Robust and efficient estimation of multivariate scatter and location

Ricardo A. Maronna[(1)] and Víctor J. Yohai[(2)]

[(1)]University of La Plata

[(2)]University of Buenos Aires and CONICET

1 Overview

Our goal is to propose estimators of multivariate location and scatter that are

- Highly robust
- Highly efficient
- Computable in practical times for large dimension
- Equivariant

We consider four families of estimators which are candidates for the above goals:

- non-monotonic S-estimators (Rocke 1996)
- MM-estimators (Tatsuoka and Tyler 2000)
- τ-estimators (Lopuhaa 1991)
- The Stahel (1981) - Donoho (1982) estimator.

Although their asymptotic properties have been studied, little is known about their finite-sample behavior and implementation.

All of them allow tuning to control efficiency.

We determine the tuning constants required, and perform an extensive simulation study to compare their behaviors

We introduce two elements to help achieve the above goals, which have not been used before in this context.

- The subsampling approach usually employed to compute starting values is very expensive for large dimension. We employ a semi-deterministic equivariant procedure, proposed initially by Peña and Prieto (2007) for outlier detection, that improves both the computing times and the statistical performances.

- Besides the popular bisquare weights, we consider a weight function that has shown to have certain optimality properties for regression.

The simulation study shows that the Rocke and MM estimators, with adequate weights, tuning and starting values, can simultaneously attain high efficiency and high robustness.

Former proposals

The most frequently employed estimators do not combine efficiency and robustness.

The Minimum Volume Ellipsoid estimator (MVE) (Rousseeuw 1985) is very bias-robust, but has a very low efficiency.

The Minimum Covariance Determinant (MCD) estimator also has a low efficiency.

S-Estimators (Davies 1987) with a monotonic weight function like the bisquare have a low efficiency for small p.

Rocke (1996) (and more generally Tyler (1994)) showed that their efficiency tends to one with increasing p.

Unfortunately, this advantage is paid for with a serious loss of robustness for large p.

2 Brief review of the estimators.

For a sample $\mathbf{X} = \{\mathbf{x_1}, ..., \mathbf{x}_n\} \subset R^p$ we deal with location and scatter estimators $\widehat{\mu} \in R^p$ and $\widehat{\mathbf{\Sigma}} \in R^{p \times p}$.

For $\mathbf{x}, \mu \in R^p$ and $\mathbf{\Sigma} \in R^{p \times p}$ define the (squared) Mahalanobis distance as

$$d(\mathbf{x}, \mu, \mathbf{\Sigma}) = (\mathbf{x} - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu).$$

All estimators in this study can be seen as weighted means and covariances.

In particular, MM-, $\tau-$ and Rocke estimators satisfy the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} W\left(\frac{d_i}{S}\right)(\mathbf{x}-\mu)(\mathbf{x}-\mu)' = \mathbf{\Sigma},$$

$$\frac{1}{n}\sum_{i=1}^{n} W\left(\frac{d_i}{S}\right)(\mathbf{x}-\mu) = 0,$$

Here W is a nonnegative "weight function" and S is a scale to be defined in each case, and for brevity we put

$$d_i = d(\mathbf{x}_i, \mu, \mathbf{\Sigma}).$$

2.1 "Monotonic" M-estimators (Maronna 1976)

They are defined as solutions of the estimating equations, with $S = 1$ and W nonincreasing.

The uniqueness of the solutions requires that $W(d)\, d$ be nondecreasing.

Unfortunately, this implies (Maronna, 1976) that their breakdown point is $\leq 1/(p+1)$, which makes these estimators unreliable except for small p.

Besides, this fact holds even if μ is known,

.....while the asymptotic breakdown point of $\widehat{\mu}$ with known $\mathbf{\Sigma}$ is 0.5 with an adequate W.

This shows that the main problem to attain high robustness is the scatter matrix.

2.2 S-estimators

Let $S = S(d_1, .., d_n)$ be a scale M-estimator defined as solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{d_i}{S}\right) = \delta,$$

where $\delta \in (0, 1)$ controls the breakdown point, and $\rho(t) \in [0, 1]$ is smooth and nondecreasing in $t \geq 0$, with $\rho(0) = 0$ and $\max \rho = 1$.

Then S-estimators (Davies 1987) are defined by the minimization of $S(d_1, .., d_n)$ with $\det(\mathbf{\Sigma}) = \mathbf{1}$.

It is easy to show that S-estimators satisfy the above estimating equations with weight function $W = \rho'$.

Here, since ρ is bounded $W(d)\,d$ is not a nondecreasing function, and therefore this case is different from monotonic M-estimators.

In particular, the asymptotic breakdown point equals δ.

For the bisquare, the weight function is

$$W(d) = 3\,(1-d)^2\,\mathsf{I}\,(d \leq 1)$$

(where $\mathsf{I}\,(.)$ denotes the indicator), which is decreasing.

It would seem intuitive that the weights of the observations should decrease with their "outlyingness".

However it will be seen in the next Section that monotonicity is not necessarily favorable.

2.3 S-estimators with a non-monotonic weight function

Rocke (1996) showed that if W is nonincreasing, the efficiency of the estimator tends to one when $p \longrightarrow \infty$.

A similar result was derived by Kent and Tyler (1996) for their constrained M-estimators.

The table shows the efficiencies (to be defined later) of the bisquare S-estimator of scatter for normal p-dimensional data.

p	2	5	10	20	30	40	50
Efficiency	0.427	0.793	0.930	0.976	0.984	0.990	0.992

However, it will be seen that the price for this increase in efficiency is a decrease in robustness.

More precisely, although the breakdown point does not decrease with increasing p, the bias caused by contamination grows rapidly with p.

This fact suggests that we need estimators with a *controllable* efficiency.

But while in regression the efficiency has to be controlled to make it higher, here we need to prevent it from becoming "too high".

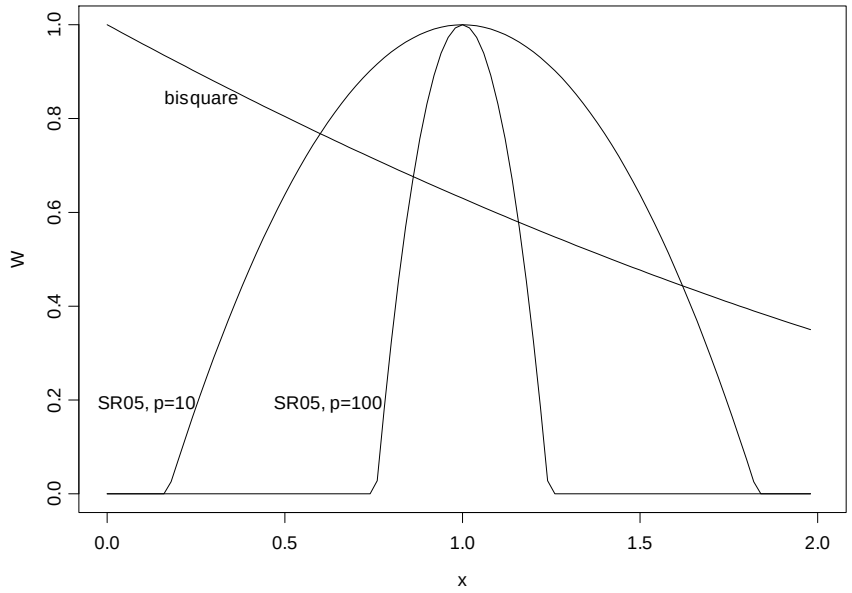
Based on the fact that for large p the p-variate standard normal distribution $\mathrm{N}_p(\mathbf{0}, \mathbf{I})$ is concentrated "near" the spherical shell with radius $\sqrt{p}$, Rocke (1996) proposed estimators with a non-monotonic weight functions.

Maronna et al. (2006) proposed a modification of Rocke's "biflat" function, namely

$$W(d) = \left[1 - \left(\frac{d-1}{\gamma}\right)^2\right] \mathsf{I}(1-\gamma \le d \le 1+\gamma)$$

where γ, which depends on α and p, controls the efficiency.

The picture shows W for $\alpha = 0.05$ and $p = 10$ and 100, and the bisquare W for comparison.



When $p \longrightarrow \infty$, the ρ of the Rocke estimator tends to the jump function that is theρ corresponding to the MVE.

Unfortunately, Maronna et al (2006, Sec. 6.8) dealt only with location.

The performance of the respective scatter matrix will be studied below.

2.4 $\tau-$estimators

$\tau-$estimators were proposed by Yohai and Zamar (1988) to obtain robust regression estimators with controllable efficiency.

Later Lopuhaä (1991) employed the same approach for multivariate estimation.

This approach requires two functions ρ_1 and ρ_2. For given $(\mu, \mathbf{\Sigma})$ call $\sigma_0(\mu, \mathbf{\Sigma})$ the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{d(\mathbf{x}_i, \mu, \mathbf{\Sigma})}{\sigma_0}\right) = \delta.$$

Then the estimator minimizes the "τ-scale"

$$\sigma(\mu, \mathbf{\Sigma}) = \sigma_0(\mu, \mathbf{\Sigma}) \frac{1}{n} \sum_{i=1}^{n} \rho_2 \left(\frac{d(\mathbf{x}_i, \mu, \mathbf{\Sigma})}{\sigma_0(\mu, \mathbf{\Sigma})} \right).$$

Here

$$\rho_2(t) = \rho_1 \left(\frac{t}{c} \right)$$

where c is chosen to regulate the efficiency.

Originally, τ-estimators were proposed to obtain estimators with higher efficiency than S-estimators for small p, which required $c > 1$.

But for large p we need $c < 1$ in order to *decrease* the efficiency.

2.5 MM-estimators

MM-estimators were initially proposed by Yohai (1987) to obtain regression estimators with a controllable efficiency.

This approach has been used in the multivariate setting by Lopuhaä (1992) and Tatsuoka and Tyler (2000).

Here we give a simplified version of the latter.

Let $\left(\widehat{\mu}_0,\widehat{\Sigma}_0\right)$ be an initial very robust although possibly inefficient estimator (e.g. the MVE).

Let S be an M-scale of $d_i^0 = d\left(\mathbf{x}_i, \widehat{\mu}_0, \widehat{\mathbf{\Sigma}}_0\right)$, $i = 1, .., n$:

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{d_i^0}{S}\right) = \delta.$$

The estimator is defined by $\left(\widehat{\mu}, \widehat{\mathbf{\Sigma}}\right)$ with $|\widehat{\mathbf{\Sigma}}| = 1$ such that

$$\sum_{i=1}^{n} \rho\left(\frac{d_i}{cS}\right) = \text{min},$$

where $d_i = d\left(\mathbf{x}_i, \widehat{\mu}, \widehat{\mathbf{\Sigma}}\right)$ and the constant c is chosen to control efficiency.

It can be shown that the solution satisfies the estimating equations
with $W = \rho'$,

Like τ-estimators, MM estimators were originally proposed to obtain estimators with higher efficiency than S-estimators for small p;

.....but here for large p the constant has to chosen to prevent the efficiency becoming *too high.*

2.6 The Stahel-Donoho estimator

This estimator is a weighted mean vector and covariance matrix,

$$\widehat{\mu} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \mathbf{x}_i,$$
$$\widehat{\mathbf{\Sigma}} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left(\mathbf{x} - \mu\right)\left(\mathbf{x} - \mu\right)',$$

with weights $w_i = W(r_i)$ where r_i is an "outlyingness measure" of $\mathbf{x}_i$.

Stahel (1981) proposed an approximate algorithm based on subsampling, the cost of which increases rapidly with p.

3 Choosing ρ (or W) for MM-, τ and Stahel-Donoho estimators

The most popular ρ in robust methods seems to be the bisquare.

Yohai and Zamar (1997) proposed a ρ for regression with certain optimality properties.

A simplified variant of this function is given by Muler and Yohai (2002).

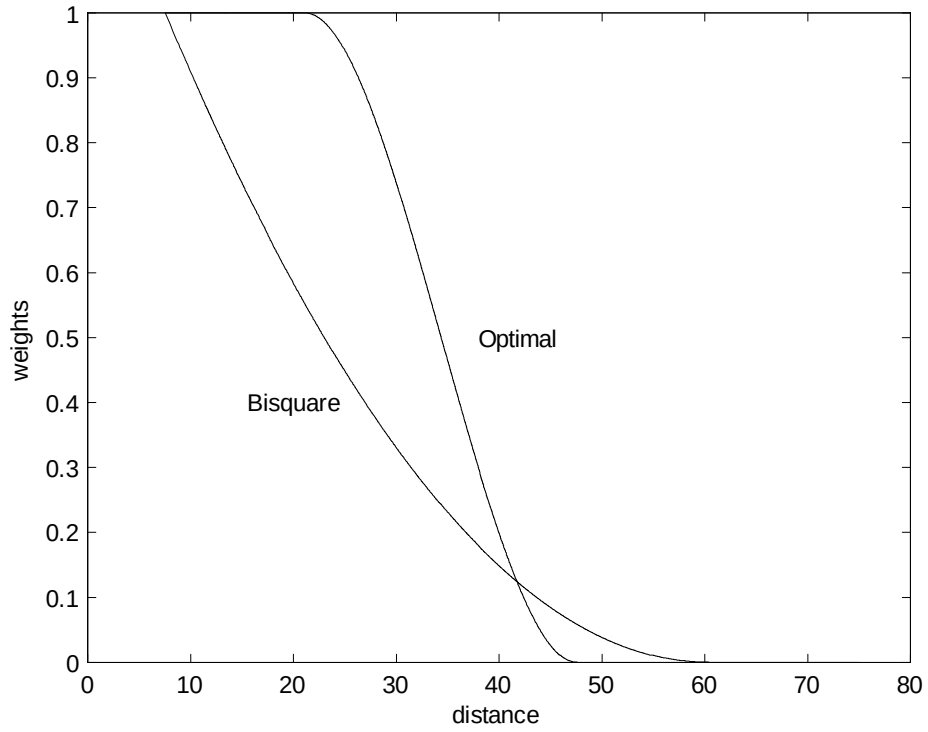
Its version for multivariate estimation has weight function

$$W_{\text{opt}}(d) = \begin{cases} 1 & \text{if} \quad d \leq 4 \\ q(d) & \text{if} \quad 4 < d \leq 9 \\ 0 & \text{if} \quad d > 9 \end{cases},$$

where $q(d)$ is a third-order polynomial such that W_{opt} is continuous and differentiable at $d = 4$ and $d = 9$.

The figure shows the bisquare and "optimal" weight functions, scaled with their respective tuning constants for the MM-estimator with 90% efficiency and $p = 30$.

It is seen that the "optimal" W yields a smaller cutoff point, and is more similar to a smooth version of hard rejection.



Bisquare and “optimal” weight functions.

4 Starting values for MM- $\tau-$ and Rocke estimators

MM- $\tau-$ and S-estimators are computed as iterative reweighted means and covariances, starting from an initial estimator.

Since in all cases we attempt to minimize a non-convex function, the initial estimator is an essential part of the procedure.

The MVE seems a good choice for this task, due to its high bias-robustness.

The standard way to compute a robust and equivariant starting point such as the MVE is subsampling.

However, ensuring a high enough breakdown point with large p may require an impractically large number of subsamples.

For these reason we need a faster and more reliable starting point.

Peña and Prieto (2007) proposed an equivariant and semi-deterministic procedure for outlier detection, which combines the projections on

- a set of $2p$ deterministic directions that are extrema of the kurtosis

- a set of random "'specific directions" obtained by stratified sampling, aimed at detecting outliers.

Although this method was originally meant for data analysis, it offers two further uses.

- First, the resulting projections can be employed to compute the Stahel-Donoho estimator;

- Second, the method yields a robust (but probably inefficient) estimator that can be used as a starting point for the iterative computing of the estimators described above.

We shall call this estimator "kurtosis plus specific directions" (KSD).

In the present setting it would not be competitive with the other estimators because its efficiency cannot be tuned,
but we shall use it as an initial estimator competing with the sampling-based MVE.

It can be proved that the asymptotic breakdown point of KSD for point-mass contamination at elliptical distributions is 0.5.

Besides, simulations by Peña and Prieto (2007) indicate that it can yield reliable results even with 40% of outliers.

5 Simulation

As a reference distribution we take the p-variate normal $\mathsf{N}_p\left(\mu_0, \Sigma_0\right)$.

In order to measure the performance of a given estimator $\left(\widehat{\mu}, \widehat{\Sigma}\right)$ we need a measure of "distance" between an estimator and the true value. Kullback-Leibler divergence

In the normal family, for μ with known Σ we have

$$D = \left(\widehat{\mu} - \mu_0\right)' \Sigma_0^{-1} \left(\widehat{\mu} - \mu_0\right),$$

and for Σ with known μ we have

$$D = \mathsf{trace}\left(\Sigma_0^{-1}\widehat{\Sigma}\right) - \mathsf{log}\left|\Sigma_0^{-1}\widehat{\Sigma}\right| - p$$

Since all estimators are equivariant we may in the simulations take without loss of generality $(\mu_0, \mathbf{\Sigma}_0) = (\mathbf{0}, \mathbf{I})$.

Each estimator is evaluated by

$$\overline{D} = \text{Monte Carlo average of the Kullback} - \text{Leibler divergences } D.$$

We generate $N = 500$ samples $\mathbf{X} = [\mathbf{x}_{ij}]$ of size n from $N_p(\mathbf{0}, \mathbf{I})$.

The estimators compared are:

- Rocke with tuning constant α;
- MM with bisquare and "optimal" ρ, with tuning constant c;
- τ with bisquare and "optimal" ρ, with tuning constant c
- Stahel-Donoho with weight function $W(r) = W_{\text{opt}}(r/c)$ where W_{opt} is the "optimal" function.

For all estimators we employed both the MVE and KSD estimators as starting values.

The tuning constants were chosen to attain an efficiency of 0.9.

All estimators have theoretical asymptotic breakdown point 0.5.

We add for completeness four other estimators with uncontrollable efficiency:

- The S-estimator (S-E) with biweight ρ. The "optimal" ρ yielded similar results.
- The MVE estimator
- The KSD estimator
- The MCD with one-step reweighting.

All scatter estimators are corrected to make the "size" consistent for normal data.

5.1 No contamination

Call $\mathbf{C}$ the sample covariance matrix. For each estimator $\hat{\mathbf{\Sigma}}$ we define

$$\text{efficiency} = \frac{\overline{D}\left(\mathbf{C}\right)}{\overline{D}\left(\hat{\mathbf{\Sigma}}\right)}.$$

The constants for each estimator are chosen to attain finite-sample efficiencies of 0.90.

To this end we computed for each estimator its tuning constants for $n = Kp$ with $K = 5$, 10 and 20 and p between 5 and 50, and then fitted the constants as functions of n and p.

For $p = 15$ the maximum efficiency of the Rocke estimator is 0.876 for all αs, and is still lower for smaller p.

The explanation is that when α tends to zero, the estimator does not tend to the covariance matrix unless p is large enough.

5.2 Contamination

For contamination rate ε, let $m = [n\varepsilon]$.

A cluster of m outliers is formed by shifting and scattering the first coordinate of the first m data points:

Given the shift K and the scatter γ we replace the first coordinate:

$$x_{i1} \longleftarrow \gamma x_{i1} + K,\ i = 1, ..., m$$

The outlier size K is varied between 1 and 12 in order to find the maximum $\overline{D}$.

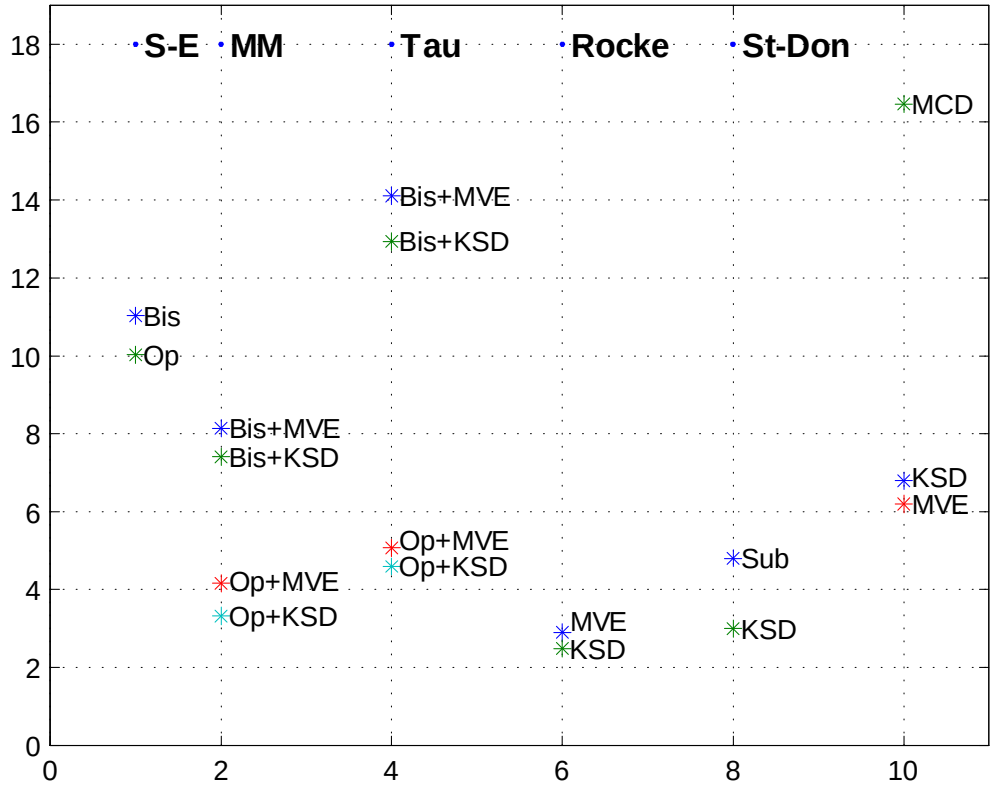
The constant γ determines the scatter of the outliers.

We employed the values $\varepsilon = 0.1$ and 0.2, and $\gamma = 0$ and 0.5.

The simulations were run for $p = 5$, 10, 15, 20 and 30, and $n = mp$ with $m = 5$, 10 and 20.

For brevity, we show the results the maximum mean $\overline{D}$ for scatter with $p = 20$, $n = 10$, $\varepsilon = 0.10$ and $\gamma = 0$.

"Bis" and "Op" stand for the bisquare and "optimal" ρs, and "Sub" stands for subsampling.



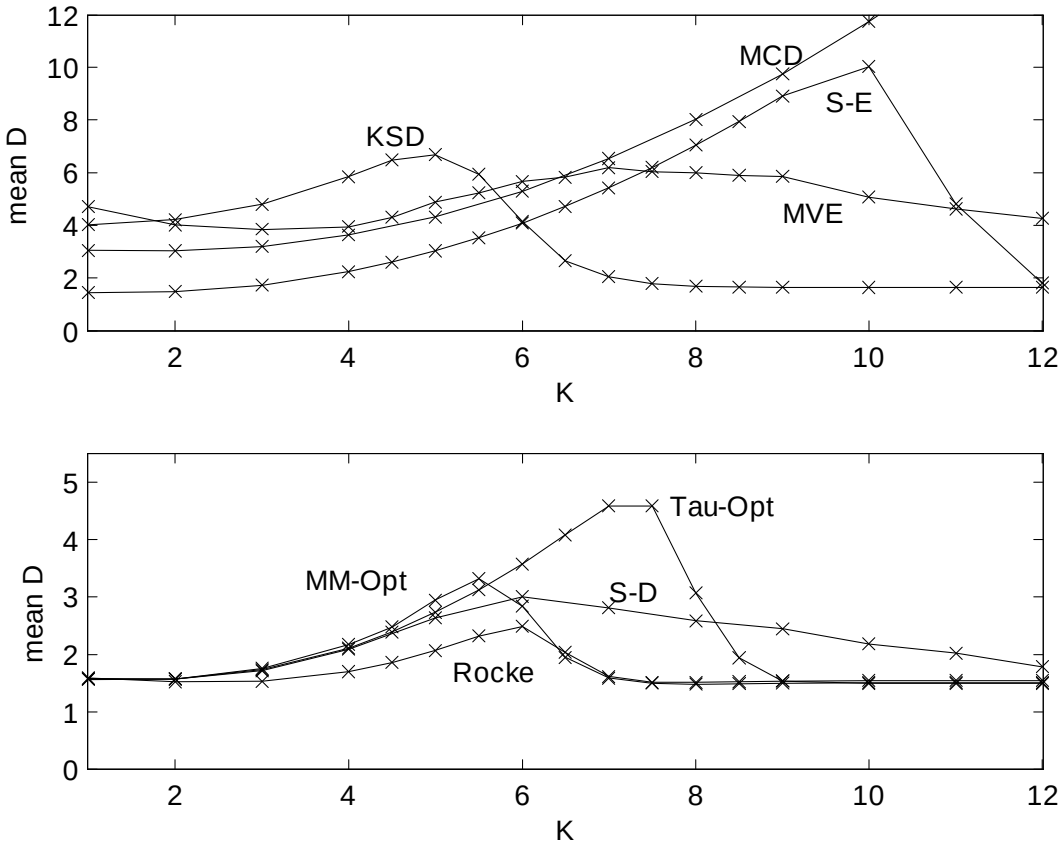
Maximum mean $\overline{D}$ for scatter with $p = 20$, $n = 200$, $\varepsilon = 0.10$ and $\gamma = 0$.

The figure shows the superiority of "Optimal" over Bisquare and of KSD over MVE (or Subsampling for S-D).

The next figure shows the values of $\overline{D}$ as a function of the outlier size K for some of the scatter estimators in the same scenario ($p = 20$, $n = 200$, $\varepsilon = 0.1$ and $\gamma = 0$).

Here "MM-Opt" stands for "MM with 'optimal' ρ".

All estimators in the lower panel start from KSD.



Mean D of of scatter estimators for $p = 20$, $n = 200$, $\varepsilon = 0.1$ and $\gamma = 0$ as a function of the outlier size K. All iterative estimators start from KSD.

The plot confirms the superiority of Rocke+KSD.

Examination of the complete simulations results shows that

- The price paid for the high efficiency of S-E is a large loss of robustness.
- KSD is always better than MVE as a starting estimator for MM and τ.
- KSD is generally better than subsampling for S-D.
- The "optimal" ρ is always better than the bisquare ρ for both MM and τ

- In all situations, the best estimators are MM and τ with "optimal" ρ, Rocke, and S-D, all starting from KSD.

- Although the results for $\gamma = 0$ and 0.5 are different, the comparisons among estimators are almost the same.

- The relative performances of the estimators for location and scatter are similar.

- The relative performances of the estimators for $n = 5p$, $10p$ and $20p$ are similar.

The following table shows a reduced version of the results for scatter, for all p, $n = 10p$ and $\gamma = 0$, and the maximum $\overline{D}$s of the scatter estimators corresponding to MM and τ (both with "optimal" ρ), Rocke and S-D, all starting from KSD.

For completion we add S-E with KSD start, and the reweighted MCD.

The results for estimators with efficiency less than 0.9 are shown in italics.

p	ε	MM Opt +KSD	τ Opt +KSD	Rocke +KSD	S-D +KSD	S-E +KSD	MCD
5	0.1	0.85	0.89	*1.26*	0.99	*1.09*	*1.99*
	0.2	2.27	2.46	*3.74*	4.53	*4.38*	*17.58*
10	0.1	1.67	1.77	*1.53*	1.61	*3.54*	*6.66*
	0.2	3.88	4.53	*1.67*	7.94	*11.26*	*21.89*
15	0.1	2.38	2.98	1.95	2.26	*6.68*	*12.53*
	0.2	5.68	7.85	4.47	12.31	*19.82*	*28.33*
20	0.1	3.32	4.59	2.49	3.00	*10.03*	*16.46*
	0.2	7.90	12.62	3.17	17.09	*25.41*	*32.04*
30	0.1	5.34	8.56	3.03	4.64	*18.39*	*17.66*
	0.2	14.21	20.71	5.61	29.66	*49.14*	*34.02*

It is seen that

- The performance of S-D is competitive for $\varepsilon = 0.1$, but is poor for $\varepsilon = 0.2$.
- For $p \leq 10$, MM has the best overall performance.
- For $p \geq 15$, Rocke has the best overall performance.

5.3 Computing times

We compare the computing times of the Rocke estimator with MVE and KSD starts.

The results are the average of 100 runs in Matlab with normal samples, on a PC with Intel TM12 Duo CPU and 3.01 GHz.

The values of n were $5p$, $10p$ and $20p$, with p between 10 and 100.

The number of subsamples N_{sub} for the MVE taken as $N_{\mathsf{sub}} = 1000$ for $p \leq 20$.

For $p > 20$ it was chosen to ensure a (probabilistic) breakdown point of 0.15, namely:

$$N_{\mathsf{sub}} = \frac{|\log \gamma|}{(1-\varepsilon)^{p+1}}, \quad \text{with } \varepsilon = 0.15,\ \gamma = 0.01.$$

The table displays the results, where for brevity we show only the values for $p = 20$, 50, 80 and 100.

p	n	N_{sub}	Rocke+MVE	Rocke+KSD
20	100	1000	0.62	0.06
	200		0.98	0.079
	400		1.31	0.15
50	250	18298	35.03	0.51
	500		42.54	1.22
	1000		81.72	3.07
80	400	2.39×10^6	—	6.43
	800		—	14.90
	1600		—	22.48
100	500	6.19×10^7	—	59.18
	1000		—	91.63
	2000		—	113.54

6 Conclusions

The Rocke estimator has a controllable efficiency for $p \geq 15$.

With equal efficiencies, the Rocke estimator with KSD start outperforms all its competitors for shift contamination

Its computing time is competitive for $p < 100$, and can probably be improved upon.

It can therefore be recommended for estimation with $p \geq 15$.

For $p < 15$ we can recommend MM with "optimal" ρ and KSD start.

References

Davies, P.L. Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. Ann. Statist. 15, 1269–1292 (1987).

Donoho, D. L. Breakdown Properties of Multivariate Location Estimators, Ph. D. Qualifying paper, Harvard University. (1982).

Kent, J.T. and Tyler, D.E. Constrained M-estimation for multivariate location and scatter. Ann. Statist., 24, 1346-1370 (1996).

Lopuhaä, H.P. Multivariate τ-estimators for location and scatter. Canad. J. Statist. 19, 307–321 (1991).

Lopuhaä, H.P. Highly efficient estimators of multivariate location with high breakdown point. Ann Stat., 20 398-413 (1992).

Maronna, R.A. Robust M-Estimators of Multivariate Location and Scatter. Ann. Statist., 4 51-67 (1976).

Maronna, R.A., Martin, R.D. and Yohai, V.J. Robust Statistics: Theory and Methods. John Wiley and Sons, New York (2006).

Maronna, R. A. and Yohai, V. J. The behavior of the Stahel–Donoho robust multivariate estimator. J. Amer. Statist. Ass., 90, 330–341 (1995).

Muler, N. and Yohai, V.J. Robust estimates for ARCH processes. J. Time Ser. Anal. 23, 341–375 (2002).

Peña, D. and Prieto, F.J. Combining Random and Specific Directions for Robust Estimation of High-Dimensional Multivariate Data. J. Comput. & Graph. Statist. 16, 228-254 (2007).

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

Rousseeuw P.J. Multivariate estimation with high breakdown point. In: Grossmann W., Pflug G., Vincze I., and Wertz W. (Eds.), Mathematical Statistics and Applications, Vol. B, 283–297. Reidel, Dordrecht (1985).

Stahel, W. A. Breakdown of covariance estimators, Research report 31, Fachgruppe für Statistik, E.T.H. Zürich (1981),

Tatsuoka, K.S. and Tyler, D.E. On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. Ann. Statist. 28, 1219–1243 (2000).

Tyler, D. E. Finite-sample breakdown points of projection-based multivariate location and scatter statistics. Ann. Statist., 22, 1024–1044 (1994).

Yohai, V.J. High Breakdown-Point and High Efficiency Robust Estimates for Regression. Ann. Statist., 15, 642–656 (1987).

Yohai, V.J. and Zamar, R. Optimal locally robust M-estimates of regression. J. Statist. Plan. & Inference., 57, 73-92 (1997).

Yohai, V.J. and Zamar, R. High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale. J. Amer. Statist. Assoc. 83, 406–413 (1988).