Statistics in Biosciences
https://doi.org/10.1007/s12561-022-09350-w

A Unified Bayesian Framework for Bi-overlapping-Clustering Multi-omics Data via Sparse Matrix Factorization

Fangting Zhou[1,2] · Kejun He[1] · James J. Cai[3] · Laurie A. Davidson[4,5] · Robert S. Chapkin[4,5] · Yang Ni[2]

Received: 10 June 2020 / Revised: 10 April 2021 / Accepted: 6 June 2022

Abstract
The advances of modern sequencing techniques have generated an unprecedented amount of multi-omics data which provide great opportunities to quantitatively explore functional genomes from different but complementary perspectives. However, distinct modalities/sequencing technologies generate diverse types of data which greatly complicate statistical modeling because uniquely optimized methods are required for handling each type of data. In this paper, we propose a unified framework for Bayesian nonparametric matrix factorization that infers overlapping bi-clusters for multi-omics data. The proposed method adaptively discretizes different types of observations into common latent states on which cluster structures are built hierarchically. The proposed Bayesian nonparametric method is able to automatically determine the number of clusters. We demonstrate the utility of the proposed method using simulation studies and applications to a single-cell RNA-sequencing dataset, a combination of single-cell RNA-sequencing and single-cell ATAC-sequencing dataset, a bulk RNA-sequencing dataset, and a DNA methylation dataset which reveal several interesting findings that are consistent with biological literature.

Keywords Bayesian nonparametric prior · Data integration · Indian buffet process · Mixture model · Single-cell sequencing

✉ Kejun He
kejunhe@ruc.edu.cn

✉ Yang Ni
yni@stat.tamu.edu

Extended author information available on the last page of the article

Published online: 08 July 2022

1 Introduction

This paper proposes a unified Bayesian framework for model-based bi-overlapping-clustering of multi-omics data. The advances of modern sequencing techniques have generated an unprecedented amount of multi-omics data (e.g., DNA methylation, chromatin accessibility, and mRNA gene expression). Unlike genetic/genomic data from single modality which can only provide a partial view, multi-omics data have enabled researchers to interrogate the complex biological systems from different perspectives and characterize more comprehensive knowledge of cellular functions and activities at the molecular level. These new scientific advancements have greatly impacted public health via improved strategies of diagnosis, treatment, and prevention of genetic diseases such as cancer. However, despite the continuous efforts by the multi-omics research community, new statistical methods are still in great need to keep up the pace with fast evolving sequencing technologies. For example, though traditional bulk RNA-sequencing (RNA-seq) and relatively new single-cell RNA-seq (scRNA-seq) techniques aim to measure the same molecules, the clustering algorithms for analyzing bulk RNA-seq data is not suitable for scRNA-seq data due to its unique features (e.g., sparsity) and thus need to be redesigned.

1.1 Challenges in Model-Based Clustering of Multi-Omics Data

1.1.1 Mixed Data Types

Different modalities/sequencing technologies generate various types of data. For example, while DNA copy number variation can be coded as ordinal variables representing status loss, neutral, or gain, DNA methylation level is often modeled as beta values in the range between 0 and 1. While bulk RNA-seq data are often treated as Gaussian distributed after log-normalization, scRNA-seq data are generally regarded as zero-inflated counts. The wide range of sampling distributions of multi-omics data makes model-based clustering tasks challenging since each data type requires a special treatment.

1.1.2 Noisiness

Multi-omics data contain high levels of noise due to technical limitations, which inevitably confound with biological variations that researchers strive to investigate. Ignoring such intrinsic and experimental noise promotes susceptibility to false conclusions which will be propagated to downstream analysis, thereby hindering scientific discoveries. For instance, [26] showed a large fraction of stochastic allele-specific expression from scRNA-seq data can be explained by technical noise, especially for lowly and moderately expressed genes. Without properly accounting for technical noise, statistical modeling can easily skew biological interpretations. Many attempts have been made to address high levels of noise. For example, [49] proposed a mixture model that can eliminate noise from microarray data. [34]

introduced surrogate variable analysis to model and remove noise from sequencing data. [14] suggested to remove technical noise based on spike-in ERCC molecules in scRNA-seq data.

1.1.3 Heterogeneity

Biological samples are heterogeneous due to, e.g., genetic differences, environmental influences, or disease severity [21, 40]. Individualized characterization is necessary to account for heterogeneity and avoid spurious results derived from the homogeneous modeling assumption. For instance, tumor heterogeneity is a well-known characteristic of cancer, and therefore sequencing data collected from tumor tissues are inherently heterogeneous. An increasing number of studies have been carried out to decipher the tumor heterogeneity from (bulk) sequencing data (e.g., the feature allocation model [33] with DNA sequencing data). Sampling heterogeneity is also commonly observed in single-cell data where tissue samples often contain multiple tissue types. Various clustering methods have been proposed to identify cell types; see e.g., a recent review [28].

1.1.4 Data Integration

With the availability of multi-omics data, a joint analysis that integrates all sources information is often desired. Although single-modality clustering methods can be independently applied to one modality at a time, post hoc processing is necessary in order to achieve a consensus conclusion from potentially incompatible clustering results from each modality. Proper propagation of the uncertainty from estimation to post hoc processing is especially important for noisy biological data but remains highly non-trivial. In addition, data integration is exposed to an even more pressing challenge for single-cell multi-omics data because typically each cell can be only assayed on a single modality. The lack of matching samples across modalities renders most of existing multi-omics data integration methods inapplicable with the exception of [36] who developed a model-based approach for the integrative analysis of single-cell chromatin accessibility and gene expression. There is, therefore, a critical need to develop new integrative clustering methods to jointly cluster multi-omics data (especially single-cell multi-omics data).

1.2 Review of Existing Literature

1.2.1 Clustering

Clustering is an unsupervised learning task that seeks to divide units into mutually exclusive groups. Extensive work has been done in developing clustering methods such as algorithm-based hierarchical clustering [23] and k-means clustering [20], and model-based finite mixture [41] and infinite mixture models [43].

1.2.2 Bi-clustering

Clustering methods are useful for clustering either observations or covariates but are not directly applicable for joint clustering both observations and covariates. Bi-clustering extends clustering by simultaneously clustering rows and columns of a data matrix. Again, numerous methods have been developed for this purpose [19, 38, 56]. A common thread of clustering and bi-clustering is that the clusters have to be mutually exclusive which becomes a limitation in clustering multi-omics data because, for example, a gene can participate in multiple pathways.

1.2.3 Overlapping Clustering

Overlapping clustering, also known as feature allocation or fuzzy clustering, relaxes the restriction to mutually exclusive clusters and allocates each unit to possibly more than one cluster. Like clustering, the vast majority of the existing overlapping clustering methods [3, 4, 11, 16] do not jointly cluster observations and covariates. Moreover, they usually work well for continuous data only.

1.2.4 Sparse Matrix Factorization

Matrix factorization decomposes a high-dimensional matrix into two low-rank matrices. Sparse matrix factorization encourages the low-rank matrices to be sparse and can be interpreted as clustering. For instance, [51] proposed a latent factor model with a sparse loading matrix for continuous data. The sparsity pattern indicates an overlapping clustering of the covariates. Many specific sparse matrix factorization algorithms have been developed for non-negative matrices [22, 30, 31], count matrices [17, 61], binary matrices [59, 60], categorical matrices [45], multinomial matrices [63], and other types of matrices. In [45, 63], both low-rank matrices are assumed to be sparse and hence can be interpreted as bi-overlapping-clustering.

1.2.5 The Proposed Method

In this paper, we extend the work of [45, 63] by proposing a sparse unified matrix factorization (UMF) framework which is, in principle, applicable to any type of omic data. The proposed UMF introduces a mixture model representation of the observations through a set of latent variables to indicate the underlying state of multi-omics observations. This simple formulation of the sampling model adaptively discretizes the observations into a binary/categorical matrix, which is more robust to high levels of noises. Moreover, while the choice of mixture kernel depends on the specific type of the omic data, the latent states are universal, which allows us to impose essentially the same latent matrix factorization priors to characterize the heterogeneity among both observations and covariates for virtually any type of omic data. Furthermore, the proposed framework can also be used to integrate multi-modal single-cell sequencing data.

Particularly, we construct a hierarchical model with a combination of latent logistic model, Indian buffet process (IBP, [16]) prior, and Dirichlet-categorical prior. Using

IBP, we are able to infer an unknown number of overlapping clusters of observations. Conditional on the clusters of observations, the Dirichlet-categorical prior clusters covariates, again allowing overlaps. Through simulation studies, we demonstrate that the proposed UMF has superior performance compared to competing methods across different types of data. We subsequently illustrate UMF with applications to a mouse scRNA-seq dataset, a breast cancer bulk RNA-seq dataset, a head and neck cancer DNA methylation dataset, and an integration of human scRNA-seq gene expression and scATAC-seq chromatin accessibility dataset, which reveal some interesting results that are consistent with existing biological literature.

The remainder of the paper is organized as follows. We introduce the proposed latent matrix factorization model in Sect. 2. Posterior inference based on Markov chain Monte Carlo (MCMC) sampling is described in Sect. 3. In Sects. 4 and 5 , we illustrate our approach with simulation studies, and analyses of four real datasets. This paper is concluded with a brief discussion in Sect. 6.

2 Model

2.1 Classifying Omic Data via Adaptive Discretization

Let x_{ij} generically denote the observed value of gene j in sample i with $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Depending on the application, x_{ij} can represent gene expression, methylation, chromatin accessibility, etc. The general idea of UMF in modeling noisy sequencing data is to introduce latent indicator variables z_{ij}'s to adaptively classify observations into latent states. In this section, we focus on four types of omic data although UMF has much wider applicability: bulk RNA-seq gene expression, DNA methylation, scRNA-seq gene expression, and scATAC-seq chromatin accessibility. In the following, we describe the unique features and corresponding sampling distribution of each data type. Moreover, the proposed UMF is also applicable to mixed data types; as an illustration, we will discuss the integration of scRNA-seq and scATAC-seq data.

2.1.1 Bulk RNA-Seq Gene Expression

Gene expression measurements provide opportunities to quantitatively characterize complex genetic diseases such as cancer at the mRNA level. Identifying disease subtypes is one of the first steps in developing personalized treatments. We will work with log-transformed, centered mRNA measurements which are often treated as continuous data with heavy tails. Let $N(\mu, \sigma^2)$ denote Gaussian distribution with mean μ and variance σ^2, and $U(a, b)$ denote uniform distribution on the interval (a, b). We adopt the probability of expression model (POE, [49]) to represent gene expressions as a mixture of one Gaussian and two uniform distributions,

$$x_{ij} \sim I(z_{ij} = -1)U\left(\mu_j - k_j^-, \mu_j\right) + I(z_{ij} = 0)N\left(\mu_j, \sigma_j^2\right) + I(z_{ij} = 1)U\left(\mu_j, \mu_j + k_j^+\right).$$

The constraint $\sigma_j < \min(k_j^-, k_j^+)/k_0$ with $k_0 > 5$ is imposed to capture the heavy tails through uniform distributions. The latent variable $z_{ij} = -1$, 0, and 1 respectively indicates the case of under, normal, and over-expression of gene j in sample i. We assume $k_j^-, k_j^+ \sim \text{Gamma}(\alpha_k, \beta_k)$, $\sigma_j^2 \sim \text{IG}(\alpha_\sigma, \beta_\sigma) I(\sigma_j < \min(k_j^-, k_j^+)/k_0)$, and $\mu_j \sim N(m_\mu, \sigma_\mu^2)$.

2.1.2 DNA Methylation

DNA methylation is an essential epigenetic factor that regulates gene transcription, and plays critical roles in gene regulation. The methylation levels are often calculated as beta values that are defined as the ratios of intensities between methylated and unmethylated alleles. Beta values are between 0 and 1 with 0 being unmethylated and 1 fully methylated. Given the range of the beta values, beta distribution is appropriate to model the methylation data. Let $\text{Beta}(u, v)$ denote a beta distribution with mean u and effective sample size v (in a more conventional parameterization of beta distribution with shape parameters a and b, $u = a/(a+b)$ and $v = a+b$). We assume that observations are generated from a mixture of two beta distributions,

$$x_{ij} \sim I\big(z_{ij} = 0\big)\text{Beta}\Big(u_j^-, v_j^-\Big) + I\big(z_{ij} = 1\big)\text{Beta}\Big(u_j^+, v_j^+\Big).$$

We restrict $u_j^- < u_j^+$ so that $z_{ij} = 1$ indicates a relatively high methylation level, whereas $z_{ij} = 0$ stands for a low level. We assume $p(u_j^-, u_j^+) \propto \text{Beta}(u_j^- | \alpha_u, \beta_u)$ $\text{Beta}(u_j^+ | \alpha_u, \beta_u) I(u_j^- < u_j^+)$ and $v_j^-, v_j^+ \sim \text{Gamma}(\alpha_v, \beta_v)$ independently.

2.1.3 scRNA-Seq Gene Expression

scRNA-seq technologies catalogue transcriptomic activities in individual cells and have facilitated new biological discoveries that were until recently impossible with bulk RNA-seq, such as revealing new gene regulatory relationships and cell types at the single-cell level. However, the excessive zeros and dispersed measurements render conventional statistical analysis unsuitable in analyzing scRNA-seq data. To incorporate these unique features of scRNA-seq data, we consider a mixture of zero-inflated Poisson and negative binomial distributions,

$$x_{ij} \sim I\big(z_{ij} = 0\big)\text{ZIP}\big(\pi_j, \lambda_j\big) + I(z_{ij} = 1)\text{NB}\big(r_j, \phi_j\big), \tag{1}$$

where $\text{NB}(r, \phi)$ denotes the negative binomial distribution with mean $r > 0$ and dispersion ϕ, and $\text{ZIP}(\pi, \lambda)$ denotes the zero-inflated Poisson distribution with zero inflation probability $0 \le \pi \le 1$ and Poisson rate parameter $\lambda > 0$. The probability mass function of $\text{ZIP}(\pi, \lambda)$ is given by,

$$\Pr(X = x) = \begin{cases} \pi + (1-\pi)\exp(-\lambda) & \text{if } x = 0, \\ (1-\pi)\exp(-\lambda)\lambda^x / x! & \text{if } x > 0. \end{cases}$$

We assume $\lambda_j < r_j$ so that the latent state $z_{ij} = 1$ represents large dispersed expressions captured by the negative binomial component and $z_{ij} = 0$ represents small

(including zero) expressions captured by the zero-inflated Poisson component. We assume $\pi_j \sim \text{Beta}(\alpha_\pi, \beta_\pi)$, $p(\lambda_j, r_j) \propto \text{Gamma}(\lambda_j|\alpha_\lambda, \beta_\lambda)\text{Gamma}(r_j|\alpha_r, \beta_r)I(\lambda_j < r_j)$, and $\phi_j \sim \text{Beta}(\alpha_\phi, \beta_\phi)$.

2.1.4 scATAC-Seq Chromatin Accessibility

The single-cell assay for transposase-accessible chromatin (scATAC-seq) maps the landscape of chromatin accessibility [8], allowing characterization of chromatin variability among individual cells. The accessible chromatin is a hallmark of active DNA regulatory elements. The close-to-binary nature of scATAC-seq leads us to consider the following mixture model representation,

$$x_{ij} \sim I(z_{ij} = 0)f_-(x_{ij}) + I(z_{ij} = 1)f_+(x_{ij}). \tag{2}$$

While other choices of $f_-(\cdot)$ and $f_+(\cdot)$ can be made, we follow [36] to set $f_-(\cdot)$ to be a point mass at zero and $f_+(\cdot)$ to be a probability mass function with zero mass at zero due to the fact that scATAC-seq data are small counts with a large number of zeros. Therefore, the latent state is deterministic: $z_{ij} = 1$ ($z_{ij} = 0$) if the promoter region of gene j is (not) accessible, i.e., $x_{ij} \neq 0$ ($x_{ij} = 0$).

2.1.5 Integration of scRNA-Seq and scATAC-Seq

scRNA-seq and scATAC-seq experiments are sometimes performed jointly to measure gene expression and chromatin accessibility for the same cell population although individual cell can only be measured by one platform. An independent analysis is deemed less efficient than a joint analysis because the former ignores the biological links between these two types of data. Because each observation/cell has only one type of data, joint clustering the combined data is challenging. Our proposed UMF is able to overcome this difficulty through dichotomizing both datasets to binary indicators and matching them on the gene levels. Specifically, for each gene j, we identify its corresponding promoter region which promotes its transcription to mRNA. In other words, if the promoter region is accessible, the corresponding gene is more likely to be transcribed/expressed. More precisely, without loss of generality, let x^r_{ij} denote the scRNA-seq gene expression of gene j in the first n_1 cells and let $x^a_{\ell j}$ denote the scATAC-seq chromatin accessibility of the promoter region of gene j in the next n_2 cells. The total number of cells is $n = n_1 + n_2$. We assume that x^r_{ij} follows the sampling model (1) with indicators z^r_{ij} and that $x^a_{\ell j}$ follows the sampling model (2) with indicators $z^a_{\ell j}$. Despite the disparate sampling distributions, the latent indicators z^r_{ij} and z^a_{ij} have coherent interpretations across the modalities: if $z^r_{ij} = 1$ or $z^a_{\ell j} = 1$, gene j is expected to have high expression in cell i or ℓ. A similar strategy was used in [36]; they focused on clustering cells whereas we consider clustering both cells and genes with possible overlaps.

2.2 Bi-overlapping-Clustering Omic Data via Unified Matrix Factorization

As shown in Sect. 2.1, multi-omics data measured from different modalities can be represented by the latent variables z_{ij}'s in a unified way. Let $\boldsymbol{Z} = [z_{ij}]_{i=1,j=1}^{n,p}$. Note that for the integration of scRNA-seq and scATAC-seq data, we vertically concatenate $\boldsymbol{Z} = [\boldsymbol{Z}^r; \boldsymbol{Z}^a]$ where $\boldsymbol{Z}^r = [z_{ij}^r]_{i=1,j=1}^{n_1,p}$ and $\boldsymbol{Z}^a = [z_{\ell j}^a]_{\ell=1,j=1}^{n_2,p}$; see Figure 1 for an illustration. We introduce lower-dimensional matrices to characterize the heterogeneity of both rows and columns of $\boldsymbol{Z}$. The latent indicator matrix $\boldsymbol{Z} \in \{0,1\}^{n\times p}$ is binary except for bulk RNA-seq data for which $\boldsymbol{Z} \in \{-1,0,1\}^{n\times p}$ is categorical. We will discuss the prior distributions for categorical and binary $\boldsymbol{Z}$ respectively.

2.2.1 Categorical

We let $\boldsymbol{A} \in \{0,1\}^{n\times K}$ and $\boldsymbol{C} \in \{-1,0,1\}^{p\times K}$ denote the sample-latent and the covariate-latent matrices, of which the clustering interpretations will be given at the end of this section. The number K of columns of $\boldsymbol{A}$ and $\boldsymbol{C}$ is usually much smaller than the dimensions of the original data (n and p). We link $\boldsymbol{A}$ and $\boldsymbol{C}$ to $\boldsymbol{Z}$ by a latent multi-class logistic model,

$$z_{ij} \sim \text{Categorical}\Big\{ M^{-1}\exp\Big(\sum_k a_{ik} w_{jk}^- I(c_{jk}=-1) + \eta_j^-\Big), M^{-1},$$
$$M^{-1}\exp\Big(\sum_k a_{ik} w_{jk}^+ I(c_{jk}=1) + \eta_j^+\Big)\Big\},$$

where M is a normalizing constant. Parameters w_{jk}^- and w_{jk}^+ tie the jth covariate to the kth cluster. Parameters η_j^- and η_j^+ represent the baseline probabilities of belonging to categories -1 and $+1$.

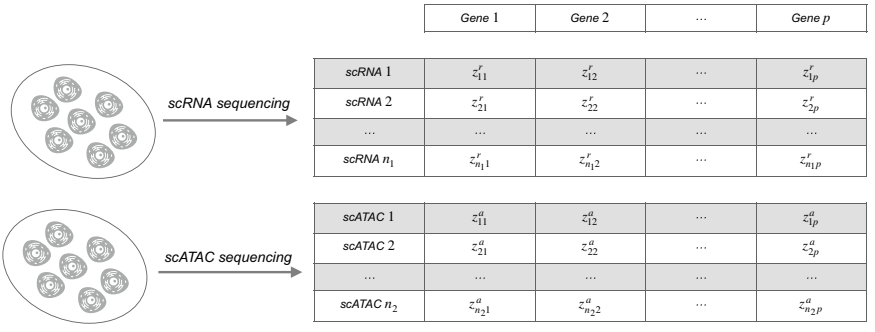


Fig. 1 Concatenation of scRNA-seq and scATAC-seq data

2.2.2 Binary

We use the same notation $\boldsymbol{A}$ for the sample-latent matrix but let $\boldsymbol{B} = (b_{jk}) \in \{0, 1\}^{p \times K}$ denote the binary covariate-latent matrix. We link $\boldsymbol{A}$ and $\boldsymbol{B}$ by a latent binary logistic model,

$$z_{ij} \sim \text{Bernoulli}\left\{ \frac{\exp\left(\sum_k a_{ik} w_{jk} b_{jk} + \eta_j\right)}{1 + \exp\left(\sum_k a_{ik} w_{jk} b_{jk} + \eta_j\right)} \right\}.$$

The interpretation of w_{jk} and η_j is similar to the categorical case.

2.2.3 Interpretations of Clustering Matrices

The sample-latent matrix $\boldsymbol{A}$ and covariate-latent matrix $\boldsymbol{B}$ or $\boldsymbol{C}$ can be interpreted as clustering of the rows and columns of $\boldsymbol{Z}$, respectively. Observation i (covariate j) belongs to cluster k if $a_{ik} \neq 0$ ($b_{jk} \neq 0$ or $c_{jk} \neq 0$). Since we do not constrain $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ to having unit row sums, clusters can have overlaps.

2.3 Indian Buffet Process and Hyperpriors

To make inference on the latent matrices, we will first impose a Bayesian nonparametric prior on $\boldsymbol{A}$ that can automatically determine the number K of clusters.

The Indian buffet process has been widely used as a Bayesian nonparametric prior on binary matrices with a potentially unbounded number of columns. To describe the matrix-generating process, we first assume a fixed number $\widetilde{K}$ of columns of $\boldsymbol{A}$ which will be relaxed later. Conditional on $\widetilde{K}$, a_{ik}'s are assumed to be independent Bernoulli random variables,

$$a_{ik}|\pi_k \sim \text{Bernoulli}(\pi_k), \quad \pi_k \sim \text{Beta}(m/\widetilde{K}, 1),$$

where m is a hyper-parameter. Marginalizing out π_k, we have

$$p(\boldsymbol{A}) = \prod_{k=1}^{\widetilde{K}} \frac{m\Gamma(r_k + m/\widetilde{K})\Gamma(n - r_k + 1)}{\widetilde{K}\Gamma(n + 1 + m\widetilde{K})}$$

where $r_k = \sum_{i=1}^{n} a_{ik}$ is the sum of the entries in the kth column of $\boldsymbol{A}$. We then take $\widetilde{K} \to \infty$, denote K as the number of non-empty columns of $\boldsymbol{A}$, and remove the columns whose entries are all zeros. The resulting matrix $\boldsymbol{A}$ follows an IBP(m) prior with probability mass function,

$$p(\boldsymbol{A}) = \frac{m^K \exp(-mH_n)}{K!} \prod_{k=1}^{K} \frac{\Gamma(r_k)\Gamma(n - r_k + 1)}{\Gamma(n + 1)},$$

where $H_n = \sum_{i=1}^{n} 1/i$ is the nth harmonic number. Moreover, the rows of $\boldsymbol{A}$ are exchangeable and the conditional probability for $a_{ik} = 1$ is $p(a_{ik} = 1|\boldsymbol{a}_{(-i)k}) = r_{(-i)k}/n$ provided $r_{(-i)k} > 0$, where $\boldsymbol{a}_{(-i)k}$ is the kth column of $\boldsymbol{A}$ excluding the ith row and

$r_{(-i)k}$ is the number of ones in $\boldsymbol{a}_{(-i)k}$. The distribution of number of new columns/clusters for each row is Poission(m/n).

Conditional on matrix $\boldsymbol{A}$ via the number of columns K, each element b_{jk} of $\boldsymbol{B}$ follows an independent beta-Bernoulli distribution $b_{jk} \sim \text{Bernoulli}(\rho)$ with $\rho \sim \text{Beta}(\alpha_\rho, \beta_\rho)$. Likewise, each element c_{jk} of $\boldsymbol{C}$ follows the Dirichlet-categorical distribution $c_{jk} \sim \text{Categorical}(\boldsymbol{\gamma})$ with $\boldsymbol{\gamma} = (\gamma_{-1}, \gamma_0, \gamma_1) \sim \text{Dirichlet}(\psi_{-1}, \psi_0, \psi_1)$. We assume $w_{jk}, w_{jk}^-, w_{jk}^+ \sim$ Gamma (a_w, b_w), $\eta_j, \eta_j^-, \eta_j^+ \sim N(\mu_\eta, \sigma_\eta^2)$, and $m \sim \text{Gamma}(\alpha_m, \beta_m)$.

3 Posterior Inference

We summarize the posterior inference by MCMC simulation for the proposed UMF with categorical covariate-latent representation; UMF with binary covariate-latent representation can be treated as a special case. The proposed UMF with categorical latent representation is parameterized by

$$\left\{\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{Z}, \{\boldsymbol{w}_j^-, \boldsymbol{w}_j^+, \eta_j^-, \eta_j^+\}_{j=1}^p, \boldsymbol{\theta}, \boldsymbol{\gamma}, m\right\},$$

where $\boldsymbol{\theta}$ generically denote the parameters of the sampling models described in Sect. 2.1, for example, $\boldsymbol{\theta} = \{\mu_j, \sigma_j^2, k_j^-, k_j^+\}_{j=1}^p$ in the bulk RNA-seq data. While most of the parameters are trivial to update with Gibbs or Metropolis-Hasting, care must be taken in updating $\boldsymbol{A}$ as its dimension can change from iteration to iteration. The details of the updating scheme of $\boldsymbol{A}$ are provided below.

We let $\boldsymbol{a}_k$ and $\boldsymbol{a}_{(-i)k}$ respectively denote the kth column of $\boldsymbol{A}$ and the kth column of $\boldsymbol{A}$ without the ith entry. Sequentially for $i = 1, \ldots, n$, we cycle through the following two steps.

Step i. Update existing non-empty columns $k = 1, \ldots, K$ of $\boldsymbol{A}$. If $\boldsymbol{a}_{(-i)k} = 0$, drop feature k. Otherwise, sample a_{ik} from the full conditional distribution

$$p(a_{ik} = 1|\cdot) \propto p\big(a_{ik} = 1|\boldsymbol{a}_{(-i)k}\big) \prod_{j=1}^{p} p\left(z_{ij} \middle| \left\{a_{ik}, c_{jk}, w_{jk}^-, w_{jk}^+\right\}_{k=1}^{K}, \eta_j^-, \eta_j^+\right).$$

If a column becomes all zeros after updating, we delete this column and reduce K to $K = K - 1$.

Step ii. After updating existing columns, we propose to add new columns. We first draw $K^* \sim \text{Poission}(m/n)$. If $K^* = 0$, we will skip this step. Otherwise, we propose a set of new parameters $\boldsymbol{c}_k^* = (c_{1k}^*, \ldots, c_{pk}^*)^\top$ and $\{w_{jk}^{-*}, w_{jk}^{+*}\}_{j=1}^p$ from their prior distributions, for $k = K + 1, \ldots, K + K^*$. We accept new features and the associated parameters with probability

$$\min\left\{1, \frac{\prod_{j=1}^{p} p\left(z_{ij} \middle| \left\{a_{ik}, c_{jk}, w_{jk}^{-}, w_{jk}^{+}\right\}_{k=1}^{K}, \left\{a_{ik}^{*}, c_{jk}^{*}, w_{jk}^{-*}, w_{jk}^{+*}\right\}_{k=K+1}^{K+K^{*}}, \eta_j^{-}, \eta_j^{+}\right)}{\prod_{j=1}^{p} p\left(z_{ij} \middle| \left\{a_{ik}, c_{jk}, w_{jk}^{-}, w_{jk}^{+}\right\}_{k=1}^{K}, \eta_j^{-}, \eta_j^{+}\right)}\right\},$$

where $a_{i,K+1} = \cdots = a_{i,K+K^*} = 1$. If new columns are accepted, we increase K to $K = K + K^*$.

In addition, due to the high correlation between latent matrix $\boldsymbol{Z}$ and uniform limits k_j^-, k_j^+ in the POE model during MCMC simulation [49], we perform a joint update of these parameters for better mixing. Details of the full MCMC algorithm are given in the Supplementary Materials.

To summarize the posterior distribution based on the Monte Carlo samples, we proceed by first calculating the maximum a posteriori estimate $\widehat{K}$ of K from the marginal posterior distribution. Conditional on estimated $\widehat{K}$, we find the point estimate of $\boldsymbol{A}$ by the following procedure. For any matrices $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \{0,1\}^{n \times \widehat{K}}$, we define a distance $d(\boldsymbol{A}, \widetilde{\boldsymbol{A}}) = \min_{\pi} D(\boldsymbol{A}, \pi(\widetilde{\boldsymbol{A}}))$, where $\pi(\widetilde{\boldsymbol{A}})$ denotes a permutation of the columns of $\widetilde{\boldsymbol{A}}$ and $D(\cdot,\cdot)$ is the Hamming distance between the two matrices. A point estimator $\widehat{\boldsymbol{A}}$ of $\boldsymbol{A}$ is then obtained as

$$\widehat{\boldsymbol{A}} = \arg\min_{\widetilde{\boldsymbol{A}}} \int d(\boldsymbol{A}, \widetilde{\boldsymbol{A}})\, \mathrm{d}p(\boldsymbol{A}|\cdot).$$

Empirically, both the integration and the optimization can be approximated using the available Monte Carlo samples. We remark that, in principle, one can obtain the maximum a posteriori (MAP) estimate of $\boldsymbol{A}$ based on its posterior samples. However, due to the discrete nature of $\boldsymbol{A}$, even the most probable estimate may only be visited once or twice during the entire course of MCMC and therefore the MAP is in general not a reliable estimate.

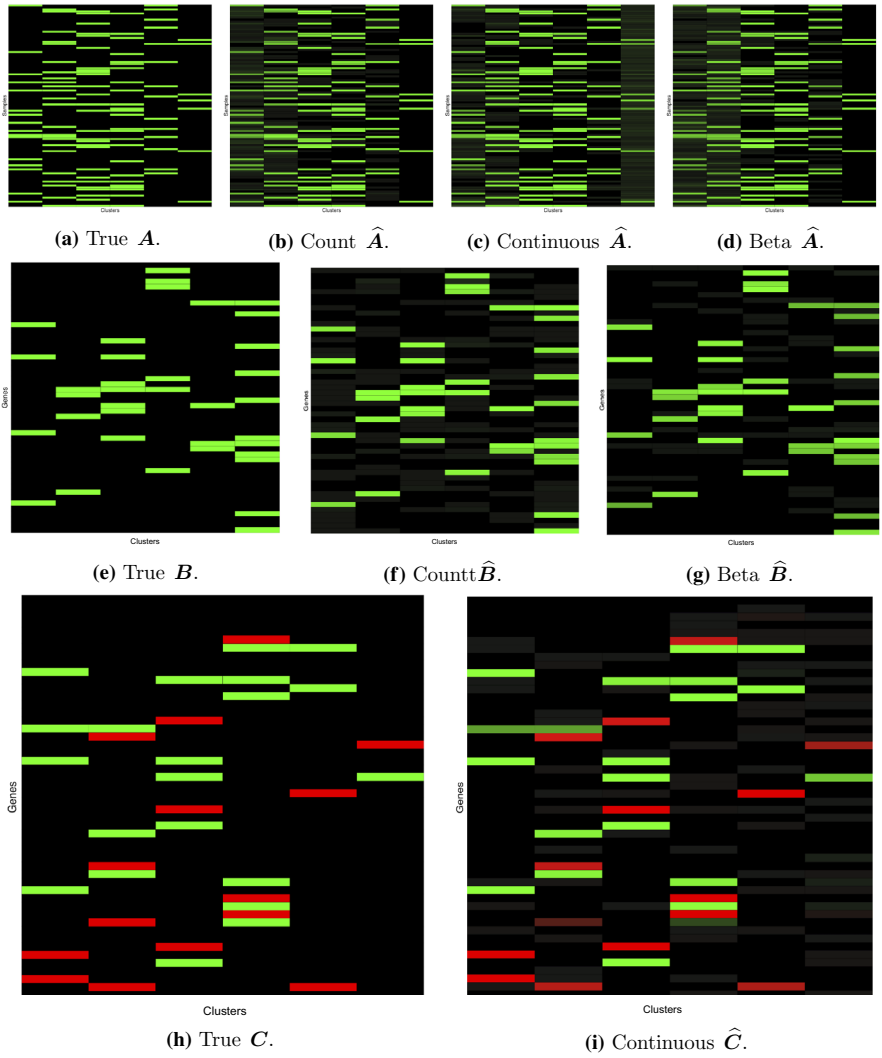
Conditional on $\widehat{\boldsymbol{A}}$, we continue to run the chain for a short period, then point estimates of other parameters are obtained as posterior means computed from those new Monte Carlo samples. Similar approaches have been adopted by [33, 45].

The code implementing the proposed models can be found in the GitHub repository at https://github.com/fangting-zhou/unified-matrix-factorization.

4 Simulation

To assess the utility of the proposed UMF, we conduct three simulation studies for sparse count data, heavy-tailed continuous data, and beta-valued data which respectively represent applications of scRNA-seq, bulk RNA-seq, and DNA methylation. We consider datasets with $n = 1000$ observations and $p = 50$ covariates. We generate the sample-latent matrix $\boldsymbol{A}$ from IBP resulting in K columns. The binary covariate-latent matrix $\boldsymbol{B}$ is generated as independent Bernoulli with success probability ρ, and the ordinal $\boldsymbol{C}$ is generated as independent categorical with

$\gamma_{-1} = \gamma_1 = \rho/2$. For each data type, we consider two scenarios $(K, \rho) = (6, 0.15)$ and $(K, \rho) = (9, 0.3)$. The generated $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ for $(K, \rho) = (6, 0.15)$ are depicted in the left panels of the first, second, and third rows of Fig. 2, respectively. We generate latent indicators z_{ij}'s from the latent (two-class or multi-class) logistic



(a) True $\boldsymbol{A}$. **(b)** Count $\widehat{\boldsymbol{A}}$. **(c)** Continuous $\widehat{\boldsymbol{A}}$. **(d)** Beta $\widehat{\boldsymbol{A}}$.

(e) True $\boldsymbol{B}$. **(f)** Countt$\widehat{\boldsymbol{B}}$. **(g)** Beta $\widehat{\boldsymbol{B}}$.

(h) True $\boldsymbol{C}$. **(i)** Continuous $\widehat{\boldsymbol{C}}$.

Fig. 2 Simulation results averaged over 50 replicates in scenario $(K, \rho) = (6, 0.15)$. **a**, **e** and **h** The true values of $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$ used to generate the simulated data. Green, black, and red cells are -1, 0, 1 respectively. **b**–**d**, **f**, **g** and **i** The estimated $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$ averaged over 50 repeated simulations. The colors of cells change gradually from red to black to green representing the values from -1 to 0 to 1. For visualization, we plot the same 100 randomly selected rows from $\boldsymbol{A}$ and $\widehat{\boldsymbol{A}}$ (Color figure online)

model described in Sect. 2.2. Finally, we generate observations x_{ij}'s from the sampling models specific to each data type as described in Sect. 2.1.

We run the MCMC algorithm for 10,000 iterations with one random initial cluster. The first 5000 iterations are discarded as burn-in and posterior samples are retained every 5th iteration after burn-in. We compare the proposed method with two competing methods. The first method is a two-step approach (TSA) that is similar to the proposed matrix factorization. But instead of joint modeling, it first dichotomizes the observations at quantile 0.25 for the sparse count data and beta-valued data or trichotomizes at quantiles (0.25, 0.75) for the heavy-tailed continuous data, similar in essence to [9] and then uses the same Bayesian nonparametric binary/categorical matrix factorization method in Sects. 2.2 and 2.3 to the discretized data. The second method that we compare to is a sparse non-negative matrix factorization (SNMF, [22]), where we regard sparsity as presence/absence of clusters. We set the number of components and sparsity parameters in this algorithm to the truth. For the continuous data, we convert them to positive numbers by taking the absolute values for SNMF.

For the sparse count data and beta-valued data, we report the estimation errors of $\boldsymbol{A}$ and $\boldsymbol{B}$. Analogously, the estimation errors of $\boldsymbol{A}$ and $\boldsymbol{C}$ for heavy-tailed continuous data are also presented. Specifically, we compute the Hamming distances between the estimated and true $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, and normalize them by the respective total number of elements in the corresponding matrices. When two matrices have different numbers of columns, we pad missing columns with zeros. The results under two sets of true values of (K, ρ) are summarized in Table 1 based on 50 repeated simulations. The performance of the proposed method is consistently better than SNMF and TSA that depends crucially on the choices of quantile cutoffs. It shows that the ad hoc choice of dichotomizing and/or trichotomizing the observations may lead to suboptimal results. The proposed UMF, on the other hand, solves this issue by adaptively discretizing the data under a Bayesian paradigm. Moreover, UMF correctly identifies the number of clusters K in 96% of the simulations whereas TSA often underestimates it when the cutoff quantile is not appropriately chosen for the simulated data. It is further worth noting that the performance of SNMF is inferior to the proposed UMF even though the number of clusters is specified to the truth, which suggests that the highly noisy and heterogeneous nature of multi-omics data poses challenges to traditional statistical methods. Figure 2 depicts the estimated sample-latent matrix $\boldsymbol{A}$ and covariate-latent matrices $\boldsymbol{B}$ or $\boldsymbol{C}$ of the proposed method averaging over repeat simulations in the scenario $(K, \rho) = (6, 0.15)$, after adjusting for label switching. These findings are visually quite close to the truth indicating that our proposed method is able to identify the overlapping clustering structures of both observations and covariates with high accuracy.

5 Applications

We illustrate the proposed method with four applications: (i) a scRNA-seq gene expression dataset with AhR intestinal stem cell-specific knockout versus wild type mouse, (ii) a combination of human B lymphocyte scRNA-seq gene expression and

Table 1 Simulation results of our unified matrix factorization and two competing approaches

Methods	$(K, \rho) = (6, 0.15)$						$(K, \rho) = (9, 0.3)$					
	Count		Continuous		Beta		Count		Continuous		Beta	
	Err $\boldsymbol{A}$	Err $\boldsymbol{B}$	Err $\boldsymbol{A}$	Err $\boldsymbol{C}$	Err $\boldsymbol{A}$	Err $\boldsymbol{B}$	Err $\boldsymbol{A}$	Err $\boldsymbol{B}$	Err $\boldsymbol{A}$	Err $\boldsymbol{C}$	Err $\boldsymbol{A}$	Err $\boldsymbol{B}$
UMF	0.047 (0.008)	0.023 (0.009)	0.052 (0.044)	0.031 (0.019)	0.056 (0.020)	0.020 (0.009)	0.054 (0.004)	0.056 (0.002)	0.082 (0.023)	0.065 (0.031)	0.093 (0.067)	0.027 (0.041)
TSA	0.014 (0.002)	0.036 (0.010)	0.059 (0.038)	0.107 (0.022)	0.015 (0.017)	0.029 (0.020)	0.143 (0.071)	0.125 (0.071)	0.133 (0.030)	0.267 (0.018)	0.171 (0.054)	0.162 (0.085)
SNMF	0.275 (0.050)	0.158 (0.036)	0.267 (0.043)	0.197 (0.033)	0.294 (0.029)	0.201 (0.035)	0.315 (0.024)	0.366 (0.042)	0.316 (0.036)	0.324 (0.053)	0.357 (0.022)	0.434 (0.036)

Average errors in estimating $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$ are quantified as the Hamming distance between the estimated and true $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$, normalized by the respective total number of elements. Numbers within the parentheses are the standard deviations over 50 replicates

scATAC-seq chromatin accessibility dataset, (iii) a TCGA breast cancer bulk RNA-seq gene expression dataset, and (iv) a TCGA head and neck cancer DNA methylation dataset.

5.1 Descriptions of Four Datasets

5.1.1 AhR Mouse scRNA-Seq Data

scRNA-seq technologies catalogue transcriptome at the single-cell level. In this study, we analyze a scRNA-seq gene expression dataset collected on two groups of mice from controlled experiments, AhR wild type (normal) and AhR knockout targeted to gastrointestinal LGR5+ stem cells. AhR, the Aryl hydrocarbon receptor, is a ligand-activated transcription factor that is capable of integrating external environmental, e.g., dietary, stimuli and host responses to modulate intestinal stem cell development, tissue regeneration, and colon cancer risk [27, 52]. We focus on a list of putative marker genes that are potentially differentially expressed between cell types. Among the list of genes, we filter out those that appear in less than 5% of samples, and discard samples that contain zero counts. The resulting dataset contains 12,812 observations with 6575 samples from experimental group and 6237 from control group, and provides expression level of 45 genes. The proposed UMF will be used to simultaneously cluster cells and genes with possible overlaps. Allowing a cell to potentially belong to more than one cluster is important in this application because cells might be undergoing dynamic cell differentiation processes at the time of measurement and therefore the transitioning cells do not belong to any definitive cell type biologically [39].

5.1.2 Human Lymphoblastoid scRNA-Seq & scATAC-Seq Data

This dataset is made of two single-cell sequencing modalities: one is the scRNA-seq and the other is the scATAC-seq, i.e., the single-cell chromatin accessibility profile. In this dataset, 7247 cells were measured with scRNA-seq and 3664 cells were measured with scATAC-seq, all from the same lymphoblastoid cell line (LCL) GM12878. One advantage of the proposed UMF is able to jointly cluster cells and genes by integrating the two single-cell sequencing modalities. We select top 25 genes with the largest variability across cells for subsequent clustering analysis, which is potentially useful for detecting previously uncharacterized cell subtypes in LCL and identifying corresponding marker genes.

5.1.3 TCGA Breast Cancer Bulk RNA-Seq Data

The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) characterized tens of thousands cancer and normal samples spanning major cancer types that include genetic, epigenetic, genomic, and proteomic features. We analyze a TCGA breast cancer dataset involving 658 primary tumor samples downloaded from the TCGA database using software TCGA-Assembler [55]. Breast cancer is the most common

cancer diagnosed among US women and is the second leading cause of cancer death among women [13]. Here, we focus our effort on clustering breast cancer patients using bulk RNA-seq gene expression data. Among more than 15,000 genes in the raw data, we select 45 genes in the ERBB signaling pathway [1], which is a crucial pathway in breast cancer development and progression. Application of the proposed UMF to this dataset can help define breast cancer subtypes and find associated marker genes.

5.1.4 TCGA Head and Neck Cancer DNA Methylation Data

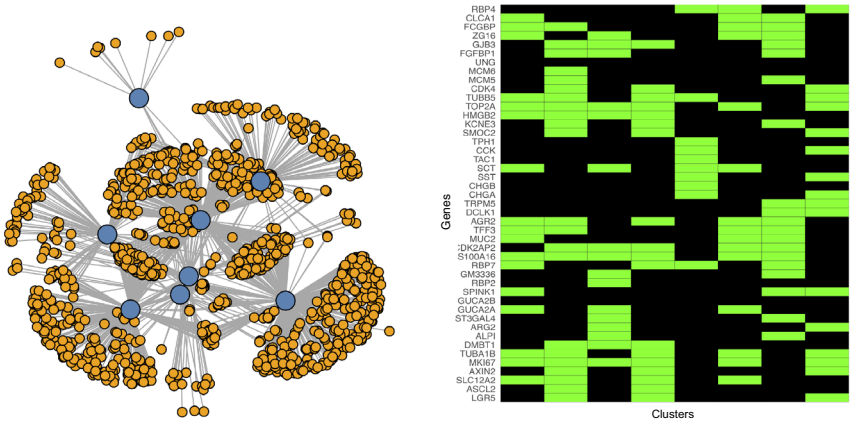
DNA methylation is an epigenetic modification which modifies the gene expression. It is known to play key roles in the carcinogenesis of head and neck squamous cell carcinoma [62] through e.g., silencing the expression of tumor suppressor genes when their promoter regions are methylated. We analyze a TCGA dataset containing 298 primary tumor samples from head and neck cancer patients. We concentrate on 21 genes from the tumor suppressor gene panels that are candidate genes frequently methylated in head and neck cancer [12]. Clustering based on methylation profiles can help elucidate the heterogeneity among head and neck cancer patients and methylated genes.

5.2 Results

We apply the proposed method to these four datasets. To check the model fit adequacy, we perform within-sample prediction that compares the observed measurements with the posterior predictive mean. The correlations between observations and predictions are 0.93, 0.98, 0.96 and 0.98 for the AhR mouse scRNA-seq data, human lymphoblastoid scRNA-seq & scATAC-seq data, TCGA breast cancer bulk RNA-seq data, and TCGA head and neck cancer DNA methylation data, respectively, which indicates adequate model fit.

5.2.1 AhR Mouse scRNA-Seq Data

We find 8 clusters as shown in Fig. 3a and b. Clusters 2 and 4 are likely subtypes of intestinal stem cells that allow for intestinal epithelium tissue repair and regeneration because they contain the well-known intestinal stem cell marker genes, LGR5, ASCL2, SLC12A2, AXIN2, SMOC2, and KCNE3 [44]. Likewise, clusters 1 and 6 can be interpreted as goblet cell subtypes as they include typical mucus components GUCA2A, MUC2, TFF3, ZG16, FCGBP, and CLCA1 which are identified as markers of goblet cells in gastrointestinal tract [7]. Cluster 5 contains marker genes of enteroendocrine cells CCK, TAC1, SCT, SST, CHGA, and CHGB [15]. This cluster also consists of neuropeptides TAC1 and CCK, which are markers of enteric neurons in the gastrointestinal tract [57]. Interestingly, it is well known that enteroendocrine cells produce gastrointestinal hormones or peptides in response to various stimuli and transmit them to the enteric nervous system to activate nervous responses [50]. Additionally, we also find some genes that are selected across several clusters. For



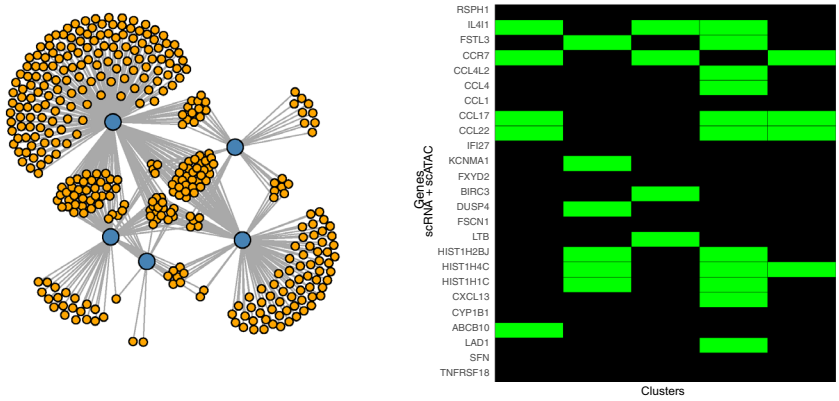
(a) Sample clusters. **(b)** $\boldsymbol{B}$: gene clusters.

Fig. 3 AhR mouse gastrointestinal scRNA-seq data. **a** Bipartite graph for clusters of randomly sampled units. Blue nodes are clusters, orange nodes are samples, and edges present sample-cluster relationship. **c** Heatmap of gene clusters. Green and black cells are 1 and 0, respectively (Color figure online)

example, TRPM5, over-expressed in clusters 7 and 8, is an intrinsic signaling component of mammalian chemosensory organs [25]. S100A16, over-expressed in all but clusters 5 and 8, acts as a novel adipogenesis promoting factor [37]. Despite the known functions of these genes, further experiments are needed to investigate the mechanism of their over-expression across clusters.

5.2.2 Human Lymphoblastoid scRNA-Seq & scATAC-Seq Data

We identify 5 clusters shown in Fig. 4a and b, representing 5 potential cell subtypes in the lymphoblastoid cell line (LCL) GM12878, which is derived from human B cells. We find that these subtypes show profiles similar to those of B cells and other antigen-presenting cells. For example, cluster 3 contains the known marker genes of B cells, LTB, CCR7, and BIRC3. Clusters 1 and 5 contain marker genes of dendritic cells, CCL22, CCL17, and CCR7. Dendritic cells are antigen-presenting cells of the mammalian immune system [2]. While the human LCLs are supposed to be a homogeneous population, our analysis suggests that there is an uncharacterized, cell subtype structure among the cells, suggesting the more heterogeneous nature of LCLs. Specifically, in GM12878, we find clusters (1, 3, and 5) with profiles similar to those of B cells and dendritic cells, two cell types well known to closely interact with each other [10, 29]. Additionally, several genes belong to multiple clusters and appear to play multiple roles in cellular processes. For example, studies have shown that IL4I1 (over-expressed in clusters 1, 3, and 4) regulates multiple steps in B cell physiology [5]. Three histone coding genes HIST1H1C, HIST1H4C, and HIST1H2BJ are involved in clusters 2 and 4, but their role in cellular differentiation has yet to be established in the literature. Further experiments need to be conducted



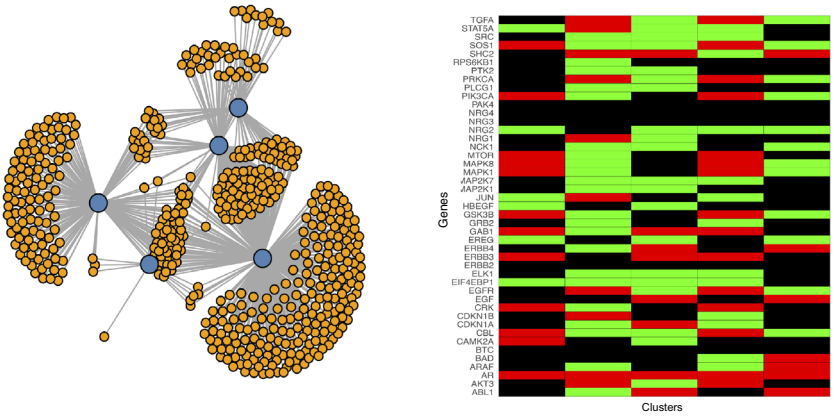
(a) Sample clusters. **(b)** $\boldsymbol{B}$: gene clusters.

Fig. 4 Human lymphoblastoid scRNA-seq & scATAC-seq data. **a** Bipartite graph for sample clusters. Blue nodes are clusters, orange nodes are 500 randomly selected samples, and edges present sample-cluster relationship. **b** Heatmap of gene clusters. Green and black cells are 1 and 0, respectively (Color figure online)

to verify the biological significance of these findings which can potentially confirm new cell subtypes in LCL.

5.2.3 TCGA Breast Cancer Bulk RNA-Seq Data

We discover 5 clusters as shown in Fig. 5a and b. Genes in MAPK signaling pathway is entirely included in cluster 2, while cluster 3 and cluster 4 also contain some



(a) Sample clusters. **(b)** $\boldsymbol{C}$: gene clusters.

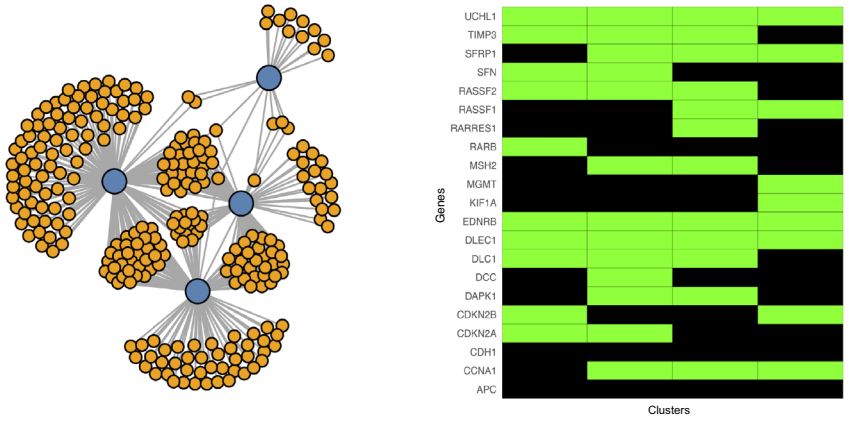
Fig. 5 TCGA breast cancer bulk RNA-seq gene expression data. **a** Bipartite graph for sample clusters. Blue nodes are clusters, orange nodes are samples, and edges present sample-cluster relationship. **b** Heatmap of gene clusters. Green, black, and red cells are 1, 0, −1 respectively (Color figure online)

of its members. The crucial role of MAPK signaling pathway in breast cancer cell growth is well established in the literature (see, for example, [18]). Primary genes in PI3K-AKT signaling pathway, which plays a significant role in cell growth and tumor proliferation in breast cancer [48], are detected in all clusters. Clinical experiments also showed that these two pathways have significant cross-talk [32]. Clusters 2 and 3 primarily contain members of the SRC/PTK2 pathway which has been identified as a promising therapeutic target in cancer [6]. The EPHB4 receptor suppresses breast cancer through the ABL1/CRK pathway [46], which is involved in cluster 2.

Genes that are significantly associated with cancer progression appear in several clusters. For instance, CBL enhances breast tumor formation and is over-expressed in human breast cancer [24]. We find CBL is under-expressed in clusters 1 &4 and over-expressed in clusters 2, 3, and 5. HBEGF is over-expressed in clusters 1 and 3 which is known as a potent inducer of cancer tumor growth and angiogenesis [47]. NCK1, over-expressed in clusters 2, 3, and 5, advances breast carcinoma cell progression and metastasis [42]. Increased (clusters 1, 3, and 5) and decreased (cluster 2) levels of STAT5A are both found in breast cancer [58]. The ERBB family includes epidermal growth factor receptor and ERBB2, ERBB3, and ERBB4 which are often mutated in cancers [54]. We find ERBB family members exhibit heterogeneous expression levels across clusters.

5.2.4 TCGA Head and Neck Cancer DNA Methylation Data

We find 4 clusters as shown in Fig. 6a and b . Genes DLC1, DLEC1, EDNRB, and UCHL1 are included in all clusters indicating their universal roles in head and neck cancer, which has been previously reported [35] in nasopharyngeal



(a) Sample clusters. **(b)** B: gene clusters.

Fig. 6 TCGA head and neck cancer DNA methylation data. **a** Bipartite graph for sample clusters. Blue nodes are clusters, orange nodes are samples, and edges present sample-cluster relationship. **b** Heatmap of gene clusters. Green and black cells are 1 and 0, respectively (Color figure online)

carcinoma (a subtype of head and neck cancer). Inactivation of CDKN2A (methylated in clusters 1 and 2) and CDKN2B (methylated in clusters 1 and 4) is frequently found in head and neck cancer [53]. They provide instructions for making P14, P15, and P16 proteins that are tumor suppressors keeping cells from growing and dividing rapidly. However, increased methylation represses their expression which can ultimately lead to malignancy and the uncontrolled tumor growth. Patients belonging to different clusters exhibit diverse methylation patterns and hence may have different responses to treatments. Co-methylated genes discovered by our method can be potential clinical targets with future experiments.

For comparison, we also apply SNMF to the same four datasets with results given in the Supplementary Materials. We tried different combinations of sparsenesses and dimensionalities, and selected the one with minimal loss. Generally speaking, the obtained clusters lack meaningful interpretations.

6 Discussion

In this paper, we developed a unified framework to simultaneously cluster observations and covariates for multi-omics data. The proposed approach accounts for the noisy, heterogeneous, sparse, and non-Gaussian nature of sequencing data, and describes the data generating process by a hierarchical Bayesian model, which allows for probabilistic characterization of latent structures via overlapping clusters through full posterior inference with natural uncertainty quantification. Using simulation studies and four real applications, we have demonstrated the proposed method is capable of identifying biologically meaningful clusters and is widely applicable to different types of multi-omics data.

There are a few directions to extend this work. First, the joint modeling approach can be used for many other tasks beyond matrix factorization. For example, gene expression networks can be inferred by replacing the matrix factorization model with a graphical model (e.g. Markov random fields or Bayesian networks) on the latent categorical/binary indicators $\boldsymbol{Z}$. Second, MCMC allows for full posterior inference but is not scalable to large and high-dimensional data. The current inference algorithm can be substantially accelerated by using consensus Monte Carlo algorithms for big-data clustering without sacrificing much accuracy. Finally, the overlapping clusters can be restricted to non-overlapping clusters if desired by considering random partition models including various extensions of the Dirichlet process.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12561-022-09350-w.

Acknowledgements Yang Ni is partially supported by the National Science Foundation, NSF DMS-1918851 and NSF DMS-2112943. Robert S. Chapkin is partially supported by the Allen Endowed Chair in Nutrition & Chronic Disease Prevention, and the National Institutes of Health (Grant Nos. R01-ES025713, R01-CA202697, R35-CA197707, and T32-CA090301). Kejun He is partially supported by the National Natural Science Foundation of China under Grant 11801560.

References

1. Arteaga CL, Moulder SL, Yakes FM (2002) HER (ERBB) tyrosine kinase inhibitors in the treatment of breast cancer. Semin Oncol 29:4–10
2. Banchereau J, Steinman RM (1998) Dendritic cells and the control of immunity. Nature 392(6673):245–252
3. Banerjee A, Krumpelman C, Ghosh J, Basu S, Mooney RJ (2005) Model-based overlapping clustering. In Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. pp 532–537
4. Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geosci 10(2–3):191–203
5. Bod L, Douguet L, Auffray C, Lengagne R, Bekkat F, Rondeau E, Molinier-Frenkel V, Castellano F, Richard Y, Prévost-Blondel A (2018) IL-4-induced gene 1: a negative immune checkpoint controlling B cell differentiation and activation. J Immunol 200(3):1027–1038
6. Bolós V, Gasent JM, Lopez-Tarruella S, Grande E (2010) The dual kinase complex FAK-SRC as a promising therapeutic target in cancer. OncoTargets Therapy 3:83
7. Brenna Ø, Furnes MW, Munkvold B, Kidd M, Sandvik AK, Gustafsson BI (2016) Cellular localization of guanylin and uroguanylin mRNAs in human and rat duodenal and colonic mucosa. Cell Tissue Res 365(2):331–341
8. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523(7561):486–490
9. Cai T, Li H, Ma J, Xia Y (2019) Differential Markov random field analysis with an application to detecting differential microbial community networks. Biometrika 106(2):401–416
10. Clark EA (1997) Regulation of B lymphocytes by dendritic cells. J Exp Med 185(5):801–804
11. Cleuziou G (2008) An extended version of the k-means method for overlapping clustering. In Proceedings of the 19th international conference on pattern recognition. pp 1–4
12. Demokan S, Dalay N (2011) Role of DNA methylation in head and neck cancer. Clin Epigenet 2(2):123
13. DeSantis CE, Ma J, Sauer AG, Newman LA, Jemal A (2017) Breast cancer statistics, 2017, racial disparity in mortality by state. CA Cancer J Clin 67(6):439–448
14. Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, Wildberg A, Wang W (2015) Normalization and noise reduction for single cell RNA-seq experiments. Bioinformatics 31(13):2225–2227
15. Engelstoft MS, Lund ML, Grunddal KV, Egerod KL, Osborne-Lawrence S, Poulsen SS, Zigman JM, Schwartz TW (2015) Research resource: a chromogranin a reporter for serotonin and histamine secreting enteroendocrine cells. Mol Endocrinol 29(11):1658–1671
16. Ghahramani Z, Griffiths TL (2006) Infinite latent feature models and the Indian buffet process. In Advances in neural information processing systems. pp 475–482
17. Gopalan P, Ruiz FJ, Ranganath R, Blei D (2014) Bayesian nonparametric Poisson factorization for recommendation systems. In Proceedings of the seventeenth international conference on artificial intelligence and statistics, pp 275–283
18. Haagenson KK, Wu GS (2010) The role of MAP kinases and MAP kinase phosphatase-1 in resistance to breast cancer treatment. Cancer Metastasis Rev 29(1):143–149
19. Hartigan JA (1972) Direct clustering of a data matrix. J Am Stat Assoc 67(337):123–129
20. Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat) 28(1):100–108
21. Heppner GH, Miller BE (1983) Tumor heterogeneity: biological implications and therapeutic consequences. Cancer Metastasis Rev 2:5–23
22. Hoyer PO (2004) Non-negative matrix factorization with sparseness constraints. J Mach Learn Res 5(11):1457–1469
23. Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32(3):241–254
24. Kang JM, Park S, Kim SJ, Hong H, Jeong J, Kim H (2012) CBL enhances breast tumor formation by inhibiting tumor suppressive activity of TGF-$\beta $ signaling. Oncogene 31(50):5123–5131
25. Kaske S, Krasteva G, König P, Kummer W, Hofmann T, Gudermann T, Chubanov V (2007) TRPM5, a taste-signaling transient receptor potential ion-channel, is a ubiquitous signaling component in chemosensory cells. BMC Neurosci 8:49

26. Kim JK, Kolodziejczyk AA, Ilicic T, Teichmann SA, Marioni JC (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat Commun 6:8687
27. Kim E, Davidson LA, Zoh RS, Hensel ME, Salinas ML, Patil BS, Jayaprakasha GK, Callaway ES, Allred CD, Turner ND, Weeks BR, Chapkin RS (2016) Rapidly cycling LGR5+ stem cells are exquisitely sensitive to extrinsic dietary factors that modulate colon cancer risk. Cell Death Dis 7(11):e2460
28. Kiselev VY, Andrews TS, Hemberg M (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 20(5):273–282
29. Kranich J, Krautler NJ (2016) How follicular dendritic cells shape the B-cell antigenome. Front Immunol 7:225
30. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
31. Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, pp 556–562
32. Lee E-R, Kim J-Y, Kang Y-J, Ahn J-Y, Kim J-H, Kim B-W, Choi H-Y, Jeong M-Y, Cho S-G (2006) Interplay between PI3K/AKT and MAPK signaling pathways in DNA-damaging drug-induced apoptosis. Biochimica et Biophysica Acta (BBA)-Mol Cell Res 1763(9):958–968
33. Lee J, Müller P, Gulukota K, Ji Y (2015) A Bayesian feature allocation model for tumor heterogeneity. Ann Appl Stat 9(2):621–639
34. Leek JT (2014) Svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res 42(21):e161
35. Li L, Tao Q, Jin H, Van Hasselt A, Poon FF, Wang X, Zeng M-S, Jia W-H, Zeng Y-X, Chan AT et al (2010) The tumor suppressor UCHL1 forms a complex with P53/MDM2/ARF to promote P53 signaling and is frequently silenced in nasopharyngeal carcinoma. Clin Cancer Res 16(11):2949–2958
36. Lin Z, Zamanighomi M, Daley T, Ma S, Wong WH (2020) Model-based approach to the joint analysis of single-cell data on chromatin accessibility and gene expression. Stat Sci 35(1):2–13
37. Liu Y, Zhang R, Xin J, Sun Y, Li J, Wei D, Zhao AZ (2011) Identification of S100A16 as a novel adipogenesis promoting factor in 3T3-L1 cells. Endocrinology 152(3):903–911
38. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinf 1(1):24–45
39. Mallik S, Zhao Z (2019) Multi-objective optimized fuzzy clustering for detecting cell clusters from single-cell expression profiles. Genes 10(8):611
40. Marusyk A, Polyak K (2010) Tumor heterogeneity: causes and consequences. Biochimica et Biophysica Acta (BBA) 1805(1):105–117
41. McLachlan GJ, Peel D (2004) Finite mixture models. Wiley, Hoboken
42. Morris DC, Popp JL, Tang LK, Gibbs HC, Schmitt E, Chaki SP, Bywaters BC, Yeh AT, Porter WW, Burghardt RC et al (2017) NCK deficiency is associated with delayed breast carcinoma progression and reduced metastasis. Mol Biol Cell 28(24):3500–3516
43. Müller P, Quintana FA, Jara A, Hanson T (2015) Bayesian nonparametric data analysis. Springer, Berlin
44. Muñoz J, Stange DE, Schepers AG, Van De Wetering M, Koo B-K, Itzkovitz S, Volckmann R, Kung KS, Koster J, Radulescu S et al (2012) The LGR5 intestinal stem cell signature: robust expression of proposed quiescent '+ 4' cell markers. EMBO J 31(14):3079–3091
45. Ni Y, Müller P, Ji Y (2019) Bayesian double feature allocation for phenotyping with electronic health records. J Am Stat Assoc 115:1–15
46. Noren NK, Foos G, Hauser CA, Pasquale EB (2006) The EPHB4 receptor suppresses breast cancer cell tumorigenicity through an ABL-CRK pathway. Nat Cell Biol 8(8):815–825
47. Ongusaha PP, Kwak JC, Zwible AJ, Macip S, Higashiyama S, Taniguchi N, Fang L, Lee SW (2004) HB-EGF is a potent inducer of tumor growth and angiogenesis. Can Res 64(15):5283–5290
48. Paplomata E, O'Regan R (2014) The PI3K/AKT/MTOR pathway in breast cancer: targets, trials and biomarkers. Therap Adv Med Oncol 6(4):154–166
49. Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E (2002) A statistical framework for expression-based molecular classification in cancer. J R Stat Soc Ser B (Statistical Methodology) 64(4):717–736
50. Rehfeld JF (1998) The new biology of gastrointestinal hormones. Physiol Rev 78(4):1087–1108
51. Ročková V, George EI (2016) Fast Bayesian factor analysis via automatic rotations to sparsity. J Am Stat Assoc 111(516):1608–1622

52. Safe S, Han H, Goldsby J, Mohankumar K, Chapkin RS (2018) Aryl hydrocarbon receptor (AhR) ligands as selective AhR modulators: genomic studies. Current Opin Toxicol 11:10–20
53. Shintani S, Nakahara Y, Mihara M, Ueyama Y, Matsumura T (2001) Inactivation of the P14ARF, P15INK4B and P16INK4A genes is a frequent event in human oral squamous cell carcinomas. Oral Oncol 37(6):498–504
54. Stern DF (2000) Tyrosine kinase signalling in breast cancer: ERBB family receptor tyrosine kinases. Breast Cancer Res 2(3):176
55. Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y (2018) TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. Bioinformatics 34(9):1615–1617
56. Xu Y, Lee J, Yuan Y, Mitra R, Liang S, Müller P, Ji Y (2013) Nonparametric Bayesian bi-clustering for next generation sequencing count data. Bayesian Anal 8(4):759
57. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, Van Der Zwan J, Häring M, Braun E, Borm LE, La Manno G et al (2018) Molecular architecture of the mouse nervous system. Cell 174(4):999–1014
58. Zeng Y, Min L, Han Y, Meng L, Liu C, Xie Y, Dong B, Wang L, Jiang B, Xu H et al (2014) Inhibition of STAT5A by NAA10P contributes to decreased breast cancer metastasis. Carcinogenesis 35(10):2244–2253
59. Zhang Z, Li T, Ding C, Zhang X (2007) Binary matrix factorization with applications. In Seventh IEEE international conference on data mining, pp 391–400
60. Zhang Z-Y, Li T, Ding C, Ren X-W, Zhang X-S (2010) Binary matrix factorization for analyzing gene expression data. Data Min Knowl Disc 20:28–52
61. Zhou M, Hannah L, Dunson D, Carin L (2012) Beta-negative binomial process and Poisson factor analysis. In Proceedings of the fifteenth international conference on artificial intelligence and statistics. pp 1462–1471
62. Zhou C, Ye M, Ni S, Li Q, Ye D, Li J, Shen Z, Deng H (2018) DNA methylation biomarkers for head and neck squamous cell carcinoma. Epigenetics 13(4):398–409
63. Zhou F, He K, Li Q, Chapkin RS, Ni Y (2021) Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization. Biostatistics

Authors and Affiliations

Fangting Zhou[1,2] · Kejun He[1] · James J. Cai[3] · Laurie A. Davidson[4,5] · Robert S. Chapkin[4,5] · Yang Ni[2]

[1] Institute of Statistics and Big Data, Renmin University of China, Beijing, China

[2] Department of Statistics, Texas A&M University, College Station, USA

[3] Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, USA

[4] Department of Nutrition and Food Science, Texas A&M University, College Station, USA

[5] Program in Integrative Nutrition and Complex Diseases, Texas A &M University, College Station, USA