Supplementary material

Yusuke Mukuta MUKUTA@MI.T.U-TOKYO.AC.JP

Graduate School of Information Science and Technology, The University of Tokyo
7–3–1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

Tatsuya Harada HARADA@MI.T.U-TOKYO.AC.JP

Graduate School of Information Science and Technology, The University of Tokyo
7–3–1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

A. The derivation of (10)

A.1. E step

It follows that

$$\begin{aligned}
\log p(z_n|x_n, y_n; W_t, \Psi_t) &= \log p(z_n) + \log(y_n|z_n, x_n) + \mathrm{C} \\
&= -\frac{z_n^T z_n}{2} - \frac{\left((W_z)_t\, z_n - (y_n - (W_x)_t\, x_n)\right)^T (\Psi_t)^{-1} \left((W_z)_t\, z_n - (y_n - (W_x)_t\, x_n)\right)}{2} + \mathrm{C} \\
&= -\frac{\delta z_t^T (\Sigma_z)_t^{-1} \delta z_t}{2} + \mathrm{C},
\end{aligned}$$

where

$$\begin{aligned}
\delta z_t &= z_n - \Sigma_z (W_z)_t^T (\Psi_t)^{-1} (y_n - (W_x)_t x_n) \\
(\Sigma_z)_t &= \left(I_{d_z} + (W_z)_t^T (\Psi_t)^{-1} (W_z)_t\right)^{-1},
\end{aligned}$$

so

$$p(z_n|x_n, y_n; W_t, \Psi_t) = \mathcal{N}\left(\Sigma_z (W_z)_t^T (\Psi_t)^{-1} (y_n - (W_x)_t x_n), \Sigma_z\right).$$

Using this,

$$Q(W_{t+1}, \Psi_{t+1}|W_t, \Psi_t) = \sum_{n=1}^N \left(-\frac{\langle z_n^T z_n \rangle_t}{2} - \frac{\langle \left(y_n - W_{t+1} \begin{pmatrix} x_n \\ z_n \end{pmatrix}\right)^T \Psi_{t+1}^{-1} \left(y_n - W_{t+1} \begin{pmatrix} x_n \\ z_n \end{pmatrix}\right) \rangle_t}{2} - \frac{1}{2} \log \det (\Psi_{t+1}) \right).$$

A.2. M step

It follows that

$$\frac{\partial Q}{\partial W_{t+1}} = \sum_{n=1}^N \langle \Psi_{t+1}^{-1} \left(y_n - W_{t+1} \begin{pmatrix} x_n \\ z_n \end{pmatrix}\right) \begin{pmatrix} x_n \\ z_n \end{pmatrix}^T \rangle_t,$$

so Q is maximized at

$$W_{t+1} = Y \begin{pmatrix} X \\ \langle Z \rangle_t \end{pmatrix}^T \begin{pmatrix} XX^T & X\langle Z \rangle_t^T \\ \langle Z \rangle_t X^T & \langle ZZ^T \rangle_t \end{pmatrix}^{-1}.$$

Also,

$$\frac{\partial Q}{\partial \Psi_{t+1}^{-1}} = -\frac{1}{2}\sum_{n=1}^{N}\left(\langle\left(y_n - W_{t+1}\left(\begin{array}{c} x_n \\ z_n \end{array}\right)\right)\left(y_n - W_{t+1}\left(\begin{array}{c} x_n \\ z_n \end{array}\right)\right)^T\rangle_t - \Psi_{t+1}^T\right),$$

so Q is maximized at

$$\Psi_{t+1}^m = \frac{1}{N}\left(Y^m Y^{mT} - \left(W_{t+1}\left(\begin{array}{c} X \\ \langle Z\rangle_t \end{array}\right) Y^T\right)_{mm}\right).$$

B. The derivation of (17)

The derivation of (17) is as follows.

$$
\begin{aligned}
\log q(z_n) &= \log p(z_n) + \sum_m \langle \log p(y_n^m | z_n, x_n, W^m, \Psi^m)\rangle_\Theta + C \\
&= -\frac{z_n^T z_n}{2} - \frac{1}{2}\sum_m \langle (y_n^m - W_x^m x_n - W_z^m z_n)^T \left(\Psi^m\right)^{-1} \left(y_n^m - W_x^m x_n - W_z^m z_n\right)\rangle_\Theta + C \\
&= -\frac{\delta z_n^T \Sigma_{z_n}^{-1} \delta z_n}{2} + C, \\
\log q(\Psi^m) &= \log p(\Psi^m) + \sum_n \langle \log p(y_n^m | z_n, x_n, W^m, \Psi^m)\rangle_{z_n, \Theta \neq \Psi^m} + C \\
&= -\frac{1}{2}\mathrm{Tr}\left((\Psi^m)^{-1} \left(K_0^m + \left(y_n^m - W_z^m z_n - W_x^m x_n\right)\left(y_n^m - W_z^m z_n - W_x^m x_n\right)^T \right)\right) \\
&- \frac{(\nu_0^m + N - d_m - 1)}{2} \log\det \Psi^m + C, \\
\log q(w_j^m) &= \langle \log p(w_j^m | \alpha)\rangle_\alpha + \sum_n \langle \log p(y_n^m | z_n, x_n, W^m, \Psi^m)\rangle_{z_n, \Theta \neq w_j^m} + C \\
&= -\frac{1}{2}(w_j^m)^T \mathrm{diag}\langle\alpha\rangle w_j^m - \frac{1}{2}(w_j^m)^T \langle (\Psi^m)^{-1}_{j,j}\rangle \begin{pmatrix} XX^T & X\langle Z\rangle^T \\ \langle Z\rangle X^T & \langle ZZ^T\rangle \end{pmatrix} w_j^m \\
&+ \left(\langle (\Psi^m)^{-1}_{j,:}\rangle Y^m \begin{pmatrix} X^T & Z^T \end{pmatrix} - \sum_{l \neq j} \langle (\Psi^m)^{-1}_{j,l}\rangle \langle W_{l,:}^m\rangle \begin{pmatrix} XX^T & X\langle Z\rangle^T \\ \langle Z\rangle X^T & \langle ZZ^T\rangle \end{pmatrix} \right) w_j^m + C, \\
\log q(\alpha_k^m) &= \log p(\alpha_k^m) + \langle \log p(W_{:,k}^m | \alpha_k^m)\rangle_{W_{:,k}^m} + C \\
&= \left(-b_0 - \frac{\langle \|W_{:,k}^m\| \rangle}{2} \right) \alpha_k^m + (a_0 + \frac{d_m}{2} - 1)\log \alpha_k^m + C,
\end{aligned}
$$

where,

$$
\begin{aligned}
\delta z_n &= z_n - \Sigma_{z_n} \sum_m \left(\langle (W_z^m)^T \rangle \langle (\Psi^m)^{-1}\rangle y_n^m - \langle (W_z^m)^T (\Psi^m)^{-1} W_x^m \rangle x_n \right), \\
\Sigma_{z_n} &= \left(I + \sum_m \langle (W_z^m)^T (\Psi^m)^{-1} W_z^m \rangle \right)^{-1}.
\end{aligned}
$$

C. The derivation of (22)

The derivation of (22) is as follows.

$$
\begin{aligned}
\log q(w_j^m) &= \langle \log p(w_j^m|\alpha)\rangle_\alpha + \sum_n \langle \log p(y_n^m|z_n, x_n, W^m, \tau^m)\rangle_{z_n, \Theta \neq w_j^m} + C \\
&= -\frac{1}{2}(w_j^m)^T \left(\mathrm{diag}\langle \alpha^m\rangle + \langle \tau^m\rangle \begin{pmatrix} XX^T & X\langle Z\rangle^T \\ \langle Z\rangle X^T & \langle ZZ^T\rangle \end{pmatrix} \right) w_j^m \\
&+ \langle \tau^m\rangle Y^m_{(j,:)} \begin{pmatrix} X^T & \langle Z^T\rangle \end{pmatrix} w_j^m + C, \\
\log q(z_n) &= \log p(z_n) + \sum_m \langle \log p(y_n^m|z_n, x_n, W^m, \tau^m)\rangle_\Theta + C \\
&= -\frac{\delta z_n^T \Sigma_z^{-1} \delta z_n}{2} + C, \\
\log q(\tau^m) &= \log p(\tau^m) + \sum_n \langle \log p(y_n^m|z_n, x_n, W^m, \tau^m)\rangle_{z_n, \Theta \neq \tau^m} + C \\
&= \left(-b_0 - \frac{1}{2}\sum_n \langle (y_n^m - W_x^m x_n - W_z^m z_n)^T (y_n^m - W_x^m x_n - W_z^m z_n)\rangle_{z_n,\Theta} \right) \tau^m \\
&+ \left(a_0 + \frac{Nd_m}{2} - 1 \right) \log \tau^m + C,
\end{aligned}
$$

where,

$$
\begin{aligned}
\delta z_n &= z_n - \Sigma_{z_n} \sum_m \langle \tau^m\rangle \langle (W_z^m)^T (y_n^m - W_x^m x_n)\rangle, \\
\Sigma_{z_n} &= \left(I + \sum_m \langle \tau^m\rangle \langle (W_z^m)^T W_z^m\rangle \right)^{-1}.
\end{aligned}
$$

The derivation of update rule for α is the same as that of (17).