Shestopalov Vlad

1. We have two absolutely identical preliminary standardized regressors $x$ and $x$. The dependent variable $y$ is centered.

In the ridge regression one minimizes the loss function

$$\text{loss}(\hat\beta) = (y-\hat y)^T(y-\hat y) + \lambda\hat\beta^T\hat\beta, \quad \hat y = \hat\beta_1 x + \hat\beta_2 x.$$

(a) Find the optimal $\hat\beta_1$ and $\hat\beta_2$ for fixed $\lambda$.

(b) What happens to the estimates when $\lambda \to \infty$?

(c) What happens to the sum $\hat\beta_1 + \hat\beta_2$ when $\lambda \to 0$?

a) $\min_{\hat\beta} \text{loss}(\hat\beta) = (y-\hat y)^T(y-\hat y) + \lambda\hat\beta^T\hat\beta \quad ; \hat y = X\hat\beta \longrightarrow \underset{[n:2]}{\begin{bmatrix} x_1 & x_1 \\ x_2 & x_2 \\ \vdots & \\ x_n & x_n \end{bmatrix}} \vee \underset{[2:1]}{\begin{bmatrix} \hat\beta_1 \\ \hat\beta_2 \end{bmatrix}}$

$$\frac{\partial \text{loss}(\hat\beta)}{\partial\hat\beta} = 2(y-X\hat\beta)^T d(y-X\hat\beta) + 2\lambda\hat\beta^T d\hat\beta$$

$$= -2(y-X\hat\beta)^T X d\hat\beta + 2\lambda\hat\beta^T d\hat\beta = [-\cancel{2}(y-X\hat\beta)^T X + \cancel{2}\lambda\hat\beta^T]d\hat\beta = 0$$

| apply operator of transposition

$$\Rightarrow \lambda\hat\beta^{T^T} = X^T(y-X\hat\beta) = X^Ty - X^TX\hat\beta \longrightarrow (\lambda - X^TX)\hat\beta = X^Ty$$

$$\hat\beta^* = (\lambda - X^TX)^{-1}X^Ty$$

b) $\lim_{\lambda\to\infty} \hat\beta^* = \lim_{\lambda\to\infty} (\lambda - X^TX)^{-1}X^Ty = 0 \quad \Rightarrow$ multicollinearity would be reduced with high $\lambda$.

c) $\lim_{\lambda\to 0} \hat\beta^* = \lim_{\lambda\to 0} (\lambda - X^TX)^{-1}X^Ty = \beta^{OLS} \quad \Rightarrow$ The sum will produce some constant, which depends on $y$

2. Consider the model $y = X\beta + u$ where $\beta$ is non-random, $\mathbb{E}(u \mid X) = 0$, the matrix $X$ of size $n\times k$ has rank $X = k$, but $\mathbb{V}\text{ar}(u \mid X) = \sigma^2 W$ with $W \neq I$. Let $\hat\beta$ be the standard OLS estimator of $\beta$.

(a) Find $\mathbb{E}(\hat\beta \mid X)$, $\mathbb{E}(\hat\beta)$.

(b) Find $\mathbb{V}\text{ar}(\hat\beta \mid X)$.

(c) How do you think, will the standard confidence interval for $\beta$ be valid in this case?

(d) Find $\mathbb{C}\text{ov}(y, \hat\beta \mid X)$.

a) $E(\hat\beta|X) = E((X^TX)^{-1}X^Ty|X) = (X^TX)^{-1}X^T E(X\beta+u|X) = \underbrace{(X^TX)^{-1}X^TX}_{I}\beta + \underbrace{\dots E(u|X)}_{0} = \beta$

$E(\hat\beta) = E(E(\hat\beta|X)) = \beta$

b) $Var(\hat\beta|X) = Var((X^TX)^{-1}X^Ty|X) = (X^TX)^{-1}X^T Var(y|X)X(X^TX)^{-1} = (X^TX)^{-1}X^T Var(\overset{\text{const}}{X\beta}+u|X)X(X^TX)^{-1} =$

$= (X^TX)^{-1}X^T\sigma^2 W X(X^TX)^{-1} = \sigma^2 (X^TX)^{-1}X^TWX(X^TX)^{-1}$

c) Without a doubt, standard CI is invalid due to heteroscedasticity ($Var(u|X) = \sigma^2 W$, $W \neq I$), errors may be correlated.

d) $Cov(y, \hat\beta|X) = Cov(y, (X^TX)^{-1}X^Ty|X) = Cov(y,y|X)(X^TX)^{-1}X^T = Var(y|X)(X^TX)^{-1}X^T =$

$= \sigma^2 W(X^TX)^{-1}X^T$

3. Consider the matrix

$$X = \begin{pmatrix} 2 & 1 \\ -1 & 2 \\ 1 & 1 \end{pmatrix}.$$

(a) Find the matrix $X^T X$ and diagonalize it.

(b) Find the SVD of $X$.

(c) Find the best approximation to $X$ with rank equal to 1.

Remark: in the principal component analysis the variables in the matrix $X$ should be standardized. If you can't do this by bare hands, feel free to use python, but provide code!

a) $X^T = \begin{pmatrix} 2 & -1 & 1 \\ 1 & 2 & 1 \end{pmatrix}$

$X^T X = \begin{pmatrix} 2 & -1 & 1 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ -1 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 1 \\ 1 & 6 \end{pmatrix} = A$

Diagonalization:

$\det(A - \lambda I) = \begin{vmatrix} 6-\lambda & 1 \\ 1 & 6-\lambda \end{vmatrix} = 35 - 12\lambda + \lambda^2 = 0 \leftrightarrow (\lambda-5)(\lambda-7) = 0$

$\lambda_1 = 5 \,;\ \lambda_2 = 7 \longrightarrow D' = \begin{pmatrix} 7 & 0 \\ 0 & 5 \end{pmatrix} \longrightarrow v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} ;\ v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

b) SVD for X: $X = \underset{\text{orthog}}{U} \cdot \underset{\text{diagonal}}{D} \cdot \underset{\text{orthog}}{V^T}$ s.t. $U^T U = I \quad V^T V = I$

$D' = \begin{pmatrix} 7 & 0 \\ 0 & 5 \\ 0 & 0 \end{pmatrix}$

$X^T X = V D^T U^T U D V^T = V D^T D V T$

$\operatorname{Det}(X^T X - \lambda I) = (\lambda - 7)(\lambda - 5) = 0$

$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \xrightarrow{\text{Normalization}} v_1 = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

$v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

$\rightarrow V = \begin{pmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$

$D = \begin{pmatrix} \sqrt{7} & 0 \\ 0 & \sqrt{5} \\ 0 & 0 \end{pmatrix}$

$XV = U \cdot D \Rightarrow \begin{pmatrix} 2 & 1 \\ -1 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = U \begin{pmatrix} \sqrt{7} & 0 \\ 0 & \sqrt{5} \\ 0 & 0 \end{pmatrix}$

$\begin{pmatrix} -1/\sqrt{3} & 3/\sqrt{2} \\ 3/\sqrt{2} & 1/\sqrt{2} \\ 0 & 2/\sqrt{2} \end{pmatrix} = \begin{pmatrix} -1/\sqrt{14} & 3/\sqrt{10} \\ 3/\sqrt{14} & 1/\sqrt{10} \\ 0 & 2/\sqrt{10} \end{pmatrix} = U$

4. The columns of $X$ are standardized. You know the SVD of the matrix $X = UDV^T$. The diagonal elements of $D$ are positive and ordered from highest to lowest, $d_{11} > d_{22} > \cdots > 0$.

Let's maximize $\|Xw\|^2$ by choosing an optimal vector $w$ subject to $\|w\|^2 = 1$.

(d) Write the Lagrangian function for this problem.

(e) Find the first order conditions. Differential is your friend!

(f) Find the optimal $w$ in terms of columns of $V$.

Hint: one may interpret the FOC in terms of eigenvalues and eigenvectors!

a) $$\begin{cases} \|xw\|^2 \to \max\limits_{w} \\ \text{s.t. } \|w\|^2 = 1 \end{cases}$$

$$\mathcal{L} = \|Xw\|^2 - \lambda[\|w\|^2 - 1] = \|UDV^Tw\|^2 - \lambda[\|w\|^2 - 1]$$

b) FOC: $d((UDV^Tw)(UDV^Tw)^T) - \lambda d(ww^T) = 2(UDV^Tw)^T d(UDV^Tw) - 2\lambda w^T dw =$

$= 2w^T V D^T U^T U D V^T dw - 2\lambda w^T dw = 2w^T dw - 2\lambda w^T dw = 2w^T dw(1-\lambda) = 0$

$\Rightarrow \lambda = 1$, $\sum w_i = 1$ (from the constraint)

c) $W = V^T w \to w = VW$

FOC: $d(UDW)(UDW)^T - \lambda d(VW) = 2(UDW)^T d(UDW) - \lambda V dW =$

$= 2W^T D^T U^T U D dW - \lambda V dW = 2W^T dW - \lambda V dW = 0$

$\to 2W^T = \lambda V \to W^T = \frac{\lambda V}{2}$

$W^T = w^T V = \frac{\lambda V}{2} \to w^T = \frac{\lambda}{2} V V^{-1}$