Lecture Notes 11: Classification with Dimensionality Reduction

Professor: Zhihua Zhang *Scribe:Tianyuan Liu, Shenjian Zhao*

1 Fisher Discriminant Analysis (Cont'd)

Recall Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ be an $n \times p$ matrix and $\mathbf{Y} = \mathbf{HXG}$ be an $n \times q$ matrix ($q < p$). Our goal is to make the *with-in covariance* $\mathbf{S}_w$ 'small' and the *between covariance* $\mathbf{S}_b$ 'large', *i.e.*

$$\max_{\mathbf{G}} \operatorname{tr}((\mathbf{G}^T\mathbf{S}_w\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{S}_b\mathbf{G}))$$

In the last lecture, we have derived that

$$\mathbf{S}_b\mathbf{G} = \mathbf{S}_w\mathbf{G}\mathbf{\Lambda}$$

Now we move to solve this augmented eigenvalue problem. Consider each column $\mathbf{g}_i$ of matrix $\mathbf{G}$

$$\begin{aligned} \mathbf{S}_b\mathbf{g}_i &= \lambda_i\mathbf{S}_w\mathbf{g}_i \\ &= \lambda_i(\mathbf{S}_t - \mathbf{S}_b)\mathbf{g}_i \\ \mathbf{S}_b\mathbf{g}_i &= \frac{\lambda_i}{1+\lambda_i}\mathbf{S}_t\mathbf{g}_i \qquad (1) \end{aligned}$$

If $\mathbf{S}_t$ is invertible, equation (1) can be solved as an ordinary eigenvalue problem

$$\mathbf{S}_t^{-1}\mathbf{S}_b\mathbf{g}_i = \lambda\mathbf{g}_i$$

where$\lambda = \frac{\lambda_i}{1+\lambda_i}$. But it's often the case that $\mathbf{S}_t$ is not invertible. We give the following theorem to show that equation (1) can be solved using the psudo-inverse $\mathbf{S}_t^\dagger$.

Theorem 1.1. *Let $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ be two $m \times m$ real matrices. Assume $\mathcal{R}(\mathbf{\Sigma}_1) \subseteq \mathcal{R}(\mathbf{\Sigma}_2)$. (Here $\mathcal{R}(\cdot)$ is the range of a matrix) Then if $(\mathbf{\Lambda}, \mathbf{A})$ are the nonzero eigenpairs of $\mathbf{\Sigma}_2^\dagger\mathbf{\Sigma}_1$, we have that $(\mathbf{\Lambda}, \mathbf{A})$ are the nonzero eigenpairs of matrix pencil $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$,* i.e.

$$\mathbf{\Sigma}_1\mathbf{A} = \mathbf{\Sigma}_2\mathbf{A}\mathbf{\Lambda} \Longleftrightarrow \mathbf{\Sigma}_2^\dagger\mathbf{\Sigma}_1\mathbf{A} = \mathbf{A}\mathbf{\Lambda}$$

Since $\mathcal{R}(\mathbf{S}_b) = \mathcal{R}(\mathbf{X}^T\mathbf{HE\Pi}^{-1}\mathbf{E}^T\mathbf{HX}) = \mathcal{R}(\mathbf{X}^T\mathbf{HE\Pi}^{-1/2})$ and $\mathcal{R}(\mathbf{S}_t) = \mathcal{R}(\mathbf{X}^T\mathbf{HHX}) = \mathcal{R}(\mathbf{X}^T\mathbf{H})$, we have $\mathcal{R}(\mathbf{S}_b) \subseteq \mathcal{R}(\mathbf{S}_t)$. By Theorem 1.1, equation (1) is equivalent to

$$\mathbf{S}_t^\dagger\mathbf{S}_b\mathbf{g} = \lambda\mathbf{g}$$

Now we move to prove Theorem 1.1.

Proof. Let $\mathbf{\Sigma}_1 = \mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}_1^T$ and $\mathbf{\Sigma}_2 = \mathbf{U}_2\mathbf{\Gamma}_2\mathbf{V}_2^T$ be the condensed SVD of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Then we have $\mathcal{R}(\mathbf{\Sigma}_1) = \mathcal{R}(\mathbf{U}_1)$ and $\mathcal{R}(\mathbf{\Sigma}_2) = \mathcal{R}(\mathbf{U}_2)$. The psudo-inverse is $\mathbf{\Sigma}_2^\dagger = \mathbf{V}_2\mathbf{\Gamma}_2^{-1}\mathbf{U}_2^T$, thus $\mathbf{\Sigma}_2\mathbf{\Sigma}_2^\dagger = \mathbf{U}_2\mathbf{U}_2^T$. Given $\mathcal{R}(\mathbf{\Sigma}_1) \subseteq \mathcal{R}(\mathbf{\Sigma}_2)$, we have $\mathcal{R}(\mathbf{U}_1) \subseteq \mathcal{R}(\mathbf{U}_2)$. Further, we can assume there is some $\mathbf{Q}$ such that $\mathbf{U}_1 = \mathbf{U}_2\mathbf{Q}$

$$\begin{aligned}
\mathbf{\Sigma}_2\mathbf{\Sigma}_2^\dagger\mathbf{\Sigma}_1 &= \mathbf{U}_2\mathbf{U}_2^T\mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}_1^T \\
&= \mathbf{U}_2\mathbf{U}_2^T\mathbf{U}_2\mathbf{Q}\mathbf{\Gamma}\mathbf{V}_1^T \\
&= \mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}_1^T \\
&= \mathbf{\Sigma}_1
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{\Sigma}_2^\dagger\mathbf{\Sigma}_1\mathbf{A} &= \mathbf{A}\mathbf{\Lambda} \\
\mathbf{\Sigma}_2\mathbf{\Sigma}_2^\dagger\mathbf{\Sigma}_1\mathbf{A} &= \mathbf{\Sigma}_2\mathbf{A}\mathbf{\Lambda} \\
\mathbf{\Sigma}_1\mathbf{A} &= \mathbf{\Sigma}_2\mathbf{A}\mathbf{\Lambda}
\end{aligned}$$

□

2 Method 2: Complete Orthogonal Decomposition

Definition 2.1 (Generalized Singular Value Decomposition)**.** *For $\mathbf{A} \in \mathbb{R}^{n\times m}$,$\mathbf{B} \in \mathbb{R}^{k\times m}$, their GSVD is given by:*

$$\begin{aligned}
&\mathbf{U}^T\mathbf{A}\mathbf{X} = \mathbf{C} = \mathrm{diag}(\alpha_1, \cdots, \alpha_m) = [\mathbf{\Sigma_A}, \mathbf{0}], \alpha_i \geq 0 \\
&\mathbf{V}^T\mathbf{B}\mathbf{X} = \mathbf{S} = \mathrm{diag}(\beta_1, \cdots, \beta_q) = [\mathbf{\Sigma_B}, \mathbf{0}], \beta_i \geq 0, q = \min(k, m),
\end{aligned}$$

where $\mathbf{U} \in \mathbb{R}^{n\times n}$, $\mathbf{V} \in \mathbb{R}^{k\times k}$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I_n}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I_k}$,$\mathbf{X} \in \mathbb{R}^{m\times m}$ is nonsingular. It holds that ${\mathbf{\Sigma_A}}^T\mathbf{\Sigma_A} + {\mathbf{\Sigma_B}}^T\mathbf{\Sigma_B} = \mathbf{I_m}$.

Proposition 2.1. *Application of GSVD. For $\mathbf{A} \in \mathbb{R}^{n\times m}$,$\mathbf{B} \in \mathbb{R}^{k\times m}$, their GSVD is*

$$\begin{aligned}
\mathbf{U}^T\mathbf{A}\mathbf{X} &= [\mathbf{\Sigma_A}, \mathbf{0}], \\
\mathbf{V}^T\mathbf{B}\mathbf{X} &= [\mathbf{\Sigma_B}, \mathbf{0}]
\end{aligned}$$

$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_m)$,*then first r vectors of $\mathbf{X}$ are the generalized eigenvectors of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{B}^T\mathbf{B}$, and the corresponding eigenvalue is $\dfrac{\alpha_i^2}{\beta_i^2}$.*

Proof: From GSVD:

$$\begin{aligned}
\mathbf{A} = \mathbf{U}[\mathbf{\Sigma_A}, \mathbf{0}]\mathbf{X}^{-1} &\Longrightarrow \mathbf{A}^T\mathbf{A} = \mathbf{X}^{-T}[\mathbf{\Sigma_A}, \mathbf{0}]^T[\mathbf{\Sigma_A}, \mathbf{0}]\mathbf{X}^{-1} \\
\mathbf{B} = \mathbf{V}[\mathbf{\Sigma_B}, \mathbf{0}]\mathbf{X}^{-1} &\Longrightarrow \mathbf{B}^T\mathbf{B} = \mathbf{X}^{-T}[\mathbf{\Sigma_B}, \mathbf{0}]^T[\mathbf{\Sigma_B}, \mathbf{0}]\mathbf{X}^{-1}
\end{aligned}$$

Because $\mathbf{X}$ is nonsingular,

$$\begin{aligned}
\mathbf{A}^T\mathbf{A}\mathbf{X} &= \mathbf{X}^{-T}\begin{pmatrix} {\mathbf{\Sigma_A}}^T\mathbf{\Sigma_A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\
\mathbf{B}^T\mathbf{B}\mathbf{X} &= \mathbf{X}^{-T}\begin{pmatrix} {\mathbf{\Sigma_B}}^T\mathbf{\Sigma_B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}
\end{aligned}$$

Let $\mathbf{x}_i, i \leq r$, then

$$\mathbf{A}^T\mathbf{A}\mathbf{x}_i = \mathbf{X}^{-T}\begin{pmatrix}\alpha_i^2\\0\\\vdots\\0\end{pmatrix}$$

$$\mathbf{B}^T\mathbf{B}\mathbf{x}_i = \mathbf{X}^{-T}\begin{pmatrix}\beta_i^2\\0\\\vdots\\0\end{pmatrix}$$

We get $\mathbf{A}^T\mathbf{A}\mathbf{x}_i = \frac{\alpha_i^2}{\beta_i^2}\mathbf{B}^T\mathbf{B}\mathbf{x}_i$. This is the solution to the generalized eigenvalue problems.

Lemma 2.1. *The CS Decomposition: Consider matrix*

$$\mathbf{Q} = \begin{pmatrix}\mathbf{Q}_1\\\mathbf{Q}_2\end{pmatrix}, \mathbf{Q}_1 \in \mathbb{R}^{m_1\times n}, \mathbf{Q}_2 \in \mathbb{R}^{m_2\times n},$$

where $m_1 \geq n$, $m_2 \geq n$, if the columns of $\mathbf{Q}$ are orthogonal, then exist orthogonal matrices $\mathbf{U}_1 \in \mathbb{R}^{m_1\times m_1}, \mathbf{U}_2 \in \mathbb{R}^{m_2\times m_2}, \mathbf{V}_1 \in \mathbb{R}^{n\times n}$, such that

$$\mathbf{U}_1^T\mathbf{Q}_1\mathbf{V}_1 = \mathbf{C}, \qquad \mathbf{U}_2^T\mathbf{Q}_2\mathbf{V}_1 = \mathbf{S}, \qquad \mathbf{C}^T\mathbf{C} + \mathbf{S}^T\mathbf{S} = \mathbf{I}_n$$

Proposition 2.2. *Using QR decomposition to solve GSVD.*

Proof: Let $\mathbf{A} \in \mathbb{R}^{n\times m}$,$\mathbf{B} \in \mathbb{R}^{k\times m}$,

$$\mathbf{C} = \begin{pmatrix}\mathbf{A}\\\mathbf{B}\end{pmatrix}$$

and $t = \text{rank}(\mathbf{C})$, perform QR to $\mathbf{C}$, gets

$$\mathbf{P}^T\mathbf{C}\mathbf{Q} = \begin{pmatrix}\mathbf{R} & \mathbf{0}\\\mathbf{0} & \mathbf{0}\end{pmatrix}$$

Where $\mathbf{Q}$ is a permutation matrix, $\mathbf{P}$ is an orthogonal matrix.$\mathbf{R}_{t\times t}$ is nonsingular. Let

$$\mathbf{P} = \begin{pmatrix}\mathbf{P}_{11} & \mathbf{P}_{12}\\\mathbf{P}_{21} & \mathbf{P}_{22}\end{pmatrix}$$

Where $\mathbf{P}_{11} \in \mathbb{R}^{n\times t}, \mathbf{P}_{21} \in \mathbb{R}^{k\times t}, ||\mathbf{P}|| \leq 1, ||\mathbf{P}_{11}|| \leq 1$.
From the lemma, first apply SVD on $\mathbf{P}_{11}$, we will have $\mathbf{U}^T\mathbf{P}_{11}\mathbf{W} = \mathbf{\Sigma_A}$, then apply QR to $\mathbf{P}_{21}\mathbf{W}$, we will have $\mathbf{P}_{21}\mathbf{W} = \mathbf{V}\mathbf{L}$.So

$$\begin{pmatrix}\mathbf{\Sigma_A}\\\mathbf{L}\end{pmatrix} = \begin{pmatrix}\mathbf{U}^T\mathbf{P}_{11}\\\mathbf{V}^T\mathbf{P}_{21}\end{pmatrix}\mathbf{W} = \begin{pmatrix}\mathbf{U}^T & \mathbf{0}\\\mathbf{0} & \mathbf{V}^T\end{pmatrix}\begin{pmatrix}\mathbf{P}_{11}\\\mathbf{P}_{21}\end{pmatrix}\mathbf{W}$$

From Lemma, we have

$$\begin{pmatrix} \mathbf{\Sigma_A}^T & \mathbf{L}^T \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma_A} \\ \mathbf{L} \end{pmatrix} = \mathbf{\Sigma_A}^T \mathbf{\Sigma_A} + \mathbf{L}^T \mathbf{L} = \mathbf{I}$$

Because $\mathbf{\Sigma_B}$ need not to be diagonal matrix, we use $\mathbf{\Sigma_B}$ to denote $\mathbf{L}$. To simplify:

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \mathbf{Q} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{11}\mathbf{R} & \mathbf{0} \\ \mathbf{P}_{21}\mathbf{R} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{U}\mathbf{\Sigma_A}\mathbf{W}^T\mathbf{R} & \mathbf{0} \\ \mathbf{V}\mathbf{\Sigma_B}\mathbf{W}^T\mathbf{R} & \mathbf{0} \end{pmatrix}$$

From above, we get

$$\mathbf{AQ} = [\mathbf{U}\mathbf{\Sigma_A}\mathbf{W}^T\mathbf{R}, \mathbf{0}]$$
$$\mathbf{BQ} = [\mathbf{V}\mathbf{\Sigma_B}\mathbf{W}^T\mathbf{R}, \mathbf{0}]$$

Change form:

$$\mathbf{U}^T\mathbf{AQ} = [\mathbf{\Sigma_A}\mathbf{W}^T\mathbf{R}, \mathbf{0}] = [\mathbf{\Sigma_A}, \mathbf{0}] \begin{pmatrix} \mathbf{W}^T\mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$
$$\mathbf{V}^T\mathbf{BQ} = [\mathbf{\Sigma_B}\mathbf{W}^T\mathbf{R}, \mathbf{0}] = [\mathbf{\Sigma_B}, \mathbf{0}] \begin{pmatrix} \mathbf{W}^T\mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

Because $\mathbf{W}^T\mathbf{R}$ is invertible, we can set

$$\mathbf{X} = \mathbf{Q} \begin{pmatrix} \mathbf{R}^{-\mathbf{1}}\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

Then come the solution of GSVD:

$$\mathbf{U}^T\mathbf{AX} = [\mathbf{\Sigma_A}, \mathbf{0}]$$
$$\mathbf{V}^T\mathbf{BX} = [\mathbf{\Sigma_B}, \mathbf{0}]$$

Example 2.1. *The Step to Solve FDA.*

1. *Compute* $\mathbf{S_t} = \mathbf{X}^T\mathbf{HHX}$, $\mathbf{S_b} = \mathbf{X}^T\mathbf{HE}\mathbf{\Pi}^{-\frac{1}{2}}\mathbf{\Pi}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{HX}$.
2. *Let* $\mathbf{A} = \mathbf{\Pi}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{HX}$,$\mathbf{B} = \mathbf{HX}$.
3. *Let*

$$\mathbf{C} = \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix}$$

4. *Apply QR to* $\mathbf{C}$*, get*

$$\mathbf{P}^T\mathbf{CQ} = \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

5. *Perform the SVD of* $\mathbf{P}_{11}$*, get* $\mathbf{U}^T\mathbf{P}_{11}\mathbf{W} = \mathbf{\Sigma_A}$*.*
6. *The corresponding vectors is first* $c-1$ *vectors of*

$$\mathbf{X} = \mathbf{Q} \begin{pmatrix} \mathbf{R}^{-\mathbf{1}}\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

.

3 Method 3: Regularized Discriminant Analysis

If $\mathbf{S}_t$ is singular, we can add perturbation to $\mathbf{S}_t$, the problem then change to the following form.

$$
\begin{aligned}
&(\mathbf{S}_t + \sigma^2 \mathbf{I}_p)^{-1}\mathbf{S}_b\mathbf{A} = \mathbf{A}\mathbf{\Lambda} \\
&(\mathbf{X}^T\mathbf{H}\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{H}\mathbf{E}\mathbf{\Pi}^{-\frac{1}{2}}\mathbf{\Pi}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{H}\mathbf{X}\mathbf{A} = \mathbf{A}\mathbf{\Lambda}
\end{aligned}
$$

For computation efficiency,If $n \gg p$, let $\mathbf{\Phi} = (\mathbf{X}^T\mathbf{H}\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{H}\mathbf{E}\mathbf{\Pi}^{-\frac{1}{2}}$.
If $p \gg n$, change it to $\mathbf{\Phi} = \mathbf{X}^T\mathbf{H}(\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{E}\mathbf{\Pi}^{-\frac{1}{2}}$. The equation can be rewrite to $\mathbf{\Phi}\mathbf{\Pi}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{H}\mathbf{X}\mathbf{A} = \mathbf{A}\mathbf{\Lambda}$. Let $\mathbf{\Psi} = \mathbf{\Pi}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{H}\mathbf{X}\mathbf{\Phi}$, $\mathbf{B} = \mathbf{\Pi}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{H}\mathbf{X}$.

Proposition 3.1. *The eigenvectors* $\mathbf{A}$ *in* $\mathbf{\Phi}\mathbf{B}\mathbf{A} = \mathbf{A}\mathbf{\Lambda}$ *is* $\mathbf{\Phi}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}{}^{-\frac{1}{2}}$, *and the eigenvalues are* $\mathbf{\Gamma}_{\mathbf{\Psi}}$*. Where* $\mathbf{\Psi} = \mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}\mathbf{V}_{\mathbf{\Psi}}^T$.

Proof:

$$
\begin{aligned}
&\mathbf{\Psi} = \mathbf{B}\mathbf{\Phi} = \mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}\mathbf{V}_{\mathbf{\Psi}}{}^T \\
&\Longrightarrow \mathbf{\Phi}\mathbf{B}\mathbf{\Phi} = \mathbf{\Phi}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}\mathbf{V}_{\mathbf{\Psi}}{}^T \\
&\Longrightarrow \mathbf{\Phi}\mathbf{B}\mathbf{\Phi}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}^{-\frac{1}{2}} = \mathbf{\Phi}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}^{-\frac{1}{2}}\mathbf{\Gamma}_{\mathbf{\Psi}} \\
&\Longrightarrow (\mathbf{\Phi}\mathbf{B})(\mathbf{\Phi}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}^{-\frac{1}{2}}) = (\mathbf{\Phi}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}^{-\frac{1}{2}})\mathbf{\Gamma}_{\mathbf{\Psi}} \\
&\Longrightarrow \mathbf{A} = \mathbf{\Phi}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}^{-\frac{1}{2}}, \mathbf{\Lambda} = \mathbf{\Gamma}_{\mathbf{\Psi}}
\end{aligned}
$$

Example 3.1. *The Step to Solve FDA.*

1. *Compute* $\mathbf{\Phi}$,$\mathbf{\Psi}$.
2. *Perform the SVD of* $\mathbf{\Psi}$*.Get* $\mathbf{\Psi} = \mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}\mathbf{V}_{\mathbf{\Psi}}^T$.
3. $\mathbf{G} = \mathbf{\Phi}\mathbf{V}_{\mathbf{\Psi}}\mathbf{\Gamma}_{\mathbf{\Psi}}^{-\frac{1}{2}}$.

4 RFDA and Rigne Regression

Let $\mathbf{Y} = [\mathbf{y}_i, \mathbf{y}_2, \cdots, \mathbf{y}_n]^T = \mathbf{E}\mathbf{\Pi}^{-1/2}\mathbf{H}_{\boldsymbol{\pi}}$, where $\mathbf{H}_{\boldsymbol{\pi}} = \mathbf{I}_c - \frac{1}{n}\sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}^T$ and $\sqrt{\boldsymbol{\pi}} = (\sqrt{n_1}, \sqrt{n_2}, \cdots, \sqrt{n_c})^T$ be a vector associated with the square root of number of nodes in each cluster.

For each row in $\mathbf{Y}$, $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{ic})$ where

$$
y_{ij} = \begin{cases} \frac{n-n_j}{n\sqrt{n_j}} & \text{if } i \in V_j \\ \frac{\sqrt{n_j}}{n} & \text{if otherwise} \end{cases}
$$

The goal is to minimize the following Lagrangian function:

$$
\begin{aligned}
\min_{\mathbf{w}_0, \mathbf{W}} L(\mathbf{w}_0, \mathbf{W}) &= \frac{1}{2}||\mathbf{Y} - \mathbf{1}_n\mathbf{w}_0^T - \mathbf{X}\mathbf{W}||_F^2 + \frac{\sigma^2}{2}\mathrm{tr}(\mathbf{W}^T\mathbf{W}) \\
&= \frac{1}{2}\sum_{i=1}^{n}||\mathbf{y}_i - \mathbf{w}_0 - \mathbf{W}^T\mathbf{x}_i||^2 + \frac{\sigma^2}{2}\mathrm{tr}(\mathbf{W}^T\mathbf{W})
\end{aligned}
$$

By taking partial derivatives, we have

$$\frac{\partial L}{\mathbf{w}_0} = n\mathbf{w}_0 + \mathbf{W}^T\mathbf{X}^T\mathbf{1}_n - \mathbf{Y}^T\mathbf{1}_n = 0 \tag{2}$$

$$\frac{\partial L}{\mathbf{W}} = (\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{I}_p)\mathbf{W} + \mathbf{X}^T\mathbf{1}_n\mathbf{w}_0^T - \mathbf{X}^T\mathbf{Y} = 0 \tag{3}$$

By solving equation (2), we have

$$\mathbf{w}_0 = -\mathbf{W}^T\mathbf{m}$$

where $\mathbf{m} = \frac{1}{n}\mathbf{X}^T\mathbf{1}_n$. Then in equation (3), we have

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{X} + \sigma^2 + \mathbf{I}_p)\mathbf{W} = \mathbf{X}^T\mathbf{1}_n\mathbf{m}^T\mathbf{W} + \mathbf{X}^T\mathbf{Y} \\
&\Longrightarrow (\mathbf{X}_T\mathbf{X} - n\mathbf{m}\mathbf{m}^T + \sigma^2\mathbf{I}_p)\mathbf{W} = \mathbf{X}^T\mathbf{Y} \\
&\Longrightarrow (\mathbf{X}^T\mathbf{H}\mathbf{X} - \sigma^2\mathbf{I}_p)\mathbf{W} = \mathbf{X}^T\mathbf{Y} \\
&\Longrightarrow \mathbf{W} = (\mathbf{X}^T\mathbf{H}\mathbf{X} - \sigma^2\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}
\end{aligned}$$

Notice that $\mathbf{Y} = \mathbf{E}\mathbf{\Pi}^{-1/2}\mathbf{H}_{\boldsymbol{\pi}} = \mathbf{E}\mathbf{\Pi}^{-1/2}(\mathbf{I}_c - \frac{1}{n}\sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}^T) = (\mathbf{I}_p - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{E}\mathbf{\Pi}^{-1/2}$, we have

$$\mathbf{W} = (\mathbf{X}^T\mathbf{H}\mathbf{X} + \sigma^2\mathbf{I}_p)^{-1}\mathbf{H}\mathbf{E}\mathbf{\Pi}^{1/2} \tag{4}$$

The result in equation (4) is exactly the same as $\mathbf{\Phi}$.