Final Exam

20213073 Donggyu Kim

June 17, 2021

1. (a) The marginal distribution $p(\boldsymbol{x};\boldsymbol{\theta})$ is

$$\begin{aligned}
p(\boldsymbol{x};\boldsymbol{\theta}) &= \int p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})\,d\boldsymbol{z} \\
&= \int p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})p(\boldsymbol{z};\boldsymbol{\theta})\,d\boldsymbol{z} \\
&= \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z}+\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)\mathcal{N}(\boldsymbol{z}|\boldsymbol{0}_h,\boldsymbol{I}_h)\,d\boldsymbol{z} \\
&= \int \frac{1}{(2\pi)^{\frac{d+h}{2}}\sigma^d}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{W}\boldsymbol{z}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{W}\boldsymbol{z}-\boldsymbol{\mu})-\frac{1}{2}\boldsymbol{z}^\top\boldsymbol{z}\right)d\boldsymbol{z}
\end{aligned}$$

The exponent part can be rewritten as follows

$$\begin{aligned}
&-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{W}\boldsymbol{z}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{W}\boldsymbol{z}-\boldsymbol{\mu})-\frac{1}{2}\boldsymbol{z}^\top\boldsymbol{z} \\
=&-\frac{1}{2\sigma^2}\left(\boldsymbol{z}^\top\boldsymbol{V}\boldsymbol{z}-2\boldsymbol{z}^\top\boldsymbol{W}^\top(\boldsymbol{x}-\boldsymbol{\mu})+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
=&-\frac{1}{2\sigma^2}\left((\boldsymbol{z}-\boldsymbol{\alpha})^\top\boldsymbol{V}(\boldsymbol{z}-\boldsymbol{\alpha})+(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})-\boldsymbol{\alpha}^\top\boldsymbol{V}\boldsymbol{\alpha}\right)
\end{aligned}$$

where $\boldsymbol{V}=\boldsymbol{W}^\top\boldsymbol{W}+\sigma^2\boldsymbol{I}_h$ and $\boldsymbol{\alpha}=\boldsymbol{V}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}-\boldsymbol{\mu})$.
Using the matrix inverse identity, we further get

$$\begin{aligned}
&(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})-\boldsymbol{\alpha}^\top\boldsymbol{V}\boldsymbol{\alpha} \\
=&(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{I}_d-\boldsymbol{W}\boldsymbol{V}^{-1}\boldsymbol{W}^\top)(\boldsymbol{x}-\boldsymbol{\mu}) \\
=&(\boldsymbol{x}-\boldsymbol{\mu})^\top\left(\boldsymbol{I}_d-(\boldsymbol{W}/\sigma)\left(\boldsymbol{I}_h+\boldsymbol{W}^\top\boldsymbol{W}/\sigma^2\right)^{-1}(\boldsymbol{W}^\top/\sigma)\right)(\boldsymbol{x}-\boldsymbol{\mu}) \\
=&(\boldsymbol{x}-\boldsymbol{\mu})^\top\left(\boldsymbol{I}_d+\boldsymbol{W}\boldsymbol{W}^\top/\sigma^2\right)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})
\end{aligned}$$

Using the fact $\det(\boldsymbol{W}\boldsymbol{W}^\top+\sigma^2\boldsymbol{I}_d)=\sigma^{2d}\det(\boldsymbol{I}_h+\boldsymbol{W}^\top\boldsymbol{W}/\sigma^2)$, we have

$$\begin{aligned}
&\int\exp\left(-\frac{1}{2}\left((\boldsymbol{z}-\boldsymbol{\alpha})^\top(\boldsymbol{V}/\sigma^2)(\boldsymbol{z}-\boldsymbol{\alpha})\right)\right)d\boldsymbol{z} \\
=&\int(2\pi)^{\frac{h}{2}}\det\left((\boldsymbol{V}/\sigma^2)^{-1}\right)^{\frac{1}{2}}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\alpha},(\boldsymbol{V}/\sigma^2)^{-1})\,d\boldsymbol{z} \\
=&(2\pi)^{\frac{h}{2}}\det\left(\boldsymbol{I}_h+\boldsymbol{W}^\top\boldsymbol{W}/\sigma^2\right)^{-\frac{1}{2}} \\
=&(2\pi)^{\frac{h}{2}}\sigma^d\det\left(\boldsymbol{W}\boldsymbol{W}^\top+\sigma^2\boldsymbol{I}_d\right)^{-\frac{1}{2}}
\end{aligned}$$

Thus, we can finally get

$$
\begin{aligned}
p(\boldsymbol{x};\boldsymbol{\theta}) = \int \frac{1}{(2\pi)^{\frac{d+h}{2}}\sigma^d} \exp\Bigg(-\frac{1}{2}\Big(& (\boldsymbol{z}-\boldsymbol{\alpha})^\top (\boldsymbol{V}/\sigma^2)(\boldsymbol{z}-\boldsymbol{\alpha}) \\
& + (\boldsymbol{x}-\boldsymbol{\mu})^\top \left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}_d\right)^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \Big)\Bigg) d\boldsymbol{z} \\
= & \frac{(2\pi)^{\frac{h}{2}}\sigma^d \det\left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d\right)^{-\frac{1}{2}}}{(2\pi)^{\frac{d+h}{2}}\sigma^d} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}_d\right)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
= & \frac{1}{(2\pi)^{\frac{d}{2}} \det\left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d\right)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}_d\right)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
= & \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d)
\end{aligned}
$$

And the conditional distribution $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})$ is

$$
\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}) &= \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} \\
&= \frac{p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})p(\boldsymbol{z};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} \\
&= \frac{\mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z}+\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_d)\mathcal{N}(\boldsymbol{z}|\boldsymbol{0}_h, \boldsymbol{I}_h)}{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d)} \\
&= \frac{1}{(2\pi)^{\frac{h}{2}}\sigma^d \det\left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d\right)^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\alpha})^\top(\boldsymbol{V}/\sigma^2)(\boldsymbol{z}-\boldsymbol{\alpha})\right) \\
&= \frac{1}{(2\pi)^{\frac{h}{2}} \det(\boldsymbol{I}_h + \boldsymbol{W}^\top\boldsymbol{W}/\sigma^2)^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\alpha})^\top(\boldsymbol{V}/\sigma^2)(\boldsymbol{z}-\boldsymbol{\alpha})\right) \\
&= \frac{1}{(2\pi)^{\frac{h}{2}} \det((\boldsymbol{V}/\sigma^2)^{-1})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\alpha})^\top(\boldsymbol{V}/\sigma^2)(\boldsymbol{z}-\boldsymbol{\alpha})\right) \\
&= \mathcal{N}(\boldsymbol{z}|\boldsymbol{\alpha}, (\boldsymbol{V}/\sigma^2)^{-1}) \\
&= \mathcal{N}(\boldsymbol{z}|\boldsymbol{V}^{-1}\boldsymbol{W}^\top(\boldsymbol{x}-\boldsymbol{\mu}), \sigma^2\boldsymbol{V}^{-1})
\end{aligned}
$$

(b) The log-likelihood $\log p(\boldsymbol{X})$ is as follows.

$$
\begin{aligned}
\log p(\boldsymbol{X}) &= \sum_{i=1}^n \log p(\boldsymbol{x};\boldsymbol{\theta}) \\
&= \sum_{i=1}^n \Bigg(-\frac{d}{2}\log 2\pi - \frac{1}{2}\log\det(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d) \\
&\qquad - \frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d\right)^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}) \Bigg)
\end{aligned}
$$

Then we can derive its maximum-likelihood estimator of the parameter $\boldsymbol{\mu}$.

$$
\begin{aligned}
\frac{\partial \log p(\boldsymbol{X})}{\partial \boldsymbol{\mu}} &= \sum_{i=1}^n (\boldsymbol{x}_i-\boldsymbol{\mu})^\top \left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d\right)^{-1} \\
&= \left(\sum_{i=1}^n \boldsymbol{x}_i - n\boldsymbol{\mu}\right)^\top \left(\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}_d\right)^{-1} = 0 \ \Rightarrow \ \boldsymbol{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i
\end{aligned}
$$

(c) The expectation of log-likelihood computed in E-step is

$$\mathbb{E}_{p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}[\log p(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})] = -\frac{d+h}{2}\log 2\pi - d\log\sigma - \frac{1}{2\sigma^2}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top(\boldsymbol{x}_i-\boldsymbol{\mu}) - \frac{1}{2\sigma^2}\mathbb{E}_{p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}\left[\boldsymbol{z}_i^\top\boldsymbol{V}\boldsymbol{z}_i - 2\boldsymbol{z}_i^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\right]$$

Let $\boldsymbol{\alpha}_{it} = \boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top(\boldsymbol{x}_i-\boldsymbol{\mu})$. Then,

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}\left[\boldsymbol{z}_i^\top\boldsymbol{V}\boldsymbol{z}_i - 2\boldsymbol{z}_i^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\right]\\
=&\mathbb{E}_{\mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\alpha}_{it},\sigma_t^2\boldsymbol{V}_t^{-1})}\left[(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})^\top\boldsymbol{V}(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it}) + 2\boldsymbol{z}_i^\top\boldsymbol{V}\boldsymbol{\alpha}_{it} - \boldsymbol{\alpha}_{it}^\top\boldsymbol{V}\boldsymbol{\alpha}_{it} - 2\boldsymbol{z}_i^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\right]\\
=&\mathbb{E}_{\mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\alpha}_{it},\sigma_t^2\boldsymbol{V}_t^{-1})}\left[(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})^\top\boldsymbol{V}(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})\right] + \boldsymbol{\alpha}_{it}^\top\boldsymbol{V}\boldsymbol{\alpha}_{it} - 2\boldsymbol{\alpha}_{it}^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\\
=&\mathbb{E}_{\mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\alpha}_{it},\sigma_t^2\boldsymbol{V}_t^{-1})}\left[\operatorname{Tr}\left(\boldsymbol{V}(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})^\top\right)\right] + \boldsymbol{\alpha}_{it}^\top\boldsymbol{V}\boldsymbol{\alpha}_{it} - 2\boldsymbol{\alpha}_{it}^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\\
=&\operatorname{Tr}\left(\mathbb{E}_{\mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\alpha}_{it},\sigma_t^2\boldsymbol{V}_t^{-1})}\left[\boldsymbol{V}(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})^\top\right]\right) + \boldsymbol{\alpha}_{it}^\top\boldsymbol{V}\boldsymbol{\alpha}_{it} - 2\boldsymbol{\alpha}_{it}^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\\
=&\operatorname{Tr}\left(\boldsymbol{V}\mathbb{E}_{\mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\alpha}_{it},\sigma_t^2\boldsymbol{V}_t^{-1})}\left[(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})(\boldsymbol{z}_i-\boldsymbol{\alpha}_{it})^\top\right]\right) + \boldsymbol{\alpha}_{it}^\top\boldsymbol{V}\boldsymbol{\alpha}_{it} - 2\boldsymbol{\alpha}_{it}^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\\
=&\sigma_t^2\operatorname{Tr}\left(\boldsymbol{V}\boldsymbol{V}_t^{-1}\right) + \boldsymbol{\alpha}_{it}^\top\boldsymbol{V}\boldsymbol{\alpha}_{it} - 2\boldsymbol{\alpha}_{it}^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})
\end{aligned}$$

Since $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ is fixed, we can also compute

$$\begin{aligned}
&\sum_{i=1}^n\left(\boldsymbol{\alpha}_{it}^\top\boldsymbol{V}\boldsymbol{\alpha}_{it} - 2\boldsymbol{\alpha}_{it}^\top\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\right)\\
=&\sum_{i=1}^n(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{V}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top(\boldsymbol{x}_i-\boldsymbol{\mu}) - 2\sum_{i=1}^n(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top(\boldsymbol{x}_i-\boldsymbol{\mu})\\
=&n\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{V}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right) - 2n\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{S}\right)
\end{aligned}$$

Therefore, the complete-data log-likelihood $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ is computed as

$$\begin{aligned}
\tilde{\mathcal{L}}(\boldsymbol{\theta}) &= \sum_{i=1}^n \mathbb{E}_{p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_t)}[\log p(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\theta})]\\
&= -\frac{n(d+h)}{2}\log 2\pi - nd\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n(\boldsymbol{x}_i-\boldsymbol{\mu})^\top(\boldsymbol{x}_i-\boldsymbol{\mu}) - \frac{n\sigma_t^2}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{V}\boldsymbol{V}_t^{-1}\right)\\
&\qquad - \frac{n}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{V}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right) + \frac{n}{\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{S}\right)\\
&= -\frac{n(d+h)}{2}\log 2\pi - nd\log\sigma - \frac{n}{2\sigma^2}\operatorname{Tr}(\boldsymbol{S}) - \frac{n\sigma_t^2}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{V}\boldsymbol{V}_t^{-1}\right)\\
&\qquad - \frac{n}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{V}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right) + \frac{n}{\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{S}\right)
\end{aligned}$$

Taking the gradient w.r.t. $\boldsymbol{W}$,

$$\frac{\partial\tilde{\mathcal{L}}(\boldsymbol{\theta})}{\partial\boldsymbol{W}} = -\frac{n\sigma_t^2}{2\sigma^2}\frac{\partial\operatorname{Tr}\left(\boldsymbol{V}\boldsymbol{V}_t^{-1}\right)}{\partial\boldsymbol{W}} - \frac{n}{2\sigma^2}\frac{\partial\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{V}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right)}{\partial\boldsymbol{W}} + \frac{n}{\sigma^2}\frac{\partial\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{S}\right)}{\partial\boldsymbol{W}}$$

First we compute

$$\begin{aligned}\frac{\partial \operatorname{Tr}\left(\boldsymbol{V}\boldsymbol{V}_t^{-1}\right)}{\partial \boldsymbol{W}} &= \frac{\partial \operatorname{Tr}\left((\boldsymbol{W}^\top \boldsymbol{W} + \sigma^2 \boldsymbol{I}_h)\boldsymbol{V}_t^{-1}\right)}{\partial \boldsymbol{W}} \\ &= \frac{\partial \operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{V}_t^{-1}\right)}{\partial \boldsymbol{W}} \\ &= 2\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\end{aligned}$$

Next,

$$\begin{aligned}\frac{\partial \operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{V}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\right)}{\partial \boldsymbol{W}} &= \frac{\partial \operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}(\boldsymbol{W}^\top \boldsymbol{W} + \sigma^2 \boldsymbol{I}_h)\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\right)}{\partial \boldsymbol{W}} \\ &= \frac{\partial \operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\right)}{\partial \boldsymbol{W}} \\ &= \left(\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1} + \left(\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\right)^\top\right)\boldsymbol{W}^\top \\ &= 2\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\end{aligned}$$

Lastly,

$$\frac{\partial \operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top \boldsymbol{S}\right)}{\partial \boldsymbol{W}} = \left(\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\right)^\top = \boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}$$

So the final form of the gradient w.r.t. $\boldsymbol{W}$ is

$$\begin{aligned}\frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\theta})}{\partial \boldsymbol{W}} &= -\frac{n\sigma_t^2}{\sigma^2}\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top - \frac{n}{\sigma^2}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top + \frac{n}{\sigma^2}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S} \\ &= -\frac{n}{\sigma^2}\boldsymbol{V}_t^{-1}\left(\left(\sigma_t^2\boldsymbol{I}_h + \boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\right)\boldsymbol{W}^\top - \boldsymbol{W}_t^\top \boldsymbol{S}\right)\end{aligned}$$

Thus, the M-step update equation for $\boldsymbol{W}_{t+1}$ is

$$\begin{aligned}\frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\theta})}{\partial \boldsymbol{W}} = 0 \;\Rightarrow\; \boldsymbol{W}_{t+1} &= \left(\left(\sigma_t^2\boldsymbol{I}_h + \boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\right)^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\right)^\top \\ &= \boldsymbol{S}\boldsymbol{W}_t\left(\sigma_t^2\boldsymbol{I}_h + \boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\boldsymbol{W}_t\right)^{-1}\end{aligned}$$

Now let's take the gradient w.r.t. σ.

$$\begin{aligned}\frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\theta})}{\partial \sigma} = &-\frac{nd}{\sigma} + \frac{n}{\sigma^3}\operatorname{Tr}(\boldsymbol{S}) - \frac{\partial}{\partial \sigma}\left(\frac{n\sigma_t^2}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{V}\boldsymbol{V}_t^{-1}\right)\right) \\ &- \frac{\partial}{\partial \sigma}\left(\frac{n}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{V}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top \boldsymbol{S}\right)\right) - \frac{2n}{\sigma^3}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top \boldsymbol{S}\right)\end{aligned}$$

Firstly,

$$\begin{aligned}\frac{\partial}{\partial \sigma}\left(\frac{n\sigma_t^2}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{V}\boldsymbol{V}_t^{-1}\right)\right) &= \frac{\partial}{\partial \sigma}\left(\frac{n\sigma_t^2}{2\sigma^2}\operatorname{Tr}\left((\boldsymbol{W}^\top \boldsymbol{W} + \sigma^2\boldsymbol{I}_h)\boldsymbol{V}_t^{-1}\right)\right) \\ &= \frac{\partial}{\partial \sigma}\left(\frac{n\sigma_t^2}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{V}_t^{-1}\right)\right) \\ &= -\frac{n\sigma_t^2}{\sigma^3}\operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{V}_t^{-1}\right)\end{aligned}$$

And then we compute

$$
\begin{aligned}
&\frac{\partial}{\partial \sigma}\left(\frac{n}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{V}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right)\right)\\
=&\frac{\partial}{\partial \sigma}\left(\frac{n}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}(\boldsymbol{W}^\top\boldsymbol{W}+\sigma^2\boldsymbol{I}_h)\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right)\right)\\
=&\frac{\partial}{\partial \sigma}\left(\frac{n}{2\sigma^2}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right)\right)\\
=&-\frac{n}{\sigma^3}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right)
\end{aligned}
$$

Using the properties of trace, we have

$$
\begin{aligned}
\frac{\partial\tilde{\mathcal{L}}(\boldsymbol{\theta})}{\partial\sigma} &= -\frac{nd}{\sigma}+\frac{n}{\sigma^3}\operatorname{Tr}(\boldsymbol{S})+\frac{n\sigma_t^2}{\sigma^3}\operatorname{Tr}\left(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{V}_t^{-1}\right)\\
&\qquad+\frac{n}{\sigma^3}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right)-\frac{2n}{\sigma^3}\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{S}\right)\\
&=-\frac{n}{\sigma^3}\Bigg(\sigma^2 d-\operatorname{Tr}(\boldsymbol{S})-\sigma_t^2\operatorname{Tr}\left(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{V}_t^{-1}\right)\\
&\qquad-\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\right)+2\operatorname{Tr}\left(\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\boldsymbol{S}\right)\Bigg)\\
&=-\frac{n}{\sigma^3}\Bigg(\sigma^2 d-\operatorname{Tr}(\boldsymbol{S})-\sigma_t^2\operatorname{Tr}\left(\boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)\\
&\qquad-\operatorname{Tr}\left(\boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)+2\operatorname{Tr}\left(\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)\Bigg)\\
&=-\frac{n}{\sigma^3}\Bigg(\sigma^2 d-\operatorname{Tr}(\boldsymbol{S})-\operatorname{Tr}\left(\sigma_t^2\boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top+\boldsymbol{W}\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)\\
&\qquad\qquad+2\operatorname{Tr}\left(\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)\Bigg)\\
&=-\frac{n}{\sigma^3}\Bigg(\sigma^2 d-\operatorname{Tr}(\boldsymbol{S})-\operatorname{Tr}\left(\boldsymbol{W}\left(\sigma_t^2\boldsymbol{I}_h+\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\boldsymbol{W}_t\right)\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)\\
&\qquad\qquad+2\operatorname{Tr}\left(\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)\Bigg)\\
&=-\frac{n}{\sigma^3}\Bigg(\sigma^2 d-\operatorname{Tr}(\boldsymbol{S})-\operatorname{Tr}\left(\boldsymbol{W}\left(\sigma_t^2\boldsymbol{I}_h+\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\boldsymbol{W}_t\right)\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)\\
&\qquad\qquad+2\operatorname{Tr}\left(\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}^\top\right)\Bigg)
\end{aligned}
$$

Substituting $\boldsymbol{W}=\boldsymbol{W}_{t+1}=\boldsymbol{S}\boldsymbol{W}_t\left(\sigma_t^2\boldsymbol{I}_h+\boldsymbol{V}_t^{-1}\boldsymbol{W}_t^\top\boldsymbol{S}\boldsymbol{W}_t\right)^{-1}$, we finally get the M-step update equation for σ_{t+1}.

$$
\begin{aligned}
\frac{\partial\tilde{\mathcal{L}}(\boldsymbol{\theta})}{\partial\sigma} &= -\frac{n}{\sigma^3}\left(\sigma^2 d-\operatorname{Tr}(\boldsymbol{S})-\operatorname{Tr}\left(\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}_{t+1}^\top\right)+2\operatorname{Tr}\left(\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}_{t+1}^\top\right)\right)\\
&=-\frac{n}{\sigma^3}\left(\sigma^2 d-\operatorname{Tr}\left(\boldsymbol{S}-\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}_{t+1}^\top\right)\right)=0\\
&\Rightarrow\ \sigma_{t+1}^2=\frac{1}{d}\operatorname{Tr}\left(\boldsymbol{S}-\boldsymbol{S}\boldsymbol{W}_t\boldsymbol{V}_t^{-1}\boldsymbol{W}_{t+1}^\top\right)
\end{aligned}
$$

2. (a) According to Problem 1-(a), we can directly get

$$
\begin{aligned}
p(\boldsymbol{z}|c,\boldsymbol{x};\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{z}|\boldsymbol{V}_c^{-1}\boldsymbol{W}_c^\top(\boldsymbol{x}-\boldsymbol{\mu}_c), \sigma_c^2\boldsymbol{V}_c^{-1}) \\
p(\boldsymbol{x}|c;\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{W}_c\boldsymbol{W}_c^\top + \sigma_c^2\boldsymbol{I}_d)
\end{aligned}
$$

where $\boldsymbol{V}_c = \boldsymbol{W}_c^\top\boldsymbol{W}_c + \sigma_c^2\boldsymbol{I}_h$.
So we can compute

$$
\begin{aligned}
p(\boldsymbol{x};\boldsymbol{\theta}) &= \sum_{j=1}^{k} p(c=j,\boldsymbol{x};\boldsymbol{\theta}) \\
&= \sum_{j=1}^{k} p(c=j;\boldsymbol{\theta})p(\boldsymbol{x}|c=j;\boldsymbol{\theta}) \\
&= \sum_{j=1}^{k} \pi_j\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{W}_j\boldsymbol{W}_j^\top + \sigma_j^2\boldsymbol{I}_d)
\end{aligned}
$$

Then we can derive

$$
\begin{aligned}
p(c|\boldsymbol{x};\boldsymbol{\theta}) &= \frac{p(c,\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} \\
&= \frac{\pi_c\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{W}_c\boldsymbol{W}_c^\top + \sigma_c^2\boldsymbol{I}_d)}{\sum_{j=1}^{k}\pi_j\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{W}_j\boldsymbol{W}_j^\top + \sigma_j^2\boldsymbol{I}_d)}
\end{aligned}
$$

(b) Since $\boldsymbol{\pi}$ is under the constraint $\sum_{j=1}^{k}\pi_j = 1$, we can introduce the Lagrangian multiplier

$$
\mathcal{L}(\boldsymbol{\theta},\lambda) = \mathcal{L}(\boldsymbol{\theta}) + \lambda\left(1 - \sum_{j=1}^{k}\pi_j\right)
$$

By taking the gradient w.r.t. π_j, we have

$$
\frac{\partial\mathcal{L}(\boldsymbol{\theta},\lambda)}{\partial\pi_j} = \sum_{i=1}^{n}\frac{r_{t,i,j}}{\pi_j} - \lambda = 0 \;\Rightarrow\; \pi_{t+1,j} = \frac{1}{\lambda}\sum_{i=1}^{n} r_{t,i,j}
$$

Using the constraint $\sum_{j=1}^{k}\pi_j = 1$ again, we can finalize the M-step update for $\boldsymbol{\pi}$.

$$
\begin{aligned}
\sum_{j=1}^{k}\pi_{t+1,j} &= \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{k} r_{t,i,j} \\
&= \frac{1}{\lambda}\sum_{i=1}^{n} 1 = \frac{n}{\lambda} = 1 \;\Rightarrow\; \lambda = n \\
\therefore \pi_{t+1,j} &= \frac{1}{n}\sum_{i=1}^{n} r_{t,i,j}
\end{aligned}
$$

By taking the gradient w.r.t. $\boldsymbol{\mu}$, we can get the M-step update for $\boldsymbol{\mu}$.

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\mu}_j} &= \sum_{i=1}^{n} r_{t,i,j} \frac{\partial}{\partial \boldsymbol{\mu}_j} \left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top (\boldsymbol{W}_j \boldsymbol{W}_j^\top + \sigma_j^2 \boldsymbol{I}_d)^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j) \right) \\
&= \sum_{i=1}^{n} r_{t,i,j} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top (\boldsymbol{W}_j \boldsymbol{W}_j^\top + \sigma_j^2 \boldsymbol{I}_d)^{-1} = \boldsymbol{0} \\
&\Rightarrow \ \boldsymbol{\mu}_{t+1,j} = \frac{\sum_{i=1}^{n} r_{t,i,j} \boldsymbol{x}_i}{\sum_{i=1}^{n} r_{t,i,j}}
\end{aligned}$$

(c) We can rewrite $q(\boldsymbol{z}_i, c_i; \boldsymbol{\theta}_t)$ as follows.

$$\begin{aligned}
q(\boldsymbol{z}_i, c_i; \boldsymbol{\theta}_t) &= p(\boldsymbol{z}_i | c_i, \boldsymbol{x}_i; \boldsymbol{\theta}_t) p(c_i | \boldsymbol{x}_i; \boldsymbol{\theta}_t) \\
&= r_{t,i,c_i} \mathcal{N}(\boldsymbol{z}_i | \boldsymbol{V}_{t,c_i}^{-1} \boldsymbol{W}_{t,c_i}^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{t,c_i}), \sigma_{t,c_i}^2 \boldsymbol{V}_{t,c_i}^{-1})
\end{aligned}$$

Then, by using the results of Problem 1-(c), we have

$$\begin{aligned}
\tilde{\mathcal{L}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \mathbb{E}_q [\log p(\boldsymbol{x}_i, \boldsymbol{z}_i, c_i; \boldsymbol{\theta})] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} r_{t,i,j} \mathbb{E}_{p(\boldsymbol{z}_i | c_i = j, \boldsymbol{x}_i; \boldsymbol{\theta}_t)} [\log p(\boldsymbol{x}_i, \boldsymbol{z}_i, c_i = j; \boldsymbol{\theta})] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} r_{t,i,j} \mathbb{E}_{p(\boldsymbol{z}_i | c_i = j, \boldsymbol{x}_i; \boldsymbol{\theta}_t)} [\log p(\boldsymbol{x}_i, \boldsymbol{z}_i | c_i = j; \boldsymbol{\theta}) + \log p(c_i = j; \boldsymbol{\theta})] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} r_{t,i,j} \left(\log \pi_j + \mathbb{E}_{p(\boldsymbol{z}_i | c_i = j, \boldsymbol{x}_i; \boldsymbol{\theta}_t)} [\log p(\boldsymbol{x}_i, \boldsymbol{z}_i | c_i = j; \boldsymbol{\theta})] \right)
\end{aligned}$$

And we already know that

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{z}_i | c_i = j, \boldsymbol{x}_i; \boldsymbol{\theta}_t)} [\log p(\boldsymbol{x}_i, \boldsymbol{z}_i | c_i = j; \boldsymbol{\theta})] \\
&= -\frac{d+h}{2} \log 2\pi - d \log \sigma_j - \frac{1}{2\sigma_j^2} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_j) - \frac{\sigma_{t,j}^2}{2\sigma_j^2} \operatorname{Tr} \left(\boldsymbol{V}_j \boldsymbol{V}_{t,j}^{-1} \right) \\
&\quad - \frac{1}{2\sigma_j^2} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{W}_{t,j} \boldsymbol{V}_{t,j}^{-1} \boldsymbol{V}_j \boldsymbol{V}_{t,j}^{-1} \boldsymbol{W}_{t,j}^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_j) \\
&\quad + \frac{1}{2\sigma_j^2} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{W}_{t,j} \boldsymbol{V}_{t,j}^{-1} \boldsymbol{W}_j^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_j)
\end{aligned}$$

where $\boldsymbol{W}_j, \boldsymbol{W}_{t,j}, \boldsymbol{V}_j, \boldsymbol{V}_{t,j}$ are defined for $c_i = j$ in the same manner as Problem 1. Let's define T_j and n_j as follows to write $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ concisely.

$$\begin{aligned}
n_j &= \sum_{i=1}^{n} r_{t,i,j} \\
T_j &= \frac{1}{n_j} \sum_{i=1}^{n} r_{t,i,j} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top
\end{aligned}$$

Then we can rewrite $\tilde{\mathcal{L}}(\boldsymbol{\theta})$

$$
\begin{aligned}
\tilde{\mathcal{L}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{j=1}^{k} r_{t,i,j}\left(\log \pi_j + \mathbb{E}_{p(\boldsymbol{z}_i|c_i=j,\boldsymbol{x}_i;\boldsymbol{\theta}_t)}[\log p(\boldsymbol{x}_i, \boldsymbol{z}_i|c_i=j;\boldsymbol{\theta})]\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} r_{t,i,j}\left(\log \pi_j - \frac{d+h}{2}\log 2\pi - d\log\sigma_j\right) \\
&\qquad - \frac{n_j}{2\sigma_j^2}\sum_{j=1}^{k}\Bigg(\operatorname{Tr}(\boldsymbol{T}_j) + \sigma_{t,j}^2 \operatorname{Tr}\left(\boldsymbol{V}_j\boldsymbol{V}_{t,j}^{-1}\right) + \operatorname{Tr}\left(\boldsymbol{W}_{t,j}\boldsymbol{V}_{t,j}^{-1}\boldsymbol{V}_j\boldsymbol{V}_{t,j}^{-1}\boldsymbol{W}_{t,j}^{\top}\boldsymbol{T}_j\right) \\
&\qquad\qquad - 2\operatorname{Tr}\left(\boldsymbol{W}_{t,j}\boldsymbol{V}_{t,j}^{-1}\boldsymbol{W}_j^{\top}\boldsymbol{T}_j\right)\Bigg) \\
&= -\sum_{j=1}^{k} n_j d\log\sigma_j \\
&\qquad - \frac{n_j}{2\sigma_j^2}\sum_{j=1}^{k}\Bigg(\operatorname{Tr}(\boldsymbol{T}_j) + \sigma_t^2 \operatorname{Tr}\left(\boldsymbol{V}_j\boldsymbol{V}_{t,j}^{-1}\right) + \operatorname{Tr}\left(\boldsymbol{W}_{t,j}\boldsymbol{V}_{t,j}^{-1}\boldsymbol{V}_j\boldsymbol{V}_{t,j}^{-1}\boldsymbol{W}_{t,j}^{\top}\boldsymbol{T}_j\right) \\
&\qquad\qquad - 2\operatorname{Tr}\left(\boldsymbol{W}_{t,j}\boldsymbol{V}_{t,j}^{-1}\boldsymbol{W}_j^{\top}\boldsymbol{T}_j\right)\Bigg) + const.
\end{aligned}
$$

which is almost same to the one in Problem 1.
Thus, by taking the gradient as in Problem 1, the M-step equations are

$$
\begin{aligned}
\boldsymbol{W}_{t+1,j} &= \boldsymbol{T}_j\boldsymbol{W}_{t,j}\left(\sigma_{t,j}^2\boldsymbol{I}_h + \boldsymbol{V}_{t,j}^{-1}\boldsymbol{W}_{t,j}^{\top}\boldsymbol{T}_j\boldsymbol{W}_{t,j}\right)^{-1} \\
\sigma_{t+1,j}^2 &= \frac{1}{d}\operatorname{Tr}\left(\boldsymbol{T}_j - \boldsymbol{T}_j\boldsymbol{W}_{t,j}\boldsymbol{V}_{t,j}^{-1}\boldsymbol{W}_{t+1,j}^{\top}\right)
\end{aligned}
$$