Robust Spectral Clustering with Rank Statistics

Joshua Cape, Xianshi Yu, Jonquil Zhongling Liao

Abstract

This talk investigates the performance of a robust spectral clustering method for latent structure recovery in noisy data matrices. We consider eigenvector-based clustering applied to a matrix of nonparametric rank statistics that is derived entrywise from the raw, original data matrix. This approach is robust in the sense that, unlike traditional spectral clustering procedures, it can provably recover population-level latent block structure even when the observed data matrix includes heavy-tailed entries and has a heterogeneous variance profile. Here, the raw input data may be viewed as a weighted adjacency matrix whose entries constitute links that connect nodes in an underlying graph or network.

Our main theoretical contributions are threefold and hold under flexible data generating conditions. First, we establish that robust spectral clustering with rank statistics can consistently recover latent block structure, viewed as communities of nodes in a graph, in the sense that unobserved community memberships for all but a vanishing fraction of nodes are correctly recovered with high probability when the data matrix is large. Second, we refine the former result and further establish that, under certain conditions, the community membership of any individual, specified node of interest can be asymptotically exactly recovered with probability tending to one in the large-data limit. Third, we establish asymptotic normality results associated with the truncated eigenstructure of matrices whose entries are rank statistics, made possible by synthesizing contemporary entrywise matrix perturbation analysis with the classical nonparametric theory of so-called simple linear rank statistics. Collectively, these results demonstrate the statistical utility of rank-based data transformations when paired with spectral techniques for dimensionality reduction. Numerical examples illustrate and support our theoretical findings. Additionally, for a dataset consisting of human connectomes, our approach yields parsimonious dimensionality reduction and improved recovery of ground-truth neuroanatomical cluster structure. We conclude with a discussion of extensions, practical considerations, and future work.

Reference: https://arxiv.org/abs/2408.10136, to appear in *Journal of Machine Learning Research*.

Author's note: As a statistician working on entrywise eigenvector perturbation analysis and with a background in applied mathematics, I am eager to engage with the numerical linear algebra community towards advancing research on topics of shared interest.