+32

Midterm exam2, High-dimensional Data Analysis, 2018 Spring [+32 points]

Not only answer but also calculation **Name:** Dong, Yi-Shian

+6 **Q1 [+6]** Consider a model

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2), \quad i = 1,\ldots,n.$$

+3 (1) [+3] Derive the Lasso estimator ($\hat{\boldsymbol{\beta}}_\lambda, \lambda = ?$) as a <u>posterior mode</u>.

(define a prior density and derive a posterior density)

Let $\beta_j \sim$ double exponential (b) $\forall j=1,2,\cdots,p \Rightarrow f(\beta_j) = \frac{1}{2b}\exp\left(\frac{-|\beta_j|}{b}\right) \ \forall j=1,2,\cdots,p$

And because of $Y_i | \beta \sim N(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}, \sigma^2) \ \forall i=1,2,\cdots,n.$

$\Rightarrow f(\beta | Y) \propto f(Y|\beta) f(\beta)$

$\propto \left[\prod_{i=1}^{n} \exp\left(\frac{-1}{2\sigma^2}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\right)\right] \times \left[\prod_{j=1}^{p}\exp\left(\frac{-|\beta_j|}{b}\right)\right]$

$= \exp\left[\frac{-1}{2\sigma^2}(y-X\beta)^T(y-X\beta) - \frac{1}{b}\sum_{j=1}^{p}|\beta_j|\right]$ (where $y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n\end{bmatrix}, \beta = \begin{bmatrix}\beta_0 \\ \vdots \\ \beta_p\end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} \cdots x_{1p} \\ \vdots \\ 1 & x_{n1} \cdots x_{np}\end{bmatrix}$)

$= \exp\left[\frac{-1}{2\sigma^2}\left((y-X\beta)^T(y-X\beta) + \frac{2\sigma^2}{b}\sum_{j=1}^{p}|\beta_j|\right)\right]$

$\Rightarrow \hat\beta_\lambda = \arg\max_\beta f(\beta|Y) = \arg\min_\beta \left[(y-X\beta)^T(y-X\beta) + \frac{2\sigma^2}{b}\sum_{j=1}^{p}|\beta_j|\right]$

$\Rightarrow$ When $\lambda = \frac{2\sigma^2}{b}$, $\hat\beta_\lambda$ are equivalent to the Lasso estimator

+3 (2) [+3] Derive the ridge estimator ($\hat{\boldsymbol{\beta}}_\lambda, \lambda = ?$) as a <u>posterior mode</u>.

(define a prior density and derive a posterior density)

Let $\beta_j \sim N(0, c)$ $\forall j = 1,2,\cdots,p$

$\because Y_i|\beta \sim N(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}, \sigma^2)$

$\therefore f(\beta|Y) \propto f(Y|\beta) f(\beta) \propto \left[\prod_{i=1}^{n}\exp\left(-\frac{(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2}{2\sigma^2}\right)\right] \times \left[\prod_{j=1}^{p}\exp\left(-\frac{\beta_j^2}{2c}\right)\right]$

$= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 - \frac{1}{2c}\sum_{j=1}^{p}\beta_j^2\right)$

$= \exp\left(\frac{-1}{2\sigma^2}(y-X\beta)^T(y-X\beta) - \frac{1}{2c}\sum_{j=1}^{p}\beta_j^2\right)$ (where $y = \begin{bmatrix}Y_1 \\ \vdots \\ Y_n\end{bmatrix}, \beta = \begin{bmatrix}\beta_0 \\ \beta_1 \\ \vdots \\ \beta_p\end{bmatrix}, X = \begin{bmatrix}1 & x_{11} \cdots x_{1p} \\ \vdots \\ 1 & x_{n1} \cdots x_{np}\end{bmatrix}$)

$= \exp\left(\frac{-1}{2\sigma^2}\left[(y-X\beta)^T(y-X\beta) + \frac{\sigma^2}{c}\sum_{j=1}^{p}\beta_j^2\right]\right)$

$\Rightarrow \hat\beta_\lambda = \arg\max_\beta f(\beta|Y) = \arg\max_\beta \exp\left(\frac{-1}{2\sigma^2}\left[(y-X\beta)^T(y-X\beta) + \frac{\sigma^2}{c}\sum_{j=1}^{p}\beta_j^2\right]\right)$

$= \arg\min_\beta \left[(y-X\beta)^T(y-X\beta) + \frac{\sigma^2}{c}\sum_{j=1}^{p}\beta_j^2\right]$

$\Rightarrow$ When $\lambda = \frac{\sigma^2}{c}$ (or $c = \frac{\sigma^2}{\lambda}$), this $\hat\beta_\lambda$ will equivalent to the ridge estimator.

Moreover, $\hat\beta_\lambda$ can be represented as $(X^TX + \lambda I)^{-1}X^Ty$.

+12 **Q2 [+12]** Consider a model <u>without an intercept</u>:

$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$, $i = 1,\dots,n$. Assume $\sum_{i=1}^{n} x_{i1} = \sum_{i=1}^{n} x_{i2} = 0$ and $\sum_{i=1}^{n} x_{i1}^2 = \sum_{i=1}^{n} x_{i2}^2 = 1$.

Answer the questions by using $r_{12} = \sum_{i=1}^{n} x_{i1}x_{i2}$, $r_{1y} = \sum_{i=1}^{n} x_{i1}Y_i$, and $r_{2y} = \sum_{i=1}^{n} x_{i2}Y_i$.

$X = \begin{bmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} \Rightarrow X^TX = \begin{bmatrix} x_{11} \cdots x_{n1} \\ x_{12} \cdots x_{n2} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} = \begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i1}x_{i2} & \sum x_{i2}^2 \end{bmatrix}$

+1 (1) [+1] $X^TX = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}_{2\times 2}$ ✓

$\Rightarrow (X^TX)^{-1} = \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$

+2 (2) [+2] Derive the LSE

$\Rightarrow$ LSE: $(X^TX)^{-1}X^Ty = \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}\begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} = \frac{1}{1-r_{12}^2}\begin{bmatrix} r_{1y} - r_{12}r_{2y} \\ r_{2y} - r_{12}r_{1y} \end{bmatrix}$

$\hat\beta_1 = \frac{r_{1y} - r_{12}r_{2y}}{1-r_{12}^2}$ ✓

$\hat\beta_2 = \frac{r_{2y} - r_{12}r_{1y}}{1-r_{12}^2}$ ✓

Below, we assume $r_{12} = r_{1y} = r_{2y} = 1/2$. $\Rightarrow X^TX = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ and LSE: $\hat\beta = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$.

+2 (3) [+2] Derive the ridge estimators $\hat\beta_{1\lambda}$ and $\hat\beta_{2\lambda}$

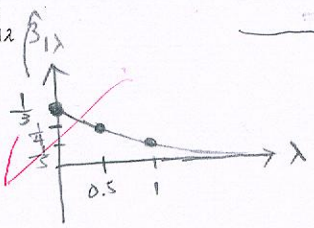
$\hat\beta_\lambda = (X^TX+\lambda I)^{-1}X^Ty = \begin{bmatrix} 1+\lambda & r_{12} \\ r_{12} & 1+\lambda \end{bmatrix}^{-1}\begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} \overset{r_{12}=r_{1y}=r_{2y}=\frac{1}{2}}{=} \begin{bmatrix} 1+\lambda & 0.5 \\ 0.5 & 1+\lambda \end{bmatrix}^{-1}\begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$

$= \frac{4}{(2\lambda+1)(2\lambda+3)}\begin{bmatrix} 1+\lambda & -0.5 \\ -0.5 & 1+\lambda \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \times \frac{1}{2} = \frac{2}{(2\lambda+1)(2\lambda+3)}\begin{bmatrix} \lambda+0.5 \\ \lambda+0.5 \end{bmatrix} = \frac{1}{2\lambda+3}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\Rightarrow \hat\beta_{1\lambda} = \hat\beta_{2\lambda} = \frac{1}{2\lambda+3}$ ✓ when $r_{12} = r_{1y} = r_{2y} = \frac{1}{2}$

+2 (4) [+2] Draw the ridge trace for $\hat\beta_{1\lambda}$

λ	0	0.5	1	$\cdots$	$\to\infty$
$\hat\beta_{1\lambda}$	$\frac{1}{3}$ $= \hat\beta_1^{LSE}$	$\frac{1}{4}$	$\frac{1}{5}$		0



this curve is decreasing and bounded at 0.

+2 (5) [+2] Derive the degrees of freedom df_λ

$df_\lambda = tr[X(X^TX+\lambda I)^{-1}X^T] = tr[X^TX(X^TX+\lambda I)^{-1}] = tr\left[\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \cdot \frac{4}{(2\lambda+1)(2\lambda+3)}\begin{bmatrix} 1+\lambda & -0.5 \\ -0.5 & 1+\lambda \end{bmatrix}\right]$

$= \frac{4}{(2\lambda+1)(2\lambda+3)}\, tr\begin{bmatrix} \lambda+0.75 & 0.5\lambda \\ 0.5\lambda & \lambda+0.75 \end{bmatrix} = \frac{4}{(2\lambda+1)(2\lambda+3)} \times (2\lambda+1.5) = \frac{8\lambda+6}{(2\lambda+1)(2\lambda+3)}$ ✓

+1 (6) [+1] Derive the degrees of freedom df_λ when $\lambda = 0$

$df_{\lambda=0} = \frac{8\lambda+6}{(2\lambda+1)(2\lambda+3)}\Big|_{\lambda=0} = \frac{6}{1\times 3} = 2$ ✓

+1 (7) [+1] Derive the degrees of freedom df_λ when $\lambda \to \infty$

$df_{\lambda\to\infty} = \lim_{\lambda\to\infty}\frac{8\lambda+6}{(2\lambda+1)(2\lambda+3)} = \lim_{\lambda\to\infty}\frac{8\lambda+6}{4\lambda^2+8\lambda+3} = \lim_{\lambda\to\infty}\frac{\frac{8}{\lambda}+\frac{6}{\lambda^2}}{4+\frac{8}{\lambda}+\frac{3}{\lambda^2}} = \frac{0}{4} = 0$ ✓

+1 (8) [+1] Choose λ by setting $df_\lambda = 1$ and $r_{12} = \frac{1}{2}$.

That is, solving that $\frac{8\lambda+6}{4\lambda^2+8\lambda+3} = 1 \Rightarrow 4\lambda^2+8\lambda+3 = 8\lambda+6$

$\Rightarrow 4\lambda^2 = 3 \Rightarrow \lambda = \frac{\sqrt{3}}{2}$ ✓ (negative value is improper) ✓

+14

Q3 [+14] Let $f(x)=\sum_{j=1}^{K}\beta_j M_j(x)$, where $M_j(x)$'s are the spline basis functions. For a knot sequence $\xi_1<\xi_2<\xi_3$ with $\Delta=\xi_2-\xi_1=\xi_3-\xi_2$, let $z_i(t)=(t-\xi_i)/\Delta$ for $i=1, 2,$ and 3.

+2 (1) [+2] Show that $\int f''(x)^2 dx$ can be written as a quadratic form for $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_K)^T$.

$$\int f''(x)^2 dx = \int \left(\sum_{i=1}^{K}\beta_i M_i''(x)\right)\left(\sum_{j=1}^{K}\beta_j M_j''(x)\right)dx = \sum_{i=1}^{K}\sum_{j=1}^{K}\beta_i\beta_j\int M_i''(x)M_j''(x)\,dx$$

$$= [\beta_1 \cdots \beta_K]\begin{bmatrix}\int M_1''(x)M_1''(x)dx & \cdots & \int M_1''(x)M_K''(x)dx\\ \vdots & & \vdots \\ \int M_K''(x)M_1''(x)dx & \cdots & \int M_K''(x)M_K''(x)dx\end{bmatrix}\begin{bmatrix}\beta_1\\ \vdots\\ \beta_K\end{bmatrix} = \beta^T\begin{bmatrix}\int M_1''(x)M_1''(x)dx & \cdots & \int M_1''(x)M_K''(x)dx\\ \vdots & & \vdots \\ \int M_K''(x)M_1''(x)dx & \cdots & \int M_K''(x)M_K''(x)dx\end{bmatrix}\beta$$

+4 (2) [+4] Calculate $\int M_2''(t)^2 dt$,

where $M_2(t)=\frac{\mathrm{I}(\xi_1\le t<\xi_2)}{2\Delta}\{7z_1(t)^3-18z_1(t)^2+12z_1(t)\}-\frac{\mathrm{I}(\xi_2\le t<\xi_3)}{2\Delta}z_3(t)^3$ $\quad \left(\frac{d}{dt}z_i(t)=\frac{1}{\Delta}\right)$

$$\rightarrow M_2'(t) = \frac{I(\xi_1\le t<\xi_2)}{2\Delta^2}\left(21z_1(t)^2-36z_1(t)+12\right) - \frac{3I(\xi_2\le t<\xi_3)}{2\Delta^2}z_3(t)^2$$

$$\rightarrow M_2''(t) = \frac{I(\xi_1\le t<\xi_2)}{2\Delta^3}\left(42z_1(t)-36\right) - 3\frac{I(\xi_2\le t<\xi_3)}{\Delta^3}z_3(t)$$

$$\Rightarrow \int M_2''(t)^2 dt = \int\frac{I(\xi_1\le t<\xi_2)}{4\Delta^6}(42z_1(t)-36)^2dt + 9\int\frac{I(\xi_2\le t<\xi_3)}{\Delta^6}z_3^2(t)\,dt$$

$$- 3\int\frac{I(\xi_1\le t<\xi_2)I(\xi_2\le t<\xi_3)}{\Delta^6}z_3(t)(42z_1(t)-36)\,dt$$

this term equal to zero $\because I(\xi_1\le t<\xi_2)I(\xi_2\le t<\xi_3)=0 \ \forall t$

$$= \frac{1}{4\Delta^6}\int_{\xi_1}^{\xi_2}(42z_1(t)-36)^2dt + \frac{9}{\Delta^6}\int_{\xi_2}^{\xi_3}z_3^2(t)\,dt$$

(Let $x=z_1(t)=\frac{t-\xi_1}{\Delta}$ $\rightarrow dt=\Delta dx$ and let $y=z_3(t)=\frac{t-\xi_3}{\Delta}$ $\rightarrow dt=\Delta dy$)

$$= \frac{1}{4\Delta^6}\int_0^1\Delta(42x-36)^2dx + \frac{9}{\Delta^6}\int_{-1}^{0}\Delta y^2dy$$

$$= \frac{1}{4\Delta^5}\times\frac{(42x-36)^3}{3\times42}\Bigg|_{x=0}^{x=1} + \frac{9}{\Delta^5}\times\frac{y^3}{3}\Bigg|_{y=-1}^{y=0}$$

$$= \frac{1}{4\Delta^5}\times\frac{6^3+36^3}{126} + \frac{9}{\Delta^5}\times\frac{1}{3} = \frac{1}{4\Delta^5}\times372+\frac{3}{\Delta^5}$$

$$= \frac{93}{\Delta^5}+\frac{3}{\Delta^5}=\frac{96}{\Delta^5}$$

+2 (3) [+2] Let Ω be a matrix with elements $\Omega_{jk} = \int M_j''(t) M_k''(t) dt$. Let $\{(Y_i, x_i),\ i = 1, \dots, n\}$ be data.

Let $\hat{f}_\lambda(x) = \sum_{j=1}^{K} \hat{\beta}_{j\lambda} M_j(x)$ be a smoothing spline. Derive the formula of $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{1\lambda}, \cdots, \hat{\beta}_{1\lambda})^T$.

RSS with penalty of roughness : $RSS(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \int f''(x)^2 dx$

$= (y - X\beta)^T (y - X\beta) + \lambda \beta^T \Omega \beta$ where $y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$, $X = \begin{bmatrix} M_1(X_1) & \cdots & M_K(X_1) \\ \vdots & & \vdots \\ M_1(X_n) & \cdots & M_K(X_n) \end{bmatrix}$

$\Rightarrow \hat{\beta}_\lambda = \arg\min_\beta RSS(\beta) = \arg\min_\beta (y - X\beta)^T (y - X\beta) + \lambda \beta^T \Omega \beta$

That equivalent to $\frac{\partial}{\partial \beta} RSS(\beta) \Big|_{\hat{\beta}_\lambda} = -2X^T(y - X\hat{\beta}_\lambda) + 2\lambda \Omega \hat{\beta}_\lambda \overset{set}{=} 0$

$\Rightarrow (X^T X + \lambda \Omega) \hat{\beta}_\lambda = X^T y \quad \Rightarrow \hat{\beta}_\lambda = (X^T X + \lambda \Omega)^{-1} X^T y$

+2 (4) [+2] Define $CV(\lambda)$.

$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{f}_\lambda(x_i)}{1 - h_i} \right)^2$,

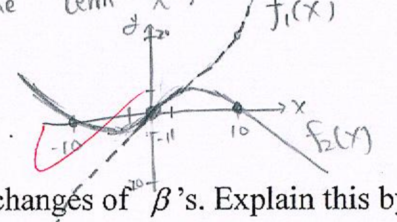
where h_i is the i-th diagonal element of $X(X^T X + \lambda \Omega)^{-1} X^T$

+2 (5) [+2] The polynomial regression is sensitive to changes of β's. Explain this by a figure.

Here compare these to function $\begin{cases} f_1(x) = x + \frac{1}{100} x^3 \\ f_2(x) = x - \frac{1}{100} x^3 \end{cases}$

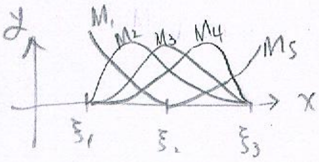
The coefficient $\frac{1}{100}$ can be treated as a small value, but it has large influence of the term x^3.

x	-10	-1	0	1	10
f_1	-20	-1.01	0	1.01	20
f_2	0	-0.99	0	0.99	0



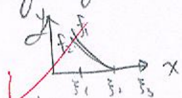
→ The performance of f_1 and f_2 are very different, this show that polynomial regression is sensitive to the changes of β's.

+2 (6) [+2] The cubic spline is not sensitive to changes of β's. Explain this by a figure.

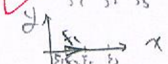


We see the figure marginally,

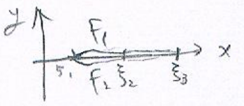
1. $\begin{cases} f_1(x) = 1.01\, M_1(x) \\ f_2(x) = 0.99\, M_1(x) \end{cases}$;



2. $\begin{cases} f_1(x) = 0.01\, M_1(x) \\ f_2(x) = -0.01\, M_2(x) \end{cases}$;



3. $\begin{cases} f_1(x) = 0.01\, M_2(x) \\ f_2(x) = -0.01\, M_2(x) \end{cases}$;



... etc.

we can see the small change of β_j's does not have obviously influence on the different model marginally, and by the continuity of cubic spline, a very small difference of each β's also no influence the result obviously (not sensitive to changes of β's)