Diversity-Promoting Bayesian Learning of Latent Variable Models – Supplementary Material

Pengtao Xie[†] PENGTAOX@CS.CMU.EDU
Jun Zhu[†‡] DCSZJ@TSINGHUA.EDU.CN
Eric P. Xing[†] EPXING@CS.CMU.EDU
[†]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[‡]Dept. of Comp. Sci. & Tech., State Key Lab of Intell. Tech. & Sys., TNList, CBICR Center, Tsinghua University, China

1. Variational Inference for LVMs with Type I MABN Prior

In this section, we present details on how to derive the variational lower bound

$$\mathbb{E}_{q(\mathbf{A})}[\log p(\mathcal{D}|\mathbf{A})] + \mathbb{E}_{q(\mathbf{A})}[\log p(\mathbf{A})] - \mathbb{E}_{q(\mathbf{A})}[\log q(\mathbf{A})] \tag{1}$$

where the variational distribution $q(\mathbf{A})$ is chosen to be

$$\begin{aligned} & q(\mathbf{A}) = \textstyle\prod_{k=1}^{K} q(\tilde{\mathbf{a}}_k) q(g_k) \\ & = \prod_{k=1}^{K} \text{vMF}(\tilde{\mathbf{a}}_k | \hat{\mathbf{a}}_k, \hat{\kappa}) \text{Gamma}(g_k | r_k, s_k) \end{aligned} \tag{2}$$

Among the three expectation terms, $\mathbb{E}_{q(\mathbf{A})}[\log p(\mathbf{A})]$ and $\mathbb{E}_{q(\mathbf{A})}[\log q(\mathbf{A})]$ are model-independent and we discuss how to compute them in this section. $\mathbb{E}_{q(\mathbf{A})}[\log p(\mathcal{D}|\mathbf{A})]$ depends on the specific LVM and a concrete example will be given in Section 2.

First we introduce some equalities and inequalities used later on.

Let $\mathbf{a} \sim \text{vMF}(\boldsymbol{\mu}, \kappa)$, then
(I) $\mathbb{E}[\mathbf{a}] = A_p(\kappa)\boldsymbol{\mu}$ where $A_p(\kappa) = \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)}$, and $I_v(\cdot)$ denotes the modified Bessel function of the first kind at order v.
(II) $\text{cov}(\mathbf{a}) = \frac{h(\kappa)}{\kappa}\mathbf{I} + (1 - 2\frac{\nu+1}{\kappa}h(\kappa) - h^2(\kappa))\boldsymbol{\mu}\boldsymbol{\mu}^T$, where $h(\kappa) = \frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)}$ and $\nu = p/2 - 1$.
Please refer to (Abeywardana, 2015) for the derivation of $\mathbb{E}[\mathbf{a}]$ and $\text{cov}(\mathbf{a})$.
(III) $\mathbb{E}[\mathbf{a}^T\mathbf{a}] = \text{tr}(\text{cov}(\mathbf{a})) + A_p^2(\kappa)\boldsymbol{\mu}^T\boldsymbol{\mu}$.

Proof

$$\begin{aligned} & \mathbb{E}[\text{tr}(\mathbf{a}^T\mathbf{a})] = \mathbb{E}[\text{tr}(\mathbf{a}\mathbf{a}^T)] = \text{tr}(\mathbb{E}[\mathbf{a}\mathbf{a}^T]) \\ & = \text{tr}(\text{cov}(\mathbf{a}) + \mathbb{E}[\mathbf{a}]\mathbb{E}[\mathbf{a}]^T) = \text{tr}(\text{cov}(\mathbf{a})) + \text{tr}(\mathbb{E}[\mathbf{a}]\mathbb{E}[\mathbf{a}]^T) \\ & = \text{tr}(\text{cov}(\mathbf{a})) + A_p^2(\kappa)\boldsymbol{\mu}^T\boldsymbol{\mu} \end{aligned} \tag{3}$$

Let $g \sim \text{Gamma}(\alpha, \beta)$, then
(IV) $\mathbb{E}[g] = \frac{\alpha}{\beta}$
(V) $\mathbb{E}[\log g] = \psi(\alpha) - \log\beta$

(VI) $\log\sum_{k=1}^{K}\exp(x_k) \leq \gamma + \sum_{k=1}^{K}\log(1+\exp(x_k-\gamma))$, where γ is a variational parameter.

$\log\int\exp(x)\text{d}x \leq \gamma + \int\log(1+\exp(x-\gamma))\text{d}x$. See (Bouchard, 2007) for the proof.
(VII) $\log(1+e^{-x}) \leq \log(1+e^{-\xi}) - \frac{x-\xi}{2} - \frac{1/2-g(\xi)}{2\xi}(x^2-\xi^2)$, $\log(1+e^x) \leq \log(1+e^{\xi}) + \frac{x-\xi}{2} - \frac{1/2-g(\xi)}{2\xi}(x^2-\xi^2)$, where ξ is a variational parameter and $g(\xi) = 1/(1+\exp(-\xi))$. See (Bouchard, 2007) for the proof.
(VIII) $\int_{\|\mathbf{y}\|_2=1} 1\text{d}\mathbf{y} = \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}$, which is the surface area[1] of p-dimensional unit sphere. $\Gamma(\cdot)$ is the Gamma function.
(IX) $\int_{\|\mathbf{y}\|_2=1}\mathbf{x}^T\mathbf{y}\text{d}\mathbf{y} = 0$, which can be shown according to the symmetry of unit sphere.
(X) $\int_{\|\mathbf{y}\|_2=1}(\mathbf{x}^T\mathbf{y})^2\text{d}\mathbf{y} \leq \|\mathbf{x}\|_2^2\frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}$.

Proof

$$\begin{aligned} & \int_{\|\mathbf{y}\|_2=1}(\mathbf{x}^T\mathbf{y})^2\text{d}\mathbf{y} \\ = & \|\mathbf{x}\|_2^2\int_{\|\mathbf{y}\|_2=1}((\tfrac{\mathbf{x}}{\|\mathbf{x}\|_2})^T\mathbf{y})^2\text{d}\mathbf{y} \\ = & \|\mathbf{x}\|_2^2\int_{\|\mathbf{y}\|_2=1}(\mathbf{e}_1^T\mathbf{y})^2\text{d}\mathbf{y} \\ & \text{(according to the symmetry of unit sphere)} \\ \leq & \|\mathbf{x}\|_2^2\int_{\|\mathbf{y}\|_2=1}1\text{d}\mathbf{y} \\ = & \|\mathbf{x}\|_2^2\frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})} \end{aligned} \tag{4}$$

Given these equalities and inequalities, we can prove Lemma 1 given in the main paper.

[1] https://en.wikipedia.org/wiki/N-sphere

Proof

$$
\begin{aligned}
&\log Z_i \\
&= \log \int \exp(\kappa(-\textstyle\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j) \cdot \tilde{\mathbf{a}}_i) \mathrm{d}\tilde{\mathbf{a}}_i \\
&\leq \gamma + \int \log(1 + \exp(\kappa(-\textstyle\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j) \cdot \tilde{\mathbf{a}}_i - \gamma) \mathrm{d}\tilde{\mathbf{a}}_i \text{(apply (VI))} \\
&\leq \gamma + \int [\log(1 + e^{-\xi}) - \frac{\kappa(\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j) \cdot \tilde{\mathbf{a}}_i + \gamma - \xi}{2} \\
&- \frac{1/2 - g(\xi)}{2\xi}((\kappa(\textstyle\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j) \cdot \tilde{\mathbf{a}}_i + \gamma)^2 - \xi^2)] \mathrm{d}\tilde{\mathbf{a}}_i \text{(apply (VII))} \\
&\leq \gamma + [\log(1 + e^{-\xi}) + \frac{\xi - \gamma}{2} + \frac{1/2 - g(\xi)}{2\xi}(\xi^2 - \gamma^2)] \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})} \\
&- \frac{1/2 - g(\xi)}{2\xi} \kappa^2 \| \textstyle\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j \|_2^2 \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})} \text{(apply (VIII,IX,X))}
\end{aligned} \tag{5}
$$

Given Lemma 1, we can derive a lower bound of $\mathbb{E}_{q(\mathbf{A})}[\log p(\mathbf{A})]$

$$
\begin{aligned}
&\mathbb{E}_{q(\mathbf{A})}[\log p(\mathbf{A})] \\
&= \mathbb{E}_{q(\mathbf{A})}[\log p(\tilde{\mathbf{a}}_1) \textstyle\prod_{i=2}^{K} p(\tilde{\mathbf{a}}_i | \{\tilde{\mathbf{a}}_j\}_{j=1}^{i-1}) \prod_{i=1}^{K} q(g_i)] \\
&= \mathbb{E}_{q(\mathbf{A})}[\log p(\tilde{\mathbf{a}}_1) \prod_{i=2}^{K} \frac{\exp(\kappa(-\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j) \cdot \tilde{\mathbf{a}}_i)}{Z_i} \prod_{i=1}^{K} \frac{\alpha_2^{\alpha_1} g_i^{\alpha_1 - 1} e^{-g_i \alpha_2}}{\Gamma(\alpha_1)}] \\
&\geq \kappa \mu_0^{\mathsf{T}} \mathbb{E}_{q(\tilde{\mathbf{a}}_1)}[\tilde{\mathbf{a}}_1] + \textstyle\sum_{i=2}^{K}(-\kappa \sum_{j=1}^{i-1} \mathbb{E}_{q(\tilde{\mathbf{a}}_j)}[\tilde{\mathbf{a}}_j] \cdot \mathbb{E}_{q(\tilde{\mathbf{a}}_i)}[\tilde{\mathbf{a}}_i] \\
&- \gamma_i - (\log(1 + e^{-\xi_i}) + \frac{\xi_i - \gamma_i}{2} + \frac{1/2 - g(\xi_i)}{2\xi_i}(\xi_i^2 - \gamma_i^2)) \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})} \\
&+ \frac{1/2 - g(\xi_i)}{2\xi_i} \kappa^2 \mathbb{E}_{q(\mathbf{A})}[\| \textstyle\sum_{j=1}^{i-1} \tilde{\mathbf{a}}_j \|_2^2] \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}) + K(\alpha_1 \log \alpha_2 \\
&- \log \Gamma(\alpha_1)) + \sum_{i=1}^{K}(\alpha_1 - 1) \mathbb{E}_{q(g_i)}[\log g_i] - \alpha_2 \mathbb{E}_{q(g_i)}[g_i] + \text{const} \\
&\geq \kappa A_p(\hat{\kappa}) \mu_0^{\mathsf{T}} \hat{\mathbf{a}}_1 + \textstyle\sum_{i=2}^{K}(-\kappa A_p(\hat{\kappa})^2 \sum_{j=1}^{i-1} \hat{\mathbf{a}}_j \cdot \hat{\mathbf{a}}_i - \gamma_i \\
&- (\log(1 + e^{-\xi_i}) + \frac{\xi_i - \gamma_i}{2} + \frac{1/2 - g(\xi_i)}{2\xi_i}(\xi_i^2 - \gamma_i^2)) \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})} \\
&+ \frac{1/2 - g(\xi_i)}{2\xi_i} \kappa^2 (A_p^2(\hat{\kappa}) \textstyle\sum_{j=1}^{i-1} \sum_{k \neq j}^{i-1} \hat{\mathbf{a}}_j \cdot \hat{\mathbf{a}}_k + \sum_{j=1}^{i-1} (\mathrm{tr}(\Lambda_j) \\
&+ A_p^2(\hat{\kappa}) \hat{\mathbf{a}}_j^T \hat{\mathbf{a}}_j)) \frac{2\pi^{(p+1)/2}}{\Gamma(\frac{p+1}{2})}) + K(\alpha_1 \log \alpha_2 - \log \Gamma(\alpha_1)) \\
&+ \textstyle\sum_{i=1}^{K}(\alpha_1 - 1)(\psi(r_i) - \log(s_i)) - \alpha_2 \frac{r_i}{s_i} + \text{const}
\end{aligned} \tag{6}
$$

where $\Lambda_j = \frac{h(\hat{\kappa})}{\hat{\kappa}} \mathbf{I} + (1 - 2\frac{\nu+1}{\hat{\kappa}} h(\hat{\kappa}) - h^2(\hat{\kappa})) \hat{\mathbf{a}}_j \hat{\mathbf{a}}_j^T$.

The other expectation term $\mathbb{E}_{q(\mathbf{A})}[\log q(\mathbf{A})]$ can be computed as

$$
\begin{aligned}
&\mathbb{E}_{q(\mathbf{A})}[\log q(\mathbf{A})] \\
&= \mathbb{E}_{q(\mathbf{A})}[\log \prod_{k=1}^{K} \mathrm{vMF}(\tilde{\mathbf{a}}_k | \hat{\mathbf{a}}_k, \hat{\kappa}) \mathrm{Gamma}(g_k | r_k, s_k)] \\
&= \sum_{k=1}^{K} \hat{\kappa} A_p(\hat{\kappa}) \| \hat{\alpha}_k \|_2^2 + r_k \log s_k - \log \Gamma(r_k) \\
&\quad + (r_k - 1)(\psi(r_k) - \log(s_k)) - r_k
\end{aligned} \tag{7}
$$

2. VI for BMEM with Type I MABN

In this section, we discuss how to derive the variational lower bound for BMEM with type I MABN. The latent variable are $\{\boldsymbol{\beta}_k\}_{k=1}^K$,$\{\boldsymbol{\eta}_k\}_{k=1}^K$,$\{z_n\}_{n=1}^N$. The joint probability of all variables is

$$
\begin{aligned}
&p(\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n, y_n, z_n\}_{n=1}^N) \\
&= p(\{y_n\}_{n=1}^N | \{\mathbf{x}_n\}_{n=1}^N, \{z_n\}_{n=1}^N, \{\boldsymbol{\beta}_k\}_{k=1}^K) \\
&\quad p(\{z_n\}_{n=1}^N | \{\mathbf{x}_n\}_{n=1}^N, \{\boldsymbol{\eta}_k\}_{k=1}^K) p(\{\boldsymbol{\beta}_k\}_{k=1}^K) p(\{\boldsymbol{\eta}_k\}_{k=1}^K)
\end{aligned} \tag{8}
$$

To perform variational inference, we employ a mean field variational distribution

$$
\begin{aligned}
Q &= q(\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{z_n\}_{n=1}^N) \\
&= \prod_{k=1}^{K} q(\boldsymbol{\beta}_k) q(\boldsymbol{\eta}_k) \textstyle\prod_{n=1}^{N} q(z_n) \\
&= \prod_{k=1}^{K} \mathrm{vMF}(\tilde{\boldsymbol{\beta}}_k | \hat{\boldsymbol{\beta}}_k, \hat{\kappa}) \mathrm{Gamma}(g_k | r_k, s_k) \mathrm{vMF}(\tilde{\boldsymbol{\eta}}_k | \hat{\boldsymbol{\eta}}_k, \hat{\kappa}) \\
&\quad \mathrm{Gamma}(h_k | t_k, u_k) \prod_{n=1}^{N} q(z_n | \boldsymbol{\phi}_n)
\end{aligned} \tag{9}
$$

Accordingly, the variational lower bound is

$$
\begin{aligned}
&\mathbb{E}_Q[\log p(\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n, y_n, z_n\}_{n=1}^N)] \\
&- \mathbb{E}_Q[\log q(\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{z_n\}_{n=1}^N)] \\
&= \mathbb{E}_Q[\log p(\{y_n\}_{n=1}^N | \{\mathbf{x}_n\}_{n=1}^N, \{z_n\}_{n=1}^N, \{\boldsymbol{\beta}_k\}_{k=1}^K)] \\
&+ \mathbb{E}_Q[\log p(\{z_n\}_{n=1}^N | \{\mathbf{x}_n\}_{n=1}^N, \{\boldsymbol{\eta}_k\}_{k=1}^K)] \\
&+ \mathbb{E}_Q[\log p(\{\boldsymbol{\beta}_k\}_{k=1}^K)]) + \mathbb{E}_Q[\log p(\{\boldsymbol{\eta}_k\}_{k=1}^K)]) \\
&- \mathbb{E}_Q[\log q(\{\boldsymbol{\beta}_k\}_{k=1}^K)] - \mathbb{E}_Q[\log q(\{\boldsymbol{\eta}_k\}_{k=1}^K)] \\
&- \mathbb{E}_Q[\log q(\{z_n\}_{n=1}^N)]
\end{aligned} \tag{10}
$$

where $\mathbb{E}_Q[\log p(\{\boldsymbol{\beta}_k\}_{k=1}^K)]$ and $\mathbb{E}_Q[\log p(\{\boldsymbol{\eta}_k\}_{k=1}^K)]$ can be lower bounded in a similar way as that in Eq.(6). $\mathbb{E}_Q[\log q(\{\boldsymbol{\beta}_k\}_{k=1}^K)]$ and $\mathbb{E}_Q[\log q(\{\boldsymbol{\eta}_k\}_{k=1}^K)]$ can be computed in a similar manner as that in Eq.(7). Next we discuss how to compute the remaining expectation terms.

Compute $\mathbb{E}_Q[\log p(\{z_n\}_{n=1}^N | \{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n\}_{n=1}^N)]$

First, $p(\{z_n\}_{n=1}^N | \{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n\}_{n=1}^N)$ is defined as

$$
\begin{aligned}
&p(\{z_n\}_{n=1}^N | \{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n\}_{n=1}^N) \\
&= \prod_{n=1}^{N} p(z_n | \mathbf{x}_n, \{\boldsymbol{\eta}_k\}_{k=1}^K) \\
&= \prod_{n=1}^{N} \frac{\prod_{k=1}^{K} [\exp(\boldsymbol{\eta}_k^{\mathsf{T}} \mathbf{x}_n)]^{z_{nk}}}{\sum_{j=1}^{K} \exp(\boldsymbol{\eta}_j^{\mathsf{T}} \mathbf{x}_n)}
\end{aligned} \tag{11}
$$

$\log p(\{z_n\}_{n=1}^N|\{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n\}_{n=1}^N)$ can be lower bounded as

$$\begin{aligned}
&\log p(\{z_n\}_{n=1}^N|\{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n\}_{n=1}^N) \\
&= \sum_{n=1}^N \sum_{k=1}^K z_{nk}\boldsymbol{\eta}_k^\mathsf{T}\mathbf{x}_n - \log(\textstyle\sum_{j=1}^K \exp(\boldsymbol{\eta}_j^\mathsf{T}\mathbf{x}_n)) \\
&= \sum_{n=1}^N \sum_{k=1}^K z_{nk}h_k\tilde{\boldsymbol{\eta}}_k^\mathsf{T}\mathbf{x}_n - \log(\textstyle\sum_{j=1}^K \exp(\boldsymbol{\eta}_j^\mathsf{T}\mathbf{x}_n)) \\
&\geq \sum_{n=1}^N \sum_{k=1}^K z_{nk}h_k\tilde{\boldsymbol{\eta}}_k^\mathsf{T}\mathbf{x}_n - c_n - \textstyle\sum_{j=1}^K \\
&\quad \log(1+\exp(\boldsymbol{\eta}_j^\mathsf{T}\mathbf{x}_n - c_n)) \text{(Using Inequality VI)} \\
&\geq \sum_{n=1}^N \sum_{k=1}^K z_{nk}h_k\tilde{\boldsymbol{\eta}}_k^\mathsf{T}\mathbf{x}_n - c_n - \textstyle\sum_{j=1}^K [\log(1+e^{-d_{nj}}) \\
&\quad -\frac{c_n-\boldsymbol{\eta}_j^\mathsf{T}\mathbf{x}_n-d_{nj}}{2} - \frac{1/2-g(d_{nj})}{2d_{nj}}((\boldsymbol{\eta}_j^\mathsf{T}\mathbf{x}_n - c_n)^2 - d_{nj}^2)] \\
&\quad \text{(Using Inequality VII)}
\end{aligned} \tag{12}$$

The expectation of $\log p(\{z_n\}_{n=1}^N|\{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n\}_{n=1}^N)$ can be lower bounded as

$$\begin{aligned}
&\mathbb{E}[\log p(\{z_n\}_{n=1}^N|\{\boldsymbol{\eta}_k\}_{k=1}^K, \{\mathbf{x}_n\}_{n=1}^N)] \\
&= A_p(\hat{\kappa}) \sum_{n=1}^N \sum_{k=1}^K \phi_{nk}\frac{t_k}{u_k}\hat{\boldsymbol{\eta}}_k^\mathsf{T}\mathbf{x}_n - c_n - \textstyle\sum_{j=1}^K [\log(1+e^{-d_{nj}}) \\
&-\frac{c_n - A_p(\hat{\kappa})\frac{t_j}{u_j}\hat{\boldsymbol{\eta}}_j^\mathsf{T}\mathbf{x}_n - d_{nj}}{2} - \frac{1/2-g(d_{nj})}{2d_{nj}}(\frac{t_j+t_j^2}{u_j^2}\mathbb{E}[\tilde{\boldsymbol{\eta}}_j^\mathsf{T}\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\tilde{\boldsymbol{\eta}}_j] \\
&-2c_nA_p(\hat{\kappa})\frac{t_j}{u_j}\hat{\boldsymbol{\eta}}_j^\mathsf{T}\mathbf{x}_n + c_n^2 - d_{nj}^2)]
\end{aligned} \tag{13}$$

where

$$\begin{aligned}
&\mathbb{E}[\tilde{\boldsymbol{\eta}}_k^\mathsf{T}\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\tilde{\boldsymbol{\eta}}_k] \\
&= \mathbb{E}[\mathrm{tr}(\tilde{\boldsymbol{\eta}}_k^\mathsf{T}\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\tilde{\boldsymbol{\eta}}_k)] \\
&= \mathbb{E}[\mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\tilde{\boldsymbol{\eta}}_k\tilde{\boldsymbol{\eta}}_k^\mathsf{T})] \\
&= \mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\mathbb{E}[\tilde{\boldsymbol{\eta}}_k\tilde{\boldsymbol{\eta}}_k^\mathsf{T}]) \\
&= \mathrm{tr}(\mathbf{x}_n\mathbf{x}_n^\mathsf{T}(\mathbb{E}[\tilde{\boldsymbol{\eta}}_k]\mathbb{E}[\tilde{\boldsymbol{\eta}}_k]^T + \mathrm{cov}(\tilde{\boldsymbol{\eta}}_k)))
\end{aligned} \tag{14}$$

Compute $\mathbb{E}[\log p(\{y_n\}_{n=1}^N|\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\mathbf{z}_n\}_{n=1}^N)]$

$p(\{y_n\}_{n=1}^N|\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\mathbf{z}_n\}_{n=1}^N)$ is defined as

$$\begin{aligned}
&p(\{y_n\}_{n=1}^N|\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\mathbf{z}_n\}_{n=1}^N) \\
&= \prod_{n=1}^N p(y_n|\mathbf{z}_n, \{\boldsymbol{\beta}_k\}_{k=1}^K) \\
&= \prod_{n=1}^N \frac{1}{\prod_{k=1}^K [1+\exp(-(2y_n-1)\boldsymbol{\beta}_k^\mathsf{T}\mathbf{x}_n)]^{z_{nk}}}
\end{aligned} \tag{15}$$

$\log p(\{y_n\}_{n=1}^N|\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\mathbf{z}_n\}_{n=1}^N)$ can be lower bounded by

$$\begin{aligned}
&\log p(\{y_n\}_{n=1}^N|\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\mathbf{z}_n\}_{n=1}^N) \\
&= -\sum_{n=1}^N \sum_{k=1}^K z_{nk}\log(1+\exp(-(2y_n-1)\boldsymbol{\beta}_k^\mathsf{T}\mathbf{x}_n)) \\
&\geq \sum_{n=1}^N \sum_{k=1}^K z_{nk}[-\log(1+e^{-e_{nk}}) + \frac{(2y_n-1)\boldsymbol{\beta}_k^\mathsf{T}\mathbf{x}_n - e_{nk}}{2} \\
&\quad +\frac{1/2-g(e_{nk})}{2e_{nk}}((\boldsymbol{\beta}_k^\mathsf{T}\mathbf{x}_n)^2 - e_{nk}^2)]
\end{aligned} \tag{16}$$

$\mathbb{E}[\log p(\{y_n\}_{n=1}^N|\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\mathbf{z}_n\}_{n=1}^N)]$ can be lower bounded by

$$\begin{aligned}
&\mathbb{E}[\log p(\{y_n\}_{n=1}^N|\{\boldsymbol{\beta}_k\}_{k=1}^K, \{\mathbf{z}_n\}_{n=1}^N)] \\
&\geq \sum_{n=1}^N \sum_{k=1}^K \phi_{nk}[-\log(1+e^{-e_{nk}}) + \frac{A_p(\hat{\kappa})\frac{r_k}{s_k}\hat{\boldsymbol{\beta}}_k^\mathsf{T}\mathbf{x}_n - e_{nk}}{2} \\
&\quad +\frac{1/2-\sigma(e_{nk})}{2e_{nk}}(\frac{r_k+r_k^2}{s_k^2}\mathbb{E}[\tilde{\boldsymbol{\beta}}_k^\mathsf{T}\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\tilde{\boldsymbol{\beta}}_k] - e_{nk}^2)] \\
&\quad \text{(Using Inequality VII)}
\end{aligned} \tag{17}$$

where $\mathbb{E}[\tilde{\boldsymbol{\beta}}_k^\mathsf{T}\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\tilde{\boldsymbol{\beta}}_k]$ can be computed in a similar way to Eq.(14).

Compute $\mathbb{E}[\log q(z_i)]$

$$\mathbb{E}[\log q(z_i)] = \sum_{k=1}^K \phi_{ik}\log\phi_{ik} \tag{18}$$

In the end, we can get a lower bound of the variational lower bound, then learn all the parameters by optimizing the lower bound via coordinate ascent: In each iteration, we pick up a parameter x and fix all other parameters, which leads to a sub-problem defined over x. Then we optimize the sub-problem w.r.t x. For some parameters, the optimal solution of the sub-problem is in closed form. If not the case, we optimize x using gradient ascent method. This process iterates until convergence. We omit the detailed derivation here since it only involves basic algebra and calculus, which can be done straightforwardly.

3. Additional Details of the Metropolis-Hastings Algorithm

Parameter Learning The mutual angular Bayesian Network (MABN) prior is parametrized by several deterministic parameters including κ, $\boldsymbol{\mu}_0$, α_1, α_2. Among them, we tune κ manually via cross validation and learn the others via an Expectation Maximization (EM) framework. Let $\mathbf{x}$ denote observed data, $\mathbf{z}$ denote all random variables and $\boldsymbol{\theta}$ denote deterministic parameters $\{\boldsymbol{\mu}_0, \alpha_1, \alpha_2\}$. EM is an algorithm aiming to learn $\boldsymbol{\theta}$ by maximize log-likelihood $p(\mathbf{x};\boldsymbol{\theta})$ of data. It iteratively performs two steps until convergence. In the E step, the posterior $p(\mathbf{z}|\mathbf{x})$ is inferred with parameters $\boldsymbol{\theta}$ fixed. In the M step, $\boldsymbol{\theta}$ is learned by optimizing a lower bound of the log-likelihood $\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})]$, where the expectation is computed w.r.t the posterior $p(\mathbf{z}|\mathbf{x})$ inferred in the E step. In our problem, we use the Metropolis-Hastings (MH) algorithm to infer the posterior $p(\mathbf{x};\boldsymbol{\theta})$ at the E step, and learn parameters $\{\boldsymbol{\mu}_0, \alpha_1, \alpha_2\}$ at the M step. The parameters $\hat{\kappa}$ and σ in proposal distributions are set manually.

Truncated Sampler for Magnitude Variable g The magnitude variable g is required to be positive, but the

proposal distribution $q(g^{(t+1)}|g^t)$ is Gaussian, which can generate non-positive values. To address this problem, we adopt a truncated sampler (Wilkinson, 2015) which repeatedly draws samples from the proposal until a positive value is obtained. Under such a truncated sampling scheme, the MH acceptance ratio needs to be modified accordingly. Please refer to (Wilkinson, 2015) for details.

4. Algorithm for Posterior Regularization of BMEM

In this section, we present the algorithmic details of posterior regularization of BMEM. Recall the problem is

$$\begin{aligned}\sup_{q(\mathbf{B},\mathbf{H},\mathbf{z})} \quad & \mathbb{E}_{q(\mathbf{B},\mathbf{H},\mathbf{z})}[\log p(\{y_i\}_{i=1}^N, \mathbf{z}|\mathbf{B},\mathbf{H})\pi(\mathbf{B},\mathbf{H})] \\ & -\mathbb{E}_{q(\mathbf{B},\mathbf{H},\mathbf{z})}[\log q(\mathbf{B},\mathbf{H},\mathbf{z})] \\ & +\lambda_1\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\beta}}_k)}[\tilde{\boldsymbol{\beta}}_k]\}_{k=1}^K) \\ & +\lambda_2\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\eta}}_k)}[\tilde{\boldsymbol{\eta}}_k]\}_{k=1}^K)\end{aligned} \tag{19}$$

where $\mathbf{B} = \{\boldsymbol{\beta}_k\}_{k=1}^K$, $\mathbf{H} = \{\boldsymbol{\eta}_k\}_{k=1}^K$ and $\mathbf{z} = \{z_i\}_{i=1}^N$ are latent variables and the post-data distribution over them is defined as $q(\mathbf{B},\mathbf{H},\mathbf{z}) = q(\mathbf{B})q(\mathbf{H})q(\mathbf{z})$. For computational tractability, we define $q(\mathbf{B})$ and $q(\mathbf{H})$ to be: $q(\mathbf{B}) = \prod_{k=1}^K q(\tilde{\boldsymbol{\beta}}_k)q(g_k)$ and $q(\mathbf{H}) = \prod_{k=1}^K q(\tilde{\boldsymbol{\eta}}_k)q(h_k)$ where $q(\tilde{\boldsymbol{\beta}}_k)$, $q(\tilde{\boldsymbol{\eta}}_k)$ are von-Mises Fisher distributions and $q(g_k)$, $q(h_k)$ are gamma distributions, and define $q(\mathbf{z})$ to be multinomial distributions: $q(\mathbf{z}) = \prod_{i=1}^N q(z_i|\boldsymbol{\phi}_i)$ where $\boldsymbol{\phi}_i$ is a multinomial vector. The priors over $\mathbf{B}$ and $\mathbf{H}$ are specified to be: $\pi(\mathbf{B}) = \prod_{k=1}^K p(\tilde{\boldsymbol{\beta}}_k)p(g_k)$ and $\pi(\mathbf{H}) = \prod_{k=1}^K p(\tilde{\boldsymbol{\eta}}_k)p(h_k)$ where $p(\tilde{\boldsymbol{\beta}}_k)$, $p(\tilde{\boldsymbol{\eta}}_k)$ are von-Mises Fisher distributions and $p(g_k)$, $p(h_k)$ are gamma distributions.

The objective in Eq.(19) can be further written as

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{B},\mathbf{H},\mathbf{z})}[\log p(\{y_i\}_{i=1}^N, \mathbf{z}|\mathbf{B},\mathbf{H})\pi(\mathbf{B},\mathbf{H})] \\ &-\mathbb{E}_{q(\mathbf{B},\mathbf{H},\mathbf{z})}[\log q(\mathbf{B},\mathbf{H},\mathbf{z})] + \lambda_1\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\beta}}_k)}[\tilde{\boldsymbol{\beta}}_k]\}_{k=1}^K) \\ &+\lambda_2\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\eta}}_k)}[\tilde{\boldsymbol{\eta}}_k]\}_{k=1}^K) \\ &= \mathbb{E}_{q(\mathbf{B},\mathbf{z})}[\log p(\{y_i\}_{i=1}^N|\mathbf{z},\mathbf{B})] + \mathbb{E}_{q(\mathbf{H},\mathbf{z})}[\log p(\mathbf{z}|\mathbf{H})] \\ &+\mathbb{E}_{q(\mathbf{H})}[\log \pi(\mathbf{H})] + \mathbb{E}_{q(\mathbf{B})}[\log \pi(\mathbf{B})] - \mathbb{E}_{q(\mathbf{B})}[\log q(\mathbf{B})] \\ &-\mathbb{E}_{q(\mathbf{H})}[\log q(\mathbf{H})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] + \lambda_1\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\beta}}_k)}[\tilde{\boldsymbol{\beta}}_k]\}_{k=1}^K) \\ &+\lambda_2\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\eta}}_k)}[\tilde{\boldsymbol{\eta}}_k]\}_{k=1}^K)\end{aligned} \tag{20}$$

Among these expectation terms, $\mathbb{E}_{q(\mathbf{B},\mathbf{z})}[\log p(\{y_i\}_{i=1}^N|\mathbf{z},\mathbf{B})]$ can be computed via Eq.(15-17), $\mathbb{E}_{q(\mathbf{H},\mathbf{z})}[\log p(\mathbf{z}|\mathbf{H})]$ can be computed via Eq.(11-14). $\mathbb{E}_{q(\mathbf{H})}[\log \pi(\mathbf{H})]$, $\mathbb{E}_{q(\mathbf{B})}[\log \pi(\mathbf{B})]$, $\mathbb{E}_{q(\mathbf{B})}[\log q(\mathbf{B})]$, $\mathbb{E}_{q(\mathbf{H})}[\log q(\mathbf{H})]$ can be computed in a way similar to Eq.(7). $\mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})]$ can be computed via Eq.(18). Given all these expectations, we can get an analytical expression of the objective in Eq.(19) and learn the parameters by optimizing this objective. Regarding how to optimize the mutual angular regularizers $\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\beta}}_k)}[\tilde{\boldsymbol{\beta}}_k]\}_{k=1}^K)$ and $\Omega(\{\mathbb{E}_{q(\tilde{\boldsymbol{\eta}}_k)}[\tilde{\boldsymbol{\eta}}_k]\}_{k=1}^K)$, please refer to (Xie et al., 2015) for details.

References

Abeywardana, Sachin. Expectation and covariance of von mises-fisher distribution. In *https://sachinruk.github.io/blog/von-Mises-Fisher/*, 2015.

Bouchard, Guillaume. Efficient bounds for the softmax function, applications to inference in hybrid models. 2007.

Wilkinson, Darren. Metropolis hastings mcmc when the proposal and target have differing support. In *https://darrenjw.wordpress.com/2012/06/04/metropolis-hastings-mcmc-when-the-proposal-and-target-have-differing-support/*, 2015.

Xie, Pengtao, Deng, Yuntian, and Xing, Eric P. Diversifying restricted boltzmann machine for document modeling. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.