Journal of Multivariate Analysis 101 (2010) 2372–2388

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Asymptotic expansion of the minimum covariance determinant estimators

Eric A. Cator, Hendrik P. Lopuhaä *

Delft University of Technology, The Netherlands

ARTICLE INFO

Article history:
Received 26 August 2009
Available online 23 June 2010

AMS 2008 subject classifications:
62G35
62E20
62H12

Keywords:
Minimum covariance determinant
Asymptotic normality
Influence function

ABSTRACT

In Cator and Lopuhaä (arXiv:math.ST/0907.0079) [3], an asymptotic expansion for the minimum covariance determinant (MCD) estimators is established in a very general framework. This expansion requires the existence and non-singularity of the derivative in a first-order Taylor expansion. In this paper, we prove the existence of this derivative for general multivariate distributions that have a density and provide an explicit expression, which can be used in practice to estimate limiting variances. Moreover, under suitable symmetry conditions on the density, we show that this derivative is non-singular. These symmetry conditions include the elliptically contoured multivariate location-scatter model, in which case we show that the MCD estimators of multivariate location and covariance are asymptotically equivalent to a sum of independent identically distributed vector and matrix valued random elements, respectively. This provides a proof of asymptotic normality and a precise description of the limiting covariance structure for the MCD estimators.

1. Introduction

The minimum covariance determinant (MCD) estimator [14] is one of the most popular robust methods to estimate multivariate location and scatter parameters. These estimators, in particular the covariance estimator, also serve as robust plug-ins in other multivariate statistical techniques, such as principal component analysis [5,16], multivariate linear regression [1,15], discriminant analysis [7], factor analysis [13], canonical correlations [17,18] and errors-in-variables models [6], among others (see also [8] for a more extensive overview). For this reason, the distributional and the robustness properties of the MCD estimators are essential for conducting inference and performing robust estimation in several statistical models.

The MCD estimators have the same high breakdown point as the minimum volume ellipsoid estimators (e.g., see [1,11]). The asymptotic properties were first studied by Butler, Davies and Jhun [2] in the framework of unimodal elliptically contoured densities; they showed that the MCD location estimator converges at $\sqrt{n}$-rate towards a normal distribution with mean equal to the MCD location functional. In the same framework, Croux and Haesbroeck [4] give the expression for the influence function of the MCD covariance functional and use this to compute limiting variances of the MCD covariance estimator. The asymptotic theory was extended and generalized by Cator and Lopuhaä [3], who studied the MCD estimators and the corresponding functional in a very general framework. They establish an asymptotic expansion of the type

$$\widehat{\theta}_n - \theta_0 = -\Lambda'(\theta_0)^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(\Psi(X_i,\theta_0) - \mathbb{E}\Psi(X_i,\theta_0)\right) + o_{\mathbb{P}}(n^{-1/2}), \tag{1.1}$$

* Corresponding address: Delft Institute of Applied Mathematics, TU Delft, Mekelweg 4, 2628 CD Delft, The Netherlands.
E-mail address: h.p.lopuhaa@tudelft.nl (H.P. Lopuhaä).
URL: http://ssor.twi.tudelft.nl/~lopuhaa/ (H.P. Lopuhaä).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2010.06.009

where $\widehat{\theta}_n$ and θ_0 denote vectors consisting of the MCD estimators and the MCD functional at the underlying distribution, respectively, and $\Psi(\cdot, \theta_0)$ is a function that we will specify later on. In principle, from this expansion a central limit theorem for the MCD estimator can be derived. However, the expansion requires the existence and non-singularity of $\Lambda'(\theta_0)$. Moreover, a more explicit expression of its inverse is desirable from a practical point of view, since it determines the limiting variances.

In this paper we show that $\Lambda'(\theta_0)$ exists as long as the underlying distribution P has a density f. Moreover, we provide an explicit expression for $\Lambda'(\theta_0)$ in Theorem 3.1. The expression offers the possibility to estimate the limiting variances of the MCD estimators in *any* model where P has a density. This extends the applicability of the MCD estimator far beyond elliptically contoured models. We will also provide sufficient symmetry conditions on f for $\Lambda'(\theta_0)$ to be non-singular. This includes the special case of elliptically contoured densities

$$f(x) = \det(\Sigma)^{-1/2} h((x-\mu)\Sigma^{-1}(x-\mu)),$$

for which we show that the MCD location and the MCD covariance estimator are asymptotically equivalent to a sum of independent vector and matrix valued random elements, respectively. This exact expansion shows that at elliptically contoured densities the MCD location and MCD covariance estimator are asymptotically independent and yields an explicit central limit theorem for both MCD estimators separately, in such a way that the limiting covariances between elements of the location and covariance estimators can be obtained directly from the covariances between elements of the summands. Furthermore, the expansion for the MCD estimators is needed to obtain the limiting distribution of robustly reweighted least squares estimators for (μ, Σ), if one uses the MCD estimators to assign the weights (see [12]).

The paper is organized as follows. In Section 2, we define the MCD estimators and MCD functionals and discuss some results from [3] that are relevant for our setup. In Section 3, we establish the expression for $\Lambda'(\theta_0)$ in terms of a linear mapping and show that this mapping is non-singular under suitable symmetry conditions. The special case of elliptically contoured densities is considered in Section 4, where we obtain an explicit expression of $\Lambda'(\theta_0)^{-1}$. From this we derive an asymptotic expansion for the estimators, prove asymptotic normality, and derive the influence function of the MCD functionals. As special cases we recover results from [2,4] under weaker conditions.

All proofs have been postponed to an Appendix at the end of the paper.

2. Definition and preliminaries

For a sample $X_1, X_2, \ldots, X_n$ from a distribution P on $\mathbb{R}^k$, the MCD estimator is defined as follows. Fix a fraction $0 < \gamma \leq 1$ and consider subsamples $S \subset \{X_1, \ldots, X_n\}$ that contain $h_n \geq \lceil n\gamma \rceil$ points. Define a corresponding trimmed sample mean and sample covariance matrix by

$$\begin{aligned} \widehat{T}_n(S) &= \frac{1}{h_n} \sum_{X_i \in S} X_i, \\ \widehat{C}_n(S) &= \frac{1}{h_n} \sum_{X_i \in S} (X_i - \widehat{T}_n(S))(X_i - \widehat{T}_n(S))'. \end{aligned} \tag{2.1}$$

Note that each subsample S determines an ellipsoid $E(\widehat{T}_n(S), \widehat{C}_n(S), \widehat{r}_n(S))$, where, for each $\mu \in \mathbb{R}^k$, Σ symmetric positive definite, and $\rho > 0$,

$$E(\mu, \Sigma, \rho) = \left\{ x \in \mathbb{R}^k : (x-\mu)'\Sigma^{-1}(x-\mu) \leq \rho^2 \right\}, \tag{2.2}$$

and

$$\widehat{r}_n(S) = \inf\left\{ s > 0 : P_n\left(E(\widehat{T}_n(S), \widehat{C}_n(S), s) \right) \geq \gamma \right\}, \tag{2.3}$$

where P_n denotes the empirical measure corresponding to the sample. Let S_n be a subsample that minimizes $\det(\widehat{C}_n(S))$ over all subsamples of size $h_n \geq \lceil n\gamma \rceil$; then the pair $(\widehat{T}_n(S_n), \widehat{C}_n(S_n))$ is an MCD estimator. Note that a minimizing subsample always exists, but it need not be unique. In [3], it is shown that a minimizing subsample S_n always has exactly $\lceil n\gamma \rceil$ points and is contained in the ellipsoid $E(\widehat{T}_n(S_n), \widehat{C}_n(S_n), \widehat{r}_n(S_n))$, which separates S_n from all other points in the sample. Note that in [2] (among others) one minimizes over subsamples of size $\lfloor n\gamma \rfloor$. This is somewhat unnatural, since it may lead to subsamples S for which $P_n(S) < \gamma$. Moreover, it may lead to situations where the trimmed subsample does not contain the majority of the points; for example, if $\gamma = 1/2$ and n is odd, then $\lfloor n\gamma \rfloor = (n-1)/2$. By considering subsamples S of size $h_n \geq \lceil n\gamma \rceil$ in definition (2.1), we always have $P_n(S) \geq \gamma$, and for any $1/2 \leq \gamma \leq 1$, the subsample contains the majority of points.

We define the MCD functionals in a similar fashion. Define a trimmed mean and covariance as follows:

$$\begin{aligned} T_P(\phi) &= \frac{1}{\int \phi \, \mathrm{d}P} \int x\phi(x) P(\mathrm{d}x), \\ C_P(\phi) &= \frac{1}{\int \phi \, \mathrm{d}P} \int (x - T_P(\phi))(x - T_P(\phi))' \phi(x) P(\mathrm{d}x) \end{aligned} \tag{2.4}$$

and define

$$r_P(\phi) = \inf\{ s > 0 : P\left(E(T_P(\phi), C_P(\phi), s)\right) \geq \gamma \}$$