A spatially smoothed MRCD estimator for local outlier detection

P. Puchhammer[1*] and P. Filzmoser[1]

[1] *TU Wien; patricia.puchhammer@tuwien.ac.at, peter.filzmoser@tuwien.ac.at.*
[*] *Presenting author*

Keywords. *Local outlier detection; Multivariate data; Spatial data; MRCD estimation.*

Many methods are available for multivariate outlier detection but until now only a hand full are developed for spatial data where there might be observations differing from their neighbors, so-called local outliers. Although there are methods based on a pairwise Mahalanobis distance approach, the type of the covariance matrices used is not yet agreed upon. For example, Filzmoser et al. [2013] propose a global covariance while Ernst and Haesbroeck [2016] suggest a very local structure by estimating one covariance matrix per observation.

To bridge the gap between the global and local approach by providing a refined covariance structure we develop spatially smoothed covariance matrices based on the MRCD estimator [Boudt et al., 2020] for pre-defined neighborhoods $a_1, \ldots, a_N$. As well known from the MCD literature, a subset of observations, the so-called H-set, is obtained by optimizing an objective function. In our case we obtain a set of optimal H-sets $\mathcal{H} = (H_1, \ldots, H_N)$ from minimizing the objective function

$$f(\mathcal{H}) = \sum_{i=1}^{N} \det \left((1-\lambda)\boldsymbol{K}_i(\mathcal{H}) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}) \right).$$

While $\boldsymbol{W} = (w_{ij})_{i,j=1,\ldots,N}$ represents the closeness of the neighborhoods, the parameter λ is essential for the degree of locality of the covariance matrices. The local covariance matrices $\boldsymbol{K}_i(\mathcal{H})$ are based on the MRCD convex combination of the sample covariance matrix of an H-set of the neighborhood a_i and a global target matrix. For the optimal set of H-sets $\mathcal{H}^* = (H_i^*)_{i=1,\ldots,N}$ of the objective function, the final covariance estimate for neighborhood a_i is defined as $\hat{\boldsymbol{\Sigma}}_{SSM,i} = (1-\lambda)\boldsymbol{K}_i(\mathcal{H}^*) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^*)$.

A heuristic algorithm based on the notion of a C-step is developed to find the optimal set of H-sets which also shows stable convergence properties in general. We demonstrate the applicability of the new covariance estimators and the importance of a compromise between locality and globality for local outlier detection with simulated and real world data, and compare the performance with other state-of-the-art methods from statistics and machine learning.

Acknowledgements: This project has received funding by the European Commission within the Horizon 2021 programme under grant agreement ID 101057741.

References

Filzmoser, P., Ruiz-Gazen, A., and Thomas-Agnan, C. (2013). Identification of local multivariate outliers. *Statistical Papers*, **55**, 29–47.

Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, **30**, 113–128.

Ernst, M. and Haesbroeck, G. (2016). Comparison of local outlier detection techniques in spatial multivariate data. *Data Mining and Knowledge Discovery*, **31**, 371–399.