

Clustering time-course gene expression data using an infinite mixture prior for marginal model

T. Baghfalaki¹, M. Ganjali²

¹Department of Statistics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran.

²Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran.

Identifying patterns of time-course gene expression data by using cluster analysis is an important problem in biomedical research. The aim of this study is to stratify the genes according to their differential temporal behavior. In this paper, we present a non-parametric Bayesian approach by considering a Dirichlet process for the mean of a marginal model using a first order autoregressive within-gene covariance matrix for the data with equally spaced time intervals. Also, the exponential covariance function as a generalization of the AR(1) model for the data with unequally spaced time intervals is applied. We call this method CTGE-IM4 for Clustering Time-course Gene Expression data using an Infinite Mixture Model prior for Marginal Model. We use the posterior similarity matrix, posterior probability that two genes belong to the same cluster, as a measure of the similarity of the gene expression profiles for clustering genes. We compare the performance of CTGE-IM4 for clustering different patterns of time-course gene expression with some benchmark methods by adjusted Rand index and the number of identified clusters in the simulation studies. To illustrate the CTGE-IM4 method, we apply it to cluster analysis of two public time-course gene expression data.

Keywords: Bayesian non-parametric; Clustering; Dirichlet process; Marginal model; Time-course gene expression data.