Advances in Clustering High Dimensional Functional Data

Thaddeus Tarpey*
Wright State University, Dayton, Ohio, thaddeus.tarpey@wright.edu

Eva Petkova
New York University, New York, NY

Interesting patterns of heterogeneity that exist in high dimensional data may be difficult to discover with classic methods of unsupervised learning. A popular approach to unsupervised learning is cluster analysis where the goal is to find an optimal partition the data into homogeneous and hopefully interpretable subgroups. Model-based clustering approaches based on finite mixtures are often based on speculative assumptions and can lead to misleading results. Alternatively, an optimal partitioning of the data via the well-known *k*-means clustering algorithm can be used instead. However, *k*-means clustering is nonparametric and susceptible to missing interesting patterns in the data. In particular, the *k*-means algorithm tends to place cluster means in the direction of primary variability which corresponds to the first few principal component directions. In this talk, we introduce advances in clustering based on determining linear transformations of the data to steer the clustering algorithm in the direction of finding interesting patterns. In particular, instead of considering the classic principal component transformation, we explore transformations based on independent component analysis. Additionally, the clustering algorithm will be enhanced by considering canonical linear transformations that minimize within cluster variability relative to between cluster variability in the iterative steps of the algorithm. This will be particularly interesting for clustering functional and image data. For instance, for longitudinal data, the basis functions used to represent trajectories over time corresponds to different linear transformations of the data. Longitudinal applications to diseases such as depression can yield trajectory cluster means corresponding to clinically distinct treatment responses over time. The clustering methodology can be further enhanced by incorporating baseline covariates and adapting the clustering methods to handle random effects.

Keywords: canonical transformation, k-means algorithm, principal components