SDS 387

Linear Models

Fall 2025

Lecture 22 - Thu, Nov 13, 2025

Instructor: Prof. Ale Rinaldo

- Today finish the proof of minimax optimality of the OLS estimator when is the model is linear and the covariates are fixed. That is, our data consist of

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \text{s.t.}$$

$$Y = \underbrace{\Phi}_{n \times d \text{ fixed matrix}} \overbrace{\beta^*}^{\text{True unknown coeff.}} + \varepsilon, \quad \varepsilon_1, \dots, \varepsilon_n \overset{\text{iid}}{\sim} (0, \underbrace{\sigma^2}_{\text{known!}})$$

- <u>Remark</u>: the extension to the random design case can be found in Mourtada's paper. The result is the same: the OLS estimator is the minimax estimator if the model is linear.

- Last time we established the following lower bound

$$\inf_{A} \sup_{\beta \in \mathbb{R}^d} \underbrace{\mathbb{E}_{\substack{\varepsilon = \begin{bmatrix}\varepsilon_1 \\ \vdots \\ \varepsilon_n\end{bmatrix} \\ \text{iid} \sim (0, \sigma^2)}} \left[R\left(A(\Phi\beta + \varepsilon) \right) \right] - \sigma^2}_{\text{excess risk of } A(\cdot)}$$

ridge estimator

$$\geq \mathbb{E}_{(\beta, Y)} \left[\| \hat{\beta}_\lambda - \beta \|^2_{\hat{\Sigma}} \right] \qquad (\text{A})$$

$$\hat{\Sigma} \to \frac{\Phi^T \Phi}{n}$$

where $\beta \sim N_d\left(0, \frac{\sigma^2}{n\lambda} I_d\right)$ prior

$$Y \mid \beta \sim N_n\left(\Phi\beta, \sigma^2 I_n\right)$$

This is a standard argument: lower bound the maximal possible risk of any estimator $A(\cdot)$ over all $\beta \in \mathbb{R}^d$ by the average risk of $A(\cdot)$ with respect to a carefully chosen distribution for β (a prior).

- Next, we have that (A) is equal

$$\mathbb{E}_{\beta \sim N_d\left(0, \frac{\sigma^2}{n\lambda} I_d\right)} \mathbb{E}_{\varepsilon \sim N_n(0, \sigma^2 I_n)} \left[\| \underbrace{(\Phi^T\Phi + n\lambda I_d)^{-1} \Phi^T (\Phi\beta + \varepsilon) - \beta}_{T_0} \|^2_{\hat{\Sigma}} \right]$$

Next, we have that $\xrightarrow{}$ exercise

$$T_0 = (\Phi^T\Phi + n\lambda I_d)^{-1} \Phi^T \varepsilon - n\lambda (\Phi^T\Phi + n\lambda I_d)\beta$$

Because $\varepsilon \perp\!\!\!\perp \beta$ the expression reduces to

$$\mathbb{E}_{\varepsilon \sim N_n(0, \sigma^2 I_n)} \left[\left\| (\hat{\Sigma} + \lambda I_d) \frac{\Phi^T}{n} \varepsilon \right\|^2_{\hat{\Sigma}} \right] + \mathbb{E}_{\beta \sim N_d(0, \frac{\sigma^2}{n\lambda} I_d)} \left[\left\| \lambda (\hat{\Sigma} + \lambda I_d)^{-1} \beta \right\|^2_{\hat{\Sigma}} \right]$$

$$= T_1 + T_2$$

It can be seen to see that

$$T_1 = \frac{\sigma^2}{n} \operatorname{tr}\left((\hat{\Sigma} + \lambda I_d)^{-2} \hat{\Sigma}^2 \right)$$

$$T_2 = \lambda^2 \, \mathbb{E}_\beta \left[\beta^T (\hat{\Sigma} + \lambda I_d)^{-1} \hat{\Sigma} \, (\hat{\Sigma} + \lambda I_d)^{-1} \beta \right]$$

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

↳ on HW

↳

$$T_1 + T_2 = \frac{\sigma^2}{n} \operatorname{tr}\left((\hat{\Sigma} + \lambda I_d)^{-1} \hat{\Sigma} \right) \qquad \forall \lambda > 0$$

This is a lower bound on the minimax risk.

EXERCISE

Notice that $\operatorname{tr}\left((\hat{\Sigma} + \lambda I_d)^{-1} \hat{\Sigma} \right) = \sum_{j=1}^{d} \frac{\hat{\lambda}_j}{\hat{\lambda}_j + \lambda}$

where $\hat{\lambda}_j$ is the j^{th} eigenvalue of $\hat{\Sigma}$.

This is decreasing in λ. So

$$\sup_\lambda T_1 + T_2 = \frac{\sigma^2}{n} \lim_{\lambda \downarrow 0} \operatorname{tr}\left((\hat{\Sigma} + \lambda I_d)^{-1} \hat{\Sigma} \right)$$

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

$$= \boxed{\frac{\sigma^2 d}{n}}$$

$\downarrow$

This is the excess risk of $\hat{\beta}$ OLS

- About minimaxity for estimation: let $\mathcal{P} = \{P_\theta, \theta \in \underbrace{\Theta}_{\text{parameter space}}\}$ be a parametric family of prob. distributions (a parametric statistical model). We are interested in estimating θ^*, the true parameter, s.t.

$$X_1, \dots, X_n \overset{iid}{\sim} P_{\theta^*}$$

For any estimator $\hat{\theta}$ of θ^* (where $\hat{\theta} = \underbrace{\hat{\theta}(X_1, \dots, X_n)}_{\text{function of data}}$)

let $L(\hat{\theta}, \theta^*)$ be the loss function associated to $\hat{\theta}$ (e.g. $L(\hat{\theta}, \theta^*) = \|\hat{\theta} - \theta^*\|^2$).

The risk of $\hat{\theta}$ is the function

$$\theta \in \Theta \longmapsto \mathbb{E}_{X_1, \dots, X_n \overset{iid}{\sim} P_\theta}\left[L(\hat{\theta}, \theta)\right] = R(\hat{\theta}, \theta)$$

- The minimax risk for this problem

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta)$$

$\downarrow$

over all estimators

A natural lower bound on the minimax risk is the Bayes risk

$$\inf_{\hat{\theta}} \underset{\theta \sim \pi}{\mathbb{E}} \left[R(\hat{\theta}, \theta) \right] = R(\pi)$$

$\pi \downarrow$ prior

a procedure attaining that infimum is called a Bayes procedure wrt π.

When the loss function is e.g. quadratic, the Bayes procedure is the posterior mean of θ.

- If $\{\pi_k\}$ is a sequence of priors s.t.

$$R(\pi_k) \to r \quad \text{as} \quad k \to \infty$$

and $\hat{\theta}$ is a procedure s.t.

$$\sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = r$$

Then $\hat{\theta}$ is minimax $\rightarrow$ HW

- Remark: if the covariate (i.e. the rows of Φ) are random, the risk of OLS $\hat{\beta}$ is

$$\frac{\sigma^2}{n} \underset{\Phi}{\mathbb{E}} \operatorname{tr}\left(\hat{\Sigma}^{-1} \Sigma \right) \quad \text{where}$$

$$\Sigma = \mathbb{E}\left[\Phi_1 \Phi_1^T\right]$$

$\hookrightarrow$ 1st row of Φ.

This is also the minimax risk.

See Mourtada's paper.

STATISTICAL INFERENCE FOR β^*.

As before we assume a well-specified linear model and fixed covariates, i.e.

$$\underset{n\times 1}{Y} = \underset{\text{fixed}}{\Phi} \overset{\text{unknown}}{\beta^*} + \varepsilon \quad \hookrightarrow \overset{iid}{\sim} (0, \sigma^2)$$

Goal: to estimate and carry out statistical inference for β^*, in fixed dimensions (i.e. d is fixed)

- Is the OLS consistent?

$$\hat{\beta} \xrightarrow{p} \beta^*$$

- Yes! Assume that $\hat{\Sigma} = \frac{\Phi^T \Phi}{n} \to \underset{d\times d}{\Sigma}$ p.d.

Then

$$\hat{\Sigma}^{-1} \frac{\Phi^T \varepsilon}{n} \xrightarrow{p} 0$$

Pf/ By WLLN $\frac{\Phi^T \varepsilon}{n} = \frac{1}{n} \sum_{i=1}^{n} \Phi_i \varepsilon_i \xrightarrow{p} 0$ (Φ_i: transpose of ith row of Φ)

To see this $\Phi_i \varepsilon_i \sim (0, \sigma^2 \Phi_i \Phi_i^T)$

So because the ε_i's are indep.

$$\mathrm{Var}\left[\frac{\Phi^T \varepsilon}{n}\right] = \frac{\sigma^2}{n^2} \Phi^T \Phi .$$

$$= \frac{\sigma^2}{n}\left(\frac{\Phi^T \Phi}{n} - \Sigma + \Sigma\right)$$

$$= \frac{\sigma^2}{n} \Sigma + \frac{\sigma^2}{n}\left(\frac{\Phi^T \Phi}{n} - \Sigma\right) \to 0$$

as $n \to \infty$

Then $\hat{\Sigma}^{-1} \frac{\Phi^T \varepsilon}{n} \xrightarrow{P} 0$ by Slutsky's theorem $\blacksquare$

- $\hat{\beta}$ is asymptotically normal

$$\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N_d(0, \sigma^2 \Sigma^{-1})$$

$\Downarrow$

$\mathbb{E}[\hat{\beta}]$