Computational Statistics and Data Analysis 81 (2015) 64–75

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

The DetS and DetMM estimators for multivariate location and scatter

Mia Hubert *, Peter Rousseeuw, Dina Vanpaemel, Tim Verdonck

KU Leuven, Department of Mathematics and LStat, Celestijnenlaan 200B, BE-3001 Heverlee, Belgium

ARTICLE INFO

Article history:
Received 11 October 2013
Received in revised form 19 July 2014
Accepted 22 July 2014
Available online 1 August 2014

Keywords:
Covariance matrix
Fast algorithm
Outliers
Robust estimation

ABSTRACT

New deterministic robust estimators of multivariate location and scatter are presented. They combine ideas from the deterministic DetMCD estimator with steps from the subsampling-based FastS and FastMM algorithms. The new DetS and DetMM estimators perform similarly to FastS and FastMM on low-dimensional data, whereas in high dimensions they are more robust. Their computation time is much lower than FastS and FastMM, which allows to compute the estimators for a range of breakdown values. Moreover, they are permutation invariant and very close to affine equivariant.

1. Introduction

The need for robust multivariate estimators that can withstand a substantial amount of contamination is well-recognized nowadays. Examples of such high-breakdown estimators of location and scatter are the MVE and MCD estimators (Rousseeuw, 1984), the Stahel–Donoho estimator (Stahel, 1981; Donoho, 1982), S-estimators (Rousseeuw and Yohai, 1984; Davies, 1987; Rousseeuw and Leroy, 1987; Lopuhaä, 1989) and MM-estimators (Yohai, 1987). They are all highly robust with breakdown value up to 50%, but they differ in terms of efficiency, bias and outlier resistance. For a recent comparison, see Hubert et al. (2014).

The computation of these estimators is challenging as their objective function is not convex and usually has several local minima. To address this problem, the current algorithms start by drawing many random subsets and then iteratively apply easy-to-compute steps that are guaranteed to decrease the objective function. Examples are FastMCD (Rousseeuw and Van Driessen, 1999), FastS (Salibian-Barrera and Yohai, 2006) and FastMM (Salibian-Barrera et al., 2006). A disadvantage of such a random sampling approach is the lack of permutation invariance: listing the observations in a different order might result in different estimates. In Hubert et al. (2012) a deterministic estimator has been proposed which turns out to be highly robust against outliers and which is very fast even in higher dimensions. As this method uses ideas from the FastMCD algorithm, it was named DetMCD.

Here we propose two estimators which are also deterministic, robust and fast, and which are inspired by the objective functions of the S and MM-estimators. For a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$, a multivariate S-estimator of location and scatter is defined as the couple $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ which minimizes $|\boldsymbol{S}|$ under the condition

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\sqrt{(\boldsymbol{x}_i - \boldsymbol{m})^t \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{m})}\right) = b \tag{1}$$

* Corresponding author. Tel.: +32 16 322023.
E-mail address: Mia.Hubert@wis.kuleuven.be (M. Hubert).
URL: http://wis.kuleuven.be/stat/robust (M. Hubert).

http://dx.doi.org/10.1016/j.csda.2014.07.013
0167-9473/

over all $(\boldsymbol{m}, \boldsymbol{S})$ where $\boldsymbol{m} \in \mathbb{R}^p$ and $\boldsymbol{S}$ is a $p \times p$ symmetric positive definite (SPD) matrix. In order to obtain positive breakdown estimates, ρ should satisfy the following conditions:

(C1) ρ is symmetric around zero and twice continuously differentiable
(C2) ρ is strictly increasing on $[0, c_0]$ for some $c_0 > 0$, constant on $[c_0, \infty[$, and $\rho(0) = 0$.

The constant b can be computed as $E_{F_0}[\rho(\|\boldsymbol{Z}\|)]$ where $F_0 = N(\boldsymbol{0}, \boldsymbol{I}_p)$ to ensure consistency at the normal model. For ρ one often chooses the function

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c^2} + \dfrac{x^6}{6c^4} & \text{for } |x| \leqslant c \\ \dfrac{c^2}{6} & \text{for } |x| > c \end{cases} \tag{2}$$

where c is an appropriate tuning constant. This is known as Tukey's bisquare ρ function, as indeed its derivative ψ contains two squares:

$$\rho'(x) = \psi(x) = \begin{cases} x\left(1 - \left(\dfrac{x}{c}\right)^2\right)^2 & \text{for } |x| \leqslant c \\ 0 & \text{for } |x| > c. \end{cases}$$

We will use this ρ-function throughout the paper. An alternative choice is the class of translated bisquare ρ-functions introduced by Rocke (1996).

Lopuhaä and Rousseeuw (1991) showed that the breakdown value of a multivariate S-estimator is $b/\rho(c)$. Under the normal model b can be computed as (Campbell et al., 1998):

$$b = \frac{p}{2}\chi^2_{p+2}(c^2) - \frac{p(p+2)}{2c^2}\chi^2_{p+4}(c^2) + \frac{p(p+2)(p+4)}{6c^4}\chi^2_{p+6}(c^2) + \frac{c^2}{6}(1 - \chi^2_p(c^2))$$

where χ^2_ν is the cdf of the χ^2 with ν degrees of freedom. For a given breakdown value between 0% and 50% we can derive the value of the corresponding tuning parameter c in (2), see Table 3 in Rousseeuw and Yohai (1984).

Subsampling-based algorithms for multivariate S-estimators were proposed by Ruppert (1992) and Campbell et al. (1998). Next Salibian-Barrera and Yohai (2006) developed the FastS algorithm for regression S-estimators, which was extended to multivariate S-estimators for location and scatter in Salibian-Barrera et al. (2006). Section 2 describes the multivariate FastS algorithm.

The MCD estimator, the FastMCD algorithm and the DetMCD estimator are explained in Section 3. In Section 4 we combine ideas from DetMCD and FastS, yielding our proposed deterministic estimator DetS. We also explain how this yields DetMM, a deterministic estimator related to the MM-estimator. Section 5 studies the performance of DetS by simulation. The DetS and DetMM methods are applied to real data in Section 6, and their permutation invariance and near-affine equivariance are studied in Section 7. Section 8 concludes with suggestions for future research.

2. The FastS algorithm for multivariate location and scatter

We start by laying out the main idea of the FastS algorithm. First, the $\boldsymbol{S}$ in (1) is written as $\sigma^2\boldsymbol{\Gamma}$ with $|\boldsymbol{\Gamma}| = 1$ and $\sigma = |\boldsymbol{\Sigma}|^{1/(2p)}$, so that the equivalent objective is to find the triplet $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Gamma}}, \hat{\sigma})$ that minimizes s under the restriction

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{\sqrt{(\boldsymbol{x}_i - \boldsymbol{m})^t\boldsymbol{G}^{-1}(\boldsymbol{x}_i - \boldsymbol{m})}}{s}\right) = b \tag{3}$$

over all $(\boldsymbol{m}, \boldsymbol{G}, s)$ where $\boldsymbol{m} \in \mathbb{R}^p$, $\boldsymbol{G}$ is a $p \times p$ SPD matrix with $|\boldsymbol{G}| = 1$ and s is a positive scalar. The location and scatter estimates are then $(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2\hat{\boldsymbol{\Gamma}})$.

The algorithm starts with N initial estimates $(\hat{\boldsymbol{\mu}}_1^{(0)}, \hat{\boldsymbol{\Gamma}}_1^{(0)}, \hat{\sigma}_1^{(0)}), \ldots, (\hat{\boldsymbol{\mu}}_N^{(0)}, \hat{\boldsymbol{\Gamma}}_N^{(0)}, \hat{\sigma}_N^{(0)})$ obtained by drawing N random subsets of size $p + 1$ that have a covariance matrix with non-zero determinant (up to numerical precision), and calculating the classical mean $\hat{\boldsymbol{\mu}}_l^{(0)}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_l^{(0)}$ of the lth subset. Then we set $\hat{\boldsymbol{\Gamma}}_l^{(0)} = |\hat{\boldsymbol{\Sigma}}_l^{(0)}|^{-1/p}\hat{\boldsymbol{\Sigma}}_l^{(0)}$ and $\hat{\sigma}_l^{(0)} = \text{med}_{i=1}^n \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_l^{(0)})^t(\hat{\boldsymbol{\Gamma}}_l^{(0)})^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_l^{(0)})}$ for all $l = 1, \ldots, N$. Next those estimates are refined by performing k so-called I-steps, resulting in

$$(\hat{\boldsymbol{\mu}}_1^{(k)}, \hat{\boldsymbol{\Gamma}}_1^{(k)}, \hat{\sigma}_1^{(k)}), \ldots, (\hat{\boldsymbol{\mu}}_N^{(k)}, \hat{\boldsymbol{\Gamma}}_N^{(k)}, \hat{\sigma}_N^{(k)}).$$

The jth I-step to refine the estimate $(\hat{\boldsymbol{\mu}}_l^{(j-1)}, \hat{\boldsymbol{\Gamma}}_l^{(j-1)}, \hat{\sigma}_l^{(j-1)})$ goes as follows:

1. Refine the scale:

$$\hat{\sigma}_l^{(j)} = \hat{\sigma}_l^{(j-1)} \sqrt{\frac{1}{nb}\sum_{i=1}^{n} \rho\left(\frac{\sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_l^{(j-1)})^t(\hat{\boldsymbol{\Gamma}}_l^{(j-1)})^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_l^{(j-1)})}}{\hat{\sigma}_l^{(j-1)}}\right)}.$$