Outlier Detection in High-Dimensional Datasets Using the MRCD Estimator

Yuanhong Wu* and Michael Pokojovy*,†,‡

*Department of Mathematical Sciences
†Data Science Program
‡Computational Science Program
The University of Texas at El Paso, El Paso, TX

Abstract

The minimum covariance determinant (MCD) estimator, first introduced by Rousseeuw (1984), is a highly robust estimator of multivariate location and scatter parameters. Multiple algorithmic implementations of the MCD estimator are known in the literature, including the well-known FastMCD algorithm of Rousseeuw & Van Driessen (1999). One major limitation of the MCD estimator is that it cannot be applied to genuinely high-dimensional datasets as the sample size (n) needs to greatly the number of variables (p). To extend the MCD estimator to the high-dimension context, i.e., $p \gg n$, Boudt et al (2020) proposed their minimum regularized covariance determinant (MRCD) estimator along with an efficient computational routine implementing the former. The MRCD estimator aims to minimize the determinant of a regularized covariance estimate over all subsamples of given size in lieu of minimizing the usual covariance determinant to avoid singularity when the dimension exceeds the desired subset size. We give a brief review of MCD and MRCD estimators, their theoretical properties as well as some known algorithmic techniques to compute these estimators. Further, we illustrate how the MRCD estimator can be applied to outlier detection using two high-dimensional real-world datasets.

Keywords: High-dimensional data analysis; outlier detection; minimum regularized covariance determinant estimator