STRUCTURED SPARSE LATENT VARIABLE MODELS FOR OVERLAPPING CLUSTERING WITH LOVE

FLORENTINA BUNEA
DEPARTMENT OF STATISTICAL SCIENCE
CORNELL UNIVERSITY

Abstract: The problem of overlapping variable clustering is that of finding overlapping sub-groups of a p-dimensional random vector X, from a sample of size n of observations on X. This problem is ubiquitous in data science, and algorithms abound. However, much less is known about the statistical guarantees of the estimated clusters, as most algorithms are not model-based.

In this talk I will introduce a novel method, **LOVE**, based on a sparse **L**atent factor model with pure variables, for **OVE**rlapping clustering with statistical guarantees. The model is used to define the population level clusters as groups of those components of X that are associated, via a sparse and structured allocation matrix, with the same unobservable latent factor, and multi-factor association is allowed. Clusters are anchored by a few components of X, called pure variables, that are respectively associated with only one latent factor. This renders structure to the allocation matrix. We prove that the existence of pure variables is a sufficient, and almost necessary, assumption for the identifiability of the allocation matrix in sparse latent factor models. Consequently, model-based clusters can be uniquely defined, and provide a bona fide estimation target. Our identifiability arguments extend to general structured sparse latent models those encountered in the related, but different, areas of non-negative matrix factorization and topic modeling.

Our model formulation allows the latent factors to be correlated. Also, both p and the unknown number of clusters K is allowed to grow with the sample size n. LOVE estimates first the set of pure variables, and the number of clusters, via a novel method that has low computational complexity, $O(p^2)$. Each cluster, anchored by pure variables, is then further populated with components of X according to the sparse estimates of the allocation matrix. The latter are obtained via a computationally efficient estimation method tailored to the structure of this problem. The combined procedure yields rate-optimal estimates of the allocation matrix and consistent estimators of the number of clusters. This analysis is performed with minimal signal strength conditions, under which we further guarantee cluster recovery with zero false positive rate, and with false negative rate control. Under stronger signal strength assumptions, the model-based clusters are recovered exactly by our method. The practical relevance of LOVE is illustrated through the analysis of an RNA-seq data set, devoted to determining the functional annotation of genes with unknown function.