Supplementary Material: Non-Gaussian Discriminative Factor Models via the Max-Margin Rank Likelihood

Xin Yuan* EIEXYUAN@GMAIL.COM
Ricardo Henao* R.HENAO@DUKE.EDU
Ephraim L. Tsalik E.T@DUKE.EDU
Duke University, Durham, NC, 27708, USA
Raymond J. Langley RLANGLEY@LRRI.ORG
Lovelace Respiratory Research Institute, Albuquerque, NM 87108, USA
Lawrence Carin LCARIN@DUKE.EDU
Duke University, Durham, NC, 27708, USA

1. Model

The full Bayesian model is:

$$\begin{aligned}
x_{i,n} &= g_i(w_{i,n}), \quad i=1,\dots,d;\ n=1,\dots,N;\\
L_i(w_{i,n}|\mathbf{w}_{i\backslash n}) &= \int \mathcal{N}(w_{i,n}-w^u_{i,n}; -\epsilon-\lambda^u_{i,n}, \lambda^u_{i,n})\\
&\quad \times \mathcal{N}(w^l_{i,n}-w_{i,n}; -\epsilon-\lambda^l_{i,n}, \lambda^l_{i,n}) d\lambda^u_{i,n} d\lambda^l_{i,n}\,,\\
L_n(y_n|\boldsymbol{\beta},\mathbf{z}_n) &= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda^c_n}}\\
&\quad \times \exp\left(-\frac{1}{2}\frac{(1-y_n\boldsymbol{\beta}^\top\mathbf{z}_n+\lambda^c_n)^2}{\lambda^c_n}\right) d\lambda^c_n\,,\\
\mathbf{z}_n &\sim \mathcal{N}(\boldsymbol{\mu}_{t(n)}, \psi^{-1}_{t(n)}\mathbf{I}_K)\,,\\
\boldsymbol{\mu}_t &\sim (\mathbf{0}, \mathbf{I}_K),\\
\psi_t &\sim \mathrm{Ga}(\psi_s, \psi_r),\\
t(n) &\sim \mathrm{Mult}(1; q_1, \dots, q_T),\\
q_t &= \nu_t \prod_{l=1}^{t-1}(1-\nu_l)\,,\\
\nu_t &\sim \mathrm{Beta}(1,\alpha), \quad \alpha \sim \mathrm{Ga}(\alpha_s, \alpha_r),\\
a_{i,k} &\sim \mathcal{N}(0, \xi_{i,k})\,, \quad \xi_{i,k} \sim \mathrm{Ga}(r_a, \eta_{i,k})\,,\\
\eta_{i,k} &\sim \mathrm{Ga}(s_a, \Phi^{(a)}_k)\,,\\
\Phi^{(a)}_k &\sim \mathrm{Ga}(1/2, \tilde{\Phi}^{(a)}), \quad \tilde{\Phi}^{(a)} \sim \mathrm{Ga}(1/2, 1),\\
\beta_k &\sim \mathcal{N}(0, b_k)\,, \quad b_k \sim \mathrm{Ga}(r_\beta, e_k)\,,\\
e_k &\sim \mathrm{Ga}(s_\beta, \Phi^{(\beta)})\,,\\
\Phi^{(\beta)} &\sim \mathrm{Ga}(1/2, \tilde{\Phi}^{(\beta)}), \quad \tilde{\Phi}^{(\beta)} \sim \mathrm{Ga}(1/2, 1).
\end{aligned}$$

2. MCMC inference

For convenience, we denote:

$$\begin{aligned}
\lambda_{i,n} &= (\lambda^l_{i,n})^{-1} + (\lambda^u_{i,n})^{-1},\\
\Delta^{(k)}_{i,n} &= \left(\frac{w^l_{i,n}+\epsilon-w_{i,n}}{\lambda^l_{i,n}} - \frac{w_{i,n}+\epsilon-w^u_{i,n}}{\lambda^u_{i,n}}\right)\\
&\quad + a_{i,k}z_{k,n}\lambda_{i,n},
\end{aligned}$$

$$\begin{aligned}
(\boldsymbol{\beta}^\top\mathbf{z}_n)_{\backslash k} &= \boldsymbol{\beta}^\top\mathbf{z}_n - \beta_k z_{k,n},\\
\Gamma_{k,n} &= \frac{y_n\beta_k[1+\lambda^c_n - y_n(\boldsymbol{\beta}^\top\mathbf{z}_n)_{\backslash k}]}{\lambda^c_n}.
\end{aligned}$$

In the following conditional posterior-distributions, "$\cdot$" refers to the conditioning parameters of the distributions, $\mathrm{IG}(a,b)$ denotes the inverse Gaussian distribution, $\mathrm{Ga}(a,b)$ the gamma distribution, and $\mathrm{GIG}(a,b,p)$ the generalized inverse Gaussian distribution.

In the *linear* case, when the Dirichlet process mixture (DPM) model is not used, the conditional posterior distributions are:

Z:

$$\begin{aligned}
p(z_{k,n}|\,\cdot) &= \mathcal{N}(\mu_{z_{k,n}}, \sigma^2_{z_{k,n}})\,,\\
\sigma^{-2}_{z_{k,n}} &= 1 + \sum_{i=1}^d a^2_{i,k}\lambda^{-1}_{i,n} + \frac{\beta^2_k}{\lambda^c_n}\,,\\
\mu_{z_{k,n}} &= \sigma^2_{z_{k,n}}\left(\sum_{i=1}^d a_{i,k}\Delta^{(k)}_{i,n} + \Gamma_{k,n}\right)\,.
\end{aligned}$$

$\mathbf{\Lambda}^l, \mathbf{\Lambda}^u, \boldsymbol{\lambda}^c$:

$$\begin{aligned}
p\left((\lambda^l_{i,n})^{-1}|\cdot\right) &= \text{IG}\left(|w^l_{i,n}+\epsilon-w_{i,n}|^{-1},1\right) \,, \\
p\left((\lambda^u_{i,n})^{-1}|\cdot\right) &= \text{IG}\left(|w_{i,n}+\epsilon-w^u_{i,n}|^{-1},1\right) \,, \\
p\left((\lambda^c_{n})^{-1}|\cdot\right) &= \text{IG}\left(|1-y_n\boldsymbol{\beta}^\top\mathbf{z}_n|^{-1},1\right) \,.
\end{aligned}$$

A:

$$\begin{aligned}
p(a_{i,k}|\,\cdot) &= \mathcal{N}(\mu_{a_{i,k}},\sigma^2_{a_{i,k}}) \,, \\
\sigma^{-2}_{a_{i,k}} &= \xi^{-1}_{i,k} + \sum_{n=1}^{N} z^2_{k,n}\lambda^{-1}_{i,n} \,, \\
\mu_{a_{i,k}} &= \sigma^2_{a_{i,k}} \sum_{n=1}^{N} z_{k,n}\Delta^{(k)}_{i,n} \,, \\
p(\xi_{i,k}|\cdot) &= \text{GIG}(2\eta_{i,k}, a^2_{i,k}, r_a - 0.5) \,, \\
p(\eta_{i,k}|\cdot) &= \text{Ga}(r_a + s_a, \xi_{i,k} + \Phi^{(a)}_k) \,, \\
p(\Phi^{(a)}_k|\cdot) &= \text{Ga}\left(\frac{1}{2} + s_a d, \tilde{\Phi}^{(a)} + \frac{1}{2}\sum_i \eta_{i,k}\right) \,, \\
p(\tilde{\Phi}^{(a)}|\cdot) &= \text{Ga}\left(1, \sum_k \Phi^{(a)}_k + 1\right) \,.
\end{aligned}$$

$\boldsymbol{\beta}$:

$$\begin{aligned}
p(\beta_k|\cdot) &= \mathcal{N}(\mu_{\beta_k},\sigma^2_{\beta_k}) \,, \quad \sigma^{-2}_{\beta_k} = b^{-1}_k + \sum_{n=1}^{N}\frac{z^2_{k,n}}{\lambda^c_n} \,, \\
\mu_{\beta_k} &= \sigma^2_{\beta_k}\sum_{n=1}^{N}\frac{y_n z_{k,n}\left[1+\lambda^c_n - y_n(\boldsymbol{\beta}^\top\mathbf{z}_n)_{\setminus k}\right]}{\lambda^c_n} \,, \\
p(b_k|\cdot) &= \text{GIG}(2e_k, \beta^2_k, r_\beta - 0.5), \\
p(e_k|\cdot) &= \text{Ga}(r_\beta + s_\beta, b_k + \Phi^{(\beta)}_k) \,, \\
p(\Phi^{(\beta)}|\cdot) &= \text{Ga}\left(\frac{1}{2} + s_\beta K, \tilde{\Phi}^{(\beta)} + \frac{1}{2}\sum_k e_k\right) \,, \\
p(\tilde{\Phi}^{(\beta)}|\cdot) &= \text{Ga}\left(1, \Phi^{(\beta)} + 1\right) \,.
\end{aligned}$$

In the *nonlinear* case, when the DPM is used:

$t(n)$ (mixture component index for n-th observation):

$$p(t(n)=t|\cdot) \propto q_t\mathcal{N}(\mathbf{z}_n;\boldsymbol{\mu}_t,\psi_t^{-1}\mathbf{I}_K) \,.$$

DPM parameters:

$$\begin{aligned}
p(\mu_{t,k}|\cdot) &= \mathcal{N}(\mu_{\mu_{t,k}},\sigma^2_{\mu_{t,k}}) \,, \quad \sigma^{-2}_{\mu_{t,k}} = 1 + \sum_{n:t(n)=t}\psi_t \,, \\
\mu_{\mu_{t,k}} &= \sigma^2_{\mu_{t,k}}\psi_t\sum_{n:t(n)=t} z_{k,n} \,, \\
p(\psi_t|\cdot) &= \text{Ga}\left(\psi_s + 0.5K,\;\; \psi_r + 0.5\sum_k \mu^2_{t,k}\right) \,, \\
p(\nu_t|\cdot) &= \text{Beta}\left(1 + \sum_{n:t(n)=t} 1,\;\; \alpha + \sum_{n:t(n)>t} 1\right) \,, \\
p(\alpha|\cdot) &= \text{Ga}\left(\alpha_s + T - 1,\;\; \alpha_r - \sum_{t=1}^{T-1}\log(1-\nu_t)\right) \,.
\end{aligned}$$

In this case, $\boldsymbol{\beta}$ in 2) and 4) should be replaced by $\boldsymbol{\beta}^{(t)}$, for $t = 1,\ldots,T$, and the summation over n in 4) will only account for $\{n : t(n) = t\}$.

3. VB inference

Since the model is fully local conjugate, the VB update equations can be obtained using the moments of the above conditional posterior distributions. Here we present the moments for the model without DPM, and for the VB inference of the DP mixture model, please refer to (Blei & Jordan, 2005). In the following expressions, $\langle\cdot\rangle$ denotes expectation, $\mathcal{K}_p(\cdot)$ is the modified Bessel function of the second kind, $\langle w_{i,n}\rangle = \langle\mathbf{a}_i^\top\rangle\langle\mathbf{z}_n\rangle$, $\langle w^l_{i,n}\rangle = \langle\mathbf{a}_i^\top\rangle\langle\mathbf{z}^l_n\rangle$ and $\langle w^u_{i,n}\rangle = \langle\mathbf{a}_i^\top\rangle\langle\mathbf{z}^u_n\rangle$.

Z:

$$\begin{aligned}
\langle z_{k,n}\rangle &= \langle\sigma^2_{z_{k,n}}\rangle\left(\sum_{i=1}^{d}\langle a_{i,k}\rangle\langle\Delta^{(k)}_{i,n}\rangle + \langle\Gamma_{k,n}\rangle\right) \,, \\
\langle\Gamma_{k,n}\rangle &= \langle(\lambda^c_n)^{-1}\rangle\left\{y_n\langle\beta_k\rangle[\langle(\lambda^c_n)^{-1}\rangle + 1\right. \\
&\qquad \left. -\langle(\lambda^c_n)^{-1}\rangle y_n(\langle\boldsymbol{\beta}^\top\rangle\langle\mathbf{z}_n\rangle)_{\setminus k}]\right\} \,, \\
\langle\sigma^2_{z_{k,n}}\rangle &= \left(1 + \sum_{i=1}^{d}\langle a^2_{i,k}\rangle\langle\lambda^{-1}_{i,n}\rangle + \langle\beta^2_k\rangle\langle(\lambda^c_n)^{-1}\rangle\right)^{-1} \,, \\
\langle z^2_{k,n}\rangle &= \langle z_{k,n}\rangle^2 + \langle\sigma^2_{z_{k,n}}\rangle.
\end{aligned}$$

$\mathbf{\Lambda}^l, \mathbf{\Lambda}^u, \boldsymbol{\lambda}^c$:

$$\begin{aligned}
\langle(\lambda^l_{i,n})^{-1}\rangle &= \left|\langle w^l_{i,n}\rangle + \epsilon - \langle w_{i,n}\rangle\right|^{-1} \,, \\
\langle(\lambda^u_{i,n})^{-1}\rangle &= \left|\langle w_{i,n}\rangle + \epsilon - \langle w^u_{i,n}\rangle\right|^{-1} \,, \\
\langle(\lambda^c_{n})^{-1}\rangle &= \left|1 - y_n\langle\boldsymbol{\beta}^\top\rangle\langle\mathbf{z}_n\rangle\right|^{-1} \,.
\end{aligned}$$

A:

$$
\begin{aligned}
\langle a_{i,k}\rangle &= \langle\sigma^2_{a_{i,k}}\rangle \sum_{n=1}^{N}\langle z_{k,n}\rangle\langle\Delta^{(k)}_{i,n}\rangle,\\
\langle\sigma^2_{a_{i,k}}\rangle &= \left(\langle\xi^{-1}_{i,k}\rangle + \sum_{n=1}^{N}\langle z^2_{k,n}\rangle\langle\lambda^{-1}_{i,n}\rangle\right)^{-1},\\
\langle a^2_{i,k}\rangle &= \langle a_{i,k}\rangle^2 + \langle\sigma^2_{a_{i,k}}\rangle,\\
\langle\Delta^{(k)}_{i,n}\rangle &= \langle(\lambda^l_{i,n})^{-1}\rangle\left(\langle w^l_{i,n}\rangle + \epsilon - \langle w_{i,n}\rangle\right)\\
&\quad - (\langle\lambda^u_{i,n})^{-1}\rangle\left(\langle w_{i,n}\rangle + \epsilon - \langle w^u_{i,n}\rangle\right)\\
&\quad + \langle a_{i,k}\rangle\langle z_{k,n}\rangle\left[\langle(\lambda^l_{i,n})^{-1}\rangle + \langle(\lambda^u_{i,n})^{-1}\rangle\right],\\
\langle\xi_{i,k}\rangle &= \frac{\sqrt{\langle a^2_{i,k}\rangle}\mathcal{K}_{r_a+0.5}\left(\sqrt{2\langle\eta_{i,k}\rangle\langle a^2_{i,k}\rangle}\right)}{\sqrt{2\eta_{i,k}}\mathcal{K}_{r_a-0.5}\left(\sqrt{2\langle\eta_{i,k}\rangle\langle a^2_{i,k}\rangle}\right)},\\
\langle\xi^{-1}_{i,k}\rangle &= \frac{\sqrt{2\langle\eta_{i,k}\rangle}\mathcal{K}_{r_a-0.5}\left(\sqrt{2\langle\eta_{i,k}\rangle\langle a^2_{i,k}\rangle}\right)}{\sqrt{\langle a^2_{i,k}\rangle}\mathcal{K}_{r_a-1.5}\left(\sqrt{2\langle\eta_{i,k}\rangle\rangle a^2_{i,k}\rangle}\right)},\\
\langle\eta_{i,k}\rangle &= \frac{r_a + s_a}{\langle\xi_{i,k}\rangle + \langle\Phi^{(a)}_k\rangle},\\
\langle\Phi^{(a)}_k\rangle &= \frac{0.5 + ds_a}{\langle\tilde{\Phi}^{(a)}\rangle + 0.5\sum_i\langle\eta_{i,k}\rangle},\\
\langle\tilde{\Phi}^{(a)}\rangle &= \frac{1}{1 + \sum_k\langle\Phi^{(a)}_k\rangle}.
\end{aligned}
$$

$\boldsymbol{\beta}$:

$$
\begin{aligned}
\langle\beta_k\rangle &= \langle\sigma^2_{\beta_k}\rangle \sum_{n=1}^{N}\Big\{y_n\langle z_{k,n}\rangle[\langle(\lambda^c_n)^{-1}\rangle + 1\\
&\qquad -\langle(\lambda^c_n)^{-1}\rangle y_n(\langle\boldsymbol{\beta}^\top\rangle\langle\mathbf{z}_n\rangle)_{\backslash k}]\Big\},\\
\langle\sigma^2_{\beta_k}\rangle &= \langle b^{-1}_k\rangle + \sum_{n=1}^{N}\langle z^2_{k,n}\rangle\langle(\lambda^c_n)^{-1}\rangle,\\
\langle\beta^2_k\rangle &= \langle\beta_k\rangle^2 + \langle\sigma^2_{\beta_k}\rangle,\\
\langle b_k\rangle &= \frac{\sqrt{\langle\beta^2_k\rangle}\mathcal{K}_{r_\beta+0.5}\left(\sqrt{2\langle e_k\rangle\langle\beta^2_k\rangle}\right)}{\sqrt{2e_k}\mathcal{K}_{r_\beta-0.5}\left(\sqrt{2\langle e_k\rangle\langle\beta^2_k\rangle}\right)},\\
\langle b^{-1}_k\rangle &= \frac{\sqrt{2e_k}\mathcal{K}_{r_\beta-0.5}\left(\sqrt{2\langle e_k\rangle\langle\beta^2_k\rangle}\right)}{\sqrt{\langle\beta^2_k\rangle}\mathcal{K}_{r_\beta-1.5}\left(\sqrt{2\langle e_k\rangle\langle\beta^2_k\rangle}\right)},\\
\langle e_k\rangle &= \frac{r_\beta + s_\beta}{\langle b_k\rangle + \langle\Phi^{(\beta)}_k\rangle},\\
\langle\Phi^{(\beta)}\rangle &= \frac{0.5 + 0.5s_\beta}{\langle\tilde{\Phi}^{(\beta)}\rangle + 0.5\sum_k\langle e_k\rangle},\\
\langle\tilde{\Phi}^{(\beta)}\rangle &= \frac{1}{\langle\Phi^{(\beta)}\rangle + 1}.
\end{aligned}
$$

4. Inferred Factor Loadings on the Handwritten Digits

We plotted the factor loadings $\mathbf{A}$ learned from USPS and MNIST datasets in Figures 2 and 1, respectively. Four models, G-L-BSVM, R-L-BSVM, G-NL-BSVM and R-NL-BSVM are used as examples. It can be be seen that the Gaussian model is trying to learn the dictionaries to reconstruct images while the rank model is trying to learning the differences (focusing on the edges).

5. Results on Gene Expression Data

We show the results of our model for gene expression data. $K = 20$ factors are used and here we only show the results generated by the proposed max-margin rank model without DP, *i.e.*, using linear Bayesian SVM as the classifier. Figure 3 shows the coefficients $\boldsymbol{\beta}$ of the learned classifiers and Figure 4 the inferred gene network from the learned factor loading matrix $\mathbf{A}$.

We list the top 200 genes of factors 14 , 9, 5, 4, 12, 7, 19, 10 in the following pages.

References

Blei, D. M. and Jordan, M. I. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

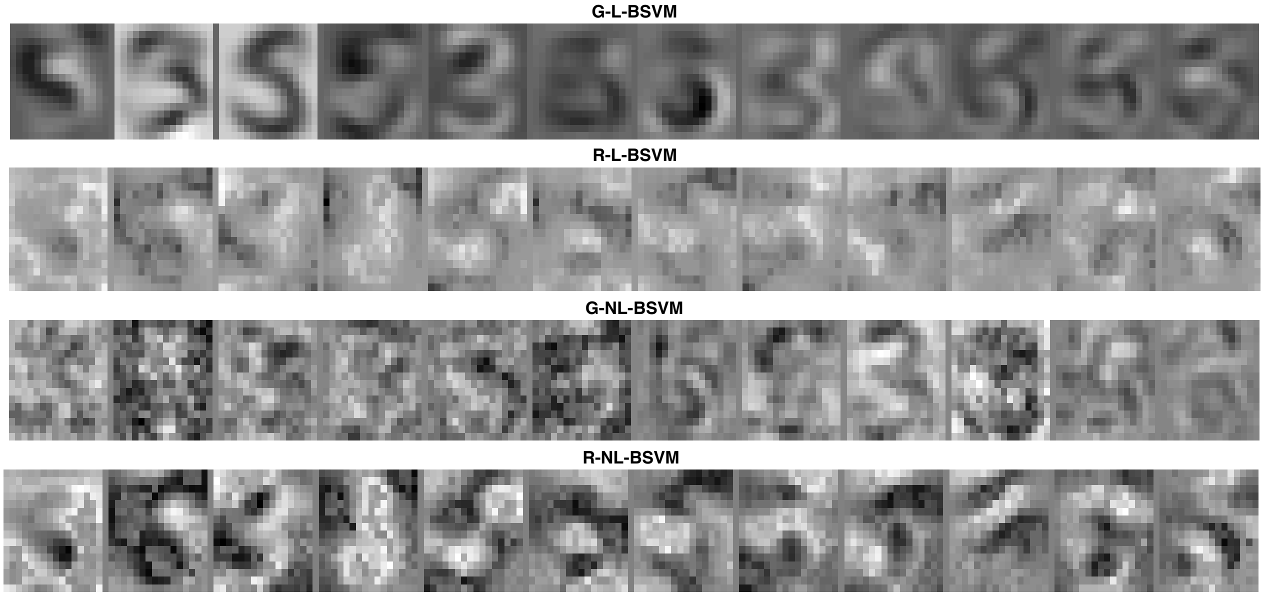

Figure 1. Inferred factor loading matrix $\mathbf{A}$ from USPS 3 vs. 5. The first 12 columns are reshaped and plotted.

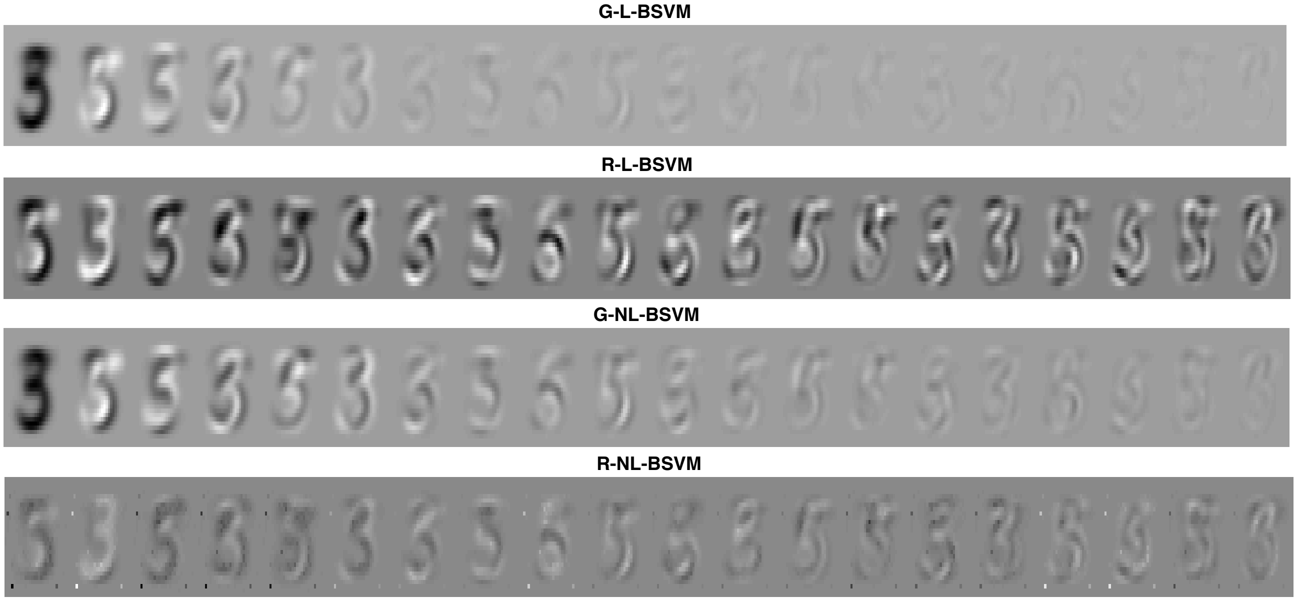

Figure 2. Inferred factor loading matrix $\mathbf{A}$ from MNIST 3 vs. 5. The 20 columns are reshaped and plotted.

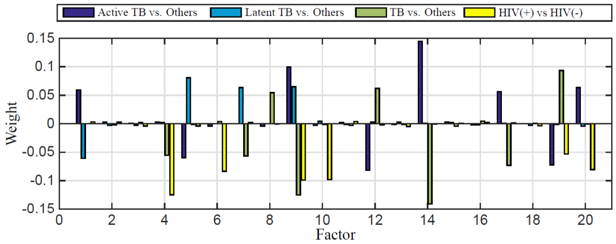

Figure 3. The learned classifier coefficients $\boldsymbol{\beta}$ of the 4 classifiers for the gene expression data.

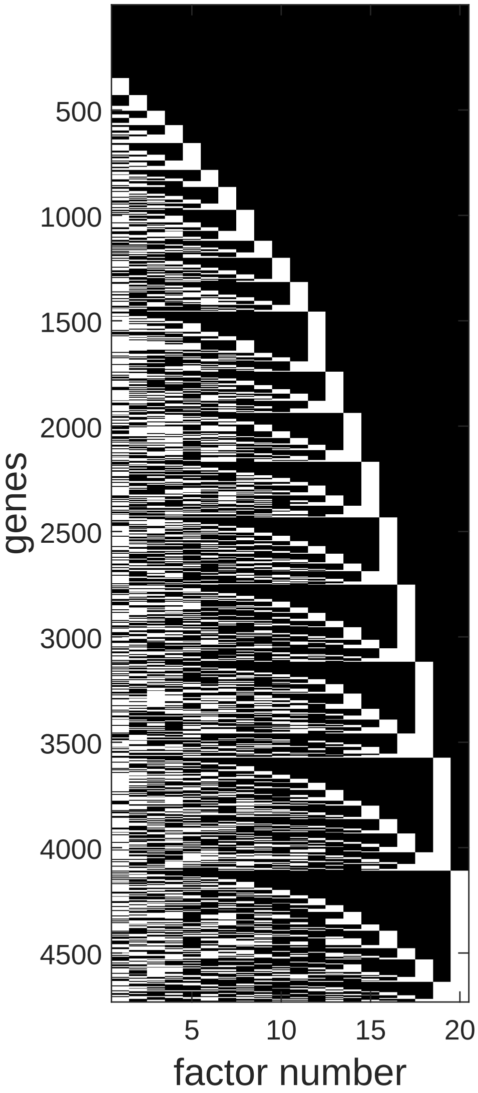

Figure 4. The learned gene network inferred from the factor loading matrix $\mathbf{A}$.