SANJEENA DANG, Carleton University

Clustering matrix-variate count data

Three-way data structures or matrix-variate data are frequent in biological studies. In RNA sequencing, three-way data structures are obtained when high-throughput transcriptome sequencing data are collected for n genes across p conditions at r occasions. Matrix variate distributions offer a natural way to model three-way data and mixtures of matrix variate distributions can be used to cluster three-way data. Clustering of gene expression data is carried out as means of discovering gene co-expression networks. In this work, a mixture of matrix variate Poisson-log normal distributions is proposed for clustering read counts from RNA sequencing. By considering the matrix variate structure, the number of covariance parameters to be estimated is reduced and the components of resulting covariance matrices provide a meaningful interpretation. We propose three different frameworks for parameter estimation - a Markov chain Monte Carlo based approach, a variational Gaussian approximation-based approach, and a hybrid approach. The models are applied to both real and simulated data, and we demonstrate that the proposed approaches can recover the underlying cluster structure. In simulation studies where the true model parameters are known, our proposed approach shows good parameter recovery.