# Chapter 12

stevenjin8

February 28, 2021

## Exercises

### Exercise 2

For simplicity, we assume that $\tilde{\mathbf{W}}_1, \ldots, \tilde{\mathbf{W}}_C$ are independently distrubted and that $\boldsymbol{\Psi}$ is known. First, we find the expected log likelihood:

$$\begin{aligned}\mathbb{E}\left[\log p(\mathbf{x}_i|\boldsymbol{\theta})|\mathbf{x}_i, \boldsymbol{\theta}^{\text{old}}\right] &= \sum_{c=1}^{C} r_{ic}\mathbb{E}\left[\log p(\mathbf{x}_i|\boldsymbol{\theta}, q_i = c)|\mathbf{x}_i, \boldsymbol{\theta}^{\text{old}}, q_i = c\right] \\ &= \sum_{c=1}^{C} r_{ic}\mathbb{E}\left[\log(p(\mathbf{x}_i|\mathbf{z}_i, q_i = c, \boldsymbol{\theta})p(\mathbf{z}_i, \boldsymbol{\theta})\pi_c)|\mathbf{x}_i, \boldsymbol{\theta}^{\text{old}}, q_i = c\right].\end{aligned}$$

We give a Gaussian prior to $\mathbf{W}_c$. The matrix normal (denoted by $\mathcal{MN}$) is given by

$$\mathcal{MN}(\mathbf{W}_c|\mathbf{U}, \mathbf{V}, \mathbf{M}) \propto \exp\left(-\frac{1}{2}\mathbf{V}^{-1}(\mathbf{W}_c - \mathbf{M})^T\mathbf{U}^{-1}(\mathbf{W}_c - \mathbf{M})\right)$$

and

$$\log \mathcal{MN}(\mathbf{W}_c|\mathbf{U}, \mathbf{V}, \mathbf{M}) = -\frac{1}{2}\mathbf{V}^{-1}(\mathbf{W}_c - \mathbf{M})^T\mathbf{U}^{-1}(\mathbf{W}_c - \mathbf{M}) + \text{constant}.$$

Next we take some derivatives (we omit the conditional for brevity):

$$\begin{aligned}\frac{\partial}{\partial \mathbf{W}_k}\mathbb{E}\left[\log p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})\right] &= r_{ik}\mathbb{E}\left[-\frac{1}{2}\frac{\partial}{\partial \tilde{\mathbf{W}}_k}(\mathbf{x}_i - \tilde{\mathbf{W}}_k\tilde{\mathbf{z}}_i)^T\boldsymbol{\Psi}^{-1}(\mathbf{x}_i - \tilde{\mathbf{W}}_k\tilde{\mathbf{z}}_i)\right] \\ &= r_{ik}\mathbb{E}\left[-\frac{1}{2}\frac{\partial}{\partial \tilde{\mathbf{W}}_k}\left(-2\mathbf{x}_i^T\boldsymbol{\Psi}^{-1}\tilde{\mathbf{W}}_k\tilde{\mathbf{z}}_i + \tilde{\mathbf{z}}_i^T\tilde{\mathbf{W}}_k\boldsymbol{\Psi}^{-1}\tilde{\mathbf{W}}_k\tilde{\mathbf{z}}_i\right)\right] \\ &= r_{ik}\mathbb{E}\left[\boldsymbol{\Psi}^{-1}\mathbf{x}_i\tilde{\mathbf{z}}_i^T - \boldsymbol{\Psi}^{-1}\tilde{\mathbf{W}}_k\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T\right] \\ &= r_{ik}\boldsymbol{\Psi}^{-1}\mathbf{x}_i\mathbb{E}\left[\tilde{\mathbf{z}}_i\right]^T - r_{ik}\boldsymbol{\Psi}^{-1}\tilde{\mathbf{W}}_k\mathbb{E}\left[\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T\right] \\ &= r_{ik}\boldsymbol{\Psi}^{-1}\mathbf{x}_i\mathbf{b}_{ik}^T - r_{ik}\boldsymbol{\Psi}^{-1}\tilde{\mathbf{W}}_k\mathbf{C}_{ik}.\end{aligned}$$

Now the prior:

$$\begin{aligned}
&\partial \log \mathcal{MN}\left(\tilde{\mathbf{W}}_k|\mathbf{U},\mathbf{V},\mathbf{M}\right) = \operatorname{tr}\left(\partial\left(\mathbf{V}^{-1}(\mathbf{M}-\tilde{\mathbf{W}}_k)^T\mathbf{U}^{-1}(\mathbf{M}-\tilde{\mathbf{W}}_k)\right)\right)\\
&= \operatorname{tr}\left(\mathbf{V}^{-1}(\partial\tilde{\mathbf{W}}_k-\mathbf{M})^T\mathbf{U}^{-1}(\tilde{\mathbf{W}}_k-\mathbf{M}) + \mathbf{V}^{-1}(\partial\tilde{\mathbf{W}}_k-\mathbf{M})^T\mathbf{U}^{-1}(\tilde{\mathbf{W}}_k-\mathbf{M})\right)\\
&= \operatorname{tr}\left(\mathbf{V}^{-1}\partial\tilde{\mathbf{W}}_k^T\mathbf{U}^{-1}\tilde{\mathbf{W}}_k + \mathbf{V}^{-1}\tilde{\mathbf{W}}_k^T\mathbf{U}^{-1}\partial\tilde{\mathbf{W}}_k\right)\\
&= \operatorname{tr}\left(\mathbf{V}^{-T}\tilde{\mathbf{W}}_k^T\mathbf{U}^{-T}\partial\tilde{\mathbf{W}}_k + \mathbf{V}^{-1}\tilde{\mathbf{W}}_k^T\mathbf{U}^{-1}\partial\tilde{\mathbf{W}}_k\right)\\
&= \operatorname{tr}\left(\left(\mathbf{V}^{-T}+\mathbf{V}^{-1}\right)\tilde{\mathbf{W}}_k^T\left(\mathbf{U}^{-T}+\mathbf{U}^{-1}\right)\partial\tilde{\mathbf{W}}_k\right).
\end{aligned}$$

Thus,

$$\frac{\partial}{\partial \mathbf{W}_k}\log\mathcal{MN}\left(\tilde{\mathbf{W}}_k|\mathbf{U},\mathbf{V},\mathbf{M}\right) = \left(\mathbf{V}^{-T}+\mathbf{V}^{-1}\right)\tilde{\mathbf{W}}_k\left(\mathbf{U}^{-T}+\mathbf{U}^{-1}\right) = \boldsymbol{\Lambda}_k$$

If you ask me, that is an aesthetic equation.

It follows that the MAP estimate of $\mathbf{W}_k$ is

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \mathbf{W}_k}\mathbb{E}\left[\log p(\mathcal{D},\boldsymbol{\theta})|\mathbf{x}_i,\boldsymbol{\theta}^{\text{old}}\right]\\
\mathbf{0} &= \boldsymbol{\Psi}^{-1}\sum_{i=1}^{N} r_{ik}\mathbf{x}_i\mathbf{b}_{ik}^T - \boldsymbol{\Psi}^{-1}\tilde{\mathbf{W}}_k\sum_{i=1}^{N} r_{ik}\mathbf{C}_{ik} + \boldsymbol{\Lambda}\\
\boldsymbol{\Psi}^{-1}\tilde{\mathbf{W}}_k\sum_{i=1}^{N} r_{ik}\mathbf{C}_{ik} &= \boldsymbol{\Psi}^{-1}\sum_{i=1}^{N} r_{ik}\mathbf{x}_i\mathbf{b}_{ik}^T + \boldsymbol{\Lambda}\\
\hat{\tilde{\mathbf{W}}}_k^{MAP} &= \left(\sum_{i=1}^{N} r_{ik}\mathbf{x}_i\mathbf{b}_{ik}^T + \boldsymbol{\Psi}\boldsymbol{\Lambda}\right)\left(\sum_{i=1}^{N} r_{ik}\mathbf{C}_{ik}\right)^{-1}.
\end{aligned}$$

Setting $\boldsymbol{\Lambda} = \mathbf{0}$ yields the solution to exercise 1. I think it is pretty cool that for exercise 1, $\boldsymbol{\Psi}$ is absent in the MLE estimate of $\hat{\tilde{\mathbf{W}}}_k$, but it is present in the MAP estimate. This makes sense as our certainty in our data should affect the influence of the prior on the estimate.

Finding the MAP estimate of all the parameters is much harder because one would need to find a conjugate prior:

$$p(\mathbf{W}_1,\ldots,\mathbf{W}_C,\boldsymbol{\Psi},\boldsymbol{\pi}).$$

## Exercise 4

a.

$$\begin{aligned}
\frac{\partial J}{\partial z_{j2}} &= 0\\
-2\mathbf{v}_2^T(\mathbf{x}_j - z_{j1}\mathbf{v}_1 - z_{j2}\mathbf{v}_2) &= 0\\
\mathbf{v}_2^T\mathbf{x}_j - z_{j1}\mathbf{v}_2^T\mathbf{v}_1 - z_{j2}\mathbf{v}_2^T\mathbf{v}_2 = 0.
\end{aligned}$$

Since $\mathbf{v}_1, \ldots, \mathbf{v}_K$ are orthonormal,

$$\mathbf{v}_2^T \mathbf{x}_j = x_{j2}.$$

b. We want to minimize $\tilde{J}$ with respect to $\mathbf{v}_2$ over $||\mathbf{v}_2|| = 1$.

$$\frac{\partial \tilde{J}}{\partial \mathbf{v}_2} = -2\mathbf{C}\mathbf{v}_2 + 2\lambda_2 \mathbf{v}_2.$$

This follows from the fact that $\mathbf{v}_1, \ldots, \mathbf{v}_K$ are orthonormal. At the same time,

$$\frac{\partial \|\mathbf{v}_2\|^2}{\partial \mathbf{v}_2} = 2\mathbf{v}_2.$$

The critical values are given by

$$\begin{aligned} -2\mathbf{C}\mathbf{v}_2 + 2\lambda_2 \mathbf{v}_2 &= a2\mathbf{v}_2 \\ \mathbf{C}\mathbf{v}_2 &= (a - \lambda_2)\mathbf{v}_2. \end{aligned}$$

It follows that $\mathbf{v}_2$ is an eigenvector of $\mathbf{C}$. To minimize $\tilde{J}$, we want maximize $\mathbf{v}_2^T \mathbf{C}\mathbf{v}_2$. Thus, we want $\mathbf{v}_2$ to have the eigenvector with the biggest possible eigenvalue. Since $\mathbf{v}_1$ already is the eigenvector with the biggest eigenvalue, $\mathbf{v}_2$ has to settle for the second largest eigenvalue.

## Exercise 5

a.

$$\begin{aligned} \|\mathbf{x}_i - \sum z_{ik}\mathbf{v}_k\|^2 &= (\mathbf{x}_i - \sum z_{ik}\mathbf{v}_k)^T(\mathbf{x}_i - \sum z_{ik}\mathbf{v}_k) \\ &= \mathbf{x}_i^T\mathbf{x}_i - 2\sum z_{ik}\mathbf{x}_i^T\mathbf{v}_k + \left(\sum z_{ik}\mathbf{v}_k\right)^T\left(\sum z_{ik}\mathbf{v}_k\right) \\ &= \mathbf{x}_i^T\mathbf{x}_i - 2\sum z_{ik}\mathbf{x}_i^T\mathbf{v}_k + \sum z_{ik}\mathbf{x}_i^T\mathbf{v}_k \\ &= \mathbf{x}_i^T\mathbf{x}_i - \sum \mathbf{v}_k^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_k. \end{aligned}$$

b.

$$\begin{aligned} J_K &= \frac{1}{n}\sum_i^n \left(\mathbf{x}_i^T\mathbf{x}_i - \sum_k^K \mathbf{v}_k^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_k\right) \\ &= \frac{1}{n}\sum_i^n \left(\mathbf{x}_i^T\mathbf{x}_i\right) - \frac{1}{n}\sum_k^K \mathbf{v}_k^T \sum_i^n \left(\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{v}_k \\ &= \sum_i^n \mathbf{x}_i^T\mathbf{x}_i - \sum_k^K \mathbf{v}_k^T\mathbf{C}\mathbf{v}_k \\ &= \sum_i^n \mathbf{x}_i^T\mathbf{x}_i - \sum_k^K \lambda_k. \end{aligned}$$

c. Since

$$\begin{aligned}
J_d &= \sum_{i}^{n} \mathbf{x}_i^T \mathbf{x}_i - \sum_{k=1}^{d} \lambda_k \mathbf{x}_i^T \mathbf{x}_i \\
&= \sum_{i}^{n} \mathbf{x}_i^T \mathbf{x}_i - \sum_{k=1}^{K} \lambda_k \mathbf{x}_i^T \mathbf{x}_i - \sum_{k=K+1}^{d} \lambda_k \mathbf{x}_i^T \mathbf{x}_i \\
&= J_K - \sum_{k=K+1}^{d} \lambda_k \mathbf{x}_i^T \mathbf{x}_i \\
&= 0,
\end{aligned}$$

we have

$$J_K = \sum_{k=K+1}^{d} \lambda_k.$$

## Exercise 6

There is a mistake in equation 12.133. $\tilde{\mathbf{C}}$ is a $d \times d$ matrix, so the equation should read

$$\tilde{\mathbf{C}} = \frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \frac{1}{n}\mathbf{X}\mathbf{X}^T - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T.$$

a.

$$\begin{aligned}
\tilde{\mathbf{C}} &= \frac{1}{n}\left(\mathbf{I} - \mathbf{v}_1\mathbf{v}_1^T\right)\mathbf{X}\mathbf{X}^T\left(\mathbf{I} - \mathbf{v}_1\mathbf{v}_1^T\right) \\
&= \left(\mathbf{C} - 2\mathbf{v}_1\mathbf{C}\mathbf{v}_1^T + \mathbf{v}_1\mathbf{v}_1^T\mathbf{C}\mathbf{v}_1\mathbf{v}_1^T\right) \\
&= \left(\mathbf{C} - 2\lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_1\mathbf{v}_1\mathbf{v}_1^T\mathbf{v}_1\mathbf{v}_1^T\right) \\
&= \left(\mathbf{C} - \lambda_1\mathbf{v}_1\mathbf{v}_1^T\right).
\end{aligned}$$

b.

Let $\mathbf{u}$ be any eigenvector of $\mathbf{C}$. Assuming that the largest eigenvalue of $\mathbf{C}$ has a multiplicity of 1, we have

$$\begin{aligned}
\tilde{\mathbf{C}}\mathbf{u} &= \mathbf{C}\mathbf{u} - \lambda\mathbf{v}_1\mathbf{v}_1^T\mathbf{u} \\
&= \lambda_i\mathbf{u} - \lambda_1\mathbf{v}_1 0.
\end{aligned}$$

when $i = 1$ (i.e. $\mathbf{u} = \mathbf{v}_1$). Thus, $\mathbf{C}$ and $\tilde{\mathbf{C}}$ have the same eigenvectors with the same eigenvalues, except for $\mathbf{v}_1$. It follows that the eigenvector with the largest eigenvalue of $\tilde{\mathbf{C}}$ is colinear with the eigenvector with the second largest eigenvalue of $\mathbf{C}$ : $\mathbf{v}_2$. Since both $\mathbf{u}$ and $\mathbf{v}_2$ are unit norm, $\mathbf{u} = \pm\mathbf{v}_2$.
c.

```
eigenvalues = []
eigenvectors = []
for _ in range(K):
    lmbd, u = f(C)
    eigenvalues.append(lmbd)
    eigenvectors.append(u)
    C = C - lmbd * u @ u.T
```

where @ denotes matrix multiplication.