Accurate and Powerful Multivariate Outlier Detection

Cerioli, Andrea
Università di Parma, Dipartimento di Economia, Sezione di Statistica e Informatica
via Kennedy 6
43125 Parma, Italy
E-mail: andrea.cerioli@unipr.it

Riani, Marco
Università di Parma, Dipartimento di Economia, Sezione di Statistica e Informatica
via Kennedy 6
43125 Parma, Italy
E-mail: mriani@unipr.it

Torti, Francesca
Università di Parma, Dipartimento di Economia, Sezione di Statistica e Informatica
via Kennedy 6
43125 Parma, Italy
E-mail: francesca.torti@nemo.unipr.it

1 Introduction

The Forward Search (FS) applied to the analysis of data divides the data into a good portion that agrees with the postulated model and a set of outliers, if any (Atkinson *et al.*, 2004, 2010). In this paper we deal with the one-population multivariate setting, where the sample is made of v-dimensional observations $y_1, \ldots, y_n$ and the postulated model states that $y_i \sim N(\mu, \Sigma)$.

The basic idea of the FS is to start from a small, robustly chosen, subset of the data and to fit subsets of increasing size, in such a way that outliers and other observations not following the general structure are revealed by diagnostic monitoring. Let m_0 be the size of the starting subset. Usually $m_0 = v + 1$ or slightly larger. Let $S^{(m)}$ be the subset of data fitted by the FS at step m ($m = m_0, \ldots, n$). At that step, the outlyingness of y_i is evaluated through its squared distance

$$\hat{d}_i^2(m) = \{y_i - \hat{\mu}(m)\}' \hat{\Sigma}(m)^{-1} \{y_i - \hat{\mu}(m)\}, \tag{1}$$

where $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$ are the estimates of μ and Σ computed from $S^{(m)}$. The squared distances $\hat{d}_1^2(m), \ldots, \hat{d}_n^2(m)$ are then ordered to obtain the fitting subset at step $m + 1$.

Whilst $S^{(m)}$ remains outlier free, the squared distances $\hat{d}_i^2(m)$ will not suffer from masking and swamping. Therefore, they are a robust estimate of the population Mahalanobis distances

$$d_i^2 = (y_i - \mu)' \Sigma^{-1} (y_i - \mu), \qquad i = 1, \ldots, n. \tag{2}$$

The main diagnostic quantity computed from the robust distances (1) at step m is $\hat{d}_{i_{\min}}^2(m)$, where

$$i_{\min} = \arg \min \hat{d}_i^2(m) \qquad i \notin S^{(m)}$$

is the observation with the minimum squared Mahalanobis distance among those not in $S^{(m)}$. The key idea is that the robust distance of the observation entering the subset at step $m + 1$ will be large if this observation is an outlier. Its peculiarity will then be revealed by a peak in the forward plot of $d_{i_{\min}}^2(m)$.

Riani *et al.* (2009) develop a formal outlier test of the null hypothesis

$$H_{0s} : \{y_1 \sim N(\mu, \Sigma)\} \cap \{y_2 \sim N(\mu, \Sigma)\} \cap \ldots \cap \{y_n \sim N(\mu, \Sigma)\}, \tag{3}$$

based on the sequence $\hat{d}^2_{i_{\min}}(m)$, $m = m_0, \ldots, n$. In this test, the values of $\hat{d}^2_{i_{\min}}(m)$ are compared to the FS envelope

$$V^2_{m,\alpha}/\sigma_T(m)^2, \tag{4}$$

where $V^2_{m,\alpha}$ is the $100\alpha\%$ cut-off point of the $(m+1)$th order statistic from the scaled F distribution

$$\frac{(m^2-1)v}{m(m-v)}F_{v,m-v}, \tag{5}$$

and the factor

$$\sigma_T(m)^2 = \frac{P(X^2_{v+2} < \chi^2_{v,m/n})}{m/n} \tag{6}$$

allows for trimming of the $n-m$ largest distances. In this factor $\chi^2_{v,m/n}$ is the m/n quantile of χ^2_v and $X^2_{v+2} \sim \chi^2_{v+2}$.

A major advantage of the FS is to strive a balance between the two enemy brothers of robust statistics: robustness against contamination and efficiency under the postulated multivariate normal model. The properties of the FS, together with the use of accurate finite-sample distributional results, lead to a detection rule for multivariate outliers that has low swamping with well behaved data and high power under a variety of contamination schemes. Extensive evidence of this behaviour is shown by Riani *et al.* (2009).

All the forward search routines for regression and multivariate analysis are contained in the FSDA toolbox for MATLAB and are freely downloadable from `http://www.riani.it/MATLAB`. The FSDA toolbox also contains a series of dynamic tools which enable the user to link the information present in different forward plots and the routines to compute S and MM estimators both in regression and in multivariate analysis.

In this paper we make a comparison with inferences that come from other popular robust multivariate techniques, including multivariate MM and S-estimators. A sketch of these estimators is given in §2, where the size of the resulting outlier tests is investigated under (3) and different tunings of control parameters. Power is then investigated in §3, where we provide direct comparison with the FS and with the reweighted-MCD outlier detection rule of Cerioli (2010).

2 Null performance of multivariate S and MM estimators

For $\tilde{\mu} \in R^v$ and $\tilde{\Sigma}$ belonging to the set of positive definite symmetric $v \times v$ matrices, S estimators of multivariate location and scatter, say $\tilde{\mu}_S$ and $\tilde{\Sigma}_S$, are defined (Rousseeuw and Leroy, 1987) to be the solution of the minimization problem $|\tilde{\Sigma}| = \min$ under the constraint

$$\frac{1}{n}\sum_{i=1}^{n}\rho(\tilde{d}_i^2) = \delta, \tag{7}$$

where

$$\tilde{d}_i^2 = (y_i - \tilde{\mu})'\tilde{\Sigma}^{-1}(y_i - \tilde{\mu}) \tag{8}$$

denotes the (robust) estimate of (2) based on $\tilde{\mu}$ and $\tilde{\Sigma}$, $\rho(x)$ is a smooth function satisfying suitable regularity and robustness properties, and

$$\delta = E\{\rho(z'z)\} \tag{9}$$

for a v-dimensional vector $z \sim N(0, I)$. Perhaps the most popular choice for the ρ function in (7) is Tukey's Biweight function

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leq c \\ \frac{c^2}{6} & \text{if } |x| > c, \end{cases} \tag{10}$$

where $c > 0$ is a tuning constant. The value of c controls the breakdown point of $\tilde{\mu}_S$ and $\tilde{\Sigma}_S$; see, e.g., Rousseeuw and Leroy (1987, pp. 135–143) for details.

Given the very robust S estimators $\tilde{\mu}_S$ and $\tilde{\Sigma}_S$, an improvement in efficiency is sometimes advocated by computing refined location and shape estimators (Salibian-Barrera, Van Aelst and Willems, 2006). These estimators, called MM estimators, are defined as the minimizers of

$$\frac{1}{n} \sum_{i=1}^{n} \rho_*(\tilde{\tilde{d}}_i^2), \tag{11}$$

where

$$\tilde{\tilde{d}}_i^2 = (y_i - \tilde{\tilde{\mu}})' \tilde{\tilde{\Sigma}}^{-1} (y_i - \tilde{\tilde{\mu}}) \tag{12}$$

and the function $\rho_*(x)$ provides higher efficiency than $\rho(x)$ at the null multivariate normal model. Minimization of (11) is carried over all $\tilde{\tilde{\mu}} \in R^v$ and all $\tilde{\tilde{\Sigma}}$ belonging to the set of positive definite symmetric $v \times v$ matrices with $|\tilde{\tilde{\Sigma}}| = 1$. The MM estimator of μ is then $\tilde{\mu}_{MM} = \tilde{\tilde{\mu}}$, while

$$\tilde{\Sigma}_{MM} = \left(|\tilde{\Sigma}_S|^{\frac{1}{v}} \right) \tilde{\tilde{\Sigma}}.$$

Most published research on the properties of multivariate S and MM estimators focuses on asymptotic efficiency comparisons. Little is known about the empirical behaviour of the robust squared distances (8) and (12) when they are used for the purpose of detecting multivariate outliers. Cerioli *et al.* (2009) and Cerioli (2010) show that asymptotic distributional results for robust distances may require considerable sample sizes in order to be applied with some confidence in practice. Therefore, we start our investigation of the empirical performance of outlier detection rules based on multivariate S and MM estimators by examining their behaviour when no outlier is present in the data.

Size estimation is performed by Monte Carlo simulation of data sets generated from the v-variate normal distribution $N(0, I)$, due to affine invariance of the squared distances (8) and (12) when computed from the robust estimators $(\tilde{\mu}_S, \tilde{\Sigma}_S)$ and $(\tilde{\mu}_{MM}, \tilde{\Sigma}_{MM})$. The estimated size of each outlier detection rule is defined to be the proportion of simulated data sets for which the null hypothesis (3) is wrongly rejected. To perform hypothesis testing, we need appropriate cut-off values for the squared robust distances, whose exact distribution is unknown. Therefore, the squared robust distances are compared to the α/n quantile of their asymptotic distribution, which is χ^2_v. The Bonferroni correction ensures that the actual size of the test of (3) will be bounded by the specified value of α if the χ^2_v approximation is adequate. In our investigation we also evaluate the effect on empirical test sizes of each of some user-defined tuning constants required for practical computation of multivariate S and MM estimators with Tukey's Biweight function (10). See, e.g., Todorov and Filzmoser (2009) for computational details. Specifically, we consider:

- `bdp`: breakdown point of the S estimators, which is inherited by the MM estimators as well (the default value is 0.5);
- `eff`: efficiency of the MM estimators (the default value is 0.95);
- `eff.shape`: dummy variable setting whether efficiency of the MM estimators is defined with respect to shape (`eff.shape`= 1, the default value) or to location (`eff.shape` = 0);

- `sampS`: number of sub-samples of dimension $(p+1)$ in the resampling algorithm for fast computation of S estimators (our default value is 100);
- `iterMM`: number of iterations in the Iterative Reweighted Least Squares algorithm for computing MM estimators (our default value is 20).

Table 1 reports the results for $n = 200$, $v = 5$ and $v = 10$, when $\alpha = 0.01$ is the nominal size for testing the null hypothesis (3) of no outliers and 5000 independent data sets are generated for each combination of parameter value. It is seen that the outlier detection rules based on the robust S and MM distances can be moderately liberal, but with estimated sizes often not too far from the nominal target. As expected, liberality increases with dimension and with the amount of trimming, both for S and MM estimators. Efficiency of the MM estimators (`eff`) is the only tuning constant which seems to have a major impact on the null behaviour of these detection rules. This is because the computation of $\tilde{\Sigma}_{MM}$ does not involve a trimming factor like (6).

All in all, we conclude that the squared robust distances computed from S and MM estimators using the default tuning parameters provide an acceptable tool for the purpose of multivariate outlier detection if we are willing to tolerate a moderate amount of liberality when (3) is true.

Table 1: Estimated size of the test of (3) for $n = 200$ and nominal test size $\alpha = 0.01$. 5000 independent data sets are generated for each combination of parameter values.

		`all parameters`	`bdp`		`eff`		`eff.shape`	`sampS`		`iterMM`	
		`default values`	0.15	0.25	0.8	0.98	0	10	500	10	500
	S	0.023	0.010	0.014	0.023	0.023	0.023	0.026	0.024	0.023	0.023
v=5	MM	0.021	0.019	0.020	0.023	0.015	0.023	0.021	0.020	0.022	0.023
	S	0.033	0.005	0.007	0.033	0.033	0.033	0.031	0.036	0.033	0.033
v=10	MM	0.038	0.035	0.028	0.068	0.019	0.038	0.0286	0.030	0.034	0.036

3 Power comparison

Having roughly the appropriate size, the outlier detection rules based on multivariate S and MM estimators can be evaluated from the point of view of power. We also include in our comparison the FS outlier detection method of Riani *et al.* (2009) and the finite-sample Reweighted MCD (RMCD) technique of Cerioli (2010). Both these additional rules use cut-off values from accurate approximations to the exact distribution of the robust distances and have very good control of the size of the test of (3) even for sample sizes considerably smaller than $n = 200$.

Average power of an outlier detection rule is defined as the proportion of contaminated observations rightly named to be outliers. We estimate it by simulation, in the case $n = 200$ and $v = 5$. Performance in higher dimension, where Tukey's Biweight function is known to have problems with contaminated data, will be studied elsewhere.

We generate v-variate observations from the location-shift contamination model

$$y_i \sim (1-\delta)N(0, I) + \delta N(0 + \lambda e, I), \qquad i = 1, \ldots, n, \tag{13}$$

where $0 < \delta < 0.5$ is the contamination rate, λ is a positive scalar and e is a column vector of ones. Again, the $0.01/n$ quantile of χ^2_v is our cut-off value for outlier detection using the squared robust distances computed from S and MM estimators. We only consider the default choices for the tuning constants in Table 1, since they provide an acceptable performance under (3). We base our estimate of average power on 5000 independent data sets for each combination of parameter values.

Table 2: Estimated average power for different shifts λ in the contamination model (13), in the case $n = 200$ and $v = 5$, when the contamination rate $\delta = 0.05$. 5000 independent data sets are generated for each combination of parameter values.

	mean shift λ						
	2	2.2	2.4	2.6	2.8	3	4
S	0.3442	0.5252	0.6960	0.8273	0.9119	0.9632	1.0
MM	0.1484	0.2801	0.4658	0.6724	0.8359	0.9354	1.0
RMCD	0.2265	0.3896	0.5741	0.7322	0.8563	0.9355	1.0
FS	0.3586	0.5665	0.7297	0.8397	0.9087	0.9528	1.0

Tables 2–4 provide the results for different values of δ. If the contamination rate is small, it is seen that the four methods behave somewhat similarly, with FS often ranking first and MM always ranking last as λ varies. However, when the contamination rate increases, the advantage of the Forward Search detection rule is paramount. In that situation both S and MM estimators become ineffective for the purpose of identifying multivariate outliers.

Table 3: Estimated average power for different shifts λ in the contamination model (13), in the case $n = 200$ and $v = 5$, when the contamination rate $\delta = 0.15$. 5000 independent data sets are generated for each combination of parameter values.

	mean shift λ							
	2	2.2	2.4	2.6	2.8	3	3.4	4
S	0.0730	0.2329	0.5322	0.7716	0.9013	0.9601	0.9955	0.9999
MM	0.0059	0.0077	0.0101	0.0123	0.0163	0.0262	0.3969	0.9936
RMCD	0.0962	0.2269	0.4283	0.6516	0.8151	0.9132	0.9880	0.9998
FS	0.5800	0.7380	0.8026	0.8781	0.9346	0.9653	0.9934	0.9999

Table 4: Estimated average power for different shifts λ in the contamination model (13), in the case $n = 200$ and $v = 5$, when the contamination rate $\delta = 0.30$. 5000 independent data sets are generated for each combination of parameter values.

	mean shift λ									
	2	2.2	2.4	2.6	2.8	3	4	6	8	10
S	0.0033	0.0043	0.0053	0.0063	0.0073	0.0088	0.0162	0.0922	0.7719	0.9757
MM	0.0021	0.0027	0.0034	0.0042	0.0047	0.0058	0.0115	0.0849	0.7694	0.9754
RMCD	0.0099	0.0494	0.1593	0.3814	0.6370	0.8393	0.9997	1	1	1
FS	0.6269	0.8657	0.9145	0.9195	0.9410	0.9669	0.9997	1	1	1

A qualitative explanation for the failure of multivariate MM estimators is shown in Figure 1 in the simple case $v = 2$. The four plots display bivariate ellipses corresponding to 0.95 probability contours at different iterations of the algorithm for computing MM estimators, for a data set simulated from the contamination model (13) with $n = 200$, $\delta = 0.15$ and $\lambda = 3$. The data can be reproduced using function `randn(200,2)` of MATLAB and putting the random number seed to 2. The contaminated units are shown with symbol $\circ$ and the two lines which intersect the estimate of the robust centroid are plotted using a dash-dot symbol. The upper left-hand panel corresponds to the first iteration (i1), where the location estimate is $\tilde{\mu}_S = (0.19, 0.18)'$ and the value of the robust correlation r derived from $\tilde{\Sigma}_S$ is 0.26. In this case the robust estimates are not too far from the true parameter values $\mu = (0, 0)'$ and $\Sigma = I$, and the corresponding outlier detection rule (i.e., the S-rule in Table 3) can be expected

to perform reasonably well. On the contrary, as the algorithm proceeds, the ellipse moves its center far from the origin and the variables artificially become more correlated. The value of r in the final iteration (i8) is 0.47 and the final centroid $\tilde{\tilde{\mu}} = \tilde{\mu}_{MM}$ is $(0.37; 0.32)'$. These features increase the bias of the parameter estimates and can contribute to masking in the supposedly robust distances (12).

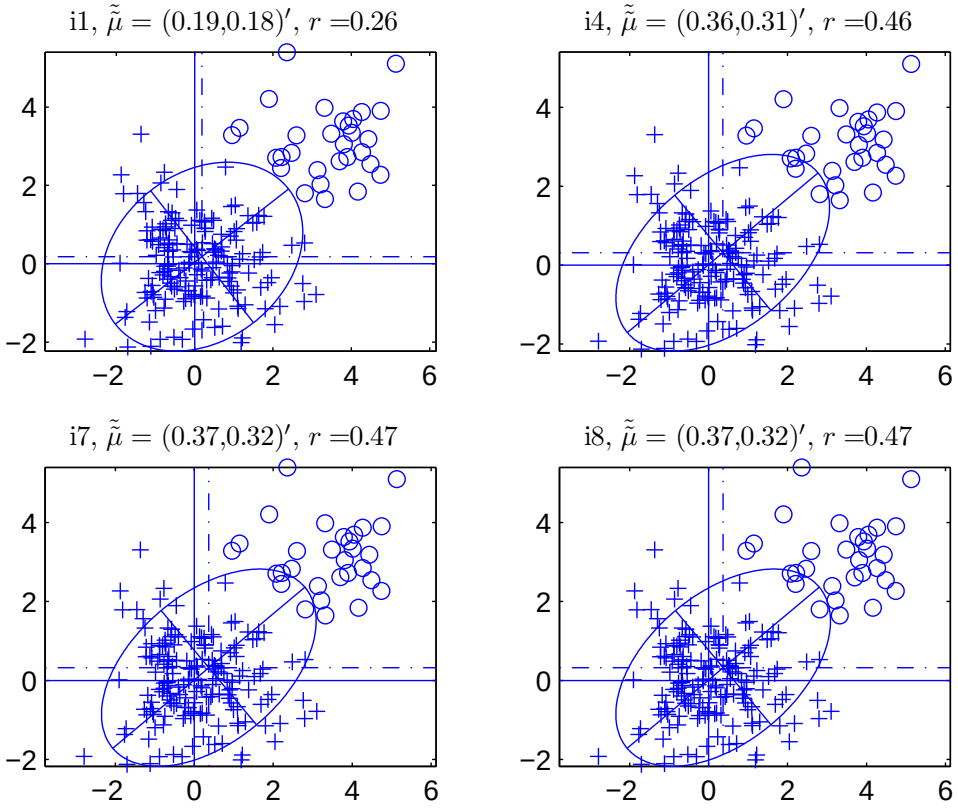

Figure 1: Ellipses corresponding to 0.95 probability contours at different iterations of the algorithm for computing multivariate MM estimators, for a data set simulated from the contamination model (13) with $n = 200$, $v = 2$, $\delta = 0.15$ and $\lambda = 3$.

REFERENCES

Atkinson, A. C., Riani, M. and Cerioli, A. (2004). *Exploring Multivariate Data with The Forward Search*. Springer, New York.

Atkinson, A. C., Riani, M. and Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the Korean Statistical Society* 39, 117–134.

Cerioli, A. (2010). Multivariate Outlier Detection With High-Breakdown Estimators. *Journal of the American Statistical Association* 105, 147–156.

Cerioli, A., Riani M., and Atkinson A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter, *Statistics and Computing* 19, 341–353.

Riani, M., Atkinson A. C. and Cerioli, A. (2009). Finding an Unknown Number of Multivariate Outliers, *Journal of the Royal Statistical Society, Ser. B* 71, 447–466.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Salibian-Barrera, M., Van Aelst, S. and Willems, G. (2006). Principal Components Analysis Based on Multivariate MM Estimators With Fast and Robust Bootstrap. *Journal of the American Statistical Association* 101, 1198–1211.

Todorov, V. and Filzmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software* 32, 1–47.