Some Perspectives on Multivariate Outlier Detection

Andrea Cerioli, Anthony C. Atkinson, and Marco Riani

Abstract We provide a selective view of some key statistical concepts that underlie the different approaches to multivariate outlier detection. Our hope is that appreciation of these concepts will help to establish a unified and widely accepted framework for outlier detection.

1 Introduction

The identification of outliers is an important step of any analysis of multivariate data. In a multivariate setting, this task poses more challenging problems than in the simpler case of a single variable for at least three basic reasons:

- outlyingness should be judged with respect to several (possibly many) dimensions simultaneously;
- there is no natural ordering of multivariate data on which 'extremeness' of an observation can be ascertained;
- simple graphical diagnostic tools like the boxplot are difficult to construct in more than one or two dimensions (Zani et al. 1998).

It is thus not surprising that the systematic study of multivariate outliers has a long history in the statistical literature and has led to remarkably different points of view. See, e.g., Hadi et al. (2009) and Morgenthaler (2006) for recent reviews on robust methods and outlier detection. In these and other reviews, it is acknowledged that the concern for outliers or grossly wrong measurements is probably as old as the experimental approach to science. The earliest reported historical references usually date back to the seventeenth century, with the first precise specifications subsequently given by Gauss and Legendre. Perhaps less known is the fact that the same concern was also present in Ancient Greece more than 2,000 years ago, as reported by Thucydides in his *History of The Peloponnesian War* (III 20, 3–4).[1]

[1] According to Thucydides, in 428 B.C. the Plataeans, besieged by the Spartans, excluded extreme measurements when estimating the height of the walls that their enemies had built around the city. In this way, they managed to break the siege.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_26, © Springer-Verlag Berlin Heidelberg 2011

In the modern statistical era, until the early 1990s, alternative methods were developed following three essentially distinct streams of research, well documented in the classical book by Barnett and Lewis (1994, Chap. 7): robust techniques for multivariate outlier accommodation; formal tests of hypotheses for precise outlier identification and, thirdly, less formal diagnostic tools for exploratory analysis including intuitive inspection of the data. With a bit of humour, the supporters of these alternative schools of outlier methodology were sometimes called the *robustniks*, the *testniks* and the *diagnostniks*, respectively. The reconciliation of different 'outlier philosophies' was already seen as an ideal by Rousseeuw and Zomeren (1990) in their rejoinder twenty years ago, but it has still to be reached. Sect. 4 provides a suggestion in that direction.

It is not the goal of this paper to provide a comprehensive overview of the wealth of methods developed for the purpose of multivariate outlier identification. Rather, the idea is to guide the reader through a few key statistical concepts that underlie the different approaches and to see how they evolved over the years. Our hope is that appreciation of these concepts will help us in establishing a unified and widely accepted framework for outlier detection.

2 Outlier Detection and Testing

In a seminal paper Wilks (1963) laid down the statistical foundations of multivariate outlier detection. Let $y = (y_1, \dots, y_n)'$ be a sample of v-dimensional observations from $N(\mu, \Sigma)$. The sample mean is $\hat{\mu}$ and the unbiased sample estimate of Σ is $\hat{\Sigma}$. Wilks derived the exact distribution of the n scatter ratios,

$$R_i = |(n-2)\hat{\Sigma}_{\{i\}}||(n-1)\hat{\Sigma}|^{-1} \qquad i = 1, \dots, n,$$

where $\hat{\Sigma}_{\{i\}}$ is the unbiased estimate of Σ computed after deleting y_i. It is easily seen (Atkinson et al. 2004, pp. 44–46) that R_i is inversely related to the squared Mahalanobis distance of observation y_i

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

so that the distributional results for R_i hold for d_i^2 as well. In particular,

$$d_i^2 = \frac{(n-1)^2}{n}(1 - R_i) \sim \frac{(n-1)^2}{n}\mathrm{Beta}\left(\frac{v}{2}, \frac{n-v-1}{2}\right) \qquad i = 1, \dots, n. \tag{2}$$

Wilks also showed how the smallest ratio $R_{(1)}$, or equivalently the largest squared distance $d_{(n)}^2$, can be used to test the outlyingness of the corresponding observation. This multivariate outlier detection rule focuses on the *intersection hypothesis* that no outlier is present in the data

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

against the alternative that one is present. The candidate outlier is the most remote observation, i.e., the observation with the largest squared Mahalanobis distance (1). The size of a test of H_{0s}, say γ, represents the proportion of good *data sets* that are wrongly declared to contain outliers. Simultaneity is dealt with in Wilks (1963) by introducing a Bonferroni bound on the probability that the test statistic exceeds a given threshold.

Simulations show that Wilks' outlier detection method, combining the scaled Beta distribution (2) with a Bonferroni bound, has very good control of the size of the test of H_{0s}. It can thus be taken as a benchmark for comparison with alternative procedures under the null hypothesis of no outliers.

However, the squared Mahalanobis distances (1) suffer from *masking*. If a few outliers contaminate the data, it is unlikely that the largest distances $d^2_{(n)}, d^2_{(n-1)}, \dots$ will be associated with the atypical observations because $\hat{\mu}$ and $\hat{\Sigma}$ will be grossly distorted by these outliers.

Wilks (1963) extended his deletion method to the case of two observations, but dealing with an unknown and possibly large number of outliers rapidly becomes infeasible. The same problem affects any other backward procedure, such as the sequential application of Wilks' test suggested by Caroni and Prescott (1992). Moving backwards, all the outliers will be missed if $d^2_{(n)}$ is masked.

3 Robust Distances from High-Breakdown Estimators

The use of high-breakdown estimators of μ and Σ in the place of the classical ones has proved to be a practical solution to the problem of masking. Popular choices for such estimators include the Minimum Covariance Determinant estimator, S-estimators and projection-based techniques. See Hubert et al. (2008) or Maronna et al. (2006, Chap. 6) for recent reviews.

Let $\tilde{\mu}$ and $\tilde{\Sigma}$ be the chosen high-breakdown estimators of μ and Σ. The corresponding squared robust Mahalanobis distances are

$$\tilde{d}_i^2 = (y_i - \tilde{\mu})' \tilde{\Sigma}^{-1} (y_i - \tilde{\mu}) \qquad i = 1, \dots, n. \tag{4}$$

The outliers in y are revealed by their large distances from the robust fit provided by $\tilde{\mu}$ and $\tilde{\Sigma}$, without suffering from masking. The key issue is that outlying observations have null or negligible weight in the computation of $\tilde{\mu}$ and $\tilde{\Sigma}$. Therefore, they cannot 'attract' these estimates and maintain a large value of $\tilde{d}_i^2$.

The need to avoid masking directs the outlier detection problem to the choice of suitable cut-offs for the robust squared distances (4), instead of the classical critical values computed from (2). However simple this step may seem, it has produced some surprising consequences.

First, the focus has shifted from the intersection hypothesis (3) to the problem of testing the n null hypotheses

$$H_{0i} : y_i \sim N(\mu, \Sigma), \qquad i = 1, \dots, n. \tag{5}$$

The most common approach has been to test all these hypotheses individually at a specified size $0.01 \leq \alpha \leq 0.05$, with $\alpha = 0.025$ being perhaps the most popular choice. This approach, which does not take multiplicity of tests into account, increases the probability of detecting truly contaminated observations, but the user must be prepared to declare at least one outlier (and often many more) in most data sets of realistic size. In other words, the user must be prepared to invest a large sum of money (if all the suspected outliers are discarded) or a large amount of time (if the suspected outliers are checked one by one) to accomplish the process of multivariate outlier detection, even when the expected number of contaminated observations is small.

The tendency of an outlier detection method to label good observations as outliers is called *swamping*. We believe that the ability to control the degree of swamping is an important property for the practical usefulness of an outlier detection method. Major application areas where even a moderate number of false outliers may have disastrous consequences include anti-fraud analysis and statistical quality control. For instance, outliers are of great interest in the analysis of trade data arising in the European Union market (Riani et al. 2008), because some of them may correspond to fraudulent transactions. Since there are hundreds of transactions to be inspected over thousands of markets, ignoring the multiplicity of tests would lead to a plethora of false signals for anti-fraud services, thus making substantial investigation of possible frauds impractical.

Another major shortcoming of the use of the squared robust distances $\tilde{d}_i^2$ is that their exact distribution is unknown. The required cut-offs are then usually computed from their asymptotic χ_v^2 distribution, although the adequacy of this approximation can be very poor even in moderately large samples, especially when the number of dimensions increases. This behaviour has been shown in many simulation studies: see, e.g., Becker and Gather (2001), Cerioli et al. (2009), Hardin and Rocke (2005) and Riani et al. (2009).

The liberality of the χ_v^2 distribution for the purpose of approximating the squared robust distances $\tilde{d}_i^2$ adds further swamping to the individual testing framework of the n hypotheses H_{0i}. It also makes the simultaneous testing of these hypotheses in (3) even more problematic, because the corresponding cut-offs lie in the extreme tail of the true but unknown distribution of the robust distances. This behaviour is in sharp contrast with the excellent null performance of the classical Mahalanobis distances (1). It also obviously calls for better approximations to the finite sample distribution of the squared robust distances $\tilde{d}_i^2$ when no outlier is present in the data.

Hardin and Rocke (2005) suggest a way to approximate the first two moments of the distribution of the squared robust distances $\tilde{d}_i^2$. However, simultaneous testing of the n hypothesis (5) requires cut-off values which are in the extreme tail of the distribution. In that case information on $E(\tilde{d}_i^2)$ and $var(\tilde{d}_i^2)$ is often not enough to obtain reliable rejection regions and it is preferable to estimate the cut-offs directly. A good finite sample approximation to the required thresholds under the intersection hypothesis (3) is proposed by Cerioli et al. (2009). Their idea is to calibrate the asymptotic cut-off values by Monte Carlo simulation. Calibration is first performed for some representative values of n and v and then extended to any n and v by parametric non-linear interpolation. The resulting outlier detection rule has very good control of the size of the test of no outliers even in situations where space is very sparsely filled (e.g., $n = 50, v = 10$). The method is very general, as in principle it can be applied to any choice of $\tilde{\mu}$ and $\tilde{\Sigma}$, and also easy to implement, once the parameters of the interpolation function are made available. However, the power may be rather low, since the technique does not allow for the variability of distances in the tail of the distribution.

Cerioli (2010) provides a power improvement by introducing an accurate approximation to the distribution of one-step reweighted robust distances. This approximation is based on a scaled Beta distribution mimicking (2) for the units not suspected of being outliers, and on a scaled F distribution for the units which are trimmed in the reweighting step. Also this method provides good control of the simultaneous size of the n outlier tests (5). Therefore, it can be useful in all the application fields where allowing for the multiplicity of tests is an important issue.

Attaining the right size through distributional results yields more powerful rules than through calibration of cut-off values. Furthermore, a substantial increase in power can be obtained by controlling the number of false discoveries only when all the data come from the prescribed null distribution. Cerioli (2010) suggests an outlier identification rule that tolerates some degree of swamping, but only when there is strong evidence that some contamination is present in the data. This follows the idea that the level of swamping provided by repeated testing of (5), although deleterious in ‘good’ data sets, may still be acceptable in a contaminated framework. Such a view is often sensible when the probability of observing a contaminated sample is small. On the contrary, if the sample is predicted to have some contamination with high probability, but the expected number of contaminants is small, other approaches could be followed. As shown in Cerioli and Farcomeni (2011), by controlling the False Discovery Rate it is possible to develop outlier identification rules for which the acceptable number of false discoveries depends explicitly on the number of outliers found.

4 Is a Reconciliation Possible?

Wilks’ outlier test and the high-breakdown identification rules described by Hubert et al. (2008) have opposite attitudes towards two basic issues of multivariate outlier detection: the null hypothesis to be tested in order to label an observation as an outlier and the approach towards the control of the number of false discoveries. A reconciliation of these alternative philosophies can be found in the Forward Search method of Atkinson et al. (2004, 2010).

The basic idea of the Forward Search (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 clearly revealed by diagnostic monitoring. Let m_0 be the size of the starting subset. Usually $m_0 = v + 1$ or slightly larger. Let $S_*^{(m)}$ denote the subset of data fitted by the FS at step m ($m = m_0, \ldots, n$). At that step, outlyingness of each observation y_i can be evaluated through the squared distance

$$d_{i*}^2(m) = \{y_i - \hat{\mu}_*(m)\}' \hat{\Sigma}_*(m)^{-1} \{y_i - \hat{\mu}_*(m)\}, \tag{6}$$

where $\hat{\mu}_*(m)$ and $\hat{\Sigma}_*(m)$ are the estimates of μ and Σ computed from $S_*^{(m)}$. The squared distances $d_{1*}^2(m), \ldots, d_{n*}^2(m)$ are then ordered to obtain the fitting subset at step $m+1$. Usually one observation enters the subset at each step, but sometimes two or more, when one or more then leave. Such occurrences are indicative of changes in structure or of clusters of outliers entering the subset.

Whilst $S_*^{(m)}$ remains outlier free, the squared distances $d_{i*}^2(m)$ will not suffer from masking and swamping. Therefore, they are a robust version of the classical Mahalanobis distance d_i^2. The main diagnostic quantity computed from these robust distances is $d_{i_{\min}*}^2(m)$, where

$$i_{\min} = \arg\ \min d_{i*}^2(m) \qquad i \notin S_*^{(m)}$$

is the observation with the minimum squared Mahalanobis distance among those not in $S_*^{(m)}$. The main idea is that the distance of the closest observation entering the subset at step $m+1$ will be large if this observation is an outlier. Its peculiarity will be clearly revealed by a peak in the forward plot of $d_{i*}^2(m)$.

The early developments of the FS aimed essentially to provide powerful plots for investigating the structure of regression and multivariate data, using quantities such as $d_{i_{\min}*}^2(m)$. Therefore, the FS might be seen a contribution of the 'diagnostic' school of outlier detection. However, it is paramount that any diagnostic quantity can result in a formal test if its null distribution is known and appropriate thresholds can be defined. The statistic $d_{i_{\min}*}^2(m)$ can be treated as a squared deletion distance on $m-1$ degrees of freedom, whose distribution is (Atkinson et al. 2004, pp. 43–44)

$$\frac{(m^2-1)v}{m(m-v)} F_{v,m-v}, \tag{7}$$

while $S_*^{(m)}$ remains outlier free. This statistic is based on $\hat{\Sigma}_*(m)$, which is a biased estimate of Σ, being calculated from the m observations in the subset that have been chosen as having the m smallest distances. As a result, Riani et al. (2009) propose a formal outlier test based on the FS by making use of the envelopes

$$V_{m,\alpha}/\sigma_T(m), \tag{8}$$

where $V_{m,\alpha}$ is the $100\alpha\%$ cut-off point of the $(m+1)$th order statistic from the scaled F distribution (7) and the factor

$$\sigma_T(m)^{-1} = \frac{m/n}{P(X_{v+2}^2 < \chi_{v,m/n}^2)} \tag{9}$$

allows for trimming of the $n - m$ largest distances. In (9) $\chi^2_{v,m/n}$ is the m/n quantile of χ^2_v and $X^2_{v+2} \sim \chi^2_{v+2}$.

The FS test for multivariate outlier identification based on thresholds derived from (8) does not require computation of the high-breakdown estimators $\tilde{\mu}$ and $\tilde{\Sigma}$. Furthermore, like the methods described in Sect. 2, this is a simultaneous test which has good control of the size of the test of no outliers (3). This property is made possible by the use of accurate finite sample distributional results for the squared Mahalanobis distances computed along the search. Nevertheless, the FS test does not suffer from masking, because it is the algorithm itself which is robust. Thus the FS test can cope with the same contamination rate as the high-breakdown methods sketched in Sect. 3.

We conclude that the Forward Search can provide a reconciliation of the three classical approaches to outlier detection introduced in Sect. 1. Being based on a flexible strategy in which the proportion of trimming is determined by the data, it enjoys high power. A further bonus of the Forward Search is its suitability for being easily adapted to cope with many different methodologies, including other multivariate techniques, linear and non-linear regression and correlated data modelling. For instance, by allowing a level of trimming smaller than 0.5, we believe that the Forward Search has the greatest potential among robust techniques to become a comprehensive approach through which cluster analysis and outlier detection could be performed under the same umbrella.

Acknowledgements The Authors are grateful to Spyros Arsenis and Domenico Perrotta for pointing out the reference to Thucydides.

References

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer–Verlag.

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

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Becker, C., & Gather, U. (2001). The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Computational Statistics and Data Analysis*, *36*, 119–127.

Caroni, C., & Prescott, P. (1992). Sequential application of Wilks's multivariate outlier test. *Applied Statistics*, *41*, 355–364.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, *105*(489), 147–156.

Cerioli A., & Farcomeni, A. (2011). Error rates for multivariate outlier detection, Computational Statistics and Data Analysis, 55, 544–553.

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

Hadi, A. S., Rahmatullah Imon, A. H. M., & Werner, M. (2009). Detection of outliers. *WIREs Computational Statistics*, *1*, 57–70.

Hardin, J., & Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, *14*, 910–927.

Hubert, M., Rousseeuw, P. J., & Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, *23*, 92–119.

Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. New York: Wiley.

Morgenthaler, S. (2006). A survey of robust statistics. *Statistical Methods and Applications*, *15*, 271–293 (Erratum *16*, 171–172).

Riani, M., Cerioli, A., Atkinson, A., Perrotta, D., & Torti, F. (2008). Fitting mixtures of regression lines with the forward search. In F. Fogelman-Soulié, D. Perrotta, J. Piskorski, & R. Steinberger (Eds.), *Mining massive data sets for security* (pp. 271–286). Amsterdam: IOS Press.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, *71*, 447–466.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. With discussion. *Journal of the American Statistical Association*, *85*, 633–651.

Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhya A*, *25*, 407–426.

Zani, S., Riani, M., & Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis*, *28*, 257–270.