SEMINARIO

Lucía Trapote Reglero

Universidad de Valladolid

Robust clustering based on trimming with increasing dimensionality

Abstract: Outliers are known to significantly distort the results of many widely applied Cluster Analysis methods. While increasing the number of clusters to be detected might seem like a straightforward way to accommodate outliers, this strategy is often ineffective and frequently impractical. To address this issue, robust clustering techniques have been developed, which not only mitigate the influence of outliers on clustering results but also help detect meaningful anomalies in data, especially when subpopulations are naturally present.

This presentation focuses on robust clustering methods based on trimming, with particular emphasis on TCLUST, an extension of the Minimum Covariance Determinant (MCD) method designed for Cluster Analysis. TCLUST is highly effective for low-dimensional datasets, but its performance deteriorates in high-dimensional settings due to the complexity involved in estimating the scatter matrices of multiple components. Although constraints on the eigenvalues of the scatter matrices can help, such an approach forces clusters to be nearly spherical with equal dispersion.
An alternative approach is the Robust Linear Grouping (RLG) algorithm, which assumes that clusters lie around lower-dimensional affine subspaces, combining clustering with dimensionality reduction. However, the RLG approach assumes errors orthogonal to the approximating subspaces and can face challenges with intersecting subspaces.

A new approach will also be presented that offers a compromise between TCLUST and RLG. This compromise results from robustifying the High Dimensional Data Clustering (HDDC) approach, already available in the literature, through the implementation of trimming and the enforcement of additional constraints on eigenvalues. The HDDC approach assumes parsimonious conditions on the scatter matrices, making it tractable in higher dimensions, but can benefit from convenient robustification. An algorithm implementing this novel robust approach will be introduced, along with illustrative examples and diagnostics for detecting outlyingness of interest. In this algorithm, it is important to provide adequate random initialization and to robustly estimate the intrinsic dimensions of the approximating affine subspaces.

Seminario del IMUVA, edificio LUCIA
Martes 1 de Julio de 2025 (10:35)