Bayesian semi-nonnegative tri-matrix factorization to identify associations between pathways and sub-types in cancer data

Sunho Park, Nabhonil Kar and Tae Hyun Hwang

Quantitative Health Sciences Cleveland Clinic,
9500 Euclid Ave. Cleveland, OH 44195
{parks, karn2, hwangt}@ccf.org
https://www.lerner.ccf.org/qhs/

This supplementary material provides more details about our proposed method. We first summarize the probability distributions used in our main paper in the following table.

Distribution	PDF	mean	variance	note
Bernoulli$(z\|\rho)$	$\rho^z(1-\rho)^{(1-z)}$	ρ	$\rho(1-\rho)$	$z \in \{0,1\},\ \rho \in [0,1]$
Gamma$(\tau\|a,b)$	$\frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-bz}$	$\frac{a}{b}$	$\frac{a}{b^2}$	$\tau > 0,\ a > 0,\ b > 0$
Exponetial$(s\|\lambda)$	$\lambda e^{-\lambda s}$	λ^{-1}	λ^{-2}	$s \in [0, \infty]$
$\mathcal{N}(x\|\mu,\sigma)$	$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma}}$	μ	σ	$x \in \mathbb{R}$
$\mathcal{TN}(x\|\mu,\sigma)$	$\frac{\mathcal{N}(x\|\mu,\sigma)}{1-\Phi(-\frac{\mu}{\sqrt{\sigma}})}$	$\mu + \sqrt{\sigma} h_1(-\frac{\mu}{\sqrt{\sigma}})$	$\sqrt{\sigma}\Big[1 - h_2(-\frac{\mu}{\sqrt{\sigma}})\Big]$	$x \in \mathbb{R}_+$, $\mathrm{h}_1(x) = \frac{\mathcal{N}(x\|0,1)}{1-\Phi(x)}$, $h_2(x) = h_1(x)[h_1(x) - x]$

1. Model Summary

The observation matrix is decomposed into the sub-matrices in the following way:

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{S}\overline{\boldsymbol{V}}^\top = \boldsymbol{U}\boldsymbol{S}(\boldsymbol{Z} \circ \boldsymbol{V})^\top, \tag{1}$$

where $\circ$ stands for an element-wise multiplication operator. Denoting all the latent variables by $\Theta \triangleq \{\boldsymbol{S}, \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{G}\}$, the joint probability distributions of the model is given as follows:

$$p(\boldsymbol{X}, \Theta) = p(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{V}, \tau) p(\tau) p(\boldsymbol{S}) p(\boldsymbol{V}, \boldsymbol{Z}|\boldsymbol{G}) p(\boldsymbol{G}). \tag{2}$$

where

$$p(\boldsymbol{X}|\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{V}, \gamma) = \prod_{(i,j)\in\Omega} p\big(X_{ij}|\boldsymbol{u}_i^\top \boldsymbol{S}(\boldsymbol{z}_j \circ \boldsymbol{v}_j), \tau\big) \tag{3}$$

$$p(\boldsymbol{S}) = \prod_{k=1}^{K}\prod_{r=1}^{R} p\big(S_{kr}\big), \tag{4}$$

$$p(\boldsymbol{V}, \boldsymbol{Z}|\boldsymbol{G}) = \prod_{r=1}^{R}\prod_{j=1}^{D} p\big(V_{jr} Z_{jr}|G_{jr}\big), \tag{5}$$

$$p(\boldsymbol{G}) = \prod_{r=1}^{R} p(\vec{\boldsymbol{g}}_r|\boldsymbol{m}_r, \boldsymbol{L}), \tag{6}$$

where $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ is a normalized Laplacian matrix and $\boldsymbol{A}$ is an adjacency matrix driven from a protein-protein interaction network ($A_{ij} = 1$ if $i \neq j$ and there is a connection between the genes i and j on the network, and otherwise $A_{ij} = 0$). Note that, the mean vector $\boldsymbol{m}_r$ is set according to the membership information encoded in the pathways Z^0: $m_{jr} = \xi_+$ if $Z^0_{jr} = 1$, otherwise $m_{jr} = \xi_-$, where $\xi_+ > 0$ and $\xi_- < 0$ (in our all experiments, we use $\xi_+ = 3$ and $\xi_- = -5$). The form of each probability distribution in (2) is given as follows:

$$\begin{aligned}
p(X_{ij}|\boldsymbol{u}_i^\top \boldsymbol{S}(\boldsymbol{z}_j \circ \boldsymbol{v}_j), \gamma) &= \mathcal{N}(X_{ij}|\boldsymbol{u}_i^\top \boldsymbol{S}(\boldsymbol{z}_j \circ \boldsymbol{v}_j), \gamma) && (7)\\
p(\gamma) &= \mathrm{Gamma}(\gamma|\alpha_a^0, \alpha_b^0), && (8)\\
p(S_{kr}) &= \mathrm{Exponential}(S_{kr}|\lambda_{kr}^{S0}), && (9)\\
p(V_{jr}, Z_{jr}|G_{jr}) &= \mathcal{N}(Z_{jr}V_{jr}|0, \sigma_{jr}^{V0})\big(\rho_{jr}(G_{jr})\big)^{Z_{jr}}\big(1 - \rho_{jr}(G_{jr})\big)^{(1-Z_{jr})}. && (10)
\end{aligned}$$

2. Variational Inference

The posterior distributions over the latent variables are approximately computed in the framework of variational inference. The variational distributions that approximate the true posterior distributions over the latent variables are assumed to be factorized as follows:

$$q(\Theta) = q(\gamma)\Big(\prod_{k=1}^{K}\prod_{r=1}^{R} q(S_{kr})\Big)\Big(\prod_{j=1}^{D}\prod_{r=1}^{R} q(V_{jr}, Z_{jr})q(G_{jr})\Big), \tag{11}$$

where

$$\begin{aligned}
q(\gamma) &= \mathrm{Gamma}(\gamma|\alpha_a, \alpha_b), && (12)\\
q(S_{kr}) &= \mathcal{TN}(S_{kr}|\mu_{kr}^S, \sigma_{kr}^S), && (13)\\
q(V_{jr}, Z_{jr}) &= \mathcal{N}\Big(V_{jr}|Z_{jr}\mu_{jr}^V, Z_{jr}\sigma_{jr}^V + (1 - Z_{jr})\sigma_{jr}^{V0}\Big)\hat{\rho}_{jr}^{Z_{jr}}(1 - \hat{\rho}_{jr})^{(1-Z_{jr})}, && (14)\\
q(G_{jr}) &= \mathcal{N}(G_{jr}|\mu_{jr}^g, \sigma_{jr}^g). && (15)
\end{aligned}$$

The variational distributions can be computed by maximizing the lower bound with respect to (w.r.t.) the variational distributions. Denoting a set of all the latent variables by $\Theta = \{\gamma, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{G}\}$, we can show that the log-likelihood can be decomposed as follows:

$$\log p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p)) \tag{16}$$

where

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\Theta) \log \frac{p(\boldsymbol{X}, \Theta)}{q(\Theta)} d\Theta, && (17)\\
\mathrm{KL}(q||p) &= -\int q(\Theta) \log \frac{p(\Theta|\boldsymbol{X})}{q(\Theta)} d\Theta. && (18)
\end{aligned}$$

where $\mathrm{KL}(q||p)$ is Kullback-Leibler (KL) divergence between the variational distribution and the true posterior distribution and is always nonnegative ($\mathrm{KL}(q||p) = 0$ if and only if $q = p$). Thus, we can easily see that the log-likehood is lower-bound by the variational lower bound $\mathcal{L}(q)$ and thus the variational distributions can be updated by maximizing $\mathcal{L}(q)$ w.r.t. their parameters. Note that, the variational bound of our model is expressed as follows:

$$
\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_{q(\Theta)}\left[\log \frac{p(\boldsymbol{X}, \Theta)}{q(\Theta)}\right] \\
&= \mathbb{E}_{q(\Theta)}\Big[\log p(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{V}, \gamma)p(\gamma)p(\boldsymbol{S})p(\boldsymbol{V}, \boldsymbol{Z}|\boldsymbol{G})p(\boldsymbol{G})\Big] \\
&\quad - \mathbb{E}_{q(\Theta)}\Big[\log q(\tau)q(\boldsymbol{S})q(\boldsymbol{V}, \boldsymbol{Z})q(\boldsymbol{G})\Big]. \qquad (19)
\end{aligned}
$$

The variational distributions $q(\gamma)$, $\{q(S_{kr})\}$ and $\{q(V_{jr}, Z_{jr})\}$ can be updated in closed from. Letting Θ_l be the variable we want to update at each turn and $\Theta^{\backslash l}$ be the remaining variables, the optimal solution of $q(\Theta_l)$ can be given by the stationary condition for the factor $q(\Theta_l)$ in the maximization problem, i.e., maximize$_{q(\Theta_l)}\mathcal{L}(q)$:

$$
\log q(\Theta_l) \propto \mathbb{E}_{q(\Theta^{\backslash l})}[\log(p(\boldsymbol{X}, \Theta))]. \qquad (20)
$$

On the other hand, 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. We provide detailed derivations of each update in the following subsections.

2.1. *Update of the variational distributions over γ and $\boldsymbol{S}$*

The variational distribution over the precision γ can be updated as follows:

$$
q(\tau) = \mathrm{Gamma}(\tau|\widehat{\alpha}_a, \widehat{\alpha}_b) \qquad (21)
$$

where

$$
\begin{aligned}
\widehat{\alpha}_a &= \alpha_a^0 + \frac{|\Omega|}{2}, \qquad (22) \\
\widehat{\alpha}_b &= \alpha_b^0 + \frac{1}{2}\sum_{(i,j)\in\Omega} \mathbb{E}_q\left[\Big(X_{ij} - \boldsymbol{u}_i^\top \boldsymbol{S}\big(\boldsymbol{z}_j \circ \boldsymbol{v}_j\big)\Big)^2\right], \qquad (23)
\end{aligned}
$$

where Ω is a set of indices of the observations and

$$
\begin{aligned}
&\mathbb{E}_{q(\Theta)}\left[\Big(X_{ij} - \boldsymbol{u}_i^\top \boldsymbol{S}\big(\boldsymbol{z}_j \circ \boldsymbol{v}_j\big)\Big)^2\right] \\
&= \left(X_{ij} - \sum_{k=1}^{K}\sum_{r=1}^{R} \widehat{U}_{ik}\langle S_{kr}\rangle\langle Z_{jr}V_{jr}\rangle\right)^2 + \sum_{k=1}^{K}\sum_{r=1}^{R}\Big[\widehat{U}_{ik}^2\langle S_{kr}^2\rangle\langle Z_{jr}V_{jr}^2\rangle - \widehat{U}_{ik}^2\langle S_{kr}\rangle^2\langle Z_{jr}V_{jr}\rangle^2\Big] \\
&+ \sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{k'\neq k}^{K}\Big[\widehat{U}_{ik}\langle S_{kr}\rangle\Big(\langle Z_{jr}V_{jr}^2\rangle - \langle Z_{jr}V_{jr}\rangle^2\Big)\widehat{U}_{ik'}\langle S_{k'r}\rangle\Big], \qquad (24)
\end{aligned}
$$

where $\langle Z_{jr}V_{jr}\rangle = \rho_{jr}\mu_{jr}^V$ and $\langle Z_{jr}^2 V_{jr}^2\rangle = \langle Z_{jr}V_{jr}^2\rangle = \rho_{jr}\big(\sigma_{jr}^V + (\mu_{jr}^V)^2\big)$.

The variable $\boldsymbol{S}$ can be updated as follows:

$$
q(S_{kr}) = \mathcal{TN}(S_{kr}|\mu_{kr}^S, \sigma_{ij}^S), \qquad (25)
$$

where

$$\sigma_{kr}^{S} = \left(\langle\gamma\rangle \sum_{(i,j)\in\Omega} \widehat{U}_{ik}^{2} \big\langle Z_{jr} V_{jr}^{2} \big\rangle \right)^{-1}, \tag{26}$$

$$\begin{aligned} \mu_{kr}^{S} = \sigma_{kr}^{S} \Bigg[& - \lambda_{kr}^{S} + \langle\gamma\rangle \sum_{(i,j)\in\Omega} \Bigg(\Big(X_{ij} - \sum_{(k',r')\neq(k,r)} \widehat{U}_{ik'} \big\langle S_{k'r'} \big\rangle \big\langle Z_{jr'} V_{jr'} \big\rangle \Big) \widehat{U}_{ik} \big\langle Z_{jr} V_{jr} \big\rangle \\ & - \widehat{U}_{ik} \Big(\big\langle Z_{jr} V_{jr}^{2} \big\rangle - \big\langle Z_{jr} V_{jr} \big\rangle^{2} \Big) \sum_{k'\neq k} \widehat{U}_{ik'} \big\langle S_{k'r} \big\rangle \Bigg) \Bigg]. \end{aligned} \tag{27}$$

2.2. *Update of the variational distributions over Z and V*

Each pair of elements, $\{Z_{jr}, V_{jr}\}$, can be updated by the inference method in.[1] From the stationary condition for $q(Z_{jr}, V_{jr})$ when maximizing the variational bound $\mathcal{L}$ in (19), we have

$$q(V_{jr}, Z_{jr}) = \frac{1}{\mathcal{Z}} \exp \Big\{ \big\langle \log p(X|\Theta) \big\rangle \mathcal{N}(Z_{jr} V_{jr} | 0, \sigma_{jr}^{V0}) \big\langle \Phi(G_{jr}) \big\rangle^{Z_{jr}} \big\langle \big(1 - \Phi(G_{jr})\big) \big\rangle^{(1-Z_{jr})} \Big\}, \tag{28}$$

where $\mathcal{Z}$ is a normalization constant. We can see that $q(V_{jr}, Z_{jr})$ can be factorized as

$$q(V_{jr}, Z_{jr}) = q(V_{jr}|Z_{jr}) q(Z_{jr}). \tag{29}$$

The marginal probability distribution over the binary variable Z_{jr} can be calculated as follows:

$$q(Z_{jr} = 1) = \rho_{jr} = \frac{1}{1 + \exp\{-\xi_{jr}\}}, \tag{30}$$

where

$$\begin{aligned} \xi_{jr} &= \log q(Z_{jr} = 1) - \log q(Z_{jr} = 0) \\ &= \big\langle \log \Phi(G_{jr}) \big\rangle - \big\langle \log(1 - \Phi(G_{jr})) \big\rangle - \frac{1}{2} \log \sigma_{jr}^{V0} + \frac{1}{2} \frac{(\mu_{jr}^{V})^{2}}{\sigma_{jr}^{V}} + \frac{1}{2} \log \sigma_{jr}^{V}. \end{aligned} \tag{31}$$

where the expectations in the second equality are approximated using Jensen's inequality:

$$\big\langle \log \Phi(G_{jr}) \big\rangle \approx \log \Phi\Big(\frac{\mu_{jr}^{g}}{\sqrt{1 + \sigma_{jr}^{g}}} \Big), \tag{32}$$

$$\big\langle \log\big(1 - \Phi(G_{jr})\big) \big\rangle = \big\langle \log\big(\Phi(-G_{jr})\big) \big\rangle \approx \log \Phi\Big(\frac{-\mu_{jr}^{g}}{\sqrt{1 + \sigma_{jr}^{g}}} \Big) \tag{33}$$

The conditional variational distribution of V_{jr} given Z_{jr} is given by

$$q(V_{jr}|Z_{jr} = 0) = \mathcal{N}(V_{jr}|0, \sigma_{jr}^{V0}), \tag{34}$$

$$q(V_{jr}|Z_{jr} = 1) = \mathcal{N}(V_{jr}|\mu_{jr}^{V}, \sigma_{jr}^{V}), \tag{35}$$

where

$$\sigma_{jr}^{V} = \left[(\sigma_{jr}^{V0})^{-1} + \langle\tau\rangle \sum_{i\in\Omega_j} \left(\Big(\sum_{k=1}^{K} \widehat{U}_{ik} \big\langle S_{kr} \big\rangle \Big)^{2} + \sum_{k=1}^{K} \widehat{U}_{ik}^{2} \Big(\big\langle S_{kr}^{2} \big\rangle - \big\langle S_{kr} \big\rangle^{2} \Big) \right) \right]^{-1}, \tag{36}$$

$$\mu_{jr}^{V} = \sigma_{jr}^{V} \left[\frac{\mu_{jr}^{V0}}{\sigma_{jr}^{V0}} + \langle\tau\rangle \sum_{i\in\Omega_j} \left(\Big(X_{ij} - \sum_{k=1}^{K} \sum_{r'\neq r} \widehat{U}_{ik} \big\langle S_{kr'} \big\rangle \big\langle Z_{jr'} V_{jr'} \big\rangle \Big) \sum_{k=1}^{K} \widehat{U}_{ik} \big\langle S_{kr} \big\rangle \right) \right]. \tag{37}$$

As a summary, the joint probability distribution is simply rewritten as follows:

$$q(V_{jr}, Z_{jr}) = \mathcal{N}\Big(V_{jr}|Z_{jr}\mu^{V}_{jr},\ \ Z_{jr}\sigma^{V}_{jr} + (1-Z_{jr})\sigma^{V0}_{jr}\Big)\rho_{jr}^{Z_{jr}}(1-\rho_{jr})^{Z_{jr}}. \tag{38}$$

2.3. *Update of the variational distributions over G*

The optimization problem (19) can be reduced as follows:

$$\text{maximize}_{q(\boldsymbol{G})}\mathcal{L}_g, \tag{39}$$

where $\mathcal{L}_g$ is a function including only terms which are related to the variable $\boldsymbol{G}$:

$$\mathcal{L}_g = \mathbb{E}_{q(\boldsymbol{Z})q(\boldsymbol{G})}\Big[\log p(\boldsymbol{Z}|\boldsymbol{G})p(\boldsymbol{G})\Big] - \mathbb{E}_{q(\boldsymbol{G})}\Big[\log q(\boldsymbol{G})\Big]. \tag{40}$$

The first term of $\mathcal{L}_g$ in eq. (39) can be calculated as follows:

$$\begin{aligned}&\mathbb{E}_{q(\boldsymbol{Z})q(\boldsymbol{G})}\Big[\log p(\boldsymbol{Z}|\boldsymbol{G})\Big]\\ &= \sum_{j,r}\big\langle Z_{jr}\big\rangle\big\langle\log\Phi(G_{jr})\big\rangle + \big\langle(1-Z_{jr})\big\rangle\big\langle\log(1-\Phi(G_{jr}))\big\rangle\\ &\approx \sum_{j,r}\rho_{jr}\log\Phi\Big(\frac{\mu^g_{jr}}{\sqrt{1+\sigma^g_{jr}}}\Big) + (1-\rho_{jr})\log\Phi\Big(\frac{-\mu^g_{jr}}{\sqrt{1+\sigma^g_{jr}}}\Big),\end{aligned} \tag{41}$$

where we have used the same techniques (using Jensen's inequality) as in the previous subsection. We then calculate the third term, a sum of entropy terms of R Gaussian distributions:

$$-\mathbb{E}_{q(\boldsymbol{G})}\Big[\log q(\boldsymbol{G})\Big] = \sum_{r=1}^{R}\mathrm{H}\big(q(\vec{\boldsymbol{g}}_r)\big) = \frac{1}{2}\sum_{r=1}^{R}\log\Big(\prod_{j=1}^{N}\sigma^g_{jr}\Big) + c \tag{42}$$

where c is a constant, which is independent of the variable $\boldsymbol{G}$. The second term is a cross entropy between two Gaussian distributions, $p(G)$ and $q(G)$, calculated as follows:

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{G})}\Big[\log p(\boldsymbol{G})\Big] &= \sum_{r=1}^{R} -\mathrm{H}\big(q(\vec{\boldsymbol{g}}_r)\big) - \mathrm{KL}\big(q(\vec{\boldsymbol{g}}_r)|p(\vec{\boldsymbol{g}}_r)\big)\\ &= -\frac{1}{2}\sum_{r=1}^{R}\Bigg(\Big((\boldsymbol{\mu}^g_r - \boldsymbol{m}_r)^{\top}\boldsymbol{L}^{-1}(\boldsymbol{\mu}^g_r - \boldsymbol{m}_r)\Big) + \Big(\sum_{j=1}^{D}\sigma^g_{jr}[\boldsymbol{L}^{-1}]_{jj}\Big)\Bigg).\end{aligned} \tag{43}$$

The gradient of $\mathcal{L}_g$ w.r.t. the parameters $\{\mu^g_{jr}, \sigma^g_{jr}\}$ also can be easily calculated. We update these parameters using limited-memory BFGS in our experiments

3. Experimental settings

We here explain how to initialize our factorization model, specifically the parameters of the prior distributions (referred to as prior hyperparameters) and those of the variational distributions (referred to as variational parameters). Regarding the variational parameters (e.g., the mean and variance for the case where the variational distribution is Gaussian) note that the variational distributions are updated cyclically, i.e., for each step, we update one variational distributions, fixing the others. Therefore, we need to initialize some (though not all) variational parameters as well as the prior hyperparameters.

The prior hyperparameters are set as follows:

- (For the noise precision γ) $\alpha_a^0 = \alpha_b^0 = 0.1$
- (For each association S_{kr}) $\lambda_{kr}^{S0} = 10$ for all k, r
- (For each element in the centroid, V_{jr}) $\mu_{jr}^{V0} = 0$ and $\sigma_{jr}^{V0} = 1$
- (For the prior mean vectors of the GPs, $\boldsymbol{m}_r$) $\xi_+ = 5$ and $\xi_- = -5$

Based on experience from experimentation on synthetic and various gene expression datasets, the factorization results of the method are generally not sensitive to the most parameters' initial settings. However, we should note that ξ_+ and ξ_- represent the prior belief in the initial pathway membership information $\boldsymbol{Z}^0$. If we set $\xi_+ = \xi_- = 0$, the prior probability of the on-off binary variable $Z_{jr} = 1$ is 0.5 regardless of whether the rth pathway includes the jth gene or not, i.e., for both cases $Z_{jr}^0 = 1$ and $Z_{jr}^0 = 0$. The more extreme values ξ_+ and ξ_- have, i.e. $\xi_+ >> 0$ and $\xi_- << 0$, the stronger the prior belief we place on the initial pathway information $\boldsymbol{Z}^0$. The setting we use in our experiments ($\xi_+ = 5$ and $\xi_- = -5$) usually gives satisfactory factorization results. However, users can adjust them according to their prior belief in the pathway information.

The variational parameters are set as follows:

- (For S_{kr}) $\mu_{kr}^S \sim \text{Uniform}([0, 1])$, for all k, r
- (For $\boldsymbol{V}$) each column is set to the centroid from K-means ran on the input data $\boldsymbol{X}$ if $R < N$ and to the randomly chosen sample (row) from the input matrix $\boldsymbol{X}$ otherwise.
- (For G_{jr}) $\mu_{jr}^g = m_{jr}$ and $\sigma_{jr}^g = \exp(-\zeta)$, where $\zeta \sim \mathcal{N}(0, 0.1^2)$

4. Bayesian semi-nonnegative- vs Point estimate non-negative factorization

As explained in the main paper, many types of genomic data are given in a form of a real-valued matrix after relevant normalization or transformation steps. However, the NMF formulation does not permit negative values in the inputted observation matrix. Thus, one of the standard ways to handle negative values for the NMF formulation is to fold the original matrix by columns:[2] every column (gene) will be represented in two new columns in a new observation matrix, one of which contains only the positive values (upregulations) and the other column only the magnitudes of the negative values (downregulations). However, the folding approach incurs increased computational complexities: the number of columns in the input matrix is doubled, and the gene-gene interaction network is 2^2 times larger. On the other hand, our approach (motivated by semi-nonnegative factorization[3]) allows the centroid matrix to have negative values but still imposes nonnegativity constraints on the encoding matrix. Furthermore, we implement our semi-nonnegative tri-matrix factorization algorithm in the framework of Bayesian learning. In the following subsections, we provide two specific examples that show the superiority of our method over non-negative tri-matrix factorization (NTriPath[4]) which is implemented using the folding approach to deal with negative values in the input matrix.

Before presenting the details of the simulated experiments, we first provide a brief introduction to NTriPath. The objective of the method is again to approximate the input matrix $\boldsymbol{X}$ as a product of the three small matrices, $\boldsymbol{U}$ (the sub-type indicator matrix), $\boldsymbol{S}$ (the association matrix) and $\overline{\boldsymbol{V}}$ (the centroid matrix), i.e., $\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{S}\overline{\boldsymbol{V}}^\top$. With $\boldsymbol{U}$ fixed, $\boldsymbol{S}$ and $\overline{\boldsymbol{V}}$ are estimated by minimizing the following objective function under the non-negativity constraints:

$$\text{minimize}_{\boldsymbol{S}\geq 0, \overline{\boldsymbol{V}}\geq 0} \frac{1}{2} f(\boldsymbol{S}, \overline{\boldsymbol{V}}), \tag{44}$$

where the objective function is define as:

$$f(\boldsymbol{S}, \overline{\boldsymbol{V}}) = \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\overline{\boldsymbol{V}}^{\top}\|_F^2 + \lambda_S\|\boldsymbol{S}\|_1^2 + \lambda_V\|\overline{\boldsymbol{V}}\|_1^2 + \lambda_Z\|\overline{\boldsymbol{V}} - \boldsymbol{Z}^0\|_F^2 + \lambda_{V_L}\text{tr}\{\overline{\boldsymbol{V}}^{\top}\boldsymbol{L}\overline{\boldsymbol{V}}\}. \tag{45}$$

The matrices $\boldsymbol{S}$ and $\overline{\boldsymbol{V}}$ are updated by multiplicative rules to ensure the non-negativity constraints.[4] Note that NTriPath involves 4 regularization parameters which should be specified by the user. Identification of associations between sub-types and pathways from input data is clearly an unsupervised learning problem since true associations generally are unknown. Therefore, it is unclear how to tune the regularization parameters of NTriPath for the given input data. In addition, the method incurs a high computational burden for large scale datasets when the best combination of the hyperparameters is searched among a set of candidate values (the search space being a 4D grid space) by cross validation. For simplicity, we fix $\lambda_{V_L} = \lambda_Z = 1$ as in our previous work. We tune only λ_S and λ_V which are related to the sparseness of the metrics. We select the best regularization constants from 2D grid space (the grid space along each dimension being defined as $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$) by finding the combination which gives the least reconstruction error. Meanwhile, our Bayesian method is able to automatically tune model complexity, including the noise precision, by integrating over all the latent variables.

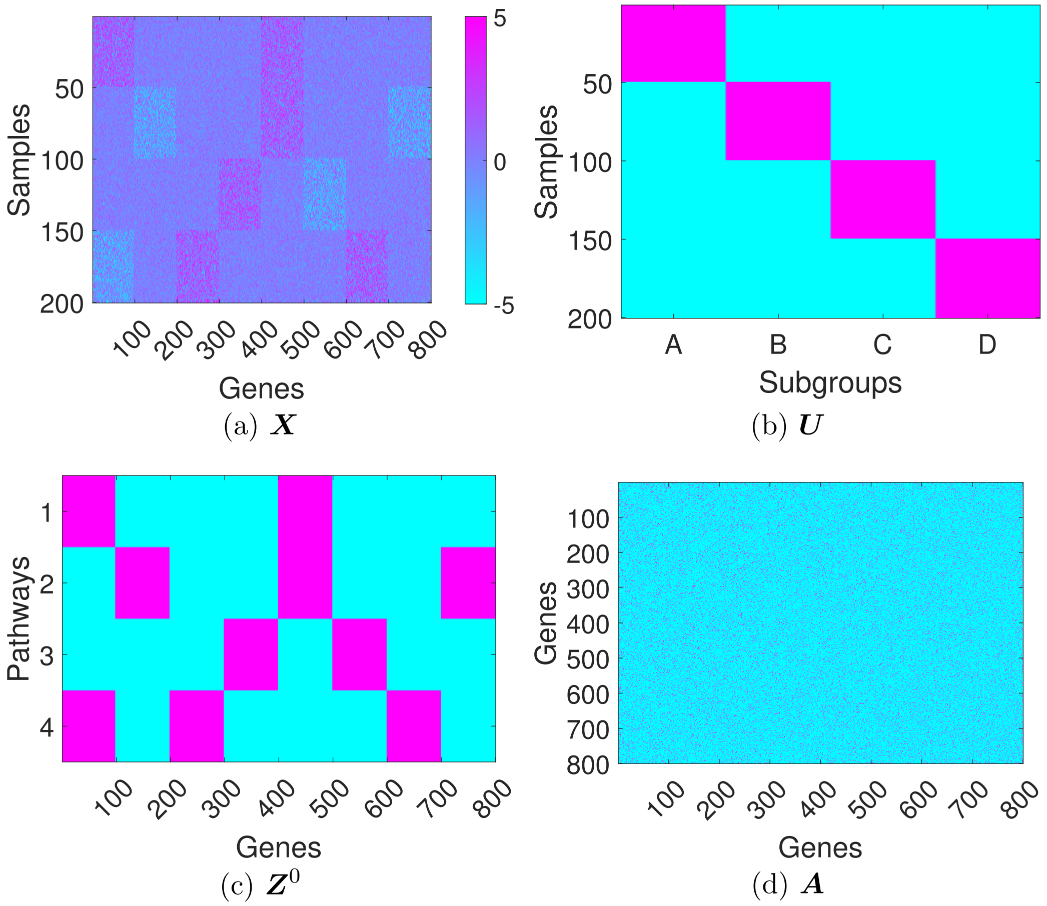
4.1. *Baseline example*

We begin by presenting a simple example that contains a basic structure in the observation matrix and other inputs. We discuss the results of this straightforward settings before examining the two cases in which our proposed Bayesian framework provides clear advantages over the folding approach. A detailed overview of data generation is now presented for our first example.

Inspired by the biological setting of the gene expression application in the main paper, we use the same terminology here as in the main script (i.e. subgroups, genes, pathways) in discussing the problem formulation and results of our simulated experiments. As in the main application, these experiments attempt to decompose patterns of upregulated and downregulated genes within different patient subgroup. In this preliminary example, as well as in subsequent experiments, the observation matrix $\boldsymbol{X} \in \mathbb{R}^{200\times 800}$ consists of 4 subgroups (each containing 50 samples) with some defined pattern among the 800 genes (which are grouped into sets of 100). Within each subgroup, a set of 100 genes can represent upregulation, downregulation or background noise. For upregulated and downregulated genes, samples are drawn from a Gaussian $\mathcal{N}(1, 2)$ or Gaussian $\mathcal{N}(-1, 2)$, respectively. For background noise, samples are drawn from a Gaussian $\mathcal{N}(0, 1)$. A simple block structure was determined with each subsample containing 2-3 "selected" (either upregulated or downregulated) gene sets (Fig. 3a). Subgroups are encoded in $\boldsymbol{U} \in \mathbb{R}_+^{200\times 4}$ using simple 1-of-K encoding ($U_{ij} \in \{0, 1\}$ and $\Sigma_j U_{ij} = 1$) (Fig. 1b). Gene-pathway prior knowledge is encoded in the matrix $\boldsymbol{Z}^0 \in \mathbb{R}_+^{800\times 4}$ and is initialized to contain similar structure to the observation $\boldsymbol{X}$. That is, we allow $\boldsymbol{Z}^0$ to contain 4 pathways that reflect the same pattern of selected genes within the subgroups of $\boldsymbol{X}$ by setting

$Z^0_{ij} = 1$ for all the genes i that are upregulated or downregulated within the subgroup that we have designated to pathway j (Fig. 1c). The gene-gene interaction network $\boldsymbol{A} \in \mathbb{R}^{800\times800}$ is initialized with approximately 10% sparcity and contains random symmetric connections (with no self connections; Fig. 3c). Our motivation for this simple design is to allow our method to arrive at a simple and predictable solution for the model's learned factors, namely, the subgroup-pathway association matrix $\boldsymbol{S} \in \mathbb{R}^{4\times4}_{+}$, the real-valued pathway-gene association matrix $\boldsymbol{V} \in \mathbb{R}^{800\times4}$ and the updated pathway-gene binary membership matrix $\boldsymbol{Z} \in \mathbb{R}^{800\times4}_{+}$.

Fig. 2 and Fig. 3 show the factorization results of our method and NTriPath, respectively. Both methods produce correct estimates of the true association matrix $\boldsymbol{S}$ although our method's estimate is more clearly separable (Fig. 2a).



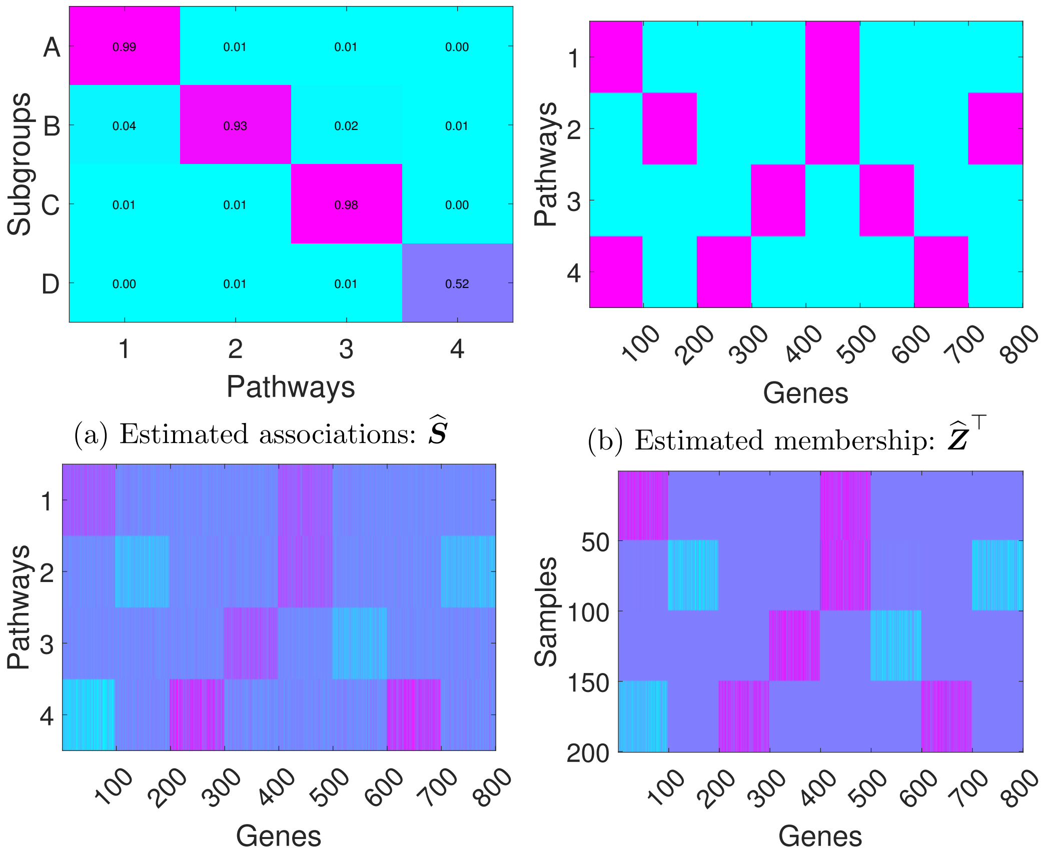
(a) $\boldsymbol{X}$

(b) $\boldsymbol{U}$

(c) $\boldsymbol{Z}^0$

(d) $\boldsymbol{A}$

Fig. 1. Inputs to both algorithms.



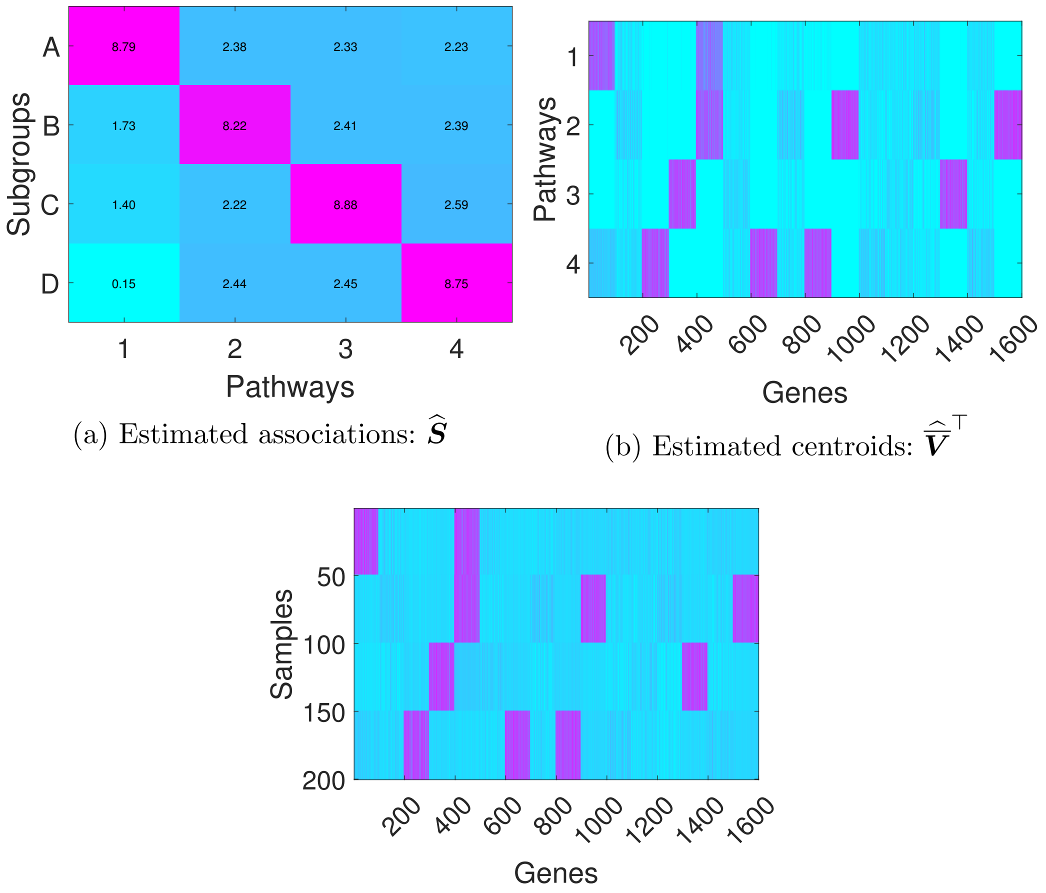
(a) Estimated associations: $\widehat{\boldsymbol{S}}$

(b) Estimated membership: $\widehat{\boldsymbol{Z}}^{\top}$

(c) Estimated centroids: $\widehat{\boldsymbol{V}}^{\top}$

(d) Reconstructed input: $\widehat{\boldsymbol{X}}$

Fig. 2. Factorization result from our method.



(a) Estimated associations: $\widehat{\boldsymbol{S}}$

(b) Estimated centroids: $\widehat{\widehat{\boldsymbol{V}}}^{\top}$

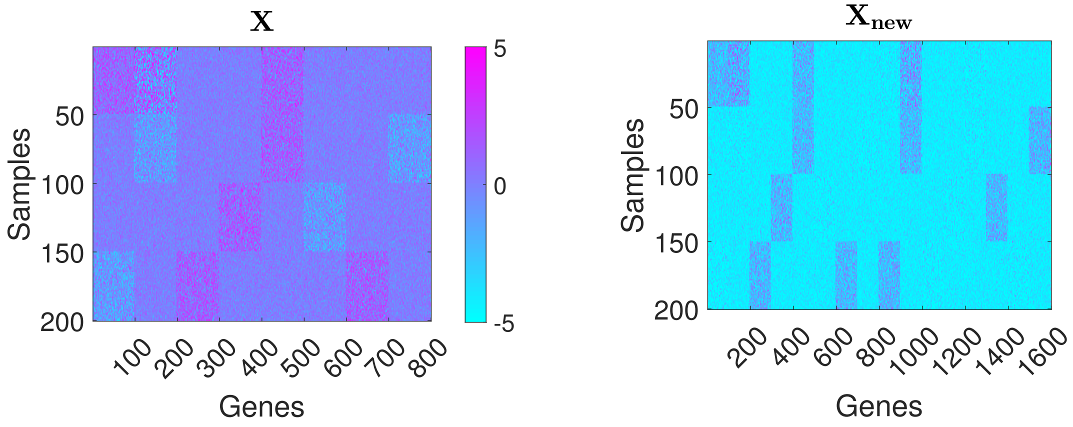
(c) Reconstructed input: $\widehat{\boldsymbol{X}}$

Fig. 3. Factorization results from NTriPath.

4.2. *Limitations of NMF methods based on the folding approach*

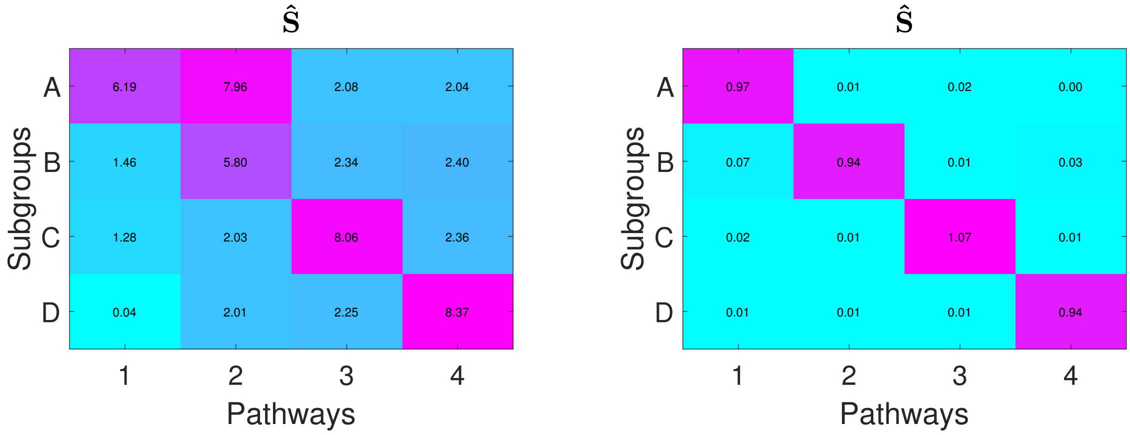
We here provide a simple example where NTriPath, employing the folding approach to deal with negative values, fails to correctly estimate associations between sub-types and pathways from a real-valued input matrix. The main reason for this incorrectly learned association matrix is that the folding approach breaks the original underlying patterns of the input matrix by separating non-negative and negative values. In fact, this issue is problematic not only for NTriPath but for all NMF methods that are based on the folding approach. Note that, however, our factorization method is free from this issue due to the semi-nonnegative modeling, which is one of the primary advantages of our method compared to NTriPath and other NMF based methods using the folding approach.

Fig. 4 shows how the input matrix $\boldsymbol{X}$ is generated based on the baseline example in the previous subsection (Fig. 4a) and how the new non-negative input matrix $\boldsymbol{X}_{new}$ is constructed by the folding approach i.e., $X_{\text{new}} \triangleq [\max(X, 0), \quad \max(-X, 0)]$ (Fig. 4b). We assume that expression values at the noisy block of genes in the first sub-type samples are drawn from i.i.d. Gaussian distributions (white Gaussian noise) with a high variance, i.e., $X_{ij} \sim \mathcal{N}(0, 5^2)$. In other words, this data block (represented as $X[1 : 50, 101 : 200]$ in MATLAB language) contains just random noise values. However, these negative and non-negative noisy elements become strong signal blocks when the input matrix is transformed by the folding approach (see $X_{new}[1:50, 101:200]$ and $X_{new}[1:50, 901:1000]$ in Fig. 4b). Thus, NTriPath tries to fit both noisy blocks, which should be ignored as noise, by adjusting the association matrix $\boldsymbol{S}$ and other factorization results. Fig. 5a clearly supports this discussion. We can see that the estimated association matrix $\widehat{\boldsymbol{S}}$ from NTriPath fails to recover the expected subgroup-pathway associations and thus the top pathways associated with each sub-type are also incorrect. However, as mentioned before, our factorization method based on the semi-nonnegative modelling yields the correct estimate for associations without being affected by the presence of the noise block (Fig. 5b).



(a) (b)

Fig. 4. A simple example where NMF methods based on the folding approach fail to correctly estimate true association between sub-types and pathway from a real-valued input matrix: a) There is a noisy block in the original input matrix, i.e., $X[1:50, 101:200]$; b) the negative and non-negative values in this block become strong signals after transformed by the folding approach.



(a) (b)

Fig. 5. The estimated association matrices from (a) NTriPath and from (b) our method. The presence of the noisy block causes NTriPath to make an incorrect association identification.

4.3. *Robustness against noise*

We compare the performance of our Bayesian factorization method and of NTriPath in the case where the input matrix is contaminated by background noise of different noise levels. Our objective here is to show whether each method is robust against increasing noise. We assume that the observation matrix $\widetilde{\boldsymbol{X}}$ is generated by adding white Gaussian noises to the data input matrix $\boldsymbol{X}$ defined in Section 4.1, i.e., $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \widetilde{\boldsymbol{E}}$, where $\widetilde{E}_{ij} \sim \mathcal{N}(0, \gamma_n^{-1})$ and the noise variances γ_n^{-1} increases incrementally from 1^2 to 10^2. We train both methods on the noisy observation matrix $\widetilde{\boldsymbol{X}}$ to test how each method performs in correctly identifying the associations between the sub-types and the pathways.

We report the performance of both methods in Fig. 6. Since we know the ground truth associations for this dataset, we can calculate the accuracy of each method based on how many associations each method correctly predicts. We repeat each experiment 20 times at each noise variance. As we can see in the figure, our method shows overall stable performance in the entire range of the noise variance. Note that our method shows the slightly worse performance at the first three noise levels but maintains almost the same performance as the noise level increases. We hypothesize that the sub-optimal performance of our method at the first three noise levels is caused by improperly randomized initialization points. However, our Bayesian factorization method still works well at high noise levels, where the performance of NTriPath dramatically diminishes. This supports our claim of the greater robustness of our method against noise relative to NTriPath.

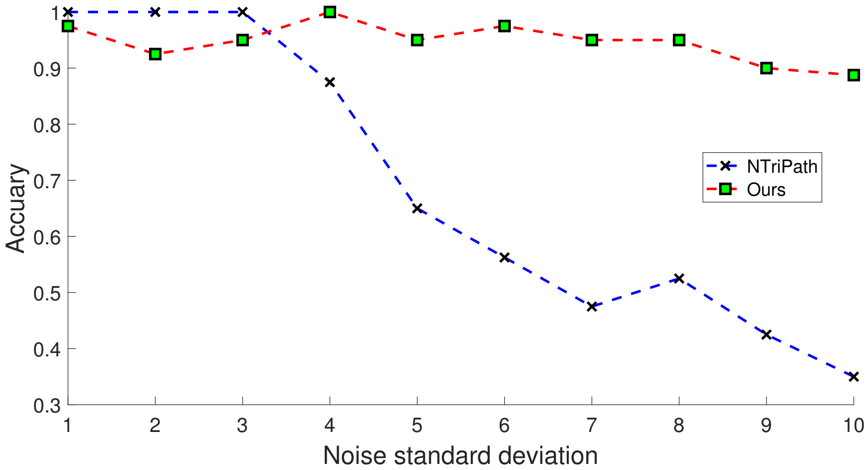


Fig. 6. Robustness of both methods against noise: the prediction performance of our method is compared to NTriPath as the noise variance increases.

5. TCGA gastric cancer and metastatic gastric cancer immunotherapy clinical-trial datasets: additional information

We here include the list of the selected pathways from both data sets in the experimental results section in the main text. Please see Table 2 for the TCGA gastric cancer data and

Table 3 for the metastatic gastric cancer immunotherapy clinical-trial data.

Table 2. Summary of the top 3-ranked pathways associated with the molecular sub-types obtained from the TCGA gastric cancer dataset.

sub-types	rank	#members	member genes
CIN	1	12	ADAMTS4,CELA1,CTRB1,DERL1,DERL2,DERL3,KLK5,MFI2, MMP26,PRSS1,PRSS3,SERPINA1
	2	14	COL2A1,COL3A1,COL9A1,COL9A2,COL9A3,COMP,FN1,MAG, MAP1B,MBP,NGFR,PLP1,PRNP,RTN4R
	3	13	BARD1,BRCA1,CSTF1,CSTF2,CSTF3,FEZ1,HTATSF1,IKBKAP, MED21,PIN1,POLR2A,RBBP8,SUPT5H
EBV	1	12	ADAMTS4,CELA1,CTRB1,DERL1,DERL2,DERL3,KLK5,MFI2, MMP26,PRSS1,PRSS3,SERPINA1
	2	13	C3,F2,F2RL3,FCER2,HP,ICAM2,ICAM4,ITGAM,ITGAX,ITGB2, JAM2,JAM3,TJP1
	3	14	CD44,EED,FN1,ICAM4,ITGA4,ITGAE,ITGB1,ITGB7,LGALS8, MADCAM1,PXN,TLN1,VCAM1,VCAN
GS	1	10	CALM1,CPE,GCG,GLP1R,GPRASP1,GRM5,MEP1A,MEP1B, OPRM1,VIPR1
	2	1	GNAQ
	3	2	CD200,CD200R1
MSI	1	3	CCDC67,CCDC85B,EIF3E
	2	1	GNAQ
	3	3	NUP155,NUPL2,ZFYVE9

Table 3. Summary of the top 3-ranked pathways associated with the treatment response obtained from the metastatic gastric cancer immunotherapy clinical-trial dataset.

sub-types	rank	#members	member genes
responder	1	14	ATN1,ECM1,ELN,FBLN1,FBLN2,FBN1,FBN2,FN1,HSPG2,ITGB1,LTBP1,MFAP2,PRELP,VCAN
	2	11	CCL19,CCL21,CCL5,CCR3,CXCL11,CXCL13,CXCL9,DPP4,IGFBP7,PF4,VCAN
	3	10	CCL11,CCL5,CCR3,CPAMD8,CXCL11,CXCL13,CXCL9,DPP4,FAP,PF4
non-responder	1	3	SLC1A4,SLC1A5,TBC1D17
	2	3	CCDC85B,KRTAP4-12,LMO2
	3	3	NMU,NMUR1,NMUR2

References

1. M. K. Titsias and M. Lázaro-Gredilla, Spike and slab variational inference for multi-task and multiple kernel learning, in *Advances in Neural Information Processing Systems (NIPS)*, eds. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger 2011 pp. 2339–2347.
2. P. Kim and B. Tidor, B. subsystem identification through dimensionality reduction of large-scale gene expression dat, *Genome Research* **13**, p. 17061718 (2003).
3. C. Ding, T. Li and M. I. Jordan, *Convex and Semi-Nonnegative Matrix Factorizations*, Tech. Rep. 60428, Lawrence Berkeley National Lab (2006).
4. 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).