Lecture Note 5: Dimensionality Reduction

Professor: Zhihua Zhang *Scribes: Yubo Xie, Binbin Li, Luo Luo*

5 Dimensionality Reduction

5.1 Population Principal Component Analysis (PCA)

Definition 5.1. If $\mathbf{x} \in \mathbb{R}^p$ is a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then the principal component transformation is

$$\mathbf{x} \to \mathbf{y} = \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu}),$$

where $\mathbf{U}$ is orthogonal, $\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U} = \boldsymbol{\Lambda}$ is diagonal and $\boldsymbol{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$).

Note that the decomposition $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ is a spectral decomposition. Based on the definition above, we have the following theorem:

Theorem 5.1. *If* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then* $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$. *Moreover,*
(a) $\mathbb{E}(y_i) = 0$;
(b) $\mathrm{Var}(y_i) = \lambda_i$;
(c) $\mathrm{Cov}(y_i, y_j) = 0$, *for* $i \neq j$;
(d) $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \cdots \geq \mathrm{Var}(y_p)$.

5.2 Sample Principal Component Analysis

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ be an $n \times p$ sample data matrix. Then the sample mean $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$, and the sample covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}^T\mathbf{H}\mathbf{X}$, where $\mathbf{H}$ is the centering matrix and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. Now consider the spectral decomposition

$$\mathbf{S} = \mathbf{G}\mathbf{L}\mathbf{G}^T,$$

where $\mathbf{G}$ and $\mathbf{L}$ are both $p \times p$ matrices, and the diagonal entries of $\mathbf{L} = \mathbf{diag}(l_1, l_2, \ldots, l_p)$ satisfy $l_1 \geq l_2 \geq \cdots \geq l_p \geq 0$. Letting $\mathbf{G}_q$ denote the matrix consisting of only the first q columns of $\mathbf{G}$ (obviously $q < p$), which makes $\mathbf{G}_q$ a $p \times q$ matrix, we have $\mathbf{G}_q^T\mathbf{G}_q = \mathbf{I}_q$. For each $\mathbf{x}_i$ ($1 \leq i \leq n$), we can perform the transformation

$$\mathbf{y}_i = \mathbf{G}_q^T(\mathbf{x}_i - \overline{\mathbf{x}}),$$

obtaining a vector that lies in $\mathbb{R}^q$. Putting all the $\mathbf{y}_i$'s together, we get an $n \times q$ matrix

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^T = [\mathbf{y}_1^{(q)}, \mathbf{y}_2^{(q)}, \ldots, \mathbf{y}_q^{(q)}],$$

where $\mathbf{y}_i^{(q)}$ ($1 \leq i \leq q$) is the i-th principal component of the sample data $\mathbf{X}$.

Often we need a threshold to decide q. Let the threshold be denoted by τ. We can choose some q such that $\sum_{i=1}^q l_i \Big/ \sum_{i=1}^p l_i \geq \tau$.

5.3 Principal Coordinate Analysis (PCO)

Still consider the sample covariance matrix $\mathbf{S} = \mathbf{X}^T\mathbf{H}\mathbf{X}$ (we can drop the constant $1/n$ here). Note that the centering matrix $\mathbf{H}$ is *idempotence*, which means $\mathbf{H}^2 = \mathbf{H}$. Therefore, we can rewrite the covariance matrix as $\mathbf{S} = \mathbf{X}^T\mathbf{H}\mathbf{H}\mathbf{X}$. Switching the two parts $\mathbf{X}^T\mathbf{H}$ and $\mathbf{H}\mathbf{X}$, we get an $n \times n$ matrix

$$\mathbf{T} = \mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}.$$

It can be shown that $\mathbf{S}$ and $\mathbf{T}$ actually have the same non-zero eigenvalues. Let $\mathbf{H}\mathbf{X}$ has SVD $\mathbf{H}\mathbf{X} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T$, where $\mathbf{U}$ is $n \times p$ column orthogonal, $\mathbf{\Gamma}$ is $p \times p$ diagonal and $\mathbf{V}$ is $p \times p$ orthogonal. Then we have $\mathbf{S} = \mathbf{V}\mathbf{\Gamma}^2\mathbf{V}^T$ and $\mathbf{T} = \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^T$ which implies $\mathbf{V}$ is $\mathbf{G}$, $\mathbf{\Gamma}^2$ is $\mathbf{L}$ and $\mathbf{T}\mathbf{U}\mathbf{\Gamma} = \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^T\mathbf{U}\mathbf{\Gamma} = \mathbf{U}\mathbf{\Gamma}\mathbf{L}$. Since $\mathbf{S} = \mathbf{G}\mathbf{L}\mathbf{G}^T$, $\mathbf{S}$ and $\mathbf{T}$, the non-zero eigenvalues of $\mathbf{S}$ and $\mathbf{T}$ are the diagonal entries of $\mathbf{L}$.

Definition 5.2. Let $\mathbf{z}_i$ be the i-th eigenvector of $\mathbf{T}$, *i.e.*, $\mathbf{T}\mathbf{z}_i = \lambda_i\mathbf{z}_i$, and it is normalized such that $\mathbf{z}_i^T\mathbf{z}_i = \lambda_i$ $(1 \le i \le p)$. For fixed k $(1 \le k \le p)$, the rows of $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k)$ are called the principal coordinates of $\mathbf{X}$ in k-dimension.

5.4 Kernel PCA

Suppose there exists a kernel $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{\Phi}(\mathbf{x}), \mathbf{\Phi}(\mathbf{y}) \rangle = \mathbf{\Phi}(\mathbf{x})^T\mathbf{\Phi}(\mathbf{y})$ such that $\mathbf{\Phi}(\cdot)$ projects a vector $\mathbf{x}$ in $\mathbb{R}^p$ into some higher dimensional space $\mathbb{R}^r$. Applying this to some sample data $\mathbf{X}$ (an $n \times p$ matrix), we have the transformation

$$\mathbf{X} \to \mathbf{F},$$

where $\mathbf{F}$ is an $n \times r$ matrix. We usually have no idea about $\mathbf{F}$, but we do know the kernel K, which means we know $\mathbf{K} = \mathbf{F}\mathbf{F}^T$. If we wish to do PCA with respect to the transformed data $\mathbf{F}$, explicitly calculating the data in the higher dimensional space can be avoidable. More specifically, suppose we want to figure out one eigenvector of the covariance matrix $\mathbf{F}^T\mathbf{H}\mathbf{F}$ (there is no loss of generality since solving others is the same), denoted by $\mathbf{u}$ and apparently satisfying $\mathbf{F}^T\mathbf{H}\mathbf{F}\mathbf{u} = \lambda\mathbf{u}$, which in turn can be rewritten as

$$\mathbf{F}^T\mathbf{H}\mathbf{H}\mathbf{F}\mathbf{u} = \lambda\mathbf{u},$$

where λ is the corresponding eigenvalue (assume it's positive). Like before, we consider the slightly different form of the covariance matrix and perform the decomposition

$$\mathbf{H}\mathbf{F}\mathbf{F}^T\mathbf{H} = \mathbf{H}\mathbf{K}\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{U}^T.$$

λ is also the eigenvalue of $\mathbf{H}\mathbf{K}\mathbf{H}$, and we know the corresponding eigenvector is $\mathbf{v}$. That is,

$$\mathbf{H}\mathbf{K}\mathbf{H}\mathbf{v} = \lambda\mathbf{v}, \tag{1}$$

and $\mathbf{v}$ is normalized such that $\mathbf{v}^T\mathbf{v} = 1$. We now show the relationship between $\mathbf{u}$ and $\mathbf{v}$.

Multiplying Equation (1) on both sides with $\mathbf{F}^T\mathbf{H}$, we get

$$\mathbf{F}^T\mathbf{H}\mathbf{H}\mathbf{K}\mathbf{H}\mathbf{v} = \lambda\mathbf{F}^T\mathbf{H}\mathbf{v}.$$

Substituting $\mathbf{K}$ with $\mathbf{F}\mathbf{F}^T$ and using the fact that $\mathbf{H}$ is idempotence, we obtain

$$\mathbf{F}^T\mathbf{H}\mathbf{F}\mathbf{F}^T\mathbf{H}\mathbf{v} = \lambda\mathbf{F}^T\mathbf{H}\mathbf{v}.$$

It can be seen that $\mathbf{F}^T\mathbf{H}\mathbf{v}$ is the eigenvector of $\mathbf{F}^T\mathbf{H}\mathbf{F}$ associated with λ.

We can normalize $\mathbf{F}^T\mathbf{H}\mathbf{v}$ to get

$$\begin{aligned}
\mathbf{u} &= \frac{\mathbf{F}^T\mathbf{H}\mathbf{v}}{\|\mathbf{F}^T\mathbf{H}\mathbf{v}\|_2} \\
&= \frac{\mathbf{F}^T\mathbf{H}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{H}\mathbf{F}\mathbf{F}^T\mathbf{H}\mathbf{v}}} \\
&= \frac{\mathbf{F}^T\mathbf{H}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{H}\mathbf{K}\mathbf{H}\mathbf{v}}} \\
&= \frac{\mathbf{F}^T\mathbf{H}\mathbf{v}}{\sqrt{\lambda\mathbf{v}^T\mathbf{v}}} \\
&= \lambda^{-1/2}\mathbf{F}^T\mathbf{H}\mathbf{v}.
\end{aligned}$$

Now take one sample $\mathbf{f}_i$ to do the principal component transformation:

$$\begin{aligned}
\left(\mathbf{f}_i - \frac{1}{n}\sum_{j=1}^{n}\mathbf{f}_j\right)^T \mathbf{u} &= \left(\mathbf{f}_i - \frac{1}{n}\mathbf{F}^T\mathbf{1}_n\right)^T \lambda^{-1/2}\mathbf{F}^T\mathbf{H}\mathbf{v} \\
&= \lambda^{-1/2}\left(\mathbf{f}_i^T\mathbf{F}^T - \frac{1}{n}\mathbf{1}_n^T\mathbf{F}\mathbf{F}^T\right)\mathbf{H}\mathbf{v} \\
&= \lambda^{-1/2}\left(\mathbf{f}_i^T\mathbf{F}^T - \frac{1}{n}\mathbf{1}_n^T\mathbf{K}\right)\mathbf{H}\mathbf{v},
\end{aligned}$$

where $\mathbf{f}_i^T\mathbf{F}^T$ can be computed using the kernel, *i.e.*,

$$\begin{aligned}
\mathbf{f}_i^T\mathbf{F}^T &= \left[\mathbf{f}_i^T\mathbf{f}_1, \mathbf{f}_i^T\mathbf{f}_2, \ldots, \mathbf{f}_i^T\mathbf{f}_i, \ldots, \mathbf{f}_i^T\mathbf{f}_n\right] \\
&= \left[K(\mathbf{x}_i, \mathbf{x}_1), K(\mathbf{x}_i, \mathbf{x}_2), \ldots, K(\mathbf{x}_i, \mathbf{x}_i), \ldots, K(\mathbf{x}_i, \mathbf{x}_n)\right].
\end{aligned}$$

5.5 The Eckart-Young Theorem

Theorem 5.2. (Eckart & Young) *Let* $\mathbf{A}$ *be a given* $n \times p$ *matrix and*

$$\begin{aligned}
\Phi(\mathbf{V}, \mathbf{Z}) &= \frac{1}{2}\mathrm{tr}\left((\mathbf{A} - \mathbf{Z}\mathbf{V}^T)(\mathbf{A} - \mathbf{Z}\mathbf{V}^T)^T\right) \\
&= \frac{1}{2}\left\|\mathbf{A} - \mathbf{Z}\mathbf{V}^T\right\|_F^2,
\end{aligned}$$

where $\mathbf{V} \in \mathbb{R}^{p\times q}$ *satisfies* $\mathbf{V}^T\mathbf{V} = \mathbf{I}_q$ *and* $\mathbf{Z} \in \mathbb{R}^{n\times q}$. *The minimum of* Φ *is obtained when* $\mathbf{V}$ *is a* $p \times q$ *matrix of orthogonal eigenvectors associated with the* q *largest eigenvalues of* $\mathbf{A}^T\mathbf{A}$ *and* $\mathbf{Z} = \mathbf{A}\mathbf{V}$.

Proof. The Lagrange multiplier is

$$\mathcal{L} = \frac{1}{2}\mathrm{tr}\left((\mathbf{A} - \mathbf{Z}\mathbf{V}^T)(\mathbf{A} - \mathbf{Z}\mathbf{V}^T)^T\right) - \mathrm{tr}\left(\mathbf{C}(\mathbf{V}^T\mathbf{V} - \mathbf{I}_q)\right),$$

where $\mathbf{C}$ is symmetric.

Taking derivative with respect to $\mathbf{Z}$, we have

$$\begin{aligned}\mathrm{d}\mathcal{L} &= \frac{1}{2}\mathrm{tr}\left((-\mathrm{d}\mathbf{Z}\mathbf{V}^T)(\mathbf{A} - \mathbf{Z}\mathbf{V}^T)^T + (\mathbf{A} - \mathbf{Z}\mathbf{V}^T)(-\mathbf{V}\mathrm{d}\mathbf{Z}^T)\right) \\ &= -\mathrm{tr}\left((\mathbf{A} - \mathbf{Z}\mathbf{V}^T)\mathbf{V}\mathrm{d}\mathbf{Z}^T\right).\end{aligned}$$

Therefore

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}} = -(\mathbf{A} - \mathbf{Z}\mathbf{V}^T)\mathbf{V} = \mathbf{0}. \tag{2}$$

Taking derivative with respect to $\mathbf{V}$, we have

$$\begin{aligned}\mathrm{d}\mathcal{L} &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{Z}\mathrm{d}\mathbf{V}^T(\mathbf{A} - \mathbf{Z}\mathbf{V}^T)^T + (\mathbf{A} - \mathbf{Z}\mathbf{V}^T)\mathrm{d}\mathbf{V}\mathbf{Z}^T\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}(\mathrm{d}\mathbf{V}^T\mathbf{V} + \mathbf{V}^T\mathrm{d}\mathbf{V})\right) \\ &= -\mathrm{tr}\left((\mathbf{A} - \mathbf{Z}\mathbf{V}^T)^T\mathbf{Z}\mathrm{d}\mathbf{V}^T\right) - \mathrm{tr}\left(\mathbf{V}\mathbf{C}\mathrm{d}\mathbf{V}^T\right) \\ &= -\mathrm{tr}\left((\mathbf{A}^T\mathbf{Z} - \mathbf{V}\mathbf{Z}^T\mathbf{Z} + \mathbf{V}\mathbf{C})\mathrm{d}\mathbf{V}^T\right).\end{aligned}$$

Therefore

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = -(\mathbf{A}^T\mathbf{Z} - \mathbf{V}\mathbf{Z}^T\mathbf{Z} + \mathbf{V}\mathbf{C}) = \mathbf{0},$$

i.e.,

$$\mathbf{A}^T\mathbf{Z} - \mathbf{V}\mathbf{Z}^T\mathbf{Z} + \mathbf{V}\mathbf{C} = \mathbf{0}. \tag{3}$$

Also, taking derivative with respect to $\mathbf{C}$, we have $\partial\mathcal{L}/\partial\mathbf{C} = \mathbf{V}^T\mathbf{V} - \mathbf{I}_q = \mathbf{0}$. It can be deduced from Equation (2) that $\mathbf{Z} = \mathbf{A}\mathbf{V}$. Multiplying Equation (3) on both sides with $\mathbf{V}^T$, we obtain

$$\mathbf{V}^T\mathbf{A}^T\mathbf{Z} - \mathbf{V}^T\mathbf{V}\mathbf{Z}^T\mathbf{Z} + \mathbf{V}^T\mathbf{V}\mathbf{C} = \mathbf{0}.$$

Using the fact that $\mathbf{Z} = \mathbf{A}\mathbf{V}$, we further deduce that $\mathbf{C} = \mathbf{0}$.

Combine all the results we already have, and we get

$$\begin{cases} \mathbf{Z} = \mathbf{A}\mathbf{V}, & (4) \\ \mathbf{A}^T\mathbf{Z} - \mathbf{V}\mathbf{Z}^T\mathbf{Z} = \mathbf{0}. & (5) \end{cases}$$

Substitute Equation (4) into Equation (5), and then

$$\mathbf{A}^T\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{V}^T\mathbf{A}^T\mathbf{A}\mathbf{V}. \tag{6}$$

Equation (6) along with Equation (4) should have given the optimal solution, but we don't know whether this gives the minimum or the maximum. Therefore we need to further take second derivatives, which would be too lengthy to describe here. Thus we won't give a strict proof. Just suppose

$$\hat{\mathbf{V}}^T\mathbf{A}^T\mathbf{A}\hat{\mathbf{V}} = \mathbf{\Gamma} = \mathbf{diag}(\gamma_1, \gamma_2, \ldots, \gamma_q),$$

and we are going to verify that, to actually achieve the minimum, $\gamma_1, \gamma_2, \ldots, \gamma_q$ should be the q largest eigenvalues of $\mathbf{A}^T\mathbf{A}$.

Consider the object function (we have dropped the constant)

$$\begin{aligned}
\operatorname{tr}\left((\mathbf{A}-\hat{\mathbf{Z}}\hat{\mathbf{V}}^T)(\mathbf{A}-\hat{\mathbf{Z}}\hat{\mathbf{V}}^T)^T\right) &= \operatorname{tr}\left(\mathbf{A}(\mathbf{I}-\hat{\mathbf{V}}\hat{\mathbf{V}}^T)(\mathbf{I}-\hat{\mathbf{V}}\hat{\mathbf{V}}^T)\mathbf{A}^T\right) \\
&= \operatorname{tr}\left(\mathbf{A}(\mathbf{I}-\hat{\mathbf{V}}\hat{\mathbf{V}}^T)\mathbf{A}^T\right) \\
&= \operatorname{tr}\left(\mathbf{A}\mathbf{A}^T\right) - \operatorname{tr}\left(\mathbf{A}\hat{\mathbf{V}}\hat{\mathbf{V}}^T\mathbf{A}^T\right) \\
&= \operatorname{tr}\left(\mathbf{A}\mathbf{A}^T\right) - \operatorname{tr}\left(\hat{\mathbf{V}}^T\mathbf{A}^T\mathbf{A}\hat{\mathbf{V}}\right) \\
&= \sum_{i=1}^{p}\gamma_i - \sum_{j=1}^{q}\gamma_{\pi_j}.
\end{aligned}$$

It can be seen that the minimum is achieved when the second term on the right of the last equality is maximized, which can be obtained by choosing the q largest eigenvalues. □

6 Probabilistic PCA

Traditional PCA is often formulated as a projection from the data space onto some linear subspace. However, 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 $\mathbf{x} \in \mathbb{R}^p$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ and $\mathbf{z} \perp\!\!\!\perp \boldsymbol{\epsilon}$. $\mathbf{z}$ is a latent variable and $\mathbf{W}$ is a loading matrix. Also we have $(\mathbf{x}-\boldsymbol{\mu} \mid \mathbf{z}) \sim \mathcal{N}(\mathbf{W}\mathbf{z}, \sigma^2\mathbf{I}_p)$.

We want to figure out the distribution of $\mathbf{x}-\boldsymbol{\mu}$. Consider

$$\begin{aligned}
\mathbf{x}-\boldsymbol{\mu} &= \mathbf{W}\mathbf{z} + \sigma\boldsymbol{\epsilon} \\
&= \begin{bmatrix}\mathbf{W} & \sigma\mathbf{I}_p\end{bmatrix}\begin{bmatrix}\mathbf{z} \\ \boldsymbol{\epsilon}\end{bmatrix},
\end{aligned}$$

where $\begin{bmatrix}\mathbf{z} \\ \boldsymbol{\epsilon}\end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+q})$. Therefore, $(\mathbf{x}-\boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}_p$.

Now that we have known $p(\mathbf{z})$, $p(\mathbf{x}-\boldsymbol{\mu} \mid \mathbf{z})$ and $p(\mathbf{x}-\boldsymbol{\mu})$, we can deduce the distribution of $\mathbf{z}$ given $\mathbf{x}-\boldsymbol{\mu}$. Let $\mathbf{M} = \sigma^2\mathbf{I}_q + \mathbf{W}^T\mathbf{W}$. In particular,

$$\begin{aligned}
p(\mathbf{z} \mid \mathbf{x}-\boldsymbol{\mu}) &= \frac{p(\mathbf{x}-\boldsymbol{\mu} \mid \mathbf{z})p(\mathbf{z})}{p(\mathbf{x}-\boldsymbol{\mu})} \\
&\propto \exp\Big(-\frac{1}{2\sigma^2}\big(\mathbf{x}-\boldsymbol{\mu}-\mathbf{W}\mathbf{z}\big)^T\big(\mathbf{x}-\boldsymbol{\mu}-\mathbf{W}\mathbf{z}\big)\Big)\exp\Big(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\Big) \\
&\propto \exp\Big(-\frac{1}{2\sigma^2}\big(\mathbf{z}^T(\sigma^2\mathbf{I}_q + \mathbf{W}^T\mathbf{W})\mathbf{z} - 2\mathbf{z}^T\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu})\big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\big(\mathbf{z}^T\sigma^{-2}\mathbf{M}\mathbf{z} - 2\sigma^{-2}\mathbf{z}^T\mathbf{M}\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu})\big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\big(\mathbf{z}^T(\sigma^2\mathbf{M}^{-1})^{-1}\mathbf{z} - 2\mathbf{z}^T(\sigma^2\mathbf{M}^{-1})^{-1}\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu})\big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\big(\mathbf{z}-\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu})\big)^T(\sigma^2\mathbf{M}^{-1})^{-1}\big(\mathbf{z}-\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu})\big)\Big).
\end{aligned}$$

Thus, we can conclude that $(\mathbf{z} \mid \mathbf{x} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$.

Now we can compute the optimized result using Maximum Likelihood Estimation. If we have a sample set $\{\mathbf{x}_1, \dots \mathbf{x}_n\} \subseteq \mathbb{R}^p$, where $\mathbf{x}_i = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon}$, and the $\mathbf{x}_i$'s are independent and identically distributed. Then the likelihood function is

$$\begin{aligned} L &= \prod_{i=1}^{n} \frac{1}{|\mathbf{C}|^{\frac{1}{2}}} \exp\left(\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right) \\ &= \frac{1}{|\mathbf{C}|^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right). \end{aligned}$$

And the log-likelihood is

$$\log L = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}).$$

Take the negative, and let

$$\begin{aligned} f = -\log L &= \frac{n}{2}\log|\mathbf{C}| + \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \\ &= \frac{n}{2}\log|\mathbf{C}| + \frac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right). \end{aligned}$$

When $\mathrm{d}f/\mathrm{d}\boldsymbol{\mu} = \mathbf{0}$, we have $\sum_{i=1}^{n}\mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{0}$, *i.e.*, $\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{0}$, from which we can obtain $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$.

Also we have

$$\begin{aligned} f &= \frac{n}{2}\log|\mathbf{C}| + \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right)^T\mathbf{C}^{-1}\left(\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right) \\ &= \frac{n}{2}\log|\mathbf{C}| + \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right)\left(\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right)^T\right) \\ &= \frac{n}{2}\log|\mathbf{C}| + \frac{n}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right)\left(\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right)^T\right) \\ &= \frac{n}{2}\log|\mathbf{C}| + \frac{n}{2}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S}). \end{aligned}$$