Statistica Sinica **6**(1996), 367-374

CROSS-CHECKING USING THE MINIMUM VOLUME ELLIPSOID ESTIMATOR

Xuming He and Gang Wang

University of Illinois and Depaul University

Abstract. We show that for a wide class of elliptic models the minimum volume ellipsoid estimator is strongly consistent and the estimating functional is continuous with respect to a weak metric. We also propose to compute an efficient estimator cross-checked by the minimum volume ellipsoid estimator. The former is taken if both estimators stay close to each other based on an affine invariant discrepancy measure. Otherwise, a high breakdown point procedure is called for. This allows us to retain good efficiency for uncontaminated data and at the same time protect against gross errors.

Key words and phrases: Breakdown point, strong consistency, efficiency, elliptic distribution, multivariate location and scatter.

1. Introduction

The need for good robust estimators of multivariate location and scatter has stimulated a great deal of interest, as they are important for identifying multiple outliers, obtaining trustworthy parameter estimates, making stable inferences and carrying out further data analyses. It is fair to say that multivariate location and scatter are cornerstones of general multivariate statistics. For example, Campbell (1980) and Devlin, Gnanadesikan and Kettening (1981) proposed to robustify principle component analysis based on a robust location-scatter estimate. It is also the basis for robust canonical variate analysis, factor analysis and cluster analysis. For bounded influence and high breakdown point estimation of multiple regression, robust versions of location-scatter in the design space are useful (see Simpson, Ruppert and Carroll (1992)). A prominent affine-equivariant high breakdown point estimator is the minimum volume ellipsoid estimator (MVE) introduced by Rousseeuw (1985). The MVE is usually applied as a safe starting point for further analysis or a reliable basis on which robust distances can be computed for diagnostics.

The first part of the present paper is concerned with some basic properties of the MVE estimator. Davies (1992b) proved that the MVE functional satisfies a local Hölder condition of order 1/2 and also obtained a non-Gaussian limiting distribution of the estimator, confirming a long standing conjecture that the rate

of convergence for the MVE estimator is $n^{-1/3}$. We establish strong consistency and functional continuity of the MVE estimator for a broader class of elliptically symmetric models. These properties indicate that the estimator should perform reasonably well if the underlying distribution is close to an elliptically symmetric one. If elliptical symmetry does not hold to any reasonable extent then an estimate of the scatter matrix would not be very meaningful after all.

A major criticism of the minimum volume ellipsoid estimator is its inefficiency and large variability. It is recognized that the slow rate of convergence has practical implications. A common belief is that some appropriate follow-up can lead to estimators with good efficiency while retaining high breakdown. Reweighting was suggested by Rousseeuw and Leroy (1987) and again by Rousseeuw and van Zomaren (1990). However, recent work of He and Portnoy (1992) indicates that reweighting does not necessarily improve on the rate of convergence. Furthermore, blindly applying a high breakdown point estimator (even if it has a good asymptotic efficiency) could entail the risk of (finite-sample) efficiency loss for uncontaminated samples. In this article, we propose a method of cross-checking using two estimators: an efficient one such as the classical sample mean-covariance (or an M-estimator) as well as the MVE (or any other high breakdown point estimator). We use an affine invariant measure of discrepancy for these two estimators to tell us whether the former is being badly influenced by outliers. The efficient one is taken if both estimators stay close to each other. A high breakdown point procedure is called for when there is evidence that the former has been distorted. With proper choices of the cut-off values for our discrepancy numbers, we are assured that the chance of false alarm is very small. The main advantage of such a cross-checking procedure is that it allows us to avoid much of the efficiency loss for "good" data and at the same time protect against gross errors. The general idea of cross-checking is given in Section 3 where it is shown that the resulting estimator is asymptotically equivalent to the more efficient estimator in the model and inherits the high breakdown point of the other. A specific proposal is given as to how the discrepancies can be measured , along with an illustration using the sample mean-covariance and the MVE estimator for the trivariate normal model. The same procedure is then applied to the well-known stackloss data (see Rousseeuw and Leroy (1987)). Proofs of all theorems of this paper are available in a more detailed technical report of He and Wang (1992).

2. Consistency and Continuity of MVE

Suppose that we have a sample $X_1, \ldots, X_n$ in p dimensions and want to estimate its "center" and "scatter" by a p dimensional vector t and a $p \times p$ matrix C. All matrices in this paper are assumed to be symmetric and positive definite unless otherwise stated. The most common estimators are the sample mean

vector and sample covariance matrix. They enjoy optimality in various senses if the data are believed to come from a p-variate normal distribution. However, these estimators are quite sensitive to outliers in the sample. Rousseeuw (1985) introduced a highly robust estimator, the minimum volume ellipsoid estimator (MVE) (t_n, C_n), where t_n is taken to be the center of the minimum volume ellipsoid covering at least half of the observations, and C_n is a p by p matrix representing the shape of the ellipsoid. In the present paper, we use $|S|$ and $\|S\|$ to denote the determinant and the L_2 norm of any square matrix S. Also, define $E(a, S) = \{x : \ (x-a)'S^{-1}(x-a) \le 1\}$ to be the ellipsoid centered at a with scatter matrix S and "radius" 1. Then, the pair (t_n, C_n) is determined by minimization of $|C_n|$ subject to

$$\#\{i; \ X_i \in E(t_n, C_n)\} \ge [(n+1)/2]. \tag{2.1}$$

In some recent literature, the right hand side of (2.1) is often replaced by $[(n+p+1)/2]$ for a maximum breakdown point. We use (2.1) in this article, but all our results apply equally well to the variant mentioned above.

The main merit of the MVE is that it has a high breakdown point close to 1/2 (see Rousseeuw and Leroy (1987)). The MVE is arguably one of the simplest high breakdown point estimators conceptually and computationally, even though it is not computationally easy (see Rousseeuw and van Zomeren (1990) for approximate algorithms). To study the continuity property of the MVE, we need to rely on the MVE functional and a topology on the space of all probability distributions.

Given any distribution G, the MVE functional $T(G) = (t_G, C_G)$ is given by minimizing $|C|$ among all pairs (t, C) such that $P_G(E(t, C)) \ge 1/2$, where P_G is the probability measure induced by G. For unimodal and elliptic distributions, the existence and uniqueness of the MVE functional are obvious. More important is its behavior in a metric neighborhood of such distributions. For two distributions F_1, F_2, define

$$d(F_1, F_2) = \inf\Big\{\eta > 0 : P_{F_1}(E(a,S)) \le P_{F_2}(E(a,(1+\eta)S)) + \eta, \\ P_{F_2}(E(a,S)) \le P_{F_1}(E(a,(1+\eta)S)) + \eta, \text{ for all } (a,S)\Big\}.$$

This metric, in an equivalent form, was first used in Davies (1992b). It is affine invariant and coarser than the usual total variation metric.

Suppose that the distribution $F = F(\mu, \Sigma)$ is elliptically symmetric with probability density function

$$f(x; \mu, \Sigma) = |\Sigma|^{-1/2} g((x-\mu)'\Sigma^{-1}(x-\mu)), \tag{2.2}$$

where g is a univariate function. In the specification of (2.2), Σ is identifiable only up to a multiplicative constant. In the rest of the section, we assume that Σ is so chosen that $P(E(\mu, \Sigma)) = 1/2$.

We now introduce a technical condition for convenience. Consider any elliptic distribution in the form of (2.2). Let $\mathcal{D} = \mathcal{D}(\gamma, \delta)$ denote the collection of all pairs (a, S) such that $0 < |S| \leq (1+\gamma)^p |\Sigma|$ and either $\|a-\mu\| > \delta$ or $\|S-\Sigma\| > \delta$.

Condition A. *For any* $\delta>0$, *there exist* $\gamma>0$ *and* $\varepsilon>0$ *such that* $\sup_{\mathcal{D}} P(E(a,S)) < P(E(\mu,\Sigma)) - \varepsilon$.

A more directly interpretable condition is

Condition B (uniqueness). (μ, Σ) *is the unique solution of minimizing* $|S|$ *subject to* $P(E(a,S)) = 1/2$.

It was shown by He and Wang (1992) that Conditions A and B are actually equivalent for any $F(\mu, \Sigma)$. The formulation of Condition A allows us to simplify the proofs of theorems stated below, as it carries a more transparent geometric interpretation. The Condition C below is also related to uniqueness, but it is independent of Condition A or B and cannot be removed in our Theorem 2.2.

Condition C. *For any* $c > 1$, $P(E(\mu, c\Sigma)) > 1/2$.

First, we establish existence of the MVE functional in a neighborhood of the model distribution.

Theorem 2.1. *Let* F *be any elliptic distribution of the form* (2.2). *Suppose that* $\{F_\varepsilon : \varepsilon > 0\}$ *is a family of distributions such that* $d(F, F_\varepsilon) < \varepsilon$. *Then there exists* $\varepsilon_0 > 0$ *such that for all* $\varepsilon < \varepsilon_0$, *there is at least one pair* $(t_\varepsilon, S_\varepsilon)$ *that minimizes* $|S|$ *among all pairs* (a, S) *with* $P_{F_\epsilon}(E(a,S)) \geq 1/2$.

Under Conditions A and C, the MVE functional changes continuously with the amount of contamination. Let $T(F_\epsilon)$ be any MVE functional at F_ϵ.

Theorem 2.2. *Under the Conditions* A *and* C, $T(F_\varepsilon) \to T(F)$ *as* $\epsilon \to 0$. *In particular, the MVE estimator* $(t_n, C_n) \to (\mu, \Sigma)$ *almost surely* (*with regard to* F) *as* $n \to \infty$.

In general, $T(F_\varepsilon)$ may not be unique. The result of Theorem 2.2 holds for every MVE functional. Under a stronger condition on g, that is, if g is nonincreasing and has non-zero derivative at 1, Davies (1992b) proved that the MVE functional is actually Hölder continuous of order exactly 1/2. By contrast, we do not require g to be monotone and positive everywhere.

3. Using MVE for Cross-Checking

The MVE estimator is known to have poor efficiency. Several methods of combining efficiency and high breakdown have been proposed and investigated in the literature. They include reweighting, one-step M-estimator and cross-checking. We refer to Rousseeuw and Leroy (1987), He and Portnoy (1992) and

Davies (1992a) for more details. As usual, we restrict ourselves to estimators that are affine equivariant.

For a typical user of statistics, an appealing method to obtain high breakdown point without giving up efficiency is to compute two estimators at the same time. The first is efficient such as the classical mean-covariance or the bounded influence GM-estimator. The other is a consistent high breakdown estimator like the MVE. If the two are close to each other, the former (and actually either) can be trusted. If the two are far apart, further investigation may be necessary. Quite likely, the first estimator has been distorted by outliers. We propose automatic cross-checking procedures using an affine equivariant discrepancy of two location-scatter estimators. A specific proposal is considered below.

Let $T_{n,1} = (t_{n,1}, C_{n,1})$ be a consistent estimator with high efficiency, and $T_{n,2} = (t_{n,2}, C_{n,2})$ a consistent high breakdown point estimator. For any positive definite matrix S, use $\lambda_{min}(S)$ and $\lambda_{max}(S)$ to denote its smallest and largest eigenvalues. We define the new estimator $T_n = T_{n,1}$ if $D_1 := (t_{n,1} - t_{n,2})' C_{n,2}^{-1}(t_{n,1} - t_{n,2}) \le d_1$, $D_2 := 1 - \lambda_{min}(C_{n,1})/\lambda_{min}(C_{n,2}) \le d_2$, and $D_3 := |\lambda_{max}(C_{n,1})/\lambda_{max}(C_{n,2}) - 1| \le d_3$ for some constants $d_1 > 0, 0 < d_2 < 1$ and $d_3 > 0$, and $T_n = T_{n,2}$ otherwise. The relationship among their functional versions is self-evident.

Theorem 3.1. *The estimator T_n defined above satisfies the following properties for any elliptic distribution.*
(1) *It is affine-equivariant if both $T_{n,i}$ $(i = 1, 2)$ are.*
(2) *It is asymptotically equivalent to $T_{n,1}$, that is, with probability one (with regard to F) there exists N such that $T_n = T_{n,1}$ for $n > N$.*
(3) *It has the same influence function as $T_{n,1}$ does, provided that the corresponding functionals $T_i((1-\epsilon)F+\epsilon\delta_x)$ $(i=1,2)$ are continuous in ϵ for each x.*
(4) *It has the same breakdown point as $T_{n,2}$ does.*

The tuning constants d_i in the definition of T_n can be determined by two considerations: how much the two estimators differ for a typical uncontaminated sample and how much deviation from a high breakdown point estimator one can tolerate. The purpose of Theorem 3.1 is to provide a sound basis for the methodology involved. A similar idea in the linear regression setup was given in He (1991).

The arbitrariness of those tuning constants is inconvenient in practice. For small to modest sample sizes, we consider using $d_i = \max\{q_i(n, \alpha_i), q_i(N_0, \alpha_i)\}$ where $q_i(n, \alpha)$ is the upper α-th quantile of the corresponding discrepancy statistic D_i, and N_0 is a fixed sample size which is taken to be 50 in our study. The quantity $q_i(n, \alpha)$ can be (approximately) computed for a model distribution (say normal) and each given n by Monte Carlo. When α is a small number (say 0.05 or 0.01), this choice of cut-off ensures that for uncontaminated data our cross-

checking will choose the efficient estimator most of the time. Note that $q_i(N_0, \alpha_i)$ is the amount of tolerance we have for deviation from a high breakdown point estimator when a larger sample size is available. The tolerance is set to be larger for small sample sizes. Also note that determining such cut-off values may take more than a few minutes, but they do not change with the sample. For the same model distribution, those cut-off values are computed once for all.

We now consider an example where $T_{n,1}$ is the sample mean and covariance matrix and $T_{n,2}$ the minimum volume ellipsoid estimator for trivariate normal models. We used an approximate algorithm for the MVE as described in Rousseeuw and Leroy (1986, p. 259) without using finite-sample correction factors. In this case (as in most other cases), there is no need to use the discrepancy number based on the smallest eigenvalues, because the scatter matrix estimate from $T_{n,1}$ does not break down to singularity even with 50% contamination. For a given n, we generated 200 trivariate normal samples of size n and computed both estimators. 1000 subsamples were used in the MVE computation. Two discrepancy numbers D_1 and D_3 were then computed at each sample. The 95 and 99 percentiles of these distance measures are taken as d_1 and d_3 respectively. If the 99 percentiles are used for both discrepancies, we get (d_1^*, d_3) in place of (d_1, d_3).

Table 1. Cut-off constants for trivariate normal model

n	15	18	20	30	40	50+
d_1	13.65	6.92	4.39	1.74	0.98	0.67
d_3	1.26	0.90	0.71	0.70	0.70	0.67
d_1^*	28.27	12.05	8.77	3.27	2.19	0.96

Because the discrepancy numbers D_1 and D_3 are highly correlated, our simulation study shows that with these choices of d_i's, the mean-covariance estimator will be chosen over 90% of the time for an uncontaminated normal sample. If d_1^* is taken, this percentage is higher. The error rate decreases with $n \geq 50$. With presence of outliers, we expect one or both D_i's to become large, forcing the high breakdown point estimator to be used. Our experiments indicate that the discrepancy based on the largest eigenvalues of the scatter matrix estimators is often most sensitive to outliers. We computed this discrepancy number for 200 standard trivariate normal samples of size $n = 30$ with the first component of two observations fixed at 10.0. It turned out that the cut-off value of $d_3 = 0.70$ was broken 95% of the time.

A replication of 200 times in any Monte Carlo calculation would be deemed very small. A larger number can be used for a more accurate determination of the tuning constants here. On the other hand, we find that the value of d_3, the most useful one for watching outliers, is surprisingly stable at uncontaminated

samples. Presence of serious outliers normally drive this discrepancy number to exceed 1.0, a value that is rarely observed in our experiment for uncontaminated data for n as small as 20 in the case of $p = 3$. This gives us reassurance that even a small number of Monte Carlo replications is useful.

For the well-studied stackloss data (see Rousseeuw and Leroy (1987, p. 76)), we computed the two discrepancy numbers when all 21 points are included. They turned out to be 0.94 and 1.64, clearly indicating that the classical mean-covariance estimate is being driven by outliers. If the first three points (outliers) are removed, these numbers become 9.78 and 0.36, suggesting that the mean estimate may have been shifted too much by outlier(s). In this case, the effect of the remaining outlier (point no. 21) is less serious. If we used d_1^* instead of d_1, our final estimate would pick the classical mean and covariance matrix.

By Theorem 3.1, the resulting estimator in this example is asymptotically (100%) efficient at the normal model and has a breakdown point nearly 1/2. If the mean-covariance estimator is replaced by a GM-estimator, one would get a bounded influence function at the cost of slight efficiency loss.

4. Concluding Remarks

The minimum volume ellipsoid estimator of multivariate location-scatter is not only consistent but also continuous with respect to an affine invariant weak metric for a large class of elliptically symmetric model distributions. These results are obtained independently of Davies (1992b) under weaker conditions. Our approach is largely based on a geometric characterization for the uniqueness of the MVE estimator. These properties are sufficient for the estimate to be used in the cross-checking method discussed in Section 3. It is worth noting that the method of cross-checking and our Theorem 3.1 to ensure robustness without giving up high efficiency is not limited to the minimum volume ellipsoid estimator. Our work does not imply in any way that the MVE is the only or the best high breakdown point estimator available today. S-estimators or reweighted versions of the MVE may be used to reduce the finite sample variability of the high breakdown estimator. In the latter case, the cross-checking has some similarity to the classical delete-k diagnostics. The targeted subset of k outliers is determined by the MVE estimator.

The cross-checking method is designed to provide information as to whether a classical procedure can be trusted. Each estimator captures some feature of the data better than others. The general idea here is to utilize information from different estimators. The specific strategy that we considered in this paper is merely one implementation that we have found quite successful. Our experiment is limited in scale, and further studies may give rise to better discrepancy statistics for the same purpose.

Acknowledgement

We benefited from helpful discussions with Laurie Davies, Peter Rousseeuw and Victor Yohai at the workshop on Data Analysis and Robustness at Ascona, Switzerland in the summer of 1992. The research of Gang Wang was partially supported by a summer research grant from DePaul University.

References

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. Appl. Statist. **29**, 231-237.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. Ann. Statist. **15**, 1269-1292.

Davies, P. L. (1990). The asymptotics of S-estimators in the linear regression model. Ann. Statist. **18**, 1651-1675.

Davies, P. L. (1992a). An efficient Fréchet differentiable high breakdown multivariate location and dispersion estimator. J. Multivariate. Anal. **40**, 311-327.

Davies, P. L. (1992b). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. Ann. Statist. **20**, 1828-1843.

Devlin, S. J., Gnanadesikan, R. and Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. J. Amer. Statist. Assoc. **76**, 354-362.

He, X. (1991). A local breakdown property of robust tests in linear regression. J. Multivariate. Anal. **38**, 294-305.

He, X. and Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator. Ann. Statist. **20**, 2161-2167.

He, X. and Wang, G. (1992). On properties and applicability of the minimum volume ellipsoid estimators. Technical Report, University of Illinois.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In Mathematical Statistics and Applications (Vol. B, Edited by W. Grossmann et al.), 283-297, Reidel Publishing.

Rousseeuw, P. J. and Leroy, A. M. (1987). Robust Regression and Outlier Detection. John Wiley, New York.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. J. Amer. Statist. Assoc. **85**, 633-639.

Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). On one-step GM estimates and stability of inferences in linear regression. J. Amer. Statist. Assoc. **87**, 439-450.

Department of Statistics, University of Illinois, Champaign, IL 61820, U.S.A.

Department of Mathematics, Depaul University, Chicago, IL 60614, U.S.A.

(Received April 1993; accepted April 1995)