

## 1 PageRank

### 1.1 Network Described by Matrix

A network of web pages and links with directed edge from one page to another means that first page contains at least one link to the second. For example,

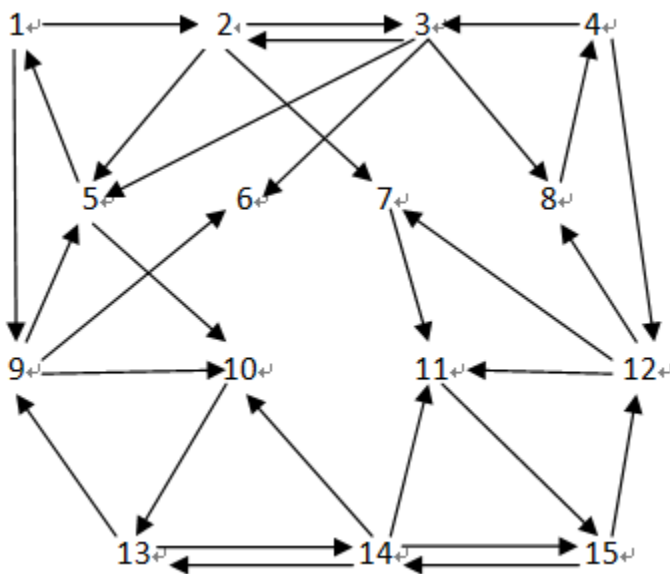


Figure 1: A sample network

The graph can be described as an Adjacency Matrix.

**Definition 1.1** *Adjacency Matrix is a matrix where  $a_{ij} = 1$  iff there is a link from  $A_i$  to  $A_j$  and 0 otherwise.*

### 1.2 Google Matrix

The crawler sits at page  $i$  with probability  $p_i$ . Next, it either moves to a random page (with fixed probability  $q$ ), or with probability  $1 - q$  clicks randomly on a link from the current page  $i$ .

The probability of the crawler moving from page  $i$  to page  $j$  is

$$\frac{q}{n} + \frac{(1 - q)A_{ij}}{n_i}$$

Where  $n_i$  is the sum of the  $i$ th row of  $\mathbf{A}$ .  
Thus,

$$p_j = \sum_{i=1}^n \left( \frac{qp_i}{n} + (1-q) \frac{p_i}{n_i} A_{ij} \right)$$

We define an  $n \times n$  matrix  $\mathbf{G}$  satisfies

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} = \mathbf{G}\mathbf{p} \quad \Rightarrow \quad \mathbf{p} = \mathbf{G}\mathbf{p}$$

**Proposition 1.1**

$$\mathbf{G}^T \mathbf{1} = \mathbf{1} \quad \text{where} \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

**Proof:**

$$\begin{aligned} \sum_{i=1}^n G_{ij} &= \sum_{j=1}^n \left( \frac{q}{n} + (1-q) \frac{A_{ij}}{n_i} \right) \\ &= q + \frac{1-q}{n_i} \sum_{j=1}^n A_{ij} = q + 1 - q = 1 \quad \square \end{aligned}$$

Thus,  $\mathbf{G}$  is a column stochastic matrix and 1 is the eigenvalue. Since  $\mathbf{G}$  is positive,

$$\mathbf{G}\mathbf{x} = \rho(\mathbf{G})\mathbf{x} \quad \Rightarrow \quad \mathbf{G}\mathbf{x} = \mathbf{x}$$

where  $\rho(\mathbf{G})$  is the spectral radius of  $\mathbf{G}$ , *i.e.*,  $\rho(\mathbf{G}) = \{ \max_i |\lambda_i| \}$ .

### 1.3 Properties of PageRank

With  $q = 0.15, p = (0.0268, 0.0299, 0.0299, 0.026, 0.0396, 0.0396, 0.0396, 0.0746, 0.1063, 0.1063, 0.0740, 0.1251, 0.1163, 0.1251)^T$ , Notice that although page 10 and page 11 have the biggest indegree, page 13 and page 15 gain the largest PageRank. It is somewhat reasonable since page 10 and page 11 - the giants have link to them respectively.

## 2 Principal Component Analysis (PCA)

### 2.1 PC Transformation

**Definition 2.1** If  $\mathbf{x} \in \mathbb{R}^p$  is a random vector with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ , then the PC transformation is the transformation

$$\mathbf{x} \rightarrow \mathbf{y} = \boldsymbol{\Gamma}^T (\mathbf{x} - \boldsymbol{\mu})$$

where  $\boldsymbol{\Gamma}$  is orthogonal,  $\boldsymbol{\Gamma}^T \boldsymbol{\Sigma} \boldsymbol{\Gamma} = \boldsymbol{\Lambda}$  is diagonal and  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ .

We have  $\mathbf{y} \in \mathbb{R}^q$  where  $q \leq p$  and  $y_i = \boldsymbol{\gamma}_i^T (\mathbf{x} - \boldsymbol{\mu})$ . We say  $y_i$  is the  $i$ th PC of  $\mathbf{x}$  and  $\boldsymbol{\gamma}_i$  is the  $i$ th column of  $\boldsymbol{\Gamma}$ , *i.e.*, the  $i$ th vector of PC loadings. Also,  $\boldsymbol{\Gamma}^T \boldsymbol{\Sigma} \boldsymbol{\Gamma} = \boldsymbol{\Lambda}$  implies  $\boldsymbol{\Sigma} = \boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Gamma}^T$ .

## 2.2 Sample PCA

Let  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$  which is an  $n \times p$  matrix. Sample variance

$$\begin{aligned} \mathbf{S}_x &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \\ &= \frac{1}{n} (\mathbf{x}_1 - \bar{\mathbf{x}}, \mathbf{x}_2 - \bar{\mathbf{x}}, \dots, \mathbf{x}_n - \bar{\mathbf{x}}) \begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_n - \bar{\mathbf{x}})^T \end{pmatrix} \end{aligned}$$

where

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathbf{X}^T \mathbf{1}$$

Then,

$$\begin{aligned} &\frac{1}{n} [(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) - (\bar{\mathbf{x}}, \bar{\mathbf{x}}, \dots, \bar{\mathbf{x}})] \\ &= \frac{1}{n} (\mathbf{X}^T - \frac{1}{n} \mathbf{X}^T \mathbf{1} \mathbf{1}^T) = \frac{1}{n} \mathbf{X}^T (\mathbf{I}_n - \frac{1}{n} \mathbf{1} \mathbf{1}^T) \\ &= \frac{1}{n} \mathbf{X}^T \mathbf{H} \end{aligned}$$

where  $\mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1} \mathbf{1}^T$  is center matrix, which satisfies  $\mathbf{H} \mathbf{1} = \mathbf{0}$ .

**Proposition 2.1**  $\mathbf{H}$  is idempotent, i.e.,  $\mathbf{H} \mathbf{H} = \mathbf{H}$ .

**Proof:**

$$\begin{aligned} \mathbf{H} \mathbf{H} &= (\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) (\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) \\ &= \mathbf{I}_n - \frac{2}{n} \mathbf{1}_n \mathbf{1}_n^T + \frac{1}{n^2} \mathbf{1}_n (\mathbf{1}_n^T \mathbf{1}_n) \mathbf{1}_n^T = \mathbf{H} \quad \square \end{aligned}$$

Thus,

$$\begin{aligned} \mathbf{S}_x &= (\frac{1}{n} \mathbf{X}^T \mathbf{H}) (\mathbf{X}^T \mathbf{H})^T = (\frac{1}{n} \mathbf{X}^T \mathbf{H}) (\mathbf{H} \mathbf{X}) \\ &= \frac{1}{n} \mathbf{X}_{(p \times n)}^T \mathbf{H} \mathbf{X}_{(n \times p)} = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^T \end{aligned}$$

Since  $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)^T = (\mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}^T) \mathbf{\Gamma} = \mathbf{H} \mathbf{X} \mathbf{\Gamma}$ ,

$$\begin{aligned} \mathbf{S}_Y &= \frac{1}{n} \mathbf{Y}^H \mathbf{H} \mathbf{Y} \\ &= \frac{1}{n} \mathbf{\Gamma}^T \mathbf{X}^T \mathbf{H} \mathbf{H} \mathbf{H} \mathbf{X} \mathbf{\Gamma} \\ &= \frac{1}{n} \mathbf{\Gamma}^T \mathbf{X}^T \mathbf{H} \mathbf{X} \mathbf{\Gamma} \\ &= \mathbf{\Gamma}^T \mathbf{S}_x \mathbf{\Gamma} = \mathbf{\Lambda} \end{aligned}$$

### 3 Principal Coordinate Analysis (PCO)

#### 3.1 Principal Coordinate

In PCA we have  $\mathbf{S} = \mathbf{X}^T \mathbf{H} \mathbf{H} \mathbf{X} = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^T$ , while in PCO we consider  $\mathbf{B} = \mathbf{H} \mathbf{X} \mathbf{X}^T \mathbf{H}$ .

**Definition 3.1** Let  $\mathbf{v}_i$  be the  $i$ th eigenvector of  $\mathbf{B}$ , i.e.,  $\mathbf{B} \mathbf{v}_i = \lambda_i \mathbf{v}_i$ , where  $\mathbf{v}_i$  normalized by  $\mathbf{v}_i^T \mathbf{v}_i = \lambda_i$ . For fixed  $k (1 \leq k \leq p)$ , the rows of  $\mathbf{V}_{(n \times k)} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$  are called the principal coordinates of  $\mathbf{X}$  in  $k$  dimension.

**Note:** Since  $\mathbf{E} \mathbf{F}$  and  $\mathbf{F} \mathbf{E}$  always have the same non-zero eigenvalues,  $\lambda_i$  is also the eigenvalue of  $\mathbf{S}$ .

**Theorem 3.1 (Duality between PCA and PCO)** The principal-coordinates of  $\mathbf{X}$  in  $k$  dimensions are given by the centered scores of the  $n$  objects on the first  $k$  principal components. i.e.,  $\mathbf{V} = \mathbf{Y}_{(PCA)} = \mathbf{H} \mathbf{X} \mathbf{\Gamma}$ .

**Proof:** Use SVD to get  $\mathbf{H} \mathbf{X} = \mathbf{P} \mathbf{D} \mathbf{Q}^T$ . We have

$$\mathbf{B}_{(PCO)} = \mathbf{H} \mathbf{X} \mathbf{X}^T \mathbf{H} = \mathbf{P} \mathbf{D} \mathbf{D} \mathbf{P}^T = \mathbf{P} \mathbf{D}^2 \mathbf{P}^T \tag{1}$$

$$\mathbf{\Sigma}_{(PCA)} = \mathbf{X}^T \mathbf{H} \mathbf{H} \mathbf{X} = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}^T = \mathbf{Q} \mathbf{D}^2 \mathbf{Q}^T \tag{2}$$

From (1),

$$\begin{aligned} \mathbf{B} = \mathbf{H} \mathbf{X} \mathbf{X}^T \mathbf{H} = \mathbf{P} \mathbf{D}^2 \mathbf{P}^T &\Rightarrow \mathbf{B} \mathbf{P} = \mathbf{P} \mathbf{D}^2 \\ &\Rightarrow \mathbf{B} \mathbf{P} \mathbf{D} = \mathbf{P} \mathbf{D}^2 \mathbf{D} \end{aligned}$$

Since  $\mathbf{B} \mathbf{V} = \mathbf{V} \mathbf{\Lambda}$ , we have  $\mathbf{V} = \mathbf{P} \mathbf{D}$  and  $\mathbf{\Lambda} = \mathbf{D}^2 = \mathbf{\Gamma}^T \mathbf{\Sigma} \mathbf{\Gamma}$ .

From (2), we have  $\mathbf{\Gamma} = \mathbf{Q}$ . Thus,

$$\mathbf{Y}_{(PCA)} = \mathbf{H} \mathbf{X} \mathbf{\Gamma} = \mathbf{P} \mathbf{D} \mathbf{Q}^T \mathbf{Q} = \mathbf{P} \mathbf{D} = \mathbf{V} \quad \square$$

#### 3.2 Comparison of PCA and PCO

If  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$  which is an  $n \times p$  matrix, where  $\mathbf{x}_i$  represents a data vector with  $p$  features. It is more efficient to use PCA when  $p < n$ , and PCO otherwise.

### 4 Classical Multidimensional Scaling (MDS)

**Definition 4.1** A distance matrix  $D$  is called Euclidean if there exists a configuration of points in some Euclidean space whose inter-points distances are given by  $D = [d_{rs}^2]$ . Suppose  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbf{R}^p$  such that

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^T (\mathbf{x}_r - \mathbf{x}_s) = \|\mathbf{x}_r - \mathbf{x}_s\|_2^2$$

**Definition 4.2** A  $n$  by  $n$  matrix  $\mathbf{A}$  is said to be metric if

- $a_{ii} = 0$  for all  $i = 1, 2, \dots, n$

- $a_{ij} + a_{ik} \geq a_{jk}$  for all triples  $(i,j,k)$

**Proposition 4.1** *A has following properties*

- $a_{i,j} \geq 0$
- *A is symmetric*

**Corollary 4.1** *A Euclidean Distance matrix is metric, but a metric matrix is not necessarily Euclidean*

**Example 4.1** *Consider 4 points situation,  $P_1, P_2, P_3, P_4$ . Their distance matrix is as follows*

$$\begin{bmatrix} 0 & 2 & 2 & 1.1 \\ 2 & 0 & 2 & 1.1 \\ 2 & 2 & 0 & 1.1 \\ 1.1 & 1.1 & 1.1 & 0 \end{bmatrix}$$

*We can view  $P_1, P_2, P_3$  as vertices of equilateral triangle and  $P_4$  as a point with the same distance to other points. It's easy to show that  $P_4$  satisfying above distance matrix does not exist. However, the distance matrix itself is metric.*

**Theorem 4.1** *Let  $\mathbf{A} = [a_{rs}] = [-\frac{1}{2}d_{rs}^2]$ ,  $\mathbf{D}$  be a distance matrix and define  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$ , then  $\mathbf{D}$  is Euclidean iff  $\mathbf{B}$  is positive semi-definite(p.s.d).*

**Proof:** (a) If  $\mathbf{D}$  is the matrix of Euclidean inter-point distances for a configuration  $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)^T$ , then  $b_{rs} = (\mathbf{z}_r - \bar{\mathbf{z}})^T(\mathbf{z}_s - \bar{\mathbf{z}})$ ,  $s = 1, \dots, n$ , where  $\bar{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i$

$$\begin{aligned} \mathbf{B} &= \mathbf{H}\mathbf{A}\mathbf{H} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{A}(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T) \\ &= \mathbf{A} - \frac{1}{n}\mathbf{A}\mathbf{1}_n\mathbf{1}_n^T - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{A} + \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^T\mathbf{A}\mathbf{1}_n\mathbf{1}_n^T \\ &= \mathbf{A} - \bar{a}_r.\mathbf{1}_n^T - \mathbf{1}_n\bar{a}_{.s} + \bar{a}_{..}\mathbf{1}_n\mathbf{1}_n^T \end{aligned}$$

where  $\bar{a}_r. = \frac{1}{n} \sum_{j=1}^n a_{rj}$ ,  $\bar{a}_{.s} = \frac{1}{n} \sum_{i=1}^n a_{is}$  and  $\bar{a}_{..} = \frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n a_{rs}$

The element-wise representation of  $\mathbf{B}$  is  $b_{rs} = a_{rs} - \bar{a}_r. - \bar{a}_{.s} + \bar{a}_{..}$ . Substitute  $a_{rs}$  and  $d_{rs}$  with  $a_{rs} = -\frac{1}{2}d_{rs}^2$  and  $d_{rs}^2 = (\mathbf{z}_r - \mathbf{z}_s)^T(\mathbf{z}_r - \mathbf{z}_s)$  we get to the conclusion.

(b)If  $\mathbf{B}$  is p.s.d of rank P, then a configuration corresponding to  $\mathbf{B}$  can be constricted as follows:

Let  $\lambda_1 \geq \lambda_2 \geq \dots \lambda_p > 0$  denote the positive eigenvalues of  $\mathbf{B}$ , with corresponding eigenvectors  $\mathbf{X} = (\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(p)})$  normalized by  $\mathbf{x}_i^T \mathbf{x}_i = \lambda_i$ , where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ .

$$\mathbf{B} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T = \mathbf{\Gamma}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Gamma}^T$$

$$\begin{aligned}\mathbf{B}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}} &= \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T.\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}} \\ &= \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\frac{1}{2}} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}\end{aligned}$$

Let  $\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}$ , then  $\mathbf{B} = \mathbf{X}\mathbf{X}^T$ . Since  $b_{rs} = \mathbf{x}_r^T \mathbf{x}_s$ ,  $b_{rs} = a_{rs} - \bar{a}_r - \bar{a}_s + \bar{a}_.$  we have

$$\begin{aligned}(\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s) &= \mathbf{x}_r^T \mathbf{x}_r - 2\mathbf{x}_r^T \mathbf{x}_s + \mathbf{x}_s^T \mathbf{x}_s \\ &= a_{rr} - 2a_{rs} + a_{ss} = -2a_{rs} = d_{rs}^2.\end{aligned}$$

Plus, since  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$ ,  $\mathbf{B}\mathbf{1}_n = \mathbf{H}\mathbf{A}\mathbf{H}\mathbf{1}_n = \mathbf{0}\mathbf{1}_n$  which means that 0 is an eigenvalue of  $\mathbf{B}$  and  $\mathbf{1}_n$  is the corresponding eigenvector. For eigenvectors corresponding to different eigenvalues, they are orthogonal. So we have  $\mathbf{1}_n^T \mathbf{X} = \mathbf{0}$ .

## 5 Fisher Linear Discrimination Analysis

Let  $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \subset \mathbf{R}^p$ ,  $\mathcal{X} = \cup_{i=1}^c \mathcal{X}_i$ ,  $\mathcal{X}_i \neq \emptyset$ ,  $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ . Suppose  $\mathcal{X}$  is classified in  $c$  classes,  $n_i = |\mathcal{X}_i|$ ,  $\bar{\mathcal{X}}_j = \frac{1}{n_j} \sum_{\mathbf{x}_i \in \mathcal{X}_j} \mathbf{x}_i$ , then  $\sum_{i=1}^c n_i = n$  and  $\bar{\mathcal{X}} = \frac{1}{n} \sum_{i=1}^c \mathcal{X}_i = \sum_{j=1}^c \frac{n_j}{n} \bar{\mathcal{X}}_j$

**Definition 5.1** *between-class scatter matrix*  $\mathbf{S}_b = \sum_{j=1}^c \frac{n_j}{n} (\bar{\mathcal{X}}_j - \bar{\mathcal{X}})(\bar{\mathcal{X}}_j - \bar{\mathcal{X}})^T$

**Definition 5.2** *within-class scatter matrix*  $\mathbf{S}_w = \frac{1}{n} \sum_{j=1}^c \sum_{\mathbf{x}_i \in \mathcal{X}_j} (\mathbf{x}_i - \bar{\mathcal{X}}_j)(\mathbf{x}_i - \bar{\mathcal{X}}_j)^T$

From definition, we have

$$\mathbf{S} = \mathbf{S}_w + \mathbf{S}_b = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathcal{X}})(\mathbf{x}_i - \bar{\mathcal{X}})^T$$

A good solution to this classification problem is one with a small within-class difference and a big between-class difference. Thus we can solve the problem by minimizing  $\text{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b)$ .

Besides classification, we also want to reduce the dimension of data, which means that we want to find  $\mathbf{Y} = \mathbf{X}\mathbf{A}$  with a small value of  $\text{tr}(\mathbf{S}_{Yw}^{-1} \mathbf{S}_{Yb})$ , where  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$ . Since

$$\mathbf{S}_{Yw} = \mathbf{A}^T \mathbf{S}_{Xw} \mathbf{A} \text{ and } \mathbf{S}_{Yb} = \mathbf{A}^T \mathbf{S}_{Xb} \mathbf{A}$$

the whole problem can be transformed as the following optimization problem:

$$\underset{\mathbf{A}}{\text{argmax}} [\text{tr}((\mathbf{A}^T \mathbf{S}_{Xw} \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{S}_{Xb} \mathbf{A}))]$$

We have  $f(\mathbf{A}) : \mathbf{R}^{p \times q} \mapsto \mathbf{R}$   $f(\mathbf{A}) = \text{tr}((\mathbf{A}^T \mathbf{S}_{Xw} \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{S}_{Xb} \mathbf{A}))$ . Solving equation  $\frac{\partial f}{\partial \mathbf{A}} = \mathbf{0}$  gives the answer.

**Lemma 5.1** *For matrix  $\mathbf{B}$ , we have  $d\mathbf{B}^{-1} = -\mathbf{B}^{-1}(d\mathbf{B})\mathbf{B}^{-1}$*

**Proof:**  $\mathbf{B}\mathbf{B}^{-1} = \mathbf{I}$ ,  $d\mathbf{B}\mathbf{B}^{-1} + \mathbf{B}d\mathbf{B}^{-1} = \mathbf{0}$   $d\mathbf{B}^{-1} = -\mathbf{B}^{-1}(d\mathbf{B})\mathbf{B}^{-1}$

**Lemma 5.2** *For matrix  $\mathbf{A}$ ,  $\mathbf{B}$ , we have  $\frac{\partial \text{tr}(\mathbf{B}\mathbf{A})}{\partial \mathbf{A}^T} = \mathbf{B}$*

Denote  $\mathbf{S}_{Xb}, \mathbf{S}_{Xw}$  as  $\mathbf{S}_1$  and  $\mathbf{S}_2$ .

$$\begin{aligned} df(\mathbf{A}) &= \text{tr}(d(\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A} + (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} d(\mathbf{A}^T \mathbf{S}_1 \mathbf{A})) \\ &= \text{tr}(d(\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A}) + \text{tr}((\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} d(\mathbf{A}^T \mathbf{S}_1 \mathbf{A})) \end{aligned}$$

$$\begin{aligned} \text{tr}((\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} d(\mathbf{A}^T \mathbf{S}_1 \mathbf{A})) &= \text{tr}(\mathbf{S}_1 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} d\mathbf{A}^T) + \text{tr}((\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_2 d\mathbf{A}) \\ &= 2\text{tr}(\mathbf{S}_1 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} d\mathbf{A}^T) \end{aligned}$$

$$\begin{aligned} \text{tr}(d(\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A}) &= -\text{tr}((\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} d(\mathbf{A}^T \mathbf{S}_2 \mathbf{A}) (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A}) \\ &= -\text{tr}((\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} (d\mathbf{A}^T \mathbf{S}_2 \mathbf{A} + \mathbf{A}^T \mathbf{S}_2 d\mathbf{A}) (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A}) \\ &= -\text{tr}((\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} (d\mathbf{A}^T \mathbf{S}_2 \mathbf{A}) (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A}) - \\ &\quad \text{tr}((\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_2 d\mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A}) \\ &= -2\text{tr}(\mathbf{S}_2 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A} d\mathbf{A}^T) \end{aligned}$$

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{A}} &= \frac{\partial(-2\text{tr}(\mathbf{S}_2 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A} d\mathbf{A}^T) + 2\text{tr}(\mathbf{S}_1 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} d\mathbf{A}^T))}{\partial \mathbf{A}} \\ &= 2\text{tr}(-\mathbf{S}_2 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} + \mathbf{S}_1 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1}) = 0 \end{aligned}$$

$$\mathbf{S}_1 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} = \mathbf{S}_2 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1}$$

$$\begin{aligned} \mathbf{S}_1 \mathbf{A} &= \mathbf{S}_2 \mathbf{A} (\mathbf{A}^T \mathbf{S}_2 \mathbf{A})^{-1} \mathbf{A}^T \mathbf{S}_1 \mathbf{A} \\ &= \mathbf{S}_2 \mathbf{A} (\mathbf{U} \mathbf{U}^T)^{-1} \mathbf{U} \mathbf{D} \mathbf{U}^T \\ &= \mathbf{S}_2 \mathbf{A} \mathbf{U}^{-T} \mathbf{D} \mathbf{U}^T \end{aligned}$$

$$\mathbf{S}_1 \mathbf{A} \mathbf{U}^{-T} = \mathbf{S}_2 \mathbf{A} \mathbf{U}^{-T} \mathbf{D}$$

Let  $\mathbf{A} \mathbf{U}^{-T} = \mathbf{B}$  The problem is transformed into a equivalent generalized eigenvalue problem  $\mathbf{S}_1 \mathbf{B} = \mathbf{S}_2 \mathbf{B} \mathbf{D}$ .