Abstract

Multi-group or multi-source data, in which observations are partitioned into groups by external variables, arise in a wide range of disciplines. Examples include spatial data grouped by proximity, country borders, or geological units; medical data categorized by diagnosis, disease, or age; and temporal data structured by days, months, or years. These groupings are typically associated with continuous variables and reflect inherent relationships among the groups – making separate analysis inappropriate.

Outliers can have a substantial impact on classical, non-robust statistical methods, often distorting results and leading to misleading interpretations if not properly addressed. This issue becomes particularly critical in complex data structures such as multi-group or spatial data, where outliers may remain hidden and bias outcomes more easily. Detecting both classical outliers and those specific to the multi-group or spatial context is essential for producing reliable estimates. Moreover, analyzing these outliers can offer valuable insights, such as the detection of mislabeling or, in the case of geochemical spatial data, the identification of regions of potential mineralization.

This thesis develops and adapts robust statistical methods for application in multi-group settings. Key contributions include the development of a robust, smoothed covariance estimator for spatial and multi-source data – applied to local outlier detection – and its use in geochemical exploration. Furthermore, a sparse multi-group principal component analysis (PCA) framework is proposed, enabling joint analysis of global and group-specific features. Finally, a cellwise robust Gaussian mixture model (GMM) is introduced for the multi-group context, allowing for the detection of transitional group outliers. These theoretical and methodological advances significantly extend the robust statistics toolbox, providing improved analytical frameworks for multi-group data and demonstrating strong performance in both simulation studies and real-world applications.