Subspace Recovery in Winsorized PCA

Sangil Han
Seoul National University

We explore the theoretical properties of subspace recovery using Winsorized Principal Component Analysis (WPCA), utilizing a common data transformation technique that caps extreme values to mitigate the impact of outliers. Despite the widespread use of winsorization in various tasks of multivariate analysis, its theoretical properties, particularly for subspace recovery, have received limited attention. We provide a detailed analysis of the accuracy of WPCA, showing that increasing the number of samples while decreasing the proportion of outliers guarantees the consistency of the sample subspaces from WPCA with respect to the true population subspace. Furthermore, we establish perturbation bounds that ensure the WPCA subspace obtained from contaminated data remains close to the subspace recovered from pure data. Additionally, we extend the classical notion of breakdown points to subspace-valued statistics and derive lower bounds for the breakdown points of WPCA. Our analysis demonstrates that WPCA exhibits strong robustness to outliers while maintaining consistency under mild assumptions. A toy example is provided to numerically illustrate the behavior of the upper bounds for perturbation bounds and breakdown points, emphasizing winsorization's utility in subspace recovery.