Statistica Sinica **32** (2022), 1835-1856
doi:https://doi.org/10.5705/ss.202020.0337

NONPARAMETRIC BAYESIAN TWO-LEVEL CLUSTERING FOR SUBJECT-LEVEL SINGLE-CELL EXPRESSION DATA

Qiuyu Wu and Xiangyu Luo

Renmin University of China

Abstract: The advent of single-cell sequencing opens new avenues for personalized treatment. In this study, we address a *two-level clustering* problem of simultaneous subject subgroup discovery (*subject level*) and cell type detection (*cell level*) for single-cell expression data from multiple subjects. Current statistical approaches either cluster cells without considering the subject heterogeneity, or group subjects without using the single-cell information. To bridge the gap between cell clustering and subject grouping, we develop a nonparametric Bayesian model, Subject and Cell clustering for Single-Cell expression data (SCSC) model, to achieve subject and cell grouping simultaneously. The SCSC model does not need to prespecify the subject subgroup number or the cell type number. It automatically induces subject subgroup structures and matches cell types across subjects. Moreover, it directly models the single-cell raw count data by deliberately considering the data's dropouts, library sizes, and over-dispersion. A blocked Gibbs sampler is proposed for the posterior inference. Simulation studies and an application to a multi-subject induced pluripotent stem cell single-cell RNA sequencing data set validate the ability of the SCSC model to simultaneously cluster subjects and cells.

Key words and phrases: Markov chain Monte Carlo, mixture of mixtures, model-based clustering, nonparametric Bayes, single-cell RNA sequencing.

1. Introduction

Advancements in biological sequencing technology, such as single-cell RNA-sequencing (scRNA-seq), have enabled the expression profiling of single cells. ScRNA-seq data are often organized into a data matrix, illustrated in Figure 1(a), where the columns are cells and the rows represent genes. Based on the scRNA-seq data matrix, discovering cell types is simply formulated as a clustering problem. Going further, if we can integrate the scRNA-seq data from multiple subjects, this presents unprecedented opportunities to investigate subject heterogeneity at the single-cell resolution. Subject heterogeneity refers to human subpopulations, patient disease subtypes, or other differentiable human biological characteristics, according to different contexts. Using disease subtypes as an

---

Corresponding author: Xiangyu Luo, Institute of Statistics and Big Data, Renmin University of China, Haidian, Beijing 100872, China. E-mail: xiangyuluo@ruc.edu.cn.