# SDS 387
# Linear Models

Fall 2024

Lecture 20 - Tue, Nov 7, 2024

Instructor: Prof. Ale Rinaldo

- Last time: minimax lower bound for OLS $\hat{\beta}$ assuming a well specified model and deterministic and full-rank design matrix $\underset{n \times d}{\Phi}$:

$$Y = \Phi \beta^* + \varepsilon$$

↳ n-dimensional vector with mean 0 and variance matrix $\sigma^2 I_n$

We know that, in this setting,

$$\sup_{\beta^*} \mathbb{E}_{\substack{Y = \Phi\beta^* + \varepsilon \\ \varepsilon \sim (0, \sigma^2 I_n)}} \left[ R(\hat{\beta}) \right] - \sigma^2 = \sigma^2 \frac{d}{n}$$

Now want to establish a lower bound on the largest possible expected excess risk that holds regardless of the choice of the estimator.

So we are interested in lower-bounding

$$\inf_{A} \sup_{P \in \mathcal{P}} \mathbb{E}_{P}\Big[R\big(A(\underbrace{\Phi\beta^*(P) + \varepsilon}_{Y})\big)\Big] - \sigma^2$$

where $\mathcal{P} = \{$ probability distributions for $Y = \Phi\beta^* + \varepsilon$, $\varepsilon \sim (0, \sigma^2 I_n)$, $\beta^* \in \mathbb{R}^d \}$

This expression is, in turn, lower bounded by

$$\inf_{A} \sup_{P \in \mathcal{P}_{Gauss}} \mathbb{E}_{P}\Big[R\big(A(\underbrace{\Phi\beta^*(P) + \varepsilon}_{Y})\big)\Big] - \sigma^2$$

algorithm that takes as input $Y \in \mathbb{R}^n$ and return an estimator $\tilde{\beta} \in \mathbb{R}^d$

where $\mathcal{P}_{Gaussian} = \{ Y \sim N_n(\Phi\beta^*, \sigma^2 I_n),\ \beta^* \in \mathbb{R}^d \}$

Remark: $\sigma^2$ is known and $\underset{n \times d}{\Phi}$ is known and deterministic

The last expression can be written

$$\inf_{A} \sup_{\beta^* \in \mathbb{R}^d} \mathbb{E}_{\varepsilon \sim N(0, \sigma^2 I_n)} \Big[R\big(A(\Phi\beta^* + \varepsilon)\big)\Big] - \sigma^2$$

only random term

We are going to take a "Bayesian" approach and lower bound the last expression by

$$\inf_{A} \mathbb{E}_{\beta^* \sim \pi} \mathbb{E}_{\varepsilon \sim N(0, \sigma^2 I_n)} \Big[R\big(A(\Phi\beta^* + \varepsilon)\big)\Big] - \sigma^2$$

$\hookrightarrow$ prior

We can pick any prior $\pi$. We pick a prior $\pi$ that is analytically convenient. Chose as prior $\pi$ the distribution $\beta^* \sim N\left(0, \frac{\sigma^2 I_n}{\lambda n}\right)$ where $\lambda > 0$.

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

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

Next, recall that

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

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

So the expression becomes

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

$$= \int_{\mathbb{R}^n} \int_{\mathbb{R}^d} \| A(y) - \beta^* \|^2_{\hat{\Sigma}} \; d\underbrace{P(\beta^* | y)}_{\substack{\text{posterior of} \\ \beta^* \text{ given } Y}} \; dP(y)$$

A standard calculation gives that the posterior is

Exercise

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

where $\hat{\beta}_\lambda = (\hat{\Sigma} + \lambda I_d)^{-1} \frac{\Phi^T \varepsilon}{n}$. Next,

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

$$\geq \inf_A \mathbb{E}_{\beta^* \mid Y}\left[\|A(Y) - \beta^*\|^2_{\hat{\Sigma}}\right]$$

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

because $\mathbb{E}_{\beta^* \mid Y}\left[\|A(Y) - \beta^*\|^2_{\hat{\Sigma}}\right]$ is minimized when

$A(Y) = \hat{\beta}_\lambda$, the posterior mean. Exercise!

Putting everything together, we have found the following lower bound for the minimax risk:

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

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

We have that

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

Exercise

Because $\beta^\star \perp\!\!\!\perp \varepsilon$ the expression is

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

$$= T_1 + T_2$$

We have that

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

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

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

Exercise

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

$$\sum_{j=1}^{d}\frac{\hat{\lambda}_j}{\hat{\lambda}_j+\lambda}$$

$\hat{\lambda}_j$ $j$-th eigenvalue of $\hat{\Sigma}$

$\hookrightarrow$ lower bound on the minimax risk.

The above bound holds for any $\lambda > 0$. Of course $\operatorname{tr}\left((\hat{\Sigma}+\lambda I_d)^{-1}\hat{\Sigma}\right)$ is $\downarrow$ in $\lambda$.

So the final lower bound:

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

$$\frac{\sigma^2}{n} \lim_{\lambda \downarrow 0} (\hat{\Sigma} + \lambda I_d)^{-1} \hat{\Sigma} =$$

Recall $\hat{\Sigma}$ is invertible by assumption

$$\frac{\sigma^2}{n} \operatorname{tr}(I_d) = \boxed{\sigma^2 \frac{d}{n}}$$

expected excess risk of $\hat{\beta}$ (OLS)

So $\hat{\beta}$ (OLS) is the minimax estimator

# STATISTICAL INFERENCE FOR $\beta^*$

- As usual the model is

$$Y = \Phi \beta^* + \varepsilon$$

full column rank and deterministic

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

<u>Goal</u>: statistical inference for $\beta^*$

- Is $\hat{\beta}$ (OLS) <u>consistent</u>, meaning $\hat{\beta} \xrightarrow{P} \beta^*$ ?

when $\hat{\beta}$ is computed using data $Y = \Phi\beta^* + \varepsilon$

Yes! To see this:

$$\hat{\beta} = (\Phi^T\Phi)^{-1}\Phi^T Y = \beta^* + \hat{\Sigma}^{-1}\frac{\Phi^T\varepsilon}{n}$$

<u>Claim</u>: if $\boxed{\hat{\Sigma} = \frac{\Phi^T\Phi}{n} \to \underset{d\times d}{\Sigma}}$ then

$$\hat{\Sigma}^{-1}\frac{\Phi^T\varepsilon}{n} \xrightarrow{P} 0 \quad \text{or} \quad \hat{\Sigma}^{-1}\frac{\Phi^T\varepsilon}{n} = o_p(1)$$

Pf/ By WLLN, $\frac{\Phi^T\varepsilon}{n} \xrightarrow{P} 0$ because

$$\frac{\Phi^T\varepsilon}{n} = \sum_{i=1}^{n}\frac{\Phi_i\varepsilon_i}{n}$$ where $\Phi_i$ is transpose of $i$th row of $\Phi$

and $\Phi_i\varepsilon_i \sim (0, \sigma^2\Phi_i\Phi_i^T)$

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

So by Chebyshev $\frac{\Phi^T\varepsilon}{n} \xrightarrow{P} 0$

Next $\hat{\Sigma}^{-1} \to \Sigma^{-1}$ so by Slutsky's theorem

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

∎

Also $\hat{\beta}$ is asymptotically normal:

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

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

We have that $\sqrt{n}(\hat{\beta}-\beta^*) = \underbrace{\hat{\Sigma}^{-1}}_{\to \Sigma^{-1}} \sqrt{n} \frac{\Phi^T \varepsilon}{n}$

So we only need to show that $\sqrt{n} \frac{\Phi^T \varepsilon}{n} \xrightarrow{d} N(0, \sigma^2 \Sigma)$

and the result will follow by Slutsky

- Next time we will going to verify this using the LF conditions, which holds if

$$\max_i \frac{\|\Phi_i\|}{\sqrt{n}} \to 0$$