DSSV 2023
Data Science, Statistics & Visualisation

ECDA 2023
European Conference on Data Analysis

L0 Regularized Cellwise Outlier Detection and Covariance Estimation

M. Mayrhofer[1], C. Rieser[1], and P. Filzmoser[1]

[1]TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria

Robust parameter estimation and outlier detection are of critical importance in statistics and computer science. Commonly, multivariate outlier detection methods focus on detecting outlying rows in the data matrix. The more recent cellwise contamination paradigm describes the settings where individual cells deviate from the values they should have had [1, 2].

We assume that the multivariate observations $\boldsymbol{x} \in \mathbb{R}^p$ are generated by the model

$$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{b} \odot \boldsymbol{z}, \tag{1}$$

where $\odot$ denotes the coordinatewise product. The random vectors $\boldsymbol{y} \in \mathbb{R}^p, \boldsymbol{b} \in \mathbb{R}^p$, and $\boldsymbol{z} \in \mathbb{R}^p$ in model (1) are independent, $\boldsymbol{y}$ has a multivariate normal distribuiton with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the cellwise contamination indicators $b_j, j = 1, \ldots, p$ are Bernoulli iid with probability q, and the entries $z_j, j = 1, \ldots, p$ are iid and have a non-informative outlier generating distribuiton. We propose a method for cellwise outlier detection and cellwise robust location and covariance estimation based on the cellwise contamination model (1). From this model, we can deduce the constrained optimization problem (2), which is based on maximum a-posteriori estimation (MAP). Let $\boldsymbol{x}_i \in \mathbb{R}^p$ denote the observations, $\boldsymbol{\mu} \in \mathbb{R}^p$ the location, $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ the covariance, and $\boldsymbol{s}_i = (s_{i1}, \ldots, s_{ip})' \in \mathbb{R}^p$ the cellwise corrections, then the optimization problem is given as

$$\min_{\boldsymbol{s}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}} n \log(\det(\boldsymbol{\Sigma})) + \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{s}_i)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{s}_i) + \lambda \sum_{i=1}^{n} \sum_{j=1}^{p} \mathbb{1}\{s_{ij} \neq 0\}, \tag{2}$$

where $\mathbb{1}(A)$ is one if condition A is fulfilled and zero otherwise. Since the optimization problem is non-convex, solving it poses computational challenges. However, we propose an approach in which we alternate between location and covariance estimation, determining the cellwise corrections based on recently developed ideas in l0 penalized regression solved by a coordinate descent approach as in [3]. Since we alternate between estimation and correction steps we can employ regularization techniques for the covariance estimation to provide robust estimates for high dimensional settings.

We want to highlight the versatility of deducing robust estimators from model (1) using MAP. For example, if we consider the case where $\boldsymbol{b}$ is a vector of either all ones or zeros in model (1), we obtain a casewise contamination model. With an appropriate data-driven choice of λ, one can then obtain a version of the Minimum Covariance Determinant (MCD) estimator [4] from problem (2).

In the presentation, we provide illustrations of the performance of this procedure based on simulated and real-world datasets.

[1] F. Alqallaf, S. Van Aelst, V. J. Yohai, and R. H. Zamar, "Propagation of outliers in multivariate data," *The Annals of Statistics*, pp. 311–331, 2009.

[2] P. J. Rousseeuw and W. V. D. Bossche, "Detecting deviating data cells," *Technometrics*, vol. 60, no. 2, pp. 135–145, 2018.

[3] H. Hazimeh and R. Mazumder, "Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms," *Operations Research*, vol. 68, no. 5, pp. 1517–1537, 2020.

[4] P. Rousseeuw, "Multivariate estimation with high breakdown point," *Mathematical Statistics and Applications Vol. B*, pp. 283–297, 01 1985.