Choosing a Multivariate Estimate for High Dimensional Data

R.A. Maronna\textsuperscript{1}, V.J. Yohai\textsuperscript{2} and A.J. Villar\textsuperscript{2}

\textsuperscript{1} Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, C.C. 172, La Plata 1900, Argentina

\textsuperscript{2} Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 1, 1426 Buenos Aires, Argentina

\textbf{Keywords:} S estimates, Robust multivariate estimation.

\textbf{Abstract}

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 $\text{ave} \{\rho(z/\sigma)\} = \delta$, where $\rho$ is nondecreasing, $\rho(0) = 0$ and $\rho(\infty) = 1$, and $\delta \in (0, 1)$. The median corresponds to $\rho(z) = I(z > 1)$ and $\delta = 0.5$, where $I(\cdot)$ is the indicator function. An S estimate $(\hat{\mu}, \hat{\Sigma})$ of the $p$-dimensional data set $\{x_1, ..., x_n\}$ is the solution of $\sigma(d_1, ..., d_n) = \min$, where $d_i = (x_i - \hat{\mu})'\hat{\Sigma}^{-1}(x_i - \hat{\mu})$ and $|\hat{\Sigma}| = 1$. 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 $\rho$ 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 $\rho$ is 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.

\textbf{References}