
SANJEENA DANG, Carleton University

Clustering compositional data with a logistic normal multinomial mixture model with an underlying latent factor structure

The human microbiome plays a crucial role in health and disease. Advances in next-generation sequencing technologies have made it possible to quantify microbiome composition with high resolution. Clustering microbiome data can uncover meaningful patterns across samples, offering insights into biological variability and disease mechanisms. However, this task presents several challenges. Microbiome data are typically high-dimensional, over-dispersed, and compositional, reflecting relative abundances. As such, analyzing such compositional data presents many challenges because they are restricted to a simplex, which complicates standard statistical analysis. Here, we develop a family of logistic normal multinomial factor analyzers (LNM-FA) by incorporating a factor analyzer structure. The family of models is suitable for high-dimensional microbiome data, as the number of parameters in LNM-FA can be greatly reduced by assuming that the underlying latent factors are small. Parameter estimation is done using a computationally efficient variant of the alternating expectation conditional maximization algorithm that utilizes variational Gaussian approximation. The proposed method is illustrated using simulated and real datasets.