Statistics and Its Interface Volume 17 (2024) 219–230

Multi-way overlapping clustering by Bayesian tensor decomposition

Zhuofan Wang*, Fangting Zhou*, Kejun He†, and Yang Ni†

The development of modern sequencing technologies provides great opportunities to measure gene expression of multiple tissues from different individuals. The three-way variation across genes, tissues, and individuals makes statistical inference a challenging task. In this paper, we propose a Bayesian multi-way clustering approach to cluster genes, tissues, and individuals simultaneously. The proposed model adaptively trichotomizes the observed data into three latent categories and uses a Bayesian hierarchical construction to further decompose the latent variables into lower-dimensional features, which can be interpreted as overlapping clusters. With a Bayesian nonparametric prior, i.e., the Indian buffet process, our method determines the cluster number automatically. The utility of our approach is demonstrated through simulation studies and an application to the Genotype-Tissue Expression (GTEx) RNA-seq data. The clustering result reveals some interesting findings about depression-related genes in human brain, which are also consistent with biological domain knowledge. The detailed algorithm and some numerical results are available in the online Supplementary Material, http://intlpress.com/site/pub/files/_supp/sii/2024/0017/0002/sii-2024-0017-0002-s001.pdf.

AMS 2000 subject classifications: Primary 62H30; secondary 62F15.

Keywords and phrases: Bayesian nonparametric prior, Gene expression data, Indian buffet process, Low-rank tensor, Mixture model.

1. INTRODUCTION

High-throughput sequencing technologies such as RNA-sequencing and microarrays have made it possible to measure the activity or expression of thousands of genes at once. Gene expression data are useful for identifying molecular subgroups, which can potentially help scientists understand signaling pathways better and ultimately design targeted treatments for genetic diseases [50, 8, 41]. A commonly used approach for identifying novel molecular subtypes is clustering, which is an unsupervised method for finding homogeneous subgroups with similar patterns of features and/or observations from heterogeneous data [52, 21, 12].

Gene expression data often exhibit heterogeneity across both features (genes) and observations (tissues or subjects). Hence, biclustering [20], which jointly clusters both features and observations, has become popular for analyzing gene expression data. Compared to clustering methods that cluster observations or features only, biclustering methods can relate a subset of genes to a certain group of subjects, which enhances interpretation. [10] was one of the early works to apply the method of biclustering to gene expression data. Subsequently, many biclustering models were proposed and found successful applications in RNA-sequencing and microarray data, including [28] for a plaid model, [5] for a combination method on row and column clustering, [29] for regularized singular value decomposition (SVD), among many others. For microbiome data, [53] developed a biclustering method via an identifiable Bayesian multinomial matrix factorization model, which was extended to the multi-omic data [54].

One common thread of biclustering approaches is that they are only applicable to matrices. However, this work is motivated by a brain-tissue gene expression dataset from the Genotype-Tissue Expression project [GTEx, 31], which contains gene expression measurements across subjects, brain tissues, and genes, and hence is a three-way tensor. To simultaneously characterize the heterogeneity along all the dimensions of a tensor, a multi-way clustering method is required.

Several attempts have been made to deal with general high-order tensors for the purpose of clustering. [48] proposed a tensor block model as the form of Tucker decomposition [46] where the core tensor represents the mean of subgroups and the factor matrices represent the cluster memberships. The idea of using factor matrices of Tucker decomposition to represent the cluster memberships was further exploited by [11], where the authors proposed to relax the original combinatorial optimization problem to its convex surrogate for computational efficiency. [18] developed a high-order spectral clustering (HSC) method and extended the Lloyd algorithm to a high-order version (HLloyd) to recover the membership matrices in the tensor block model. Alternative to the Tucker decomposition, the CANDECOMP/PARAFAC (CP) decomposition [9, 19, 23] has

*The first two authors contribute equally to this work.
†Corresponding authors.

also been applied for three-way tensor clustering [47]. Although these methods have been proven useful, there still exist some drawbacks. First, they do not allow overlapping clusters, which, however, are quite plausible for gene expression data. For example, genes can be active in more than one biological process by regulating or coding proteins through multiple pathways [2]. Second, a pre-specified number of clusters is often required for existing methods. The number of clusters is, however, usually unknown in real-word applications. Finally, gene expression data are often observed with high level of noises as well as missing values due to technical limitations. Ignoring these features of gene expression data may lead to erroneous (mixing up noises with the true expression variations) and/or inefficient (removing observations with missing values) scientific discoveries.

In this work, we propose a Bayesian multi-way clustering (BayMC) approach that addresses all the aforementioned limitations of existing tensor clustering methods. To be robust to the high-level noises of the gene expression data, a latent categorical tensor is introduced to adaptively trichotomize gene expression into one of the three categories, namely, over/normal/underexpression, coded as $1/0/-1$, respectively, for each gene, tissue, and subject. To simultaneously cluster genes, tissues, and subjects, the latent ternary categorical tensor is decomposed into three sparse lower-dimensional feature matrices, each of which can be interpreted as cluster indicators along one of the three dimensions of the tensor. This decomposition can be thought of as a probabilistic CP decomposition. As a by-product, the low-dimensionality of the feature matrices allows natural handling of missing data. The low-dimensional feature matrices are modeled hierarchically by the Indian Buffet Process (IBP) prior, the beta-Bernoulli prior, and the Dirichlet-categorical prior. This set of prior specifications gives rise to the desirable overlapping clusters and eliminates the need to fix the number of clusters a priori.

The rest of this paper is organized as follows. The preliminary knowledge of tensor and IBP is introduced in Section 2. We present our method in Section 3, including the proposed multi-way clustering model, the discussion on model identifiability, and the posterior inference. In Sections 4 and 5, we respectively demonstrate the utility of our model with simulation studies and the analysis of GTEx RNA-seq dataset. This paper is summarized in Section 6 with a brief concluding remark.

2. NOTATIONS AND PRELIMINARIES

In this section, we briefly introduce the notations and preliminaries of tensor and IBP, which will be used to build our multi-way clustering model hierarchically.

2.1 Tensor

A tensor is a multi-dimensional array, and the order K of a tensor is its number of dimensions, also known as ways or modes [25]. For a K-way tensor $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$, we let $x_{i_1 i_2 \ldots i_K}$ denote the $(i_1, i_2, \ldots, i_K)$-th element of $\boldsymbol{X}$, $i_k = 1, \ldots, I_k$, $k = 1, \ldots, K$, where $(i_1, i_2, \ldots, i_K)$ is called the access indices.

Rank-one tensors. We say a K-way tensor $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$ is of rank one if it can be written as the outer product of K vectors, i.e.

$$\boldsymbol{X} = \boldsymbol{a}^{(1)} \circ \boldsymbol{a}^{(2)} \circ \cdots \circ \boldsymbol{a}^{(K)},$$

where $\boldsymbol{a}^{(k)} = (a_1^{(k)}, \ldots, a_{I_k}^{(k)})^\intercal \in \mathbb{R}^{I_k}$, $k = 1, \ldots, K$, and the symbol $\circ$ represents the vector outer product. This means that each element of the tensor is the product of the corresponding vector elements, i.e., $x_{i_1 i_2 \cdots i_K} = a_{i_1}^{(1)} a_{i_2}^{(2)} \ldots a_{i_K}^{(K)}$ for $i_k = 1, \ldots, I_k$ and $k = 1, \ldots, K$.

CANDECOMP/PARAFAC decomposition. Rank-one tensors are key components of the CANDECOMP/PARAFAC (CP) decomposition, which decomposes a tensor into a sum of rank-one tensors. More precisely, we write $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$ as

$$\boldsymbol{X} \approx \sum_{r=1}^{R} \boldsymbol{a}_r^{(1)} \circ \boldsymbol{a}_r^{(2)} \circ \cdots \circ \boldsymbol{a}_r^{(K)},$$

where R is a positive integer and $\boldsymbol{a}_r^{(k)} \in \mathbb{R}^{I_k}$ for $k = 1, \ldots, K$ and $r = 1, \ldots, R$.

Fibers and matricization. Subarrays of a tensor are formed when a subset of the access indices is fixed. In particular, a mode-k fiber of a K-way tensor refers to a vector defined by fixing all access indices but the one of the k-th mode. Therefore, the mode-k fibers can be seen as the higher-order analogue of the rows or columns of a matrix. As for matricization, it is the process of reordering the elements of a K-way tensor into a matrix. In particular, the mode-k matricization of a tensor $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$, denoted by $\boldsymbol{X}_{(k)}$, arranges the mode-k fibers to be the columns of the resulting matrix. The arrangement follows the original order of the modes such that the $(i_1, i_2, \cdots, i_K)$-th element of $\boldsymbol{X}$ maps to the (i_k, j)-th element of $\boldsymbol{X}_{(k)}$, where

$$j = 1 + \sum_{d=1, d \neq k}^{K} (i_d - 1) J_d \quad \text{with} \quad J_d = \prod_{m=1, m \neq k}^{d-1} I_m.$$

This process of matricization is also known as unfolding or flattening.

Matrix Kronecker, Khatri-Rao, and Hadamard products. The Kronecker product of two generic matrices $\boldsymbol{A} = (a_{ij})_{I \times J} \in \mathbb{R}^{I \times J}$ and $\boldsymbol{B} = (b_{kl})_{K \times L} \in \mathbb{R}^{K \times L}$ is denoted by $\boldsymbol{A} \otimes \boldsymbol{B}$. The result is a matrix of size $(IK) \times (JL)$ and defined by

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1J}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2J}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}\boldsymbol{B} & a_{I2}\boldsymbol{B} & \cdots & a_{IJ}\boldsymbol{B} \end{bmatrix}.$$

The Khatri-Rao product is defined as a column-wise Kronecker product for two matrices with the same column number [43]. More precisely, let $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_L) \in \mathbb{R}^{I\times L}$ and $\boldsymbol{C}' = (\boldsymbol{c}'_1, \ldots, \boldsymbol{c}'_L) \in \mathbb{R}^{J\times L}$ be two generic matrices with the same number of columns, their Khatri-Rao product $\boldsymbol{C} \odot \boldsymbol{C}' \in \mathbb{R}^{IJ\times L}$ is defined as

$$\boldsymbol{C} \odot \boldsymbol{C}' = [\boldsymbol{c}_1 \otimes \boldsymbol{c}'_1 \quad \boldsymbol{c}_2 \otimes \boldsymbol{c}'_2 \quad \cdots \quad \boldsymbol{c}_L \otimes \boldsymbol{c}'_L],$$

where $\otimes$ denotes the Kronecker product.

The Hadamard product is the elementwise matrix product for the matrices of the same size. Given matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, both of size $I \times J$, we let $\boldsymbol{A} * \boldsymbol{B}$ denote their Hadamard product. The product $\boldsymbol{A} * \boldsymbol{B}$ is also a matrix of size $I \times J$ and is defined by

$$\boldsymbol{A} * \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \cdots & a_{1J}b_{1J} \\ a_{21}b_{21} & a_{22}b_{22} & \cdots & a_{2J}b_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}b_{I1} & a_{I2}b_{I2} & \cdots & a_{IJ}b_{IJ} \end{bmatrix},$$

where a_{ij} and b_{ij} are the (i, j)-th element of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively.

2.2 Indian buffet process

The Indian buffet process [IBP, 17] has been widely used as a Bayesian nonparametric prior on binary matrices with a finite number of rows and potentially an unbounded number of columns. The generative process of IBP is as follows. For the first row, we select the first Poisson(m) entries to be 1, where m is a hyperparameter. Then sequentially, for the i-th row, $i \geq 2$, we let m_r be the column sum of the r-th column from the current matrix with $i - 1$ rows. For all r such that $m_r > 0$, we set the r-th entry of the i-th row to 1 with probability m_r/i. Once having exhausted all r such that $m_r > 0$, we additionally set the next Poisson(m/i) number of entries to be 1.

The IBP can be represented as an alternative (but equivalent) generative process. We first assume the binary matrix $\boldsymbol{A} = [\alpha_{ir}]$ to be generated has n rows and $\widetilde{R}$ columns. Conditional on $\widetilde{R}$, α_{ir}'s are assumed to be beta-Bernoulli random variables, $\alpha_{ir}|\pi_r \overset{\text{ind}}{\sim} \text{Ber}(\pi_r)$ and $\pi_r \sim \text{Beta}(m/\widetilde{R}, 1)$, $r = 1, \ldots, \widetilde{R}$, where m is again a hyperparameter. Marginalizing out π_r, we can obtain the probability mass function for $\boldsymbol{A}$ as

$$p(\boldsymbol{A}) = \prod_{r=1}^{\widetilde{R}} \frac{m\Gamma\left(s_r + m/\widetilde{R}\right)\Gamma\left(n - s_r + 1\right)}{\widetilde{R}\,\Gamma\left(n + 1 + m/\widetilde{R}\right)}, \tag{1}$$

where $s_r = \sum_{i=1}^{n} \alpha_{ir}$ is the sum of the r-th column of $\boldsymbol{A}$. We then take $\widetilde{R} \to \infty$ and remove the columns where all the entries are zeros. Let R denote the number of remaining columns. It can be shown that $\boldsymbol{A}$ follows IBP(m) (without a specific ordering of columns) such that the probability mass function (1) can be rewritten as

$$p(\boldsymbol{A}) = \frac{m^R \exp(-mH_n)}{R!} \prod_{r=1}^{R} \frac{\Gamma\left(s_r\right)\Gamma\left(n - s_r + 1\right)}{\Gamma(n+1)},$$

where $H_n = \sum_{i=1}^{n} 1/i$ is the n-th Harmonic number. Moreover, the conditional probability for $\alpha_{ir} = 1$ is $p\left(\alpha_{ir} = 1 \mid \boldsymbol{\alpha}_{(-i)r}\right) = s_{(-i)r}/n$ provided $s_{(-i)r} > 0$, where $\boldsymbol{\alpha}_{(-i)r}$ is the r-th column of $\boldsymbol{A}$ excluding the i-th row and $s_{(-i)r}$ is the sum of $\boldsymbol{\alpha}_{(-i)r}$. The distribution of number of new columns for each row is Poisson (m/n). These properties are essential for the posterior inference in Section 3.3.

3. METHOD

3.1 Multi-way clustering model

Suppose the RNA-seq data is collected from G genes across T tissues and D donors/subjects. We focus on the log-transformed, centered messenger RNA (mRNA) measurements which are often treated as continuous data with heavy tails. We denote the obtained mRNA measurement from the g-th gene, t-th tissue, and d-th donor as y_{dtg}, $d = 1, \ldots, D$, $t = 1, \ldots, T$, and $g = 1, \ldots, G$. The overall measurements of gene expression form a three-way tensor of size $D \times T \times G$ which is denoted as $\boldsymbol{Y}$. According to [38], we use the probability of expression model (POE) to represent the normalized gene expression measurements as a mixture of one Gaussian and two uniform distributions. In other words, for $1 \leq d \leq D$, $1 \leq t \leq T$, and $1 \leq g \leq G$, we write

$$\begin{aligned} y_{dtg} \sim\ & I(z_{dtg} = -1)\text{U}(\eta_t + \mu_g - \kappa_g^-,\ \eta_t + \mu_g) \\ & + I(z_{dtg} = 0)\text{N}(\eta_t + \mu_g, \sigma_g^2) \\ & + I(z_{dtg} = 1)\text{U}(\eta_t + \mu_g,\ \eta_t + \mu_g + \kappa_g^+), \end{aligned} \tag{2}$$

where the latent categorical variable $z_{dtg} = -1$, 0, and 1 respectively indicate the case of under-expression, normal, and over-expression of gene g in the t-th tissue and the d-th donor. In (2), η_t and μ_g represent the main effect of the t-th tissue and g-th gene, respectively. σ_g is the gene specific standard deviation of normal component and the parameters κ_g^+ and κ_g^- provide the limits of the uniform components of the mixture. We also set $\min\{\kappa_g^+, \kappa_g^-\} > \kappa_0\sigma_g$ with $\kappa_0 > 5$ to provide heavier tail probability than the normal when the deviation away from the mean is not too large.

The likelihood function is thus derived from (2) as

$$
\begin{aligned}
&p(\boldsymbol{Y}|\boldsymbol{Z}, \{\eta_t\}_{t=1}^T, \{\mu_g, \sigma_g^2, \kappa_g^-, \kappa_g^+\}_{g=1}^G) \\
&\quad = \prod_{d=1}^{D}\prod_{t=1}^{T}\prod_{g=1}^{G} p(y_{dtg}|z_{dtg}, \eta_t, \mu_g, \sigma_g^2, \kappa_g^-, \kappa_g^+) \\
&\quad = \prod_{d=1}^{D}\prod_{t=1}^{T}\prod_{g=1}^{G} \Big(f_{-1}(y_{dtg})^{I(z_{dtg}=-1)} \\
&\qquad\qquad \times f_0(y_{dtg})^{I(z_{dtg}=0)} f_1(y_{dtg})^{I(z_{dtg}=1)} \Big)
\end{aligned} \tag{3}
$$

where $f_{-1}(\cdot)$, $f_0(\cdot)$, and $f_1(\cdot)$ are the density functions of $\mathrm{U}(\eta_t+\mu_g-\kappa_g^-, \eta_t+\mu_g)$, $\mathrm{N}(\eta_t+\mu_g, \sigma_g^2)$, and $\mathrm{U}(\eta_t+\mu_g, \eta_t+\mu_g+\kappa_g^+)$, respectively.

We assign the priors on the unknown parameters in (2) as $\mu_g, \eta_t \sim \mathrm{N}(\mu, \sigma^2)$, $\kappa_g^-, \kappa_g^+ \sim \mathrm{Gamma}(a_\kappa, b_\kappa)$, and $\sigma_g^2 \sim \mathrm{IG}(a_\sigma, b_\sigma)\, I(\sigma_g < \min\{\kappa_g^-, \kappa_g^+\}/\kappa_0)$, where $\mathrm{IG}(a_\sigma, b_\sigma)$ is the inverse-Gamma distribution with parameters a_σ and b_σ. Using the mixture model (2) and the priors on its parameters, the adaptive trichotomization takes into account noises due to the sequencing errors and the technical instability [38], and therefore, the latent variable z_{dtg} can be viewed as a denoised version of the gene expression measurement y_{dtg}. We denote $\boldsymbol{Z} = (z_{dtg}) \in \{-1, 0, 1\}^{D\times T\times G}$.

Although mixture model (2) helps classify the status of gene g in the t-th tissue and the d-th donor, the obtained $\boldsymbol{Z}$ remains the same dimensions of $\boldsymbol{Y}$ as a three-way categorical (ternary) tensor. To reduce the dimensionality and achieve the purpose of clustering, we introduce lower-dimensional matrices $\boldsymbol{C}_1 \in \{0,1\}^{D\times R}$, $\boldsymbol{C}_2 \in \{0,1\}^{T\times R}$, and $\boldsymbol{C}_3 \in \{-1,0,1\}^{G\times R}$ to characterize the heterogeneity along the three modes of $\boldsymbol{Z}$. In particular, with R setting to be the number of clusters, which is often much smaller than the dimensions of $\boldsymbol{Z}$ (i.e., D, T and G), $c_{1d}^r = 1$ represents that d-th donor is in cluster r and $c_{1d}^r = 0$ otherwise. Similarly, we can define membership matrices of tissues and genes as $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$, respectively. Note that $\boldsymbol{C}_3$ is a ternary matrix whose elements take values at $\{1, 0, -1\}$ to represent gene over-/normal/under-expression in cluster r. We link z_{dtg} with $\boldsymbol{C}_1, \boldsymbol{C}_2$, and $\boldsymbol{C}_3$ by a multi-class logistic model:

$$
\begin{aligned}
z_{dtg} \sim \ \mathrm{Categorical} \Big\{ & M^{-1}\exp(\theta_{dtg}^-), M^{-1}, \\
& M^{-1}\exp(\theta_{dtg}^+) \Big\},
\end{aligned} \tag{4}
$$

where M is a normalizing constant and parameters $(\theta_{dtg}^-, \theta_{dtg}^+)$ are formulated as

$$
\begin{aligned}
\theta_{dtg}^- &= \sum_{r=1}^{R} c_{1d}^r c_{2t}^r \omega_{gr}^- I\left(c_{3g}^r = -1\right) + b^- \quad \text{and} \\
\theta_{dtg}^+ &= \sum_{r=1}^{R} c_{1d}^r c_{2t}^r \omega_{gr}^+ I\left(c_{3g}^r = 1\right) + b^+.
\end{aligned} \tag{5}
$$

In (5), parameters ω_{gr}^+ and ω_{gr}^- tie the g-th gene to the r-th cluster and are assumed to be positive for identifiability. Parameters b^+ and b^- control the baseline probabilities of z_{dtg} being $+1$ and -1, respectively. We denote $\boldsymbol{\omega}^+ \in \mathbb{R}^{G\times R}$, where its (g,r)-th element is ω_{gr}^+ ($\boldsymbol{\omega}^-$ with ω_{gr}^- analogously) and $\boldsymbol{B}^+ \in \mathbb{R}^{D\times T\times G}$ with all elements being b^+ ($\boldsymbol{B}^-$ with b^- analogously). Let $\boldsymbol{\Theta}^+$ be a tensor of size $D\times T\times G$ whose (d,t,g)-th element is θ_{dtg}^+ ($\boldsymbol{\Theta}^-$ with θ_{dtg}^- analogously). Further set $\widetilde{\boldsymbol{C}}_3^+ = \boldsymbol{\omega}^+ * I(\boldsymbol{C}_3 = 1)$ and $\widetilde{\boldsymbol{C}}_3^- = \boldsymbol{\omega}^- * I(\boldsymbol{C}_3 = -1)$ as the Hadamard product of $\boldsymbol{\omega}^+$ and $\boldsymbol{\omega}^-$ respectively with $I(\boldsymbol{C}_3 = 1)$ and $I(\boldsymbol{C}_3 = -1)$, where the indicator function applies on each element of $\boldsymbol{C}_3$. We can then rewrite (5) in a tensor form,

$$
\begin{aligned}
\boldsymbol{\Theta}^- &= \sum_{r=1}^{R} \boldsymbol{c}_1^r \circ \boldsymbol{c}_2^r \circ \widetilde{\boldsymbol{c}}_3^{\,r-} + \boldsymbol{B}^- \quad \text{and} \\
\boldsymbol{\Theta}^+ &= \sum_{r=1}^{R} \boldsymbol{c}_1^r \circ \boldsymbol{c}_2^r \circ \widetilde{\boldsymbol{c}}_3^{\,r+} + \boldsymbol{B}^+,
\end{aligned} \tag{6}
$$

where $\boldsymbol{c}_1^r$, $\boldsymbol{c}_2^r$, $\widetilde{\boldsymbol{c}}_3^{\,r-}$, and $\widetilde{\boldsymbol{c}}_3^{\,r+}$ are the r-th columns of $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\widetilde{\boldsymbol{C}}_3^-$, and $\widetilde{\boldsymbol{C}}_3^+$, respectively. The proposed model (6) has the same form of CP decomposition but with a nice interpretation of simultaneous clustering of donors, tissues, and genes, indicated by binary membership matrices $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $I(\boldsymbol{C}_3 = 1)$, and $I(\boldsymbol{C}_3 = -1)$.

Although we focus on three-way tensors because of the motivating GTEx data, the proposed model (2)–(6) can be generalized to tensors of higher order (i.e., K-way tensors with $K > 3$). To this end, suppose the observed is a K-way ($K > 3$) tensor $\boldsymbol{Y}$ of size $D_1 \times D_2 \times \cdots \times D_K$ where the last mode is the main variable of interest (e.g., the gene expression in the motivating GTEx data). We assume that each entry of the observed tensor can be represented as a mixture of three components, namely, abnormally low, normal, and abnormally high according to the last mode. We use a latent ternary variable with values in $\{-1, 0, 1\}$ to indicate its mixture component as in (2). The membership matrices $\boldsymbol{C}_k \in \{0,1\}^{D_k\times R}$, $k = 1, \ldots, K-1$, and $\boldsymbol{C}_K \in \{-1,0,1\}^{D_K\times R}$ are used to characterize the heterogeneity along the modes of $\boldsymbol{Z}$ and to achieve the purpose of clustering. The link between the ternary tensor and the membership matrices is a multi-class logistic model just as in (4) with parameters $(\theta^-_{d_1 d_2\cdots d_K}, \theta^+_{d_1 d_2\cdots d_K})$. These parameters can be decomposed similarly as (5) but with K-way outer products.

The priors in the latent model (5) are assigned as follows. Each element of $\boldsymbol{C}_1$ is assumed to follow a beta-Bernoulli distribution, i.e., $c_{1d}^r \overset{\text{ind}}{\sim} \mathrm{Ber}(\rho)$ with $\rho \sim \mathrm{Beta}(a_\rho, b_\rho)$. Matrix $\boldsymbol{C}_2$ follows an IBP, $\boldsymbol{C}_2 \sim \mathrm{IBP}(m)$, which automatically determines the number R of clusters. Note that the IBP prior can be assigned to either the first or the second mode. For faster computation, we suggest assigning it to the mode with a smaller dimension. In

our application, the number of subjects (the first mode) is larger than that of tissues (the second mode). Therefore, we impose the IBP prior on the membership matrix of the second mode. Each element of $\boldsymbol{C}_3$ follows the Dirichlet-categorical distribution, i.e., $c_{3g}^r \overset{\text{ind}}{\sim} \text{Categorical}(\boldsymbol{\gamma})$ with $\boldsymbol{\gamma} = (\gamma_{-1}, \gamma_0, \gamma_1) \sim \text{Dirichlet}(\psi_{-1}, \psi_0, \psi_1)$. We also assume independently $\omega_{gr}^+, \omega_{gr}^- \sim \text{Gamma}(a_\omega, b_\omega)$, $b^+, b^- \sim N(\mu_b, \sigma_b^2)$, and $m \sim \text{Gamma}(a_m, b_m)$.

3.2 Identifiability

Unlike the matrix decomposition where its uniqueness is often not guaranteed, the condition to achieve the uniqueness of the CP decomposition of higher-order tensors is weaker. Let $k_{\boldsymbol{C}_1}$, $k_{\boldsymbol{C}_2}$, $k_{\widetilde{\boldsymbol{C}}_3^+}$, and $k_{\widetilde{\boldsymbol{C}}_3^-}$ denote the k-ranks of matrices $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\widetilde{\boldsymbol{C}}_3^+$, and $\widetilde{\boldsymbol{C}}_3^-$, respectively. Here, the k-rank of a generic matrix $\boldsymbol{C}$ is defined as the maximum value k such that any k columns of $\boldsymbol{C}$ are linearly independent, denoted by $k_{\boldsymbol{C}}$. One commonly-used sufficient condition for CP decomposition uniqueness, according to [27], is the Kruskal's condition, i.e.,

$$k_{\boldsymbol{C}_1} + k_{\boldsymbol{C}_2} + \min\{k_{\widetilde{\boldsymbol{C}}_3^+}, k_{\widetilde{\boldsymbol{C}}_3^-}\} \geq 2R + 2.$$

Note that the Kruskal's condition is sufficient for the unique CP decomposition of tensor $\boldsymbol{\Theta}^+$ (respectively $\boldsymbol{\Theta}^-$) into real-valued $\boldsymbol{C}_1, \boldsymbol{C}_2$, and $\widetilde{\boldsymbol{C}}_3^+$ (respectively $\widetilde{\boldsymbol{C}}_3^-$). Furthermore, due to the binary nature of $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, we provide an alternative sufficient condition for the identifiability of our model which has less restrictions on k-ranks compared with the Kruskal's condition.

Proposition 1. *For model (6), if there exist integer matrices $\boldsymbol{U}_1 \in \mathbb{Z}^{R\times D}$ and $\boldsymbol{U}_2 \in \mathbb{Z}^{R\times T}$ such that $\boldsymbol{U}_1\boldsymbol{C}_1 = \boldsymbol{I}_R$ and $\boldsymbol{U}_2\boldsymbol{C}_2 = \boldsymbol{I}_R$, then $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$ are uniquely identifiable up to column permutation.*

Proof. We present the proof for $\boldsymbol{\Theta}^+$; the proof for $\boldsymbol{\Theta}^-$ can be obtained similarly. Denote $\boldsymbol{\Theta}_{(2)}^+$, $\boldsymbol{B}_{(2)}^+ \in \mathbb{R}^{T\times(DG)}$ as the mode-2 unfolding of tensors $\boldsymbol{\Theta}^+$ and $\boldsymbol{B}^+$, respectively. Let

$$\boldsymbol{C}_{-2}^+ = \widetilde{\boldsymbol{C}}_3^+ \odot \boldsymbol{C}_1 \in \mathbb{R}^{DG\times R}, \tag{7}$$

where $\odot$ is the Khatri-Rao product introduced in Section 2.1. We then have

$$\boldsymbol{\Theta}_{(2)}^+ = \boldsymbol{C}_2(\boldsymbol{C}_{-2}^+)^\intercal + \boldsymbol{B}_{(2)}^+. \tag{8}$$

According to [53], identifiability holds for $\boldsymbol{C}_2$ and $\boldsymbol{C}_{-2}^+$ by the matrix form (8) under the assumption that there exists an integer matrix $\boldsymbol{U}_2 \in \mathbb{Z}^{R\times T}$ with $\boldsymbol{U}_2\boldsymbol{C}_2 = \boldsymbol{I}_R$. The uniqueness of $\boldsymbol{C}_1$ and $\boldsymbol{C}_{-1}^+$ hold by similar arguments due to $\boldsymbol{U}_1\boldsymbol{C}_1 = \boldsymbol{I}_R$, where $\boldsymbol{C}_{-1}^+ = \widetilde{\boldsymbol{C}}_3^+ \odot \boldsymbol{C}_2 \in \mathbb{R}^{TG\times R}$. Next, observed from (7), the identifiability for $\widetilde{\boldsymbol{C}}_3^+$ also holds given the uniqueness of $\boldsymbol{C}_{-2}^+$ and $\boldsymbol{C}_1$. It is because the non-uniqueness of $\widetilde{\boldsymbol{C}}_3^+$ only occurs when there exists a column of $\boldsymbol{C}_1$ to be all zeros according to the definitions of Khatri-Rao and Kronecker products in Section 2.1, which contradicts with the assumption $\boldsymbol{U}_1\boldsymbol{C}_1 = \boldsymbol{I}_R$. Finally, the uniqueness of $\boldsymbol{C}_3$ can be obtained by noticing that all elements in $\boldsymbol{\omega}^+$ are positive and $\boldsymbol{C}_3$ can be equivalently written as $\text{sgn}(\widetilde{\boldsymbol{C}}_3^+)$, where $\text{sgn}(\cdot)$ applies on each element of $\widetilde{\boldsymbol{C}}_3^+$. □

Remark 1. *The condition used in Proposition 1 is mild since it can be satisfied, taking $\boldsymbol{U}_1\boldsymbol{C}_1 = \boldsymbol{I}_R$ as an example, if for any $r = 1, \cdots, R$, there exists $d = 1, \cdots, D$ such that $\boldsymbol{c}_d = \boldsymbol{e}_r$ where $\boldsymbol{c}_d$ is the d-th row of $\boldsymbol{C}_1$ and $\boldsymbol{e}_r$ is a unit vector with 1 at its r-th entry (in this case, $\boldsymbol{U}_1$ is just a binary matrix that acts to pick out those R rows of $\boldsymbol{C}_1$). This implies that, the condition can be satisfied if for any cluster r, there exists at least one member of this cluster that does not belong to any other clusters. The identifiability result of Proposition 1 can be generalized to the general K-way tensor model. In particular, if there exist integer matrices $\boldsymbol{U}_k \in \mathbb{Z}^{R\times D_k}$ such that $\boldsymbol{U}_k\boldsymbol{C}_k = \boldsymbol{I}$, $k = 1, 2, \cdots, K-1$, then $\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_K$ are uniquely identifiable up to column permutations. Its proof is similar to that of 3-way model and thus omitted.*

3.3 Posterior inference

The proposed Bayesian multi-way clustering (BayMC) model in Section 3.1 includes the following parameters and hyperparameters

$$\{\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{Z}, \{\eta_t\}_{t=1}^T, \{\mu_g, \sigma_g^2, \kappa_g\}_{g=1}^G, \boldsymbol{\omega}^+, \boldsymbol{\omega}^-, m, \rho, \boldsymbol{\gamma}, b^+, b^-\}.$$

We will use a Markov chain Monte Carlo (MCMC) algorithm to sample these model parameters from the analytical intractable posterior distribution.

One important step in the MCMC algorithm is updating the cluster configuration through IBP construction. The identities (7) and (8) are very useful in the sampling scheme since they transfer the CP parameters of a tensor into a matrix form. Denote $\boldsymbol{c}_{2,-t}^r$ as the rth column of $\boldsymbol{C}_2$ without the t-th entry, $r = 1, \ldots, R$ and $t = 1, \ldots, T$. For $t = 1, \ldots, T$, we cycle through the following main steps to sample $\boldsymbol{C}_2$. Other parameters can be sampled using Gibbs or Metropolis-Hasting and the details can be found in Section S.1.1 of the Supplementary Material.

1. Update existing (non-empty) columns $r = 1, \ldots, R$ of $\boldsymbol{C}_2$. In particular, we sample the binary c_{2t}^r, $r = 1, \ldots, R$, from the full conditional distribution,

$$p(c_{2t}^r|\cdot) \propto p(c_{2t}^r|\boldsymbol{c}_{2,-t}^r)\, p(\boldsymbol{z}_{(2)t}|c_{2t}^r, \boldsymbol{C}_{-2}^+, \boldsymbol{C}_{-2}^-, b^+, b^-),$$

where $\boldsymbol{z}_{(2)t}$ is the t-th row of $\boldsymbol{Z}_{(2)}$, $\boldsymbol{C}_{-2}^+ = \widetilde{\boldsymbol{C}}_3^+ \odot \boldsymbol{C}_1$, and $\boldsymbol{C}_{-2}^- = \widetilde{\boldsymbol{C}}_3^- \odot \boldsymbol{C}_1$. In the above, $p(c_{2t}^r = 1|\boldsymbol{c}_{2,-t}^r) = \sum_{i\neq t} c_{2i}^r/T$ according to the generating process of IBP

(in Section 2.2), and

$$p(\boldsymbol{z}_{(2)t}|c^r_{2t}, \boldsymbol{C}^+_{-2}, \boldsymbol{C}^-_{-2}, b^+, b^-) = \prod_{d=1}^{D}\prod_{g=1}^{G} p(z_{dtg}|c^r_{2t}, \boldsymbol{C}^+_{-2}, \boldsymbol{C}^-_{-2}, b^+, b^-).$$

2. Propose new clusters. After all existing columns are updated, we propose to add new columns. We draw $R^* \sim \text{Poi}(m/T)$. If $R^* = 0$, we proceed to the next step. Otherwise, we propose a set of new parameters for the new clusters constructed in $\boldsymbol{C}^{*+}_{-2}$ and $\boldsymbol{C}^{*-}_{-2}$ from their prior distributions. We accept new columns and the associated new parameters with probability $\min(1, \gamma)$ with

$$\gamma = \frac{p(\boldsymbol{z}_{(2)t}|\boldsymbol{c}_{2t}, \boldsymbol{1}_{R^*}, \boldsymbol{C}^+_{-2}, \boldsymbol{C}^-_{-2}, \boldsymbol{C}^{*+}_{-2}, \boldsymbol{C}^{*-}_{-2}, b^+, b^-)}{p(\boldsymbol{z}_{(2)t}|\boldsymbol{c}_{2t}, \boldsymbol{C}^+_{-2}, \boldsymbol{C}^-_{-2}, b^+, b^-)},$$

where $\boldsymbol{c}_{2t}$ is the t-th row of $\boldsymbol{C}_2$ and $\boldsymbol{1}_{R^*}$ is a vector of length R^* with each entry being 1. If new columns are accepted, we increase R to $R + R^*$.

To summarize the posterior distribution based on the Monte Carlo samples, we proceed by first calculating the maximum a posteriori estimate $\widehat{R}$ of R from the marginal posterior distribution. Conditional on estimated $\widehat{R}$, we first find the point estimate of $\boldsymbol{C}_2$ by the following procedure. For any matrices $\boldsymbol{C}_2$ and $\widetilde{\boldsymbol{C}}_2 \in \{0,1\}^{T\times\widehat{R}}$, we define a distance $d(\boldsymbol{C}_2, \widetilde{\boldsymbol{C}}_2) = \min_\pi D(\boldsymbol{C}_2, \pi(\widetilde{\boldsymbol{C}}_2))$, where $\pi(\widetilde{\boldsymbol{C}}_2)$ denotes a permutation of the columns of $\widetilde{\boldsymbol{C}}_2$ and $D(\cdot,\cdot)$ is the Hamming distance between the two matrices. A point estimator $\widehat{\boldsymbol{C}}_2$ of $\boldsymbol{C}_2$ is then obtained as

$$\widehat{\boldsymbol{C}}_2 = \arg\min_{\widetilde{\boldsymbol{C}}_2} \int d(\boldsymbol{C}_2, \widetilde{\boldsymbol{C}}_2) \mathrm{d}p(\boldsymbol{C}_2 \mid \text{data}).$$

The integral as well as the optimization can be approximated using the available posterior samples. Conditional on $\widehat{\boldsymbol{C}}_2$, we continue to run the Markov chain for a while. Afterwards the point estimates of other parameters are obtained as the posterior means computed from the new Monte Carlo samples. Similar approaches have been used in [32, 54, 53].

4. SIMULATION

In our simulation study, we compare our BayMC method with three alternative approaches: (i) a sparse unified matrix factorization [UMF, 54], which is a degenerated matrix version (i.e., biclustering) of our method, (ii) the HLloyd method [18], and (iii) the MultiCluster method [47]. Since HLloyd and MultiCluster both assume each object belongs to only one cluster while our model allows overlaps, we set up two simulation cases: overlapped clusters and non-overlapped clusters for fair comparison.

Data generation. We consider a three-way tensor of $D = 200$, $T = 15$, and $G = 15$, which has a similar size to our real data application (in Section 5). The number of clusters is set to be $R = 3$. For the overlapped case, the tissue membership matrix $\boldsymbol{C}_2$ is generated from an IBP process with $m = 1$. As for the donor membership matrix $\boldsymbol{C}_1$, elements are generated from Bernoulli distribution with probability $\rho = 0.3$. The gene membership matrix $\boldsymbol{C}_3$ are generated from categorical distribution with categories $\{-1, 0, 1\}$ and event probabilities $\boldsymbol{p} = \{0.15, 0.7, 0.15\}$. We also set the weight matrices $\boldsymbol{\omega}^+$ and $\boldsymbol{\omega}^-$ to be the same and denoted as $\boldsymbol{\omega}$ below. Note that $\boldsymbol{\omega}$ can be regarded as the term controlling the signal strength of the CP decomposition. We first set each row of $\boldsymbol{\omega}$ to $(6.0, 6.5, 7.0)$ and will change the values to inspect how the performance will vary with different levels of signal strength. The results for various $\boldsymbol{\omega}$'s are presented in Section S.2.1 of the Supplementary Material. Both b^- and b^+ are set as $\log 0.1$. For the non-overlapped case, each row of the binary membership matrices $\boldsymbol{C}_1 \in \{0,1\}^{200\times 3}$ and $\boldsymbol{C}_2 \in \{0,1\}^{15\times 3}$ is independently generated from a unit-trail multinomial distribution with event probabilities $\boldsymbol{p} = (0.3, 0.3, 0.4)$. As for $\boldsymbol{C}_3 \in \{-1,0,1\}^{15\times 3}$, we first generate its rows as for $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, and afterwards we additionally set half of the 1's in $\boldsymbol{C}_3$ to be -1 in random. The weight matrix $\boldsymbol{\omega}$ and baseline parameters (b^+, b^-) are generated as in the overlapped case. In both cases, the latent indicator tensor $\boldsymbol{Z}$ is then generated according to (4) and (5). Finally, we generate the observations y_{dtg}'s through the sampling model, i.e., from $U(-5, 0)$ when $z_{dtg} = -1$, from $N(0, 1)$ when $z_{dtg} = 0$, and from $U(0, 5)$ when $z_{dtg} = 1$, for $d = 1, \ldots, D$, $t = 1, \ldots, T$, and $g = 1, \ldots, G$. For each case, 50 simulated datasets are generated independently.

Implementation of methods. To apply our approach, we run the MCMC algorithm for 5,000 iterations with one random initial cluster. The first 2,500 iterations are discarded as burn-in and posterior samples are retained every 5th iteration after burn-in. For the HLloyd and MultiCluster methods, since they need a pre-specified number of clusters, we plug in the true number of clusters when we apply them. Furthermore, since MultiCluster only provides three real-valued matrices from the CP decomposition, we additionally apply a k-means step on these matrices to obtain the membership information for the clustering purpose. As for the matrix form of our method, we transform the original tensor to a matrix by taking the average on the mode of donors.

Comparison of performance. We first summarize the correct number of clusters estimated by BayMC and UMF. It shows that our method can estimate the cluster number accurately at least 94% of the 50 simulated replicates for two cases while the matrix biclustering UMF only accurately estimates less than 10% of the replicates for the number of clusters. The failure of biclustering UMF may because it loses the information of heterogeneity from the mode of donor by taking averages. To further evaluate the performance of various method for clustering, we calculate the error between the estimated and the true membership

Table 1. Estimation errors for various methods in the simulation study. Reported are the relative Hamming distances between the estimated and true $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$. The numbers in parentheses are the standard errors

		BayMC	HLloyd	UMF	MultiCluster
Overlapped	Error of $\boldsymbol{C}_1$	**0.013** (0.009)	0.568 (0.010)	-	0.566 (0.014)
	Error of $\boldsymbol{C}_2$	**0.029** (0.021)	0.939 (0.019)	1.602 (0.011)	1.040 (0.011)
	Error of $\boldsymbol{C}_3$	**0.088** (0.022)	1.154 (0.009)	0.730 (0.033)	0.981 (0.017)
Non-overlapped	Error of $\boldsymbol{C}_1$	0.019 (0.010)	**≤0.001** (≤0.001)	-	0.052 (0.029)
	Error of $\boldsymbol{C}_2$	0.011 (0.007)	**≤0.001** (≤0.001)	0.965 (0.034)	0.013 (0.013)
	Error of $\boldsymbol{C}_3$	**0.067** (0.006)	0.893 (0.010)	0.553 (0.031)	0.872 (0.019)

matrices through Hamming distance normalized by the respective total number of elements. If the estimated number of clusters differs from the truth, we pad the smaller matrix with columns of zeros to make them comparable. Note that HLloyd and MultiCluster cannot distinguish the over-/under-expressed genes as 1 and -1 in $\boldsymbol{C}_3$, thus we set -1's in true $\boldsymbol{C}_3$ to 1's when calculating the Hamming distance for these two methods. The average results of two cases based on 50 simulated replicates are summarized in Table 1. In the overlapped case, the BayMC method outperforms the other competing methods in estimating the membership matrices. For HLloyd and MultiCluster, although they take the true number of clusters as inputs, their performance of estimating the membership matrices is significantly inferior to BayMC, which may be explained by their model assumptions without overlapping. In the non-overlapped case, which is in favor of HLloyd and MultiCluster, HLloyd has the best performance in estimating $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. However, BayMC outperforms all the competing methods in estimating $\boldsymbol{C}_3$, and its performance in estimating $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ is the second best. Note again that HLloyd and MultiCluster directly inputs the true number of clusters. Overall, the proposed BayMC method shows its advantage of clustering. In both cases, UMF shows unsatisfactory results, which may because it loses its efficiency by ignoring the heterogeneity of donors.

Sensitivity analysis. We perform sensitivity analysis under the setting of overlapped case regarding the choice of all the hyperparameters: (a_m, b_m) for m, (a_ρ, b_ρ) for ρ, $(\psi_{-1}, \psi_0, \psi_1)$ for $\boldsymbol{\gamma}$, (μ, σ^2) for $\{\eta_t\}_{t=1}^T$ and $\{\mu_g\}_{g=1}^G$, (a_κ, b_κ) for $\{\kappa_g\}_{g=1}^G$, (a_ω, b_ω) for elements in $\boldsymbol{\omega}^+$ and $\boldsymbol{\omega}^-$, and (μ_b, σ_b^2) for b^+ and b^-. Results show that the inference under the proposed BayMC model is relatively robust. Details can be found in Section S.2.2 of the Supplementary Material.

Additional simulation with missing observations. It is commonly seen that the observed gene expression data are accompanied by some missing values. For example, roughly 50% observations in GTEx RNA-seq dataset are missing due to lack of measurements for random tissues in some donors. Our low-tensor-rank Bayesian hierarchical model can directly handle the missing observations without performing data imputation. In particular, when some y_{dtg}'s are missing, the corresponding likelihood function and latent model will have similar forms as in (3) and (4) but restrict (d, t, g) in Ω, where Ω represents the non-missing index subset. We conduct an additional experiment to demonstrate the performance of our proposed BayMC when there exists a proportion of missing values. More precisely, from the data generated in the overlapped case, we randomly delete the gene expression of 50% tissues among all donors to mimic the missing mechanism of the GTEx data. In this additional setting, the BayMC method could correctly identify the number of clusters on 70% of the 50 replicates. The results of the average estimated membership matrices are depicted in the third row of Figure 1. As a comparison, we also plot the true membership matrices (the first row) and the average estimation from the complete observations (the second row) in Figure 1. Figure 1 shows BayMC has similar performance for missing and complete observations, and visibly close to the truth.

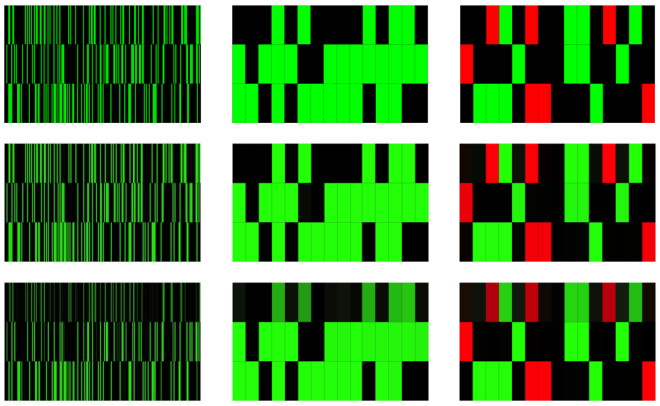
5. REAL DATA

We apply the proposed BayMC method to the GTEx v6 gene expression data[1] and compare with two tensor-based alternatives, HLloyd and MultiCluster. The data consist of RNA-seq samples collected from 544 individuals across 53 human tissues. For each donor, some clinical information like gender and age is also recorded.

5.1 Data preprocessing

Since the raw data are the gene read counts of RNA-seq, a preprocessing procedure is needed before the clustering methods are applied.

[1]Available at https://www.gtexportal.org/home/datasets.



(a) True C_1 vs. Estimated C_1. (b) True C_2 vs. Estimated C_2. (c) True C_3 vs. Estimated C_3.

Figure 1. The average simulation results based on 50 replicates in the overlapped case for BayMC. The green, black, and red cells represent 1, 0, *and* -1, *respectively. The first row: plots of the true values of* $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, *and* $\boldsymbol{C}_3$. *The second row: plots of the average estimations of* $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, *and* $\boldsymbol{C}_3$ *with the complete observations. The third row: plots of the average estimations of* $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, *and* $\boldsymbol{C}_3$ *with* 50% *proportion of missing observations.*

Quality control. The first step of preprocessing data is to normalize the raw data to eliminate the systematic variation and guarantee the expression levels are comparable between genes and samples. Similar to [48], we use estimateSizeFactor (in R package DESeq2), which takes into account both sequencing depth and RNA composition, to achieve normalization. We then take a logarithm of the processed data.

Subset selection. To derive an interpretable clustering results, we lay our interest on the brain subset of GTEx-RNA data. It is well-known that depression (major depressive disorder or clinical depression) is a common but serious mood disorder [49, 51]. It causes severe symptoms that affect how people feel, think, and handle daily activities, such as sleeping, eating, and working [4, 26, 30]. We are particularly curious about genes related to depression and how they take effect. We thus focus on the 13 brain tissues and 15 depression-related genes. These genes are selected according to Atlas of the Developing Human Brain (the corresponding database is available at http://www.brainspan.org/ish) and listed in Table 2. There are 193 donors related to the brain tissues. Therefore, we get a 193 (donors) $\times$ 13 (tissues) $\times$ 15 (genes) tensor for analysis.

We run the MCMC algorithm for 10,000 iterations with one random initial cluster. The first 5,000 iterations are discarded as burn-in and posterior samples are retained every 10th iteration after burn-in. We summarize R, $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$ using the same procedure as described in Section 3.3. To apply HLloyd and MultiCluster, note, again, the number of clusters need to be pre-specified. We plug in 6 as the number of clusters for these two methods according to the suggestion in [47]. Also, the missing values in the original data are imputated by k nearest neighbor (KNN) for HLloyd and MultiCluster as elaborated in [47].

5.2 Results

For the proposed BayMC method, we depict the trace plot of the number of clusters over the course of MCMC in Figure 2, which indicates 3 clusters. The point estimates of the membership matrices of donors, tissues, genes are depicted from left to right in Figure 3, respectively. The obtained clustering results can be interpreted by the spatial information in brain for tissues, the gene ontology (GO) enrichment analyses among both the overexpressed (+1 in $\boldsymbol{C}_3$) and underexpressed (-1 in $\boldsymbol{C}_3$) genes, and the hypergeometric test on donors [40, 3].

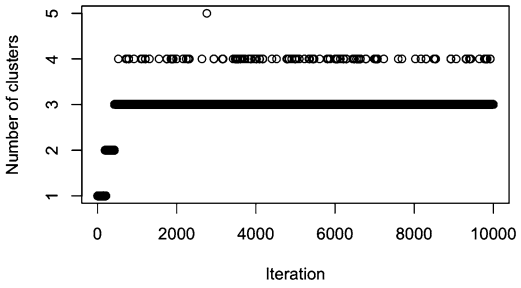


Figure 2. The trace plot of the number of clusters for the proposed BayMC method.

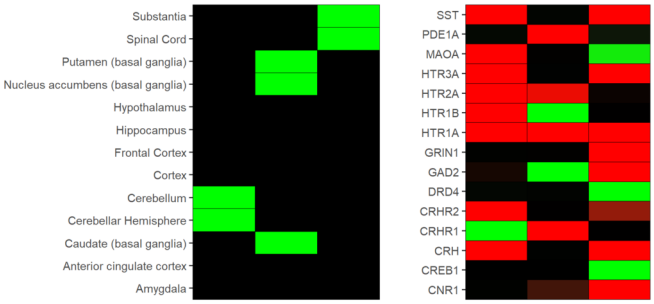


Figure 3. From left to right are the estimated membership matrices of tissues and genes using BayMC. The green, black, and red cells represent 1, 0, *and* -1, *respectively. The membership matrix of donors is presented in Section S.3 of the Supplementary Material.*

In particular, the first cluster can be treated as a cerebellum related cluster, since its tissue membership only contains cerebellar hemisphere and cerebellum [44]. Genes underexpressed in this cluster are highly correlated to regulation of serotonin secretion ($p = 3.67 \times 10^{-9}$ under Bonferroni correction), which is consistent with the scientific finding in [39, 24, 37, 36]. We also summarize the age and gender effect of donors on the first cluster by conducting hypergeometric test. The results show that females are en-

Table 2. The depression-related genes according to Atlas of the Developing Human Brain

NAME	DESCRIPTION
CNR1	cannabinoid receptor 1 (brain)
CREB1	cAMP responsive element binding protein 1
CRH	corticotropin releasing hormone
CRHR1	corticotropin releasing hormone receptor 1
CRHR2	corticotropin releasing hormone receptor 2
DRD4	dopamine receptor D4
GAD2	glutamate decarboxylase 2 (pancreatic islets and brain, 65kDa)
GRIN1	glutamate receptor, ionotropic, N-methyl D-aspartate 1
HTR1A	5-hydroxytryptamine (serotonin) receptor 1A, G protein-coupled
HTR1B	5-hydroxytryptamine (serotonin) receptor 1B, G protein-coupled
HTR2A	5-hydroxytryptamine (serotonin) receptor 2A, G protein-coupled
HTR3A	5-hydroxytryptamine (serotonin) receptor 3A, ionotropic
MAOA	monoamine oxidase A
PDE1A	phosphodiesterase 1A, calmodulin-dependent
SST	omatostatin

riched in this cluster ($p = 0.076$), confirmed by the research of [35], which may be a factor relevant to the lower incidence of major unipolar depression in males. The second cluster is a basal ganglia related cluster, since its tissue membership contains three basal ganglia subtissues: caudate, putamen, and nucleus accumbens [44]. The underexpressed genes in this cluster are highly correlated to regulation of amine transport ($p = 5.67 \times 10^{-4}$ under Bonferroni correction), whereas the overexpressed genes in this cluster are highly related to neurotransmitter secretion ($p = 2.96 \times 10^{-2}$ under Bonferroni correction). These results are confirmed by [16], which shows that basal ganglia have been found to contain remarkably high levels of many of the neurotransmitters. Hypergeometric test shows no significant effects of age and gender on this cluster. Comparing the results on the first and second clusters, the proposed BayMC method distinguishes abnormally high and low expressions of some genes across these clusters, which may be interpreted through the identified subjects and tissues in each cluster. For example, BayMC finds that gene HTR1B has low expression in the first cluster and high expression in the second cluster. This finding that HTR1B has different gene expression levels between cerebellum and basal ganglia related tissues is consistent with existing biological knowledge [13, 14]. The third cluster consists of spinal cord and substantia nigra [1]. The underexpressed genes in this cluster are highly correlated to anterograde trans-synaptic signaling ($p = 8.37 \times 10^{-7}$ under Bonferroni correction), whereas the overexpressed genes in this cluster are highly related to dopamine metabolic process ($p = 1.196 \times 10^{-2}$ under Bonferroni correction). We can find similar results in [42], which shows the key role of dopamine in spinal cord. Hypergeometric test shows females ($p = 0.017$) are enriched in this cluster, which may suggest a gender effect on this cluster [6].

For the results of applying HLloyd and MultiCluster methods on GTEx RNA-seq dataset, we depict their estimated membership matrices in Figures 4 and 5, respectively. These figures show that three of their clusters with respect to tissues are consistent with ours among their obtained 6 clusters. However, neither method allows overlaps, which may not be suitable for gene expression analysis since genes are known to function through multiple pathways. Moreover, the Tucker decomposition used for HLloyd lacks interplay information among the three modes of tensor. In other words, they can not associate the clustering information of identified genes with that of the brain subregions.

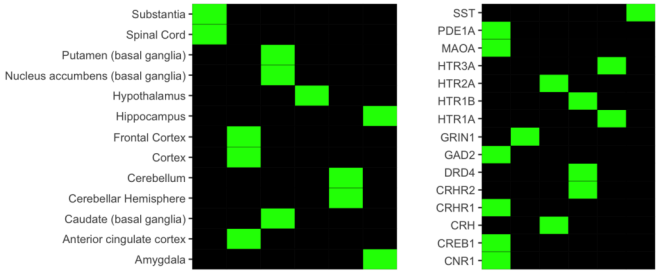


Figure 4. From left to right are the estimated membership matrices of tissues and genes using HLloyd. The green and black cells represent 1 *and* 0*, respectively.*

To further compare the proposed BayMC with HLloyd and MultiCluster, we check the model fit adequacy (measure of "lack-of-fit") for various methods. More precisely, we calculate the relative errors and correlations between the observed tensor measurements $\boldsymbol{Y}$ and the estimated tensor $\widehat{\boldsymbol{Y}}$ from three methods. The results are summarized in Table 3. It shows that the in-sample error is 7.4% and in-sample correlation between fitted value and observation is 96% for BayMC, which outperforms the other two methods, especially HLloyd. This result is consistent with our simulation study that we can benefit from adaptive learning on the latent indicator tensor, interactive multi-way clustering model, and allowing overlaps in clusters.

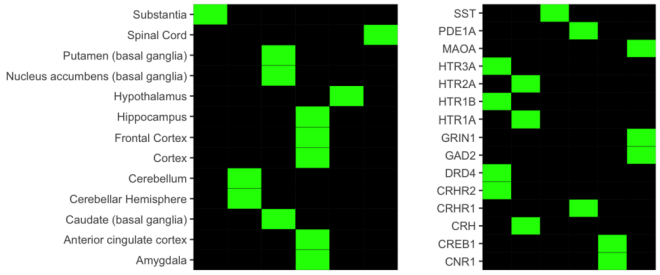


Figure 5. From left to right are the estimated membership matrices of tissues and genes using MultiCluster. The green and black cells represent 1 *and* 0*, respectively.*

Table 3. The model fit adequacy of GTEx RNA-seq data

	BayMC	HLloyd	MultiCluster
In-sample error	7.4%	16.6%	7.6%
In-sample correlation	96%	90%	95.6%

6. DISCUSSION

In this paper, we have proposed a novel identifiable multiway clustering approach for higher-order tensor data. With the form of CP decomposition, our model can fully explore the tensor structure, cluster all the modes simultaneously, and characterize the interaction among the modes. Using Bayesian hierarchical model and a nonparametric Bayesian prior, our approach can also automatically determine the number of clusters from the posterior samples and allow overlapping clusters. Applying the proposed method on GTEx RNA-seq data, we discovered three gene expression modules within brain region, which may further assist in uncovering disease mechanism.

The proposed BayMC is also conceptually related to community detection based on high-order Stochastic Block Models (SBMs), such as [34, 15, 22], for which tensor decomposition on the latent features and clustering on hypergraphs are often implemented. Our model is different from the approach of SBM in several folds. First, SBM requires the binary/ternary tensor, which represents the relationships between entities, to be observed; whereas our model treats the ternary tensor as latency and learns it adaptively from the real-valued data. Second, to implement SBM, at least two modes of the tensor should have the identical number of dimensions since these two modes of the tensor both represent all entities. Finally, SBM needs an additional step to achieve clustering analysis.

There are some directions for future work to generalize our model. First, the current posterior inference is based on MCMC, and the computation is quite expensive for data of large scale. We may explore faster algorithms such as consensus Monte Carlo algorithms [33] and variational inference [7]. Second, extending our model to multi-omics data by integrating multi-omics information, for example, combining DNA methylation and protein abundance data [45], is also of interest.

ACKNOWLEDGEMENTS

He's research was supported by the NSFC No.11801560 and by Public Computing Cloud, Renmin University of China. Ni's research was supported by NSF DMS-2112943 and NIH 1R01GM148974-01.

Received 27 September 2022

REFERENCES

[1] Aldred, E. M. (2009). *Pharmacology: A handbook for complementary healthcare professionals*. Elsevier, Amsterdam, Netherlands.

[2] Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S. and Mooney, R. J. (2005). Model-based overlapping clustering. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* 532–537.

[3] Beal, S. (1976). Fisher's hypergeometric test for a comparison in a finite population. *The American Statistician* **30** 165–168.

[4] Beck, A. T. and Greenberg, R. L. (1979). *Coping with depression*. Institute for Rational Living, New York.

[5] Bergmann, S., Ihmels, J. and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E* **67** 031902.

[6] Beyer, C., Pilgrim, C. and Reisert, I. (1991). Dopamine content and metabolism in mesencephalic and diencephalic cell cultures: Sex differences and effects of sex steroids. *Journal of Neuroscience* **11** 1325–1333.

[7] Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112** 859–877. MR3671776

[8] Braun, T., Bober, E., Winter, B., Rosenthal, N. and Arnold, H. (1990). Myf-6, a new member of the human gene family of myogenic determination factors: Evidence for a gene cluster on chromosome 12. *The EMBO Journal* **9** 821–831.

[9] Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35** 283–319.

[10] Cheng, Y. and Church, G. M. (2000). Biclustering of Expression Data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology* **8** 93–103.

[11] Chi, E. C., Gaines, B. R., Sun, W. W., Zhou, H. and Yang, J. (2020). Provable convex co-clustering of tensors. *Journal of Machine Learning Research* **21** 8792–8849. MR4209500

[12] De Souto, M. C., Costa, I. G., De Araujo, D. S., Ludermir, T. B. and Schliep, A. (2008). Clustering cancer gene expression data: A comparative study. *BMC Bioinformatics* **9** 1–14.

[13] Duan, J., Sanders, A., Vander Molen, J., Martinolich, L., Mowry, B., Levinson, D., Crowe, R., Silverman, J. and Gejman, P. (2003). Polymorphisms in the 5'-untranslated region of the human serotonin receptor 1B (HTR1B) gene affect gene expression. *Molecular Psychiatry* **8** 901–910.

[14] Fujita, T., Aoki, N., Mori, C., Fujita, E., Matsushima, T., Homma, K. J. and Yamaguchi, S. (2020). The dorsal arcopallium of chicks displays the expression of orthologs of mammalian fear related serotonin receptor subfamily genes. *Scientific Reports* **10** 1–16.

[15] Ghoshdastidar, D. and Dukkipati, A. (2014). Consistency of spectral partitioning of uniform hypergraphs under planted partition model. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* **1** 397–405. MR3670495

[16] Graybiel, A. M. (1990). Neurotransmitters and neuromodulators in the basal ganglia. *Trends in Neurosciences* **13** 244–254.
[17] Griffiths, T. L. and Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. In *Proceedings of the 18th International Conference on Neural Information Processing Systems* 475–482.
[18] Han, R., Luo, Y., Wang, M. and Zhang, A. R. (2020). Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*. MR4382029
[19] Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics* **16** 1–84.
[20] Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association* **67** 123–129. MR0405726
[21] Jiang, D., Tang, C. and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering* **16** 1370–1386.
[22] Ke, Z. T., Shi, F. and Xia, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*.
[23] Kiers, H. A. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics* **14** 105–122.
[24] Kish, S. J., Furukawa, Y., Chang, L.-J., Tong, J., Ginovart, N., Wilson, A., Houle, S. and Meyer, J. H. (2005). Regional distribution of serotonin transporter protein in postmortem human brain: Is the cerebellum a SERT-free brain region? *Nuclear Medicine and Biology* **32** 123–128.
[25] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review* **51** 455–500. MR2535056
[26] Kraft, U. (2006). Burned out. *Scientific American Mind* **17** 28–33.
[27] Kruskal, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications* **18** 95–138. MR0444690
[28] Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica* **12** 61–86. MR1894189
[29] Lee, M., Shen, H., Huang, J. Z. and Marron, J. S. (2010). Biclustering via sparse singular value decomposition. *Biometrics* **66** 1087–1095. MR2758496
[30] Lewinsohn, P. M., Muñoz, R. F., Youngren, M. A. and Zeiss, A. M. (1986). *Control your depression (Rev'd Ed)*. Prentice-Hall, New York.
[31] Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., Consortium, T. G., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G. and Guigó, R. (2015). The human transcriptome across tissues and individuals. *Science* **348** 660–665.
[32] Ni, Y., Müller, P. and Ji, Y. (2020). Bayesian double feature allocation for phenotyping with electronic health records. *Journal of the American Statistical Association* **115** 1620–1634. MR4189742
[33] Ni, Y., Müller, P., Diesendruck, M., Williamson, S., Zhu, Y. and Ji, Y. (2020). Scalable Bayesian nonparametric clustering and classification. *Journal of Computational and Graphical Statistics* **29** 53–65. MR4085863
[34] Nickel, M., Tresp, V. and Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* 809–816.
[35] Nishizawa, S., Benkelfat, C., Young, S., Leyton, M., Mzengeza, S. d., De Montigny, C., Blier, P. and Diksic, M. (1997). Differences between males and females in rates of serotonin synthesis in human brain. *Proceedings of the National Academy of Sciences* **94** 5308–5313.
[36] Oostland, M. and Hooft, J. A. v. (2016). Serotonin in the cerebellum. In *Essentials of Cerebellum and Cerebellar Disorders* 243–247. Springer, Cham.
[37] Oostland, M. and Van Hooft, J. (2013). The role of serotonin in cerebellar development. *Neuroscience* **248** 201–212.
[38] Parmigiani, G., Garrett, E. S., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 717–736. MR1979385
[39] Pazos, A. and Palacios, J. (1985). Quantitative autoradiographic mapping of serotonin receptors in the rat brain. I. Serotonin-1 receptors. *Brain Research* **346** 205–230.
[40] Rivals, I., Personnaz, L., Taing, L. and Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23** 401–407.
[41] Schnell, P. M., Tang, Q., Offen, W. W. and Carlin, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* **72** 1026–1036. MR3591587
[42] Sharples, S. A., Koblinger, K., Humphreys, J. M. and Whelan, P. J. (2014). Dopamine: A parallel pathway for the modulation of spinal locomotor networks. *Frontiers in Neural Circuits* **8** 55.
[43] Smilde, A., Bro, R. and Geladi, P. (2005). *Multi-way analysis: Applications in the chemical sciences*. John Wiley & Sons, West Sussex, UK.
[44] Standring, S. (2008). *Gray's anatomy: The anatomical basis of clinical practice*. Churchill Livingstone, St Louis, MO.
[45] Stranger, B. E., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W. F., Lonsdale, J. T., McDonald, A. et al. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease The eGTEx Project. *Nature Genetics* **49** 1664–1670.
[46] Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** 279–311. MR0205395
[47] Wang, M., Fischer, J. and Song, Y. S. (2019). Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *The Annals of Applied Statistics* **13** 1103–1127. MR3963564
[48] Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* 715–725.
[49] Wise, T., Radua, J., Via, E., Cardoner, N., Abe, O., Adams, T. M., Amico, F., Cheng, Y., Cole, J., de Azevedo Marques Périco, C. et al. (2017). Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: Evidence from voxel-based meta-analysis. *Molecular Psychiatry* **22** 1455–1463.
[50] Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A. and Staudt, L. M. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences* **100** 9991–9996.
[51] Yang, Y., Liu, S., Jiang, X., Yu, H., Ding, S., Lu, Y., Li, W., Zhang, H., Liu, B., Cui, Y. et al. (2019). Common and specific functional activity features in schizophrenia, major depressive disorder, and bipolar disorder. *Frontiers in Psychiatry* **10** 52.
[52] Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics* **17** 309–318.
[53] Zhou, F., He, K., Li, Q., Chapkin, R. S. and Ni, Y. (2022a). Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization. *Biostatistics* **23** 891–909. MR4454961
[54] Zhou, F., He, K., Cai, J. J., Davidson, L. A., Chapkin, R. S. and Ni, Y. (2022b). A unified Bayesian framework for bi-overlapping-clustering multi-omics data via sparse matrix factorization. *Statistics in Biosciences* **To appear**. MR4454961

Zhuofan Wang
The Center for Applied Statistics
Institute of Statistics and Big Data
Renmin University of China
Beijing 100872
China
E-mail address: zf.wang@ruc.edu.cn

Fangting Zhou
The Center for Applied Statistics
Institute of Statistics and Big Data
Renmin University of China
Beijing 100872
China
Department of Statistics
Texas A&M University
College Station TX 77843
USA
E-mail address: 2017000834@ruc.edu.cn

Kejun He
The Center for Applied Statistics
Institute of Statistics and Big Data
Renmin University of China
Beijing 100872
China
E-mail address: kejunhe@ruc.edu.cn

Yang Ni
Department of Statistics
Texas A&M University
College Station TX 77843
USA
E-mail address: yni@stat.tamu.edu