SDS 391-3

Lecture 20, Thu Nov 6

- GAUSS MARKOV THEOREM

Assume a linear model with fixed covariates, i.e. $\underset{n\times d}{\Phi}$ is deterministic:

$$\underset{n\times 1}{Y} = \underset{n\times d}{\Phi}\,\underset{d\times 1}{\beta^*} + \underset{n\times 1}{\varepsilon}$$

↳ iid errors $\sim (0, \sigma^2)$ ↓ homoschedastic errors

Any estimator of β^* of the form $\underset{d\times n}{A}Y$ (deterministic) is a linear unbiased estimator of β^* if

$$\mathbb{E}[AY] = \beta^* \qquad (\text{for all } \beta^* \in \mathbb{R}^d)$$

↳ $\hat\beta = (\Phi^T\Phi)^{-1}\Phi^T Y$ is a linear unbiased estimator.

- How good is $\hat\beta$ compared to all linear unbiased estimators?

<u>Gauss - Markov Thm</u> $\hat\beta$ is the BLUE! (Best Linear Unbiased Estimator)

<u>Remark</u>: "Best" means that, for any other linear unbiased estimator AY,

$$\mathrm{Var}[\hat\beta] \preceq \mathrm{Var}[AY]$$

where for 2 psd matrices M_1 and M_2 of the same size

$$M_1 \preceq M_2 \iff M_2 - M_1 \succeq 0$$

psd order

↳ "is positive semidefinite"

This means that, for any $x \in \mathbb{R}^d$,

$$x^T \operatorname{Var}[\hat{\beta}]\, x \leq x^T \operatorname{Var}[AY]\, x$$

choose x to be the i^{th} standard basis vector $e_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \to i^{th}$ position

to conclude that

the i^{th} element of diagonal of $\operatorname{Var}[\hat{\beta}]$ is

smaller that i^{th} element $\operatorname{Var}[AY]$

$$\hookrightarrow \operatorname{Var}[\hat{\beta}_i] \leq \operatorname{Var}[\tilde{\beta}_i] \qquad \tilde{\beta} = AY$$

In fact this implies

$$\operatorname{Var}[\underbrace{c^T\hat{\beta}}_{\text{contrast}}] \leq \operatorname{Var}[c^T\tilde{\beta}] \quad \text{for all } c \in \mathbb{R}^d$$

Pf/ Let $\tilde{\beta} = AY$ s.t. $\mathbb{E}[\tilde{\beta}] = \beta^*$. Then

$$\beta^* = \mathbb{E}[\underset{d\times n}{A}Y] = \mathbb{E}[A\Phi\beta^* + A\varepsilon] = A\Phi\beta^* \quad \forall \beta^*$$

$\Downarrow$

$$A\Phi = I_{d\times d}$$

Next $$\operatorname{Var}[\tilde{\beta}] = A \underbrace{\operatorname{Var}[Y]}_{\sigma^2 I_n} A^T = \sigma^2 AA^T$$

Let $D = A - (\Phi^T\Phi)^{-1}\Phi^T$ so

$$\operatorname{Var}[\tilde{\beta}] = \operatorname{Var}[AY] = \sigma^2 \left(D + (\Phi^T\Phi)^{-1}\Phi^T\right)\left(D + (\Phi^T\Phi)^{-1}\Phi^T\right)^T$$

But $\underbrace{\left(D + (\Phi^T\Phi)^{-1}\Phi^T\right)}_{A}\Phi = I_d$ so

$$\mathrm{Var}[\tilde{\beta}] = \underbrace{\sigma^2 DD^T}_{\succeq 0} + \underbrace{\sigma^2 (\Phi^T\Phi)^{-1} \cancel{\Phi^T \Phi (\Phi^T\Phi)^{-1}}}_{\sigma^2 (\Phi^T\Phi)^{-1} = \mathrm{Var}[\hat{\beta}]}$$

$$\downarrow$$

$$\mathrm{Var}[\tilde{\beta}] \succeq \mathrm{Var}[\hat{\beta}] \quad \blacksquare$$

- ftw : extension to random Φ !

RIDGE REGRESSION

- Chapter 3 for Bach's book Learning theory from first principles

- Suppose d is large compared to n, almost close to n

  in this case $\Phi^T\Phi$ may not be well-conditioned

  i.e. $\dfrac{\lambda_{max}(\Phi^T\Phi)}{\lambda_{min}(\Phi^T\Phi)}$ is large and $(\Phi^T\Phi)^{-1}$ is unstable

  $\Rightarrow$ $\mathrm{Var}[\hat{\beta}] = \sigma^2 (\Phi^T\Phi)^{-1}$ is unstable

- One approach is to regularize, to solve a penalized least squares problem that includes a penalty for not choosing "good" solutions.

- The ridge regression estimator arises as the solution to this problem:

For a value $\lambda \geq 0$, let
↳ penalty parameter

ridge estimator $\leftarrow \hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \underbrace{\frac{\| Y - \Phi\beta \|^2}{n} + \lambda \|\beta\|^2}_{f(\beta)}$

$$= \left(\frac{\Phi^T \Phi}{n} + \lambda I_d\right)^{-1} \frac{\Phi^T Y}{n} = \left(\hat{\Sigma} + \lambda I_d\right)^{-1} \frac{\Phi^T Y}{n}$$

when $\lambda = 0$ this reduces to OLS $\hat{\beta}$ (if $\hat{\Sigma}$ is invertible or the min-norm estimator $\Phi^+ Y$ if it is not)

Pf/ The objective function is strictly convex so the first order optimality conditions are

$$0 = \nabla f(\hat{\beta}_\lambda) = \frac{2}{n} \Phi^T (\Phi\hat{\beta}_\lambda - Y) + 2\lambda\hat{\beta}_\lambda$$

↳ solve for $\hat{\beta}_\lambda$ ∎

This solution exists, is unique even if $d > n$ and regardless of how poorly conditioned $\Phi^T\Phi$ is!

Remarks 1) if $\lambda \to 0$ then

$\hat{\beta}_\lambda \to \hat{\beta}_{\text{MIN-NORM}}$ or $\hat{\beta}$ if $\Phi^T\Phi$ is invertible

2) Alternative expression:

$$\hat{\beta}_\lambda = \left(\frac{\Phi^T\Phi}{n} + \lambda I_d\right)^{-1} \frac{\Phi^T Y}{n} = \frac{\Phi^T}{n}\left(\frac{\Phi\Phi^T}{n} + \lambda I_n\right)^{-1} Y$$

3) Let $\underset{n\times d}{\Phi} = U \Sigma V^T$ Then

$$\hat{Y} = \Phi \hat{\beta}_\lambda = \sum_{j=1}^{\mathrm{rank}(\Phi)} \underset{\substack{\downarrow \\ j^{th} \text{ column of } U}}{u_j} \langle Y, u_j \rangle \frac{\sigma_j^2}{\sigma_j^2 + \lambda}$$

σ_j j^{th} singular value of Φ

in contrast $$\Phi \hat{\beta} = \sum_{j=1}^{\mathrm{rank}(\Phi)} u_j \langle Y, u_j \rangle$$

- <u>Thm 3.7</u> The excess risk of $\hat{\beta}_\lambda$ is

$$\mathbb{E}\left[R(\hat{\beta}_\lambda)\right] - \underbrace{R(\beta^*)}_{\sigma^2} = \lambda^2 \beta^{*T} (\hat{\Sigma} + \lambda I_d)^{-2} \hat{\Sigma} \beta^* + \frac{\sigma^2}{n} \, \mathrm{tr} \, \underbrace{\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I_d)^{-2} \right)}_{\sum_j \frac{\hat{\lambda}_j^2}{(\hat{\lambda}_j + \lambda)^2}}$$

$\hat{\lambda}_j$ jth eigenv. of $\hat{\Sigma}$

$=$ Bias term + Variance term

- Bias is $\uparrow$ in λ variance is $\downarrow$ in λ.
- Task : choose optimal λ ! In Prop. 3.8 of Bach's book you will see a convenient choice of λ that minimizes an upper bound on the excess risk

In particular with $\lambda = \dfrac{\sigma \, \mathrm{tr}(\hat{\Sigma})}{\|\beta^*\| \sqrt{n}}$

$$\mathbb{E}\left[R(\hat{\beta}_\lambda)\right] - R(\beta^*) \leq \frac{\sigma \, \mathrm{tr}(\hat{\Sigma}) \, \|\beta^*\|_2}{\sqrt{n}}$$

- <u>Remark</u> : i) this is a "slow" rate

ii) we do not know σ or $\|\beta^*\|$

|

iii) in practice you will do cross validation

- MINIMAX
  LOWER BOUND ON EXCESS RISK (Section 3.7 in Bach's book)