SDS 387
Linear Models

Fall 2024

Lecture 18 - Thu, Oct 31, 2024

Instructor: Prof. Ale Rinaldo

GAUSS MARKOV THEOREM

We are in the setting where $\underset{n\times d}{\Phi}$ is a deterministic design matrix and the linear model is well-specified: rank$(\Phi) = d < n$

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

$$\varepsilon \sim (0, \sigma^2 I_n)$$

Any estimator of β^* of the form $\underset{d\times n}{A}Y$, where A is a deterministic matrix (possibly depending on Φ) is said to be linear and is unbiased if

$$\mathbb{E}[AY] = \beta^*, \quad \forall \beta^*$$

- We know that $\hat{\beta} = (\Phi^T\Phi)^{-1}\Phi^T Y$ is an unbiased linear estimator of β^*.

- GAUSS-MARKOV THEOREM: $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of β^*

This means that, among all unbiased linear estimators of β^* of the form $\underset{d\times n}{A}Y$,

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

where we say that a $d\times d$ matrix B is $\succeq 0$ (non-negative in the positive-semidefinite or Lowener partial order) if B is psd. Then we can write $\underset{d\times d}{A} \succeq \underset{d\times d}{B}$ to mean that $A - B \succeq 0$.

<u>Remark</u> This is a partial order, meaning that it is possible that A and B are not comparable meaning neither $A \preceq B$ nor $B \preceq A$ holds. In particular $A \npreceq B$ does not mean $A \succeq B$.

Pf/ Let AY be an unbiased estimator of β^*. Then

$$\beta^* = \mathbb{E}[AY] = A\Phi\beta^* + A\underbrace{\mathbb{E}[\varepsilon]}_{=0} = A\Phi\beta^*$$

$\Rightarrow$ this is true for any β^* so $A\Phi = I_{d\times d}$

Now $\mathrm{Var}[AY] = \sigma^2 AA^T$ (because $\mathrm{Var}[Y] = \mathrm{Var}[\varepsilon] = \sigma^2 I_n$)

Next let $D = A - (\Phi^T\Phi)^{-1}\Phi^T$. So

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

Because $A\Phi = I$, $\left(D + (\Phi^T\Phi)^{-1}\Phi^T\right)\Phi = I$ or

$D\Phi + I = I \Rightarrow D\Phi = 0$

As a result

$$\mathrm{Var}[AY] = \sigma^2 DD^T + \sigma^2 (\Phi^T\Phi)^{-1}\cancel{\Phi^T\Phi}\cancel{(\Phi^T\Phi)^{-1}}$$

+ cross terms that are zero

$$= \underbrace{\sigma^2 DD^T}_{\succeq 0} + \mathrm{Var}[\hat{\beta}]$$

$\Downarrow$

$$\mathrm{Var}[AY] - \mathrm{Var}[\hat{\beta}] \succeq 0 \quad \text{or} \quad \mathrm{Var}[AY] \succeq \mathrm{Var}[\hat{\beta}]$$

In fact $c^T\hat{\beta}$ for any $c \in \mathbb{R}^d$ is the BLUE of $c^T\beta^*$

Learning Theory from First Principles by Francis Bach

Chapter 3

RIDGE REGRESSION

- Suppose that d is large compared to n, meaning that $\frac{d}{n}$ is close to 1. There are of course computational issues because $\Phi^T\Phi$ is not well conditioned and statistical issues because $\mathrm{Var}[\hat{\beta}] = \sigma^2 \underbrace{(\Phi^T\Phi)^{-1}}_{\text{poorly conditioned}}$

- One approach is to <u>regularize</u>, to solve a penalized least-squares problem that encourages "good" properties of the solutions. Ridge regression is an important example (and so is Lasso).

<u>Note</u>: here we are interested in minimizing the prediction risk or the excess risk.

The ridge least squares estimator is

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{\|Y - \Phi\beta\|^2}{n} + \lambda \|\beta\|^2 \qquad \lambda > 0$$

regularization parameter

if $\lambda = 0$ this gives us $\hat{\beta}$, the OLS

$\hat{\beta}_\lambda$ is uniquely defined no matter Φ and is equal to

$$\hat{\beta}_\lambda = \underbrace{\left(\frac{\Phi^T\Phi}{n} + \lambda I_d\right)^{-1}}_{\text{always invertible}} \frac{\Phi^T}{n} Y = (\hat{\Sigma} + \lambda I)^{-1} \frac{\Phi^T Y}{n}$$

Pf/ Let $F_\lambda(\beta) = \frac{1}{n} \|Y - \Phi\beta\|^2 + \lambda\|\beta\|^2$ is strictly convex so $\hat{\beta}_\lambda$ is found by checking first order optimality conditions:

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

Remarks HW!

1) $\lim_{\lambda \downarrow 0} \hat{\beta}_\lambda = \hat{\beta}_{MN}$ HW

2) Alternative expression:

$$\hat{\beta}_\lambda = \frac{\Phi^T}{n} \Big(\underbrace{\frac{\Phi\Phi^T}{n}}_{n\times n} + \lambda I_n \Big)^{-1} Y$$

better numerically if $\mathrm{rank}(\Phi) = n < d$

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

$$\Phi\hat{\beta}_\lambda = \sum_{i=1}^{\mathrm{rank}(\Phi)} u_i \langle Y, u_i \rangle \frac{\sigma_i^2}{\sigma_i^2 + \lambda}$$

u_i: ith column of U

σ_i ith singular value

compare to

$$\Phi\hat{\beta}_{MN} = \sum_{i=1}^{\mathrm{rank}(\Phi)} u_i \langle Y, u_i \rangle$$

Statistical properties of $\hat{\beta}_\lambda$ (recall we want to minimize the risk!)

Proposition 3.7 The excess risk of $\hat{\beta}_\lambda$

$$\mathbb{E}[R(\hat{\beta}_\lambda)] - R(\beta^*) = \lambda^2 \beta^{*T} (\hat{\Sigma} + \lambda I_d)^{-2} \hat{\Sigma} \beta^* +$$

$$\frac{\sigma^2}{n} \operatorname{tr}\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I_d)^{-1}\right)$$

$$= B + V$$

B is a bias term increasing in λ and V is the variance term, decreasing in λ

<u>Note</u> if $\hat{\Sigma}$ is invertible, when $\lambda = 0$ we recove the risk of $\hat{\beta}$ (which is $\sigma^2 \frac{d}{n}$).

Pf/ Recall the decomposition of the excess risk of any estimator $\tilde{\beta}$:

$$\mathbb{E}[R(\tilde{\beta})] - R(\beta^*) = \| \mathbb{E}[\tilde{\beta}] - \beta^* \|^2_{\hat{\Sigma}} + \mathbb{E}\left[\| \tilde{\beta} - \mathbb{E}[\tilde{\beta}] \|^2_{\hat{\Sigma}}\right]$$

σ^2

Now replace $\tilde{\beta}$ by $\hat{\beta}_\lambda$. For the bias:

$$\mathbb{E}[\hat{\beta}_\lambda] = (\hat{\Sigma} + \lambda I)^{-1} \frac{\Phi^T}{n} \Phi \beta^* + (\hat{\Sigma} + \lambda I)^{-1} \frac{\Phi^T}{n} \underbrace{\mathbb{E}[\varepsilon]}_{0}$$

$$= (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} \beta^*$$

Exercise

$$= \beta^* - \lambda (\hat{\Sigma} + \lambda I_d)^{-1} \beta^*$$

$$\hookrightarrow \| \mathbb{E}[\hat{\beta}_\lambda] - \beta^* \|^2_{\hat{\Sigma}} = \lambda^2 \beta^{*T} (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-1} \beta^*$$

$$= \lambda^2 \beta^{*T} (\hat{\Sigma} + \lambda I_d)^{-2} \hat{\Sigma} \beta^*$$

because $(\hat{\Sigma} + \lambda I)^{-1}$ and $\hat{\Sigma}$ commute. HW

As for the variance term:

$$\mathbb{E}\left[\|\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda]\|^2_{\hat{\Sigma}}\right] = \mathbb{E}\left[\left\|(\hat{\Sigma} + \lambda I_d)^{-1} \frac{\Phi^T \varepsilon}{n}\right\|^2_{\hat{\Sigma}}\right]$$

$$= \frac{1}{n^2} \mathbb{E}\left[\varepsilon^T \Phi (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-1} \Phi^T \varepsilon\right]$$

$$= \frac{1}{n^2} \mathbb{E}\left[\operatorname{tr}\left(\Phi^T \varepsilon \varepsilon^T \Phi \; (\hat{\Sigma} + \lambda)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-1}\right)\right]$$

$$= \frac{\sigma^2}{n} \operatorname{tr}\left(\hat{\Sigma} \; (\hat{\Sigma} + \lambda)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda)^{-1}\right)$$

$$= \frac{\sigma^2}{n} \underbrace{\operatorname{tr}\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I)^{-2}\right)}_{\sum_i \frac{\hat{\lambda}_i^2}{(\hat{\lambda}_i + \lambda)^2}}$$

$\hat{\lambda}_i$ ith eigenvalue of $\hat{\Sigma}$

degrees of freedom