Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Q1 Robust and efficient estimation of multivariate scatter and location

Q2 Ricardo A. Maronna [a,*], Victor J. Yohai [b]

[a] *Department of Mathematics, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Argentina*

[b] *Department of Mathematics, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Argentina*

ARTICLE INFO

Article history:
Received 23 March 2016
Received in revised form 6 November 2016
Accepted 7 November 2016
Available online xxxx

Keywords:
MM-estimator
τ-estimator
S-estimator
Stahel–Donoho estimator
Kullback–Leibler divergence

ABSTRACT

Several equivariant estimators of multivariate location and scatter are studied, which are highly robust, have a controllable finite-sample efficiency and are computationally feasible in large dimensions. The most frequently employed estimators are not quite satisfactory in this respect. The Minimum Volume Ellipsoid (MVE) and the Minimum Covariance Determinant (MCD) estimators are known to have a very low efficiency. S-estimators with a monotonic weight function like the bisquare have a low efficiency when the dimension p is small, and their efficiency tends to one with increasing p. Unfortunately, this advantage is outweighed by a serious loss in robustness for large p. Four families of estimators with controllable efficiencies whose performance for moderate to large p has not been explored to date are studied: S-estimators with a non-monotonic weight function, MM-estimators, τ-estimators, and the Stahel–Donoho estimator. Two types of starting estimators are employed: the MVE computed through subsampling, and a semi-deterministic procedure previously proposed for outlier detection, based on the projections with maximum and minimum kurtosis. A simulation study shows that an S-estimator with non-monotonic weight function can simultaneously attain high efficiency and high robustness for $p \geq 15$, while an MM-estimator with a particular weight function can be recommended for $p < 15$. For both recommended estimators, the initial values are given by the semi-deterministic procedure mentioned above.

1. Introduction

Sample means and covariances are basic elements of most procedures in multivariate analysis. They are the maximum likelihood estimators for multivariate normal data. It is known, however, that even a small proportion of atypical observations may seriously affect them. A large number of approaches have been proposed – beginning with Gnanadesikan and Kettenring (1972) – to replace the sample mean vector and covariance matrix by a "location vector" and "scatter matrix" that are "robust", in the sense that they are not seriously affected by a small proportion of atypical observations. At the same time, it is desired that they are "efficient" in the sense that for normal data they should be near the classical sample mean and covariance matrix.

Both robustness and efficiency can be measured in different ways. The simplest robustness measure is the breakdown point which, although important, does not give enough information about the estimator's behavior under contamination. A

[*] Correspondence to: Departamento de Matemática, Facultad de Ciencias Exactas, C.C. 172, La Plata 1900, Argentina.
E-mail address: rmaronna@retina.ar (R.A. Maronna).

http://dx.doi.org/10.1016/j.csda.2016.11.006
0167-9473/© 2016 Elsevier B.V. All rights reserved.

more informative measure is the maximum bias under contamination; see Maronna et al. (2006) for details. The efficiency of scatter matrices can also be measured in several ways. The existence of different estimators poses the problem of choosing the most adequate ones. Several criteria must be taken into account for this purpose. The first two are naturally the robustness and the efficiency. It will be argued that they are not independent of each other; more precisely, that unless the number p of variables is small, in order to ensure an adequate degree of robustness it is necessary to control the efficiency. Controlling the efficiency is also necessary to make the different estimators comparable.

Another criterion, which is especially important for large p is the computing time. Most robust estimators are computed iteratively starting from some initial estimator, whose computation is the most time-consuming part of the procedure. For all affine-equivariant estimators, the initial estimator requires some form of subsampling, and the number of subsamples required to attain a desirable degree of robustness grows rapidly with p. It is therefore desirable to find alternatives to the usual forms of subsampling.

Therefore, the authors' goal is to select the most satisfactory estimators of multivariate location and scatter, in the sense that they (a) are highly robust, in the sense of having not only a high breakdown point but also a relatively low contamination bias; (b) have a controllable finite-sample efficiency, and (c) are computationally feasible for large p.

The most frequently employed estimators do not combine efficiency and robustness. The Minimum Volume Ellipsoid estimator (MVE) (Rousseeuw, 1985) is highly bias-robust, but has a very low efficiency: namely, its asymptotic efficiency is zero. The Minimum Covariance Determinant (MCD) estimator also has low efficiency, unless the user accepts a very low breakdown point. S-Estimators (Davies, 1987) with a monotonic weight function like the bisquare have a low efficiency for small dimension p. Rocke (1996) – and more generally Tyler (1994) – showed that their efficiency tends to one with increasing p. Unfortunately, this advantage is outweighed by a serious increase of the contamination bias for large p.

There are several families of estimators that might possess properties (a)–(b)–(c) above. But although their asymptotic properties have been studied theoretically, little is known about their finite-sample behavior and implementation. Four families of estimators, all of which are known to fulfill goals (a) and (b) are chosen for this study : S-estimators with a non-monotonic weight function which depends on p (Rocke, 1996); *MM*-estimators (Tatsuoka and Tyler, 2000); τ-estimators (Lopuhaä, 1991); and the estimator proposed by Stahel (1981) and Donoho (1982). All of them can be tuned to control efficiency without affecting the breakdown point. More details about them are given in the next section.

To help achieve the above goals, two elements are introduced, which although not new, have not been used before in this context, The first one concerns the starting values, which strongly affect the performance of an estimator. The subsampling approach usually employed for computing the starting values is very expensive for large dimensions. This study demonstrates that a semi-deterministic equivariant procedure, initially proposed by Peña and Prieto (2007) for outlier detection, dramatically improves both the computing times and the statistical performances of the estimators. The second one concerns the loss function ρ. Besides the popular bisquare ρ, a ρ-function is introduced that has been shown to possess certain favorable properties for time series (Muler and Yohai, 2002). It is demonstrated that it outperforms the bisquare function in most cases.

This study only considers equivariant estimators. There exist many non-equivariant proposals (see e.g. Hubert et al., 2015, and references therein); but the comparison between equivariant and non-equivariant estimators is difficult. In particular, a non-equivariant estimator is more difficult to tune for a given efficiency, since the latter depends on the correlations.

An extensive simulation study shows that the Rocke estimator, with adequate choices for ρ, tuning and starting values, outperforms its competitors for dimensions $\geq$ 15, and the same can be said of the MM estimator for dimensions $<$15.

Summing up, the contribution of this article is to study several estimators of multivariate location and scatter, to calculate the tuning constants that allow to control their efficiency, to introduce a non-standard loss function which improves on their performances, to introduce a non-standard initial estimator that yields both lower computing times and a better statistical performance, and to give specific recommendations for the choice of an estimator.

The contents of the paper are as follows. Section 2 reviews the estimators to be considered in this study. In Section 3 we discuss the choice of the ρ-function for MM- and τ-estimators. Section 4 deals with computational details. In Section 5 the estimators are compared through a simulation study. In Section 6 the estimators are applied to a real dataset. Finally Section 7 summarizes results. An Appendix is available online as Supplementary material (see Appendix A). It contains approximations for the tuning constants, results on the robustness of the Peña–Prieto procedure, details on the computation of the MVE and the full results of the simulations.

2. Review of estimators

This section describes the three families of estimators that will be considered in this study. The first are M- and MM-estimators, which are based on the minimization of a loss function of the Mahalanobis distances. They can be seen as weighted classical estimators with weights depending on the Mahalanobis distances. The second are the estimators based on the minimization of a robust scale of Mahalanobis distances. If the scale is "smooth", these estimators satisfy the same type of estimating equations as M-estimators, and can therefore also be seen as weighted classical estimators. The third type is the Stahel–Donoho estimator, which is a weighted classical estimator with weights depending on an "outlyingness" measure; but unlike the two former families, here the outlyingness depends on the one-dimensional data projections instead of the Mahalanobis distances.

Please cite this article in press as: Maronna, R.A., Yohai, V.J., Robust and efficient estimation of multivariate scatter and location. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.11.006