

# An Adaptive Method for Multivariate Outlier Detection

P. Filzmoser<sup>1</sup>

<sup>1</sup> Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria. E-mail: P.Filzmoser@tuwien.ac.at

**Keywords:** Outliers, Robustness, Visualization, Adaptive method.

## Abstract

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  $\pm 2$ -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  $\chi^2$  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  $\chi^2$  plot, which draws the empirical distribution function of the robust Mahalanobis distances against the  $\chi^2$  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.

In an application with data from geochemistry the usefulness of the proposed method is demonstrated. Moreover, we propose a new plot for visualizing multivariate outliers of spatial data.

## References

- R.G. Garrett. The chi-square plot: A tool for multivariate outlier recognition. *Journal of Geochemical Exploration*, 32, 319–341, 1989.
- D. Gervini. A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, 84, 116–144, 2003.
- P.J. Rousseeuw and B.C. Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633–651, 1990.