Detecting anomalous data cells

P.J. Rousseeuw[1*] and W. Van den Bossche[1]

[1] *KU Leuven, Belgium; peter@rousseeuw.net; W.VandenBossche@wis.kuleuven.be*
[*] *Presenting author*

Keywords. *Cellwise outliers; High dimension; Missing values; Multivariate statistics; Rowwise outliers.*

A multivariate dataset consists of n cases in d dimensions, and is often stored in an n by d data matrix. It is well-known that real data may contain outliers. Depending on the circumstances, outliers may be (a) undesirable errors which can adversely affect the data analysis, or (b) valuable nuggets of unexpected information. In statistics and data analysis the word outlier usually refers to a row of the data matrix, and the methods to detect such outliers only work when at most 50% of the rows are contaminated. But often only one or a few cells (coordinates) in a row are outlying, and they may not be found by looking at each variable (column) separately. We propose the first method to detect anomalous data cells which takes the correlations between the variables into account. It has no restriction on the number of contaminated rows, and can deal with high dimensions. Other advantages are that it provides estimates of the 'expected' values of the outlying cells, while imputing the missing values at the same time. We illustrate the method on several real data sets, where it uncovers more structure than found by purely columnwise methods or purely rowwise methods. Following the approach of **?**, the proposed method can also serve as an initial step for estimating multivariate location and scatter matrices.

References

Agostinelli, C., Leung, A., Yohai, V.J. & Zamar, R.H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, **24**, 441–461.

Rousseeuw, P.J. & Van den Bossche, W. (2016). Detecting anomalous data cells. *arXiv*:1601.07251 .