A MULTIVARIATE OUTLIER DETECTION METHOD

P. Filzmoser
Department of Statistics and Probability Theory
Vienna, AUSTRIA
e-mail: P.Filzmoser@tuwien.ac.at

Abstract

A method for the detection of multivariate outliers is proposed which accounts for the data structure and sample size. The cut-off value for identifying outliers is defined by a measure of deviation of the empirical distribution function of the robust Mahalanobis distance from the theoretical distribution function. The method is easy to implement and fast to compute.

1 Introduction

Outlier detection belongs to the most important tasks in data analysis. The outliers describe the abnormal data behavior, i.e. data which are deviating from the natural data variability. Often outliers are of primary interest, for example in geochemical exploration they are indications for mineral deposits. The cut-off value or threshold which divides anomalous and non-anomalous data numerically is often the basis for important decisions.

Many methods have been proposed for univariate outlier detection. They are based on (robust) estimation of location and scatter, or on quantiles of the data. A major disadvantage is that these rules are independent from the sample size. Moreover, by definition of most rules (e.g. mean ±2· scatter) outliers are identified even for "clean" data, or at least no distinction is made between outliers and extremes of a distribution.

The basis for multivariate outlier detection is the Mahalanobis distance. The standard method for multivariate outlier detection is robust estimation of the parameters in the Mahalanobis distance and the comparison with a critical value of the χ^2 distribution (Rousseeuw and Van Zomeren, 1990). However, also values larger than this critical value are not necessarily outliers, they could still belong to the data distribution.

In order to distinguish between extremes of a distribution and outliers, Garrett (1989) introduced the χ^2 plot, which draws the empirical distribution function of the robust Mahalanobis distances against the χ^2 distribution. A break in the tail of the distributions is an indication for outliers, and values beyond this break are iteratively deleted.

The approach of Garrett (1989) needs a lot of interaction of the analyst with the data since this method is not an automatic procedure. We propose a method which computes the outlier threshold adaptively from the data. By investigation of the tails of the difference between the empirical and a hypothetical distribution function we find an adaptive threshold value which increases with sample size if there are no outliers and which is bounded in presence of outliers.

2 Methods for Multivariate Outlier Detection

The shape and size of multivariate data are quantified by the covariance matrix. A well-known distance measure which takes into account the covariance matrix is the Mahalanobis distance. For a p-dimensional multivariate sample x_i ($i = 1, \ldots, n$) the Mahalanobis distance is defined as

$$\mathrm{MD}_i = \Big((x_i - t)^T C^{-1} (x_i - t)\Big)^{1/2} \qquad \text{for} \qquad i = 1, \ldots, n \tag{1}$$

where t is the estimated multivariate location and C the estimated covariance matrix. Usually, t is the multivariate arithmetic mean, and C is the sample covariance matrix. For multivariate normally distributed data the values are approximately chi-square distributed with p degrees of freedom (χ^2_p). Multivariate outliers can now simply be defined as observations having a large (squared) Mahalanobis distance. For this purpose, a quantile of the chi-squared distribution (e.g., the 97.5% quantile) could be considered. However, this approach has several shortcomings. The Mahalanobis distances need to be estimated by a robust procedure in order to provide reliable measures for the recognition of outliers. Single extreme observations, or groups of observations, departing from the main data structure can have a severe influence to this distance measure, because both location and covariance are usually estimated in a non-robust manner. Many robust estimators for location and covariance have been introduced in the literature. The minimum covariance determinant (MCD) estimator is probably most frequently used in practice, partly because a computationally fast algorithm is available (Rousseeuw and Van Driessen, 1999). Using robust estimators of location and scatter in formula (1) leads to so-called robust distances (RD). Rousseeuw and Van Zomeren (1990) used these RDs for multivariate outlier detection. If the squared RD for an observation is larger than, say, $\chi^2_{p;0.975}$, it can be declared a candidate outlier.

This approach, however has shortcomings: It does not account for the sample size n of the data, and, independently from the data structure, observations could be flagged as outliers even it they belong to the data distribution. A better procedure than using a fixed threshold is to adjust the threshold to the data set at hand. Garrett (1989) used the chi-square plot for this purpose, by plotting the squared Mahalanobis distances (which have to be computed at the basis of robust estimations of location and scatter) against the quantiles of χ^2_p, the most extreme points are deleted until the remaining points follow a straight line. The deleted points are the identified outliers. This method, however, is not automatic, it needs user interaction and experience on the part of the analyst. Moreover, especially for large data sets, it can be time consuming, and also to some extent it is subjective. In the next section a procedure that does not require analyst intervention, is reproducible and therefore objective, into consideration is introduced.

3 An Adaptive Method

The chi-square plot is useful for visualizing the deviation of the data distribution from multivariate normality in the tails. This principle is used in the following. Let $G_n(u)$

denote the empirical distribution function of the squared robust distances RD_i^2, and let $G(u)$ be the distribution function of χ^2_p. For multivariate normally distributed samples, G_n converges to G. Therefore the tails of G_n and G can be compared to detect outliers. The tails will be defined by $\delta = \chi^2_{p;1-\alpha}$ for a certain small α (e.g., $\alpha = 0.025$), and

$$p_n(\delta) = \sup_{u \geq \delta} \Big(G(u) - G_n(u) \Big)^+$$

is considered, where "+" indicates the positive differences. In this way, $p_n(\delta)$ measures the departure of the empirical from the theoretical distribution only in the tails, defined by the value of δ. $p_n(\delta)$ can be considered as a measure of outliers in the sample. Gervini (2003) used this idea as a reweighting step for the robust estimation of multivariate location and scatter. The efficiency of the estimator could be improved considerably.

$p_n(\delta)$ will not be directly used as a measure of outliers. The threshold should be infinity in case of multivariate normally distributed data, i.e. extreme values or values from the same distribution should not be declared as outliers. Therefore a critical value p_{crit} is introduced, which helps to distinguish between outliers and extremes. The measure of outliers in the sample is then defined as

$$\alpha_n(\delta) = \begin{cases} 0 & \text{if} \quad p_n(\delta) \leq p_{crit}(\delta, n, p) \\ p_n(\delta) & \text{if} \quad p_n(\delta) > p_{crit}(\delta, n, p). \end{cases}$$

The threshold value which will be called *adjusted quantile* is then determined as $c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta))$. The critical value for distinguishing between outliers and extremes can be derived by simulation, and the result is approximately

$$p_{crit}(\delta, n, p) = \frac{0.24 - 0.003 \cdot p}{\sqrt{n}} \quad \text{for} \quad \delta = \chi^2_{p,0.975} \text{ and } p \leq 10$$

and

$$p_{crit}(\delta, n, p) = \frac{0.252 - 0.0018 \cdot p}{\sqrt{n}} \quad \text{for} \quad \delta = \chi^2_{p,0.975} \text{ and } p > 10$$

(see Filzmoser, Reimann, and Garrett, 2003).

4 Example

We simulate a data set in two dimensions in order to simplify the graphical visualization. 85 data points follow a bivariate standard normal distribution. Multivariate outliers are introduced by 15 points coming from a bivariate normal distribution with mean $(2, 2)^T$ and covariance matrix $\mathrm{diag}(1/10, 1/10)$. The data are presented in Figure 1. The MCD estimator is applied and the robust distances are computed. Figure 2 shows in more detail how the adaptive outlier detection method works. The plot shows the empirical distribution function of the ordered squared distances (points) and the theoretical distribution function of χ^2_2 (solid line). The dotted line $\chi^2_{2;0.975}$ identifies only four outliers which are presented in the left part of Figure 3 as dark points. The adjusted quantile gives a more realistic rule, and the identified outliers are shown in the right part of Figure 3.

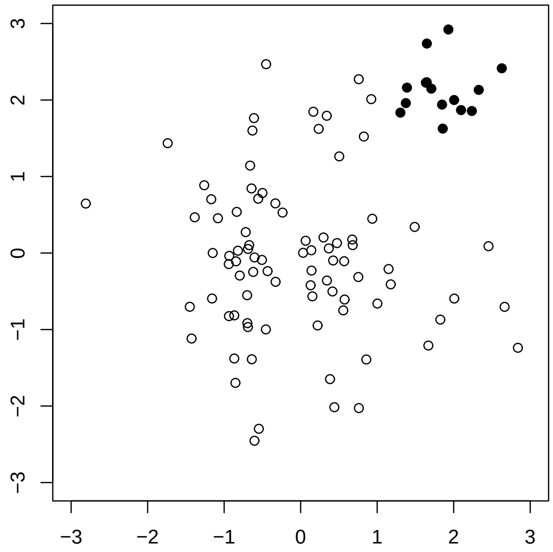

Figure 1: Simulated bivariate data with outliers shown as dark points.

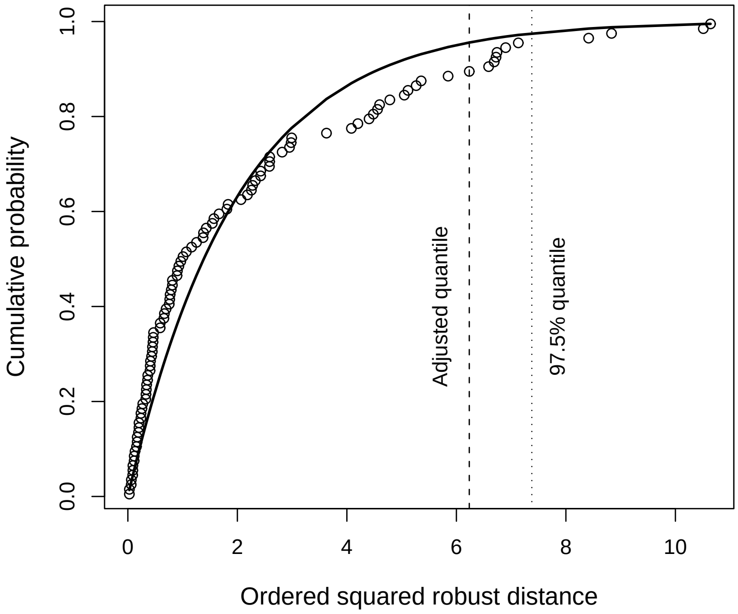

Figure 2: Empirical distribution function of the ordered squared distances (points) and theoretical distribution function of χ^2_2 (solid line).

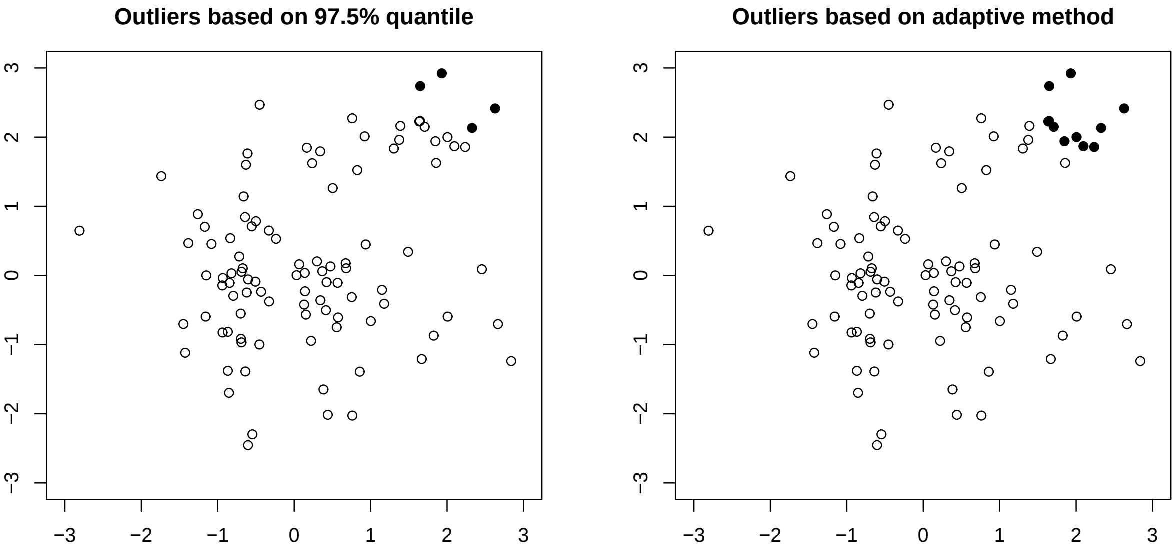

Figure 3: Resulting outliers (dark points) according to $\chi^2_{2;0.975}$ (left) and the adaptive quantile (right).

5 Summary

An automated method to identify outliers in multivariate space was developed. The method compares the difference between the empirical distribution of the squared robust distances and the distribution function of the chi-square distribution. The method accounts not only for different dimension of the data but also for different sample size.

References

[1] Filzmoser P., Reimann C., Garrett R.G. (2003). Multivariate outlier detection in exploration geochemistry. Technical report TS 03-5, Department of Statistics, Vienna University of Technology, Austria. Dec. 2003.

[2] Garrett R.G. (1989). The chi-square plot: A tool for multivariate outlier recognition. *Journal of Geochemical Exploration*. Vol. **32**, pp. 319-341.

[3] Gervini D. (2003). A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *Journal of Multivariate Analysis*. Vol. **84**, pp. 116-144.

[4] Rousseeuw P.J., Van Driessen K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. Vol. **41**, pp. 212-223.

[5] Rousseeuw P.J., Van Zomeren B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*. Vol. **85**(411), pp. 633-651.