Outlier Detection with Application to Geochemistry

Peter Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology

Keywords: Outliers, Robustness, Multivariate methods, Extremes

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 $\pm 2\cdot$ scatter) outliers are identified even for "clean" data, or at least no distinction is made between outliers and extremes of a distribution (Reimann, Filzmoser, and Garrett, 2005).

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 tails of the distribution is an indication for outliers, and values beyond this break are iteratively deleted. Gervini (2003) used this idea and compared theoretical and empirical distribution function in the tails to define the proportion of outliers in the data. In a further development, Filzmoser, Garrett, and Reimann (2005) adjusted the adaptive method of Gervini (2003) to sample size and dimensionality. It turns out that the resulting outlier detection method is not very sensitive with respect to the choice of tuning parameters (Filzmoser, 2005). The method has been implemented in `R` in the package *mvoutlier*.

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.

P. Filzmoser. Identification of multivariate outliers: a performance study. *Austrian Journal of Statistics*, 34(2):127–138, 2005.

P. Filzmoser, R.G. Garrett, and C. Reimann. Multivariate outlier detection in exploration geochemistry. *Computers and Geosciences*, 31:579–587, 2005.

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.

C. Reimann, P. Filzmoser, and R.G. Garrett. Background and threshold: critical comparison of methods of determination. *Science of the Total Environment*, 346:1–16, 2005.

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.