Homework 2 (Lecture: Bayesian Learning)

Solve Exercises 1–10, and submit an answer sheet to MAR4.034 by 18.2.2016. Exercises 1–5 (70P) are mandatory. For Exercises 6-10, you can choose one of the following.

- Exercises 6 (15P) and 7 (15P).
- Exercises 8 (20P) and 9 (10P).
- Exercises 10 (30P).

Answering more than one can be a plus. (Do more than one if you think your score in the first homework is not sufficient, or you want a better score/grade.)

1 Notation

We use the following notation for basic distribution (density) function.

$$\mathrm{Norm}_M(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right)}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}}, \tag{1}$$

$$\mathrm{Gamma}(x; \alpha, \beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x), \tag{2}$$

$$\mathrm{Wishart}_M(\boldsymbol{X}; \boldsymbol{V}, \nu) \equiv \frac{|\boldsymbol{X}|^{\frac{\nu-M-1}{2}}\exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)}, \tag{3}$$

$$\mathrm{Multinomial}_{K,N}(\boldsymbol{x}; \boldsymbol{\theta}) \equiv N! \prod_{k=1}^{K} (x_k!)^{-1}\theta_k^{x_k}, \tag{4}$$

$$\mathrm{Dirichlet}_K(\boldsymbol{x}; \boldsymbol{\phi}) \equiv \frac{\Gamma(\sum_{k=1}^{K}\phi_k)}{\prod_{k=1}^{K}\Gamma(\phi_k)} \prod_{k=1}^{K} x_k^{\phi_{k-1}}. \tag{5}$$

Moments of them can be found in the pdf file for the introduction of this lecture (BayesianLearningIntroduction.pdf).

2 Conditional Conjugacy

Assume that the parameters $\boldsymbol{w}$ of a model likelihood $p(\mathcal{D}|\boldsymbol{w})$ can be divided into S groups $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_S)$, so that the following *conditional conjugacy* holds: There exists a decomposable prior

$$p(\boldsymbol{w}) = \prod_{s=1}^{S} p(\boldsymbol{w}_s)$$

such that, for any $s = 1, \ldots, S$, the posterior

$$p(\boldsymbol{w}_s | \{\boldsymbol{w}_{s'}\}_{s' \neq s}, \mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w}) p(\boldsymbol{w}_s)}{p(\mathcal{D}|\{\boldsymbol{w}_{s'}\}_{s' \neq s})}$$

on $\boldsymbol{w}_s$ given the other parameters $\{\boldsymbol{w}_{s'}\}_{s' \neq s}$ fixed as constants is in the same distribution family as the prior $p(\boldsymbol{w}_s)$. We also assume that computing moments of, as well as drawing samples from, this family is not hard.

3 Variational Bayesian (VB) learning

Let r be an approximate posterior distribution, and define the free energy by

$$F(r) = \left\langle \log \frac{r(\boldsymbol{w})}{p(\mathcal{D}|\boldsymbol{w}) p(\boldsymbol{w})} \right\rangle_{r(\boldsymbol{w})}. \tag{6}$$

Exercise 1 (10P): Prove that minimizing the free energy (6) amounts to minimizing the Kullback Leibler divergence to the Bayes posterior.

Under the assumed conditional conjugacy, variational Bayesian learning solves the following problem:

$$\min_r F(r) \quad \text{s.t.} \quad r(\boldsymbol{w}) = \prod_{s=1}^{S} r_s(\boldsymbol{w}_s). \tag{7}$$

Exercise 2 (20P): Derive the following general update rule for variational Bayesian learning:

$$r_s(\boldsymbol{w}_s) \propto p(\boldsymbol{w}_s) \exp \langle \log p(\mathcal{D}|\boldsymbol{w}) \rangle_{\prod_{s' \neq s} r_{s'}(\boldsymbol{w}_{s'})}. \tag{8}$$

4 Gibbs sampling

Gibbs sampling simply iterates the following sampling process for $s = 1, \ldots, S$ in turn to get a Markov chain:

$$\text{Draw } \boldsymbol{w}_s^{\text{new}} \text{ from } p(\boldsymbol{w}_s | \{\boldsymbol{w}_{s'}^{\text{old}}\}_{s' \neq s}, \mathcal{D}).$$

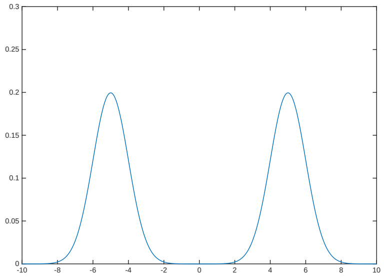

Figure 1: A distribution with two modes.

5 Difference between variational Bayes (VB) and expectation propagation (EP)

Variational Bayes minimizes

$$\mathrm{KL}\left(r(\boldsymbol{w})||p(\boldsymbol{w}|\mathcal{D})\right) = \int r(\boldsymbol{w}) \log \frac{r(\boldsymbol{w})}{p(\boldsymbol{w}|\mathcal{D})} d\boldsymbol{w}, \tag{9}$$

while expectation propagation minimizes

$$\mathrm{KL}\left(p(\boldsymbol{w}|\mathcal{D})||r(\boldsymbol{w})\right) = \int p(\boldsymbol{w}|\mathcal{D}) \log \frac{p(\boldsymbol{w}|\mathcal{D})}{r(\boldsymbol{w})} d\boldsymbol{w}. \tag{10}$$

Exercise 3 (10P): Assume that the Bayes posterior has two modes like in Fig.1. It is observed that the variational Bayes posterior tends to cover one of the modes, while the expectation propagation tends to cover both modes with a broader distribution. Explain intuitively (and qualitatively) why this tendency is observed.

6 Matrix factorization

Consider the matrix factorization model

$$\begin{aligned} p(\boldsymbol{V}|\boldsymbol{A},\boldsymbol{B}) &= \prod_{l=1}^{L}\prod_{m=1}^{M} \mathrm{Norm}_1(V_{l,m}; \widetilde{\boldsymbol{b}}_l^{\top} \widetilde{\boldsymbol{a}}_m, \sigma^2) \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{V} - \boldsymbol{B}\boldsymbol{A}^{\top}\|_{\mathrm{Fro}}^2\right)}{(2\pi\sigma^2)^{LM/2}}, \qquad (11) \\ p(\boldsymbol{A}) &= \prod_{m=1}^{M} \mathrm{Norm}_H(\widetilde{\boldsymbol{a}}_m; \boldsymbol{0}, \boldsymbol{C}_A) \end{aligned}$$

$$= \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{C}_A^{-1}\boldsymbol{A}^\top\right)\right)}{(2\pi)^{MH/2}|\boldsymbol{C}_A|^{M/2}}, \tag{12}$$

$$p(\boldsymbol{B}) = \prod_{l=1}^{L} \mathrm{Norm}_H(\widetilde{\boldsymbol{b}}_l; \boldsymbol{0}, \boldsymbol{C}_B)$$

$$= \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{B}\boldsymbol{C}_B^{-1}\boldsymbol{B}^\top\right)\right)}{(2\pi)^{LH/2}|\boldsymbol{C}_B|^{L/2}}, \tag{13}$$

where $\boldsymbol{V} \in \mathbb{R}^{L\times M}$ is an observed matrix, $\boldsymbol{A} \in \mathbb{R}^{M\times H}, \boldsymbol{B} \in \mathbb{R}^{L\times H}$ for $H \leq \min(L, M)$ are the parameters to be estimated. A column vector is denoted by a bolded-faced small letter, and a row vector is denoted by a bolded-faced small letter with tilde, e.g.,

$$\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_H) = (\widetilde{\boldsymbol{a}}_1, \ldots, \widetilde{\boldsymbol{a}}_M)^\top \in \mathbb{R}^{M\times H}.$$

The noise variance σ^2 is treated as a hyperparameter, i.e., considered as a constant or point-estimated by the empirical Bayesian procedure. The prior covariances

$$\boldsymbol{C}_A = \mathbf{Diag}(c_{a_1}^2, \ldots, c_{a_H}^2), \qquad \boldsymbol{C}_B = \mathbf{Diag}(c_{b_1}^2, \ldots, c_{b_H}^2) \tag{14}$$

are assumed to be diagonal, and also treated as hyperparameters.

6.1 VB for MF

We solve the following problem to get a local search algorithm for VB posterior:

$$\min_r F(r) \quad \text{s.t.} \quad r(\boldsymbol{A}, \boldsymbol{B}) = r_A(\boldsymbol{A}) r_B(\boldsymbol{B}),$$

where

$$F(r) = \left\langle \log \frac{r(\boldsymbol{A}, \boldsymbol{B})}{p(\boldsymbol{V}|\boldsymbol{A}, \boldsymbol{B}) p(\boldsymbol{A}) p(\boldsymbol{B})} \right\rangle_{r(\boldsymbol{A},\boldsymbol{B})}.$$

In this model, the stationary condition (8) is written as

$$r_A(\boldsymbol{A}) \propto p(\boldsymbol{A}|\boldsymbol{C}_A) \exp \left\langle \log p(\boldsymbol{V}|\boldsymbol{A}, \boldsymbol{B}) \right\rangle_{r_B(\boldsymbol{B})}, \tag{15}$$

$$r_B(\boldsymbol{B}) \propto p(\boldsymbol{B}|\boldsymbol{C}_B) \exp \left\langle \log p(\boldsymbol{V}|\boldsymbol{A}, \boldsymbol{B}) \right\rangle_{r_A(\boldsymbol{A})}. \tag{16}$$

Substituting the model likelihood (11) and the prior (12) on $\boldsymbol{A}$ into the stationary condition (15), and focusing on the dependency on $\boldsymbol{A}$, we have

$$r_A(\boldsymbol{A}) \propto \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{C}_A^{-1}\boldsymbol{A}^\top\right)\right)}{(2\pi)^{MH/2}|\boldsymbol{C}_A|^{M/2}} \cdot \exp\left\langle \log\left(\frac{\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{V}-\boldsymbol{B}\boldsymbol{A}^\top\|_{\mathrm{Fro}}^2\right)}{(2\pi\sigma^2)^{LM/2}}\right)\right\rangle_{r_B(\boldsymbol{B})}$$

$$\propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{C}_A^{-1}\boldsymbol{A}^\top\right) - \frac{1}{2\sigma^2}\left\langle \|\boldsymbol{V} - \boldsymbol{B}\boldsymbol{A}^\top\|_{\mathrm{Fro}}^2 \right\rangle_{r_B(\boldsymbol{B})}\right)$$

$$
\begin{aligned}
&\propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{C}_A^{-1}\boldsymbol{A}^\top + \frac{1}{\sigma^2}\left\langle -2\boldsymbol{V}^\top\boldsymbol{B}\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{A}^\top\right\rangle_{r_B(\boldsymbol{B})}\right)\right)\\
&\propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}\left(\frac{\left\langle\boldsymbol{B}^\top\boldsymbol{B}\right\rangle_{r_B(\boldsymbol{B})}}{\sigma^2} + \boldsymbol{C}_A^{-1}\right)\boldsymbol{A}^\top - \frac{2\boldsymbol{V}^\top\langle\boldsymbol{B}\rangle_{r_B(\boldsymbol{B})}\boldsymbol{A}^\top}{\sigma^2}\right)\right)\\
&\propto \exp\left(-\frac{\mathrm{tr}\left((\boldsymbol{A}-\widehat{\boldsymbol{A}})\widehat{\boldsymbol{\Sigma}}_A^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})^\top\right)}{2}\right), \qquad (17)
\end{aligned}
$$

where

$$
\widehat{\boldsymbol{A}} = \sigma^{-2}\boldsymbol{V}^\top\langle\boldsymbol{B}\rangle_{r_B(\boldsymbol{B})}\widehat{\boldsymbol{\Sigma}}_A, \qquad (18)
$$

$$
\widehat{\boldsymbol{\Sigma}}_A = \sigma^2\left(\left\langle\boldsymbol{B}^\top\boldsymbol{B}\right\rangle_{r_B(\boldsymbol{B})} + \sigma^2\boldsymbol{C}_A^{-1}\right)^{-1}. \qquad (19)
$$

Similarly, by substituting the model likelihood (11) and the prior (13) on $\boldsymbol{B}$ into the stationary condition (16), and focusing on the dependency on $\boldsymbol{B}$, we have

$$
\begin{aligned}
r_B(\boldsymbol{B}) &\propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{B}\boldsymbol{C}_B^{-1}\boldsymbol{B}^\top\right) - \frac{1}{2\sigma^2}\left\langle\|\boldsymbol{V}-\boldsymbol{B}\boldsymbol{A}^\top\|_{\mathrm{Fro}}^2\right\rangle_{r_A(\boldsymbol{A})}\right)\\
&\propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{B}\boldsymbol{C}_B^{-1}\boldsymbol{B}^\top + \frac{1}{\sigma^2}\left\langle -2\boldsymbol{V}\boldsymbol{A}\boldsymbol{B}^\top + \boldsymbol{B}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{B}^\top\right\rangle_{r_A(\boldsymbol{A})}\right)\right)\\
&\propto \exp\left(-\frac{\mathrm{tr}\left((\boldsymbol{B}-\widehat{\boldsymbol{B}})\widehat{\boldsymbol{\Sigma}}_B^{-1}(\boldsymbol{B}-\widehat{\boldsymbol{B}})^\top\right)}{2}\right), \qquad (20)
\end{aligned}
$$

where

$$
\widehat{\boldsymbol{B}} = \sigma^{-2}\boldsymbol{V}\langle\boldsymbol{A}\rangle_{r_A(\boldsymbol{A})}\widehat{\boldsymbol{\Sigma}}_B, \qquad (21)
$$

$$
\widehat{\boldsymbol{\Sigma}}_B = \sigma^2\left(\left\langle\boldsymbol{A}^\top\boldsymbol{A}\right\rangle_{r_A(\boldsymbol{A})} + \sigma^2\boldsymbol{C}_B^{-1}\right)^{-1}. \qquad (22)
$$

Eqs.(17) and (20) imply that

$$
\begin{aligned}
\langle\boldsymbol{A}\rangle_{r_A(\boldsymbol{A})} &= \widehat{\boldsymbol{A}},\\
\left\langle\boldsymbol{A}^\top\boldsymbol{A}\right\rangle_{r_A(\boldsymbol{A})} &= \widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}} + M\widehat{\boldsymbol{\Sigma}}_A,\\
\langle\boldsymbol{B}\rangle_{r_B(\boldsymbol{B})} &= \widehat{\boldsymbol{B}},\\
\left\langle\boldsymbol{B}^\top\boldsymbol{B}\right\rangle_{r_B(\boldsymbol{B})} &= \widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + L\widehat{\boldsymbol{\Sigma}}_B.
\end{aligned}
$$

Substituting above into Eqs.(18)–(22) gives

$$
\widehat{\boldsymbol{A}} = \sigma^{-2}\boldsymbol{V}^\top\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_A, \qquad (23)
$$

$$\widehat{\boldsymbol{\Sigma}}_A = \sigma^2 \left(\widehat{\boldsymbol{B}}^\top \widehat{\boldsymbol{B}} + L\widehat{\boldsymbol{\Sigma}}_B + \sigma^2 \boldsymbol{C}_A^{-1}\right)^{-1}, \tag{24}$$

$$\widehat{\boldsymbol{B}} = \sigma^{-2} \boldsymbol{V}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\Sigma}}_B, \tag{25}$$

$$\widehat{\boldsymbol{\Sigma}}_B = \sigma^2 \left(\widehat{\boldsymbol{A}}^\top \widehat{\boldsymbol{A}} + M\widehat{\boldsymbol{\Sigma}}_A + \sigma^2 \boldsymbol{C}_B^{-1}\right)^{-1}. \tag{26}$$

Starting from some initial values, iterating Eqs.(23)–(26) (substituting the right-hand side into the left-hand side) until convergence forms a local search algorithm of VB in MF.

6.2 Empirical VB for MF

The free energy can be written as a function of $\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\Sigma}}_A, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Sigma}}_B, \boldsymbol{C}_A, \boldsymbol{C}_B$ and σ^2:

$$\begin{aligned}2F = {} & LM\log(2\pi\sigma^2) + \frac{\|\boldsymbol{V} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{A}}^\top\|^2_{\mathrm{Fro}}}{\sigma^2} + M\log\frac{|\boldsymbol{C}_A|}{|\widehat{\boldsymbol{\Sigma}}_A|} + L\log\frac{|\boldsymbol{C}_B|}{|\widehat{\boldsymbol{\Sigma}}_B|} \\ & - (L+M)H + \mathrm{tr}\left\{\boldsymbol{C}_A^{-1}\left(\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}} + M\widehat{\boldsymbol{\Sigma}}_A\right) + \boldsymbol{C}_B^{-1}\left(\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + L\widehat{\boldsymbol{\Sigma}}_B\right)\right. \\ & \left. + \sigma^{-2}\left(-\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}}\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + \left(\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}} + M\widehat{\boldsymbol{\Sigma}}_A\right)\left(\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + L\widehat{\boldsymbol{\Sigma}}_B\right)\right)\right\},\end{aligned} \tag{27}$$

where $|\cdot|$ is the determinant of a matrix. Remember that $\boldsymbol{C}_A$ and $\boldsymbol{C}_B$, defined in Eq.(14), are diagonal.

Exercise 4 (20P): Compute the partial derivatives of Eq.(27) with respect to each diagonal element $c^2_{a_h}$ of $\boldsymbol{C}_A$, each diagonal element $c^2_{b_h}$ of $\boldsymbol{C}_B$ and the noise variance σ^2. (Advise) Confirm that the derivatives are consistent with the updated rules (28)–(30), which are derived as the stationary condition.

From the derivatives, we obtain the following update rules:

$$c^2_{a_h} = \|\widehat{\boldsymbol{a}}_h\|^2/M + \left(\widehat{\boldsymbol{\Sigma}}_A\right)_{h,h}, \tag{28}$$

$$c^2_{b_h} = \|\widehat{\boldsymbol{b}}_h\|^2/L + \left(\widehat{\boldsymbol{\Sigma}}_B\right)_{h,h}, \tag{29}$$

$$\sigma^2 = \frac{\|\boldsymbol{V}\|^2_{\mathrm{Fro}} - \mathrm{tr}\left(2\boldsymbol{V}^\top\widehat{\boldsymbol{B}}\widehat{\boldsymbol{A}}^\top\right) + \mathrm{tr}\left((\widehat{\boldsymbol{A}}^\top\widehat{\boldsymbol{A}} + M\widehat{\boldsymbol{\Sigma}}_A)(\widehat{\boldsymbol{B}}^\top\widehat{\boldsymbol{B}} + L\widehat{\boldsymbol{\Sigma}}_B)\right)}{LM}. \tag{30}$$

Iterating Eqs.(23)–(26) and Eqs.(28)–(30) gives a local solution of empirical VB learning.

6.3 Gibbs sampling for MF

Let $(\boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)})$ be the current sample in the Markov chain. In Gibbs sampling, we draw a new sample from the following Gaussian conditional posterior:

$$\boldsymbol{A}^{(t+1)} \sim p(\boldsymbol{A}|\boldsymbol{B}^{(t)}, \boldsymbol{V}) \propto p(\boldsymbol{V}|\boldsymbol{A}, \boldsymbol{B}^{(t)})p(\boldsymbol{A})$$

$$
\begin{aligned}
&\propto \exp\left(-\frac{\mathrm{tr}\left((\boldsymbol{A}-\widehat{\boldsymbol{A}})\widehat{\boldsymbol{\Sigma}}_A^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})^\top\right)}{2}\right),\\
\boldsymbol{B}^{(t+1)} \sim p(\boldsymbol{B}|\boldsymbol{A}^{(t+1)},\boldsymbol{V}) &\propto p(\boldsymbol{V}|\boldsymbol{A}^{(t+1)},\boldsymbol{B})p(\boldsymbol{B})\\
&\propto \exp\left(-\frac{\mathrm{tr}\left((\boldsymbol{B}-\widehat{\boldsymbol{B}})\widehat{\boldsymbol{\Sigma}}_B^{-1}(\boldsymbol{B}-\widehat{\boldsymbol{B}})^\top\right)}{2}\right).
\end{aligned}
$$

Exercise 5 (10P): Derive $\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\Sigma}}_A, \widehat{\boldsymbol{B}}$, and $\widehat{\boldsymbol{\Sigma}}_B$.

7 MF with missing entries

Let Λ be the set of indices, at which the element of $\boldsymbol{V}$ are observed. Then, the model likelihood is written as

$$
\begin{aligned}
p(\boldsymbol{V}|\boldsymbol{A},\boldsymbol{B}) &= \prod_{(l,m)\in\Lambda} \mathrm{Norm}_1(V_{l,m}; \widetilde{\boldsymbol{b}}_l^\top \widetilde{\boldsymbol{a}}_m)\\
&= \frac{\exp\left(-\frac{1}{2\sigma^2}\|\mathcal{P}_\Lambda(\boldsymbol{V})-\mathcal{P}_\Lambda(\boldsymbol{B}\boldsymbol{A}^\top)\|_{\mathrm{Fro}}^2\right)}{(2\pi\sigma^2)^{|\Lambda|/2}},
\end{aligned}
\tag{31}
$$

where $\mathcal{P}_\Lambda(\boldsymbol{V}) : \mathbb{R}^{L\times M} \mapsto \mathbb{R}^{L\times M}$ maps the observed entries to the observed values and the unobserved entries to zero, i.e.,

$$
\left(\mathcal{P}_\Lambda(\boldsymbol{V})\right)_{l,m} = \begin{cases} V_{l,m} & \text{if } (l,m)\in\Lambda,\\ 0 & \text{otherwise,}\end{cases}
$$

and $|\Lambda|$ denotes the number of the observed entries Λ.

We use the priors (12) and (13).

7.1 VB for MF with missing entries

We can start from the stationary conditions (15) and (16). Substituting the model likelihood (31) and the prior (12) into the stationary condition (15), and focusing on the dependency on $\boldsymbol{A}$, we have

$$
\begin{aligned}
r_A(\boldsymbol{A}) &\propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{C}_A^{-1}\boldsymbol{A}^\top\right) - \frac{\left\langle\|\mathcal{P}_\Lambda(\boldsymbol{V})-\mathcal{P}_\Lambda(\boldsymbol{B}\boldsymbol{A}^\top)\|_{\mathrm{Fro}}^2\right\rangle_{r_B(\boldsymbol{B})}}{2\sigma^2}\right)\\
&\propto \exp\Big(-\tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{C}_A^{-1}\boldsymbol{A}^\top\right)\\
&+\frac{\sum_{(l,m)\in\Lambda}\left\langle -2V_{l,m}\sum_{h=1}^H B_{l,h}A_{m,h} + \sum_{h=1}^H\sum_{h'=1}^H B_{l,h}B_{l,h'}A_{m,h}A_{m,h'}\right\rangle_{r_B(\boldsymbol{B})}}{\sigma^2}\Big)
\end{aligned}
$$

$$\propto \exp\left(-\frac{\sum_{m=1}^{M}\left((\widetilde{\boldsymbol{a}}_m - \widehat{\widetilde{\boldsymbol{a}}}_m)^\top \widehat{\boldsymbol{\Sigma}}_{A,m}^{-1} (\widetilde{\boldsymbol{a}}_m - \widehat{\widetilde{\boldsymbol{a}}}_m)\right)}{2} \right), \tag{32}$$

where

$$\widehat{\widetilde{\boldsymbol{a}}}_m = \sigma^{-2} \widehat{\boldsymbol{\Sigma}}_{A,m} \sum_{l;(l,m)\in\mathit{\Lambda}} V_{l,m} \left\langle \widetilde{\boldsymbol{b}}_l \right\rangle_{r_B(\boldsymbol{B})}, \tag{33}$$

$$\widehat{\boldsymbol{\Sigma}}_{A,m} = \sigma^2 \left(\sum_{l;(l,m)\in\mathit{\Lambda}} \left\langle \widetilde{\boldsymbol{b}}_l \widetilde{\boldsymbol{b}}_l^\top \right\rangle_{r_B(\boldsymbol{B})} + \sigma^2 \boldsymbol{C}_A^{-1} \right)^{-1}. \tag{34}$$

Here, $\sum_{(l,m)\in\mathit{\Lambda}}$ is the sum over the indices (l,m) condtained in the set $\mathit{\Lambda}$, and $\sum_{l;(l,m)\in\mathit{\Lambda}}$ is the sum over l that satisfies $(l,m)\in\mathit{\Lambda}$ given m.

Eq.(32) implies that the posterior on $\boldsymbol{A}$ is the Gaussian

$$\begin{aligned} r_A(\boldsymbol{A}) &= \prod_{m=1}^{M} \mathrm{Norm}_H(\widetilde{\boldsymbol{a}}_m; \widetilde{\widehat{\boldsymbol{a}}}_m, \widehat{\boldsymbol{\Sigma}}_{A,m}) \\ &= \prod_{m=1}^{M} \frac{\exp\left(-\frac{(\widetilde{\boldsymbol{a}}_m - \widetilde{\widehat{\boldsymbol{a}}}_m)^\top \widehat{\boldsymbol{\Sigma}}_{A,m}^{-1} (\widetilde{\boldsymbol{a}}_m - \widetilde{\widehat{\boldsymbol{a}}}_m)}{2} \right)}{(2\pi)^{H/2} |\widehat{\boldsymbol{\Sigma}}_{A,m}|^{1/2}} \end{aligned} \tag{35}$$

with mean $\widetilde{\widehat{\boldsymbol{a}}}_m$ and covariance $\widehat{\boldsymbol{\Sigma}}_{A,m}$ satisfying Eqs.(33) and (34).

Similarly, we have

$$\begin{aligned} r_B(\boldsymbol{B}) &\propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{B}\boldsymbol{C}_B^{-1}\boldsymbol{B}^\top \right) - \frac{\left\langle \|\mathcal{P}_{\mathit{\Lambda}}(\boldsymbol{V}) - \mathcal{P}_{\mathit{\Lambda}}(\boldsymbol{B}\boldsymbol{A}^\top)\|_{\mathrm{Fro}}^2 \right\rangle_{r_A(\boldsymbol{A})}}{2\sigma^2} \right) \\ &\propto \exp\left(-\frac{\sum_{l=1}^{L}\left((\widetilde{\boldsymbol{b}}_m - \widehat{\widetilde{\boldsymbol{b}}}_l)^\top \widehat{\boldsymbol{\Sigma}}_{B,l}^{-1} (\widetilde{\boldsymbol{b}}_l - \widehat{\widetilde{\boldsymbol{b}}}_l)\right)}{2} \right), \end{aligned} \tag{36}$$

where

$$\widehat{\widetilde{\boldsymbol{b}}}_l = \sigma^{-2} \widehat{\boldsymbol{\Sigma}}_{B,l} \sum_{m;(l,m)\in\mathit{\Lambda}} V_{l,m} \left\langle \widetilde{\boldsymbol{a}}_m \right\rangle_{r_A(\boldsymbol{A})}, \tag{37}$$

$$\widehat{\boldsymbol{\Sigma}}_{B,l} = \sigma^2 \left(\sum_{m;(l,m)\in\mathit{\Lambda}} \left\langle \widetilde{\boldsymbol{a}}_m \widetilde{\boldsymbol{a}}_m^\top \right\rangle_{r_A(\boldsymbol{A})} + \sigma^2 \boldsymbol{C}_B^{-1} \right)^{-1}. \tag{38}$$

Eq.(36) implies that the posterior on $\boldsymbol{B}$ is Gaussian

$$r_B(\boldsymbol{B}) = \prod_{l=1}^{L} \mathrm{Norm}_H(\widetilde{\boldsymbol{b}}_l; \widetilde{\widehat{\boldsymbol{b}}}_l, \widehat{\boldsymbol{\Sigma}}_{B,l})$$

$$= \prod_{l=1}^{L} \frac{\exp\left(-\frac{(\widetilde{\boldsymbol{b}}_l - \widetilde{\widehat{\boldsymbol{b}}}_l)^\top \widehat{\boldsymbol{\Sigma}}_{B,l}^{-1} (\widetilde{\boldsymbol{b}}_l - \widetilde{\widehat{\boldsymbol{b}}}_l)}{2}\right)}{(2\pi)^{H/2} |\widehat{\boldsymbol{\Sigma}}_{B,l}|^{1/2}} \tag{39}$$

with mean $\widetilde{\widehat{\boldsymbol{b}}}_m$ and covariance $\widehat{\boldsymbol{\Sigma}}_{B,l}$ satisfying Eqs.(37) and (38).

Exercise 6 (15P): Compute the expectations in the right-hand sides of Eqs.(33), (34), (37), **and**(38) **based on Eqs.**(35) **and** (39). **Then, derive update rules for the means and the covariances $\widetilde{\widehat{\boldsymbol{a}}}_m, \widehat{\boldsymbol{\Sigma}}_{A,m}, \widetilde{\widehat{\boldsymbol{b}}}_l, \widehat{\boldsymbol{\Sigma}}_{B,l}$. (Advise: Confirm that the update rules coincide to Eqs.**(23)–(26) **when there is no missing entries, i.e., Λ contains all the elements of $\boldsymbol{V}$.)**

7.2 Gibbs sampling for MF with missing entries

Let $(\boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)})$ be the current sample in the Markov chain. In Gibbs sampling, we draw a new sample from the following Gaussian conditional posterior: for each row of $\boldsymbol{A}^{(t)} = (\widetilde{\boldsymbol{a}}_1^{(t)}, \ldots, \widetilde{\boldsymbol{a}}_M^{(t)})$ and $\boldsymbol{B}^{(t)} = (\widetilde{\boldsymbol{b}}_1^{(t)}, \ldots, \widetilde{\boldsymbol{b}}_L^{(t)})$,

$$\begin{aligned}
\widetilde{\boldsymbol{a}}_m^{(t+1)} \sim p(\widetilde{\boldsymbol{a}}_m | \boldsymbol{B}^{(t)}, \boldsymbol{V}) &\propto p(\boldsymbol{V}|\boldsymbol{A}, \boldsymbol{B}^{(t)}) p(\boldsymbol{A}) \\
&\propto \mathrm{Norm}_H(\widetilde{\boldsymbol{a}}_m; \widetilde{\widehat{\boldsymbol{a}}}_m, \widehat{\boldsymbol{\Sigma}}_{A,m}), \\
\widetilde{\boldsymbol{b}}_l^{(t+1)} \sim p(\widetilde{\boldsymbol{b}}_l | \boldsymbol{A}^{(t+1)}, \boldsymbol{V}) &\propto p(\boldsymbol{V}|\boldsymbol{A}^{(t+1)}, \boldsymbol{B}) p(\boldsymbol{B}) \\
&\propto \mathrm{Norm}_H(\widetilde{\boldsymbol{b}}_l; \widetilde{\widehat{\boldsymbol{b}}}_l, \widehat{\boldsymbol{\Sigma}}_{B,l}).
\end{aligned}$$

Exercise 7 (15P): Derive $\widetilde{\widehat{\boldsymbol{a}}}_m, \widehat{\boldsymbol{\Sigma}}_{A,m}, \widetilde{\widehat{\boldsymbol{b}}}_l$, and $\widehat{\boldsymbol{\Sigma}}_{B,l}$.

8 Mixture of Gaussians (MoG)

Consider the following Gaussian mixture model:

$$p(\boldsymbol{z}|\boldsymbol{\alpha}) = \mathrm{Multinomial}_{K,1}(\boldsymbol{z}; \boldsymbol{\alpha}), \tag{40}$$

$$p(\boldsymbol{x}|\boldsymbol{z}, \{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{k=1}^{K} \{\mathrm{Norm}_M(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{I}_M)\}^{z_k}, \tag{41}$$

$$p(\boldsymbol{\alpha}|\boldsymbol{\phi}) = \mathrm{Dirichlet}_K(\boldsymbol{\alpha}; (\phi, \ldots, \phi)^\top), \tag{42}$$

$$p(\{\boldsymbol{\mu}_k\}_{k=1}^K | \sigma_0^2) = \prod_{k=1}^{K} \mathrm{Norm}_M(\boldsymbol{\mu}_k | \boldsymbol{0}, \sigma_0^2 \boldsymbol{I}_M). \tag{43}$$

For N observations $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$, the complete likelihood for the parameters $\boldsymbol{w} = (\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)$ and the hidden variables $\mathcal{H} = \{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(N)}\}$

are given as

$$p(\mathcal{D}, \{\boldsymbol{z}^{(n)}\}_{n=1}^{N}|\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^{K}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left\{\alpha_k \mathrm{Norm}_M(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{I}_M)\right\}^{z_k^{(n)}}. \tag{44}$$

8.1 VB for MoG

To obtain an approximate posterior $r(\mathcal{H}, \boldsymbol{w})$ on the parameters and the hidden variables, we solve the following problem

$$\widehat{r} = \operatorname*{argmin}_{r} F(r) \qquad \text{s.t.} \qquad r(\mathcal{H}, \boldsymbol{w}) = r_{\mathcal{H}}(\mathcal{H})r_w(\boldsymbol{w}). \tag{45}$$

Under this constraint, the free energy is written as

$$\begin{aligned} F(r) &= \left\langle \log \frac{r_{\mathcal{H}}(\mathcal{H})r_w(\boldsymbol{w})}{p(\mathcal{D}, \mathcal{H}|\boldsymbol{w})p(\boldsymbol{w})} \right\rangle_{r_{\mathcal{H}}(\mathcal{H})r_w(\boldsymbol{w})} \\ &= \sum_{\mathcal{H}} \int r_{\mathcal{H}}(\mathcal{H})r_w(\boldsymbol{w}) \log \frac{r_{\mathcal{H}}(\mathcal{H})r_w(\boldsymbol{w})}{p(\mathcal{D}, \mathcal{H}|\boldsymbol{w})p(\boldsymbol{w})} d\boldsymbol{w}. \end{aligned} \tag{46}$$

By applying the variational method to the problem above, we obtain the following as stationary conditions, which corresponds to Eq.(8):

$$r_{\mathcal{H}}(\mathcal{H}) \propto \exp \left\langle \log p(\mathcal{D}, \mathcal{H}|\boldsymbol{w})\right\rangle_{r_w(\boldsymbol{w})}, \tag{47}$$

$$r_w(\boldsymbol{w}) \propto p(\boldsymbol{w}) \exp \left\langle \log p(\mathcal{D}, \mathcal{H}|\boldsymbol{w})\right\rangle_{r_{\mathcal{H}}(\mathcal{H})}. \tag{48}$$

By substituting (44) into Eq.(47), and focusing on the dependency on $\mathcal{H} = \{\boldsymbol{z}^{(n)}\}_{n=1}^{N}$, we have

$$\begin{aligned} &r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^{N}) \\ &\quad\propto \exp\left\langle \log \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\alpha_k \frac{\exp\left(-\frac{\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\|^2}{2}\right)}{(2\pi)^{M/2}}\right)^{z_k^{(n)}}\right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha},\{\boldsymbol{\mu}_k\}_{k=1}^{K})} \\ &\quad\propto \prod_{n=1}^{N}\prod_{k=1}^{K}\exp\left\langle z_k^{(n)}\left(\log\alpha_k - \frac{1}{2}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\|^2\right)\right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha},\{\boldsymbol{\mu}_k\}_{k=1}^{K})} \\ &\quad\propto \prod_{n=1}^{N}\prod_{k=1}^{K}\exp\Bigg(z_k^{(n)}\Big(\left\langle\log\alpha_k\right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha},\{\boldsymbol{\mu}_k\}_{k=1}^{K})} \\ &\qquad\qquad\qquad - \frac{1}{2}\left\langle\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\|^2\right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha},\{\boldsymbol{\mu}_k\}_{k=1}^{K})}\Big)\Bigg) \\ &\quad\propto \prod_{n=1}^{N}\left(\prod_{k=1}^{K}\left(\overline{z}_k^{(n)}\right)^{z_k^{(n)}}\right), \end{aligned} \tag{49}$$

where

$$\overline{\boldsymbol{z}}_k^{(n)} = \exp\left(\langle \log \alpha_k \rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)} - \frac{1}{2} \left\langle \|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\|^2 \right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)} \right). \quad (50)$$

Eq.(49) implies that

$$r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N) = \prod_{n=1}^N \mathrm{Multinomial}_{K,1}\left(\boldsymbol{z}^{(n)}; \widehat{\boldsymbol{z}}^{(n)}\right), \quad (51)$$

where

$$\widehat{z}_k^{(n)} = \frac{\overline{z}_k^{(n)}}{\sum_{k'=1}^K \overline{z}_{k'}^{(n)}}. \quad (52)$$

On the other hand, by substituting the model likelihood (44) and the priors (42) and (43) into Eq.(48), and focusing on the dependency on $\boldsymbol{w} = (\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)$, we have

$$\begin{aligned}
& r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K) \\
& \quad \propto \left(\prod_{k=1}^K \alpha_k^{\phi-1} \frac{\exp\left(-\frac{\|\boldsymbol{\mu}_k\|^2}{2\sigma_0^2}\right)}{(2\pi\sigma_0^2)^{M/2}} \right) \\
& \qquad \cdot \exp \left\langle \log \prod_{n=1}^N \prod_{k=1}^K \left(\alpha_k \frac{\exp\left(-\frac{\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\|^2}{2}\right)}{(2\pi)^{M/2}} \right)^{z_k^{(n)}} \right\rangle_{r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N)} \\
& \quad \propto \prod_{k=1}^K \alpha_k^{\phi-1} \exp\left\{ -\frac{\|\boldsymbol{\mu}_k\|^2}{2\sigma_0^2} \right. \\
& \qquad \left. + \sum_{n=1}^N \left(\log \alpha_k - \frac{1}{2}\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\|^2 \right) \left\langle z_k^{(n)} \right\rangle_{r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N)} \right\} \\
& \quad \propto \prod_{k=1}^K \alpha_k^{\overline{N}_k+\phi-1} \exp\left(-\frac{(\overline{N}_k + \sigma_0^{-2})\|\boldsymbol{\mu}_k - \frac{\overline{N}_k \overline{\boldsymbol{x}}_k}{\overline{N}_k + \sigma_0^{-2}}\|^2}{2} \right),
\end{aligned} \quad (53)$$

where

$$\overline{N}_k = \sum_{n=1}^N \langle z_k^{(n)} \rangle_{r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N)}, \quad (54)$$

$$\overline{\boldsymbol{x}}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^N \boldsymbol{x}^{(n)} \langle z_k^{(n)} \rangle_{r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N)}. \quad (55)$$

Eq.(53) implies that the posterior is the (independent) product of a Dirichlet distribution on $\boldsymbol{\alpha}$ and the Gaussian distribution on $\{\boldsymbol{\mu}_k\}_{k=1}^K$, i.e.,

$$r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K) = r_\alpha(\boldsymbol{\alpha}) r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K),$$

$$\text{where} \qquad r_\alpha(\boldsymbol{\alpha}) = \mathrm{Dirichlet}\left(\boldsymbol{\alpha}; \widehat{\boldsymbol{\alpha}}\right), \tag{56}$$

$$r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{k=1}^K \mathrm{Norm}_M\left(\boldsymbol{\mu}_k; \widehat{\boldsymbol{\mu}}_k, \widehat{\sigma}_k^2 \boldsymbol{I}_M\right), \tag{57}$$

$$\widehat{\alpha}_k = \overline{N}_k + \phi, \tag{58}$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\overline{N}_k \overline{\boldsymbol{x}}_k}{\overline{N}_k + \sigma_0^{-2}}, \tag{59}$$

$$\widehat{\sigma}_k^2 = \frac{1}{\overline{N}_k + \sigma_0^{-2}}. \tag{60}$$

We can use the following expectation over multinomial, Dirichlet, and Gaussian distribution:

$$\begin{aligned}
\langle z_k^{(n)} \rangle_{r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N)} &= \widehat{z}_k^{(n)}, \\
\langle \log \alpha_k \rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)} &= \langle \log \alpha_k \rangle_{r_\alpha(\boldsymbol{\alpha})} \\
&= \mathit{\Psi}(\widehat{\alpha}_k) - \mathit{\Psi}\left(\sum_{k'=1}^K \widehat{\alpha}_{k'}\right), \\
\left\langle \|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\|^2 \right\rangle_{r_{\alpha,\mu}(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K)} &= \left\langle \|(\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k) + (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)\|^2 \right\rangle_{r(\boldsymbol{\mu}_k)} \\
&= \|\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\|^2 + M\widehat{\sigma}_k^2,
\end{aligned}$$

where

$$\mathit{\Psi}(x) = \frac{\partial \log \mathit{\Gamma}(x)}{\partial x}$$

is called the Digamma function.

By substituting above into Eqs.(50), (54), and (55), we conclude that the VB posterior is given by

$$\begin{aligned}
r(\{\boldsymbol{z}^{(n)}\}_{n=1}^N, \boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^K) &= r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N) r_\alpha(\boldsymbol{\alpha}) r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K), \\
\text{where} \qquad r_z(\{\boldsymbol{z}^{(n)}\}_{n=1}^N) &= \prod_{n=1}^N \mathrm{Multinomial}_{K,1}\left(\boldsymbol{z}^{(n)}; \widehat{\boldsymbol{z}}^{(n)}\right), \\
r_\alpha(\boldsymbol{\alpha}) &= \mathrm{Dirichlet}\left(\boldsymbol{\alpha}; \widehat{\boldsymbol{\alpha}}\right), \\
r_\mu(\{\boldsymbol{\mu}_k\}_{k=1}^K) &= \prod_{k=1}^K \mathrm{Norm}_M\left(\boldsymbol{\mu}_k; \widehat{\boldsymbol{\mu}}_k, \widehat{\sigma}_k^2 \boldsymbol{I}_M\right).
\end{aligned}$$

Here, $\{\widehat{\boldsymbol{z}}^{(n)}\}_{n=1}^N$, $\widehat{\boldsymbol{\alpha}}$ and $\{\widehat{\boldsymbol{\mu}}_k, \widehat{\sigma}_k^2\}_{k=1}^K$ satisfies

$$\widehat{z}_k^{(n)} = \frac{\overline{z}_k^{(n)}}{\sum_{k'=1}^K \overline{z}_{k'}^{(n)}}, \tag{61}$$

$$\widehat{\alpha}_k = \overline{N}_k + \phi, \tag{62}$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\overline{N}_k \overline{\boldsymbol{x}}_k}{\overline{N}_k + \sigma_0^{-2}}, \tag{63}$$

$$\widehat{\sigma}_k^2 = \frac{1}{\overline{N}_k + \sigma_0^{-2}}, \tag{64}$$

where

$$\begin{aligned} \overline{\boldsymbol{z}}_k^{(n)} &= \exp\left(\Psi(\widehat{\alpha}_k) - \Psi\left(\sum_{k'=1}^{K} \widehat{\alpha}_{k'}\right) - \frac{1}{2}\left\|\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\right\|^2 + M\widehat{\sigma}_k^2\right) \\ &= \exp\left(\Psi(\widehat{\alpha}_k) - \frac{1}{2}\left\|\boldsymbol{x}^{(n)} - \widehat{\boldsymbol{\mu}}_k\right\|^2 + M\widehat{\sigma}_k^2 + \text{const.}\right), \end{aligned} \tag{65}$$

$$\overline{N}_k = \sum_{n=1}^{N} \widehat{z}_k^{(n)}, \tag{66}$$

$$\overline{\boldsymbol{x}}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^{N} \boldsymbol{x}^{(n)} \widehat{z}_k^{(n)}. \tag{67}$$

Applying Eqs.(61)–(64) in turn (Eqs.(65)–(67) are also used when necessary), we can obtain a local solution for VB learning, which *monotonically* minimizes the free energy:

$$\begin{aligned} F &= \left\langle \log \frac{r_{\mathcal{H}}(\mathcal{H}) r_w(\boldsymbol{w})}{p(\boldsymbol{w})} \right\rangle_{r_{\mathcal{H}}(\mathcal{H}) r_w(\boldsymbol{w})} - \left\langle \log p(\mathcal{D}, \mathcal{H}|\boldsymbol{w}) \right\rangle_{r_{\mathcal{H}}(\mathcal{H}) r_w(\boldsymbol{w})} \\ &= \left\langle \log \frac{(\widehat{z}_k^{(n)})^{z_k^{(n)}} \frac{\Gamma(\sum_{k=1}^{K} \widehat{\alpha}_k)}{\prod_{k=1}^{K} \Gamma(\widehat{\alpha}_k)} \prod_{k=1}^{K} \alpha_k^{\widehat{\alpha}_k - 1} \frac{\exp\left(-\frac{\|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k\|^2}{2\widehat{\sigma}_k^2}\right)}{(2\pi\widehat{\sigma}_k^2)^{M/2}}}{\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K} \prod_{k=1}^{K} \alpha_k^{\phi - 1} \frac{\exp\left(-\frac{\|\boldsymbol{\mu}_k\|^2}{2\sigma_0^2}\right)}{(2\pi\sigma_0^2)^{M/2}}} \right\rangle_{r_{\mathcal{H}}(\mathcal{H}) r_w(\boldsymbol{w})} \\ &\qquad - \left\langle \log \prod_{n=1}^{N} \prod_{k=1}^{K} \left\{ \alpha_k \frac{\exp\left(-\frac{\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\|^2}{2}\right)}{(2\pi)^{M/2}} \right\}^{z_k^{(n)}} \right\rangle_{r_{\mathcal{H}}(\mathcal{H}) r_w(\boldsymbol{w})} \\ &= \log\left(\frac{\Gamma(\sum_{k=1}^{K} \widehat{\alpha}_k)}{\prod_{k=1}^{K} \Gamma(\widehat{\alpha}_k)}\right) - \log\left(\frac{\Gamma(K\phi)}{(\Gamma(\phi))^K}\right) + \frac{M}{2} \sum_{k=1}^{K} \log \frac{\sigma_0^2}{\widehat{\sigma}_k^2} - \frac{KM}{2} \\ &\quad + \sum_{n=1}^{N} \sum_{k=1}^{K} \widehat{z}_k^{(n)} \log \widehat{z}_k^{(n)} + \sum_{k=1}^{K} \left(\widehat{\alpha}_k - \phi - \overline{N}_k\right) \left(\Psi(\widehat{\alpha}_k) - \Psi(\textstyle\sum_{k'=1}^{K} \widehat{\alpha}_{k'})\right) \\ &\quad + \sum_{k=1}^{K} \frac{\|\widehat{\boldsymbol{\mu}}_k\|^2 + M\widehat{\sigma}_k^2}{2\sigma_0^2} + \sum_{k=1}^{K} \frac{\overline{N}_k \left(M \log(2\pi) + M\widehat{\sigma}_k^2\right)}{2} \\ &\quad + \sum_{k=1}^{K} \frac{\overline{N}_k \|\overline{\boldsymbol{x}}_k - \widehat{\boldsymbol{\mu}}_k\|^2 + \sum_{n=1}^{N} \widehat{\boldsymbol{z}}_k^{(n)} \|\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}_k\|^2}{2}. \end{aligned} \tag{68}$$

Note that const. in (65) does not affect Eq.(61), and can be replaced with 0.

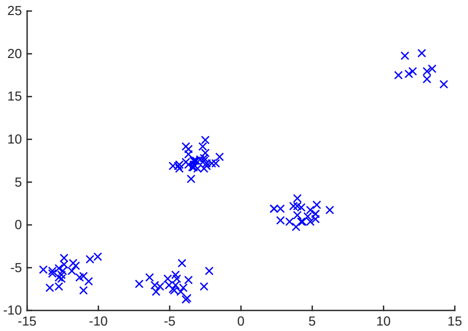

Figure 2: Scatter plot of “data2.txt”.

Exercise 8 (20P): Implement the VB algorithm for MoG. $N = 100$ **samples in** $L = 2$ **dimensional space are given in a separate file “data2.txt”. Fig.2 shows a scatter plot of the data points. Set** $K = 10, \phi = 1, \sigma_0^2 = 10000$**, Use the initial values for** $\{\widehat{z}_k^{(n)}\}$**, given as a** $K \times N$ **matrix in a separate file “initial.txt”. Since the initialization for other unknowns are not given, the update rules should be applied in the following order, Eqs.**(62), (63), (64), (61)**. Iterate the whole update rules** $T = 20$ **times.**

Draw the data samples with different colors according to the $\mathrm{argmax}_k(\widehat{z}_k^{(n)})$ **for** $t = 2, 5, 10$**, where** $t = 1, \ldots, T$ **is the number of iterations. Also, draw** $\widehat{\alpha}_k$ **after sorting in decreasing order. Fig.3 show them for** $t = 1$ **and** $t = 20$ **(Estimated clustering centers** $\{\widehat{\boldsymbol{\mu}}_k\}$ **are also plotted with circles). Any programming language or software can be used. Only submit printed figures. (Advises) When confirming the monotonic decrease of the free energy** (68) **(this is not mandatory), don’t compute products of gamma functions but the sum of the log of gamma functions, which is supported by popular programming languages, e.g., the *gammaln* command in matlab. Fig.4 shows a typical curve of the decreasing free energy in VB learning.**

8.2 Gibbs sampling for MoG

Let $(\boldsymbol{\alpha}^{(t)}, \{\boldsymbol{\mu}_k^{(t)}\}_{k=1}^K, \{\boldsymbol{z}^{(n,t)}\}_{n=1}^N)$ be the current sample in the Markov chain. In the *naive* Gibbs sampling, we draw a new sample from the following distributions:

$$\{\boldsymbol{z}^{(n,t+1)}\}_{n=1}^N \sim p(\{\boldsymbol{z}^{(n)}\}_{n=1}^N | \boldsymbol{\alpha}^{(t)}, \{\boldsymbol{\mu}_k^{(t)}\}_{k=1}^K, \mathcal{D}) = \prod_{n=1}^N \mathrm{Multinomial}_{K,1}\left(\boldsymbol{z}^{(n)}; \widehat{\boldsymbol{z}}^{(n)}\right), \tag{69}$$

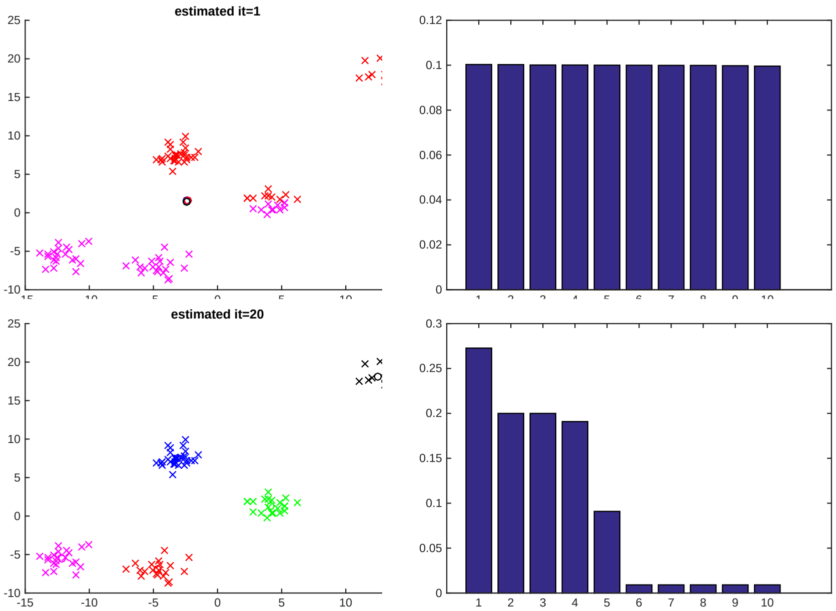

Figure 3: Scatter plot (left) with different colors based on the clustering result, and the estimated weight parameters (right) for $t = 1$ (top) and $t = 20$ (bottom).

$$\boldsymbol{\alpha}^{(t)} \sim p(\boldsymbol{\alpha}|\{\boldsymbol{z}^{(n,t+1)}\}_{n=1}^{N}, \mathcal{D}) = \mathrm{Dirichlet}\left(\boldsymbol{\alpha}; \widehat{\boldsymbol{\alpha}}\right), \tag{70}$$

$$\{\boldsymbol{\mu}_k^{(t)}\}_{k=1}^{K} \sim p(\{\boldsymbol{\mu}_k\}_{k=1}^{K}|\{\boldsymbol{z}^{(n,t+1)}\}_{n=1}^{N}, \mathcal{D}) = \prod_{k=1}^{K} \mathrm{Norm}_M\left(\boldsymbol{\mu}_k; \widehat{\boldsymbol{\mu}}_k, \widehat{\sigma}_k^2 \boldsymbol{I}_M\right). \tag{71}$$

Exercise 9 (10P): Derive $\widehat{\boldsymbol{z}}^{(n)}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\mu}}_k$, and $\widehat{\sigma}_k^2$.

Based on the conditional conjugacy, one can integrate the parameters $\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k\}_{k=1}^{K}$ out, and obtain the posterior on the hidden variables $\{\boldsymbol{z}^{(n)}\}_{n=1}^{N}$. Then, the posterior of $\boldsymbol{z}^{(n)}$ conditioned on $\{\boldsymbol{z}^{(n')}\}_{n' \neq n}$ can be used for sampling. This method is called collapsed Gibbs sampling.

9 Latent Dirichlet allocation (LDA)

Assume that a corpus of M documents are given. Each document m consists of $N^{(m)}$ tokens $\{\boldsymbol{w}^{(n,m)}\}_{n=1}^{N^{(m)}}$, and we express L different words with the one-of-L expression $\boldsymbol{w}^{(n,m)} \in \{\boldsymbol{e}_l\}_{l=1}^{L}$ ($\boldsymbol{e}_l$ is a L-dimensional vector with only a single element equal to one, and the others equal to zero).

In latent Dirichlet allocation, we assume that each token belongs to a latent topic $\boldsymbol{z}^{(n,m)} \in \{\boldsymbol{e}_h\}_{h=1}^{H}$. Each document has a specific topic distributions $\widetilde{\boldsymbol{\theta}}_m$,

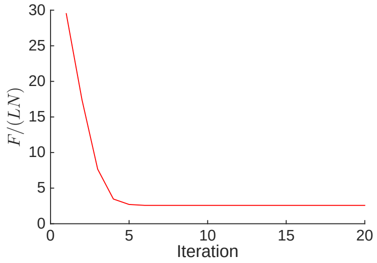

Figure 4: Free energy.

and each topic has a specific word distribution $\boldsymbol{\beta}_h$. The model is described as follows:

$$p(\boldsymbol{z}^{(n,m)}|\widetilde{\boldsymbol{\theta}}_m) = \mathrm{Multinomial}_{H,1}(\boldsymbol{z}^{(n,m)}; \widetilde{\boldsymbol{\theta}}_m), \tag{72}$$

$$p(\boldsymbol{w}^{(n,m)}|\boldsymbol{z}^{(n,m)}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H) = \prod_{h=1}^{H} \left\{\mathrm{Multinomial}_{L,1}(\boldsymbol{w}^{(n,m)}; \boldsymbol{\beta}_h)\right\}^{z_h^{(n,m)}}. \tag{73}$$

We use the Dirichlet priors for $\widetilde{\boldsymbol{\theta}}_m$ and $\boldsymbol{\beta}_h$:

$$p(\widetilde{\boldsymbol{\theta}}_m|\boldsymbol{\alpha}) = \mathrm{Dirichlet}_H(\widetilde{\boldsymbol{\theta}}_m; \boldsymbol{\alpha}), \tag{74}$$

$$p(\boldsymbol{\beta}_h|\boldsymbol{\eta}) = \mathrm{Dirichlet}_L(\boldsymbol{\beta}_h; \boldsymbol{\eta}). \tag{75}$$

We summarize the document parameters as $\boldsymbol{\Theta} = (\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_M)^\top \in \mathbb{R}^{M \times H}$, and the topic parameters as $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H) \in \mathbb{R}^{L \times H}$. The joint distribution of the observed data $\mathcal{D} = \{\{\boldsymbol{w}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^{M}$, and the hidden variables $\mathcal{H} = \{\{\boldsymbol{z}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^{M}$ is written as

$$\begin{aligned} p(\mathcal{D}, \mathcal{H}|\boldsymbol{\Theta}, \boldsymbol{B}) &= \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} p(\boldsymbol{w}^{(n,m)}|\boldsymbol{z}^{(n,m)}, \boldsymbol{\beta}_h) p(\boldsymbol{z}^{(n,m)}|\widetilde{\boldsymbol{\theta}}_m) \\ &= \prod_{m=1}^{M} \prod_{n=1}^{N^{(m)}} \prod_{h=1}^{H} \left(\Theta_{m,h} \prod_{l=1}^{L} B_{l,h}^{w_l^{(n,m)}}\right)^{z_h^{(n,m)}}. \end{aligned} \tag{76}$$

Exercise 11 (30P): Derive the variational Bayesian algorithm for estimating the posterior on the hidden variables $\mathcal{H} = \{\{\boldsymbol{z}^{(n,m)}\}_{n=1}^{N^{(m)}}\}_{m=1}^{M}$ and the parameters $\boldsymbol{w} = (\boldsymbol{\Theta}, \boldsymbol{B})$, fixing the hyperparameters $\boldsymbol{\kappa} = (\boldsymbol{\alpha}, \boldsymbol{\eta})$ as constants.