Practical Session 1. AMMI, 18/05/2020

Exercise 1

(a) $f(x) = xA \quad \in \mathbb{R}^{m \times n}$

$\nabla f \in \mathbb{R}^{m \times n} \qquad \nabla f_{ij} = f_{ij}'(x) \qquad f_{ij}(x) = f(x)_{ij} = A_{ij}\, x$

$\hookrightarrow \nabla f_{ij} = A_{ij}$ $\qquad \boxed{\nabla f = A}$

(b) $f(x) = \underset{m,n\ n}{A}\, x \in \mathbb{R}^m$

m dimensions, n variables : $\nabla f \in \mathbb{R}^{m \times n}$.

$\nabla f_{ij} = \dfrac{\partial f_i}{\partial x_j} \qquad f_i(x) = A_i^T x = \sum_k A_{ik} x_k.$

$\dfrac{\partial f_i}{\partial x_j} = A_{ij}$ $\qquad \boxed{\nabla f = A}$

(c) $f(x) = \mathrm{Tr}(A^T X) \in \mathbb{R} \qquad X \in \mathbb{R}^{m,n}$

1 dimension, $m \times n$ ~~dimension~~) variables : $\nabla f \in \mathbb{R}^{m \times n}$

$f(x) = \sum A_{ij} X_{ij}$

$\nabla f_{ij} = \dfrac{\partial f}{\partial X_{ij}} = A_{ij}$ $\qquad \boxed{\nabla f = A}$

Exercise 2

$X \in \mathbb{R}^{n \times p}$, $\lambda > 0$ $\qquad M = X^T X + \lambda I_p$

→ Let $M' := X^T X \in \mathbb{R}^{p \times p}$, we will show that M' has non-negative eigenvalues.

Let μ be an eigenvalue of M', and y an eigenvector

$$M'y = \mu y \quad \Rightarrow \quad y^T M' y = \mu y^T y = \mu \|y\|^2$$

$$y^T M' y = y^T X^T X y = (Xy)^T Xy = \|Xy\|^2$$

Therefore $\mu = \dfrac{\|Xy\|^2}{\|y\|^2} \geqslant 0$

→ The eigenvalues of M are equal to λ + the eigenvalues of M'

$$\mu \in Eig(M') \iff \mu + \lambda \in Eig(M) \qquad (1)$$

Therefore the eigenvalues of M are $\geqslant \lambda$ and > 0 $(\neq)$

so M is invertible

(M is invertible $\Leftrightarrow$ its eigenvalues are $\neq 0$)

proof of (1):

- M' is symmetric ($M'^T = M'$) so there exists $U \in \mathbb{R}^{p \times p}$ orthogonal ($U^T U = I$) such that $M' = U^T D U$ $\qquad D = Diag(\lambda_1, \dots, \lambda_n)$

  The λ_i's are the eigenvalues of M'

- $M = M' + \lambda I$ $\qquad$) $U^T U = I$

$$= U^T D U + \lambda U^T U$$

$$= U^T (D + \lambda I) U$$

$$= U^T \begin{pmatrix} \lambda_1 + \lambda & & 0 \\ & \ddots & \\ 0 & & \lambda_n + \lambda \end{pmatrix} U$$

↳ The eigenvalues of M are $\lambda_i + \lambda$

Exercise 3

(a), (b), (c) : seen in class

(d) → pseudo-inverse

→ regularize with small λ : $(X^TX)^{-1}$ becomes $(X^TX + \lambda_n I)^{-1}$ (Ridge Regression) which is invertible (Ex2)

Exercise 4

For $A \in \mathbb{R}^p$, $Var(A) \in \mathbb{R}^{p \times p}$ is the covariance matrix.

$$Var(A) = \mathbb{E}\left((A - \mathbb{E}(A))(A - \mathbb{E}(A))^T \right)$$

$$Var(A)_{ij} = Cov(A_i, A_j)$$
$$= \mathbb{E}(A_i A_j) - \mathbb{E}(A_i)\mathbb{E}(A_j)$$

→ In this exercise, $\mathbb{E}\varepsilon = 0$ and $Var(\varepsilon) = \sigma^2 I$

so $Cov(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$

and $Var(\varepsilon_i) = \sigma^2$

→ Define $P = (X^TX)^{-1}X^T$.

$$\hat{\beta}^{OLS} = PY$$
$$= (X^TX)^{-1}X^TX\beta^* + P\varepsilon \quad \text{(} Y = X\beta^* + \varepsilon \text{ (assumption))}$$
$$= \beta^* + P\varepsilon.$$

$$\mathbb{E}(\hat{\beta}^{OLS}) = \beta^* + \mathbb{E}(P\varepsilon) \quad \text{(} \mathbb{E}(P\varepsilon) = P\mathbb{E}(\varepsilon) = 0 \text{, } \mathbb{E} \text{ is linear)}$$
$$= \beta^*$$

→ $\hat{\beta}^{OLS}$ is unbiased

Exercise 4 (Continued)

→ $\mathrm{Var}(\hat{\beta}^{OLS}) = \mathrm{Var}(\hat{\beta}^{OLS} - \beta^*)$

$= \mathrm{Var}(P\varepsilon)$

$= P\,\mathrm{Var}(\varepsilon)P^T$

$= \sigma^2 I$

$\mathrm{Var}(\hat{\beta}^{OLS}) = \sigma^2 P P^T$

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

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

plugging back $P = (X^TX)^{-1}X^T$

$P^T = X(X^TX)^{-1}$

$((M^{-1})^T = (M^T)^{-1})$

Exercise 5 (slide 17)

Gauss Markov Theorem: Least Squares estimator is Best Linear Unbiased Estimator

Assumptions: $Y = X\beta^* + \varepsilon$ $\begin{cases} E\varepsilon = 0 \\ \mathrm{Var}(\varepsilon) = \sigma^2 I \end{cases}$

Among linear unbiased estimators, $\hat{\beta}^{OLS}$ has the lowest variance

In mathematical terms:

If $\tilde{\beta} = CY$ is an unbiased estimator,

then $\mathrm{Var}(\hat{\beta}^{OLS}) \leq \mathrm{Var}(\tilde{\beta})$

(⚠ this condition does not mean $\mathrm{Var}(\hat{\beta}^{OLS})_{ij} \leq \mathrm{Var}(\tilde{\beta})_{ij}$ $\forall i, j$)

i.e. $M = \mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta}^{OLS})$ is a positive semidefinite matrix

i.e. $\forall x_0,\ x_0^T M x_0 \geq 0$

PROOF: Let $\tilde{\beta} = CY$ be an unbiased estimator, $C \in \mathbb{R}^{p \times n}$

Define $D := C - (X^TX)^{-1}X^T$

We are going to show that $\mathrm{Var}(\hat{\beta}^{OLS}) \leq \mathrm{Var}(\tilde{\beta})$

→ Step 1: show that $DX = 0$

$\tilde{\beta}$ is unbiased : $\mathbb{E}(\tilde{\beta}) = \beta^*$ (1)

$$\begin{aligned}\mathbb{E}(\tilde{\beta}) &= \mathbb{E}(CY) \\ &= C\,\mathbb{E}(Y) \qquad \swarrow\ \mathbb{E} \text{ is linear} \\ &= CX\beta^* + C\underbrace{\mathbb{E}(\varepsilon)}_{=0} \qquad \swarrow\ Y = X\beta^* + \varepsilon \\ &= CX\beta^* \quad (2)\end{aligned}$$

Combining (1) and (2): $\beta^* = CX\beta^*$

so $(CX - I)\beta^* = 0$ $\qquad DX = CX - I$

so $DX\beta^* = 0$. This is true for all β^*

↳ $DX = 0$ □

→ Step 2: show that $\mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(\hat{\beta}^{OLS}) + \sigma^2 DD^T$

→ We reuse the notation $P = (X^TX)^{-1}X^T$

since $DX = 0$, $DP^T = PD^T = 0$

→
$$\begin{aligned}\tilde{\beta} &= BCY \\ &= (D+P)(X\beta^* + \varepsilon) \\ &= \underbrace{DX\beta^*}_{=0} + D\varepsilon + \underbrace{PX}_{=I}\beta^* + P\varepsilon \\ &= \beta^* + (D+P)\varepsilon\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}((D+P)\varepsilon) \\ &= (D+P)\underbrace{\mathrm{Var}(\varepsilon)}_{=\sigma^2 I}(D+P)^T \\ &= \sigma^2\left(DD^T + \underbrace{PD^T}_{=0} + \underbrace{DP^T}_{=0} + \underbrace{PP^T}_{=(X^TX)^{-1}}\right) \\ &= \sigma^2 DD^T + \underbrace{\sigma^2(X^TX)^{-1}}_{=\mathrm{Var}(\hat{\beta}^{OLS})}\end{aligned}$$

□

note: $Y \in \mathbb{R}^p$ random vector, $C \in \mathbb{R}^{p\times p}$ fixed.

Then $\mathbb{E}(CY) = C\,\mathbb{E}(Y)$ and $\mathrm{Var}(CY) = C\,\mathrm{Var}(Y)\,C^T$

→ Step 3 (final) $M = Var(\tilde{\beta}) - Var(\hat{\beta}^{OLS})$ is positive semi-definite (pd)

. $M = Var(\tilde{\beta}) - Var(\hat{\beta}^{OLS})$ ↙ Step 2.

$= \sigma^2 DD^T$

. Let $x_0 \in \mathbb{R}^n$: $x_0^T M x_0 = \sigma^2 x_0^T DD^T x_0$

$= \sigma^2 (D^T x_0)^T (D^T x_0)$

$= \sigma^2 \|D^T x_0\|^2 \geq 0$

This is true $\forall x_0$: M is p.d.

Finally: $\underline{Var(\tilde{\beta}) \geq Var(\hat{\beta}^{OLS})}$

$\tilde{\beta}$ has a "larger" variance than $\hat{\beta}^{OLS}$.

WHY DOES THIS MATTER?

Suppose we observe measurements (what we want to estimate) $y = \mu + \underbrace{\varepsilon}_{\text{noise}}$

n measurements $y_i = \mu + \varepsilon_i$ $\quad E(\varepsilon_i) = 0$, $Var(\varepsilon) = \sigma^2 I$

We want an estimator $\hat{\mu}$ of μ.

Which is the best estimator?:

① $\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} y_i$

② $\hat{\mu}_2 = y_1$

③ $\hat{\mu}_3 = \frac{1}{n} \sum y_i + \frac{1}{n}$.

→ BIAS: . $E(\hat{\mu}_1) = E(\hat{\mu}_2) = \mu$ $\quad \hat{\mu}_1$ and $\hat{\mu}_2$ are unbiased.

. $E(\hat{\mu}_3) = \mu + \frac{1}{n} \neq \mu$ $\quad \hat{\mu}_3$ is biased

→ VARIANCE: . $Var(\hat{\mu}_1) = Var(\hat{\mu}_3) = \frac{\sigma^2}{n}$

. $Var(\hat{\mu}_2) = \sigma^2$

→ ERROR: $Er(\hat{\mu}) = bias^2(\hat{\mu}) + Var(\hat{\mu})$

$$\begin{cases} E_n(\hat{\mu}_1) = \dfrac{\sigma^2}{n} \\ E_n(\hat{\mu}_2) = \sigma^2 \\ E_n(\hat{\mu}_3) = \dfrac{1}{n^2} + \dfrac{\sigma^2}{n} \end{cases}$$

→ $\hat{\mu}_1$ and $\hat{\mu}_2$ are <u>both</u> unbiased, but $\hat{\mu}_1$ is (much) better than $\hat{\mu}_2$

→ $\hat{\mu}_3$ is biased, and $\hat{\mu}_2$ is not, but $\hat{\mu}_3$ is (much) better than $\hat{\mu}_2$

↳ <u>Variance matters</u> just as much as bias.

<u>Exercise 6</u> slide 33. RIDGE REGRESSION.

loss $J(\beta) = MSE(\beta) + \lambda \|\beta\|^2$. $\lambda \geq 0$

$= \frac{1}{n} (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|^2$.

gradient

$\nabla J(\beta) = \frac{2}{n} X^T (X\beta - y) + 2\lambda\beta$.

$\nabla J(\hat{\beta}_\lambda^{ridge}) = 0 \iff \hat{\beta}_\lambda^{ridge} = (X^T X + \lambda n I)^{-1} X^T y$

<u>NB</u>: when $\lambda > 0$, $(X^T X + \lambda n I)^{-1}$ <u>always</u> exists (exercise 2)

<u>Assumption</u>: $X^T X = nI$ (orthogonal vectors x_i)

↳ $\hat{\beta}_\lambda^{ridge} = \frac{1}{n(1+\lambda)} X^T y$ ⟩ $y = X\beta^* + \varepsilon$

$= \frac{1}{n(1+\lambda)} (\underbrace{X^T X}_{=nI} \beta^* + X^T \varepsilon)$

$\hat{\beta}_\lambda^{ridge} = \frac{1}{(1+\lambda)} \beta^* + \frac{1}{n(1+\lambda)} X^T \varepsilon$

Taking expectations: $\mathbb{E}(\hat{\beta}_\lambda^{ridge}) = \frac{1}{1+\lambda} \beta^* + \frac{1}{n(1+\lambda)} X^T \underbrace{\mathbb{E}(\varepsilon)}_{=0}$

$= \frac{1}{1+\lambda} \beta^*$

→ Hence, $\text{bias}\left(\hat{\beta}_\lambda^{\text{ridge}}\right) = \mathbb{E}\left(\hat{\beta}_\lambda^{\text{ridge}}\right) - \beta^*$

$$= -\frac{\lambda}{1+\lambda}\beta^*$$

→ large λ = large bias
↳ small λ = small bias.

→ $\text{Var}\left(\hat{\beta}_\lambda^{\text{ridge}}\right) = \text{Var}\left(\frac{1}{n(1+\lambda)} X^T \varepsilon\right)$

$$= \frac{1}{n^2(1+\lambda)^2} X^T \underbrace{\text{Var}(\varepsilon)}_{\sigma^2 I} X$$

$$= \frac{\sigma^2}{n^2(1+\lambda)^2} \underbrace{X^T X}_{nI}$$

$$= \frac{\sigma^2}{n(1+\lambda)^2} I.$$

→ large λ = small variance
↳ small λ = large variance.

under our assumptions, $\text{Var}\left(\hat{\beta}^{\text{OLS}}\right) = \sigma^2 (X^T X)^{-1} = \frac{\sigma^2}{n} I$

so $\text{Var}\left(\hat{\beta}_\lambda^{\text{ridge}}\right) = \frac{1}{(1+\lambda)^2} \text{Var}\left(\hat{\beta}^{\text{OLS}}\right)$

$\lambda \to 0$: $\hat{\beta}_\lambda^{\text{ridge}} \to \hat{\beta}^{\text{OLS}}$: no bias, large variance

$\lambda \to +\infty$: $\hat{\beta}_\lambda^{\text{ridge}} \to 0$: large bias, no variance.

Exercise 7 : Ridge performance

$$f(\lambda) = \underset{S, x_0}{\mathbb{E}} \left[\text{bias}^2 \left(x_0^T \hat{\beta}_\lambda^{\text{ridge}} \right) + \text{Var} \left(x_0^T \hat{\beta}_\lambda^{\text{ridge}} \right) \right] \quad \Big| \begin{array}{l} \mathbb{E} x_0 = 0 \\ \mathbb{E} x_0 x_0^T = I \end{array}$$

= expectation of total error.

→ Let us compute the two terms:

- $\text{bias}^2 \left(x_0^T \hat{\beta}_\lambda \right) = \left(x_0^T \, \text{bias} \left(\hat{\beta}_\lambda \right) \right)^2$

$$= \frac{\lambda^2}{(1+\lambda)^2} \left(x_0^T \beta^* \right)^2$$

$$= \frac{\lambda^2}{(1+\lambda)^2} \beta^{*T} x_0 x_0^T \beta^*$$

↓ $\text{bias}(\hat{\beta}) = \beta^*$

↓ $\left(x_0^T \beta^* \right)^2 = \sum x_{0,i} x_{0,j} \beta_i^* \beta_j^* = \beta^{*T} x_0 x_0^T \beta^*$

- $\text{Var} \left(x_0^T \hat{\beta}_\lambda^{\text{ridge}} \right) = x_0^T \, \text{Var} \left(\hat{\beta}_\lambda^{\text{ridge}} \right) x_0$

$$= \frac{\sigma^2}{n(1+\lambda)} x_0^T x_0$$

→ Taking expectations, with $\mathbb{E}(x_0 x_0^T) = I$, we get:

$$f(\lambda) = \frac{\lambda^2}{(1+\lambda)^2} \|\beta^*\|^2 + \frac{\sigma^2 p}{n(1+\lambda)^2}.$$

$$\mathbb{E}(x_0^T x_0) = \sum_{i=1}^{p} \mathbb{E}(x_{0,i}^2) = p$$

→ Finding the optimum λ^*:

Define $a = \|\beta^*\|^2 = f(+\infty)$ $\quad b = \frac{\sigma^2 p}{n} = f(0)$

$$f(\lambda) = \frac{1}{(1+\lambda)^2} (a\lambda^2 + b)$$

$$f'(\lambda) = \frac{2}{(1+\lambda)^3} (a\lambda - b)$$

$$f'(\lambda^*) = 0 \iff \lambda^* = \frac{b}{a} = \frac{\sigma^2 p}{n \|\beta^*\|^2}$$

---

keeping our notations: $f(\lambda^*) = \frac{1}{(1+\frac{b}{a})^2} \left(a \left(\frac{b}{a} \right)^2 + b \right) = \frac{ab}{a+b}$

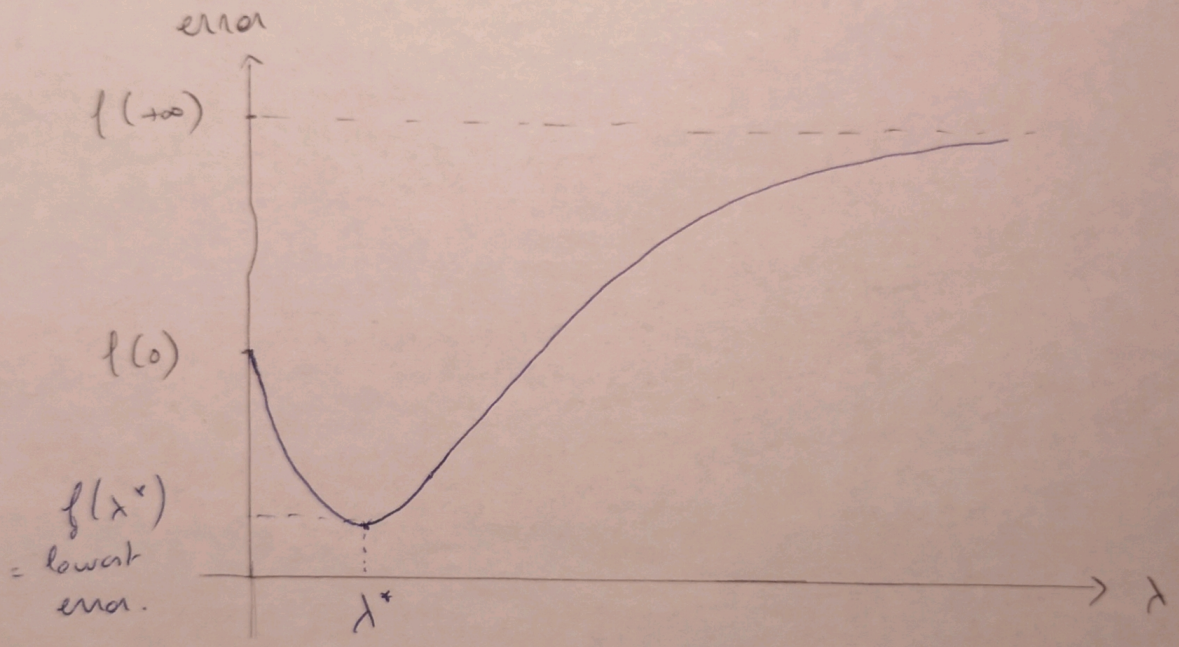
$$f(\lambda^*) = \frac{ab}{a+b} = \frac{f(0)\,f(+\infty)}{f(0)+f(+\infty)}$$

$$a \geq 0 \Rightarrow \frac{ab}{a+b} \leq \frac{a(a+b)}{a+b} = a$$

$$b \geq 0 \Rightarrow \frac{ab}{a+b} \leq \frac{(a+b)b}{a+b} = b$$

$$\Big\} \quad \frac{ab}{a+b} \leq \min(a,b)$$

So finally, $f(\lambda^*) \leq \min\left(f(0), f(+\infty)\right)$



(can be)

CONCLUSION : Some regularization is better than none and better than the estimator 0

Exercise 8.

In notebook of practical session 2 in more detail

Exercise 8

$$\beta^{new} = \beta^{old} - \left[\nabla^2_\beta J(\beta^{old})\right]^{-1} \nabla_\beta J(\beta^{old})$$

lemma:

β^{new} is also the solution to $\min_\beta J_q(\beta)$

where $J_q(\beta) := J(\beta^{old}) + \nabla J(\beta^{old})^T (\beta - \beta^{old}) + \frac{1}{2} (\beta - \beta^{old})^T \nabla^2_\beta J(\beta^{old}) (\beta - \beta^{old})$

J_q is the quadratic approximation to J in β^{old}.

(analogy in 1 dimension: $J_q(x) = J(x_0) + J'(x_0)(x - x_0) + \frac{1}{2} J''(x_0)(x - x_0)^2$)

proof: $\nabla J_q(\beta) = \nabla J(\beta^{old}) + \nabla^2_\beta J(\beta^{old})(\beta - \beta^{old})$

$= 0 \iff \beta = \beta^{old} - \left[\nabla^2_\beta J(\beta^{old})\right]^{-1} \nabla_\beta J(\beta^{old})$

$= \beta^{new}$

Writing all the terms of J_q that depend on β:

$\rightarrow \beta^T \nabla J(\beta^{old}) = \frac{1}{n} \beta^T X^T P y + 2\lambda \beta^T \beta^{old}$

$\rightarrow \frac{1}{2} \beta^T \nabla^2_\beta J(\beta^{old}) \beta = \frac{1}{2n} \beta^T X^T W X \beta + \lambda \beta^T \beta$

$\rightarrow -\beta^T \nabla^2_\beta J(\beta^{old}) \beta^{old} = -\frac{1}{n} \beta^T X^T W X \beta^{old} - 2\lambda \beta^T \beta^{old}$.

Putting it all together:

$$J_q(\beta) = -\frac{1}{n} \beta^T X^T W X \beta^{old} + \frac{1}{n} \beta^T X^T P y + \frac{1}{2n} \beta^T X^T W X \beta + \lambda \beta^T \beta + C$$

$$= -\frac{1}{n} \beta^T X^T W \underbrace{(X\beta^{old} - W^{-1} P y)}_{:= z} + \frac{1}{2n} \beta^T X^T W X \beta + \lambda \beta^T \beta + C$$

$$= \frac{1}{2n} (X\beta - z)^T W (X\beta - z) + \lambda \|\beta\|^2 + C.$$