Robust Maximum Likelihood in the Gaussian Case

Vandev,D.L. [1]
and Neykov,N.M. [2]

Abstract

Some widely used robust estimators of location and scale are derived from two robust modifications of the maximum likelihood principle in the multivariate normal case.

1 Introduction

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 to the MLE have been developed in the last decades. Some of the widely known robust estimators of multivariate location and scatter matrix are M–estimators, Stahel – Donoho estimator, the Minimum Volume Ellipsoid estimator (MVE), the Minimum Covariance Determinant estimator (MCD) and S–estimators. For detailed introduction see Huber (1981), Hampel et al. (1986), and Rousseeuw and Leroy (1987).

Neykov and Neytchev (1990), following the definitions of the Least Median of Squares estimator (LMS) and the Least Trimmed sum of Squares (LTS) regression estimator of Rousseeuw (1984), introduced two classes of estimators for the parameters of any unimodal distribution with regular density function as an extension of the MLE. Vandev (1993) developed a technique for computing the breakdown point of these estimators. He proved that the breakdown point is not less than $(n-k)/n$, where k is a tuning constant of the estimators which can be chosen by the user within some reasonable range of values.

In this paper we extend the definition of Neykov and Neytchev (1990) to the multivariate normal case and illustrate the technique of Vandev (1993) concerning the breakdown point. The basic results concerning the connection of the finite-sample breakdown point, dimensionality of the Gaussian distribution and the notion of d–fullness introduced by Vandev (1993) are presented in section 3 and 4. It is shown that both MVE and MCD estimators introduced by Rousseeuw (1986) are obtained as a particular case.

2 Definitions and notations

Let $x_1, x_2, ..., x_n$ be a sample of n independent observations in the p-dimensional euclidean space E. The most popular estimator of the unknown location μ and the inverse scatter matrix S is the normal maximum likelihood estimator MLE which is defined as

$$\operatorname*{argmax}_{\mu,S} \prod_{i=1}^{n} \phi(x_i, \mu, S) = \operatorname*{argmin}_{\mu,S} \sum_{i=1}^{n} - \ln(\phi(x_i, \mu, S)),$$

[1] Inst. of Mathematics., Bulgarian Academy of Sciences, P.O. Box 373,1113 Sofia, Bulgaria
[2] Inst. of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66 Trakia Blvd., 1184 Sofia, Bulgaria

where $\phi(x, \mu, S)$ denotes the multivariate normal density,

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

In the following we shall recall the basic definition introduced by Neykov and Neytchev (1990) concerning the robustified MLE estimators.

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

Definition 1. The Least Median of log density values Estimator (LME) of θ is defined as

$$\text{LME}(x_1, x_2, ..., x_n) = \underset{\theta}{\text{argmin}} \underset{i}{\text{med}}(-\ln\psi(x_i, \theta)).$$

The least trimmed log-likelihood estimator (LTE) of θ is defined as

$$\text{LTE}(k)(x_1, x_2, ..., x_n) = \underset{\theta}{\text{argmin}} \sum_{i=1}^{k}\{-\ln(\psi(x, \theta))_{(i)}\},$$

where $(\psi(x, \theta))_{(1)} \geq (\psi(x, \theta))_{(2)}, ..., \geq (\psi(x, \theta))_{(n)}$ are the ordered density values. Instead of taking the median of the ordered density values we shall also consider the k-th order-statistics $(\psi(x, \theta))_{(k)}$, where $k > n/2$, and minimize $-\ln(\psi(x, \theta))_{(k)}$. This estimator will be denoted by LME(k).

The next definition is due to Vandev (1993).

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 all functions in this subset is a subcompact function.

We remined the reader that a real valued function $g(z)$ defined on a topological space Z is called subcompact, if its Lebesgue sets $L(M) = \{z : g(z) \leq M\}$ are compact (or empty) for any all M. The present definition is a little bit more general than the one given in Vandev (1993), where the positivity of the functions is required. But, note that by Comment 2 of Vandev (1993) the estimator LME(k) does not require this positivity.

The breakdown properties of both LME(k) and LTE(k) estimators in the Gaussian case are of primery interest in this paper. Vandev (1993) showed that the breakdown point of both LME(k) and LTE(k) is not less than $(n-k)/n$ if the set of n positive functions $-\ln\psi(x_i, \theta)$ is d–full and $(n+d)/2 < k < (n-d)$. In order to apply this result, it remains to determine the value of d for a particular family of densities.

3 Basic Result

In this section we shall study the d–fullness of the set $f(x_1, \mu, S), ..., f(x_n, \mu, S)$, where

$$f(x, \mu, S) = -\ln(\phi(x, \mu, S)) = (x-\mu)'S(x-\mu)/2 - (\ln\det(S))/2.$$

THEOREM 1. If $x_1, ..., x_n$ is a Gaussian sample, then the functions $\{f(x_1, \mu, S), ..., f(x_n, \mu, S)\}$ form a $(p+1)$–full set with probability one.

For the proof we need the following lemma.

LEMMA. For any $p+1$ points $x_1, ..., x_{p+1}$ in general position, the function

$$I(\mu, S) = \max_{i \in \{1,...,p+1\}} f(x_i, \mu, S)$$

is subcompact in μ and S.
Proof: We will use the following inequalities:

$$(1) \qquad 2I(\mu, S) \geq \frac{1}{p+1}\sum_{i=1}^{p+1}(x_i-\mu)'S(x_i-\mu) - \ln\det(S)$$

$$\geq \frac{1}{p+1}\sum_{i=1}^{p+1}(x_i-\bar{x})'S(x_i-\bar{x}) - \ln\det(S) + (\mu-\bar{x})'S(\mu-\bar{x}).$$

Here $\bar{x}$ is the mean of $x_1, ..., x_{p+1}$. We will postpone the investigation of the last term (which is nonnegative) on the right-hand side of and work on the inequalities for the first two terms:

$$\begin{aligned} 2I(\mu, S) &\geq \frac{1}{p+1}\sum_{i=1}^{p+1}(x_i-\bar{x})'S(x_i-\bar{x}) - \ln\det(S) \\ &\geq tr(SB) - \ln\det(SB) + \ln\det(B), \end{aligned}$$

where $B = \sum_{i=1}^{p+1}(x_i-\bar{x})(x_i-\bar{x})'/(p+1)$, which according to the assumption on $x_1, ..., x_{p+1}$ is positive definite. Denote by $\alpha_1, ..., \alpha_p$ the eigenvalues of SB. The inequalities

$$H = tr(SB) - \ln\det(SB) \geq \sum_{i=1}^{p}(\alpha_i - \ln(\alpha_i)) \geq p$$

impose compactness on the elements of the positive definite matrix SB, because $H \geq \alpha_i - \ln(\alpha_i)$ implies

$$e^{-H} < \alpha_i < e.min(1, H/(e-1)).$$

The left-hand inequality is obvious. The right-hand one follows easily from the simple inequality $\ln(x) < 1 + (1/e)(x-e)$ which holds for $e < x$.

These bounds on the eigenvalues of SB together with our assumption involving B show that S has eigenvalues that are bounded from below and above. Therefore, S lies in a compact set.

Now let us come back to the second term of (1). Denote by H the minimal value of the already estimated first two terms. Then we have the following

$$2I(\mu, S) - H - \ln\det(B) \geq (\mu-\bar{x})'S(\mu-\bar{x}).$$

Since the determinant of S is nonzero it follows that μ also belongs to a compact set. **Q.E.D.**
Proof (of the theorem): Consider a sample of size n. It is clear that with probability equal to 1 any $p+1$ points of the sample are in general position. By the lemma, the function $I(\mu, S)$ is subcompact in μ, S for any $p+1$ points. **Q.E.D.**

4 The Breakdown Point of the MVE and MCD estimators

Let us recall some definitions and properties of the well known MVE and MCD estimators introduced by Rousseeuw. For details see Rousseeuw and Leroy (1988) and Lopuhaa and Rousseeuw (1991).

The MVE estimator for multivariate location is defined as the center of the minimum volume ellipsoid covering at least h points, where $h = (n + p + 1)/2$. The corresponding MVE estimator of the covariance matrix is given by the same ellipsoid, multiplied by a suitable constant to obtain asymptotic consistency for multivariate normal data.

The MCD of the location μ is defined as the mean of the h points for which the determinant of the covariance matrix is minimal. The optimal choice for h is equal to $(n+p+1)/2$. The corresponding MCD estimator of the covariance matrix is given by the classical covariance matrix based on the selected h observations, multiplied by a suitable constant to obtain consistency for multivariate normal data.

Both the MVE estimator and the MCD estimator possess a breakdown point of 0.5 (the best that can be achieved). Furthermore, it can be shown that the MCD is $n^{1/2}$ - consistent and asymptotically normal while the MVE is $n^{1/3}$ - consistent. It is possible, however, to construct a robust estimator that retains the high breakdown point of the MVE and achieves $n^{1/2}$ - consistency (Lopuhaa and Rousseeuw,1991).

4.1 THE MVE ESTIMATOR

Theorem 1 of this paper and the result of Vandev (1993) applied to the Gaussian case imply that if one chooses k within the bounds, $(n + p + 1)/2 < k < n - p - 1$, then the LME($k$) estimator has breakdown point not less than $(n - k)/n$.

Let μ, S minimize the LME(k) criterion. This means that in the sample there are k observations $x_{l(n)}, x_{l(n-1)}, \ldots, x_{l(n-k-1)}$, such that $f(x_{l(n)}, \mu, S) \leq f(x_{l(n-1)}, \mu, S) \leq ... \leq f(x_{l(n-k-1)}, \mu, S)$, and all other observations have values greater than $f(x_{l(n-k-1)}, \mu, S)$, so that they will lay outside the ellipsoid with center μ and scatter $R = S^{-1}$. It is clear that with probability 1 the upper $p + 1$ values will be equal. It is well known that the minimal volume ellipsoid around any simplex has an uniquely defined volume, i.e. value of the determinant of R, and that its center coincides with the barycenter of the vertices (see Vandev 1993). That is these $p + 1$ points will lay on the surface of this ellipsoid and it will have minimal volume. This shows that the estimator LME(k) of μ and S for the Gaussian law is, therefore constructed in the following way:

1. Find the Minimal Volume Ellipsoid containing not less than k data points. Denote its parameters by b and R:

$$E(b, R) \quad = \quad \{x : (x - b)'R(x - b) \leq 1\}.$$

2. The value of b is then the estimate of μ. The matrix qR is the estimate of S^{-1}, where q is a correction constant that gurantees asymptotic unbiasedness.

4.2 The MCD estimator

Theorem 1 of this paper and the result of Vandev (1993) applied to the Gaussian case imply that if one chooses k within the bounds, $(n + p + 1)/2 < k < n - p - 1$, then the LTE($k$) estimator has breakdown point not less than $(n - k)/n$. Note that in this case one should bound from below the eigenvalues of the covariance matrix in order to use the result of Vandev.

Let $\{x_1, x_2, ..., x_k\}$ be the subset of the observations over which the LTE(k) criterion is minimized. A simple argument, which we leave to the reader, shows that the minimum of

$$(2) \qquad J(\mu, R) = \sum_{j=1}^{k} f(x_j, \mu, R^{-1})$$

is reached when

$$\mu = \frac{1}{k}\sum_{j=1}^{k} x_j \qquad \text{and} \qquad R = \frac{1}{k}\sum_{j=1}^{k}(x_j - \mu)(x_j - \mu)'.$$

The minimum of $J(\mu, R)$ may be represented as

$$\min_{\mu,R} J(\mu, R) = k.p/2 + (\ln\det(R))/2 + (k.p/2)\ln(2\pi).$$

This shows that among all sets of observations of fixed size k the minimum of (2) is reached when $\det(R)$ is minimal. Then the estimator LTE(k) of μ and S based on the Gaussian law is found in the following way:

1. Find the k points with minimal determinant of R, where R denote their covariance matrix.

2. The estimate of μ is then the average of these points and the estimate of S is pR^{-1}/q, where q is the correction factor.

The correction constant q for both MVE and MCD estimators can be obtained in the same way as in Lopuhaa and Rousseeuw (1991).

Acknowledgements. The authors would like to thank Professor S.Morgenthaler for helpful comments improving the presentation. N.M. Neykov would also like to thank to the Swiss National Science Foundation for sponsoring his participation in the Workshop on Data Analysis and Robustness, June 29 to July 3, 1992, held at the Centro Stefano Franscini, Monte Verita, Ascona, Switzerland.

References

[1] Hampel, F.R. Ronchetty, E.M. Rousseeuw, P.J. and Stahel, W.A.(1986). *Robust Statistics: The Approach Based on influence Functions*. John Wiley and Sons, New York.

[2] Huber, P.J.(1981). *Robust Statistics*. John Wiley and Sons, New York.

[3] Lopuhaa, H.P. and Rousseeuw, P.J.(1991). "Breakdown Point of Affine Equivariant Estimators of Multivariate Location and Covariance Matrix", *Ann. Statist.* vol.19, pp. 229-248.

[4] Neykov, N.M. and Neytchev, P.N. (1990). A Robust Alternative of the ML estimators. Proc. COMPSTAT'90, Short communications, Dubrovnik, Yugoslavia, pp.99-100.

[5] Rousseeuw, P.J.(1984). Least Median of Squares Regression, *J. Am. Statist. Assoc.*, vol 79., pp.871-880.

[6] Rousseeuw,P.J. (1986). "Multivariate Estimation with High Breakdown Point". In: *Mathematical Statistics and Applications* Vol. B, W.Grossmann, G.Pflug, I.Vincze, and W.Wertz (eds.), Dordrecht: Reidel Publishing Company, 283-297.

[7] Rousseeuw,P.J. and Leroy,A.(1987). *Robust Regression and Outlier Detection*, John Wiley and Sons, New York.

[8] Vandev D.L.(1992). "A Minimal Volume Ellipsoid around a Simplex", *C. R. Acad. Bulgare Sci.*, 45, (6), pp.37-40.

[9] Vandev D.L.(1993). "A Note on Breakdown Point of the Least Median of Squares and Least Trimmed Estimators", *Statistical and Probability Letters*, 16, pp.117-119.