estimators):

— Algorithm 1: (`OUTDV1`)

In the first step some good observations on the base of a low confidence level (.65) will be detected by using conventional estimators.

```
xm=mahal(x,x);
mq=chi2inv(gamma1^(1/n),m);
J=find(xm<mq);  %good observations
```

— Algorithm 2: (`OUTDV2`)

In the first step a number (prescribed by user) of observations with higher Mahalanobis distance are deleted.

```
xm=mahal(x,x);
i=max([n-io;m+1])
xx=sort(xm);mq=xx(i,1);J=find(xm<mq);
xm=mahal(x,x(J,:));
xx=sort(xm);mq=xx(i,1);J=find(xm<mq);
```

Both algorithms proceed in the same manner in the second step: the Mahalanobis distances of all observations to good observations will be calculated and compared with the quintile of the distribution of maximum of n independent chi-squares.

```
xm=mahal(x,x(J,:));
mq=chi2inv(gamma^(1/n),m);
J=find(xm>mq);
```

When the expected number of outliers is a considerable percent of the sample $(10 - 50\%)$, a special attention should be made on the first step.

A typical distribution of simulated data with contamination – the problem of cutoff value is also important in this case.

MCD and MVE

In the famous book [9] the Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) estimators were proposed for robust estimation of the covariance matrix. Both estimators became usable when fast versions of the initial algorithms were developed. Rousseeuw [11] proposed fast version of MCD.

3. Stochastic Optimization

The famous Robins-Monro [8] procedure, later extended by Kiefer and Wolfowitz [4], when applied to the problem of minimizing the function $F(\theta)$ consists in the following. Let start with some $\theta = \theta_0$. Let now calculate the gradient $grad(F(\theta))$

at this point. It may be randomly disturbed by some random variable with zero expectation. At the step i the parameter will be changed according the following formula:

$$\theta_{i+1} = \theta_i - \gamma_i * \frac{grad(F(\theta_i)}{||grad(F(\theta_i))||}. \tag{8}$$

The sequence $\{\gamma_i, i = 1, 2, ...\}$ is chosen to satisfy the relations: $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$, $\sum_{i=1}^{\infty} \gamma_i = \infty$. Here the only difference with the the standard method as described in Wasan [14] is the normalizing of the gradient.

4. The Proposed Algorithm

Let F be set of functions of size n defined on p-dimensional Euclidean space E. Let $A(\theta) = \{f(\theta, x), x \in X\}$ be the (increasingly) ordered set of the values of f at a fixed point θ. Denote by $M(k, \theta)$ the k-smallest number in the set $A(\theta)$ and by $T(k, \theta)$ - the sum of k smallest numbers. Denote by:

$$LME(k) = \arg\min_{\theta} M(k, \theta) = \arg\min_{\theta} f_{(k)}(\theta), \tag{9}$$

$$LTE(k) = \arg\min_{\theta} T(k, \theta) = \arg\min_{\theta} \sum_{i=1}^{k} f_{(i)}(\theta), \tag{10}$$

where $f_{(1)}(\theta) \leq f_{(2)}(\theta) \leq \cdots \leq \cdots \leq f_{(n)}(\theta)$. As usual here the subindex denote the element of the corresponding permutation which depends on the value of θ.

```
Step 0.  SET number of iterations maxi, set i=1, set \delta.
Step 1.  Chose at random 10 indexes among the numbers from 1 to n.
         Calculate these 10 functions.  Sort their values.
Step 2.  Chose  the  value j, such that (j/10=k/n)
         and the  function which  produces that value (say f).
Step 3.  Calculate the normalized gradient D(f) of f.
Step 4.  SET B:=B - D(f)*\delta /i.  Set i=i+1.
         IF i < maxi THEN GOTO STEP 1.
```

5. MATLAB program

Here we present a MATLAB program able to handle the stochastic approximation algorithm in robust statistics.

```
function [theta] = soaml(x,theta0,FUN,pr,delta,iter)
[n,m]=size(x); theta=theta0;
for k=1:iter
    gama=delta/k;                          %   new gama
    J=round(ones(kkk,1)/2+rand(kkk,1)*n);  %   10 random in (1:n) numbers
    eval(['[Y,X]=' FUN '(x(J,:),theta);']);% residuals, gradient
    [dum,list]=sort(Y);      %   sort 25 values
%================LME or LTE ==================
    jj=list(pr,1);        %  jj=list(1:pr,1);
    s=X(jj,:)';           %  s=sum(X(jj,:))';
%==============================================
    w=sqrt(s'*s);
    theta=theta-s*(gama/w);
end
```

Pgm. 1: LME (and LTE) program

Note between commented lines the minor changes needed to transform this program for work in the LTE case.

The user defined function `[Y,X]=FUN(x,theta)` should produce in `Y` the values corresponding to observations in `x` and in `X` – corresponding gradients. Below we present some examples of such functions for various estimators:

```
function [Y,X]=gradmea(x,a)
[n,m]=size(x);
aa=a';
X=x-aa(ones(n,1),:);
Y= diag(X*X');
X=-2*X.*x;
```

Pgm. 2: Location

```
function [Y,X]=gradreg(x,a)
[n,m]=size(x);
xx=[ones(n,1),x(:,2:m)];
Y=x(:,1)-xx*a;
X= -2*(Y(:,ones(1,m)).*xx);
Y=Y.*Y;
```

Pgm. 3: Regression

```
function [Y,X]=gradnor(x,a)
[n,dum]=size(x);
mu=a(ones(n,1),1);
si=exp(a(2,1));
Y=(x-mu)/si;
X=[-Y/(2*si),(ones(n,1)-Y.*Y)];
Y=Y.*Y/2+ones(n,1)*a(2);
```

Pgm. 4: $N(\mu,\sigma)$ in R^1

6. Examples of application

Here we present several simulated examples. In all cases we use 1000 observations generated and 20% contamination, when not mentioned other.

6.1. Location

The 6-dimensional mean, 100 hundred repetitions, 6/10 LME:

Table 1: LME estimate of location

True	56.1761	0.9569	2.0455	3.0177	4.0263	4.9971
Est.	56.1019	0.9668	2.0476	2.9981	4.0225	4.9860
Err.	0.6054	0.0915	0.0936	0.0964	0.0870	0.0942

6.2. Simple regression

The first regression model was chosen to illustrate the robust properties of the used version of maximum likelihood. The response Y is generated by the following model:

$$y = 5 - 2 * x + e.$$

Here e is a standard normal random variable. The sample consists of 1000 observations. It was corrupted by destroying 30% of the observations. The algorithm was used with number of iterations equal to 150 and $\delta = 10$.
On Fig.1 a random solution is presented for the estimator 6/10. For a comparison the unique least squares solution is also plotted.

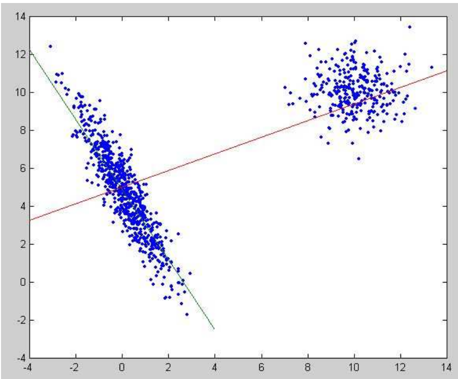

Figure 1: LME and LSQ regression

6.3. Multiple Regression

The model is:

$$y = 2 - 2 * x_1 + 5 * x_2 - 5 * x_3 + x_4 + e.$$

The aim was to test the performance of different estimators of the same kind (LME) when the percent of contamination changes.

In this case we each time generate totally new data set of 4000 uniform random numbers for x and 1000 normal for e. Each experiment was performed 100 times in order to estimate the variance.

The results are presented in the following table. The number of contaminated observations is shown in the first column. The form of used estimator is in the second column. Each cell in the table contains the average (with the sample standard error below) for 100 simulated with the same model data sets. In the next 4 columns are the results for the parameters of the model. The last column represents the obtained value of the functional.

Table 2: Simulation results for multiple regression

Cont.	Est.	$a_0 = 2$	$a_1 = -2$	$a_2 = 5$	$a_3 = -5$	LME
100	9/10	1.9235	-1.9421	4.9492	-4.8944	3.3777
		.1149	.1569	.1249	.1265	2.4862
	8/10	1.9644	-1.9821	4.9029	-4.9164	1.2839
		.0990	.0987	.1401	.1136	.1512
	7/10	1.9390	-2.0412	4.9467	-4.8380	1.1343
		.1596	.1834	.1959	.1705	.2168
	6/10	1.9823	-1.9756	4.7355	-4.7443	.9773
		.2313	.2328	.3107	.2601	.2017
200	8/10	1.9664	-2.0136	4.7889	-4.7541	5.9930
		.1534	.1828	.1808	.2410	3.7446
	7/10	1.9103	-1.9337	4.9010	-4.8629	1.3670
		.2338	.2833	.2781	.2783	.5853
	6/10	1.9484	-1.9812	4.8867	-4.9113	1.0957
		.1811	.2229	.2409	.1701	.2407
300	7/10	1.7643	-1.7374	4.4975	-4.5186	7.7961
		.4012	.3970	.6973	.7480	4.5630
	6/10	1.8873	-1.8956	4.8093	-4.7899	1.5153
		.3159	.2421	.5369	.4467	.8834
400	6/10	1.5886	-1.6556	4.2696	-4.1614	9.5648
		.5058	.4803	.7968	.9176	4.7157

What is easily seen in this table are the good results of 7/10 estimator for 10% contamination and 6/10 estimator – for 20%.

7. Mean and covariance

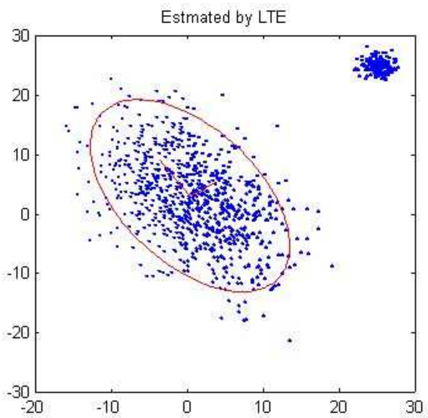

Figure 2: Location and scale

The estimating of variance needs special attention because it has to be positive. In the one-dimensional case the problem is solved using new parameter $\ln \sigma$ (see Pgm.4).In the multi-dimensional case however such approach is not easy. Before explaining difficulties let us present one unsuccessful example of two-dimensional estimate of mean and the covariance:

Table 3: Location and scale

Original mean	0.6242	2.5444
Estimated mean	0.9204	3.0970
Original Cova	37.0107	-20.4700
	-20.4700	51.5504
Estimated Cova	35.1363	-10.3004
	-10.3004	43.6298

7.1. The problem of gradient

In the simultaneous estimation of the mean and covariance the main problem consists in calculation of the gradient of $Q = -\log L(x, m, \Sigma)$:

$$Q = \log\det(\Sigma)^{1/2} + (x-\mu)'\Sigma^{-1}(x-\mu). \tag{11}$$

Let denote $M = \Sigma^{-1}$. Then it is easy to show that

$$\frac{dQ}{dM} = -M^{-1} + (x-\mu)(x-\mu)'. \tag{12}$$

Let us replace $M = \exp(L)$ as in the univariate case and try to use the formal relation

$$\frac{dQ}{dL} = \frac{dQ}{dM} \bigotimes \frac{dM}{dL}.$$

Consider the standard expansion of $\exp(L)$

$$M = \exp L = I + L + L^2/2! + L^3/3! + \ldots$$

The question now is how to represent $\frac{dM}{dL}$. We tried the following approximation of this $(m \times m)^2$ tensor:

$$\frac{dM}{dL} = (I + L/18) \bigotimes (I + L/18)$$

Thus we come to the result:

$$\frac{dQ}{dL} = (I + L/18)'((x-\mu)(x-\mu)' - M^{-1})(I + L/18)$$

Note that we are not sure how exact is this approximation.

7.2. The Simulation Results

These were obtained using 20% contamination of 1000 observations and MLE 6/10.

Table 4: Means

Original	18.0293	0.9973	-1.9745	3.0041	-6.0700	3.3209
Estimated	18.0253	1.0166	-2.0165	2.9931	-5.9780	3.3387
S.E.	0.1845	0.1340	0.1453	0.1536	0.1419	0.1077

Table 5: Original covariance matrix

6.0332	0.4005	-0.6253	1.0875	-1.9673	1.0251
0.4005	1.0741	-0.0554	-0.0257	0.0529	0.0359
-0.6253	-0.0554	0.8846	-0.0332	0.0724	0.0294
1.0875	-0.0257	-0.0332	0.9997	-0.0110	0.0234
-1.9673	0.0529	0.0724	-0.0110	2.2645	0.0299
1.0251	0.0359	0.0294	0.0234	0.0299	0.4159

Table 6: Estimated covariance matrix

4.0793	0.2168	-0.3825	0.6135	-1.2302	0.6816
0.2168	0.6719	-0.0042	0.0088	-0.0132	0.0119
-0.3825	-0.0042	0.6703	-0.0101	0.0233	-0.0072
0.6135	0.0088	-0.0101	0.6832	-0.0281	0.0238
-1.2302	-0.0132	0.0233	-0.0281	1.4797	-0.0266
0.6816	0.0119	-0.0072	0.0238	-0.0266	0.3330

While the estimation of mean is excellent (see Table 4), the bias of the covariance is obvious on Table 6. Thus the proposed algorithm was not successful with estimation of covariance matrix. The reason is that the true unbiased gradient is not easy to obtain.

Editor Note

This paper is compiled by the draft version and presentation of Dimitar Vandev. Presumably, it can not be considered as completed paper but it contains original ideas which we ought to present to statistical college.

REFERENCES

[1] F. R. Hampel, E. Ronchetti, P. J. Rousseeuw, W. A. Stahel. Robust Statistics. The Approach Based on Influence Functions. Wiley, New York, 1986.

[2] D. M. Hawkins. A feasible solution algorithm for the minimum volume ellipsoid estimator. *Comput. Statist.*, **9** (1993) 95–107.

[3] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, **35** (1964) 73–101.

[4] J. Kiefer, J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **23** (1952) 462–466.

[5] R. D. Martin, C. J. Masreliez. Robust estimation via stochastic approximation. *IEEE Trans. Inf. Theory* **21** (1975) 263–271.

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

[7] D. Peña, V. Yohai. A fast procedure for outlier diagnostics in large regression problems. *J. Amer. Statist. Assoc.*, **94** (1999) 434–445.

[8] H. Robbins, S. Monroe. On a Stochastic Approximation Technique, *Ann. Math. Statist.*, **22** (1951) 400–407.

[9] P. J. Rousseeuw, A. Leroy. Robust Regression and Outlier Detection. Wiley, New York, 1987.

[10] P. J. Rousseeuw. Least median of squares regression. *J. Amer. Statist. Assoc.*, **79** (1984) 871–880.

[11] P. J. Rousseeuw, K. Van Driessen. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41** (1999) 212–223.

[12] J. W. Tukey. A survey of sampling from contaminated distributions. In: *Contributions to Probability and Statistics* (ed. I. Olkin), Stanford Univ. Press, Stanford, (1960) 448–485.

[13] D. Vandev. A note on breakdown point of the least median squares and least trimmed squares. *Statist. Probab. Lett.*, **16** (1993) 117–119.

[14] M. T. Wasan. Stochastic approximation. Cambridge Univ.Press, 1969.

[15] M. Werner. Identification Of Multivariate Outliers In Large Data Sets. PhD Thesis, Univrsity of Colorado at Denver, 2003.

[16] `http://www.agoras.ua.ac.be/` Antwerp Group on Robust & Applied Statistics. (2004).

Dimitar L. Vandev
Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 J. Bourchier blvd.
1164 Sofia, Bulgaria