Pacific Symposium on Biocomputing 25:427-438(2020)

Bayesian semi-nonnegative matrix tri-factorization to identify pathways associated with cancer phenotypes

Sunho Park[1], Nabhonil Kar[1], Jae-Ho Cheong[2] and Tae Hyun Hwang[1,*]

[1] *Quantitative Health Sciences, Cleveland Clinic, 9500 Euclid Ave., Cleveland, OH 44195*
[2] *Department of Biomedical Systems Informatics Yonsei University College of Medicine, 250 Seongsanno Seodaemun-gu Seoul, 120-752 Korea*
[*] *Email: hwangt@ccf.org*

Accurate identification of pathways associated with cancer phenotypes (e.g., cancer subtypes and treatment outcomes) could lead to discovering reliable prognostic and/or predictive biomarkers for better patients stratification and treatment guidance. In our previous work, we have shown that non-negative matrix tri-factorization (NMTF) can be successfully applied to identify pathways associated with specific cancer types or disease classes as a prognostic and predictive biomarker. However, one key limitation of non-negative factorization methods, including various non-negative bi-factorization methods, is their limited ability to handle negative input data. For example, many types of molecular data that consist of real-values containing both positive and negative values (e.g., normalized/log transformed gene expression data where negative values represent down-regulated expression of genes) are not suitable input for these algorithms. In addition, most previous methods provide just a single point estimate and hence cannot deal with uncertainty effectively.

To address these limitations, we propose a Bayesian semi-nonnegative matrix tri-factorization method to identify pathways associated with cancer phenotypes from a real-valued input matrix, e.g., gene expression values. Motivated by semi-nonnegative factorization, we allow one of the factor matrices, the centroid matrix, to be real-valued so that each centroid can express either the up- or down-regulation of the member genes in a pathway. In addition, we place structured spike-and-slab priors (which are encoded with the pathways and a gene-gene interaction (GGI) network) on the centroid matrix so that even a set of genes that is not initially contained in the pathways (due to the incompleteness of the current pathway database) can be involved in the factorization in a stochastic way specifically, if those genes are connected to the member genes of the pathways on the GGI network. We also present update rules for the posterior distributions in the framework of variational inference. As a full Bayesian method, our proposed method has several advantages over the current NMTF methods, which are demonstrated using synthetic datasets in experiments. Using the The Cancer Genome Atlas (TCGA) gastric cancer and metastatic gastric cancer immunotherapy clinical-trial datasets, we show that our method could identify biologically and clinically relevant pathways associated with the molecular subtypes and immunotherapy response, respectively. Finally, we show that those pathways identified by the proposed method could be used as prognostic biomarkers to stratify patients with distinct survival outcome in two independent validation datasets. Additional information and codes can be found at https://github.com/parks-cs-ccf/BayesianSNMTF.

Keywords: Phenotype-pathway association, Bayesian learning, Semi-non-negative tri-matrix factorization, Structured slab-and-spike distribution, Variational inference

© 2019 The Authors. Open Access chapter published by World Scientific Publishing Company and

1. Introduction

Accurate identification of pathways associated with cancer phenotypes (e.g., cancer subtypes and treatment outcomes) enables us to understand better molecular biology processes in cancer and could lead to discovering reliable prognostic and/or predictive biomarkers for better patients stratification and treatment guidance. Non-negative matrix tri-factorization (NMTF) models can provide an intuitive and efficient way to identify associations between two different entities by simultaneously clustering rows and columns of the data matrix.[1] In our previous work[2] (referred to as NTriPath), we use NMTF to identify pathways associated with cancer types from mutation data: the mutation data matrix is decomposed into the cancer-type indicator matrix, the association matrix between cancer types and pathways, and the centroid matrix (each centroid corresponds to the pattern of gene mutations within each pathway). Pathway membership information, e.g., gene-pathway annotations from Kegg pathway database, and a gene-gene interaction (GGI) network are incorporated into the factorization model through the framework of regularized optimization. It is shown from the The Cancer Genome Atlas (TCGA) data that the top pathways ranked by the method are closely related to clinical outcomes.[2] However, this approach has several limitations. First, the input matrix is restricted to be non-negative and hence cannot readily model many types of genomic data, including copy number alteration and normalized/log transformed gene expressions, which are real-valued. Second, the method provides just a single point estimate of the model's parameters and thus cannot deal with uncertainty well. Moreover, it involves many hyper-parameters, e.g., regularization constants, which should be tuned carefully. However, since the association identification from the input (mutation) matrix is clearly an unsupervised problem, i.e., there is no corresponding output for the input matrix, it is not clear how to find the optimal hyper-parameter values for the given input data.

To address the aforementioned limitations of NTriPath, we propose a novel Bayesian semi-nonnegative matrix factorization model, where the biological prior knowledge represented by a pathway database and a GGI network is incorporated into the factorization through structured spike-and-slab sparse priors.[3] First, in order to handle real-valued input data, e.g., gene expression values, we allow one of the latent (factor) matrices, the centroid matrix, to have positive and negative values so that each centroid (corresponding to a pathway) can express the up-regulation or the down-regulations of the member genes in the pathway. Second, we encode pathway membership information and a GGI network into the factorization model through the framework of Bayesian learning. Specifically, we model the priors over the centroid matrix matrix using the structured spike-and-slab distributions, where our prior knowledge of the sparsity pattern is encoded into the prior distributions thorough underlying Gaussian processes (GPs).[3] To conclude the prior modeling for the centroid matrix, we define the mean vectors and covaraince matrices of the GPs using the pathway membership information and the GGI network. As a result, even non-member genes of the pathways can be involved in the factorization in a stochastic manner. Note that our method is a full Bayesian approach: priors

distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

are placed on the model's parameters (the latent matrices) and hyper-parameters (e.g., the noise precision) and are updated by observations (resulting in the posteriors). Thus, in contrast to NTriPath, which relies on only the single most probable setting of the model's parameters and hyper-parameters (regularization constants), our method produces more robust factorization results by averaging over all possible settings. Finally, we propose the update rules for the posterior distributions by utilizing the framework of variational inference. Using experiments on synthetic datasets, we show the superiority of our proposed method over NTriPath (where a folding approach[4] is used to deal with negative values in the input matrix). Using TCGA gastric cancer and metastatic gastric cancer immunotherapy clinical-trial datasets,[5] we show that the proposed method could identify biologically and clinically-relevant pathways associated with TCGA gastric cancer molecular subtypes and immunotherapy response. Finally we show that those pathways identified by our method could be used as prognostic biomarkers to stratify patients with distinct survival outcome in two independent validation datasets.

Notations: For a matrix $\boldsymbol{A}$, $\boldsymbol{a}_i$ represents its ith row vector, i.e., $(\boldsymbol{A}_{i,:})^\top$. Similarly, $\vec{\boldsymbol{a}}_j \triangleq \boldsymbol{A}_{:,j}$ refers to its jth column vector. The (i,j)th element of the matrix $\boldsymbol{A}$ is expressed by A_{ij}.

2. Background

Non-negative matrix factorization (NMF), which here refers to the matrix bi-factorization (decomposing a matrix into two smaller matrices), has been applied to many different biological problems as a tool for clustering, dimensionality reduction and visualization (please see references herein[6]). It provides a parts-based local representation, making NMF unique compare to other linear dimensionality reduction methods such as principal component analysis (PCA). However, NMF is limited to non-negative input data. When the input matrix contains positive and negative values, a natural way is to decompose the input matrix into a centroid matrix (assumed to be real-valued) and a cluster membership indicator matrix (assumed to be non-negative). This approach is the main motivation of semi-nonnegative factorization,[7] and we use this same idea to allow our method to find patterns from real-valued input data.

The spike-and-slab prior is the standard approach for sparse learning, which is the selection of a subset of features from high-dimensional input data. It can be expressed as a mixture of a point mass at 0 (spike) and a continuous distribution (slab):

$$\overline{V}_{ij} \sim \rho_{ij}\mathcal{N}(\overline{V}_{ij}|0,\sigma^{\overline{V}}_{jr}) + (1-\rho_{ij})\delta_0(\overline{V}_{ij}) \tag{1}$$

where $\mathcal{N}(\cdot)$ is a Gaussian distribution, $\rho_{ij} \in [0,1]$ is a mixing coefficient, and $\delta_0(\cdot)$ is Dirac delta funciton, i.e., $\delta_0(\overline{V}_{ij}) = 1$ at $\overline{V}_{ij} = 0$, and 0 elsewhere. The mixture structure of the spike-and-slab prior can produce a bi-separation effect where the posterior distributions over the coefficients for *irrelevant* features are peaked at zero while those over the coefficients of *relevant* features have a large probability of being non-zero. The spike-and-slab prior (1) can be equivalently rewritten with a binary variable, and the posterior mean of this binary variable indicates how the corresponding coefficient is actually different from zero.

3. Bayesian Semi-Nonnegative Tri-Matrix Factorization (Bayesian SNTMF)

We propose a Bayesian method to identify associations between cancer phenotypes (e.g., molecular subtypes) and pathways from human cancer genomic data. In this work, we consider

only gene expression data, but our method can be applied to other data types that can be formed into real-valued matrices, e.g., copy number and miRNA expression. We develop a semi-nonnegative matrix tri-factorization method in the framework of Bayesian learning, where the prior knowledge represented by a pathway membership information and a GGI network is taken into account in the factorization through structured spike and slab prior distributions.[3]

3.1. *Model formulation*

We assume that observations are given in the form of a matrix $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ where X_{ij} represents the ith patient's expression value for the jth gene, and N and D are the number of samples and genes, respectively. We assume that pathway information is also given in a form of a matrix $\boldsymbol{Z}^0 \in \mathbb{R}^{D \times R}$, where R is the number of the pathways and each element represents the membership of a gene to a pathway, i.e., $Z^0_{jr} = 1$ if the jth gene is a member of the rth pathway, and $Z^0_{jr} = 0$ otherwise. Our main objective is to approximate $\boldsymbol{X}$ as a product of three latent matrices added with residuals $\boldsymbol{E} \in \mathbb{R}^{N \times D}$:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\overline{\boldsymbol{V}}^{\top} + \boldsymbol{E} \tag{2}$$

where $\boldsymbol{U} \in \mathbb{R}_+^{N \times K}$, $\boldsymbol{S} \in \mathbb{R}_+^{K \times R}$, $\overline{\boldsymbol{V}} \in \mathbb{R}^{D \times R}$ and K is the number of the subtypes. We assume that the matrix $\boldsymbol{U}$ is constructed from patient clinical data: K is the number of subtypes we are interested in, and $U_{ij} = 1$ indicates that the ith patient is of the jth subtype (1-of-K encoding, i.e., $U_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} U_{ik} = 1$). The real-valued matrix $\overline{\boldsymbol{V}}$ consists of R basis vectors, and its rth column is a pattern associated with a corresponding pathway: only few elements (corresponding to the member genes of a pathway, i.e., $\{j | Z^0_{jr} = 1\}$) would have non-zero values, representing either over-expression ($\overline{V}_{jr} > 0$) or under-expression ($\overline{V}_{jr} < 0$), and all other elements are set to zero. Then, the non-negative matrix $\boldsymbol{S}$ encodes associations between the subtypes and the pathways, where each element S_{ij} represents the association between the ith subtype and the jth pathway. Once $\boldsymbol{S}$ is learned, we can easily identify pathways related to a certain subtype by selecting the top pathways that have the largest values in the corresponding row in $\boldsymbol{S}$. As all the latent variables are learned in the Bayesian learning framework, the likelihood of the model and the prior distribution over the latent variables are defined according to our model assumptions.

Assuming the residuals E_{ij} in eq. (2) to be sampled from i.i.d. Gaussian distributions with mean zero and precision γ, we can specify the likelihood of the factorization model:

$$X_{ij} \sim \mathcal{N}(X_{ij} | \boldsymbol{u}_i^{\top} \boldsymbol{S} \overline{\boldsymbol{v}}_j^{\top}, \gamma^{-1}), \tag{3}$$

where the precision γ (the inverse of the variance) is sampled from a Gamma distribution.

The following discusses how we define the priors over the latent variables. For $\boldsymbol{S}$, each element is assumed to be sampled from an Exponential distribution to ensure its non-negativity:

$$S_{kr} \sim \text{Exponential}(S_{kr} | \lambda^{S0}_{kr}). \tag{4}$$

For $\overline{\boldsymbol{V}}$, the simplest inference approach would be to calculate the posterior distributions (with Gaussian distribution priors) over only the elements in the matrix that are corresponding to the member genes in the pathways, i.e., $\mathcal{M} \triangleq \{(j, r) | Z^0_{jr} = 1\}$, and leave the other elements as zero. However, it is widely accepted that pathway databases are not complete, that there

are unknown missing genes in a pathway. To include unknown missing member genes in the pathways into the factorization, we use the concept of sparse learning, where sparse prior distributions (e.g., spike-and-slab or Laplace distributions) are placed over all the elements of $\overline{\boldsymbol{V}}$ and only few elements (including those in the set $\mathcal{M}$) are encouraged to have non-zero values. We make use of a gene-gene interaction network as well as of the pathway information $\boldsymbol{Z}^0$ to determine the the positions of non-zero elements in $\overline{\boldsymbol{V}}$ based on the assumption that two connected genes in the graph would more likely to be active together in a pathway. Denote a gene-gene interaction network by $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, where $A_{jj'} = 1$ if genes j and j' are connected on the network, and $A_{jj'} = 0$ otherwise, and assume that there is no self connection, i.e., $A_{jj} = 0$. We then will show that the priors incorporating $\boldsymbol{Z}^0$ and $\boldsymbol{A}$ can be defined using the structured spike and slab prior model[3] which imposes spatial constraints on spike-and-slab probabilities through a Gaussian process (GP). We define a GP for each pathway and encode the mean vector and covariance matrix of the GP using our prior knowledge given by $\boldsymbol{Z}^0$ and $\boldsymbol{A}$.

With reparametrization of the variable $\overline{V}_{jr} = V_{jr}Z_{jr}$ (Z_{jr} is assumed to be a binary variable, i.e., $Z_{jr} \in \{0, 1\}$), where $V_{jr} \sim \mathcal{N}(V_{jr}|0, \sigma_{jr}^{V0})$ and $Z_{jr} \sim \text{Bernoulli}(\rho_{jr})$, the spike-and-slab prior over $\overline{V}_{jr}$ in (1) can be equivalently written for the new variables V_{jr} and Z_{jr}:

$$V_{jr}, Z_{jr} \sim \mathcal{N}(V_{jr}Z_{jr}|0, \sigma_{jr}^{V0})\rho_{jr}^{Z_{jr}}(1-\rho_{jr})^{1-Z_{jr}}. \tag{5}$$

We can consider the binary variable Z_{jr} as a on-off switch which determines whether V_{jr} is included into the factorization model. To connect $\boldsymbol{Z}^0$ and $\boldsymbol{A}$ to Z_{jr}, we define the parameter of the Bernoulli distribution ρ_{jr} in the following hierarchical way based on the frame of GP:

$$\rho_{jr} = \Phi(G_{jr}), \tag{6}$$

$$\vec{\boldsymbol{g}}_r|\boldsymbol{Z}^0, \boldsymbol{A} \sim \mathcal{N}(\vec{\boldsymbol{g}}_r|\boldsymbol{m}_r, \boldsymbol{L}) \tag{7}$$

where $\Phi(w_1) = \int_{-\infty}^{w_1} \mathcal{N}(w|0, 1)$ is a cumulative standard Gaussian distribution function and $\vec{\boldsymbol{g}}_r = [G_{1r}, G_{2r}, ..., G_{Dr}]^\top$. Each element of the mean vector $\boldsymbol{m}_r$ is set according to the membership information encoded in $\boldsymbol{Z}^0$: $m_{jr} = \xi_+$ where $\xi_+ > 0$ if $\boldsymbol{Z}_{jr}^0 = 1$, and $m_{jr} = \xi_-$ where $\xi_- < 0$ otherwise (the more negative value ξ_- is, the more sparse prior we get). The covariance matrix $\boldsymbol{L}$ is set to a normalized Laplacian matrix $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D}$ is a diagonal matrix whose ith diagonal element is a summation of the ith row of the matrix $\boldsymbol{A}$. Combining all these assumptions, we can see that if gene i (a nonmember of the rth pathway) has connections to the member genes on the network, then G_{ir} would become high and its on-off binary variable Z_{ir} is more likely to be one. Note that $\overline{\boldsymbol{V}} = \boldsymbol{Z} \circ \boldsymbol{V}$. The binary matrix $\boldsymbol{Z}$ is determined by a stochastic process, and thus the elements in $\boldsymbol{V}$ that even are not in the set $\mathcal{M}$ (originally not in the pathways) can contribute to the factorization model.

As a result, our factorization model can be summarized as follows:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}(\boldsymbol{Z} \circ \boldsymbol{V})^\top + \boldsymbol{E}, \tag{8}$$

$$E_{ij} \sim N(0, \gamma), \quad \forall i, j \tag{9}$$

$$\gamma \sim \text{Gamma}(\gamma|\alpha_a^0, \alpha_b^0), \tag{10}$$

$$S_{kr} \sim \text{Exponential}(S_{kr}|\lambda_{kr}^{S0}), \quad \forall k, r \tag{11}$$

$$V_{jr}, Z_{jr}|G_{ij} \sim \mathcal{N}(V_{jr}Z_{jr}|0, \sigma_{jr}^{V0})\Phi(G_{jr})^{Z_{jr}}(1-\Phi(G_{jr}))^{1-Z_{jr}}, \quad \forall j, r \tag{12}$$

$$\vec{\boldsymbol{g}}_r|\boldsymbol{Z}^0, \boldsymbol{A} \sim \mathcal{N}(\vec{\boldsymbol{g}}_r|\boldsymbol{m}_r, \boldsymbol{L}), \quad \forall r. \tag{13}$$

The conceptual view of our method is depicted in Figure 1.

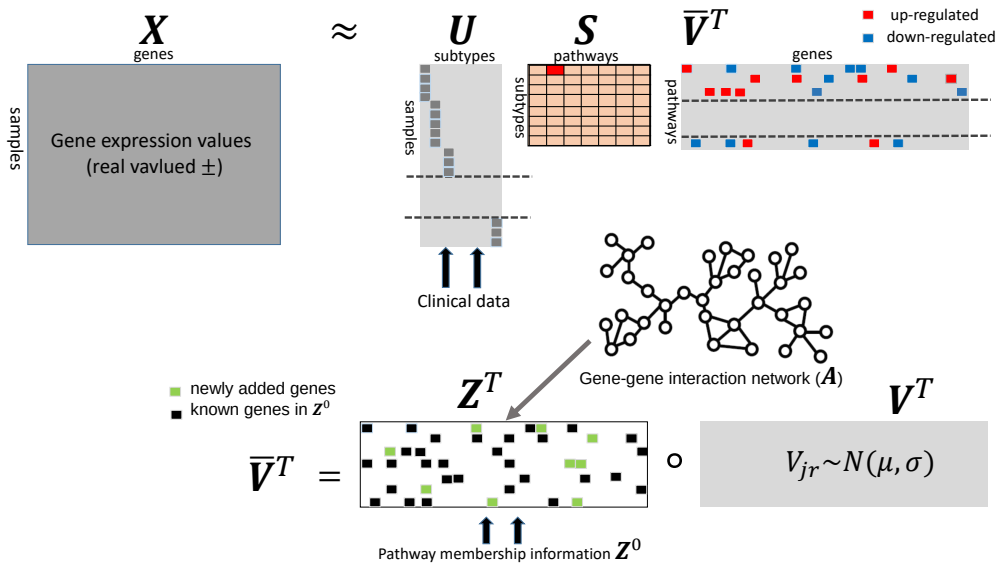


Fig. 1. The input matrix is decomposed into $\boldsymbol{U}$ (samples $\times$ subtypes), $\boldsymbol{S}$ (subtypes $\times$ pathways), and $\overline{\boldsymbol{V}}$ (genes $\times$ pathways). The centroid matrix $\overline{\boldsymbol{V}}$ is further decomposed into the binary indicator matrix $\boldsymbol{Z}$ and the genome-wide pattern matrix $\boldsymbol{V}$. We encode the pathway membership information $\boldsymbol{Z}^0$ and the GGI network $\boldsymbol{A}$ into the binary matrix $\boldsymbol{Z}$ through the structure spike-and-slab priors.

3.2. *Variational inference*

We approximate the posterior distributions over all the latent variables in the variational inference framework as their close form expressions are not available. We assume that the variational distributions are factorized as follows:

$$q(\gamma, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{G}) = q(\gamma)\Big(\prod_{k=1}^{K}\prod_{r=1}^{R} q(S_{kr})\Big)\Big(\prod_{j=1}^{D}\prod_{r=1}^{R} q(V_{jr}, Z_{jr})q(G_{jr})\Big). \tag{14}$$

Note that the elements in the latent matrices ($\boldsymbol{S}$, $\overline{\boldsymbol{V}} = \boldsymbol{Z} \circ \boldsymbol{V}$, and $\boldsymbol{G}$) are assumed to be fully factorized. The form of each variational distribution is assumed to be as follows

$$q(\gamma) = \text{Gamma}(\gamma|\alpha_a, \alpha_b), \tag{15}$$
$$q(S_{kr}) = \mathcal{TN}(S_{kr}|\mu^S_{kr}, \sigma^S_{kr}), \tag{16}$$
$$\begin{aligned} q(V_{jr}, Z_{jr}) &= q(V_{jr}|Z_{jr})q(Z_{jr}) \\ &= \mathcal{N}\Big(V_{jr}|Z_{jr}\mu^V_{jr}, Z_{jr}\sigma^V_{jr} + (1-Z_{jr})\sigma^{V0}_{jr}\Big)\hat{\rho}^{Z_{jr}}_{jr}(1-\hat{\rho}_{jr})^{(1-Z_{jr})}, \end{aligned} \tag{17}$$
$$q(G_{jr}) = \mathcal{N}(G_{jr}|\mu^g_{jr}, \sigma^g_{jr}), \tag{18}$$

where $\mathcal{TN}(s|\mu, \sigma)$ represents a truncated Normal distribution defined on the nonnegative region $s \geq 0$, i.e., $\mathcal{TN}(s|\mu, \sigma) = \frac{\sqrt{1/(2\pi\sigma)}\exp\{-\frac{1}{2\sigma}(s-\mu)^2\}}{1-\Phi(-\mu/\sqrt{\sigma})}$ if $s \geq 0$, and $\mathcal{TN}(s|\mu, \sigma) = 0$ otherwise. Denoting a set of all the latent variables by $\Theta = \{\gamma, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{G}\}$, the variational distribution, $q(\Theta)$, can

be obtained by maximizing the variational lower bound with respect to $q(\Theta)$:[8]

$$\text{maximize}_q \mathcal{L}(q) \triangleq \int q(\Theta) \log \frac{p(\boldsymbol{X}, \Theta)}{q(\Theta)} d\Theta. \tag{19}$$

Note that the variational bound $\mathcal{L}$ is a lower bound on the log-likelihood, i.e., $\log p(\boldsymbol{X}) = \mathcal{L}(q) + \text{KL}(q(\Theta)||p(\Theta|\boldsymbol{X}))$, where the second term in RHS is the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior distribution and always nonnegative. Thus, finding the optimal variational distributions by solving the optimization problem (19) can be easily justified. For each step, we update one variational distribution, fixing the others, and we then proceed to cyclically update all variational distributions in this manner. Based on combining the inference methods for Bayesian non-negative matrix tri-factorziation in[9] and for spike-and-slab prior distributions, the variational distributions $q(\gamma)$, $\{q(S_{kr})\}$ and $\{q(V_{jr}, Z_{jr})\}$, can be updated in closed form. For $\{q(G_{jr})\}$, their means and variances can be updated by any iterative gradient-based optimization methods, e.g., limited-memory BFGS used in our experiments. More detailed derivations are found in our supplementary material available at https://github.com/parks-cs-ccf/BayesianSNMTF.

4. Experimental results

We conduct experiments on both simulation and real-world datasets: 1) using the simulation datasets, we show how our method works and display the superiority of our method over NtriPath (which is a point estimate method); 2) using the two gastric cancer datasets, we demonstrate that the our method can identify biologically and clinically-relevant pathways associated with the molecular subtypes in gastric cancer as well as immunotherapy response and validate these results on independent validation datasets.

We here discuss how to find pathways closely associated with each subtype based on the factorization results from our method, as the final outputs of our method are the variational distributions (the approximate posteriors) over the latent variables, including the association matrix $\boldsymbol{S}$. Specifically, we simply use the posterior mean of each variable as its estimate. We denote the estimate of each latent matrix $\boldsymbol{M}$ by $\widehat{\boldsymbol{M}}$, where each element represents the posterior mean of the corresponding element in the matrix $\boldsymbol{M}$ (please refer to our supplementary material to see how to calculate the mean value of each posterior distribution). For the estimate association matrix $\widehat{\boldsymbol{S}}$, which is always non-negative, we can easily see that the larger $\widehat{S}_{ij}$ is, the stronger association between the ith subtype and the jth pathway. Lastly, we explain how to initialize some variables in our model. For the mean vectors of the GPs ($\boldsymbol{G}$), we set $\xi_+ = 5$ and $\xi_- = -5$ for all the experiments, which means that we assume a strong prior belief on the initial pathway information $\boldsymbol{Z}^0$. However, as we will see from the experiment with simulation datasets, our method is able to recover missing pathway membership. The detailed information on the initialization for our method is included in the supplementary material.

4.1. *Simulation datasets*

With this simple example, we first show how our method works in the case of incomplete pathway membership information. We generate the observation matrix $\boldsymbol{X} \in \mathbb{R}^{300 \times 400}$, where the matrix contains 3 subtypes and each subtype shows a unique pattern, one or two blocks

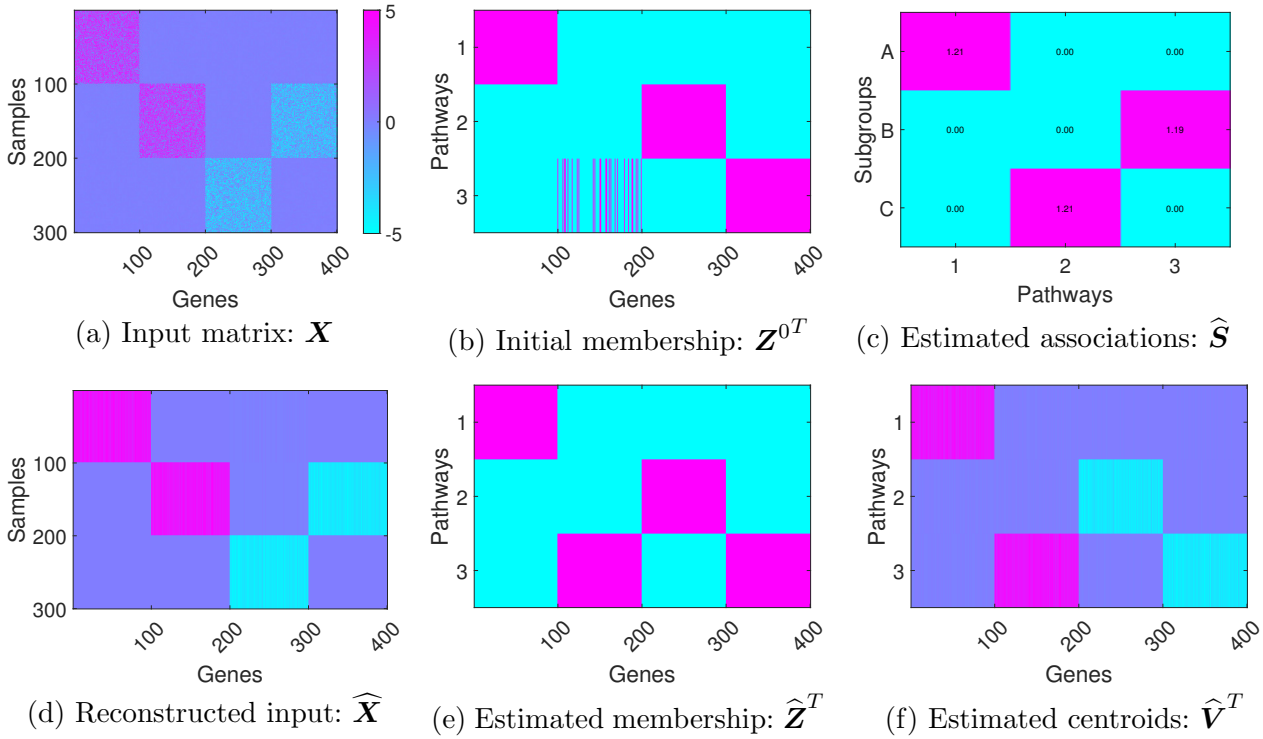
of up- or down-regulated genes in each subtype ($\boldsymbol{X}$ in Figure 2-(a)). Elements in the pattern blocks are drawn from either $\mathcal{N}(2,2)$ for the up-regulation case or $\mathcal{N}(-2,2)$ for the down-regulation case, but elements in the non-pattern blocks are assumed to be background noise and are sampled from $\mathcal{N}(0,0.1^2)$. We construct the subtype indicator matrix $\boldsymbol{U}$ based on our knowledge on the subtype information. We generate a pathway membership matrix $\boldsymbol{Z}_0$ according to the block structure of the input matrix $\boldsymbol{X}$ such that the true associations between the subtypes and the pathways can be easily identifiable ($\boldsymbol{Z}^{0^T}$ in Figure 2 (b)). Note that we assume the pathway membership matrix $\boldsymbol{Z}^0$ incomplete: we randomly remove 80% of member genes from one of the blocks in the 3rd pathway. For the gene-gene interaction network, we randomly connect two genes on the network with probability 0.1.

Figure 2 (c)-(f) shows that our factorization method works well even with the incomplete pathway information. Figure 2 (c) indicates that our method can accurately estimate true associations between subtypes and pathways. For example, the pathway associated with the 2nd subtype (which includes the samples 101 to 200 in the input data) is the 3rd pathway as we designed, and we can easily confirm this association from the estimate association matrix $\widehat{\boldsymbol{S}}$ because only $\widehat{S}_{23}$ has a significantly high value and the others, $\widehat{S}_{21}$ and $\widehat{S}_{22}$, are zero. This result is the same for the other subtypes. we also see that our method can successfully recover the pathway membership information from the data ($\widehat{\boldsymbol{Z}}^\top$ in Figure 2 (e)). This is a promising result considering current pathway databases might be incomplete as our knowledge on molecular biology processes is incomplete. Finally, we can see that our method can correctly find the up/down regulation patterns from the real-valued input data ($\widehat{\boldsymbol{V}}^\top$ in Figure 2 (f)).

We also test our method on an additional simulation dataset to show the superiority of our method over NTriPath. For non-negative factorization methods, one of standard ways to deal with negative values in the input matrix is to fold the matrix by columns:[4] every column will be represented in two new columns in a new matrix, one of which contains only positive values and the other only the magnitudes of negative values. This approach doubles the number of columns in the original matrix and thus causes additional computational burdens, e.g., the GGI network becomes 2^2 times larger. Moreover, it breaks the original patterns in the input matrix because non-negative and negative values are separately processed. In addition, we can see that our method is more robust against noise in general than NTriPath, as Bayesian methods deal with uncertainty more effectively than point estimate methods which rely on a single most probable setting of the model's parameters. Detailed information about this experiment is included in our supplementary material.

4.2. *TCGA gastric cancer and metastatic gastric cancer immunotherapy clinical-trial datasets*

We first identify the top pathways associated with: 1) molecular subtypes in the TCGA gastric cancer (GC) data; and 2) response/non-response in the metastatic gastric cancer (mGC) immunotherapy clinical-trial data.[5] We then validate the pathways identified by our method in both datasets by investigating if these pathways could be used as prognostic biomarkers to stratify patients from two validation datasets, ACRG[10] and MDACC,[11] into groups with distinct survival outcomes.



(a) Input matrix: $\boldsymbol{X}$ (b) Initial membership: $\boldsymbol{Z}^{0T}$ (c) Estimated associations: $\widehat{\boldsymbol{S}}$

(d) Reconstructed input: $\widehat{\boldsymbol{X}}$ (e) Estimated membership: $\widehat{\boldsymbol{Z}}^T$ (f) Estimated centroids: $\widehat{\boldsymbol{V}}^T$

Fig. 2. Factorization results of the simulation data under the assumption that the pathway membership information might be incomplete: multiple member genes in one of the pathways are missed (b). The results indicate that our method can successfully recover the membership information (e).

We provide brief descriptions of the datasets with the notations used in Section 3. For TCGA GC data ($N = 277$), we download the normalized gene expression (mRNA) data[a]. The samples are divided into $K = 4$ groups according to their molecular subtypes: Epstein-Barr virus (EBV), microsatellite instability (MSI), genomically stable (GS), and chromosomal instability (CIN). For the immunotherapy response for mGC data ($N = 45$), we download the gene expression data from,[5] which is normalized by FPKM, and additionally apply log-transformation and standardization. The data includes the patients' treatment outcomes, which are categorized into 4 subtypes: complete response (CR), partial response (PR), progressive disease (PD), and stable disease (SD). In order to find more distinguishable patterns between groups, we here divide the samples into just $K = 2$ groups: responders (CR+PR) and non-responders (PD+SD). Next, we download a GGI network ($\boldsymbol{A}$) from [b] and use $R = 4,620$ sub-networks from[12] to define the pathway membership matrix $\boldsymbol{Z}^0$. After combining all these different data sources, the numbers of the input genes are $D_1 = 14,787$ and $D_2 = 15,347$ for TCGA gastric cancer data and the immunotherapy response data, respectively. The information of both datasets is summarized in Table 1.

After training our factorization model on each dataset, we select the top 3 ranked path-

[a]The data was downloaded from CBioportal (http://www.cbioportal.org/). The downloading option was 'TCGA_stad_rna_seq_v2_mrna' (RNASeq V2 RSEM normalized expression values).
[b]https://thebiogrid.org/. The version is BIOGRID-ORGANISM-Homo_sapiens-3.4.153.

Table 1. Summary of the two datasets, TCGA GC and mGC datasts.

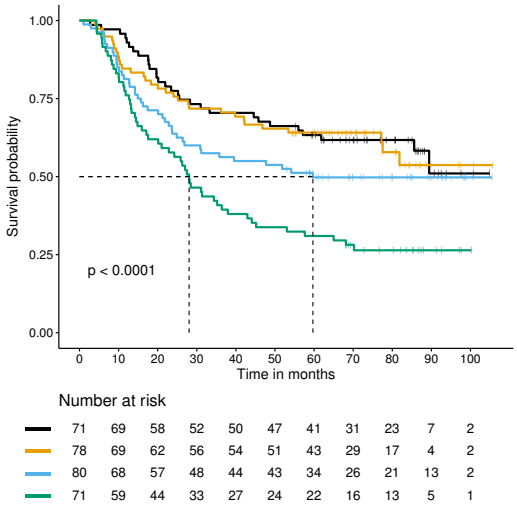
data	N	D	K	phenotypes
TCGA gastric cancer	277	14,787	4	{CIN vs EBV vs GS vs MSI}
Immnotherapy response	45	15,347	2	{responder vs non-responder}

ways for each subtype based on the estimated association matrix $\widehat{\boldsymbol{S}}$ ($12 = 3 \times 4$ pathways consisted of 83 genes are selected for TCGA GC data, and $6 = 3 \times 2$ pathways consisted of 36 genes for the immunotherapy response for mGC data). To assess biological relevance of identified top pathways from TCGA GC and immunotherapy for mGC datasets, we perform gene set enrichment analysis using PANTHER (http://www.pantherdb.org). We find that genes in the pathways identified by our method are enriched with biologically relevant pathways that are associated with cancer phenotypes. For example, 36 genes from mGC immunotherapy response data are enriched with positive regulation of TGFbeta pathway, T-cell migration, etc. Specifically, member genes of 36 gene signatures such as FN1 and FBLN1, involved with TGF-beta regulation are down-regulated and CCL5, CCL21, and CXCL13 which are involved with T-cell migration are up-regulated in response group compared to non-response group, respectively. Activation of TGFbeta pathway serves as a central mechanism to suppress the immune system, thus deactivation of TGFbeta may increase response to immunotherapy.[13] Active T-cell migration into tumor microenvironment could increase response rates to immunotherapy and increase survival.[14] These indicate that our proposed method utilizing real-valued input data could successfully identify down and/or up-regulated pathways that are biologically relevant to and associated with immunotherapy response. It is worth noting that these findings were not reported in the original work.[5] Further details of pathway analysis are available at https://github.com/parks-cs-ccf/BayesianSNMTF.

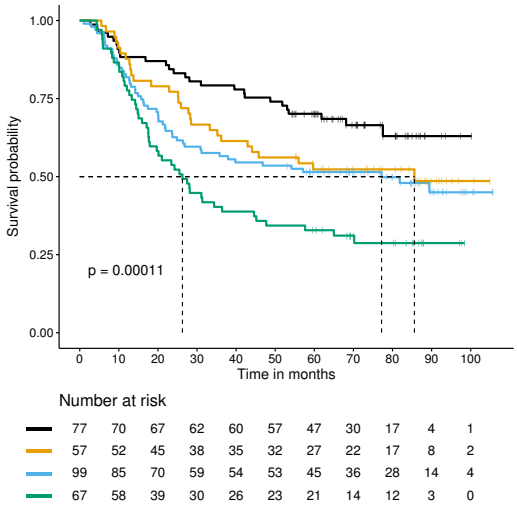
To evaluate prognostic utility of 83 and 36 genes in the top 3 pathways from TCGA GC and mGC immunotherapy datasets, we perform consensus clustering to stratify gastric cancer patients using two validation cohorts ACRG ($N = 300$) and MDACC ($N = 267$), respectively. Setting the number of clusters to 4, we run a consensus clustering method (500 NMF repetition with bootstrapping[15]) on gene expression values of the selected genes in each dataset and generate Kaplan-Meier (KM) plots using overall survival. Figure 3 shows that subtypes identified by 83 and 36 genes from TCGA GC and mGC immunotherapy datasets have distinct survival outcomes which suggests that the pathways identified by our method can serve as prognostic biomarkers to stratify GC patients.

5. Conclusion

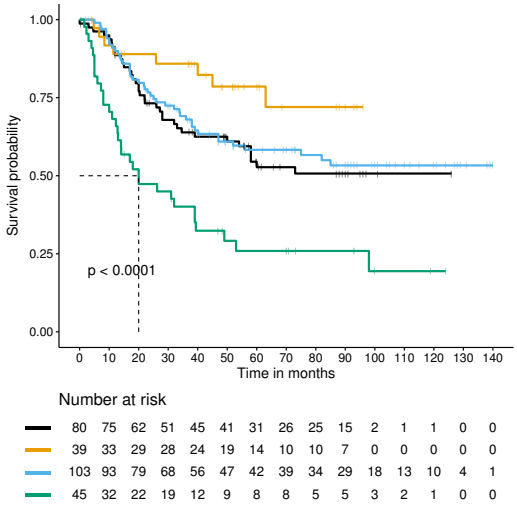
We have proposed a Bayesian semi-nonnegative matrix tri-factorization method to identify associations between cancer phenotypes e.g., molecular subtypes or immunotherapy response, and pathways from the real-valued input matrix, e.g., gene expressions. Motivated by semi-nonnegative factorization,[7] we allow the centroid matrix to be real-valued so that each centroid



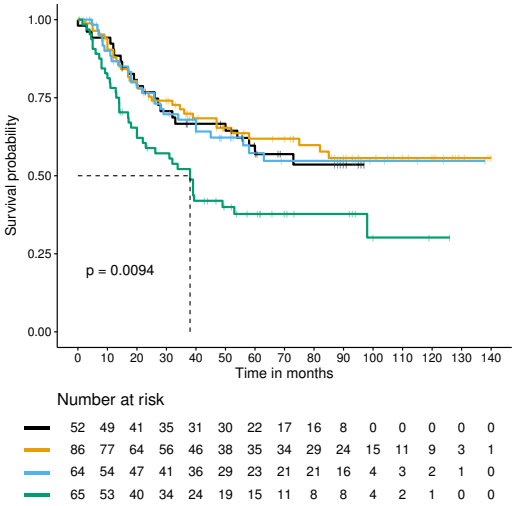
(a1) ACRG subtypes based on 83 genes



(b1) ACRG subtypes based on 36 genes



(a2) MDACC subtypes based on 83 genes



(b2) MDACC subtypes based on 36 genes

Fig. 3. KM plots from ACRG and MDACC cohorts. In each of ACRG and MDACC validation cohorts, four subtypes clustered based on gene expression values of the 83 and 36 gene signatures from TCGA GC and the mGC immunotherapy response datasets, respectively. KM plots with log-rank test indicate that the subtypes identified by the 83 and 36 gene signatures have statistically significant different survival outcomes.

vector can capture the up/down-regulated patterns of member genes in the pathways. We incorporate pathway membership information and a GGI network into the factorization model using the framework of Bayesian learning through structured spike-and-slab priors.[3] We also present efficient variational update rules for the posterior distributions. We show the usefulness of our methods on the synthetic and the gastric cancer data sets. To get a more complete understanding of molecular biology processes, it is necessary to integrate multiple types of genomic data, e.g., gene expression, copy number, miRNA, etc. We believe that data integration can be easily implemented in our factorization model, as similarly done in.[16]

References

1. C. Ding, T. Li, W. Peng and H. Park, Orthogonal nonnegative matrix tri-factorizations for clustering, in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, (Philadelphia, PA, 2006).
2. S. Park, S.-J. Kim, D. Yu, S. Pea-Llopis, J. Gao, J. S. Park, B. Chen, J. Norris, X. Wang, M. Chen, M. Kim, J. Yong, Z. Wardak, K. Choe, M. Story, T. Starr, J.-H. Cheong and T. H. Hwang, An integrative somatic mutation analysis to identify pathways linked with survival outcomes across 19 cancer types, *Bioinformatics* **32**, 1643 (2016).
3. M. R. Andersen, O. Winther and L. K. Hansen, Bayesian inference for structured spike and slab priors, in *Advances in Neural Information Processing Systems (NIPS)*, eds. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger 2014 pp. 1745–1753.
4. P. Kim and B. Tidor, B. subsystem identification through dimensionality reduction of large-scale gene expression dat, *Genome Research* **13**, p. 17061718 (2003).
5. S. T. Kim, R. Cristescu, A. J. Bass, K.-M. Kim, J. I. Odegaard, K. Kim, X. Q. Liu, X. Sher, H. Jung, M. Lee, S. Lee, S. H. Park, J. O. Park, Y. S. Park, H. Y. Lim, H. Lee, M. Choi, A. Talasaz, P. S. Kang, J. Cheng, A. Loboda, J. Lee and W. K. Kang, Comprehensive molecular characterization of clinical responses to pd-1 inhibition in metastatic gastric cancer, *Nature medicine* **24**, p. 14491458 (2018).
6. K. Devarajan, Nonnegative matrix factorization: An analytical and interpretive tool in computational biology, *PLoS Computational Biology* **4** (2008).
7. C. Ding, T. Li and M. I. Jordan, *Convex and Semi-Nonnegative Matrix Factorizations*, Tech. Rep. 60428, Lawrence Berkeley National Lab (2006).
8. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).
9. T. Brouwer and P. Lio', Fast bayesian non-negative matrix factorisation and tri-factorisation, in *NIPS 2016 Workshop: Advances in Approximate Bayesian Inference*, 2016.
10. R. Cristescu, J. Lee, M. Nebozhyn, K.-M. Kim, J. Ting, S. S. Wong, J. Liu, Y. Gang Yue, J. Wang, K. Yu, X. Ye, I.-G. Do, S. Liu, L. Gong, J. Fu, J. Gang Jin, M.-G. Choi, T. Sung Sohn, J. Ho Lee and A. Aggarwal, Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes, *Nature medicine* **21** (04 2015).
11. B. H. Sohn, J.-E. Hwang, H.-J. Jang, H.-S. Lee, S. C. Oh, J.-J. Shim, K.-W. Lee, E. H. Kim, S. Y. Yim, S. H. Lee, J.-H. Cheong, W. Jeong, J. Y. Cho, J. Kim, J. Chae, J. Lee, W. K. Kang, S. Kim, S. H. Noh, J. A. Ajani and J.-S. Lee, Clinical significance of four molecular subtypes of gastric cancer identified by the cancer genome atlas project, *Clinical Cancer Research* (2017).
12. S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie and A. J. Butte, Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets, *PLoS Comput Biol* **6**, p. e1000662 (2010).
13. K. Ganesh and J. Massague, TGF- Inhibition and Immunotherapy: Checkmate, *Immunity* **48**, 626 (04 2018).
14. L. L. van der Woude, M. A. J. Gorris, A. Halilovic, C. G. Figdor and I. J. M. de Vries, Migrating into the Tumor: a Roadmap for T Cells, *Trends Cancer* **3**, 797 (11 2017).
15. S. Monti, P. Tamayo, J. P. Mesirov and T. R. Golub, Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data., *Machine Learning* **52**, 91 (2003).
16. T. Brouwer and P. Lio', Bayesian hybrid matrix factorisation for data integration, in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.