Chapter 21. Variational inference

21.1.

a.

$$\frac{\partial}{\partial \mu} \log p(\mu, l|\mathcal{D}) = \frac{-2}{-2\sigma^2}(n(\bar{x}-\mu)) = \frac{n(\bar{x}-\mu)}{\sigma^2}.$$

$$\frac{\partial}{\partial l} \log p(\mu, l|\mathcal{D}) = \frac{\partial \sigma}{\partial l}(-\frac{n}{\sigma}+\frac{1}{\sigma^3}(ns^2+n(\bar{x}-\mu)^2)) = \sigma(-\frac{n}{\sigma}+\frac{1}{\sigma^3}(ns^2+n(\bar{x}-\mu)^2))$$

$$= -n + \frac{ns^2+n(\bar{x}-\mu)^2}{\sigma^2}.$$

b.

$$\frac{\partial^2}{\partial \mu^2} \log p(\mu, l|\mathcal{D}) = \frac{\partial}{\partial \mu}\frac{n(\bar{x}-\mu)}{\sigma^2} = -\frac{n}{\sigma^2}.$$

$$\frac{\partial^2}{\partial \mu \partial l} \log p(\mu, l|\mathcal{D}) = \frac{\partial}{\partial l}\frac{n(\bar{x}-\mu)}{\sigma^2} = \frac{\partial \sigma}{\partial l}(-2)\frac{n(\bar{x}-\mu)}{\sigma^3} = -2n\frac{\bar{x}-\mu}{\sigma^2}.$$

$$\frac{\partial^2}{\partial l^2} \log p(\mu, l|\mathcal{D}) = \frac{\partial}{\partial l}[-n + \frac{ns^2+n(\bar{x}-\mu)^2}{\sigma^2}]$$

$$= \frac{\partial \sigma}{\partial l}(-2)\frac{ns^2+n(\bar{x}-\mu)^2}{\sigma^3} = -\frac{2}{\sigma^2}(ns^2+n(\bar{x}-\mu)^2).$$

c. Let $\boldsymbol{\theta} = (\mu, l)$. Then the posterior mode for $\boldsymbol{\theta}$ occurs at:

$$\frac{\partial}{\partial \mu} \log p(\mu, l|\mathcal{D}) = \frac{n(\bar{x}-\mu)}{\sigma^2} = 0 \Rightarrow \mu = \bar{x}.$$

$$\frac{\partial}{\partial l} \log p(\mu, l|\mathcal{D}) = -n + \frac{ns^2+n(\bar{x}-\mu)^2}{\sigma^2} = 0 \Rightarrow \sigma^2 = s^2.$$

Therefore, $\boldsymbol{\theta}^* = (\bar{x}, \log s)$.
Evaluating the inverse Hessian at this parameter values gives:

$$\mathbf{H}^{-1}|_{\boldsymbol{\theta}^*} = (\mathbf{H}|_{\boldsymbol{\theta}^*})^{-1} = \begin{pmatrix} -\frac{n}{s^2} & 0 \\ 0 & -2n \end{pmatrix}^{-1} = \begin{pmatrix} -\frac{s^2}{n} & 0 \\ 0 & -\frac{1}{2n} \end{pmatrix}.$$

Therefore, the Laplace approximation to the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is:

$$\mathcal{N}(\boldsymbol{\theta}|(\bar{x}, \log s), \begin{pmatrix} -\frac{s^2}{n} & 0 \\ 0 & -\frac{1}{2n} \end{pmatrix})$$

.

21.2.
Let $\tilde{\boldsymbol{\Sigma}} = \log \boldsymbol{\Sigma}$, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = e^{-\tilde{\boldsymbol{\Sigma}}}$.

$$l = \log p(\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}|\mathcal{D}) = -\frac{N}{2} \log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \text{const}.$$

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = N\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}).$$

$$\frac{\partial l}{\partial \tilde{\boldsymbol{\Sigma}}} = \frac{\partial \boldsymbol{\Lambda}}{\partial \tilde{\boldsymbol{\Sigma}}} [\frac{N}{2}\boldsymbol{\Lambda}^{-1} - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T] = \frac{N}{2} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T.$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\mu}^2} = -N\boldsymbol{\Sigma}^{-1}.$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\mu} \partial \boldsymbol{\Sigma}} = \frac{\partial \boldsymbol{\Lambda}}{\partial \tilde{\boldsymbol{\Sigma}}} N(\bar{\mathbf{x}} - \boldsymbol{\mu}) = -N\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}).$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\Sigma}^2} = \frac{\partial \boldsymbol{\Lambda}}{\partial \tilde{\boldsymbol{\Sigma}}} [-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T] = \frac{1}{2}\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}})$. Then the posterior mode for $\boldsymbol{\theta}$ occurs at:

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = 0 \Rightarrow \boldsymbol{\mu} = \bar{\mathbf{x}}.$$

$$\frac{\partial l}{\partial \boldsymbol{\Sigma}} = 0 \Rightarrow \boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

Therefore, $\boldsymbol{\theta}^* = (\bar{\mathbf{x}}, \log[\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T])$.
Evaluating the inverse Hessian at this parameter values gives:

$$\mathbf{H}^{-1}|_{\boldsymbol{\theta}^*} = (\mathbf{H}|_{\boldsymbol{\theta}^*})^{-1} = \begin{pmatrix} -N^2[\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T]^{-1} & 0 \\ 0 & -\frac{N}{2}\mathbf{I} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} -\frac{1}{N^2}\sum_{i=1}^N(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T & 0 \\ 0 & -\frac{2}{N}\mathbf{I} \end{pmatrix}.$$

Therefore, the Laplace approximation for $p(\boldsymbol{\mu}, \log \boldsymbol{\Sigma}|\mathcal{D})$ is a joint Gaussian of

$$\mathcal{N}(\boldsymbol{\mu}|\bar{\mathbf{x}}, -\frac{1}{N^2}\sum_{i=1}^N(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T$$

and

$$\mathcal{N}(\log \boldsymbol{\Sigma}|\log[\frac{1}{N}\sum_{i=1}^N(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T], -\frac{2}{N}\mathbf{I}).$$

21.3.

$$\log p(\mathcal{D}|\mu,\lambda) = \frac{N}{2}\log\lambda - \frac{\lambda}{2}\sum_{i=1}^N(x_i-\mu)^2 + \text{const}$$

$$\mathbb{E}[\log p(\mathcal{D}|\mu,\lambda)] = \frac{N}{2}\mathbb{E}[\log\lambda] - \frac{1}{2}\mathbb{E}[\lambda]\sum_{i=1}^N\mathbb{E}[(x_i-\mu)^2] + \text{const}$$

$$= \frac{N}{2}(\phi(a_N)-\log b_N) - \frac{Na_N}{2b_N}((\mu_N-\bar{x})^2+\hat{\sigma}^2+\frac{1}{\kappa_N}) + \text{const}$$

$$\log p(\lambda) = (a_0-1)\log\lambda - b_0\lambda + \text{const}$$

$$\mathbb{E}[\log p(\lambda)] = (a_0-1)\mathbb{E}[\log\lambda] - b_0\mathbb{E}[\lambda] + \text{const}$$

$$= (a_0-1)(\phi(a_N)-\log b_N) - b_0\frac{a_N}{b_N} + \text{const}$$

$$\log p(\mu|\lambda) = \frac{1}{2}\log(\lambda) - \frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2 + \text{const}$$

$$\mathbb{E}[\log p(\mu|\lambda)] = \frac{1}{2}\mathbb{E}[\log(\lambda)] - \frac{\kappa_0}{2}\mathbb{E}[\lambda(\mu-\mu_0)^2] + \text{const}$$

$$= \frac{1}{2}(\phi(a_N)-\log b_N) + \frac{\kappa_0}{2}[\frac{a_N}{b_N}]((\mu_N-\mu_0)^2+\frac{1}{\kappa_N}) + \text{const}$$

$$\Rightarrow L(q) = \mathbb{E}[\log p(\mathcal{D}|\mu,\lambda)] + \mathbb{E}[\log p(\mu|\lambda)] + \mathbb{E}[\log p(\lambda)] - \mathbb{E}[\log q(\mu)] - \mathbb{E}[\log q(\lambda)]$$

$$= \frac{1}{2}\log\frac{1}{\kappa_N} + \log\Gamma(a_N) - a_N\log b_N + \text{const}.$$

21.4.

$$L(q) = \sum_{\mathbf{z}} \int q(\mathbf{z}, \boldsymbol{\theta}) \log \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{q(\mathbf{z}, \boldsymbol{\theta})} d\boldsymbol{\theta}$$
$$= \mathbb{E}[\log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}[\log p(\mathbf{z}|\boldsymbol{\pi})] + \mathbb{E}[\log p(\boldsymbol{\pi})] + \mathbb{E}[\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda})]$$
$$-\mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log q(\boldsymbol{\pi})] - \mathbb{E}[\log q(\boldsymbol{\mu}, \boldsymbol{\Lambda})].$$

$$\log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_i \sum_k z_{ik} \log \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})$$
$$= \sum_i \sum_k z_{ik}[\frac{1}{2}\log|\boldsymbol{\Lambda}_k| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\mathbf{x}_i - \boldsymbol{\mu}_k)] + \text{const}$$

$$\mathbb{E}[\log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]$$
$$= \sum_k N_k \frac{1}{2}\mathbb{E}[\log|\boldsymbol{\Lambda}_k|] - \sum_i \sum_k \frac{1}{2} r_{ik} \mathbb{E}[(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\mathbf{x}_i - \boldsymbol{\mu}_k)] + \text{const}$$
$$= \sum_k \frac{1}{2} N_k \log \tilde{\Lambda}_k - \sum_i \sum_k \frac{1}{2} r_{ik}(D\beta_k^{-1} + \nu_k(\mathbf{x}_i - \mathbf{m}_k)^T \boldsymbol{L}_k (\mathbf{x}_i - \mathbf{m}_k)) + \text{const}$$
$$= \frac{1}{2}\sum_k N_k[\log \tilde{\Lambda}_k - D\beta_k^{-1}] - \sum_i \sum_k \frac{1}{2} r_{ik}\nu_k(\mathbf{x}_i - \mathbf{m}_k)^T \boldsymbol{L}_k (\mathbf{x}_i - \mathbf{m}_k) + \text{const}$$
$$= \frac{1}{2}\sum_k N_k[\log \tilde{\Lambda}_k - D\beta_k^{-1}] - \sum_i \sum_k \frac{1}{2} r_{ik}\nu_k \text{tr}[(\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T \boldsymbol{L}_k] + \text{const}$$
$$= \frac{1}{2}\sum_k N_k[\log \tilde{\Lambda}_k - D\beta_k^{-1}] - \sum_i \sum_k \frac{1}{2} r_{ik}\nu_k \text{tr}[(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T \boldsymbol{L}_k]$$
$$+ \sum_i \sum_k \frac{1}{2} r_{ik}\nu_k \text{tr}[(\bar{\mathbf{x}}_k - \mathbf{m}_k)(\bar{\mathbf{x}}_k - \mathbf{m}_k)^T \boldsymbol{L}_k] + \text{const}$$
$$= \frac{1}{2}\sum_k N_k[\log \tilde{\Lambda}_k - D\beta_k^{-1} - \nu_k \text{tr}(\mathbf{S}_k \mathbf{L}_k) - \nu_k(\bar{\mathbf{x}}_k - \mathbf{m}_k)^T \mathbf{L}_k (\bar{\mathbf{x}}_k - \mathbf{m}_k)] + \text{const}.$$

$$\log p(\mathbf{z}|\boldsymbol{\pi}) = \sum_i \sum_k z_{nk} \log \pi_k$$

$$\mathbb{E}[\log p(\mathbf{z}|\boldsymbol{\pi})] = \sum_i \sum_k r_{ik} \log \tilde{\pi}_k.$$

$$\log p(\boldsymbol{\pi}) = \log \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = \log C_{\mathrm{dir}}(\boldsymbol{\alpha}_0) + \sum_k (\alpha_0 - 1) \log \pi_k$$

$$\mathbb{E}[\log p(\boldsymbol{\pi})] = \log C_{\mathrm{dir}}(\boldsymbol{\alpha}_0) + (\alpha_0 - 1) \sum_k \log \tilde{\pi}_k.$$

$$\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_k [\log \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) + \log \mathrm{Wi}(\boldsymbol{\Lambda}_k|\mathbf{L}_0, \nu_0)]$$

$$\mathbb{E}[\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_k [\frac{1}{2} \log \mathbb{E}[|\beta_0 \boldsymbol{\Lambda}_k|] - \frac{1}{2} \mathbb{E}[(\boldsymbol{\mu}_k - \mathbf{m}_0)^T \beta_0 \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0)]$$

$$+ \log C_{\mathrm{Wi}}(\mathbf{L}_0, \nu_0) + \frac{\nu_0 - D - 1}{2} \mathbb{E}[\log \boldsymbol{\Lambda}_k] - \frac{1}{2} \mathbb{E}[\mathrm{tr}(\mathbf{L}_0^{-1} \boldsymbol{\Lambda}_k)]] + \mathrm{const}$$

$$= \sum_k [\frac{1}{2}[D \log \beta_0 + \log \tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k} - \beta_0 \nu_k (\mathbf{m}_k - \mathbf{m}_0)^T \mathbf{L}_k (\mathbf{m}_k - \mathbf{m}_0)]$$

$$+ \log C_{\mathrm{Wi}}(\mathbf{L}_0, \nu_0) + \frac{\nu_0 - D - 1}{2} \log \tilde{\Lambda}_k - \frac{1}{2} \nu_k \mathrm{tr}(\mathbf{L}_0^{-1} \mathbf{L}_k)] + \mathrm{const}.$$

$$\log q(\mathbf{z}) = \sum_i \sum_k z_{ik} \log r_{ik}$$

$$\mathbb{E}[\log q(\mathbf{z})] = \sum_i \sum_k r_{ik} \log r_{ik}.$$

$$\log q(\boldsymbol{\pi}) = \log \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \log C_{\mathrm{dir}}(\boldsymbol{\alpha}) + \sum_k (\alpha_k - 1) \log \pi_k$$

$$\mathbb{E}[\log q(\boldsymbol{\pi})] = \log C_{\mathrm{dir}}(\boldsymbol{\alpha}) + \sum_k (\alpha_k - 1) \log \tilde{\pi}_k.$$

$$\log q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_k [\log \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) + \log \mathrm{Wi}(\boldsymbol{\Lambda}_k|\mathbf{L}_k, \nu_k)]$$

$$\mathbb{E}[\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_k [\frac{1}{2}\log \mathbb{E}[|\beta_k \boldsymbol{\Lambda}_k|] - \frac{1}{2}\mathbb{E}[(\boldsymbol{\mu}_k - \mathbf{m}_k)^T \beta_k \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_k)]$$
$$-\mathbb{H}[q(\boldsymbol{\Lambda}_k)]] + \text{const}$$
$$= \sum_k [\frac{1}{2}\log \tilde{\Lambda}_k + \frac{D}{2}\log \beta_k - \mathbb{H}[q(\boldsymbol{\Lambda}_k)]] + \text{const.}$$

21.5.
We use the two properties:
The first one is that each marginal of Dirichlet distribution is Beta distribution with: $\pi_k \sim \mathrm{B}(\alpha_k, \sum_{k'} \alpha_{k'} - \alpha_k)$.
The second one is that for Beta distribution, the logarithm of the geometric mean is given by $\mathbb{E}[\ln X] \sim \phi(\alpha) - \phi(\alpha + \beta)$ for $X \sim \mathrm{B}(\alpha, \beta)$.
Combining these, we have: $\mathbb{E}[\log \pi_k] = \phi(\alpha_k) - \phi(\sum_{k'} \alpha_{k'})$.

21.6.
By the property of probability mass function:

$$q_i(1) + q_i(-1) = 1$$

By the property of mean:

$$\mu_i = 1 \cdot q_i(1) + (-1) \cdot q_i(-1)$$

Combining these, we get:

$$q_i(x_i) = \frac{1 + x_i \mu_i}{2}$$

We aim to minimize the KL divergence of q with respect to p, to get:

$$\mathbb{KL}(q(x_i)||p(x_i)) = -\sum_{j \in \text{nbr}(i)} W_{ij}\mu_i\mu_j - \sum_{x_i} L_i(x_i) q_i(x_i) - \sum_{x_i} q(x_i)\log q(x_i)$$

$$= -\sum_{j \in \text{nbr}(i)} W_{ij}\mu_i\mu_j - (\frac{L_i^+(1+\mu_i)}{2} + \frac{L_i^-(1-\mu_i)}{2}) - (\frac{1+\mu_i}{2}\log\frac{1+\mu_i}{2} + \frac{1-\mu_i}{2}\log\frac{1-\mu_i}{2})$$

Differentiating this with respect to μ_i, we get:

$$\frac{\partial \mathbb{KL}(q(x_i)||p(x_i))}{\partial \mu_i} = -\sum_{j \in \text{nbr}(i)} W_{ij}\mu_j - \frac{L_i^+ - L_i^-}{2} + \log\frac{\mu_i}{1-\mu_i} = 0$$

$$\Rightarrow \mu_i = \tanh\Big(\sum_{j\in \mathrm{nbr}(i)} W_{ij}\mu_j + \frac{L_i^+ - L_i^-}{2}\Big).$$

21.7.

$$\mathbb{KL}(p||q) = \sum_x\sum_y p(x,y)\log p(x,y) - \sum_x\sum_y p(x,y)\log q(x) - \sum_y\sum_x p(x,y)\log q(y)$$
$$= \sum_x\sum_y p(x,y)\log p(x,y) - \sum_x p(x)\log q(x) - \sum_y p(y)\log q(y)$$
$$= \mathbb{H}(p(x,y)) - \mathbb{H}(p(x)) - \mathbb{H}(p(y)) + \mathbb{KL}(p(x)||q(x)) + \mathbb{KL}(p(y)||q(y))$$
$$= \mathbb{KL}(p(x)||q(x)) + \mathbb{KL}(p(y)||q(y)) + \mathrm{const}$$

To minimize this quantity, we should set $q(x) = p(x)$ and $q(y) = p(y)$.
For reverse KL divergence, the support of q must be a subset of p. Because q is factorized, its support should always be the Cartesian product of supports of marginals $q(x)$ and $q(y)$. So, we have three choices for the support of q:
- $\{1,2\} \times \{1,2\}$
- $\{(3,3)\}$
- $\{(4,4)\}$

Thus the reverse KL for this p has three distinct minima.
For the first case, we have $q(x) = (\frac{1}{2}, \frac{1}{2}, 0, 0)$, $q(y) = (\frac{1}{2}, \frac{1}{2}, 0, 0)$. Computing the KL divergence, we get $\mathbb{KL}(q||p) = \log 2$.
For the second case, we have $q(x) = q(y) = (0,0,1,0)$. Computing the KL divergence, we get $\mathbb{KL}(q||p) = \log 4$.
For the third case, we have $q(x) = q(y) = (0,0,0,1)$. Computing the KL divergence, we get $\mathbb{KL}(q||p) = \log 4$.
If we set $q(x,y) = p(x)p(y)$, the reverse KL divergence becomes infinite, because we have $p(1,3) = 0$ but $q(1,3) = 1/16$.

21.8.
To minimize the KL divergence, we have $\mathbb{E}[E] = \mathbb{E}[E_q]$, which gives:

$$\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[(\mathbf{y}_t - \sum_m \mathbf{W}_m\mathbf{x}_{tm})^T\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \sum_m \mathbf{W}_m\mathbf{x}_{tm})]$$
$$-\sum_m \mathbb{E}[\mathbf{x}_{1m}^T\tilde{\boldsymbol{\pi}}_m] - \sum_{t=2}^{T}\sum_m \mathbb{E}[\mathbf{x}_{tm}^T\tilde{\mathbf{A}}_m\mathbf{x}_{t-1,m}]$$

$$= -\sum_{t=1}^{T}\sum_{m}\mathbb{E}[\mathbf{x}_{tm}^T\tilde{\boldsymbol{\xi}}_{tm}] - \sum_{m}\mathbb{E}[\mathbf{x}_{1m}^T\tilde{\boldsymbol{\pi}}_m] - \sum_{t=2}^{T}\sum_{m}\mathbb{E}[\mathbf{x}_{tm}^T\tilde{\mathbf{A}}_m\mathbf{x}_{t-1,m}]$$

Therefore, we have:

$$\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[(\mathbf{y}_t - \sum_m \mathbf{W}_m\mathbf{x}_{tm})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_t - \sum_m \mathbf{W}_m\mathbf{x}_{tm})] = -\sum_{t=1}^{T}\sum_m \mathbb{E}[\mathbf{x}_{tm}^T\tilde{\boldsymbol{\xi}}_{tm}].$$

$$\Rightarrow \frac{1}{2}\mathbb{E}[(\mathbf{y}_t - \sum_m \mathbf{W}_m\mathbf{x}_{tm})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_t - \sum_m \mathbf{W}_m\mathbf{x}_{tm})] = -\sum_m \mathbb{E}[\mathbf{x}_{tm}^T\tilde{\boldsymbol{\xi}}_{tm}]$$

Let $\bar{\mathbf{y}}_{tm} = \mathbf{y}_t - \sum_{l\neq m}\mathbf{W}_l\mathbf{x}_{tl}$. Then we have:

$$\mathbb{E}[\frac{1}{2}\bar{\mathbf{y}}_{tm}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}_{tm} - \bar{\mathbf{y}}_{tm}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}_m\mathbf{x}_{tm} + \frac{1}{2}\mathbf{x}_{tm}^T\mathbf{W}_{tm}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}_{tm}\mathbf{x}_{tm}] = -\sum_m \mathbb{E}[\mathbf{x}_{tm}^T\tilde{\boldsymbol{\xi}}_{tm}].$$

Setting $\mathbf{x}_{tm} = \mathbf{1}$, we have:

$$\mathbb{E}[\frac{1}{2}\bar{\mathbf{y}}_{tm}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}_{tm}] - \tilde{\mathbf{y}}_{tm}\boldsymbol{\Sigma}^{-1}\mathbf{W}_m + \frac{1}{2}\mathrm{diag}(\mathbf{W}_{tm}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}_{tm}) = -\tilde{\boldsymbol{\xi}}_{tm} - \sum_{l\neq m}\mathbb{E}[\mathbf{x}_{tl}^T\tilde{\boldsymbol{\xi}}_{tl}].$$

Setting $\mathbf{x}_{tm} = \mathbf{0}$, we have:

$$\mathbb{E}[\frac{1}{2}\bar{\mathbf{y}}_{tm}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}_{tm}] = -\sum_{l\neq m}\mathbb{E}[\mathbf{x}_{tl}^T\tilde{\boldsymbol{\xi}}_{tl}].$$

Comparing the two equations above, we have:

$$\boldsymbol{\xi}_{tm} = e^{\mathbf{W}_m^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{y}}_{tm} - \frac{1}{2}\boldsymbol{\delta}_m}.$$

21.9.

First we note that the conditional distribution for x_{ij} can be written as

$$p(x_{ij}|\mathbf{z}_i,\boldsymbol{\theta}) = \mathrm{Ber}(x_{ij}|\mathrm{sigm}(\eta_{ij})) = (\frac{1}{1+e^{-\eta_{ij}}})^{x_{ij}}(1-\frac{1}{1+e^{-\eta_{ij}}})^{1-x_{ij}}$$

$$= e^{\eta_{ij}x_{ij}}\frac{e^{-\eta_{ij}}}{1+e^{-\eta_{ij}}} = e^{\eta_{ij}x_{ij}}\mathrm{sigm}(-\eta_{ij}).$$

Applying the JJ bound gives:

$$p(x_{ij}|\mathbf{z}_i,\boldsymbol{\theta}) = e^{\eta_{ij}x_{ij}}\mathrm{sigm}(-\eta_{ij}) \geq e^{\eta_{ij}x_{ij}}\mathrm{sigm}(\xi_{ij})e^{-\frac{\eta_{ij}+\xi_{ij}}{2}-\lambda(\xi_{ij})(\eta_{ij}^2-\xi_{ij}^2)}.$$

Therefore, we have:

$$\log p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) = \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) + \log p(\mathbf{z}_i)$$

$$\geq \log p(\mathbf{z}_i) + \boldsymbol{\eta}_i^T \mathbf{x}_i + \sum_j [\log \operatorname{sigm}(\xi_{ij}) - \frac{1}{2}\xi_{ij} - \lambda(\xi_{ij})(\eta_{ij}^2 - \xi_{ij}^2)] - \frac{1}{2}\boldsymbol{\eta}_i^T \mathbf{1}$$

$$= -\frac{1}{2}\mathbf{z}_i^T \mathbf{z}_i + \boldsymbol{\eta}_i^T (\mathbf{x}_i - \frac{1}{2}\mathbf{1}) - \sum_j \lambda(\xi_{ij})(\eta_{ij}^2) + \text{const}$$

$$= -\frac{1}{2}(\mathbf{z}_i - \boldsymbol{\mu}_i)^T (\mathbf{I} + 2\sum_j \lambda(\xi_{ij})\mathbf{w}_j \mathbf{w}_j^T)(\mathbf{z}_i - \boldsymbol{\mu}_i) + \text{const.}$$

Therefore, the posterior approximation $q(\mathbf{z}_i)$ is Gaussian with covariance

$$\boldsymbol{\Sigma}_i = [\mathbf{I} + 2\sum_j \lambda(\xi_{ij})\mathbf{w}_j \mathbf{w}_j^T]^{-1}$$

Substituting this, we get the mean

$$\boldsymbol{\mu}_i = \boldsymbol{\Sigma}_i [\sum_j (x_{ij} - \frac{1}{2} + 2\lambda(\xi_{ij})\beta_j)\mathbf{w}_j]$$

Before fitting the model, we have to determine the variational parameters $\boldsymbol{\xi}_i$ by maximizing the lower bound on the marginal likelihood using the EM algorithm.
In the E step, we use $\boldsymbol{\xi}_{i,\text{old}}$ to compute the posterior distribution $q(\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
In the M step, we maximize the expected complete data log likelihood

$$Q(\boldsymbol{\xi}_i, \boldsymbol{\xi}_{i,\text{old}}) = \mathbb{E}[\sum_j (\log \operatorname{sigm}(\xi_{ij}) - \frac{1}{2}\xi_{ij} - \lambda(\xi_{ij})(\eta_{ij}^2 - \xi_{ij}^2))] + \text{const.}$$

Setting $\frac{\partial Q}{\partial \xi_{ij}} = 0$, we get

$$\lambda'(\xi_{ij})(\mathbf{w}_j^T \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^T]\mathbf{w}_j + 2\beta_j \mathbf{w}_j^T \mathbb{E}[\mathbf{z}_i] + \beta_j^2 - \xi_{ij}^2) = 0$$

$$\begin{aligned} \Rightarrow \xi_{ij}^2 &= \mathbf{w}_j^T \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^T]\mathbf{w}_j + 2\beta_j \mathbf{w}_j^T \mathbb{E}[\mathbf{z}_i] + \beta_j^2 \\ &= \mathbf{w}_j^T (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T)\mathbf{w}_j + 2\beta_j \mathbf{w}_j^T \boldsymbol{\mu}_i + \beta_j^2. \end{aligned}$$

For optimizing model parameters, we again use EM to increase the variational approximated likelihood with respect to $\mathbf{w}_j$ and β_j.
Let $\tilde{\mathbf{w}}_j = (\mathbf{w}_j; \beta_j)^T$, $\tilde{\mathbf{z}}_i = (\mathbf{z}_i; 1)$.
In the E step, we use $\tilde{\mathbf{w}}_{j,\text{old}}$ to compute the posterior distribution $q(\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
In the M step, we maximize the expected complete data log likelihood

$$Q(\tilde{\mathbf{w}}_j, \tilde{\mathbf{w}}_{j,\text{old}}) = \mathbb{E}[\boldsymbol{\eta}_i^T(\mathbf{x}_i - \frac{1}{2}\mathbf{1}) - \sum_j \lambda(\xi_{ij})(\eta_{ij}^2)] + \text{const}.$$

Setting $\frac{\partial Q}{\partial \tilde{\mathbf{w}}_j} = 0$, we get

$$\sum_j (x_{ij} - \frac{1}{2})\mathbb{E}[\tilde{\mathbf{z}}_i] - \sum_j 2\lambda(\xi_{ij})\mathbb{E}[\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T]\tilde{\mathbf{w}}_j = 0$$

$$\Rightarrow \tilde{\mathbf{w}}_j = [2\sum_j \lambda(\xi_{ij})\mathbb{E}[\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T]]^{-1}[\sum_j (x_{ij} - \frac{1}{2})\mathbb{E}[\tilde{\mathbf{z}}_i]].$$

where

$$\mathbb{E}[\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T] = \begin{pmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^T & 1 \end{pmatrix}$$

and

$$\mathbb{E}[\tilde{\mathbf{z}}_i] = (\boldsymbol{\mu}_i; 1).$$

21.10.
Update factor $q(z_i)$:
The log of the optimized factor is given by

$$\log q(z_i) = \mathbb{E}[\log p(y_i, z_i, \mathbf{w}, \mathbf{x}_i)] + \text{const}$$

$$= \log p(y_i|z_i) + \mathbb{E}[\log p(z_i|\mathbf{x}_i, \mathbf{w})] + \text{const}$$

$$= y_i \log \mathbf{1}_{z_i > 0} + (1 - y_i)\log \mathbf{1}_{z_i \leq 0} - \frac{1}{2}\mathbb{E}[(z_i - \mathbf{w}^T\mathbf{x}_i)^2] + \text{const}$$

Without loss of generality, we assume $y_i = 1$ to get:

$$= \log \mathbf{1}_{z_i > 0} - \frac{1}{2}z_i^2 + z_i\mathbb{E}[\mathbf{w}^T\mathbf{x}_i] + \text{const}$$

Setting $\mu_i = \mathbb{E}[\mathbf{w}^T \mathbf{x}_i]$, we obtain:

$$q(z_i) \propto \mathbf{1}_{z_i>0} \mathcal{N}(z_i|\mu_i, 1).$$

Considering the remaining case $y_i = 0$, we get the truncated univariate gaussian:
$q(z_i) = \mathcal{N}(z_i|\mu_i, 1)\mathbf{1}_{z_i>0}$ if $y_i = 1$, $q(z_i) = \mathcal{N}(z_i|\mu_i, 1)\mathbf{1}_{z_i\leq 0}$ if $y_i = 0$.

Update factor $q(\mathbf{w})$:
We assume a prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{V}_0)$ on $\mathbf{w}$.
The log of the optimized factor is given by

$$\log q(\mathbf{w}) = \mathbb{E}[\log p(\mathbf{y}, \mathbf{z}, \mathbf{w}, \mathbf{X})] + \text{const}$$

$$= \mathbb{E}[\log p(\mathbf{z}|\mathbf{X}, \mathbf{w})] + \mathbb{E}[\log p(\mathbf{w})] + \text{const}$$

$$= -\frac{1}{2}\mathbb{E}[(\mathbf{z} - \mathbf{w}^T\mathbf{X})^2] - \frac{1}{2}\mathbf{w}^T\mathbf{V}_0^{-1}\mathbf{w} + \text{const}$$

$$= \mathbf{w}^T\mathbb{E}[\mathbf{X}^T\mathbf{z}] - \frac{1}{2}\mathbf{w}^T(\mathbf{V}_0^{-1} + \mathbb{E}[\mathbf{X}^T\mathbf{X}])\mathbf{w} + \text{const}.$$

Hence, we get the multivariate normal:
$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{V})$, where

$$\mathbf{V} = (\mathbf{V}_0^{-1} + \mathbb{E}[\mathbf{X}^T\mathbf{X}])^{-1}$$

$$\mathbf{m} = \mathbf{V}\mathbb{E}[\mathbf{X}^T\mathbf{z}].$$

Update factor $q(\mathbf{x}_i)$:
We assume a prior $\mathcal{N}(\mathbf{x}_i|\mathbf{0}, \mathbf{V}_i)$ on $\mathbf{x}_i$.
The log of the optimized factor is given by

$$\log q(\mathbf{x}_i) = \mathbb{E}[\log p(y_i, z_i, \mathbf{w}, \mathbf{x}_i)] + \text{const}$$

$$= \mathbb{E}[\log p(z_i|\mathbf{x}_i, \mathbf{w})] + \mathbb{E}[\log p(\mathbf{x}_i)] + \text{const}$$

$$= -\frac{1}{2}\mathbb{E}[(z_i - \mathbf{w}^T\mathbf{x}_i)^2] - \frac{1}{2}\mathbf{x}_i^T\mathbf{V}_i^{-1}\mathbf{x}_i + \text{const}$$

$$= \mathbf{x}_i^T\mathbb{E}[z_i\mathbf{w}] - \frac{1}{2}\mathbf{x}_i^T(\mathbf{V}_i^{-1} + \mathbb{E}[\mathbf{w}^T\mathbf{w}])\mathbf{x}_i + \text{const}.$$

Hence, we get the multivariate normal:
$q(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i|\mathbf{m}_i, \mathbf{U}_i)$, where

$$\mathbf{U}_i = (\mathbf{V}_i^{-1} + \mathbb{E}[\mathbf{w}^T\mathbf{w}])^{-1}$$

$$\mathbf{m}_i = \mathbf{U}_i\mathbb{E}[z_i]\mathbb{E}[\mathbf{w}].$$

Computing the expectations:

$$\mathbb{E}[\mathbf{w}] = \mathbf{m}$$
$$\mathbb{E}[\mathbf{w}^T\mathbf{w}] = \mathbf{m}^T\mathbf{m} + \mathrm{tr}(\mathbf{V})$$
$$\mathbb{E}[\mathbf{x}_i] = \mathbf{m}_i$$
$$\mathbb{E}[\mathbf{x}_i^T\mathbf{x}_i] = \mathbf{m}_i^T\mathbf{m}_i + \mathrm{tr}(\mathbf{U}_i)$$

$\mathbb{E}[z_i] = \mu_i + \frac{\phi_i}{1-\Phi_i}$ if $y_i = 1$, $\mathbb{E}[z_i] = \mu_i - \frac{\phi_i}{\Phi_i}$ if $y_i = 0$.