Homework I: DATA620013

Advanced Statistical Learning

黄之豪

20110980005

Oct 25th, 2020

1 Problem 1

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

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

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

$$\begin{aligned}
\operatorname{Var}\left(d^T y\right) &= \operatorname{Var}\left(c^T\hat{\beta} + \left(d^T y - c^T\hat{\beta}\right)\right) \\
&= \operatorname{Var}\left(c^T\hat{\beta}\right) + \operatorname{Var}\left(d^T y - c^T\hat{\beta}\right) + 2\operatorname{Cov}\left(c^T\hat{\beta}, d^T y - c^T\hat{\beta}\right)
\end{aligned}$$

Now, we want to show that $\operatorname{Cov}\left(c^T\hat{\beta}, d^T y - c^T\hat{\beta}\right) = 0$

$$\begin{aligned}
\operatorname{Cov}\left(c^T\hat{\beta}, d^T y - c^T\hat{\beta}\right) &= \operatorname{Cov}\left(c^T\left(X^T X\right)^{-1} X^T y, d^T y - c^T\left(X^T X\right)^{-1} X^T y\right) \\
&= \operatorname{Cov}\left(d^T X\left(X^T X\right)^{-1} X^T y, d^T y - d^T X\left(X^T X\right)^{-1} X^T y\right) \\
&= d^T X\left(X^T X\right)^{-1} X^T \operatorname{Cov}(y, y)\left[d^T\left(I - X\left(X^T X\right)^{-1} X^T\right)\right]^T \\
&= d^T X\left(X^T X\right)^{-1} X^T I\left(I - X\left(X^T X\right)^{-1} X^T\right)^T d \\
&= d^T(X\left(X^T X\right)^{-1} X^T - X\left(X^T X\right)^{-1} X^T)d \\
&= 0
\end{aligned}$$

which use $Var(y) = Var(X\beta + \varepsilon) = E(X\beta + \varepsilon - X\beta)(X\beta + \varepsilon - X\beta)^T = I$ and $Cov(Ay, Ay) = ACov(y, y)A^T$ so that, $Var(d^T y) \leq Var(c^T\hat{\beta})$, where the equal sign works when and only when $\operatorname{Var}\left(d^T y - c^T\hat{\beta}\right) = 0$, which means $d^T y = l^T y$ almost everywhere. From the condition of problem 1, $d \neq l$, so

$$Var(d^T y) > Var(l^T y)$$

2 Problem 2

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

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

the likelihood of the data is

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

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

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

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

so that (1) equals

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

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 $\mathrm{d}f = tr(\frac{\partial f^T}{\partial x}\mathrm{d}x)$ (*), and $d|X| = |X|\,\mathrm{tr}\left(X^{-1}dX\right)$ (**)

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

the derivative of first item of (3) is

$$\sum_{i=1}^{N} \mathrm{d}log|\Sigma| = \sum_{i=1}^{N} |\Sigma|^{-1}\mathrm{d}|\Sigma| \overset{(**)}{=} \sum_{i=1}^{N} tr(\Sigma^{-1}\mathrm{d}\Sigma) = tr(N * \Sigma^{-1}\mathrm{d}\Sigma) \tag{4}$$

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

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

the derivative of second item of (3) is

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

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

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

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

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

moreover, from (*), we have

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

let (8) equals 0, we have

$$
\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{K}S_k = \sum_{k=1}^{K}\frac{n_k-1}{N}(\frac{1}{n_k-1}S_k)
$$

$$
S_k = \sum_{j=1}^{n_k}(x_j-\hat{\mu_k})(x_j-\hat{\mu_k})^T
$$

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 λ is the eigenvalue of Σ, and α_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 λ to be the max of all eigenvalues of Σ , which is λ_1

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

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 α_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 λ_2 is the second largest eigenvalue, and

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

Then use Recursion method, we have

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

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

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

$$
\begin{aligned}
\arg\max_{\beta} p(x|\beta)p(\beta) &= \arg\max_{\beta} \prod_{i=1}^{n} p(x_i|\beta)p(\beta) \\
&= \arg\max_{\beta} \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{(1-y_i)}p(\beta) \\
&= \arg\max_{\beta} \sum_{i=1}^{n} y_i log\pi(x_i) + (1-y_i)log(1-\pi(x_i)) + logp(\beta) \\
&= \arg\max_{\beta} \sum_{i=1}^{n} y_i log\pi(x_i) + (1-y_i)log(1-\pi(x_i)) - \frac{1}{2\alpha^2}\beta^T\beta + logC_0 \\
&= \arg\max_{\beta} \sum_{i=1}^{n} y_i log\pi(x_i) + (1-y_i)log(1-\pi(x_i)) - \frac{1}{2\alpha^2}\beta^T\beta \\
&= \arg\min_{\beta} -\sum_{i=1}^{n} [y_i log\pi(x_i) + (1-y_i)log(1-\pi(x_i)) - \frac{1}{2\alpha^2}\beta^T\beta]
\end{aligned}
$$

so the loss function can be supposed to

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