SDS 387
Linear Models

Fall 2025

Lecture 21 - Tue, Nov 11, 2025

Instructor: Prof. Ale Rinaldo

- Today: minimax lower bound for linear regression when the model is well-specified (i.e. linear) and the covariates are fixed.

- See Section 3.7 of Bach's book
  - Exact minimax risk for linear least squares and the lower tail of sample covariance matrices by Jaouad Mourtada

AoS paper

- Recall that (if the model is linear and the covariates are fixed) the excess risk of the OLS is

$$\sigma^2 \frac{d}{n} = \mathbb{E}\left[\|\hat\beta - \beta^*\|^2_{\hat\Sigma_n} \right]$$

Question: is this risk any good?

- Answer: yes! The OLS is minimax optimal (in these settings)

1

- MINIMAX ESTIMATION: Suppose we are interested in estimating a parameter θ^*, which is a functional of the data generating distribution P^*. We will write this as $\theta^* = \theta(P^*)$. (In linear regression settings with random covariates we may take θ functional.

$$\theta(P^*) = \theta^* = \mathbb{E}[\Phi\Phi^\top]^{-1}\,\mathbb{E}[\Phi \cdot Y]$$

and P^* is the distribution of $(\Phi, Y) \in \mathbb{R}^d \times \mathbb{R}$.)

- We also need to specify a collection $\mathcal{P}$ of distributions containing P^*. → statistical model

- Example if

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

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

$$Y \sim N(\Phi\beta^*, \sigma^2 I_n)$$

$$\mathcal{P} = \{ N(\Phi\beta, \sigma^2 I_n),\ \beta \in \mathbb{R}^d \}$$

if P^* is $N(\Phi\beta^*, \sigma^2 I_n) \in \mathcal{P}$ then

$$\theta^* = \theta(P^*) = \beta^*$$

We call this model $\mathcal{P}_{\text{Gaussian}}$

If on the other hand $\varepsilon \sim (\underset{n\times 1}{0}, \sigma^2 I_n)$

Then $\mathcal{P}_{\text{well specified}} = \{$ set of all distributions for Y s.t. $\mathbb{E}[Y] = \Phi\beta$

$\text{Var}[Y] = \sigma^2 I_n$

$\beta \in \mathbb{R}^d \}$

<u>Remark</u> σ^2 is <u>known</u>

- Of course $\mathcal{P}_{\text{Gaussian}} \subset \mathcal{P}_{\text{well specified}}$

- In order to measure how well an estimator (a measurable function of the data!), say $\tilde{\beta}$,

does we consider its risk

$$R(\underset{\text{estimator}}{\tilde{\beta}}, \underset{\text{target parameter}}{\beta^*})$$

The estimator $\tilde{\beta}$ is computed using an iid sample from P^*.

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

In our case

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

$$\text{excess risk} = \mathbb{E}[R(\tilde{\beta})] - \underset{R(\beta^*)}{\sigma^2}$$

where $R(\beta)$ is the predictive risk

$$R(\beta) = \mathbb{E}\left[\frac{\| Y_{new} - \Phi\beta\|^2}{n}\right]$$

Y_{new} = new draw from P^*.

- We say that an estimator $\tilde{\beta}$ is minimax optimal if it minimizes

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\|\tilde{\beta} - \theta(P)\|^2_{\hat{\Sigma}}\right]$$

$\theta(P)$ is the regression parameter for P

- The quantity sample size

$$\inf_{\tilde{\beta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\|\tilde{\beta} - \theta(P)\|^2_{\hat{\Sigma}}\right] = R_{minimax,n}$$

for us $\mathcal{P}$ is P well specified

where inf is over all estimators is called the MINIMAX RISK for estimating $\theta(\cdot)$ over $\mathcal{P}$.

- For an estimator $\tilde{\beta}$ of β^* let $R_n^{sup}(\tilde{\beta})$ be st

$$R(\tilde{\beta}, \beta^*) \leq R_n^{sup}(\tilde{\beta})$$

Then $\tilde{\beta}$ is minimax rate optimal if

$$\limsup_{n \to \infty} \frac{R_n^{sup}(\tilde{\beta})}{R_{minimax,n}} \leq C$$

≥ 1

- $\tilde{\beta}$ is sharp minimax rate optimal if $C=1$
- $\tilde{\beta}$ is exact minimax optimal if

$$R_n^{\sup}(\tilde{\beta}) = R_{\text{minimax}, n}$$

<u>Thm</u> $\hat{\beta}$, the OLS estimator, is exact minimax optimal for $\mathcal{P}_{\text{Gaussian}}$. It is also minimax optimal for $\mathcal{P}_{\text{well-defined}}$

~~Proof~~ For our problem we are interested in a lower bound on the quantity:

$$\inf_{A} \sup_{\beta \in \mathbb{R}^d} \mathbb{E}_{\underset{n \times 1}{\varepsilon} \sim (0, \sigma^2 I_n)} \left[R\left(A(\overbrace{\Phi\beta + \varepsilon}^{Y})\right)\right] - \sigma^2$$

A: algorithm that takes as inputs $\underset{n\times 1}{Y}$ and $\underset{n\times d}{\Phi}$ → fixed matrix

and returns a vector in $\mathbb{R}^d$

$$\geq \inf_{A} \sup_{\beta \in \mathbb{R}^d} \mathbb{E}_{\varepsilon \sim N_n(0, \sigma^2 I_n)} \left[R(A(\Phi\beta + \varepsilon))\right] - \sigma^2$$

↓ We have replaced $\sup\limits_{p \in \mathcal{P}_{\text{well defined}}}$ by $\sup\limits_{p \in \mathcal{P}_{\text{Gaussian}}}$

$$\geq \inf_{A} \ \mathbb{E}_{\beta \sim \pi} \ \mathbb{E}_{\varepsilon \sim N_n(0, \sigma^2 I_n)} \ R\left[\left(A(\Phi\beta + \varepsilon)\right)\right] - \sigma^2$$

$\beta \sim \pi \to$ prior distribution over $\mathbb{R}^d$

π here is any distribution on $\mathbb{R}^d$. Not a formal Bayesian argument. We choose a prior that is mathematically convenient. The prior for β is $N_d\left(0, \frac{\sigma^2}{\lambda n} I_d\right)$ where $\lambda > 0$.

Then $(\beta, \Phi\beta + \varepsilon) \in \mathbb{R}^d \times \mathbb{R}^n$ is jointly Gaussian with mean $0 \in \mathbb{R}^d \times \mathbb{R}^n$ and covariance matrix

$$\frac{\sigma^2}{n\lambda} \begin{array}{cc} \begin{bmatrix} I_d & \Phi^T \\ \Phi & \Phi\Phi^T + n\lambda I_n \end{bmatrix} & \begin{array}{c} d \\ n \end{array} \\ \begin{array}{cc} d & n \end{array} & \end{array}$$

Recall that

$$R\left(A(\Phi\beta + \varepsilon)\right) - \sigma^2 = \| A(\Phi\beta + \varepsilon) - \beta \|^2_{\hat{\Sigma}}$$

$$\hat{\Sigma} = \frac{1}{n} \Phi^T \Phi$$

So we want to minimize wrt the choice of A the quantity:

$$\mathbb{E}_{(\beta, \underset{\underset{\Phi\beta+\varepsilon}{\downarrow}}{Y})} \left[\| A(Y) - \beta \|^2_{\hat{\Sigma}} \right] =$$

where (β, Y) are jointly Gaussian.

$$= \int_{\mathbb{R}^n} \int_{\mathbb{R}^d} \| A(y) - \beta \|^2_{\hat{\Sigma}} \, dP(\beta | y) \, dP(y)$$

$dP(\beta | y) \downarrow$ posterior distribution of β given y.

A standard calculation gives that

$$\beta \,|\, Y = y \sim N_d\left(\hat{\beta}_\lambda , \frac{\sigma^2}{n} \left(\hat{\Sigma} + \lambda I_d \right)^{-1} \right)$$

$\hat{\beta}_\lambda \downarrow$ ridge estimator

where $\hat{\beta}_\lambda = \left(\hat{\Sigma} + \lambda I_d \right)^{-1} \dfrac{\Phi^T y}{n}$.

Next

$$\int_{\mathbb{R}^d} \| A(y) - \beta \|^2_{\hat{\Sigma}} \, dP(\beta | y) = \mathbb{E}_{\beta | Y} \left[\| A(Y) - \beta \|^2_{\hat{\Sigma}} \right]$$

$$\geq \mathbb{E}_{\beta \mid Y}\left[\| \hat{\beta}_\lambda - \beta \|^2_{\hat{\Sigma}} \right]$$

conditional expectation of β given Y.