Lecture Notes 6: Probabilistic PCA and EM algorithm

Professor: Zhihua Zhang *Scribe:Xin Wang, Cheng Chen*

6 Probabilistic PCA

6.1 Procrustes Transformation

**Procrustes was an African bandit in Greek mythology, who stretched or squashed his visitors to fit his iron bed(eventually killing them).*

Orthogonal Procrustes Problem is a matrix approximation problem (here we regard it as an optimization problem), where given two matrices $\mathbf{X}$ and $\mathbf{Y}$, one is asked to find an **Orthogonal matrix** $\mathbf{U}$ which most closely maps $\mathbf{X}$ to $\mathbf{Y}$. Intuitively, $\mathbf{U}$ is to rotate $\mathbf{X}$ by a certain angle.

Proposition 6.1. *Let $\mathbf{X}$ and $\mathbf{Y}$ be $n \times p$ matrices, and $\mathbf{U}$ is a $p \times p$ orthogonal matrix which minimizes $||\mathbf{Y} - \mathbf{X}\mathbf{U}^T||_F^2$, s.t. $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$, where $||\cdot||$ denotes the Frobenius norm. Then, $\mathbf{U} = \mathbf{Q}\mathbf{R}^T$, where $\mathbf{R}$ and $\mathbf{Q}$ are orthogonal matrices.*

Proof.

$$\begin{aligned}
||\mathbf{Y} - \mathbf{X}\mathbf{U}^T||_F^2 &= \mathrm{tr}((\mathbf{Y} - \mathbf{X}\mathbf{U}^T)^T(\mathbf{Y} - \mathbf{X}\mathbf{U}^T)) \\
&= \mathrm{tr}(\mathbf{Y}^T\mathbf{Y}) + \mathrm{tr}(\mathbf{U}\mathbf{X}^T\mathbf{X}\mathbf{U}^T) - 2\mathrm{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{U}^T) \\
&= \mathrm{tr}(\mathbf{Y}^T\mathbf{Y}) + \mathrm{tr}(\mathbf{X}^T\mathbf{X}) - 2\mathrm{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{U}^T)
\end{aligned}$$

Note: $\mathrm{tr}(\mathbf{Y}^T\mathbf{Y})$ and $\mathrm{tr}(\mathbf{X}^T\mathbf{X})$ are constants, so in order to minimize $||\mathbf{Y} - \mathbf{X}\mathbf{U}^T||_F^2$, we need the maximum of $\mathrm{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{U}^T)$. We first perform *Single Value Decomposition*(SVD) on $\mathbf{Y}^T\mathbf{X}$,

$$\mathbf{Y}^T\mathbf{X} = \mathbf{Q}\mathbf{\Lambda}\mathbf{R}^T$$

where $\mathbf{Q}$ and $\mathbf{R}$ are orthogonal matrices.

$$\begin{aligned}
\mathrm{tr}(\mathbf{Q}\mathbf{\Lambda}\mathbf{R}^T\mathbf{U}^T) &= \mathrm{tr}(\mathbf{\Lambda}\mathbf{R}^T\mathbf{U}^T\mathbf{Q}) \\
&= \mathrm{tr}(\mathbf{\Lambda}\mathbf{Z}) \ (Let\, \mathbf{R}^T\mathbf{U}^T\mathbf{Q} = \mathbf{Z}) \\
&= \sum_{i=1}^{p} \boldsymbol{\lambda}_i \mathbf{Z}_{ii} \ (\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)) \\
&\leq \sum_{i=1}^{p} \boldsymbol{\lambda}_i \ (Since\, for\, orthonomal\, matrix,\, all\, the\, elements\, are\ \leq 1)
\end{aligned}$$

The equity is obtained when $\mathbf{Z} = \mathbf{I}$, that is,

$$\mathbf{Z} = \mathbf{I} = \mathbf{R}^T\mathbf{U}^T\mathbf{Q} \Longrightarrow \mathbf{U} = \mathbf{Q}\mathbf{R}^T$$

□

6.2 Probabilistic PCA(closed form approach)

Definition 6.1. *Probabilistic PCA is often formulated as a mapping from the latent space into the data space via*

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon}$$

where $\mathbf{x} \in \mathbb{R}^p$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_p)$, $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_q)$ *and* $\mathbf{z} \perp \boldsymbol{\epsilon}$. $\mathbf{z} \in \mathbb{R}^q$ *is a latent variable and* $\mathbf{W}$ *is a* $p \times q$ *loading matrix.*

Given a set of samples, $\{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathbb{R}^p$, we have

- Sample mean: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum\limits_{i=1}^{n} \mathbf{x}_i$
- Sample variance: $\mathbf{S} = \frac{1}{n}\sum\limits_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$

Based on the previous notes, we have the following results. Please refer to *Lecture Notes 5* for detailed proof if necessary.

1. $(\mathbf{x} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \mathbf{C})$, where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}_p$
2. $(\mathbf{z}|\mathbf{x} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$, where $\mathbf{M} = \sigma^2\mathbf{I}_q + \mathbf{W}^T\mathbf{W}$
3. Likelihood function $L = \frac{1}{|\mathbf{C}|^{\frac{n}{2}}}\exp\left(-\frac{1}{2}\sum\limits_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right)$ (Ignore the constant coefficient), and we let $f = -\log L$
4. When $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\mu}} = 0$, we have $f = \frac{n}{2}\log|C| + \frac{n}{2}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})$

Let function $\mathbf{F}$ be $\mathbf{F}(\mathbf{W}, \sigma^2) = \log|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})$. For convenience, just let $\sigma^2 = \tau$. In the following part, we are going to minimize the value of $\mathbf{F}$ and estimate $\mathbf{W}$ and τ. That is,

$$\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{W}} = 0 \quad \text{and} \quad \frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\tau} = 0$$

Before we get into detailed proof, the following techniques should be noted.

1. $\mathrm{d}\log|\mathbf{C}| = \mathrm{tr}(\mathbf{C}^{-1}\mathrm{d}\mathbf{C})$
2. $\mathrm{d}\,\mathrm{tr}(\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{x}) \implies \frac{\mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{x}^T)}{\mathrm{d}\mathbf{x}} = \mathbf{A}$
3. $\mathrm{d}\mathbf{C}^{-1} = -\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1}$

Proof. 1. We assume that $\mathbf{B}$ is the adjoint matrix of $\mathbf{C}$, and we have $\mathbf{B}\mathbf{C} = \mathbf{C}\mathbf{B} = |\mathbf{C}|\mathbf{I}$. Then we can get $|\mathbf{C}| = \sum\limits_{j=1}^{p}\mathbf{C}_{ij}(\mathbf{B}^T)_{ij}$, which indicates that $\frac{\partial|\mathbf{C}|}{\partial\mathbf{C}_{ij}} = (\mathbf{B}^T)_{ij}$.

Therefore, we have

$$\begin{aligned}
\mathrm{d}|\mathbf{C}| &= \sum_i \sum_j (\mathbf{B}^T)_{ij} (\mathrm{d}\mathbf{C})_{ij} \\
&= \mathrm{tr}(\mathbf{B}\,\mathrm{d}\mathbf{C}) \\
&= \mathrm{tr}(|\mathbf{C}|\mathbf{C}^{-1}\mathrm{d}\mathbf{C}) \\
&= |\mathbf{C}|\mathrm{tr}(\mathbf{C}^{-1}\mathrm{d}\mathbf{C}) \\
\mathrm{d}\log|\mathbf{C}| &= \frac{\mathrm{d}|\mathbf{C}|}{|\mathbf{C}|} = \mathrm{tr}(\mathbf{C}^{-1}\mathrm{d}\mathbf{C})
\end{aligned}$$

2. The proof was left as homework in last class.

3. $$\mathbf{C}\cdot \mathbf{C}^{-1} = \mathbf{I} \Rightarrow \mathrm{d}\mathbf{C}\cdot \mathbf{C}^{-1} + \mathbf{C}\cdot \mathrm{d}\mathbf{C}^{-1} = 0 \Rightarrow \mathrm{d}\mathbf{C}^{-1} = -\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1}$$

□

Now we set out to calculate $\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{W}}$ and $\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\tau}$.

$\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{W}}$:

$$\begin{aligned}
\mathrm{d}\mathbf{F} &= \mathrm{d}\log|\mathbf{C}| + \mathrm{d}\,\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S}) \\
&= \mathrm{tr}(\mathbf{C}^{-1}\mathrm{d}\mathbf{C}) + \mathrm{tr}(d\mathbf{C}^{-1}\mathbf{S}) \\
&= \mathrm{tr}(\mathbf{C}^{-1}\mathrm{d}\mathbf{C}) - \mathrm{tr}(\mathbf{C}^{-1}\mathrm{d}\mathbf{C}\mathbf{C}^{-1}\mathbf{S}) \\
&= \mathrm{tr}(\mathbf{C}^{-1}(\mathrm{d}\mathbf{W}\mathbf{W}^T + \mathbf{W}\mathrm{d}\mathbf{W}^T)) - \mathrm{tr}(\mathbf{C}^{-1}(\mathrm{d}\mathbf{W}\mathbf{W}^T + \mathbf{W}\mathrm{d}\mathbf{W}^T)\mathbf{C}^{-1}\mathbf{S})
\end{aligned}$$

Since we have

$$\mathrm{tr}(\mathbf{C}^{-1}\mathrm{d}\mathbf{W}\mathbf{W}^T) = \mathrm{tr}(\mathbf{W}\mathrm{d}\mathbf{W}^T\mathbf{C}^{-1}) = \mathrm{tr}(\mathbf{C}^{-1}\mathbf{W}\mathrm{d}\mathbf{W}^T)$$

Then we get

$$\begin{aligned}
\mathrm{d}\mathbf{F} &= 2\mathrm{tr}(\mathbf{C}^{-1}\mathbf{W}\mathrm{d}\mathbf{W}^T) - 2\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W}\mathrm{d}\mathbf{W}^T) \\
&= 2\mathrm{tr}[(\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W})\mathrm{d}\mathbf{W}^T]
\end{aligned}$$

Therefore,

$$\frac{1}{2}\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{W}} = \mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W}$$

If we let $\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\mathbf{W}} = 0$, we can achieve that $\mathbf{W} = \mathbf{S}\mathbf{C}^{-1}\mathbf{W}$.
Substitute $\mathbf{C} = \tau\mathbf{I}_p + \mathbf{W}\mathbf{W}^T$ to the equation above. We can prove the following two equations are equivalent. The proof is omitted here.

$$\begin{aligned}
\mathbf{S}(\tau\mathbf{I}_p + \mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W} &= \mathbf{W} \\
\mathbf{S}\mathbf{W}(\tau\mathbf{I}_q + \mathbf{W}^T\mathbf{W})^{-1} &= \mathbf{W}
\end{aligned}$$

We perform Eigendecomposition on $\mathbf{W}^T\mathbf{W}$, which is $\mathbf{W}^T\mathbf{W} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, we get

$$\begin{aligned}
\mathbf{S}\mathbf{W}\mathbf{V}(\tau\mathbf{I}_q + \mathbf{\Lambda})^{-1} &= \mathbf{W}\mathbf{V} \\
\mathbf{S}\mathbf{W}\mathbf{V} &= \mathbf{W}\mathbf{V}(\tau\mathbf{I}_q + \mathbf{\Lambda})
\end{aligned}$$

where the diagonal of $\tau\mathbf{I}_q+\mathbf{\Lambda}$ is composed of eigenvalues of $\mathbf{S}$ and $\mathbf{WV}$ is composed of eigenvectors. However, $\mathbf{WV}$ may not be orthonormal. Hence, we need to normalize $\mathbf{WV}$:

$$\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^T\mathbf{W}^T\mathbf{W}\mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{I}$$

Therefore, we have

$$\mathbf{SWV}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{WV}\mathbf{\Lambda}^{-\frac{1}{2}}(\tau\mathbf{I}_q + \mathbf{\Lambda})$$

Let $\Phi_q = \mathbf{WV}\mathbf{\Lambda}^{-\frac{1}{2}}, \mathbf{\Gamma}_q = \tau\mathbf{I}_q + \mathbf{\Lambda}$, We can know that $\Phi_q^T\Phi_q = \mathbf{I}_q$ and the diagonal of $\mathbf{\Gamma}_q$ is composed of eigenvalues of $\mathbf{S}$. We can also substitute $\mathbf{\Gamma}_q$ into $\mathbf{F}$, and get the minimum value of $\mathbf{F}$ when the diagonal of $\mathbf{\Gamma}_q$ is composed of top q largest eigenvalues of $\mathbf{S}$.
Together we have

$$\begin{aligned}
\mathbf{S}\Phi_q &= \Phi_q\mathbf{\Gamma}_q \\
\mathbf{W} &= \Phi_q(\mathbf{\Gamma}_q - \tau\mathbf{I}_q)^{\frac{1}{2}}\mathbf{V}^T
\end{aligned}$$

Remark 6.1. *We can assign $\mathbf{V} = \mathbf{I}_q$ without loss of generality, since $\mathbf{V}$ can be arbitrary orthogonal rotation matrix. Moreover, Φ_q is a $p \times q$ matrix with q column vectors as the principal eigenvectors of $\mathbf{S}$. Now, we have known something about $\mathbf{W}$ and should continue to estimate τ.*

$\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\tau}$:

$$\begin{aligned}
d\mathbf{F} &= \mathrm{tr}(\mathbf{C}^{-1}d\tau\mathbf{I}) - \mathrm{tr}(\mathbf{C}^{-1}d\tau\mathbf{C}^{-1}\mathbf{S}) \quad (d\tau \text{ is a scalar}) \\
&= [\mathrm{tr}(\mathbf{C}^{-1}) - \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1})]d\tau
\end{aligned}$$

which implies

$$\frac{d\mathbf{F}}{d\tau} = \mathrm{tr}(\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}) = 0$$

Recall another condition

$$\mathbf{C}^{-1} = (\tau\mathbf{I}_p + \mathbf{W}\mathbf{W}^T)^{-1} = \tau^{-1}\mathbf{I}_p - \tau^{-1}\mathbf{W}(\tau\mathbf{I}_q + \mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T$$

Multiply $\mathbf{S}$ on both sides, we have

$$\begin{aligned}
\mathbf{S}\mathbf{C}^{-1} &= \tau^{-1}\mathbf{S} - \tau^{-1}\mathbf{SW}(\tau\mathbf{I}_q + \mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T \\
&= \tau^{-1}\mathbf{S} - \tau^{-1}\mathbf{W}\mathbf{W}^T \quad (\text{Since } \mathbf{SW}(\tau\mathbf{I}_q + \mathbf{W}^T\mathbf{W})^{-1} = \mathbf{W})
\end{aligned}$$

$$\begin{aligned}
\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1} - \mathbf{C}^{-1} &= \mathbf{C}^{-1}\tau^{-1}\mathbf{S} - \mathbf{C}^{-1}\tau^{-1}\mathbf{W}\mathbf{W}^T - \mathbf{C}^{-1} \\
&= \tau^{-1}\mathbf{C}^{-1}\mathbf{S} - \mathbf{C}^{-1}\tau^{-1}\mathbf{W}\mathbf{W}^T - \mathbf{C}^{-1} \\
&= \tau^{-1}(\tau^{-1}\mathbf{S} - \tau^{-1}\mathbf{W}\mathbf{W}^T) - \mathbf{C}^{-1}\tau^{-1}\mathbf{W}\mathbf{W}^T - \mathbf{C}^{-1} \\
&= \tau^{-1}(\tau^{-1}\mathbf{S} - \tau^{-1}\mathbf{W}\mathbf{W}^T) - \tau^{-1}\mathbf{C}^{-1}(\mathbf{C} - \tau\mathbf{I}_p) - \mathbf{C}^{-1} \\
&= \tau^{-1}(\tau^{-1}\mathbf{S} - \tau^{-1}\mathbf{W}\mathbf{W}^T) - \tau^{-1}\mathbf{I}_p \\
&= \tau^{-2}\mathbf{S} - \tau^{-2}\mathbf{W}\mathbf{W}^T - \tau^{-1}\mathbf{I}_p
\end{aligned}$$

$$\frac{d\mathbf{F}}{d\tau} = \mathrm{tr}(\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}) = \tau^{-2}\mathrm{tr}(\mathbf{S} - \mathbf{W}\mathbf{W}^T - \tau\mathbf{I}_p) = 0$$

Then we have

$$\begin{aligned}
\tau\mathrm{tr}(\mathbf{I}_p) &= \mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{W}^T\mathbf{W}) \\
p\tau &= \mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{\Gamma}_q - \tau\mathbf{I}_q) \\
&= \mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{\Gamma}_q) + q\tau \\
(p-q)\tau &= \mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{\Gamma}_q) \\
\tau &= \frac{\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{\Gamma}_q)}{p-q} \\
&= \frac{1}{p-q}\sum_{j=q+1}^{p} \mathbf{\Gamma}_j
\end{aligned}$$

The condition above is referred as the first order condition, which has a clear interpretation as the variance "lost" in the projection, averaged over the lost dimensions.

Summary:

$$\tau = \frac{1}{p-q}\sum_{j=q+1}^{p} \mathbf{\Gamma}_j \tag{1}$$

$$\mathbf{W} = \Phi_q(\mathbf{\Gamma}_q - \tau\mathbf{I}_q)^{\frac{1}{2}}\mathbf{V}^T \tag{2}$$

In practice, to find the most likely model given $\mathbf{S}$, we should first estimate τ from Equation (1), and then $\mathbf{W}$ from Equation (2), where for simplicity, we would effectively ignore $\mathbf{V}$(choose $\mathbf{V} = \mathbf{I}$).

6.3 Expectation-Maximization(EM algorithm)

Definition 6.2. *Given a statistical model consisting of a set $\mathbf{X}$ of observed data, a set of unobserved latent data or missing values $\mathbf{Z}$, and a vector of unknown parameters $\boldsymbol{\theta}$, along with a likelihood function $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$, the maximum likelihood estimation (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data*

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax} \log p(\mathbf{X}|\boldsymbol{\theta}) = \mathrm{argmax} \log \int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{z}$$

However, this quantity is often intractable.

Intuition of EM:

- Introduce an auxiliary variable $\mathbf{z}$ and the pairs $(\mathbf{x}, \mathbf{z})$ are referred as *complete data.*

- $$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})\mathrm{d}\mathbf{z} \\
\log p(\mathbf{x}|\boldsymbol{\theta}) &= \log \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})d\mathbf{z} \\
&= \log \int \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} q(\mathbf{z})d\mathbf{z} \quad (q(\mathbf{z}) \text{ is a distribution of } \mathbf{z}) \\
&\geq \int \left(\log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \right) q(\mathbf{z})d\mathbf{z} \quad (\log \text{ is a concave function})
\end{aligned}
$$

- Assign $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})$, where t denotes the t^{th} iteration.

- $$
\begin{aligned}
\int \left(\log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \right) q(\mathbf{z})d\mathbf{z} &= \int \left(\log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})} \right) p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})d\mathbf{z} \\
&= \int (\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}))p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})d\mathbf{z} - \int (\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}^{(t)}))p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})d\mathbf{z}
\end{aligned}
$$

  where the second entry is a constant and the first entry is a function of $\boldsymbol{\theta}$. Specifically,

  $$\mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int (\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}))p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})d\mathbf{z}$$

Steps of EM:
The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

- **Expectation step (E step)**: Calculate the expected value of the log likelihood function, with respect to the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$: $\mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}}[\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})]$

- **Maximization step (M step)**: Find the parameter that maximizes this quantity:

  $$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

6.4 Probabilistic PCA(EM approach)

In the EM approach to maximizing the likelihood for PPCA, we consider the latent variables $\{\mathbf{z}_i\}$ to be "missing" data and the "complete" data to comprise the observations together with these latent variables. The corresponding complete-data log-likelihood is then:

$$L_c = \sum_{i=1}^{n} \log p(\mathbf{x}_i, \mathbf{z}_i)$$

where, in PPCA, $\log p(\mathbf{x}_i, \mathbf{z}_i)$ is

$$
\begin{aligned}
\log p(\mathbf{x}_i, \mathbf{z}_i) &= \log[p(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i)] \\
&= \log\left[\frac{1}{(2\pi)^{\frac{p}{2}}\tau^{\frac{p}{2}}}\exp(-\frac{1}{2\tau}||\mathbf{x}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_i||^2)exp(-\frac{1}{2}||\mathbf{z}_i||^2)\right] \\
&= -\frac{p}{2}\log\tau - \frac{1}{2\tau}||\mathbf{x}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_i||^2 - \frac{1}{2}||\mathbf{z}_i||^2 \quad \text{(Omit the constant coefficient)} \\
&= -\frac{p}{2}\log\tau - \frac{1}{2\tau}||\mathbf{x}_i - \boldsymbol{\mu}||^2 - \frac{1}{2\tau}||\mathbf{W}\mathbf{z}_i||^2 + \frac{1}{\tau}(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{W}\mathbf{z}_i) - \frac{1}{2}||\mathbf{z}_i||^2
\end{aligned}
$$

Note: The former notion of parameter $\boldsymbol{\theta}$ in EM algorithm is $\{\tau, \mathbf{W}\}$ in this case.

- **E-Step**: we take the expectation of L_c with respect to the distribution of $p(\mathbf{z}_i|\boldsymbol{\theta}_i)$:

$$
\begin{aligned}
\mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{n}\int(\log p(\mathbf{x}_i, \mathbf{z}_i))p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i \\
&= \sum_{i=1}^{n}\Big\{-\int\frac{p}{2}\log\tau p(\mathbf{z}_i|\boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i - \int\frac{1}{2\tau}||\mathbf{x}_i - \boldsymbol{\mu}||^2 p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i \\
&\quad - \int\frac{1}{2\tau}||\mathbf{W}\mathbf{z}_i||^2 p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i + \int\frac{1}{\tau}(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{W}\mathbf{z}_i)p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i \\
&\quad - \int\frac{1}{2}||\mathbf{z}_i||^2 p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i\Big\} \\
&= -\frac{np}{2}\log\tau - \frac{n}{2\tau}\mathrm{tr}(\mathbf{S}) \\
&\quad - \sum_{i=1}^{n}\int\left[\frac{1}{2\tau}||\mathbf{W}\mathbf{z}_i||^2 - \frac{1}{\tau}(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{W}\mathbf{z}_i) + \frac{1}{2}||\mathbf{z}_i||^2\right]p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i
\end{aligned}
$$

(The first two entries are independent from $\mathbf{z}_i$, and $\mathrm{tr}(\mathbf{S}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$)

We can define:

$$\langle\mathbf{z}_i\rangle = \int\mathbf{z}_i p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i = \mathbf{M}_{(t)}^{-1}\mathbf{W}_{(t)}^T(\mathbf{x}_i - \boldsymbol{\mu})$$

(Since $(\mathbf{z}|\mathbf{x} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$, where $\mathbf{M} = \sigma^2\mathbf{I}_q + \mathbf{W}^T\mathbf{W}$)

$$\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle = \int\mathbf{z}_i\mathbf{z}_i^T p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})\mathrm{d}\mathbf{z}_i = \tau_{(t)}\mathbf{M}_{(t)}^{-1} + \langle\mathbf{z}_i\rangle\langle\mathbf{z}_i\rangle^T$$

(Since $Cov(\mathbf{z}_i) = \mathbb{E}(\mathbf{z}_i\mathbf{z}_i^T) - \mathbb{E}(\mathbf{z}_i)\mathbb{E}(\mathbf{z}_i^T)$)

Hence, $\mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is

$$\mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = -\frac{np}{2}\log\tau - \frac{n}{2\tau}\mathrm{tr}(\mathbf{S}) - \sum_{i=1}^{n}\left\{\frac{1}{2\tau}\mathrm{tr}(\mathbf{W}^T\mathbf{W}\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle) - \frac{1}{\tau}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{W}\langle\mathbf{z}_i\rangle + \frac{1}{2}\mathrm{tr}(\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle)\right\}$$

- **M-Step:** Maximize $\mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$
  (a):

$$\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{W}} = \frac{1}{\tau}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})\langle\mathbf{z}_i^T\rangle - \frac{1}{\tau}\sum_{i=1}^{n}\mathbf{W}\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle = 0$$

$$\mathbf{W}\sum_{i=1}^{n}\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle = \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})\langle\mathbf{z}_i^T\rangle$$

$$\mathbf{W}^{(t+1)} = \left(\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})\langle\mathbf{z}_i^T\rangle\right)\left(\sum_{i=1}^{n}\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle\right)^{-1} \tag{3}$$

  (b):

$$\frac{\partial\mathbf{Q}}{\partial\tau} = -\frac{np}{2}\frac{1}{\tau} + \frac{n}{2\tau^2}\mathrm{tr}(\mathbf{S}) + \frac{1}{2\tau^2}\sum_{i=1}^{n}\mathrm{tr}(\mathbf{W}^T\mathbf{W}\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle) - \frac{1}{\tau^2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{W}\langle\mathbf{z}_i^T\rangle = 0$$

$$\tau^{(t+1)} = \frac{1}{p}[\mathrm{tr}(\mathbf{S}) + \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}(\mathbf{W}^T\mathbf{W}\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle) - \frac{2}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{W}\langle\mathbf{z}_i\rangle]$$

Since $\mathbf{W}\sum_{i=1}^{n}\langle\mathbf{z}_i, \mathbf{z}_i^T\rangle = \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})\langle\mathbf{z}_i^T\rangle$, we can combine the last two entries.

$$\tau^{(t+1)} = \frac{1}{p}[\mathrm{tr}(\mathbf{S}) - \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{W}^{(t+1)}\langle\mathbf{z}_i\rangle] \tag{4}$$

  In practice, we can first calculate $\mathbf{W}^{(t+1)}$ using Equation (3) and then substitute to Equation (4) to get $\tau^{(t+1)}$.

Summary:
The pros of EM approach is that it reduces the computational complexity by avoiding performing full SVD, which is used in the previous closed form approach.