SDS 387
Linear Models

Fall 2024

Lecture 14 - Thu, Oct 15, 2024

Instructor: Prof. Ale Rinaldo

- Announcement: no class on Tue, Oct 22.
- Q2 in HW3: I will clarify the question and update the HW after class

Projection of random variables

(Chapter 11 of Van der Vaart's book on Asymptotic statistics)

Let T and $\{S, S \in \mathcal{S}\}$ (collection of r.v.'s) be random variables with finite second moments.

We want to "projectu" T on $\mathcal{S}$. A r.v. $\hat{S} \in \mathcal{S}$ is a L_2-projection of T onto $\mathcal{S}$ when $\hat{S}$ minimizes

space of all r.v.'s that have finite variance

$$S \in \mathcal{S} \longmapsto \mathbb{E}\left[(T-S)^2\right]$$

$\hat{S}$ is not necessarily unique!

- <u>Aside</u> : think of L_2 (the space of all r.v.'s with finite 2 moments) as a Hilbert space wrt inner product

$$S_1, S_2 \in L_2 \quad \to \quad \langle S_1, S_2 \rangle = \mathbb{E}[S_1 S_2]$$

<u>Note</u>: this is not a space of r.v.'s but of equivalent classes of random variables, where 2 r.v.'s are in the same equivalence class when they are identical with prob. 1.

If $\mathcal{S}$ is a vector space (closed wrt to addition and scalar multiplication) then $\hat{S}$ is the projection of T onto $\mathcal{S}$ iff $T - \hat{S}$ and S are orthogonal

i.e. $\mathbb{E}[(T-\hat{S})S] = 0 \quad \forall S \in \mathcal{S}$:

<u>Thm 11.1</u> $\hat{S}$ is the projection of T onto $\mathcal{S}$ iff

i) $\hat{S} \in \mathcal{S}$ and ii) $\mathbb{E}[(T-\hat{S})S] = 0 \quad \forall S \in \mathcal{S}$

The projection is unique (in the sense that if $\hat{S}'$ is another projection then $P(\hat{S} \neq \hat{S}') = 0$).

if $\mathcal{S}$ contains the constant functions, then

$$\mathbb{E}[\hat{S}] = \mathbb{E}[T] \quad \text{and} \quad \operatorname{cov}(T-\hat{S}, S) = 0 \quad \forall S \in \mathcal{S}.$$

Pf/ The condition $\mathbb{E}[(T-\hat{S})S] = 0$ is called <u>orthogonality</u>

Assume orthogonality. Then, $\forall S \in \mathcal{S}$,

$$\mathbb{E}[(T-S)^2] = \mathbb{E}[(T-\hat{S})^2] + 2\underbrace{\mathbb{E}[(T-\hat{S})(\hat{S}-S)]}_{=0 \text{ by orthogonality because } \hat{S}-S \in \mathcal{S}} + \mathbb{E}[(\hat{S}-S)^2]$$

$$\geq \mathbb{E}\left[(T-\hat{S})^2\right]$$

Above, we have an equality iff $\mathbb{E}\left[(\hat{S}-S)^2\right]=0$ iff $\mathbb{P}(\hat{S}=S)=1$.

Conversely suppose $\hat{S}$ is a projection. Then $\forall \alpha \in \mathbb{R}$

$$0 \leq \mathbb{E}\left[(T-\hat{S}-\alpha S)^2\right] - \mathbb{E}\left[(T-\hat{S})^2\right] = \alpha^2\, \mathbb{E}[S^2] - 2\alpha\, \mathbb{E}[(T-\hat{S})S]$$

As a function of α, the RHS is a parabola that has to stay above the x-axis. The zeros of this parabola are $\alpha=0$ and $\alpha = \frac{2\,\mathbb{E}[(T-\hat{S})S]}{\mathbb{E}[S^2]}$

$$\hookrightarrow \quad \mathbb{E}[(T-\hat{S})S]=0 \quad \forall S \in \mathcal{S}$$

Furthermore, if $\mathcal{S}$ contains the constan r.v.'s then by orthogonality

$$\mathbb{E}\left[(T-\hat{S})\cdot c\right]=0 \quad \forall c$$

$$\hookrightarrow \quad \mathbb{E}[T] = \mathbb{E}[\hat{S}]$$

■

Corollary Pythagore theorem for r.v.'s

$$\mathbb{E}[T^2] = \mathbb{E}[\hat{S}^2] + \mathbb{E}\left[(T-\hat{S})^2\right]$$

- Arguably the most important type of L_2-projection is the conditional expectation. Suppose we have 2 r.v.'s, X and Y, with finite 2^{nd} moments. I want to approximate or predict Y using X. Formally, I want to find the function g s.t. $\mathbb{E}[g^2(X)] < \infty$ and

$$\mathbb{E}\left[(Y - g(X))^2\right] \leq \mathbb{E}\left[(Y - f(X))^2\right]$$

over all (measurable) functions f s.t. $\mathbb{E}\left[f^2(X)\right] < \infty$

In this case $S = \{ f(X),\ f \text{ measurable and } \mathbb{E}[f^2(X)] < \infty \}$

It turns out that

$$g(X) = \mathbb{E}[Y \mid X]$$

you can see this by verifying orthogonality:

$$\mathbb{E}\left[\left(Y - \mathbb{E}[Y|X]\right) f(X)\right] = \mathbb{E}\left[Y \cdot f(X)\right] - \mathbb{E}\left[\mathbb{E}[Y|X] f(X)\right]$$

$\forall f$ $\qquad = 0$

by law of iterated expectation

aka tower property of conditional expectation

- Aside: a more direct way to see this, without ortogonality is to use the fact that

$$\underset{c \in \mathbb{R}}{\arg\min}\ \mathbb{E}\left[(X - c)^2\right] = \mathbb{E}[X].$$

So,

$$\mathbb{E}\left[(Y - f(X))^2\right] \geq \mathbb{E}\left[\left(Y - \mathbb{E}[Y|X]\right)^2\right].$$

$\forall f$ Exercise!

- <u>Remark</u>: the conditional expectation is well-defined even without a second moment!

This is the more general measure-theoretic definition of conditional expectation

LINEAR MODELS

(I will follow Bach's book draft for the next several lectures)

- General regression setting: let Y is a univariate random variable called the response variable. Let $X \in \mathbb{R}^d$ a random vector of covariates or features or explanatory variables

Our goal is to "learn" about Y using X.

- In its most general form "learn" refers to learning the regression function: not random

$$x \in \mathbb{R}^d \longmapsto \mathbb{E}[Y \mid X = x]$$

- Assuming that Y and X have finite second moments this can be cast as the problem of minimizing

$$\mathbb{E}[(Y - f(X))^2]$$

over all f (with 2nd moment)

↓

MSE (mean squared error)

↓

prediction task!

- There are specific instances of this problem, introduced below in increasing order of generality

  - linear regression models

$$\mathbb{E}[Y \mid X = x] = \alpha + \beta^T x \quad \text{some } \alpha \in \mathbb{R},\ \beta \in \mathbb{R}^d$$

Note that

$$\mathbb{E}[Y \mid X = x] = \alpha + \beta^T \varphi(x) \quad \text{also linear}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^{d'}$

is a feature mapping

For example, polynomial regression:

$$\mathbb{E}[Y \mid X = x] = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

- <u>non-parametric regression</u>:

$$Y = \underset{\text{some unknown } f}{f(X)} + \underset{\text{error or noise}}{\varepsilon} \quad \text{where} \quad \mathbb{E}[\varepsilon \mid X] = 0$$

(in fact, often times $\varepsilon \perp\!\!\!\perp X$) $\mathbb{E}[\varepsilon] = 0$

without any assumptions on f this is a hopeless task. Typically f is assumed to belong to a class of well-behaved (ie smooth) functions.

- <u>most general form of regression</u> (assumption free setting)

$$Y = \mathbb{E}[Y \mid X] + \underbrace{Y - \mathbb{E}[Y \mid X]}_{\varepsilon}$$

<u>Note</u>: $\mathbb{E}[\varepsilon \mid X] = 0$

but, of course $\varepsilon \not\perp\!\!\!\perp X$

While you may not be able to estimate $\mathbb{E}[Y \mid X]$ you may be interested in approximating it using a simple, e.g. linear, model! → mis-specified model

To summarize, the model can be

linear $\mathbb{E}[Y \mid X = x] = x^T\beta$

non-linear $\mathbb{E}[Y \mid X = x] = f(x)$

mis-specified approximate $\mathbb{E}[Y \mid X]$ with a simple model

The covariate X can be treated as

- deterministic : X is not random
- random : X is random

The error term $\varepsilon = Y - \mathbb{E}[Y \mid X]$ can be

- parametric $\varepsilon \sim N_d(0, \sigma^2 I_d)$ → homoschedastic

  $N_d(0, \Sigma)$

  ↳ heteroschedastic

- $\varepsilon \perp\!\!\!\perp X$
- $\varepsilon \not\perp\!\!\!\perp X$