SDS 387
Linear Models

Fall 2024

Lecture 23 - Tue, Nov 19, 2024

Instructor: Prof. Ale Rinaldo

- Last time: linear regression with random design and well-specified model

$$(\Phi_1, Y_1), \ldots, (\Phi_n, Y_n) \overset{iid}{\sim} P_{\Phi, Y}$$

$$Y_i = \Phi_i^T \beta^* + \varepsilon_i$$

$\Sigma = \mathbb{E}[\Phi_i \Phi_i^T]$ assumed invertible

$\hookrightarrow \varepsilon_1, \ldots, \varepsilon_n \mid \Phi_1, \ldots, \Phi_n \overset{iid}{\sim} (0, \sigma^2)$

- Exercise: what happens when the linear model is not true?

- We saw that the expected excess risk of $\hat{\beta}$ (OLS) is

$$\sigma^2 \, \mathbb{E}_{\Phi_1, \ldots, \Phi_n}\left[\operatorname{tr}\left(\Sigma \hat{\Sigma}^{-1} \right) \right]$$

and

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \Phi_i \Phi_i^T$$

which is assumed to be invertible with prob. 1.

- Question: is this optimal?

Theorem 1 of Mourtada (AOS 2022, 20(4), 2157-2178)

i) Assume that either $d > n$ or the distribution of the Φ_i's is degenerate (supported on an affine subspace of $\mathbb{R}^d$)

Then the minimax risk is infinity!

ii) If $n \geq d$ and the distribution of the Φ_i's is not degenerate ($\hat{\Sigma}$ is invertible wp 1) then

$$\inf_{\tilde{\beta}} \sup_{\beta^*} \mathbb{E}\left[R(\tilde{\beta})\right] = \frac{\sigma^2}{n} \mathbb{E}\left[tr(\Sigma \hat{\Sigma}^{-1})\right]$$

all estimators of β^*

excess risk

OLS estimator is minimax optimal

Above, the expectation is wrt to the distribution of $\Phi_1, \ldots, \Phi_n$ and $\varepsilon_1, \ldots, \varepsilon_n$

Pf/ Very similar to the proof of minimax optimality of OLS in fixed design setting.

$$\inf_{\tilde{\beta}} \sup_{\beta^*} \mathbb{E}_{\substack{\Phi_1, \ldots \Phi_n \\ \varepsilon_1, \ldots \varepsilon_n}} \left[R(\tilde{\beta})\right] \geq \inf_{\tilde{\beta}} \sup_{\beta^*} \mathbb{E}_{\Phi_1, \ldots, \Phi_n} \left[R(\tilde{\beta})\right]$$

$\varepsilon_1, \ldots, \varepsilon_n \mid \Phi_1, \ldots, \Phi_n \overset{iid}{\sim} N(0, \sigma^2)$

We replace $\sup_{\beta^*}$ with an average, assuming a prior distribution for $\beta^* \sim N_d(0, \frac{\sigma^2}{n\lambda} I_d)$ some $\lambda > 0$ to be chosen later

Using the same calculations in the fixed design setting, the minimax risk is lower bounded by the Bayes risk

$$\frac{\sigma^2}{n} \, \mathbb{E}_{\Phi_1, \dots, \Phi_n} \left[\mathrm{tr}\left(\left(\hat{\Sigma} + \lambda I_d \right)^{-1} \Sigma \right) \right] \qquad \forall \lambda > 0$$

Case i): $\hat{\Sigma}$ is not invertible with prob 1 (the degenerate case)

eg $d > n$ or the distribution of the Φ_i's is supported on a affine subspace

Then

$$\mathrm{tr}\left(\Sigma^{1/2} \left(\hat{\Sigma} + \lambda I_d \right)^{-1} \Sigma^{1/2} \right) \geq \overbrace{\lambda_{\max}}^{\text{largest eigenvalue}} (\cdot)$$

$$\bigstar = \frac{1}{\lambda_{\min}\left(\Sigma^{-1/2} \left(\hat{\Sigma} + \lambda I_d \right) \Sigma^{-1/2} \right)}$$

Since $\hat{\Sigma}$ is not invertible wp 1, $\exists \, u \in \underbrace{\mathbb{S}^{d-1}}_{\text{d-dimensional unit sphere}}$ possibly (random)

s.t. $u^T \Sigma^{-1/2} \left(\hat{\Sigma} + \lambda I_d \right) \Sigma^{-1/2} u = 0 + \lambda \| \Sigma^{-1/2} u \|^2$

u is chosen so that $\Sigma^{-1/2} u$ is in the null space of $\hat{\Sigma}$

$$\leq \lambda \, \lambda_{\max}\left(\Sigma^{-1} \right)$$

$$= \frac{\lambda}{\lambda_{\min}(\Sigma)}$$

So $\bigstar \geq \dfrac{\lambda_{\min}(\Sigma)}{\lambda}$

So the minimax risk is lower bounded by

$$\frac{\sigma^2}{n} \frac{d_{min}(\Sigma_i)}{\lambda} \qquad \forall \lambda > 0$$

Letting $\lambda \to 0$ the minimax lower bound $\to \infty$

in) If $\hat{\Sigma}$ is invertible with prob 1 then $tr\left((\hat{\Sigma} + \lambda I_d)^{-1} \Sigma \right)$ is $\downarrow$ in λ and converges to $tr(\hat{\Sigma}^{-1} \Sigma)$ as $\lambda \to 0$.

By monotone convergence theorem the minimax lower bound is $\frac{\sigma^2}{n} \mathbb{E}\left[tr(\hat{\Sigma}^{-1} \Sigma) \right]$ ∎

Mourtada's Corollary 2: if $n \geq d$ and the distribution of the Φ_i's is not degenerated you can further lower bound this by

$$\sigma^2 \frac{d}{n-d-1} \qquad n > (d+1)$$

If the covariates $\Phi_1, \ldots, \Phi_n \overset{iid}{\sim} N(0, I_d)$ this value is the exact minimax risk.

∎ AN EXACT ANALYSIS OF THE RISK UNDER GAUSSIAN SETTINGS

Breiman & Freedman (1983) JASA 78 (381)

Assume all the conditions above and furter suppose that $\Phi_1, \dots, \Phi_n \overset{iid}{\sim} N(0, I_d)$. Then the risk of LS is

$$\frac{\sigma^2}{n} \mathbb{E}\left[tr\left(\hat{\Sigma}^{-1} \underbrace{\Sigma}_{I_d} \right) \right] = \sigma^2 \mathbb{E}\left[\left(\underbrace{\sum_{i=1}^n \Phi_i \Phi_i^T}_{} \right)^{-1} \right]$$

Wishart distribution with parameter I_d and n degrees of freedom

its inverse is called the inverse Wishart

$$= \begin{cases} \dfrac{\sigma^2 \, tr(I_d)}{n-d-1} = \sigma^2 \dfrac{d}{n-d-1} & \text{if } n > d+1 \\ \infty & \text{if } n = d \text{ or } d+1 \end{cases}$$

At interpolation ($n = d$ or $n=d+1$) the risk explodes.

- Now, under the same settings, something interesting happens when $d > n$. See Belkin, Hsu and Xu (2020) SIAM Journal of Mathematics & Data Science

You would expect "bad" risk behavior.

Surprisingly the risk is stable and in fact it can be smaller than when $d \le n$.

When $d > n$ we will fit the min-norm estimator:

$$\hat{\beta}_{MN} = \Phi^{+} Y = (\Phi^T \Phi)^{+} \Phi^T Y = \Phi^T \underbrace{(\Phi \Phi^T)^{-1}}_{\substack{n \times n \\ \text{invertible}}} Y$$

$\underset{n \times d}{\Phi}$ with rows Φ_i^T $\qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

Let's compute the expected excess risk $\mathbb{E}\left[\|\hat{\beta}_{MN} - \beta^*\|^2_{I_d}\right]$

The first thing to notice there is a bias term:

$$\beta^* - \hat{\beta}_{MN} = \beta^* - \Phi^T (\Phi\Phi^T)^{-1} \Phi \beta^* - \Phi^T (\Phi \Phi^T)^{-1} \varepsilon$$

$$= \underbrace{\left(I_d - \underbrace{\Phi^T (\Phi \Phi^T)^{-1} \Phi}_{\Pi}\right)}_{\substack{I - \Pi \text{ orthogonal projection} \\ \text{onto the null space of} \\ \underset{n \times d}{\Phi}}} \beta^* - \Phi^T (\Phi \Phi^T)^{-1} \varepsilon$$

So

random projection matrix

$$\mathbb{E}\left[\|\beta^* - \hat{\beta}_{MN}\|^2\right] = \mathbb{E}_{\Phi}\left[\|(I - \Pi)\beta^*\|^2\right] + \mathbb{E}_{\Phi}\left[\operatorname{tr}\left((\Phi\Phi^T)^{-1} \cancel{\Phi\Phi^T} \cancel{(\Phi\Phi^T)^{-1}} \underbrace{\mathbb{E}_{\varepsilon \mid \Phi}\left[\varepsilon \varepsilon^T\right]}_{\sigma^2 I_n}\right)\right]$$

$$= \mathbb{E}_{\Phi}\left[\|(I - \Pi)\beta^*\|^2\right] + \mathbb{E}_{\Phi}\left[(\Phi\Phi^T)^{-1}\right]$$

$$= T_1 + T_2$$

Next,

$$T_1 = \|\beta^*\|^2 - \mathbb{E}\left[\|\Pi\beta^*\|^2\right]$$

To compute $\mathbb{E}\left[\|\Pi\beta^*\|^2\right]$ we will use the fact that if $Z \sim N(0, I)$ then $UZ \sim N(0, I)$ where $\underset{d\times d}{U}$ is an orthogonal matrix. Let $U_1, U_2, \dots, U_d$ be $d \times d$ orthogonal matrices s.t. $U_i\beta^* = \|\beta\| e_i$

$\downarrow$ ith standard basis vector

So, $\forall i = 1, \dots, d$

$$\|\Pi\beta^*\|^2 = \beta^{*T}\Phi^T(\Phi\Phi^T)^{-1}\Phi\beta^* \overset{d}{=} \beta^{*T}U_i^T\Phi^T(\Phi U_i U_i^T\Phi^T)^{-1}\Phi U_i\beta^*$$

$$= \|\beta^*\|^2 e_i^T \Phi^T(\Phi\Phi^T)^{-1}\Phi e_i$$

$$= \|\beta^*\|^2 \operatorname{tr}\left(\Phi^T(\Phi\Phi^T)^{-1}\Phi e_i e_i^T\right)$$

So

$$\mathbb{E}\left[\|\Pi\beta^*\|^2\right] = \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\left[\|\beta^*\|^2 \operatorname{tr}\left(\Phi^T(\Phi\Phi^T)^{-1}\Phi e_i e_i^T\right)\right]$$

$$= \|\beta^*\|^2 \frac{1}{d}\,\mathbb{E}\left[\operatorname{tr}\left(\Phi^T(\Phi\Phi^T)^{-1}\Phi \underbrace{\sum_{i=1}^{d} e_i e_i^T}_{I_d}\right)\right]$$

$$= \|\beta^*\|^2 \frac{1}{d}\,\mathbb{E}\left[\operatorname{tr}\left(\underbrace{\cancel{(\Phi\Phi^T)^{-1}}\,\cancel{\Phi\Phi^T}}_{I_n}\right)\right]$$

$$= \|\beta^*\|^2 \frac{n}{d}$$