Midterm 2: Math 6266 (Zhilova)

Peter Williams

12/07/2017

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 $R(\delta, \theta) = E_\theta(||\delta - \theta||^2)$, where $\delta = \delta(X)$ is some estimator of θ, and $||\cdot||^2$ 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:

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

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

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

2. Let $\hat{R} = p\sigma^2 + ||h(X)||^2 - 2\sigma^2\, tr(Dh(X))$, where $h = (h_1, ..., h_p)^\intercal : R^p \to R^p$ 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,

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

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

$$p\sigma^2 - 2E_\theta \sum_{i=1}^p (X_i - \theta_i)h_i(X) + ||h(X)||^2 = p\sigma^2 + ||h(X)||^2 - 2\sigma^2 E_\theta[\sum_{i=1}^p h_i'(X)] =$$

$$p\sigma^2 + ||h(X)||^2 - 2\sigma^2[\sum_{i=1}^p \frac{\partial h_i(X)}{\partial x_i}] = p\sigma^2 + ||h(X)||^2 - 2\sigma^2 tr(Dh(X)) = p\sigma^2 + ||h(X)||^2 - 2\sigma^2 tr(Dh(X)) = \hat{R}$$

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, $X = (x_1, ..., x_p)^\intercal$, we have,

$$R(\hat{\theta}_{js}, \theta) = E||\hat{\theta}_{js} - \theta||^2 = E||X - h(X) - \theta||^2 = E||(X-\theta) - h(X)||^2 = E[((X-\theta) - h(X))^\intercal((X-\theta) - h(X))] =$$

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

which by Stein's Identity reduces to,

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

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

$$h'(X) = \nabla h(X) = \frac{\partial h(X)}{\partial x_1} + ... + \frac{\partial h(X)}{\partial x_p} = (p-2)\sigma^2[\frac{(X \cdot X) - 2x_1^2}{(X \cdot X)^2} + ... + \frac{(X \cdot X) - 2x_p^2}{(X \cdot X)^2}] = \;...$$

$$= (p-2)\sigma^2[\frac{1}{(X \cdot X)^2}\sum_{i=1}^{p}[(X \cdot X) - 2x_i^2] = (p-2)\sigma^2[\frac{1}{(X \cdot X)^2}[p(X \cdot X) - 2(X \cdot X)]] = (p-2)\sigma^2[\frac{(p-2)(X \cdot X)}{(X \cdot X)^2}]$$

which reduces to $h'(X) = \frac{(p-2)^2\sigma^2}{(X \cdot X)}$. So we have $E[h'(X)] = (p-2)^2\sigma^2 E[\frac{1}{X \cdot X}]$.

Returning to the risk function, we have,

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

$$= R(\hat{\theta}_{js}, \theta) = p\sigma^2 - \sigma^4(p-2)^2 E[\frac{1}{X \cdot X}] < p\sigma^2 = R(\theta, X)$$

4. Now consider an i.i.d. sample $Y_1, ..., Y_n$ 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 $\theta = (\theta_1, ..., \theta_n)^\intercal$, and $\theta_1 = \theta_2 = ... = \theta_p$, 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,

$$R(\theta, \hat{\theta}_{js}) = E||\bar{Y} - g(Y) - \theta||^2 = E[(\bar{Y} - \theta)^\intercal(\bar{Y} - \theta) - 2(\bar{Y} - \theta)^\intercal g(Y) + g(Y)^\intercal g(Y)] =$$

$$E||\bar{Y} - \theta||^2 - 2E(\bar{Y} - \theta)^\intercal g(Y) + E||g(Y)||^2 =$$

$$p\frac{\sigma^2}{n} - 2\frac{\sigma^2}{n}E(g'(Y)) + E||g(Y)||^2 = p\frac{\sigma^2}{n} - 2(\frac{\sigma^2}{n})^2(p-2)^2 E(\frac{1}{||\bar{Y}||^2}) + (\frac{\sigma^2}{n})^2(p-2)^2 E(\frac{1}{||\bar{Y}||^2}) =$$

$$p\frac{\sigma^2}{n} - (\frac{\sigma^2}{n})^2(p-2)^2 E(\frac{1}{||\bar{Y}||^2})$$

using Stein's identity. Thus we have,

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

Exercise 2

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

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

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

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

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

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

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

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 $i = 1, ..., p$, with $z_j = \lambda_j\hat{\eta}_j$, we have $E(\alpha_j\lambda_j^{-1}z_j) = \alpha_j\lambda_j^{-1}E(z_j) = \alpha\frac{\lambda_j}{\lambda_j}E(\hat{\eta}_j) = \alpha_j\eta_j^*$. Using this, for the bias component we have,

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

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 $Var(z_j) = U_jX_j^\intercal Var(Y)X_jU_j^\intercal = \sigma^2U_jX_j^\intercal X_jU_j^\intercal = \sigma^2\lambda_j$. We have,

$$Var(\hat{\eta}A) = E||\hat{\eta}A - E(\hat{\eta}A)||^2 = E[\sum_{i=1}^{p}(\hat{\eta}A_{,j} - \alpha_j\eta_j^*)^2] = E[\sum_{i=1}^{p}(\alpha_j(\lambda_j^{-1}z_j - \eta_j^*))^2] = E\sum_{i=1}^{p}\alpha_j^2(\lambda_j^{-2}z_j^2 - 2\lambda_j^{-1}z_j\eta_j^* + (\eta_j^*)^2) =$$

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

.

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

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

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

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

By extension of $P(X_j^* = X_i|\{X_i\}_{i=1}^n) = 1/n \ \ \forall\, j = 1, ..., n$, we have $E(X_j^*|\{X_i\}_{i=1}^n) = E(X_j^*|X_1, X_2, ..., X_n) = 1/n(X_1) + 1/n(X_2) + ... + 1/n(X_n) = n^{-1}\sum_{i=1}^n X_i = \bar{X}$. By the weak law of large numbers, as $n \to \infty$, we have $(X_1 + ... + X_n)/n = nE(X_1)/n = E(X_1)$, since $E(X_1) = ... = E(X_n)$. For the more general case we have,

$$E(X_j^{*m}|\{X_i\}_{i=1}^n) = \sum_{i=1}^n \frac{1}{n} X_i^m, \text{as } n \to \infty, \frac{(X_1^m + X_2^m + ... + X_n^m)}{n} = nE(X_1^m)/n = E(X_1^m)$$

$$\to E(X_j^{*m}|\{X_i\}_{i=1}^n) \xrightarrow{P} E(X_1^m)$$

2. Show also that

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

Noting from above, $E(X_j^*|\{X_i\}_{i=1}^n) = \bar{X}$, and using empirical distribution, we can write,

$$Var(X_j^*|\{X_i\}_{i=1}^n) = E(X_j^* - E(X_j^*|\{X_i\}_{i=1}^n))^2 = \frac{1}{n}\sum_i^n (X_i - E(X_j^*|\{X_i\}_{i=1}^n))^2 = \frac{1}{n}\sum_i^n (X_i - \bar{X})^2$$

By the weak law of large numbers, we have $\bar{X} \xrightarrow{P} E(X_i)$, so we can say,

$$\text{as } n \to \infty, \frac{1}{n}\sum_i^n (X_i - \bar{X})^2 \xrightarrow{P} E(X_i - E(X_i))^2 \to Var(X_j^*|\{X_i\}_{i=1}^n) \xrightarrow{P} Var(X_1)$$