SDS 387
Linear Models

Fall 2024

Lecture 19 - Tue, Nov 5, 2024

Instructor: Prof. Ale Rinaldo

- Progress report extension: now due on Nov 15.

- Recall we are considering the well-specified, fixed-design linear regression setting:

$$\underset{n\times 1}{Y} = \underset{n\times d}{\Phi}\,\underset{d\times 1}{\beta^*} + \underset{n\times 1}{\varepsilon} \qquad \varepsilon \sim (0, \sigma^2 I_n)$$

meaning: has mean 0 and covariance $\sigma^2 I_n$

- Last time we looked at ridge regression estimator:

$$\hat{\beta}_\lambda = \left(\hat{\Sigma} + \lambda I_d\right) \frac{\Phi^\top Y}{n} \qquad \underset{\text{tuning parameter}}{\lambda > 0}$$

$$\hat{\Sigma} = \Phi^\top\Phi / n$$

- Numerically it can be more stable than OLS and, statistically, it provides a different trade-off between bias and variance

- We saw that the expected excess risk of $\hat{\beta}_\lambda$ is

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

bias term $\uparrow \lambda$ variance $\downarrow \lambda$

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

When $\lambda = 0$ this reduces to $\sigma^2 \frac{d}{n}$ the risk of OLS.

Question: what is the optimal λ?

<u>Prop 3.8 in Bach's book</u>

Setting $\lambda_{optimal} = \dfrac{\sigma \sqrt{\operatorname{tr}(\hat{\Sigma})}}{\|\beta^*\| \sqrt{n}}$ we get that

$$\mathbb{E}\left[R(\hat{\beta}_\lambda)\right] - \sigma^2 \leq \frac{\sigma \sqrt{\operatorname{tr}(\hat{\Sigma})} \, \|\beta^*\|}{\sqrt{n}}$$

<u>Remark</u>: compare this with $\sigma^2 \frac{d}{n}$ (the risk of OLS)

i) If d is <u>fixed</u> and $n \to \infty$ then OLS is better because it vanishes at a rate $O(1/n)$ while the risk of ridge vanishes at slower rate $O\left(\frac{1}{\sqrt{n}}\right)$

When d changes (increases with n) and $\hat{\Sigma}$, σ and $\|\beta^*\|$ also change with n, then ridge can be better than OLS.

ii) this is not necessarily the best choiche of λ. It is the best choice for a simpler upper bound on the risk.

iii) in practice how do you choose λ?

Use <u>cross-validation</u>.

Pf/ First, the eigenvalues of $(\hat{\Sigma} + \lambda I_d)^{-2} \lambda \hat{\Sigma} \leq 1/2$.

To see this, the eigenvalues of this matrix are of the form

$$\frac{\hat{\lambda}_i \lambda}{(\hat{\lambda}_i + \lambda)^2}$$

where $\hat{\lambda}_1, \dots, \hat{\lambda}_d$ are eigenvalues of $\hat{\Sigma}$

which is always $\leq 1/2$ because

$$\frac{ab}{(a+b)^2} \leq 1/2 \qquad a, b > 0$$

So, using this fact, the bias term is

$$\lambda \, \beta^{*T} (\hat{\Sigma} + \lambda I_d)^{-2} \lambda \hat{\Sigma} \beta^* \leq \frac{\lambda}{2} \|\beta^*\|^2$$

Exercise

As for the variance,

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

$$\overset{\text{Exercise}}{\leq} \frac{\sigma^2}{n\lambda} \underbrace{\| \hat{\Sigma} \lambda (\hat{\Sigma} + \lambda I_d)^{-2} \|_{op}}_{\leq 1/2} \operatorname{tr}(\hat{\Sigma})$$

$$\leq \frac{\sigma^2}{n\lambda} \operatorname{tr}(\hat{\Sigma})$$

So, the expected excess risk of $\hat{\beta}_\lambda$ is upper bounded by:

$$\frac{\lambda}{2} \|\beta^*\|^2 + \frac{\sigma^2 \operatorname{tr}(\hat{\Sigma})}{2n\lambda} := a\lambda + \frac{b}{\lambda}$$

$a, b > 0$

This is minimized at $\lambda = \sqrt{b/a}$, with optimal value $\sqrt{2ab}$. ∎

LOWER BOUND ON RISK OF OLS

Section 3.7 in Bach's book + paper by Mourtada

Recall that the expected excess risk of $\hat{\beta}$ (OLS) is $\sigma^2 \frac{d}{n}$ (assuming fixed covariates and a well-specified model).

Here we show that this value is optimal in a minimax sense

- Suppose we are interested in estimate a parameter θ^*, generally defined as a function(al) of a probability distribution, say P. To highlight this fact, we write $\theta^*(P)$ [Note: θ^* does not need to identify or fully specify P.] We also specify a collection, say $\mathcal{P}$, of probability distributions whose parameter θ^* is of interest. We observe data (an iid sequence of length, say, n) and construct an estimator, $\hat{\theta}_n$ of θ^* [we do not know which P in $\mathcal{P}$ has generated the data, so we do not know $\theta^* = \theta^*(P)$].
  We measure the quality of $\hat{\theta}_n$ by its risk
  $$R\left(\theta^*(P), \hat{\theta}_n\right)$$

Example: i) $\mathcal{P}$: set of distributions of Y, where

$$Y = \Phi \beta^* + \varepsilon \rightarrow \varepsilon \sim (0, \sigma^2)$$

Φ: fixed and known; β^*: unknown; σ^2: known

For each $P \in \mathcal{P}$
$\theta^*(P) = \beta^*(P)$

iii) $\mathcal{P}_{Gauss}$: $Y \sim N_n(\Phi \beta^*, \sigma^2 I_n)$

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

for $P \in \mathcal{P}_{Gauss}$ $\quad \theta^*(P) = \beta^*(P)$

Of course $\mathcal{P}_{Gauss} \subset \mathcal{P}$

Of course for any estimator $\tilde{\beta}$ of β^* the risk is

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

<u>Remark</u> think of the risk $R(\theta^*, \hat{\theta}_n)$ as a function of θ^*.

- How do we use this setting to evaluate whether an estimator is good or optimal?
- The minimax approach requires you to evaluate the minimax risk and find an estimator that at least asymptotically, achieves this risk value.
- The minimax risk is

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} R(\hat{\theta}_n, \theta^*(P))$$

where $\inf_{\hat{\theta}_n}$ is the infimum over all estimators (functions of the data)

Minimax risk measures intrinsic statistical hardness of an estimation task (or any statistical task)

For regression, this translates into

$$\inf_{\tilde{\beta}} \sup_{P \in \mathcal{P}} \left(\mathbb{E}_P \left[R(\tilde{\beta}) \right] - \sigma^2 \right)$$

- An estimator $\hat{\theta}_n$ is minimax rate-optimal if asymptotically as $n \to \infty$, its risk is of the same order as the minimax risk. Formally, let

  $R_{n,\text{minimax}}$ the value of the minimax risk and for an estimator $\hat{\theta}_n$, let $R_n^{\sup}(\hat{\theta}_n) \triangleq \sup_{\theta^*} R(\hat{\theta}_n, \theta^*)$

  Then $\hat{\theta}_n$ is minimax rate optimal if

  $$\limsup_n \underbrace{\frac{R_n^{\sup}(\hat{\theta}_n)}{R_{n,\text{minimax}}}}_{\geq 1} \leq C$$

- $\hat{\theta}_n$ is sharp minimax rate optimal if $C = 1$

- $\hat{\theta}_n$ is exact minimax optimal if

  $$R_n^{\sup}(\hat{\theta}_n) = R_{n,\text{minimax}} \quad \text{all } n.$$

<u>Thm</u> $\hat{\beta}$ (OLS) is exact minimax optimal for $\mathcal{P}_{\text{Gauss}}$

$\rightarrow$ $\hat{\beta}$ (OLS) is minimax optimal for $\mathcal{P}$

- For regression, we are interested in computing a lower boun on this quantity:

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

algorthm taking as input Y and Φ → fixed

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