STOR 767, Advanced Machine Learning
Homework 1, Theoretical Part

Zhenghan Fang

May 22, 2019

1. Answer to question 1:

   (i)
   $$Y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{1}$$
   where $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^T \\ \vdots \\ \boldsymbol{X}_n^T \end{bmatrix}$, and $\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$.

   (ii) The least square estimate (LSE) of $\boldsymbol{\beta}$ is
   $$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} (Y - \boldsymbol{X}\boldsymbol{\beta})^T(Y - \boldsymbol{X}\boldsymbol{\beta}) \tag{2}$$
   Set the derivative of $(Y - \boldsymbol{X}\boldsymbol{\beta})^T(Y - \boldsymbol{X}\boldsymbol{\beta})$ w.r.t $\boldsymbol{\beta}$ to zero, we get
   $$\boldsymbol{X}^T(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS}) = 0 \tag{3}$$
   $$\Rightarrow \quad \boldsymbol{X}^T Y - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS} = 0 \tag{4}$$
   $$\Rightarrow \quad \hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y \tag{5}$$
   The LSE of σ^2 is
   $$\hat{\sigma}^2_{LS} = \frac{1}{n}(Y - \hat{Y})^T(Y - \hat{Y}) \tag{6}$$
   $$= \frac{1}{n}(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS})^T(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS}) \tag{7}$$
   $$= \frac{1}{n}(Y - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y)^T(Y - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y) \tag{8}$$

(iii)

$$\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \tag{9}$$

$\mathbf{H}$ represents the linear transform from $\{Y_i\}_{i=1}^n$ to the estimations of $\{Y_i\}_{i=1}^n$ by the linear regression model. $\mathbf{H}$ implies that this transform is determined only by $\{\boldsymbol{X}_i\}_{i=1}^n$.

(iv)

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}_{LS}) &= \mathbb{E}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y\right) && (10)\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}\left(Y\right) && (11)\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} && (12)\\
&= \boldsymbol{\beta} && (13)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2_{LS}) &= \mathbb{E}\left(\frac{1}{n}(Y - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y)^T(Y - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y)\right) && (14)\\
&= \mathbb{E}\left(\frac{1}{n}(Y - \mathbf{H}Y)^T(Y - \mathbf{H}Y)\right) && (15)\\
&= \frac{1}{n}\left((\boldsymbol{I} - \mathbf{H})\boldsymbol{X}\boldsymbol{\beta}\right)^T\left((\boldsymbol{I} - \mathbf{H})\boldsymbol{X}\boldsymbol{\beta}\right) + \frac{\sigma^2}{n}\operatorname{Tr}\left((\boldsymbol{I} - \mathbf{H})(\boldsymbol{I} - \mathbf{H})^T\right) && (16)\\
&= \frac{\sigma^2}{n}\operatorname{Tr}\left(\boldsymbol{I} - \mathbf{H}\right) && (17)\\
&= \sigma^2\left(1 - \frac{1}{n}\operatorname{Tr}\left(\mathbf{H}\right)\right) && (18)
\end{aligned}$$

$$\begin{aligned}
\mathbf{Cov}\left(\hat{\boldsymbol{\beta}}_{LS}\right) &= \mathbf{Cov}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y\right) && (19)\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Cov}\left(Y\right)\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)^T && (20)\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Cov}\left(Y\right)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} && (21)\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Cov}\left(\epsilon\right)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} && (22)\\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} && (23)
\end{aligned}$$

(v)

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML} &= \arg\max_{\boldsymbol{\beta},\sigma^2} \Pr\left(\boldsymbol{X}, Y|\boldsymbol{\beta}, \sigma^2\right) && (24)\\
&= \arg\max_{\boldsymbol{\beta},\sigma^2} \Pr\left(Y|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right) && (25)\\
&= \arg\max_{\boldsymbol{\beta},\sigma^2} \prod_{i=1}^{n} \Pr\left(Y_i|\boldsymbol{X}_i, \boldsymbol{\beta}, \sigma^2\right) && (26)\\
&= \arg\max_{\boldsymbol{\beta},\sigma^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right] && (27)\\
&= \arg\max_{\boldsymbol{\beta},\sigma^2} \sum_{i=1}^{n} \left\{-\frac{1}{2}\log\left(\sigma^2\right) - \frac{(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right\} && (28)\\
&= \arg\max_{\boldsymbol{\beta},\sigma^2} -\frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 && (29)
\end{aligned}
$$

Differentiate the object function w.r.t. $\boldsymbol{\beta}$ and σ^2, and set the derivative to zero, we get

$$
\hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TY \tag{30}
$$

$$
\begin{aligned}
& \frac{n}{2\hat{\sigma}^2_{ML}} - \frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}_{ML})^2}{2(\hat{\sigma}^2_{ML})^2} = 0 && (31)\\
\Rightarrow \quad & \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}_{ML})^2 && (32)\\
\Rightarrow \quad & \hat{\sigma}^2_{ML} = \frac{1}{n}(Y - \mathbf{H}Y)^T(Y - \mathbf{H}Y) && (33)
\end{aligned}
$$

Proof of $\hat{\boldsymbol{\beta}}_{ML} \perp\!\!\!\perp \hat{\sigma}^2_{ML}$:

$$
\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}_{ML}) &= \mathbb{E}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TY) && (34)\\
&= \boldsymbol{\beta} && (35)
\end{aligned}
$$

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{ML}\hat{\sigma}^2_{ML}\right] &= \mathbb{E}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y\hat{\sigma}^2_{ML}\right] &(36)\\
&= \mathbb{E}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}+\epsilon)\hat{\sigma}^2_{ML}\right] &(37)\\
&= \mathbb{E}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\hat{\sigma}^2_{ML} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\epsilon\hat{\sigma}^2_{ML}\right] &(38)\\
&= \mathbb{E}\left[\boldsymbol{\beta}\hat{\sigma}^2_{ML} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\epsilon\hat{\sigma}^2_{ML}\right] &(39)\\
&= \boldsymbol{\beta}\mathbb{E}\left[\hat{\sigma}^2_{ML}\right] + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}\left[\epsilon\hat{\sigma}^2_{ML}\right] &(40)
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}\left[\epsilon\hat{\sigma}^2_{ML}\right] &= \mathbb{E}[\epsilon\boldsymbol{X}\boldsymbol{\beta}^T(\boldsymbol{I}-\mathbf{H})\boldsymbol{X}\boldsymbol{\beta} + \epsilon\epsilon^T(\boldsymbol{I}-\mathbf{H})\boldsymbol{X}\boldsymbol{\beta} &(41)\\
&\quad + \epsilon\boldsymbol{X}\boldsymbol{\beta}^T(\boldsymbol{I}-\mathbf{H})\epsilon + \epsilon\epsilon^T(\boldsymbol{I}-\mathbf{H})\epsilon] &(42)\\
&= 0 + \sigma^2(\boldsymbol{I}-\mathbf{H})\boldsymbol{X}\boldsymbol{\beta} + \sigma^2(\boldsymbol{I}-\mathbf{H})\boldsymbol{X}\boldsymbol{\beta} + 0 &(43)\\
&= 0 &(44)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{ML}\hat{\sigma}^2_{ML}\right] &= \boldsymbol{\beta}\mathbb{E}\left[\hat{\sigma}^2_{ML}\right] &(45)\\
&= \mathbb{E}\left[\hat{\boldsymbol{\beta}}_{ML}\right]\mathbb{E}\left[\hat{\sigma}^2_{ML}\right] &(46)
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}_{ML} \perp\!\!\!\perp \hat{\sigma}^2_{ML}$.

(vi) From the calculation above, we get

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{ML} &= \hat{\boldsymbol{\beta}}_{LS} &(47)\\
\hat{\sigma}^2_{ML} &= \hat{\sigma}^2_{LS} &(48)
\end{aligned}$$

This suggests that the MLE of $\boldsymbol{\beta}$ and σ^2 under the assumption that error ϵ follows Gaussian distribution yields the LSE of $\boldsymbol{\beta}$ and σ^2.

(vii) We denote the new independent sample to be predicted as $(\boldsymbol{X}_m, Y_m)$, because we have defined $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^T \\ \vdots \\ \boldsymbol{X}_n^T \end{bmatrix}$ and $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$.

$$\begin{aligned}
\mathbf{MSE}(\hat{Y}_{LS}) &= \mathbb{E}\left(Y_m - \hat{Y}_{LS}\right)^2 && (49)\\
&= \mathbb{E}\left(\boldsymbol{X}_m^T\boldsymbol{\beta} + \epsilon - \hat{Y}_{LS}\right)^2 && (50)\\
&= \mathbb{E}\left(\boldsymbol{X}_m^T\boldsymbol{\beta} - \hat{Y}_{LS}\right)^2 + \mathbb{E}(\epsilon^2) && (51)\\
&= \left[\mathbb{E}\left(\boldsymbol{X}_m^T\boldsymbol{\beta} - \hat{Y}_{LS}\right)\right]^2 + \mathbf{Cov}(\boldsymbol{X}_m^T\boldsymbol{\beta} - \hat{Y}_{LS}) + \sigma^2 && (52)\\
&= \left[\mathbb{E}\left(\boldsymbol{X}_m^T\boldsymbol{\beta} - \hat{Y}_{LS}\right)\right]^2 + \mathbf{Cov}(\hat{Y}_{LS}) + \sigma^2 && (53)
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{X}_m^T\boldsymbol{\beta} - \hat{Y}_{LS}\right) &= \boldsymbol{X}_m^T\boldsymbol{\beta} - \mathbb{E}\left(\hat{Y}_{LS}\right) && (54)\\
&= \boldsymbol{X}_m^T\boldsymbol{\beta} - \mathbb{E}\left(\boldsymbol{X}_m^T\hat{\boldsymbol{\beta}}_{LS}\right) && (55)\\
&= 0 && (56)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{Cov}(\hat{Y}_{LS}) &= \boldsymbol{X}_m^T\mathbf{Cov}(\hat{\boldsymbol{\beta}}_{LS})\boldsymbol{X}_m && (57)\\
&= \sigma^2\boldsymbol{X}_m^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_m && (58)
\end{aligned}$$

Therefore,

$$\mathbf{MSE}(\hat{Y}_{LS}) = \sigma^2\left(\boldsymbol{X}_m^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_m + 1\right) \tag{59}$$

2. Answer to question 2:

   Let

$$\boldsymbol{X}_1 = \begin{bmatrix} \boldsymbol{X}_{11}^T \\ \vdots \\ \boldsymbol{X}_{n1}^T \end{bmatrix} \tag{60}$$

and split the new independent sample $\boldsymbol{X}_m$ as

$$\boldsymbol{X}_m = \begin{bmatrix} \boldsymbol{X}_{m1} \\ \boldsymbol{X}_{m2} \end{bmatrix} \tag{61}$$

where $\boldsymbol{X}_{m1} \in \mathbb{R}^{p_1}, \boldsymbol{X}_{m2} \in \mathbb{R}^{p_2}$.

(i)

$$\mathbf{MSE}\left(\hat{Y}_{LS}^{(1)}\right) = \sigma^2 \left(\boldsymbol{X}_{m1}^T (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_{m1} + 1\right) \tag{62}$$

$$\mathbf{MSE}\left(\hat{Y}_{LS}^{(1,2)}\right) = \sigma^2 \left(\boldsymbol{X}_m^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_m + 1\right) \tag{63}$$

(ii) For $\mathbf{MSE}\left(\hat{Y}_{LS}^{(1)}\right)$, the bias of $\hat{Y}_{LS}^{(1)}$ is not zero:

$$\mathbb{E}\left(\boldsymbol{X}_m^T \boldsymbol{\beta} - \hat{Y}_{LS}^{(1)}\right) = \boldsymbol{X}_m^T \boldsymbol{\beta} - \mathbb{E}\left(\hat{Y}_{LS}^{(1)}\right) \tag{64}$$

$$= \boldsymbol{X}_m^T \boldsymbol{\beta} - \boldsymbol{X}_{m1}^T \mathbb{E}\left(\hat{\boldsymbol{\beta}}_{LS}^{(1)}\right) \tag{65}$$

where

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}_{LS}^{(1)}\right) = \mathbb{E}\left((\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T Y\right) \tag{66}$$

$$= (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X} \boldsymbol{\beta} \tag{67}$$

Therefore,

$$\mathbf{MSE}\left(\hat{Y}_{LS}^{(1)}\right) = \left(\boldsymbol{X}_m^T \boldsymbol{\beta} - \boldsymbol{X}_{m1}^T (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X} \boldsymbol{\beta}\right)^2 \tag{68}$$

$$+ \sigma^2 \left(\boldsymbol{X}_{m1}^T (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_{m1} + 1\right) \tag{69}$$

Besides,

$$\mathbf{MSE}\left(\hat{Y}_{LS}^{(1,2)}\right) = \sigma^2 \left(\boldsymbol{X}_m^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_m + 1\right) \tag{70}$$

Therefore, the condition under which

$$\mathbf{MSE}\left(\hat{Y}_{LS}^{(1,2)}\right) \leq \mathbf{MSE}\left(\hat{Y}_{LS}^{(1)}\right) \tag{71}$$

is

$$\sigma^2 \boldsymbol{X}_m^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_m \leq \tag{72}$$

$$\left(\boldsymbol{X}_m^T \boldsymbol{\beta} - \boldsymbol{X}_{m1}^T (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X} \boldsymbol{\beta}\right)^2 + \sigma^2 \boldsymbol{X}_{m1}^T (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_{m1} \tag{73}$$

3. Answer to question 3:

   (i) Let

$$\tilde{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \mathbf{b} \tag{74}$$

   Because $\boldsymbol{\Sigma}$ is real symmetric, it can be decomposed as

$$\boldsymbol{\Sigma} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T \tag{75}$$

   where $\mathbf{Q}$ is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of $\boldsymbol{\Sigma}$.
   Let

$$\mathbf{A} = (\mathbf{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}})^T \tag{76}$$

$$\mathbf{b} = -\mathbf{A}\boldsymbol{\mu} \tag{77}$$

   Then

$$\begin{aligned}
cov(\tilde{\boldsymbol{x}}) &= \mathbf{A}cov(\boldsymbol{x})\mathbf{A}^T && (78)\\
&= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T && (79)\\
&= (\mathbf{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}})^T\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T(\mathbf{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}}) && (80)\\
&= \mathbf{I} \qquad (\mathbf{Q}^T\mathbf{Q} = \mathbf{I}) && (81)\\
& && (82)\\
E(\tilde{\boldsymbol{x}}) &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} && (83)\\
&= \mathbf{0} && (84)
\end{aligned}$$

   (ii) a) If $\boldsymbol{y}^T\boldsymbol{x} \neq 0$,
   $\boldsymbol{x}\boldsymbol{y}^T$ has one nonzero eigenvalue

$$\lambda_1 = \boldsymbol{y}^T\boldsymbol{x} \tag{85}$$

   with corresponding eigenvector

$$\boldsymbol{v}_1 = \boldsymbol{x} \tag{86}$$

   and $(n-1)$ zero eigenvalues with corresponding eigenvectors

$$\left\{\boldsymbol{v} \mid \boldsymbol{v}^T\boldsymbol{y} = 0\right\} \tag{87}$$

b) If $\boldsymbol{y}^T\boldsymbol{x} = 0$ and $\boldsymbol{y} \neq \boldsymbol{0}$ and $\boldsymbol{x} \neq \boldsymbol{0}$,
$\boldsymbol{x}\boldsymbol{y}^T$ has $(n-1)$ zero eigenvalues with corresponding eigenvectors

$$\left\{\boldsymbol{v} \mid \boldsymbol{v}^T\boldsymbol{y} = 0\right\} \tag{88}$$

c) If $\boldsymbol{y} = \boldsymbol{0}$ or $\boldsymbol{x} = \boldsymbol{0}$,

$$\boldsymbol{x}\boldsymbol{y}^T = \boldsymbol{0} \tag{89}$$

(iii) Because $\boldsymbol{A}$ is real symmetric,

$$\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2 = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}) \tag{90}$$

and the eigendecomposition of $\boldsymbol{A}$ yields

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T \tag{91}$$

where $\boldsymbol{Q}$ is orthogonal and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are eigenvalues of $\boldsymbol{A}$, i.e. $\lambda_1, ..., \lambda_p$. Then,

$$\begin{align}
\boldsymbol{A}\boldsymbol{A} &= \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T \tag{92}\\
&= \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{\Lambda}\boldsymbol{Q}^T \tag{93}\\
&= \begin{bmatrix}\boldsymbol{q}_1 & ... & \boldsymbol{q}_p\end{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\Lambda} \begin{bmatrix}\boldsymbol{q}_1^T \\ \vdots \\ \boldsymbol{q}_p^T\end{bmatrix} \tag{94}\\
&= \sum_{i=1}^{p} \lambda_i^2 \boldsymbol{q}_i\boldsymbol{q}_i^T \tag{95}
\end{align}$$

where $\boldsymbol{q}_1, ..., \boldsymbol{q}_p$ are the columns of $\boldsymbol{Q}$. Therefore,

$$\begin{align}
\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2 = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}) &= \mathrm{Tr}\left(\sum_{i=1}^{p} \lambda_i^2 \boldsymbol{q}_i\boldsymbol{q}_i^T\right) \tag{96}\\
&= \sum_{i=1}^{p} \lambda_i^2 \,\mathrm{Tr}\left(\boldsymbol{q}_i\boldsymbol{q}_i^T\right) \tag{97}\\
&= \sum_{i=1}^{p} \lambda_i^2 \qquad (\boldsymbol{q}_i\text{'s are orthonormal vectors}) \tag{98}
\end{align}$$