

Midterm 2: Math 6266 (Zhilova)

Peter Williams

12/09/2017

Exercise 1 (The James-Stein estimator)

Let $X \sim N(\theta, \sigma^2 I_p)$ for some $\sigma^2 > 0$, $\theta \in R^p$; dimension ≥ 3 ; θ is an unknown true parameter. Denote the quadratic risk function as $R(\delta, \theta) = E_\theta(|\delta - \theta|)$, where $\delta = \delta(X)$ is some estimator of θ , and $|\cdot|$ is the ℓ_2 -norm in R^p .

1. Calculate the quadratic risk for $\delta = X$

With $R(\theta, \delta) = R(\theta, X) = E[\ell(\theta, X)] = E\|X - \theta\|^2$. We can calculate the quadratic risk:

$$E\|X - \theta\|^2 = E(X - \theta)^\top (X - \theta) = E[X^\top X] - 2\theta^\top E[X] + \theta^\top \theta = E[X^\top X] - \theta^\top \theta = E[X^\top X] - \|\theta\|^2$$

which for $X \sim N(\theta, \sigma^2 I_p)$, reduces to

$$E[X^\top X] - \|\theta\|^2 = \sum_{i=1}^p E[X_i^2] - \|\theta\|^2 = \sum_{i=1}^p (\theta_i^2 + \sigma^2) - \|\theta\|^2 = p\sigma^2 + \|\theta\|^2 - \|\theta\|^2 = p\sigma^2$$

2. Let $\hat{R} = p\sigma^2 + \|h(X)\|^2 - 2\sigma^2 \text{tr}(Dh(X))$, where $h = (h_1, \dots, h_p)^\top : R^p \rightarrow R^p$ is a differentiable function, s.t. all necessary moments exist. $Dh(X)$ is a $p \times p$ matrix of partial derivatives: $\{Dh(x)\}_{i,j} = \frac{\partial}{\partial x_j} h_i(x)$. Show that \hat{R} is an unbiased risk estimator for $\delta(X) = h(X)$, i.e.

$$R(\theta, X - h(X)) = E_\theta \hat{R}$$

Relying on the lecture notes from Jordan (2014) referred to in the midterm problem, we have,

$$R(\theta, X - h(X)) = E_\theta \left[\sum_{i=1}^p ((X_i - \theta_i) - h_i(X))^2 \right] = E_\theta \left[\sum_{i=1}^p (X_i - \theta_i)^2 - 2 \sum_{i=1}^p (X_i - \theta_i) h_i(X) + \sum_{i=1}^p (h_i(X))^2 \right]$$

Using Stein's identity, $E(X - \theta)h(X) = \sigma^2 E[h'(X)]$ we have,

$$\begin{aligned} p\sigma^2 - 2E_\theta \sum_{i=1}^p (X_i - \theta_i) h_i(X) + \|h(X)\|^2 &= p\sigma^2 + \|h(X)\|^2 - 2\sigma^2 E_\theta \left[\sum_{i=1}^p h'_i(X) \right] = \\ p\sigma^2 + \|h(X)\|^2 - 2\sigma^2 \left[\sum_{i=1}^p \frac{\partial h_i(X)}{\partial x_i} \right] &= p\sigma^2 + \|h(X)\|^2 - 2\sigma^2 \text{tr}(Dh(X)) = p\sigma^2 + \|h(X)\|^2 - 2\sigma^2 \text{tr}(Dh(X)) = \hat{R} \end{aligned}$$

3. Consider $h(X) = \frac{(p-2)\sigma^2}{\|X\|^2} X$ and the James-Stein estimator $X - h(X)$. Show that $R(\theta, \hat{\theta}_{JS}) < R(\theta, X)$, for all $\theta \in R^p$.

Noting, $X = (x_1, \dots, x_p)^\top$, we have,

$$\begin{aligned} R(\hat{\theta}_{JS}, \theta) &= E\|\hat{\theta}_{JS} - \theta\|^2 = E\|X - h(X) - \theta\|^2 = E\|(X - \theta) - h(X)\|^2 = E[((X - \theta) - h(X))^\top ((X - \theta) - h(X))] = \\ E[(X - \theta)^\top (X - \theta) - 2(X - \theta)^\top h(X) + (h(X))^\top (h(X))] &= E\|X - \theta\|^2 - 2E[(X - \theta)^\top h(X)] + E\|h(X)\|^2 \end{aligned}$$

which by Stein's Identity reduces to,

$$R(\hat{\theta}_{JS}, \theta) = p\sigma^2 - 2\sigma^2 E(h'(X)) + ((p-2)\sigma^2)^2 E\left\| \frac{X}{\|X\|^2} \right\|^2$$

Focusing in on $h'(X)$, we have

$$\begin{aligned} h'(X) &= \nabla h(X) = \frac{\partial h(X)}{\partial x_1} + \dots + \frac{\partial h(X)}{\partial x_p} = (p-2)\sigma^2 \left[\frac{(X \cdot X) - 2x_1^2}{(X \cdot X)^2} + \dots + \frac{(X \cdot X) - 2x_p^2}{(X \cdot X)^2} \right] = \dots \\ &= (p-2)\sigma^2 \left[\frac{1}{(X \cdot X)^2} \sum_{i=1}^p [(X \cdot X) - 2x_i^2] \right] = (p-2)\sigma^2 \left[\frac{1}{(X \cdot X)^2} [p(X \cdot X) - 2(X \cdot X)] \right] = (p-2)\sigma^2 \left[\frac{(p-2)(X \cdot X)}{(X \cdot X)^2} \right] \\ &\text{which reduces to } h'(X) = \frac{(p-2)^2\sigma^2}{(X \cdot X)}. \text{ So we have } E[h'(X)] = (p-2)^2\sigma^2 E\left[\frac{1}{X \cdot X}\right]. \end{aligned}$$

Returning to the risk function, we have,

$$\begin{aligned} R(\hat{\theta}_{js}, \theta) &= p\sigma^2 - 2\sigma^2 E(h'(X)) + ((p-2)\sigma^2)^2 E\left\| \frac{X}{\|X\|^2} \right\|^2 = p\sigma^2 - 2\sigma^4(p-2)^2 E\left[\frac{1}{X \cdot X}\right] + (p-2)^2\sigma^4 E\left[\frac{1}{X \cdot X}\right] = \\ &= R(\hat{\theta}_{js}, \theta) = p\sigma^2 - \sigma^4(p-2)^2 E\left[\frac{1}{X \cdot X}\right] < p\sigma^2 = R(\theta, X) \end{aligned}$$

4. Now consider an i.i.d. sample Y_1, \dots, Y_n where $Y_i \sim N(\theta, \sigma^2 I_p)$. Denote $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$. Calculate the risk $R(\theta, \bar{Y})$.

With $\theta = (\theta_1, \dots, \theta_n)^\top$, and $\theta_1 = \theta_2 = \dots = \theta_p$, we have,

$$R(\theta, \bar{Y}) = E \sum_{i=1}^p (\bar{Y} - \theta)^2 = pE(\bar{Y} - \theta_1)^2 = p[E(\bar{Y}^2) - \theta_1 E(\bar{Y}) + \theta_1^2] = p(\theta_1^2 + \frac{\sigma^2}{n}) - 2p\theta_1^2 + p\theta_1^2 = p\frac{\sigma^2}{n}$$

5. Consider the estimator $\hat{\theta}_{JS} = \bar{Y} - \frac{(p-2)\sigma^2}{\|\bar{Y}\|^2} \bar{Y}$. Show that $R(\theta, \hat{\theta}_{JS}) < R(\theta, \bar{Y})$ for all $\theta \in R^p$, with $\bar{Y} \sim N(\theta, \frac{\sigma^2}{n} I_p)$ Setting $g(Y) = \frac{(p-2)\sigma^2/n\bar{Y}}{\|\bar{Y}\|^2}$, we have,

$$\begin{aligned} R(\theta, \hat{\theta}_{js}) &= E\|\bar{Y} - g(Y) - \theta\|^2 = E[(\bar{Y} - \theta)^\top (\bar{Y} - \theta) - 2(\bar{Y} - \theta)^\top g(Y) + g(Y)^\top g(Y)] = E\|\bar{Y} - \theta\|^2 - 2E(\bar{Y} - \theta)^\top g(Y) + E\|g(Y)\|^2 = \\ &= p\frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} E(g'(Y)) + E\|g(Y)\|^2 = p\frac{\sigma^2}{n} - 2\left(\frac{\sigma^2}{n}\right)^2 (p-2)^2 E\left(\frac{1}{\|\bar{Y}\|^2}\right) + \left(\frac{\sigma^2}{n}\right)^2 (p-2)^2 E\left(\frac{1}{\|\bar{Y}\|^2}\right) = p\frac{\sigma^2}{n} - \left(\frac{\sigma^2}{n}\right)^2 (p-2)^2 E\left(\frac{1}{\|\bar{Y}\|^2}\right) \end{aligned}$$

using Stein's identity. Thus we have,

$$R(\theta, \hat{\theta}_{js}) = p\frac{\sigma^2}{n} - \left(\frac{\sigma^2}{n}\right)^2 (p-2)^2 E\left(\frac{1}{\|\bar{Y}\|^2}\right) < p\frac{\sigma^2}{n} = R(\theta, \bar{Y})$$

Exercise 2

Consider the linear regression model $Y_i = X_i^\top \theta^* + \varepsilon_i$, $i = 1, \dots, n$, the errors ε_i are i.i.d., $E\varepsilon_i = 0$, $\text{Var}(\varepsilon_i) = \sigma^2 > 0$. The unknown true parameter $\theta^* \in R^p$. Assume that matrix $XX^\top = \sum_{i=1}^n X_i X_i^\top$ is not invertible, i.e. some of its eigenvalues equal to zero.

1. Derive the spectral representation of the model $Y = X^\top \theta^* + \varepsilon$, i.e. show that for some $Z, \xi, \eta^* \in R^p$ the model is equivalent to $Z = \lambda \eta^* + \xi$, where $\lambda = \text{diag}\{\lambda_1, \dots, \lambda_p\}$, and $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ are eigenvalues of XX^\top .

The symmetric matrix XX^\top has spectral decomposition $XX^\top = U^\top \lambda U \rightarrow \lambda = U(XX^\top)U^\top$, with $U^\top U = I_p$. If we take the original model and multiple through by UX , we have spectral representation,

$$(UX)Y = (UX)X^\top(I_p)\theta^* + (UX)\varepsilon = (UX)Y = U(XX^\top)U^\top U\theta^* + (UX)\varepsilon = Z = \lambda \eta^* + \xi$$

with, $Z = (UX)Y$, $\eta^* = U\theta^*$, and $\xi = (UX)\varepsilon$.

2. Let $A = \text{diag}\{\alpha_1, \dots, \alpha_p\}$ for some numbers $\alpha_1, \dots, \alpha_p \in [0, 1]$. Let $\hat{\eta}_A = (\hat{\eta}_{A,1}, \dots, \hat{\eta}_{A,p})^\top$, be a shrinkage estimator of $\hat{\eta}^* = (\eta_1^*, \dots, \eta_p^*)^\top$

$$\hat{\eta}A_{\cdot,j} = \begin{cases} \alpha_j \lambda_j^{-1} z_j, & \text{if } \lambda_j \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Find the bias, variance and the quadratic risk of $\hat{\eta}A : R(\eta^*, \hat{\eta}A) = E(\|\hat{\eta}A - \eta^*\|^2)$

Using the bias-variance decomposition we have:

$$E\|\hat{\eta}A - \eta^*\|^2 = E\|\hat{\eta}A - E(\hat{\eta}A)\|^2 + \|E(\hat{\eta}A) - \eta^*\|^2$$

with, $Var(\hat{\eta}A) = E\|\hat{\eta}A - E(\hat{\eta}A)\|^2$, and $Bias^2(\hat{\eta}A) = \|E(\hat{\eta}A) - \eta^*\|^2$.

Returning to the notation above for individual coefficient estimates, we have for $i = 1, \dots, p$, with $z_j = \lambda_j \hat{\eta}_j$, we have $E(\alpha_j \lambda_j^{-1} z_j) = \alpha_j \lambda_j^{-1} E(z_j) = \alpha_j \lambda_j^{-1} E(\lambda_j \hat{\eta}_j) = \alpha_j \hat{\eta}_j^*$. Using this, for the bias component we have,

$$Bias^2(\hat{\eta}A) = \|E(\hat{\eta}A) - \eta^*\|^2 = \sum_{i=1}^p (E(\hat{\eta}A_{\cdot,i}) - \eta_j^*)^2 = \sum_{i=1}^p (E(\alpha_j \lambda_j^{-1} z_j) - \eta_j^*)^2 = \sum_{i=1}^p (\alpha_j \hat{\eta}_j^* - \eta_j^*)^2 = \sum_{i=1}^p ((\alpha_j - 1) \eta_j^*)^2$$

Thus for the bias we have $\sum_{i=1}^p |((\alpha_j - 1) \eta_j^*)|$.

For the variance component, $Var(\hat{\eta}A) = E\|\hat{\eta}A - E(\hat{\eta}A)\|^2$, using $Var(z_j) = U_j X_j^T Var(Y) X_j U_j^T = \sigma^2 U_j X_j^T X_j U_j^T = \sigma^2 \lambda_j$. We have,

$$\begin{aligned} Var(\hat{\eta}A) &= E\|\hat{\eta}A - E(\hat{\eta}A)\|^2 = E\left[\sum_{i=1}^p (\hat{\eta}A_{\cdot,i} - \alpha_j \hat{\eta}_j^*)^2\right] = E\left[\sum_{i=1}^p (\alpha_j (\lambda_j^{-1} z_j - \eta_j^*))^2\right] = E\sum_{i=1}^p \alpha_j^2 (\lambda_j^{-2} z_j^2 - 2\lambda_j^{-1} z_j \eta_j^* + (\eta_j^*)^2) = \\ &= \sum_{i=1}^p \alpha_j^2 (\lambda_j^{-2} E(z_j^2) - (\eta_j^*)^2) = \sum_{i=1}^p \alpha_j^2 (\lambda_j^{-2} (\lambda_j (\sigma^2 + \lambda_j (\eta_j^*)^2) - (\eta_j^*)^2)) = \sum_{i=1}^p \alpha_j^2 (\lambda_j^{-1} \sigma^2) = Var(\hat{\eta}A) \end{aligned}$$

Thus for the quadratic risk we have,

$$E\|\hat{\eta}A - \eta^*\|^2 = Bias^2(\hat{\eta}A) + Var(\hat{\eta}A) = \sum_{i=1}^p ((\alpha_j - 1) \eta_j^*)^2 + \sum_{i=1}^p \alpha_j^2 (\lambda_j^{-1} \sigma^2)$$

Exercise 3

Let X_1, \dots, X_n be real valued *i.i.d.* random variables. Assume $E(|X_i| |M|) < \infty$ for some $M \geq 2$. Let X_1^*, \dots, X_n^* be a bootstrap sample based on the original data X_1, \dots, X_n and obtained by the Efron's bootstrap procedure, i.e.

$$P(X_j^* = X_i | \{X_i\}_{i=1}^n) = 1/n \quad \forall j = 1, \dots, n$$

1. Show that for all integer $m \in [0, M]$

$$E(X_j^{*m} | \{X_i\}_{i=1}^n) \xrightarrow{P} E(X_1^m) \text{ for } n \rightarrow \infty.$$

(Hint 1: Use the Weak Law of Large Numbers.)

2. Show also that

$$Var(X_j^* | \{X_i\}_{i=1}^n) \xrightarrow{P} Var(X_1) \text{ for } n \rightarrow \infty.$$

(Hint 2: the 1-st bootstrap moment of X_j^* equals to $E(X_j^* | \{X_i\}_{i=1}^n) = \sum_{i=1}^n X_i/n$.)