Lecture Note 10: Spectral Clustering and FDA

Professor: Zhihua Zhang *Scribes: Yubo Xie, Yingkai Li*

1 Spectral Clustering

1.1 Spectral Relaxation

We have derived the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Y}} \quad & \mathrm{tr}(\mathbf{Y}^T\mathbf{L}\mathbf{Y}), \\ \text{s.t.} \quad & \mathbf{Y}^T\mathbf{\Pi}\mathbf{Y} = \mathbf{I}_{c-1}, \\ & \mathbf{Y}^T\mathbf{\Pi}\mathbf{1}_n = \mathbf{0}. \end{aligned}$$

Letting $\mathbf{Y}_0 = \mathbf{\Pi}^{1/2}\mathbf{Y}$, we can transform the problem above into the following one:

$$\begin{aligned} \min_{\mathbf{Y}_0} \quad & \mathrm{tr}(\mathbf{Y}_0^T\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}\mathbf{Y}_0), \\ \text{s.t.} \quad & \mathbf{Y}_0^T\mathbf{Y}_0 = \mathbf{I}_{c-1}, \\ & \mathbf{Y}_0^T\mathbf{\Pi}^{1/2}\mathbf{1}_n = \mathbf{0}. \end{aligned}$$

The following theorem will give the solution.

Theorem 1.1. *Suppose that* $\mathbf{L}$ *is a real symmetric matrix such that* $\mathbf{L}\mathbf{1}_n = \mathbf{0}$ *and suppose that the diagonal entries of* $\mathbf{\Pi}$ *are all positive. Let* $\boldsymbol{\mu}_1 = \alpha\mathbf{\Pi}^{1/2}\mathbf{1}_n$ *be the eigenvector associated with the eigenvalue* $\gamma_1 = 0$ *of* $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$*, where* $\alpha^2 = 1/(\mathbf{1}_n^T\mathbf{\Pi}\mathbf{1}_n)$*. Let the remaining eigenvalues of* $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$ *be arranged so that* $\gamma_2 \leq \cdots \leq \gamma_n$*, and let the corresponding orthonormal eigenvectors be denoted by* $\boldsymbol{\mu}_i$*,* $i = 2, \ldots, n$*. Then the solution of the problem is* $\widehat{\mathbf{Y}}_0 = \mathbf{U}\mathbf{Q}$ *where* $\mathbf{U} = [\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_c]$ *and* $\mathbf{Q}$ *is an arbitrary* $(c-1) \times (c-1)$ *orthonormal matrix, with* $\min\{\mathrm{tr}(\mathbf{Y}_0^T\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}\mathbf{Y}_0)\} = \sum_{i=2}^{c}\gamma_i$*. Furthermore, if* $\gamma_c < \gamma_{c+1}$*, then* $\widehat{\mathbf{Y}}_0$ *is a strict local minimum of* $\mathrm{tr}(\mathbf{Y}_0^T\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}\mathbf{Y}_0)$*.*

Proof. Let $\mathbf{S} = \mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$. The Lagrangian is

$$\mathcal{L}(\mathbf{Y}_0, \mathbf{A}, \mathbf{b}) = \mathrm{tr}(\mathbf{Y}_0^T\mathbf{S}\mathbf{Y}_0) - \mathrm{tr}(\mathbf{A}(\mathbf{Y}_0^T\mathbf{Y}_0 - \mathbf{I}_{c-1})) - \mathbf{b}^T\mathbf{Y}_0^T\mathbf{\Pi}^{1/2}\mathbf{1}_n,$$

where $\mathbf{A}$ is a $(c-1) \times (c-1)$ symmetric matrix and $\mathbf{b}$ is a $(c-1) \times 1$ vector.

Taking derivative with respect to $\mathbf{Y}_0$, we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Y}_0} = 2\mathbf{S}\mathbf{Y}_0 - 2\mathbf{Y}_0\mathbf{A} - \mathbf{\Pi}^{1/2}\mathbf{1}_n\mathbf{b}^T.$$

Letting $\partial\mathcal{L}/\partial\mathbf{Y}_0 = \mathbf{0}$ results in

$$2\mathbf{S}\mathbf{Y}_0 - 2\mathbf{Y}_0\mathbf{A} - \mathbf{\Pi}^{1/2}\mathbf{1}_n\mathbf{b}^T = \mathbf{0}. \tag{1}$$

Multiplying Equation (1) on both sides with $\mathbf{1}_n^T\mathbf{\Pi}^{1/2}$, we obtain

$$2\mathbf{1}_n^T\mathbf{\Pi}^{1/2}\mathbf{S}\mathbf{Y}_0 - 2\mathbf{1}_n^T\mathbf{\Pi}^{1/2}\mathbf{Y}_0\mathbf{A} - \mathbf{1}_n^T\mathbf{\Pi}\mathbf{1}_n\mathbf{b}^T = \mathbf{0},$$

which implies $\mathbf{b} = \mathbf{0}$. Therefore

$$\mathbf{S}\mathbf{Y}_0 = \mathbf{Y}_0\mathbf{A}.$$

Now consider the spectral decomposition of $\mathbf{A}$, *i.e.*, $\mathbf{A} = \mathbf{Q}^T\mathbf{\Gamma}_1\mathbf{Q}$, where $\mathbf{Q}$ is a $(c-1)\times(c-1)$ orthonormal matrix and $\mathbf{\Gamma}_1$ is a $(c-1)\times(c-1)$ diagonal matrix. Thus we have

$$\mathbf{S}\mathbf{Y}_0\mathbf{Q}^T = \mathbf{Y}_0\mathbf{Q}^T\mathbf{\Gamma}_1.$$

Note that the diagonal entries of $\mathbf{\Gamma}_1$ are the eigenvalues of $\mathbf{S}$ and the columns of $\mathbf{Y}_0\mathbf{Q}^T$ are the associated eigenvectors. Since $\mathbf{S}\mathbf{\Pi}^{1/2}\mathbf{1}_n = \mathbf{0}$, $\mathbf{\Pi}^{1/2}\mathbf{1}_n$ is the eigenvector of $\mathbf{S}$ associated with eigenvalue 0. Let $\mathbf{\Gamma}_1 = \mathbf{diag}(\gamma_2, \ldots, \gamma_c)$. Thus we have $\widehat{\mathbf{Y}}_0 = [\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_c]\mathbf{Q}$. We can see that $\widehat{\mathbf{Y}}_0$ satisfies the two constraints, *i.e.*, $\widehat{\mathbf{Y}}_0^T\widehat{\mathbf{Y}}_0 = \mathbf{I}_{c-1}$ and $\widehat{\mathbf{Y}}_0^T\mathbf{\Pi}^{1/2}\mathbf{1}_n = \mathbf{0}$.

To verify that $\widehat{\mathbf{Y}}_0$ is actually the solution of the problem, we need to work out the Hessian matrix of $\mathcal{L}$ with respect to $\mathbf{Y}_0$. Before that, it is important to notice

$$\begin{aligned}\mathrm{tr}(\mathbf{ABCD}) &= \mathrm{tr}(\mathbf{DABC}) = \mathrm{tr}(\mathbf{CDAB}) = \mathrm{tr}(\mathbf{BCDA})\\ &= (\mathbf{vec}(\mathbf{D}^T))^T(\mathbf{C}^T\otimes\mathbf{A})\mathbf{vec}(\mathbf{B}).\end{aligned}$$

Using this formula, we have

$$\mathrm{tr}(\mathbf{Y}_0^T\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}\mathbf{Y}_0) = (\mathbf{vec}(\mathbf{Y}_0^T))^T(\mathbf{S}\otimes\mathbf{I}_{c-1})\mathbf{vec}(\mathbf{Y}_0^T),$$

where $\mathbf{vec}(\mathbf{Y}_0^T) = (y_{11}, \ldots, y_{1,c-1}, y_{21}, \ldots, y_{n,c-1})^T$. Then the Hessian matrix is given by

$$\mathbf{H}(\mathbf{Y}_0) = \frac{\partial^2\mathcal{L}}{\partial\,\mathbf{vec}(\mathbf{Y}_0^T)\,\partial\,\mathbf{vec}(\mathbf{Y}_0^T)^T} = \mathbf{S}\otimes\mathbf{I}_{c-1} - \mathbf{I}_n\otimes\mathbf{A}.$$

Let $\mathbf{B}$ be an arbitrary $n\times(c-1)$ matrix such that $\mathbf{B}^T[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c] = \mathbf{0}$. We can always say that $\mathbf{B} = [\boldsymbol{\mu}_{c+1}, \ldots, \boldsymbol{\mu}_n]\mathbf{\Phi}$, where $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{c-1}]$ is an $(n-c)\times(c-1)$ matrix. If we denote $\mathbf{\Gamma}_2 = \mathbf{diag}(\gamma_{c+1}, \ldots, \gamma_n)$, then

$$\begin{aligned}&(\mathbf{vec}(\mathbf{B}^T))^T\mathbf{H}(\widehat{\mathbf{Y}}_0)\mathbf{vec}(\mathbf{B}^T)\\ =\ &(\mathbf{vec}((\mathbf{BQ})^T))^T\mathbf{H}(\widehat{\mathbf{Y}}_0)\mathbf{vec}((\mathbf{BQ})^T)\\ =\ &\mathrm{tr}(\mathbf{Q}^T\mathbf{B}^T\mathbf{SBQ}) - \mathrm{tr}(\mathbf{AQ}^T\mathbf{B}^T\mathbf{BQ})\\ =\ &\mathrm{tr}(\mathbf{B}^T\mathbf{SB}) - \mathrm{tr}(\mathbf{QAQ}^T\mathbf{B}^T\mathbf{B})\\ =\ &\mathrm{tr}(\mathbf{B}^T\mathbf{SB}) - \mathrm{tr}(\mathbf{\Gamma}_1\mathbf{B}^T\mathbf{B})\\ =\ &\mathrm{tr}(\mathbf{\Phi}^T\mathbf{\Gamma}_2\mathbf{\Phi}) - \mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Phi}^T\mathbf{\Phi})\\ =\ &\sum_{i=1}^{c-1}\boldsymbol{\phi}_i^T\mathbf{\Gamma}_2\boldsymbol{\phi}_i - \sum_{i=1}^{c-1}\gamma_{i+1}\boldsymbol{\phi}_i^T\boldsymbol{\phi}_i\\ =\ &\sum_{i=1}^{c-1}\boldsymbol{\phi}_i^T(\mathbf{\Gamma}_2 - \gamma_{i+1}\mathbf{I}_{n-c})\boldsymbol{\phi}_i\\ \geq\ &0.\end{aligned}$$

We can see that if $\gamma_c < \gamma_{c+1}$, then the inequality above is strict, which means $\widehat{\mathbf{Y}}_0$ is a strict local minimizer of the problem. □

1.2 Rounding Schemes

Now consider transforming the real-valued solution into discrete values so that we can obtain the clustering result. We present two algorithms using two rounding schemes.

1.2.1 *K*-means Rounding

The algorithm is shown in Algorithm 1.

Algorithm 1: Spectral Clustering with K-means Rounding

1 **Input**: An affinity matrix $\mathbf{W}$ and a diagonal matrix $\mathbf{\Pi}$
2 **Relaxation**: Obtain $\mathbf{Y} = \mathbf{\Pi}^{-1/2}\mathbf{U}\mathbf{Q} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$
3 **Initialization**: Choose the initial partition $\mathbf{E}$
4 **Rounding**: Repeat until convergence:
  (1) Compute $\mathbf{m}_j = \frac{1}{\sum_{i \in V_j} \pi_i} \sum_{i \in V_j} \pi_i \mathbf{y}_i$
  (2) Find $t_i = \operatorname{argmin}_j \|\mathbf{y}_i - \mathbf{m}_j\|$ and recompute $\mathbf{E}$ by allocating the i-th data point to class t_i
5 **Output**: $\{t_1, \ldots, t_n\}$

1.2.2 Procrustean Rounding

A matrix $\mathbf{Y}$ whose columns are piecewise constant with respect to $\mathbf{E}$ provides a representation of the objective function value Pcut. If we have such a matrix $\mathbf{Y}$, we can find the partition $\mathbf{E}$: Letting $t_i = \operatorname{argmin}_j\{y_{ij}\}$, allocate $\mathbf{x}_i$ to t_i-th class if $y_{i,t_i} > 0$ and to the c-th class otherwise. On the other hand, if we have the partition, we can attempt to find an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Y} = \mathbf{\Pi}^{-1/2}\mathbf{U}\mathbf{Q}$ is as close as possible to the partition. Specifically, we formulate the following Procrustes problem:

$$\min_{\mathbf{Q}} \operatorname{tr}((\mathbf{E}\mathbf{G} - \mathbf{U}\mathbf{Q})(\mathbf{E}\mathbf{G} - \mathbf{U}\mathbf{Q})^T),$$

where $\mathbf{G} = [\mathbf{I}_{c-1} - \frac{1}{c}\mathbf{1}_{c-1}\mathbf{1}_{c-1}^T, -\frac{1}{c}\mathbf{1}_n]^T$. If we let the SVD of $\mathbf{U}^T\mathbf{E}\mathbf{G}$ be $\mathbf{U}^T\mathbf{E}\mathbf{G} = \mathbf{\Theta}\mathbf{\Lambda}\mathbf{V}^T$, then the solution is given by $\mathbf{Q} = \mathbf{\Theta}\mathbf{V}^T$. The algorithm is shown in Algorithm 2.

Algorithm 2: Spectral Clustering with Procrustean Rounding

1 **Input**: An affinity matrix $\mathbf{W}$ and a diagonal matrix $\mathbf{\Pi}$
2 **Relaxation**: Obtain $\mathbf{Y} = \mathbf{\Pi}^{-1/2}\mathbf{U}\mathbf{Q}$
3 **Initialization**: Choose the initial partition $\mathbf{E}$
4 **Rounding**: Repeat until convergence:
  (1) Recompute $\mathbf{E}\mathbf{G}$, perform SVD $\mathbf{U}^T\mathbf{E}\mathbf{G} = \mathbf{\Theta}\mathbf{\Lambda}\mathbf{V}^T$ and let $\mathbf{Q} = \mathbf{\Theta}\mathbf{V}^T$
  (2) Recompute $\mathbf{Y} = \mathbf{\Pi}^{-1/2}\mathbf{U}\mathbf{Q}$, let $t_i = \operatorname{argmin}_j\{y_{ij}\}$ and recompute $\mathbf{E}$ by allocating the i-th data point to class t_i if $\max_j\{y_{ij}\} > 0$ and to class c otherwise
5 **Output**: $\{t_1, \ldots, t_n\}$

1.3 Minimum-variance Criteria

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be the observed data and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$. Then the *with-in class covariance matrix* is given by

$$\mathbf{S}_w = \frac{1}{n} \sum_{j=1}^{c} \sum_{i \in V_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T,$$

where $\mathbf{m}_j = \frac{1}{n_j} \sum_{i \in V_j} \mathbf{x}_i$. Consider the trace of $\mathbf{S}_w$

$$\mathrm{tr}(\mathbf{S}_w) = \frac{1}{n} \sum_{j=1}^{c} \sum_{i \in V_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2.$$

Clustering algorithms based on the minimization of this trace are referred to as minimum-variance methods.

We also have the *between-class covariance matrix*

$$\mathbf{S}_b = \frac{1}{n} \sum_{j=1}^{c} n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T,$$

where $\mathbf{m} = \frac{1}{n}\mathbf{X}^T \mathbf{1}_n$, and the covariance matrix $\mathbf{S}_t = \frac{1}{n}\mathbf{X}^T \mathbf{H} \mathbf{X}$.

Now we are going to express $\mathbf{S}_w$ and $\mathbf{S}_b$ in terms of matrix. Let $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_c]^T = \mathbf{D}^{-1}\mathbf{E}^T\mathbf{X}$. Dropping some constant, we have

$$\begin{aligned}
\mathbf{S}_b &= \begin{bmatrix} \mathbf{m}_1 - \mathbf{m} & \mathbf{m}_2 - \mathbf{m} & \cdots & \mathbf{m}_c - \mathbf{m} \end{bmatrix} \begin{bmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_c \end{bmatrix} \begin{bmatrix} (\mathbf{m}_1 - \mathbf{m})^T \\ (\mathbf{m}_2 - \mathbf{m})^T \\ \vdots \\ (\mathbf{m}_c - \mathbf{m})^T \end{bmatrix} \\
&= \left(\mathbf{M}^T - \frac{1}{n}\mathbf{X}^T \mathbf{1}_n \mathbf{1}_c^T \right) \mathbf{D} \left(\mathbf{M}^T - \frac{1}{n}\mathbf{X}^T \mathbf{1}_n \mathbf{1}_c^T \right)^T.
\end{aligned}$$

Next we attempt to figure out the term $(\mathbf{M}^T - \frac{1}{n}\mathbf{X}^T \mathbf{1}_n \mathbf{1}_c^T)$. Note that

$$\begin{aligned}
\mathbf{M}^T - \frac{1}{n}\mathbf{X}^T \mathbf{1}_n \mathbf{1}_c^T &= \mathbf{X}^T \mathbf{E}\mathbf{D}^{-1} - \mathbf{X}^T \frac{1}{n} \mathbf{1}_n \mathbf{1}_c^T \\
&= \mathbf{X}^T \left(\mathbf{E}\mathbf{D}^{-1} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_c^T \right) \\
&= \mathbf{X}^T \left(\mathbf{E}\mathbf{D}^{-1} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \mathbf{E}\mathbf{D}^{-1} \right) \\
&= \mathbf{X}^T \mathbf{H} \mathbf{E} \mathbf{D}^{-1}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbf{S}_b &= \mathbf{X}^T \mathbf{H} \mathbf{E} \mathbf{D}^{-1} \mathbf{D} \mathbf{D}^{-1} \mathbf{E}^T \mathbf{H} \mathbf{X} \\
&= \mathbf{X}^T \mathbf{H} \mathbf{E} \mathbf{D}^{-1} \mathbf{E}^T \mathbf{H} \mathbf{X}.
\end{aligned}$$

It can be shown that there is some relationship between the three covariance matrices $\mathbf{S}_t$, $\mathbf{S}_w$ and $\mathbf{S}_b$. Specifically, we have

$$\begin{aligned}
\mathbf{S}_t &= \sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \\
&= \sum_{j=1}^{c}\sum_{i\in V_j}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \\
&= \sum_{j=1}^{c}\sum_{i\in V_j}(\mathbf{x}_i - \mathbf{m}_j + \mathbf{m}_j - \mathbf{m})(\mathbf{x}_i - \mathbf{m}_j + \mathbf{m}_j - \mathbf{m})^T \\
&= \sum_{j=1}^{c}\sum_{i\in V_j}(\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T + \sum_{j=1}^{c}\sum_{i\in V_j}(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \\
&= \mathbf{S}_w + \mathbf{S}_b.
\end{aligned}$$

Therefore

$$\mathbf{S}_w = \mathbf{S}_t - \mathbf{S}_b = \mathbf{X}^T\mathbf{H}\mathbf{X} - \mathbf{X}^T\mathbf{H}\mathbf{E}\mathbf{D}^{-1}\mathbf{E}^T\mathbf{H}\mathbf{X}.$$

Now $\min\{\mathrm{tr}(\mathbf{S}_w)\}$ becomes

$$\begin{aligned}
\max\ \mathrm{tr}(\mathbf{S}_b) &= \max\ \mathrm{tr}(\mathbf{X}^T\mathbf{H}\mathbf{E}\mathbf{D}^{-1}\mathbf{E}^T\mathbf{H}\mathbf{X}) \\
&= \max\ \mathrm{tr}(\mathbf{E}^T\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}\mathbf{E}\mathbf{D}^{-1}) \\
&= \max\ \mathrm{tr}(\mathbf{E}^T\mathbf{H}\mathbf{K}\mathbf{H}\mathbf{E}\mathbf{D}^{-1}).
\end{aligned}$$

1.4 Label Switching

Definition 1.1. Suppose there is a set $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$. U and V are two partitions of $\mathcal{A}$ where $U = \{U_1, U_2, \ldots, U_s\}$ and $V = \{V_1, V_2, \ldots, V_r\}$. We use the notation

$$\text{Rand Index} = \frac{\alpha + \beta}{\alpha + \beta + \gamma + \rho} = \frac{\alpha + \beta}{\mathrm{C}_n^2}$$

to represent the difference between partition U and V. And $\alpha, \beta, \gamma, \rho$ are defined as follows:

$$\begin{aligned}
&\alpha = |\mathcal{A}_1|, \text{ where } \mathcal{A}_1 = \{(a_i, a_j) \mid a_i, a_j \in V_k,\ a_i, a_j \in U_l\}, \\
&\beta = |\mathcal{A}_2|, \text{ where } \mathcal{A}_2 = \{(a_i, a_j) \mid a_i \in V_{k_1},\ a_j \in V_{k_2},\ a_i \in U_{l_1},\ a_j \in U_{l_2}\}, \\
&\gamma = |\mathcal{A}_3|, \text{ where } \mathcal{A}_3 = \{(a_i, a_j) \mid a_i, a_j \in V_k,\ a_i \in U_{l_1},\ a_j \in U_{l_2}\}, \\
&\rho = |\mathcal{A}_4|, \text{ where } \mathcal{A}_4 = \{(a_i, a_j) \mid a_i \in V_{k_1},\ a_j \in V_{k_2},\ a_i, a_j \in U_l\}.
\end{aligned}$$

2 Fisher Discriminant Analysis

Let $\mathbf{X} = [x_1, x_2, \ldots, x_n]^T$ be an $n \times p$ matrix and $\mathbf{Y} = \mathbf{X}\mathbf{A}_{p\times q}$ be an $n \times q$ matrix. Since $\mathbf{S}_{xt} = \mathbf{X}^T\mathbf{H}\mathbf{X}$, then $\mathbf{S}_{yt} = \mathbf{Y}^T\mathbf{H}\mathbf{Y} = (\mathbf{X}\mathbf{A})^T\mathbf{H}\mathbf{X}\mathbf{A} = \mathbf{A}^T\mathbf{S}_{xt}\mathbf{A}$. By the same method, we can get $\mathbf{S}_{yw} = \mathbf{A}^T\mathbf{S}_{xw}\mathbf{A}$ and $\mathbf{S}_{yb} = \mathbf{A}^T\mathbf{S}_{xb}\mathbf{A}$. The object is to find matrix $\mathbf{A}$ maximizing $J_1(\mathbf{A}) = \mathrm{tr}(\mathbf{S}_{yw}^{-1}\mathbf{S}_{yb})$, *i.e.*,

$$\max_{\mathbf{A}} J_1 = \mathrm{tr}((\mathbf{A}^T\mathbf{S}_{xw}\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{S}_{xb}\mathbf{A}))$$

By letting the partial derivative equal zero, we can solve the matrix $\mathbf{A}$ satisfying the requirement:

$$\frac{\partial J_1}{\partial \mathbf{A}} = 2\mathbf{S}_{xb}\mathbf{A}(\mathbf{A}^T\mathbf{S}_{xw}\mathbf{A})^{-1} - 2\mathbf{S}_{xw}\mathbf{A}(\mathbf{A}^T\mathbf{S}_{xw}\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{S}_{xb}\mathbf{A})(\mathbf{A}^T\mathbf{S}_{xw}\mathbf{A})^{-1} = \mathbf{0},$$

which leads to

$$\mathbf{S}_{xw}\mathbf{A}(\mathbf{A}^T\mathbf{S}_{xw}\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{S}_{xb}\mathbf{A}) = \mathbf{S}_{xb}\mathbf{A}.$$

Let $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, then $\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T} = \mathbf{I}$. Let $\mathbf{L}^{-1}\mathbf{B}\mathbf{L}^{-T} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ and $\mathbf{Q} = \mathbf{L}\mathbf{U}$. Then,

$$\begin{aligned}\mathbf{B} &= \mathbf{L}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{L}^T = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T,\\ \mathbf{A} &= \mathbf{L}\mathbf{L}^T = \mathbf{L}\mathbf{U}\mathbf{U}^T\mathbf{L}^T = \mathbf{Q}\mathbf{I}\mathbf{Q}^T.\end{aligned}$$

We can transform the equation above in the same form:

$$\begin{aligned}\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}^{-T} &= \mathbf{I},\\ \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}^{-T} &= \mathbf{\Lambda}.\end{aligned}$$

Consider $\mathbf{A}^T\mathbf{S}_{xw}\mathbf{A}$ and $\mathbf{A}^T\mathbf{S}_{xb}\mathbf{A}$ to be the $\mathbf{A}$ and $\mathbf{B}$ respectively. We can get

$$\begin{aligned}\mathbf{S}_{xw}\mathbf{A}(\mathbf{Q}\mathbf{Q}^T)^{-1}\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T &= \mathbf{S}_{xb}\mathbf{A},\\ \mathbf{S}_{xw}\mathbf{A}\mathbf{Q}^{-T}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T &= \mathbf{S}_{xb}\mathbf{A},\\ \mathbf{S}_{xw}\mathbf{A}\mathbf{Q}^{-T}\mathbf{\Lambda} &= \mathbf{S}_{xb}\mathbf{A}\mathbf{Q}^{-T}.\end{aligned}$$

We can consider this as the form of augmented eigenvalue.