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

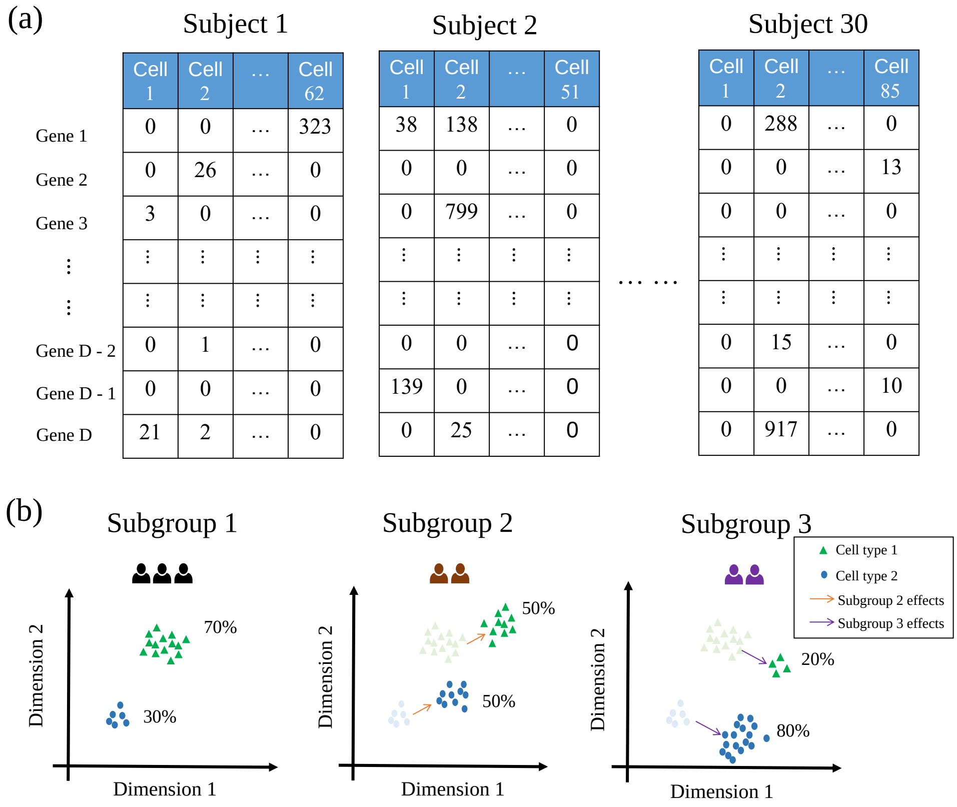
Corresponding author: Xiangyu Luo, Institute of Statistics and Big Data, Renmin University of China, Haidian, Beijing 100872, China. E-mail: xiangyuluo@ruc.edu.cn.

illustration, biological studies have found differences in tumor cell proportions among subtypes of breast cancers (Makki (2015)), lung cancers (Busch et al. (2016)), and other diseases. These subtle observations can be captured by the scRNA-seq data, but may be missed when using the traditional bulk expression data, which are the aggregated expression signals from diverse cell types. Consequently, it is imperative to employ subject-level single-expression data (Figure 1(a)) to understand cellular and subject heterogeneity.

In this study, we address a two-level clustering statistical problem by directly modeling multi-subject scRNA-seq data. An artificial demonstration of the two-level clustering is shown in Figure 1(b). At the cell level, cells with similar expression values are clustered together, and at the subject level, subjects with similar cellular distributions are grouped together. Two subjects are said to have the same cellular distribution if they share the same cell type proportions and expression levels for each cell type. In addition, to obtain valid biological results, cell types must be matched across subjects by considering the effects caused by the subject subgroups (Figure 1(b)). Note that our two-level clustering problem differs from the bi-clustering approaches (Cheng and Church (2000); Turner, Bailey and Krzanowski (2005)), which group subjects and genes using the aggregated expression data matrix.

There has been a large amount of statistical literature on cell clustering or subject clustering. On the one hand, cell clustering methods fit heterogeneous scRNA-seq data using the latent variable model (Buettner et al. (2015)), hierarchical clustering (žurauskienė and Yau (2016)), consensus approach (Kiselev et al. (2017)), or model-based mixture models (Prabhakaran et al. (2016); Sun et al. (2017); Song, Chan and Wei (2020); Liu, Warren and Zhao (2019)). Nevertheless, when applied to multi-subject scRNA-seq data, these methods do not consider subject heterogeneity, and ignore the fact that the gene expression levels may change with subjects, thus possibly leading to incorrect cell clustering results.

On the other hand, subject clustering methods are based on the aggregated expression matrix, with genes in rows and subjects in columns, where the expression vector of one subject can be viewed as the row averages of the subject's gene-cell expression matrix in Figure 1(a). Pan and Shen (2007) adopted a normal mixture model and developed an L_1-penalized expectation-maximization algorithm to distinguish subjects and detect differentially expressed (DE) genes. Wang and Zhu (2008) instead used the L_∞ and hierarchical penalties to refine the clustering results. The sparse k-means proposed by Witten and Tibshirani (2010) simultaneously extracted a few DE genes and grouped subjects by maximizing the weighted between-cluster sum-of-squares. Huo et al. (2016) subsequently



Subject 1

	Cell 1	Cell 2	…	Cell 62
Gene 1	0	0	…	323
Gene 2	0	26	…	0
Gene 3	3	0	…	0
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
Gene D - 2	0	1	…	0
Gene D - 1	0	0	…	0
Gene D	21	2	…	0

Subject 2

	Cell 1	Cell 2	…	Cell 51
Gene 1	38	138	…	0
Gene 2	0	0	…	0
Gene 3	0	799	…	0
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
Gene D - 2	0	0	…	0
Gene D - 1	139	0	…	0
Gene D	0	25	…	0

Subject 30

	Cell 1	Cell 2	…	Cell 85
Gene 1	0	288	…	0
Gene 2	0	0	…	13
Gene 3	0	0	…	0
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
Gene D - 2	0	15	…	0
Gene D - 1	0	0	…	10
Gene D	0	917	…	0

Figure 1. Artificial illustration of the data structure and study goal. (a) Subject-level single-cell expression data. (b) An illustration of a two-level clustering problem. In subgroup 1, cell type 1 proportion is 70%, shown in green triangles, and cell type 2 proportion is 30%, shown in blue dots. Compared with subgroup 1, the cellular distribution in subgroup 2 can change in two ways: cell proportions and cell locations. For a good visualization, only two gene dimensions are illustrated (expression in log scale). The orange and purple arrows represent the effects of subgroups 2 and 3, respectively, when subgroup 1 is treated as a reference.

generalized the sparse k-means to expression data from multiple studies. Luo and Wei (2019) proposed a more efficient and flexible Bayesian framework to conduct integrative subject clustering. Because these methods do not employ single-cell expression information, subtle differences (e.g., cellular composition changes) cannot be detected.

All the methods mentioned above, except that of Prabhakaran et al. (2016), require predetermination of the number of clusters and trials of multiple choices, which may be practically difficult and computationally expensive. The Dirichlet process (DP) is a nonparametric Bayesian prior (Ferguson (1973); Sethuraman

(1994)), and is well known for its flexibility in automatically selecting the number of clusters in a data-driven manner. However, the DP only addresses one-level clustering, motivating two extensions—the hierarchical DP (HDP) (Teh et al. (2006)) and the nested DP (NDP) (Rodriguez, Dunson and Gelfand (2008))—that are close to our two-level clustering problem. Unfortunately, using the terms in our context, the HDP assigns a cell mixture distribution to each subject, but with different mixture weights; thus, the subjects cannot form a group structure. Although the NDP promotes the subject group structure, subjects in different groups do not share cell components, causing difficulty in matching cell types across subjects. In other words, if two distributions from the NDP share one cell component, the two distributions must be the same almost surely, which is not realistic in our problem. To deal with the degeneracy issue of the NDP, Camerlenghi et al. (2019) developed a latent nested nonparametric prior that allows common and group-specific cell types across subject subgroups, but their method meets practical computational challenges when applied to more than two subject subgroups or to high-dimensional expression data. When more than two subject subgroups need to be considered, Beraha, Guglielmi and Quintana (2021) extended the HDP to the semi-HDP to induce subject dependence and grouped distributions using a finite-dimensional distribution over cluster indicators.

Actually, in the discussion of the NDP paper (Rodriguez, Dunson and Gelfand (2008)), James (2008) has constructed a fully nonparametric prior to combine the NDP and the HDP that can address the degeneracy problem of the NDP and achieve two-level clustering for nested data. We follow Section 4 in his discussion and call his prior the *hybrid NDP-HDP prior*. In the field of text analysis, the hybrid prior has been employed to conduct entity-topic modeling (Tekumalla, Agrawal and Bhattacharya (2015)), and its multi-level extension introduced in (Paisley et al. (2015)) allows for tree-structured topic hierarchies. Recently, Denti et al. (2020) proposed a common atoms model built upon a similar nonparamemtric prior to analyze microbiome data. The model does not introduce an additional HDP part, but constrains the common atoms of the sampled distributions.

To the best of our knowledge, there is no statistical approach to simultaneously tackle subject and cell clustering on multi-subject scRNA-seq data. For the two-level clustering part, we take advantage of the hybrid NDP-HDP prior (James (2008)), inducing shared components for cells and group structures for subjects. For the data modeling part, we exploit the zero-inflated Poisson-log-normal (ZIPLN) distribution with a Probit dropout mechanism, which accounts for the zero-inflation, over-dispersion, and count nature of the scRNA-seq data.

Integrating the nonparametric Bayesian prior with the ZIPLN distribution results in the proposed model, Subject and Cell clustering for Single-Cell expression data (SCSC) model. This model enables simultaneous subject and cell clustering for scRNA-seq raw count data, and does not require specifying the subject or cell cluster numbers in advance. For the posterior inference of the SCSC model, we designed an efficient blocked Gibbs sampler (Ishwaran and James (2001)) based on an approximation to the SCSC model. The approximation accuracy is guaranteed theoretically, as long as the truncation levels and related parameters are chosen appropriately.

The remainder of this paper is organized as follows. Section 2 presents a brief review of the DP and its two extensions, the HDP and the NDP, which are prerequisites to introducing the hybrid NDP-HDP prior that enjoys the strengths of the HDP and the NDP. In Section 3, we bring in the hybrid NDP-HDP prior, derive theoretical results about the distributions sampled from the prior, and present the SCSC model built on the hybrid prior and tailored to the scRNA-seq data. In Section 4, we introduce a truncated SCSC model to ease the posterior computing, and provide a theorem to quantify its approximation error. An efficient posterior sampling scheme for the SCSC model is discussed in Section 5, and the model is applied to synthetic and real-world data in Section 6. Finally, we conclude the paper in Section 7.

2. Preliminaries on Nonparametric Priors

Suppose that the scRNA-seq data are collected for m subjects, with subject j having n_j sequenced cells in some tissue, and in each cell, the expression levels for D genes are measured. We denote the observed read count mapped to gene g in cell i for subject j by $X_{gi}^{(j)}$. All the read counts for subject j can be wrapped up using a data matrix $\mathbf{X}^{(j)}$ with D genes in rows and n_j cells in columns. To describe the subject heterogeneity, we assume that subjects can be separated to form several subgroups, where subjects in the same subgroup share similar characteristics, and subjects in different subgroups have distinct features. We use $S^{(j)}$ to represent the subgroup to which subject j belongs. Similarly, the cell heterogeneity is characterized by cell types, and the cell type of cell i for subject j is denoted by $C_i^{(j)}$. Here, $\mathbf{X}^{(j)}$ are observed, but the subject subgroup and cell type indicators need to be estimated.

2.1. Dirichlet process

The DP mixture model (Lo (1984)) based on the DP prior (Ferguson (1973)) can be considered a generalized version of the finite-mixture model. For notational simplicity, we temporarily consider only the cell data from subject 1 and let the gene number D be one. Thus, the column vectors $\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_{n_1}^{(1)}$ of $\mathbf{X}^{(1)}$ can be simplified to univariate samples $X_1, \ldots, X_{n_1}$, and the cell type indicators $C_i^{(1)}$ simplify to C_i. The finite-mixture model allocates each cell to one of K cell types, with the probability of cell type k being π_k; that is, $\mathbb{P}(C_i = k) = \pi_k$ and $\sum_{k=1}^{K} \pi_k = 1$. Given that cell i is assigned to cell type k, X_i is assumed to be from the distribution $f(x|\mu_k)$, where f is a probability density (or mass) function, which will be specified in the next section, and μ_k is a parameter describing the cell-type-k effect. Usually, the total cell type number K is unknown to data analysts, and it is challenging to accurately estimate its value. The DP mixture overcomes this challenge by generalizing K to infinity and allowing finite nonempty components, thereby not requiring a prespecification of K.

The DP is constructed using the stick-breaking process (Sethuraman (1994)). Imagine that we have a stick of length 1 unit, and we intend to break this stick into infinite pieces. We first sample a value ψ_1 from the beta distribution $\text{Beta}(1, \alpha)$ ($\alpha > 0$), and then cut the stick at point ψ_1 away from its left endpoint. Accordingly, the piece of length $\pi_1 (:= \psi_1)$ is retained, and we continue to break the remaining stick with length $1 - \pi_1$. Once again, we generate a value ψ_2 from $\text{Beta}(1, \alpha)$, cut off ψ_2 proportion of the remaining length $1 - \pi_1$, and obtain a new piece with length $\pi_2 := (1 - \pi_1)\psi_2$. Repeating the breaking procedure on the stick, we have an infinite number of pieces, with the kth piece's length $\pi_k := (1 - \sum_{i=1}^{k-1} \pi_i) \cdot \psi_k$ ($\psi_k \sim \text{Beta}(1, \alpha)$). Each piece k is further given a mark (parameter) μ_k, sampled from a distribution H. In this way, we construct a probability measure, $P = \sum_{k=1}^{\infty} \pi_k \delta_{\mu_k}$ (δ_μ indicates the Dirac measure at μ), with infinite weights $\{\pi_k\}_{k=1}^{\infty}$ and support on infinite atoms $\{\mu_k\}_{k=1}^{\infty}$. The measure P is said to be from a DP with concentration parameter α and base distribution H, written as $P \sim \text{DP}(\alpha, H)$. Under P, each cell i has the probability π_k of being from cell type k, for any positive integer k, without a constraint K.

2.2. Hierarchical Dirichlet process and nested Dirichlet process

The DP is only applicable for one-level clustering. When another subject level exists, the HDP (Teh et al. (2006)) aims to cluster cells for each subject, and is able to match cell types in different subjects. In other words, if the cell type indicators $C_{i_1}^{(j_1)}$ and $C_{i_2}^{(j_2)}$ are equal (j_1 may not be j_2), then cell i_1 in subject

j_1 and cell i_2 in subject j_2 must be from the same cell type. Assume $G^{(j)}$ is the subject-j-specific distribution having the form $\sum_{k=1}^{\infty} \pi_k^{(j)} \delta_{\mu_k^{(j)}}$, based on which the cells in subject j are clustered. To encourage a common support set across $G^{(j)}$, the HDP adopts a hierarchy structure. At the higher level $G_0 \sim \mathrm{DP}(\alpha, H)$, and then at the lower level, $G^{(j)}$ are independent and identically distributed (i.i.d.) and generated from $\mathrm{DP}(\gamma, G_0)$. Because G_0 from $\mathrm{DP}(\alpha, H)$ is a discrete distribution and plays the role of the base distribution in $\mathrm{DP}(\gamma, G_0)$, the atoms $\mu_k^{(j)}$ of the support of $G^{(j)}$ must be consistent with those of G_0. This characteristic guarantees the shared cell types across $G^{(j)}$ in the HDP.

Nevertheless, in the HDP, any two subjects have distinct cell distributions owing to different weights (cell proportions), that is, $\mathbb{P}(G^{(j_1)} = G^{(j_2)}) = 0$ if $j_1 \neq j_2$; thus, no group structure exists among subjects (Figure 2(a)). The NDP (Rodriguez, Dunson and Gelfand (2008)) permits subject grouping while clustering cells. This is achieved by replacing the base measure G_0 in $\mathrm{DP}(\gamma, G_0)$ with a Dirichlet process $\mathrm{DP}(\alpha, H)$, written as $\mathrm{DP}(\gamma, \mathrm{DP}(\alpha, H))$. Specifically, if we let $Q = \mathrm{DP}(\gamma, \mathrm{DP}(\alpha, H))$, Q takes the form of $\sum_{k=1}^{\infty} \phi_k \delta_{G_k^*}$, where the atoms of Q are not numerical values, but distributions G_k^* from $\mathrm{DP}(\alpha, H)$. Subsequently, $G^{(j)}$ are i.i.d. sampled from Q, and $\mathbb{P}(G^{(j)} = G_k^*) = \phi_k$. Rodriguez, Dunson and Gelfand (2008) showed that there is a positive probability that two distributions $G^{(j_1)}$ and $G^{(j_2)}$ are identical, thus inducing group structures for $G^{(j)}$ (Figure 2(b)). Despite the simultaneous clustering on subjects and cells enjoyed by the NDP, its assumed continuous measure H leads to distinct supports between two subject subgroups (Figure 2(b)). The distributions of the two subjects from the NDP either share all atoms in the support and cell proportions, or lack any common atom. Specifically, if $G^{(j_1)}$ and $G^{(j_2)}$ from the NDP have one shared atom, then the whole distribution $G^{(j_1)}$ is equal to $G^{(j_2)}$ almost surely. This is called the degeneracy issue of the NDP, outlined in Camerlenghi et al. (2019), which causes the difficulty of cell-type-matching for two different subject subgroups in our study.

3. The SCSC Model

The hybrid NDP-HDP prior proposed by James (2008) succeeds in promoting subject subgroups with shared cell types. The nonparametric prior is constructed by assigning a DP prior to the base measure in the NDP,

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\alpha, H), \\ G^{(j)} &\overset{i.i.d.}{\sim} \mathrm{DP}(\nu, \mathrm{DP}(\gamma, G_0)), \quad j = 1, \ldots, m. \end{aligned} \tag{3.1}$$

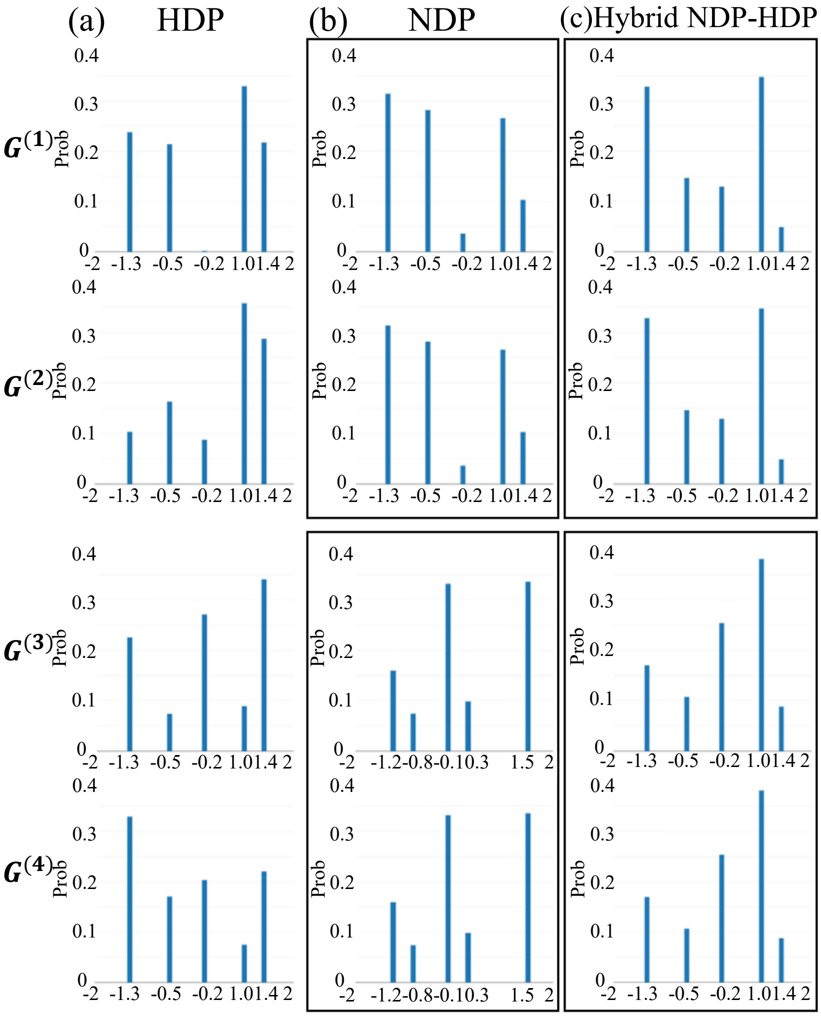


Figure 2. A simple demonstration of three nonparametric Bayesian priors: the HDP, NDP, and hybrid NDP-HDP. (a) The HDP can make subject-specific distributions $G^{(1)}$, $G^{(2)}$, $G^{(3)}$, and $G^{(4)}$ share the distribution support. However, each distribution $G^{(k)}$ has its own bar heights (weights). (b) The NDP can achieve the subject subgroup structures; however, two distributions in different subgroups do not have the same support, making it difficult to match cell types across subgroups. (c) The hybrid NDP-HDP prior not only groups subject-specific distributions, but also enables cell-type-matching between any two subject subgroups.

On the one hand, because G_0 is drawn from $\mathrm{DP}(\alpha, H)$, it has a countable support set. This property of G_0 makes the child distributions $G^{(j)}$ share the same support, thus enabling cell-type matching across subjects, an important aspect the NDP lacks. On the other hand, given G_0, the NDP helps to form

Table 1. Comparing the capabilities of the HDP, NDP, and hybrid NDP-HDP priors.

Prior	Subject subgroup structures	Shared support
HDP	×	√
NDP	√	×
Hybrid NDP-HDP prior	√	√

subgroups for subjects. Therefore, the hierarchical and nested nonparametric prior (3.1) integrates the strengths of the HDP and the NDP (Figure 2(c) and Table 1).

For the nonparametric prior (3.1), we assume the base measure H is a non-atomic probability measure on the measurable space $(U, \mathcal{B})$, where U is a D-dimensional subset of $\mathbb{R}^D$ ($U \subset \mathbb{R}^D$), $H(\{y\}) = 0$ for any $y \in U$, and $\mathcal{B}$ is the Borel σ-field of U. Denote the correlation matrix of the distribution H by $\mathbf{R}_H$. We then have the following results for the distributions $G^{(j)}$ from the prior (3.1).

Proposition 1. *For any Borel set $A \in \mathcal{B}$, we have*

(1) $\mathbb{E}\left(G^{(j)}(A)|H\right) = H(A)$.

(2) $\mathbb{V}\left(G^{(j)}(A)|H\right) = ((\alpha+\gamma+1)\,H(A)\,(1-H(A)))/((\alpha+1)\,(\gamma+1))$.

(3) $\mathrm{Cor}\left(G^{(j)}(A), G^{(j')}(A)|H\right) = (1/(1+\nu))(\nu\gamma+\alpha+\gamma+\nu+1)/(\alpha+\gamma+1)$ *for $j \neq j'$.*

(4) *When $D = 1$, let $\mu_i^{(j)}$ and $\mu_{i'}^{(j')}$ denote random variables from $G^{(j)}$ and $G^{(j')}$, respectively. The correlation between $\mu_i^{(j)}$ and $\mu_{i'}^{(j')}$ is*

$$\mathrm{Cor}\left(\mu_i^{(j)}, \mu_{i'}^{(j')}\right) = \begin{cases} \dfrac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)} & \textit{for } j = j',\ i \neq i' \\ \dfrac{\nu\gamma+\alpha+\gamma+\nu+1}{(\nu+1)(\alpha+1)(\gamma+1)} & \textit{for } j \neq j' \end{cases}.$$

(5) *When $D \geq 2$, let $\boldsymbol{\mu}_i^{(j)}$ and $\boldsymbol{\mu}_{i'}^{(j')}$ denote the random vectors from $G^{(j)}$ and $G^{(j')}$, respectively. The correlation matrix between $\boldsymbol{\mu}_i^{(j)}$ and $\boldsymbol{\mu}_i^{(j')}$ is*

$$\mathrm{Cor}\left(\boldsymbol{\mu}_i^{(j)}, \boldsymbol{\mu}_{i'}^{(j')}\right) = \begin{cases} \dfrac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}\mathbf{R}_H & \textit{for } j = j',\ i \neq i' \\ \dfrac{\nu\gamma+\alpha+\gamma+\nu+1}{(\nu+1)(\alpha+1)(\gamma+1)}\mathbf{R}_H & \textit{for } j \neq j' \end{cases}.$$

Note that when α goes to infinity, G_0 in the hybrid NDP-HDP prior approaches its centering measure H. In this limiting case $\alpha \to +\infty$, the hybrid prior degenerates to the NDP, so the results above are consistent with those for the NDP (Rodriguez, Dunson and Gelfand (2008)). The proof of the proposition can be found in Supplementary Material, Section S1.

We also tailor a zero-inflated distribution to the scRNA-seq raw count data and connect the data-modeling part to the hybrid NDP-HDP prior. One important feature of the scRNA-seq count data is that it contains a relatively large proportion of zeros compared with the bulk RNA-seq data. This zero-inflation phenomenon, also called dropouts, is mainly caused by a low number of mRNA molecules in one cell. As a result, the expression levels on some genes do not surpass the measurable threshold of the sequencing technology, thus leading to the zero observations.

To model dropout events, we assume that $Y_{gi}^{(j)}$ is the underlying true read count mapped to gene g in cell i for subject j; however, these $Y_{gi}^{(j)}$ are only partially observed through the collected data $X_{gi}^{(j)}$, owing to the dropouts. Because the probability of a dropout occurring relies on the value of $Y_{gi}^{(j)}$, (i.e., the larger the value of $Y_{gi}^{(j)}$, the less likely we are to observe a zero value), the dropout mechanism is "nonignorable," in the terminology of the field of missing data analysis,

$$X_{gi}^{(j)} = \begin{cases} 0 & \text{with probability } p(Y_{gi}^{(j)}) \\ Y_{gi}^{(j)} & \text{with probability } 1 - p(Y_{gi}^{(j)}) \end{cases}.$$

The dropout rate $p(y)$ is modeled as $\Phi(\lambda_{g0} + \lambda_{g1} \log_2(y+1))$ using a Probit link, in which $\lambda_{g1} < 0$ and Φ is the cumulative distribution function of the standard normal distribution. A negative λ_{g1} guarantees a negative correlation between y and $p(y)$, and its dependence on the gene index g accurately models the biological observation that the dropout rate may be associated with the gene's features, such as the gene length (Liu, Warren and Zhao (2019)).

Owing to the count nature and over-dispersion of the scRNA-seq data, we adopt the Poisson-log-normal (PLN) distribution for the variable $Y_{gi}^{(j)}$. The PLN distribution has two parameters, η and σ^2, corresponding to the mean and variance, respectively, of the logarithmic Poisson rate. Mathematically, $Y \sim \mathrm{PLN}(\eta, \sigma^2)$ if and only if $Y \sim \mathrm{Poi}(e^\theta)$, $\theta \sim \mathrm{N}(\eta, \sigma^2)$. The equivalence implies that the PLN accounts for the over-dispersion (Supplementary Material, Section S2).

Moreover, a technical factor that can bias the analysis of sequencing data is the *library size*, which differs from one cell to another, and is defined as the total number of mapped reads to that cell (a detailed description of the library size is given in the Supplementary Material, Section S3 and Figure S1). To consider the effect of cells' library sizes, we model $Y_{gi}^{(j)}$ using $Y_{gi}^{(j)} \sim \text{Poi}(s_i^{(j)} e^{\theta_{gi}^{(j)}})$ and $\theta_{gi}^{(j)} \sim \text{N}(\eta_{gi}^{(j)}, \sigma_g^2)$, written as $Y_{gi}^{(j)} \sim \text{PLN}(s_i^{(j)}, \eta_{gi}^{(j)}, \sigma_g^2)$ for simplicity, where $s_i^{(j)}$ is a scaling factor that considers different library sizes of cells. Specifically, if we denote the library size of cell i in subject j by $l_i^{(j)}$, $s_i^{(j)}$ is calculated as $l_i^{(j)}/median_i\, l_i^{(j)}$ and $l_i^{(j)} = \sum_{g=1}^{D} X_{gi}^{(j)}$, based on the definition of the library size. Here, $\eta_{gi}^{(j)}$ represents the effects on gene g caused by cell i and subject j, and σ_g^2 reflects the variation. We separate cell effects from subject effects, and let $\eta_{gi}^{(j)}$ be the addition of the cell-specific effect $\mu_{gi}^{(j)}$ and the subject-specific effect $\beta_g^{(j)}$.

Combining the dropout mechanism and the PLN distribution for $Y_{gi}^{(j)}$ gives the zero-inflated PLN (ZIPLN) distribution for the observed data $X_{gi}^{(j)}$, which can be expressed as $X_{gi}^{(j)} \sim \text{ZIPLN}(\lambda_{g0}, \lambda_{g1}, s_i^{(j)}, \mu_{gi}^{(j)} + \beta_g^{(j)}, \sigma_g^2)$. Finally, we assign the nonparametric prior (3.1) to the cell-specific effect vector $\boldsymbol{\mu}_i^{(j)} = (\mu_{1i}^{(j)}, \ldots, \mu_{Di}^{(j)})^\top$, and arrive at the following SCSC model,

$$\begin{aligned}
G_0 &\sim \text{DP}(\alpha, H),\\
G^{(j)} &\overset{i.i.d.}{\sim} \text{DP}(\nu, \text{DP}(\gamma, G_0)), \quad j = 1, \ldots, m,\\
\boldsymbol{\mu}_i^{(j)} &\overset{i.i.d.}{\sim} G^{(j)}, \quad i = 1, \ldots, n_j \quad \text{for each } j,\\
X_{gi}^{(j)} &\sim \text{ZIPLN}(\lambda_{g0}, \lambda_{g1}, s_i^{(j)}, \mu_{gi}^{(j)} + \beta_g^{(j)}, \sigma_g^2) \quad \text{for each } j,\ i, \text{ and } g. \qquad (3.2)
\end{aligned}$$

Here, the base measure H is a non-atomic probability measure on the measurable space $(\mathbb{R}^D, \mathcal{B})$, where $\mathbb{R}^D$ is a real coordinate space of dimension D, and $\mathcal{B}$ is the Borel σ-field of $\mathbb{R}^D$. We constrain the subject-specific effects $\beta_g^{(j_1)} = \beta_g^{(j_2)}$, for any g, if $G^{(j_1)} = G^{(j_2)}$, because subjects from the same subgroup usually exhibit the same characteristics. Moreover, to make the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ estimable, we let one subject subgroup act as the "reference" group, and constrain the subject effects $\boldsymbol{\beta}^{(j)}$ of the reference group to be zero.

4. The Truncated SCSC Model

Exact posterior sampling for the SCSC model can be performed using the Polya-urn scheme (Pitman (1996)), which marginalizes the distributions G_0 and $G^{(j)}$ ($j \geq 1$). However, the marginalization procedure introduces extra dependence among the cells, and causes the cell-type allocation update for one cell to rely on all other cells. Such a sequential update scheme results in unnecessary and heavy computations. Therefore, to enhance the posterior sampling efficiency of the SCSC model, we use the blocked Gibbs sampler (Ishwaran and James

(2001)), where the updates in each parameter block are independent, by taking a truncation strategy (Ishwaran and James (2001); Rodriguez, Dunson and Gelfand (2008)) in which we set the upper bounds L for the number of subject subgroups and K for the cell type number. Moreover, the blocked Gibbs sampler favors the use of parallel computing to further speed up posterior inference. The truncated SCSC model is

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\alpha, H), \\ G^{(j)} &\overset{i.i.d.}{\sim} \mathrm{DP}_L(\nu, \mathrm{DP}_K(\gamma, G_0)), \quad j = 1, \ldots, m, \\ \boldsymbol{\mu}_i^{(j)} &\overset{i.i.d.}{\sim} G^{(j)}, \quad i = 1, \ldots, n_j \quad \text{for each } j, \\ X_{gi}^{(j)} &\sim \mathrm{ZIPLN}(\lambda_{g0}, \lambda_{g1}, s_i^{(j)}, \mu_{gi}^{(j)} + \beta_g^{(j)}, \sigma_g^2), \quad \text{for each } j,\ i,\ \text{and } g. \end{aligned} \tag{4.1}$$

Using the stick-breaking process metaphor, $\mathrm{DP}_K(\gamma, G_0)$ indicates that we break the unit stick into K pieces, rather than infinite pieces. The following theorem states that the truncation model (4.1) is an accurate approximation to the original model (3.2), as long as the concentration parameters γ and ν and the truncation numbers L and K are selected appropriately. The choice of (ν, γ, K, L) is discussed later. See the Supplementary Material, Section S4, for the proof, which is based on Theorem B1 in the NDP paper (Rodriguez, Dunson and Gelfand (2008)).

Theorem 1. *Denote the prior distributions of the cell effects $\boldsymbol{\mu}$ from the SCSC model and the truncated SCSC model by $p^{\infty\infty}(\boldsymbol{\mu})$ and $p^{KL}(\boldsymbol{\mu})$, respectively. Based on the priors, we have the marginal distributions $p^{\infty\infty}(\mathbf{x})$ and $p^{KL}(\mathbf{x})$, respectively, for the observed data $\mathbf{x}$ by integrating out all the parameters. We then have*

$$\frac{1}{4}\int \left|p^{KL}(\mathbf{x}) - p^{\infty\infty}(\mathbf{x})\right| d\mathbf{x} \leq 1 - \left\{1 - \left(\frac{\nu}{\nu+1}\right)^{L-1}\right\}^m \left\{1 - \left(\frac{\gamma}{\gamma+1}\right)^{K-1}\right\}^{\sum_{j=1}^m n_j}.$$

If we expand the implicit distributions $G^{(j)}$ in model (4.1) in terms of the subject cluster indicators $S^{(j)}$ and the cell type indicators $C_i^{(j)}$, then we obtain a more concrete and interpretable model:

$$\begin{aligned} \boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_K) &\sim \mathrm{GEM}_K(\alpha), \\ \boldsymbol{\mu}_k &\overset{i.i.d.}{\sim} H, \quad \text{for } k = 1, \ldots, K, \\ \boldsymbol{\pi}_\ell = (\pi_{1\ell}, \ldots, \pi_{K\ell}) &\overset{i.i.d.}{\sim} \mathrm{Dir}(\gamma\xi_1, \gamma\xi_2, \ldots, \gamma\xi_K), \quad \text{for } \ell = 1, \ldots, L, \\ \boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_L) &\sim \mathrm{GEM}_L(\nu), \end{aligned}$$

$$
\begin{aligned}
S^{(j)} &\overset{i.i.d.}{\sim} \mathrm{MN}(1;\phi_1,\phi_2,\ldots,\phi_L), \quad \text{for } j=1,\ldots,m,\\
C_i^{(j)}|S^{(j)}=\ell &\overset{i.i.d.}{\sim} \mathrm{MN}(1;\pi_{1\ell},\ldots,\pi_{K\ell}), \quad \text{for } i=1,\ldots,n_j \quad \text{for each } j,\\
X_{gi}^{(j)}|S^{(j)}=\ell, C_i^{(j)}=k &\sim \mathrm{ZIPLN}(\lambda_{g0},\lambda_{g1},s_i^{(j)},\mu_{gk}+\beta_{g\ell},\sigma_g^2),\\
&\quad\ \text{for each } j,\ i,\ \text{and } g.
\end{aligned} \tag{4.2}
$$

MN is the multinomial distribution and Dir indicates the Dirichlet distribution. $\mathrm{GEM}_L(\nu)$ refers to the truncated stick-breaking process in which the stick proportions $\{\phi'_1,\phi'_2,\ldots,\phi'_{L-1}\}$ are i.i.d. from $\mathrm{Beta}(1,\nu)$, and $\phi_1=\phi'_1$, $\phi_\ell=\phi'_\ell\prod_{t=1}^{\ell-1}(1-\phi'_t)$, for $2\le\ell\le L-1$, and $\phi_L=1-\sum_{\ell=1}^{L-1}\phi_\ell$. This is similar for $\mathrm{GEM}_K(\alpha)$. Again, we note that the subgroup-one effect vector $\boldsymbol{\beta}_1$ is fixed at zero for identifiability. We prove that model (4.2) is equivalent to model (4.1) in Supplementary Material, Section S5. Subsequently, we focus on model (4.2) to perform the Bayesian inference.

Note that in the stick-breaking process, the prior expectation of the first stick's length is always larger than others and, in practice, we usually assign the first subgroup as the reference group. Thus, we need to be cautious about the choice of ν that reflects our prior belief for the relative weight of the reference group ($1/(1+\nu)$ in expectation). If we replace the truncated stick-breaking prior in model (4.2) $\boldsymbol{\phi}=(\phi_1,\phi_2,\ldots,\phi_L)\sim\mathrm{GEM}_L(\nu)$ with a finite-dimensional Dirichlet prior (Ishwaran and James (2001)) $\boldsymbol{\phi}=(\phi_1,\phi_2,\ldots,\phi_L)\sim\mathrm{Dir}(\nu/L,\nu/L,\ldots,\nu/L)$, this would mitigate the effect of the prior weight bias induced by the truncated stick-breaking process. However, this replacement breaks the equivalence between models (4.1) and (4.2).

In model (4.2), a larger ν encourages more subject subgroups, and a larger γ reflects that the cell proportions across subject subgroups have more concentration on the normalized $(\xi_1,\xi_2,\ldots,\xi_K)$, the assignments of which are determined by α. Thus, we first choose γ and ν to reflect our prior belief, and then choose K and L appropriately to guarantee a small approximation error. Throughout the paper, we use $\nu=\gamma=0.5$ and $K=L=15$, giving a small approximation error in the simulation and the real application.

Because we cluster high-dimensional expression data, it is important to conduct feature selection. Tadesse, Sha and Vannucci (2005) proposed a Bayesian variable selection method to cluster high-dimensional samples and identify discriminating variables simultaneously. Therefore, we incorporate this idea into the proposed SCSC model, resulting in a variable selection version, which we call SCSC-vs. Further details can be found in the Supplementary Material, Section S6.

5. Bayesian Posterior Inference

We next specify the priors for the unknown parameters in model (4.2). The prior for the concentration parameter α ($\alpha > 0$) is a gamma distribution, $\alpha \sim \Gamma(a_{\alpha_1}, a_{\alpha_2})$. The baseline distribution H of cell-type-k effects μ_{gk} is set as the Cartesian product of D normal distributions $\mathrm{N}(\eta_\mu, \tau_\mu^2)$, and we assign hyper-priors $\eta_\mu \sim \mathrm{N}(u_\mu, \omega_\mu^2)$ and $\tau_\mu^2 \sim \mathrm{Inv}\Gamma(b_{\mu 1}, b_{\mu 2})$ to η_μ and τ_μ^2, respectively. Similarly, we assign a normal distribution $\mathrm{N}(\eta_\beta, \tau_\beta^2)$ to the subgroup effect $\beta_{g\ell}$, and assign η_β and τ_β^2 hyper-priors $\eta_\beta \sim \mathrm{N}(u_\beta, \omega_\beta^2)$ and $\tau_\beta^2 \sim \mathrm{Inv}\Gamma(b_{\beta 1}, b_{\beta 2})$ to introduce the hierarchy for subject effects. This enables information to be borrowed across genes. The prior distribution of the variance σ_g^2 is an inverse-gamma distribution $\sigma_g^2 \sim \mathrm{Inv}\Gamma(b_{\sigma 1}, b_{\sigma 2})$, and the priors for the zero-inflation-related parameters λ_{g0} and λ_{g1} are given by the weakly informative priors $\mathrm{N}(\eta_{\lambda_{g0}}, \tau_{\lambda_{g0}}^2)$ and $\mathrm{N}(\eta_{\lambda_{g1}}, \tau_{\lambda_{g1}}^2)\mathbb{I}(\lambda_{g1} < 0)$, respectively.

Finally, given the priors and model (4.2), we use the blocked Gibbs sampler (Ishwaran and James (2001)) to perform the posterior sampling. Sampling directly from a ZIPLN distribution suffers from an intractable infinite sum and integral. Therefore, we augment the model with the auxiliary variables $\theta_{gi}^{(j)}$ and $Y_{gi}^{(j)}$(Tanner and Wong (1987)) specified in Section 3 to make the sampling for the ZIPLN feasible. The Gibbs sampling scheme is presented in detail in the Supplementary Material, Section S7. Some steps of the blocked Gibbs sampler do not correspond to tractable distributions; hence, we adopt a Metropolis-within-Gibbs framework in such cases. The proposal distributions and the calculations of the acceptance rates are contained in the Supplementary Material, Section S8. For each iteration of the Gibbs sampler, the computational complexity is $\mathcal{O}(DKL\sum_{j=1}^m n_j)$, which increases linearly with the gene number D, the total cell number $\sum_{j=1}^m n_j$, and the upper bounds K and L. Thus, the MCMC algorithm can scale well on a large volume of scRNA-seq data.

After the burn-in period, defined as the first half of the iterations, we collect the posterior samples from the last half of the iterations for statistical inference. Furthermore, we estimate the subgroup and cell-type indicators, $S^{(j)}$ and $C_i^{(j)}$, respectively, using the mode of the posterior samples to keep the integer nature. For the subgroup effects and cell-type-specific effects, $\beta_{g\ell}$ and μ_{gk}, respectively, the posterior mean is used for estimation.

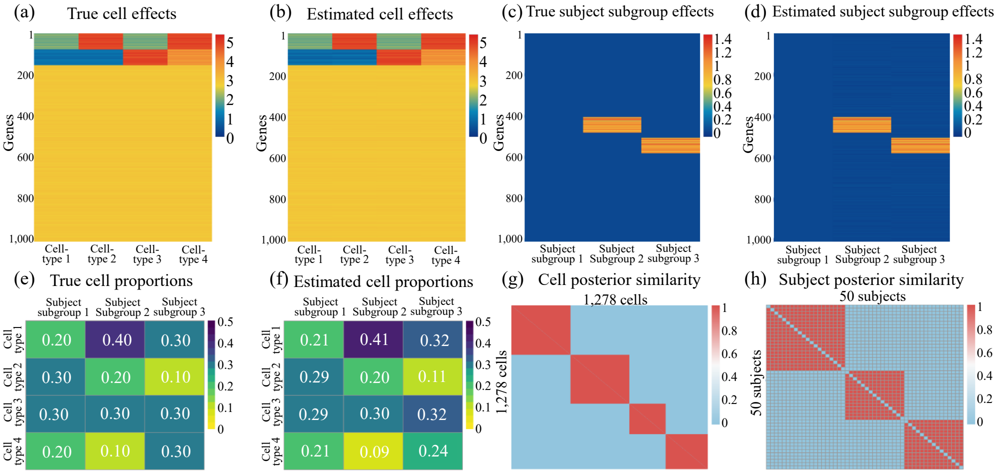


Figure 3. Performance of the SCSC model in the simulation. (a) Heatmap of the true cell effects μ_{gk}, and (b) heatmap of cell effect estimations. In both (a) and (b), one row represents one gene, and each column represents one cell type. (c) Heatmap of the true subject subgroup effects, and (d) heatmap of subject subgroup effect estimations. In both (c) and (d), one row represents one gene, and each column represents one subject subgroup. (e) Heatmap of the true cell proportions for each subgroup. (f) The cell proportion estimates. (g–h) Posterior similarity matrix heatmaps for (g) cells and (h) subjects. In the similarity matrix, the (i, j) element is the posterior probability that objects i and j are in the same cluster, for $i \neq j$.

6. Results

6.1. Simulation

We generated data following model (4.2); for a detailed description, see the Supplementary Material, Section S9. We then applied our SCSC model to this data set using $\gamma = \nu = 0.5$, the subject subgroup upper bound $L = 15$, and cell type number upper bound $K = 15$, which guarantees a small approximation error 0.0011, based on Theorem 1. We performed 10,000 iterations. By correcting the label switching (Supplementary Material, Section S10), we evaluated the estimates of the SCSC model for the cell type effects $\boldsymbol{\mu}$, subgroup effects $\boldsymbol{\beta}$, and cellular proportions for each subject subgroup $\boldsymbol{\pi}$. The comparison between the true parameter values and the estimates is shown in Figure 3(a–f), indicating that the SCSC model estimated these parameters well. Figure 3(g–h) display the posterior similarity matrices for cell clustering and subject clustering, respectively, showing clear clustering structures for cells and subjects. Hence, the SCSC model automatically and accurately identifies the underlying heterogeneity for subjects and cells.

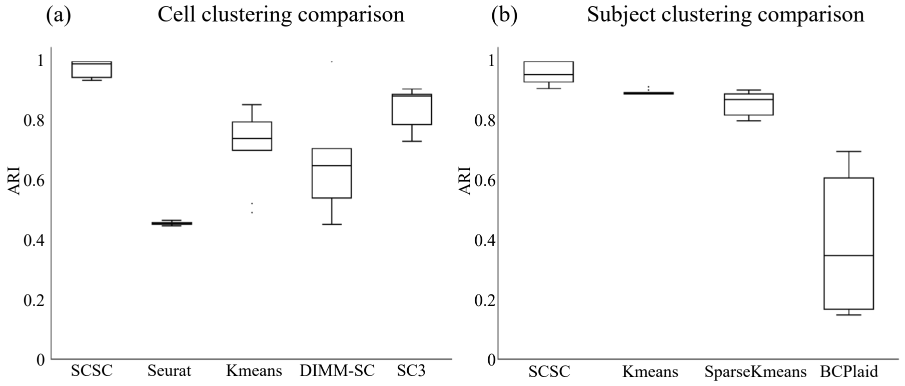


Figure 4. Clustering performance of the SCSC model and competing methods in the cell clustering and subject clustering settings based on 10 realizations. (a) ARI box plots for the SCSC model and other cell clustering approaches. (b) ARI box plots for the SCSC model and other subject clustering approaches. The implementation details of the competing methods are provided in the Supplementary Material, Section S11.

Because there is no statistical approach to simultaneously cluster subjects and cells, we compared the SCSC against several popular cell and subject clustering approaches, respectively. We selected the cell clustering approaches k-means (MacQueen (1967)), SC3 (Kiselev et al. (2017)), DIMM-SC (Sun et al. (2017)), and Seurat (Butler et al. (2018); Stuart et al. (2019)), and the subject clustering approaches kmeans (MacQueen (1967)), SparseKmeans (Witten and Tibshirani (2010)), and BCPlaid (Turner, Bailey and Krzanowski (2005)). Box plots for the adjusted Rand index (ARI) values (Hubert and Arabie (1985)) of all methods under the cell and subject clustering settings based on 10 realizations are shown in Figure 4. Overall, the SCSC model performed better in terms of both cell clustering and subject clustering. When clustering cells, the SCSC model borrows information across multiple subjects, and considers the subject differences. When grouping subjects, the model exploits the cell information of each subject to discover the subtle difference. Owing to the two-way information-sharing strategy, the SCSC model outperforms competing methods in terms of both cell clustering and subject grouping.

The performance of the SCSC and SCSC-vs on low-signal scenarios and model misspecification cases is discussed in the Supplementary Material, Section S12.

6.2. Real application

Sarkar et al. (2019) collected scRNA-seq data sets from 7,585 induced pluripotent stem cells (iPSCs) from a total of 54 Yoruba subjects in Nigeria. The data sets are publicly available with the accession code GSE118723 in GEO (Edgar, Domrachev and Lash (2002)). Although the purpose of the study (Sarkar et al. (2019)) was to detect variance QTLs, we can use the same data set to mine other interesting information, such as the cell and subject heterogeneity presented here. At the subject level, Yoruba is one of Nigeria's largest ethnic groups, and the Yorubas in the same lineage are more likely to suffer from the same genetic diseases (Olaitan et al. (2014)). Therefore, analyzing the heterogeneity of the Yorubas can clarify their family relationships or find Yoruba sub-races. At the cell level, the iPSCs are reprogrammed from the somatic cells in adult tissues, and have the ability to differentiate into several cell types. Hence, they can potentially be used to make personalized treatments for patients. The iPSCs derived from different somatic cell types may demonstrate heterogeneous differentiation abilities (Kim et al. (2011)). Our aim is to apply the SCSC model to the data set to distinguish Yoruba individuals and, at the same time, separate the iPSC heterogeneity.

Our analysis focused on the scRNA-seq counts from batch 6 in Sarkar et al. (2019), which includes 20 subjects and 1,152 cells. In the preprocessing procedure, we filtered out cells with a zero proportion of more than 80%, and genes with a zero proportion of more than 30%. We further removed subjects having less than five cells, resulting in a scRNA-seq data set with 14 subjects, 1,028 cells, and 4,178 genes. The cell numbers of the 14 selected subjects ranged from 29 to 129. During the analysis, the scaling factors were computed to adjust for the effects of the library sizes.

We then implemented the SCSC model with $(\gamma, \nu, K, L) = (0.5, 0.5, 15, 15)$, resulting in a small approximation error of 0.0009. The blocked Gibbs sampler performed 10,000 iterations, with the first half as the burn-in period, taking about 21.66 hours using 24 CPU cores. The trace plots in the Supplementary Material, Figure S2 show that the chains attained convergence during burn-in. Two Yoruba subgroups and two iPSC types were identified. The posterior similarity matrix heatmaps for cells and subjects are given in the Supplementary Material, Figure S3. Yoruba subgroup 1 contained 4 subjects and had cellular compositions of 23.68% and 76.32% for cell types 1 and 2, respectively. Yoruba subgroup 2 contained 10 subjects, with cell type compositions of 21.55% and 78.45%. The heatmaps for the logarithm-transformed and row-scaled expression

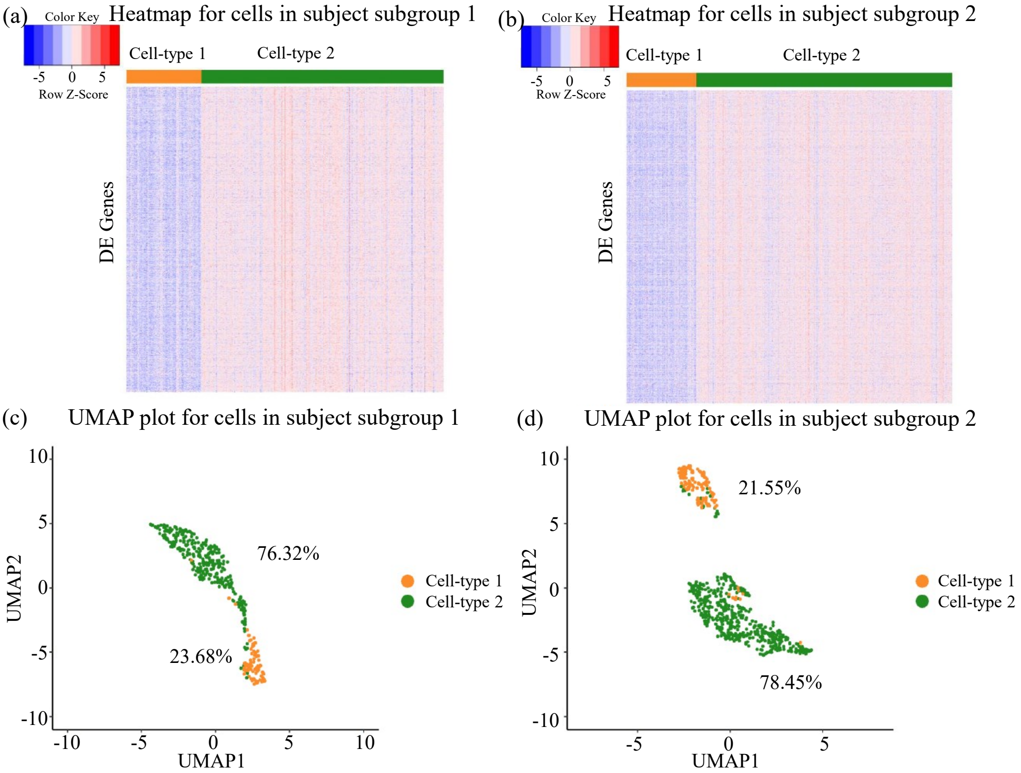


Figure 5. Performance of the SCSC model on the Yoruba iPSC scRNA-seq data. (a) Heatmap for the logarithm-transformed and row-scaled gene expression values of the cells in subject subgroup 1. There are 2,698 DE genes, 94 type-1 cells, and 303 type-2 cells. Cells under the same color are from the same cell type. (b) Heatmap for the logarithm-transformed and row-scaled gene expression values of the cells in subject subgroup 2. There are 2,698 DE genes, 136 type-1 cells, and 495 type-2 cells. (c–d) Scatter plots by projecting cells in subject subgroups 1 and 2 onto a two-dimensional space using UMAP via the R package umap (Konopka (2019)). Cells are colored by the estimated cell types: cell type 1 (orange), cell type 2 (green).

values in Yoruba subgroups 1 and 2, are shown in Figures 5(a) and 5(b), respectively. We observed clearly differential expression patterns between cell types 1 and 2 on the detected cell type DE genes, indicating the existence of heterogeneity among the iPSCs. In addition to, the cellular compositions, the estimated effects of the Yoruba subgroups also demonstrated the heterogeneity of the Yoruba individuals (Supplementary Material, Figure S4). A clear cell pattern in Yoruba subgroups 1 and 2 is observed in Figure 5(c–d): cells of type 1 (orange) and type 2 (green) are well separated. Sensitivity analyses (Supplementary Material, Section S13, and Supplementary Material, Figures S5–7) demonstrate that the clustering result obtained by the SCSC model is robust to the choices of the hyperparameters. The validation of the SCSC model clustering results are provided in the Supplementary Material, Section S14.

7. Conclusion

In this study, we developed a nonparametric Bayesian model, SCSC, to simultaneously discover subject and cell heterogeneity in a two-level clustering approach. The SCSC model has the flexibility of learning the subject subgroup or cell type number from the data without a prespecification. Unlike priors such as the HDP or the NDP, we employed the hybrid NDP-HDP prior (James (2008)) to induce group structures in subjects, cluster cells in each subject, and match cell types across subjects. The ZIPLN distribution developed in the SCSC model directly models the count nature, over-dispersion, and dropouts of the scRNA-seq data. Owing to these two features, the SCSC model achieves subject-level and cell-level clustering on the multi-subject scRNA-seq data. When clustering subjects, the SCSC model takes advantage of the cell resolution differences; when clustering cells, it borrows information across multiple subjects. The two-way information-sharing strategy enables SCSC to obtain more accurate clustering results than competing methods do in the domain of either subject clustering using bulk expression data or cell clustering based on scRNA-seq data.

To the best of our knowledge, the SCSC model is the first unified approach to address the two-level clustering for scRNA-seq data. Notably, the SCSC model bridges the methodology gap between subject clustering based on aggregated gene expression data and scRNA-seq cell clustering. The framework in the SCSC model can be further adapted to situations where the observed data are sparse and count-valued, and two-level clustering is of interest. The following are possible directions to extend the SCSC model. All distributions induced by the hybrid NDP-HDP prior have the same atoms. However, one subject subgroup may have its own cell type. For example, one tumor subtype can have a unique tumor cell subclone. Thus, incorporating the semi-HDP (Beraha, Guglielmi and Quintana (2021)) can help generate distributions in which there exist both shared and unique atoms. Additionally, the DP is a special case of the Pitman–Yor process (Pitman and Yor (1997)), which has many desirable features in practice. Thus replacing the HDP with hierarichical Pitman–Yor processes would create more realistic clustering behavior, especially in the scRNA-seq data analysis (Camerlenghi et al. (2020)).

Considering the continuous progress of sequencing technology, single-cell

RNA sequencing will become affordable and available to more persons. We thus envision that the SCSC model will be a useful method to facilitate the development of personalized treatment in a time of single-cell genomics.

Supplementary Material

Supplementary Material provides the proofs, MCMC derivations, and simulation and application results. The R package to implement the SCSC model is available on GitHub `https://github.com/WgitU/SCSC`.

Acknowledgments

We are grateful to the editor, associate editor, and referees for their invaluable suggestions which have substantially improved the quality of the paper. This work was supported by the National Natural Science Foundation of China (11901572). We thank the High-performance Computing Platform of Renmin University of China for providing computing resources.

References

Beraha, M., Guglielmi, A. and Quintana, F. A. (2021). The semi-hierarchical Dirichlet Process and its application to clustering homogeneous distributions. *Bayesian Analysis* **16**, 1187–1219.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J. et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**, 155–160.

Busch, S. E., Hanke, M. L., Kargl, J., Metz, H. E., Macpherson, D. and Houghton, A. M. (2016). Lung cancer subtypes generate unique immune responses. *Journal of Immunology* **197**, 4493–4503.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420.

Camerlenghi, F., Dumitrascu, B., Ferrari, F., Engelhardt, B. E. and Favaro, S. (2020). Nonparametric Bayesian multi-armed bandits for single cell experiment design. *The Annals of Applied Statistics* **14**, 2003–2019.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I. and Rodríguez, A. (2019). Latent nested nonparametric priors (with discussion). *Bayesian Analysis* **14**, 1303–1356.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Ismb* **8**, 93–103.

Denti, F., Camerlenghi, F., Guindani, M. and Mira, A. (2021). A common atom model for the Bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association*, 1–12.

Edgar, R., Domrachev, M. and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2**, 193–218.

Huo, Z., Ding, Y., Liu, S., Oesterreich, S. and Tseng, G. (2016). Meta-analytic framework for sparse K-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association* **111**, 27–42.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

James, L. (2008). Discussion of nested Dirichlet process paper by Rodriguez, Dunson and Gelfand. *Journal of the American Statistical Association* **483**, 1131–1154.

Kim, K., Zhao, R., Doi, A., Ng, K., Unternaehrer, J., Cahan, P. et al. (2011). Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. *Nature Biotechnology* **29**, 1117–1119.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T. et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* **14**, 483–486.

Konopka, T. (2019). *UMAP: uniform manifold approximation and projection.* R Package (Version 0.2.1.0).

Liu, Y., Warren, J. L. and Zhao, H. (2019). A hierarchical Bayesian model for single-cell clustering using RNA-sequencing data. *The Annals of Applied Statistics* **13**, 1733–1752.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics* **12**, 351–357.

Luo, X. and Wei, Y. (2019). Batch effects correction with unknown subtypes. *Journal of the American Statistical Association* **114**, 581–594.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Edited by L. M. L. Cam and J. Neyman), 281–297. University of California Press, Oakland.

Makki, J. (2015). Diversity of breast carcinoma: Histological subtypes and clinical relevance. *Clinical Medicine Insights Pathology* **8**, 23–31.

Olaitan, P. B., Odesina, V., Ademola, S. A., Fadiora, S. O., Oluwatosin, O. M. and Reichenberger, E. J. (2014). Recruitment of Yoruba families from Nigeria for genetic research: Experience from a multisite keloid study. *BMC Medical Ethics* **15**, 65.

Paisley, J., Wang, C., Blei, D. M. and Jordan, M. I. (2015). Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 256–270.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8**, 1145–1164.

Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. *Lecture Notes-Monograph Series* **30**, 245–267.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900.

Prabhakaran, S., Azizi, E., Carr, A. and Pe'er, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning* (Edited by M. F. Balcan and K. Q. Weinberger) **48**, 1070-1079.

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association* **103**, 1131–1154.

Sarkar, A. K., Tung, P.-Y., Blischak, J. D., Burnett, J. E., Li, Y. I., Stephens, M. et al. (2019). Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genetics* **15**, e1008045.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistics Sinica* **4**, 639–650.

Song, F., Chan, G. M. A. and Wei, Y. (2020). Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction. *Nature Communications* **11**, 3274.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. et al. (2019). Comprehensive integration of single cell data. *Cell* **177**, 1888–1902.

Sun, Z., Wang, T., Deng, K., XF, W., Lafyatis, R., Ding, Y. et al. (2017). DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* **34**, 139–146.

Tadesse, M. G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100**, 602–617.

Tanner, M. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

Tekumalla, L. S., Agrawal, P. and Bhattacharya, I. (2015). Nested hierarchical Dirichlet processes for multi-level non-parametric admixture modeling. *arXiv preprint arXiv: 1508.06446*.

Turner, H., Bailey, T. C. and Krzanowski, W. J. (2005). Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics & Data Analysis* **48**, 235–254.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64**, 440–448.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**, 713–726.

žurauskienė, J. and Yau, C. (2016). pcaReduce: Hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* **17**, 140.

Qiuyu Wu

Institute of Statistics and Big Data, Renmin University of China, Haidian, Beijing 100872, China.

E-mail: w.qy@ruc.edu.cn

Xiangyu Luo

Institute of Statistics and Big Data, Renmin University of China, Haidian, Beijing 100872, China.

E-mail: xiangyuluo@ruc.edu.cn

(Received August 2020; accepted February 2021)