Choosing a Multivariate Estimate for High Dimensional Data

R.A. Maronna[1], V.J. Yohai[2] and A.J. Villar[2]

[1] Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, C.C. 172, La Plata 1900, Argentina
[2] Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 1, 1426 Buenos Aires, Argentina

Keywords: S estimates, Robust multivariate estimation.

Abstact

It is known that the efficiency at the normal of M estimates of multivariate location and scatter increases with the dimension p. This fact applies for estimates based on the minimization of a smooth M-scale ("S estimates").

Recall that an M-scale of the data set $z = \{z_1, ..., z_n\}$ is the solution $\sigma = \sigma(z)$ of ave $\{\rho(z/\sigma)\} = \delta$, where ρ is nondecreasing, $\rho(0) = 0$ and $\rho(\infty) = 1$, and $\delta \in (0, 1)$. The median corresponds to $\rho(z) = \mathrm{I}(z > 1)$ and $\delta = 0.5$, where I(.) is the indicator function. An S estimate $\left(\widehat{\mu}, \widehat{\mathbf{\Sigma}}\right)$ of the $p-$dimensional data set $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ is the solution of $\sigma(d_1, ..., d_n) = \min$, where $d_i = (\mathbf{x}_i - \widehat{\mu})' \, \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_i - \widehat{\mu})$ and $\left|\widehat{\mathbf{\Sigma}}\right| = 1$. S estimates can be expressed as weighted means and covariances with weights $w_i = W(d_i)$, where $W = \rho'$.

It is shown that when applying an S estimate with continuous W to a large normal sample with large p, all observations have approximately the same weights. This implies that the efficiency tends to one when $p \to \infty$; *but* also that the asymptotic bias in contamination neighborhoods must increase.

Rocke (1996) realized this fact and proposed that ρ depend on p. He defined a "translated bisquare" family of functions. It seems that his proposal had only limited diffusion, although it has been implemented in S-Plus.

We reconsider this approach from another point of view. Adrover and Yohai (2002) showed that the maximum asymptotic bias of the MVE is remarkably constant for large p, and much lower than that of the MCD and other robust estimates. To obtain an estimate with the good bias behavior of the MVE but without its inefficiency, we propose a family of $\rho-$functions such that for large p, it approaches the step $\rho-$function corresponding to the MVE. The asymptotic efficiency and bias of this estimate is shown to be competitive with other robust estimates. This is confirmed by a simulation study.

The popular iterative reweighting algorithm ensures a decrease of the scale if ρ in concave, i.e., if W is nonincreasing. Since this property does not hold for the type of functions considered here, we propose a modification of the iterative algorithm that ensures a decrease of the scale.

The actual performance of S estimates depend crucially on the strategy used for the initial values. We propose a simple modification of the usual subsampling procedure, that greatly improves the performance of estimates based on subsampling, with only a small increase in computational cost. In particular, it yields remarkable improvements for the MVE.

References

J. Adrover and V.J. Yohai (2002). Projection estimates of multivariate location. *Ann. Statist.,* 30, 1760–1781

D.M. Rocke (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension, *Ann. Statist.,* 24, 1327-1345.