

# Homework I: DATA620013

## Advanced Statistical Learning

黄之豪

20110980005

Oct 25th, 2020

### 1 Problem 1

Assume that  $X$  is not random,  $y = X\beta + \varepsilon$

$$E(d^T y) = E(d^T (X\beta + \varepsilon)) = d^T X\beta$$

As is given unbiased,  $Ed^T y = Ec^T y = c^T \beta$ , so we have  $d^T X = c^T$

$$\begin{aligned} \text{Var}(\hat{\beta}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T \\ &= E((X^T X)^{-1} X^T (X\beta + \varepsilon) - \beta)((X^T X)^{-1} X^T (X\beta + \varepsilon) - \beta)^T \\ &= E((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T \\ &= E((X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}) \\ &= (X^T X)^{-1} \end{aligned}$$

so that,  $\text{Var}(c^T \hat{\beta}) = c^T (X^T X)^{-1} c = d^T X (X^T X)^{-1} X^T d$

As the same,  $\text{Var}(d^T y) = \text{Var}(d^T (X\beta + \varepsilon)) = E(d^T (X\beta + \varepsilon) - d^T X\beta)(d^T (X\beta + \varepsilon) - d^T X\beta)^T = d^T d$

We want to show that,  $\forall d$ , s.t.  $Ed^T y = c^T \beta$

$$\text{Var}(c^T \hat{\beta}) \leq \text{Var}(d^T y) \iff d^T d - d^T X (X^T X)^{-1} X^T d \geq 0 \iff I - X (X^T X)^{-1} X^T \succeq 0$$

To prove that  $I - X (X^T X)^{-1} X^T \succeq 0$ , we set  $I - X (X^T X)^{-1} X^T = M$

$$M^2 = (I - X (X^T X)^{-1} X^T)(I - X (X^T X)^{-1} X^T) = M$$

so that,  $I - X (X^T X)^{-1} X^T = M = M^2 = M^T M$ ,  $I - X (X^T X)^{-1} X^T$  is a positive semidefinite matrix. The proof is completed.

### 2 Problem 2

$$\theta = \{\mu_k, \Sigma\}, \quad k = 1, \dots, K$$

$$p_k(x_i; \theta) = \text{Pr}(G = k | X = x_i; \theta) = \frac{f(x|G = k) \text{Pr}(G = k)}{\text{Pr}(X = x)} \propto f(x|G = k) \text{Pr}(G = k)$$

the likelihood of the data is

$$\mathcal{L}(\theta; x) = \prod_{i=1}^N p_k(x_i; \theta) = \prod_{i=1}^N \frac{f(x|G = k) \text{Pr}(G = k)}{\text{Pr}(X = x)}$$

the log likelihood of the data is

$$\log(\mathcal{L}(\theta; x)) = \sum_{i=1}^N \log(p_{k_i}(x_i; \theta)) \propto \sum_{i=1}^N \log(f_{k_i}(x_i)\pi_{k_i}) = \sum_{i=1}^N \log(f_{k_i}(x_i)) + \log\pi_{k_i}$$

where  $x_i$  means the  $i_{th}$  sample,  $k_i$  means the class of the  $i_{th}$  sample and

$$f_k(x) = \frac{1}{(2\pi)^{\frac{P}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)}$$

(1)  $\hat{\mu}_k$

Suppose that the  $k_{th}$  class has  $n_k$  samples

$$\frac{\partial \log(\mathcal{L}(\theta; x))}{\partial \mu_k} = \frac{\partial \sum_{i=1}^N \log(f_{k_i}(x_i))}{\partial \mu_k} = \frac{\partial \sum_{j=1}^{n_k} \log(f_k(x_j))}{\partial \mu_k} = \sum_{j=1}^{n_k} \frac{\frac{\partial f_k(x_j)}{\partial \mu_k}}{f_k(x_j)} \quad (1)$$

$$\frac{\partial f_k(x_j)}{\partial \mu_k} = f_k(x_j) \left( -\frac{1}{2} (\Sigma^{-1} + (\Sigma^{-1})^T) (\mu_k - x_j) \right) = f_k(x_j) (-\Sigma^{-1}) (\mu_k - x_j)$$

so that (1) equals

$$\sum_{j=1}^{n_k} (-\Sigma^{-1}) (\mu_k - x_j) \quad (2)$$

let (2) equals to 0, we have

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} x_j$$

(2)  $\hat{\Sigma}$

From the matrix derivative formula, we know  $df = \text{tr}(\frac{\partial f^T}{\partial x} dx)$  (\*), and  $d|X| = |X| \text{tr}(X^{-1} dX)$  (\*\*)

$$\arg \max_{\Sigma} \log(\mathcal{L}(\theta; x)) = \arg \max_{\Sigma} \sum_{i=1}^N -\frac{1}{2} \log|\Sigma| - \frac{1}{2} (x_i - \mu_{k_i})^T \Sigma^{-1} (x_i - \mu_{k_i}) = \arg \min_{\Sigma} \sum_{i=1}^N \log|\Sigma| + (x_i - \mu_{k_i})^T \Sigma^{-1} (x_i - \mu_{k_i}) \quad (3)$$

the derivative of first item of (3) is

$$\sum_{i=1}^N d \log|\Sigma| = \sum_{i=1}^N |\Sigma|^{-1} d|\Sigma| \stackrel{(**)}{=} \sum_{i=1}^N \text{tr}(\Sigma^{-1} d\Sigma) = \text{tr}(N * \Sigma^{-1} d\Sigma) \quad (4)$$

As the assumption above, all K classes and  $n_k$  samples about the  $k_{th}$  class, the second item of (3) also equals

$$\sum_{k=1}^K \sum_{j=1}^{n_k} (x_j - \mu_k)^T \Sigma^{-1} (x_j - \mu_k) \quad (5)$$

the derivative of second item of (3) is

$$\begin{aligned}
\sum_{k=1}^K \sum_{j=1}^{n_k} (x_j - \mu_k)^T d\Sigma^{-1}(x_j - \mu_k) &= - \sum_{k=1}^K \sum_{j=1}^{n_k} (x_j - \mu_k)^T \Sigma^{-1} d\Sigma \Sigma^{-1} (x_j - \mu_k) \\
&= tr(- \sum_{k=1}^K \sum_{j=1}^{n_k} (x_j - \mu_k)^T \Sigma^{-1} d\Sigma \Sigma^{-1} (x_j - \mu_k)) \\
&= - \sum_{k=1}^K \sum_{j=1}^{n_k} tr((x_j - \mu_k)^T \Sigma^{-1} d\Sigma \Sigma^{-1} (x_j - \mu_k)) \\
&= - \sum_{k=1}^K \sum_{j=1}^{n_k} tr(\Sigma^{-1} (x_j - \mu_k) (x_j - \mu_k)^T \Sigma^{-1} d\Sigma) \\
&= - \sum_{k=1}^K tr(\Sigma^{-1} \sum_{j=1}^{n_k} (x_j - \mu_k) (x_j - \mu_k)^T \Sigma^{-1} d\Sigma)
\end{aligned} \tag{6}$$

let  $S_k = \sum_{j=1}^{n_k} (x_j - \mu_k) (x_j - \mu_k)^T$ , the (6) equals

$$- \sum_{k=1}^K tr(\Sigma^{-1} S_k \Sigma^{-1} d\Sigma) \tag{7}$$

from (3), (4) and (7), we have

$$d\log(\mathcal{L}) = tr(N * \Sigma^{-1} d\Sigma - \Sigma^{-1} S_k \Sigma^{-1} d\Sigma) = tr((N * \Sigma^{-1} - \Sigma^{-1} \sum_{k=1}^K S_k \Sigma^{-1}) d\Sigma)$$

moreover, from (\*), we have

$$\frac{\partial \log(\mathcal{L}(\theta; x))}{\partial \Sigma} = N * \Sigma^{-1} - \Sigma^{-1} \sum_{k=1}^K S_k \Sigma^{-1} \tag{8}$$

let (8) equals 0, we have

$$\begin{aligned}
\hat{\Sigma} &= \frac{1}{N} \sum_{k=1}^K S_k = \sum_{k=1}^K \frac{n_k - 1}{N} \left( \frac{1}{n_k - 1} S_k \right) \\
S_k &= \sum_{j=1}^{n_k} (x_j - \hat{\mu}_k) (x_j - \hat{\mu}_k)^T
\end{aligned}$$

the  $\frac{1}{n_k - 1} S_k$  is the covariance matrix of the  $k_{th}$  class, so the weight of the pooled covariance estimate is  $\frac{n_k - 1}{N}$

### 3 Problem 3

First solve the first principle component  $y_1 = \alpha_1^T \mathbf{x}$ , as  $var(y_1) = \alpha_1^T \Sigma \alpha_1$ , that is to solve the optimization problem as follow,

$$\begin{aligned}
\max_{\alpha_1} \quad & \alpha_1^T \Sigma \alpha_1 \\
\text{s.t.} \quad & \alpha_1^T \alpha_1 = 1
\end{aligned}$$

To solve the optimization problem, a Lagrange function should be defined as follow

$$\mathcal{L}(\alpha_1, \lambda) = \alpha_1^T \Sigma \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1)$$

$$\mathcal{L}'_{\alpha_1} = \Sigma \alpha_1 - \lambda \alpha_1 = 0$$

so that  $\lambda$  is the eigenvalue of  $\Sigma$ , and  $\alpha_1$  is the corresponding eigenvector. The objective function becomes

$$\alpha_1^T \Sigma \alpha_1 = \alpha_1^T \lambda \alpha_1 = \lambda \alpha_1^T \alpha_1 = \lambda$$

If we want to maximize the objective function, we should let  $\lambda$  to be the max of all eigenvalues of  $\Sigma$ , which is  $\lambda_1$

$$\text{var}(y_1) = \alpha_1^T \Sigma \alpha_1 = \lambda_1$$

Second, solve the second principle component

$$\begin{aligned} \max_{\alpha_2} \quad & \alpha_2^T \Sigma \alpha_2 \\ \text{s.t.} \quad & \alpha_2^T \Sigma \alpha_2 = 1 \\ & \alpha_1^T \Sigma \alpha_2 = 0, \alpha_2^T \Sigma \alpha_1 = 0 \end{aligned}$$

note that  $\alpha_1^T \Sigma \alpha_2 = \alpha_2^T \Sigma \alpha_1 = \alpha_2^T \lambda_1 \alpha_1 = 0$

so that

$$\alpha_2^T \alpha_1 = \alpha_1^T \alpha_2 = 0$$

we have the Lagrange function as

$$\mathcal{L}(\alpha_2, \lambda, \mu) = \alpha_2^T \Sigma \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - \mu(\alpha_2^T \alpha_1)$$

$$\mathcal{L}'_{\alpha_2} = 2\Sigma\alpha_2 - 2\lambda\alpha_2 - \mu\alpha_1 = 0$$

let  $\alpha_1^T$  left multiply the above formula, we have

$$2\alpha_1^T \Sigma \alpha_2 - 2\lambda \alpha_1^T \alpha_2 - \mu \alpha_1^T \alpha_1 = 0$$

so that  $\mu = 0$ , and we have

$$\Sigma \alpha_2 - \lambda \alpha_2 = 0$$

so that we know  $\lambda_2$  is the second largest eigenvalue, and

$$\text{var}(y_2) = \alpha_2^T \Sigma \alpha_2 = \lambda_2$$

Then use Recursion method, we have

$$\text{var}(y_k) = \alpha_k^T \Sigma \alpha_k = \lambda_k$$

where  $\lambda_k$  is the  $k_{th}$  largest eigenvalue.

## 4 Problem 4

$$p(\beta|x) = \frac{p(x|\beta)p(\beta)}{p(x)} \propto p(x|\beta)p(\beta)$$

$$\arg \max_{\beta} p(\beta|x) = \arg \max_{\beta} p(x|\beta)p(\beta)$$

let  $\Pr(G = 1 \mid X = x, \beta) = \pi(x)$ , so  $\Pr(G = 0 \mid X = x, \beta) = 1 - \pi(x)$

$$p(\beta) = \frac{1}{(2\pi)^{\frac{P}{2}} |\alpha^2 I|^{\frac{1}{2}}} e^{-\frac{1}{2} \beta^T (\alpha^2 I)^{-1} \beta}$$

$$\begin{aligned}
\arg \max_{\beta} p(x|\beta)p(\beta) &= \arg \max_{\beta} \prod_{i=1}^n p(x_i|\beta)p(\beta) \\
&= \arg \max_{\beta} \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{(1-y_i)} p(\beta) \\
&= \arg \max_{\beta} \sum_{i=1}^n y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)) + \log p(\beta) \\
&= \arg \max_{\beta} \sum_{i=1}^n y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)) - \frac{1}{2\alpha^2} \beta^T \beta + \log C_0 \\
&= \arg \max_{\beta} \sum_{i=1}^n y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)) - \frac{1}{2\alpha^2} \beta^T \beta \\
&= \arg \min_{\beta} - \sum_{i=1}^n [y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)) - \frac{1}{2\alpha^2} \beta^T \beta]
\end{aligned}$$

so the loss function can be supposed to

$$\begin{aligned}
\mathcal{L}(\beta) &= \frac{n}{2\alpha^2} \beta^T \beta - \sum_{i=1}^n [y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i))] \\
&= \frac{n}{2\alpha^2} \beta^T \beta - \sum_{i=1}^n [y_i \log \frac{\pi(x_i)}{1 - \pi(x_i)} + \log(1 - \pi(x_i))] \\
&= \frac{n}{2\alpha^2} \beta^T \beta - \sum_{i=1}^n [y_i (\beta^T x_i) - \log(1 + e^{\beta^T x_i})]
\end{aligned}$$