SDS 387
Linear Models

Fall 2024

Lecture 4 - Thu, Sep 5, 2024

Instructor: Prof. Ale Rinaldo

L_p Convergence

A sequence of r.v.'s $\{X_n\}$ converges in L_p, $p \geq 1$, to a r.v. X when

$$\|X_n - X\|_p \to 0 \quad \text{as} \quad n \to \infty$$

where, for any r.v. Y, is

$$\left(\mathbb{E}\left[|Y|^p\right]\right)^{1/p}$$

→ it is a norm over p-integrable r.v.'s when $p \geq 1$

- When $p = 2$ this yields the mean squared error criterion.

Specifically, suppose you are interested in estimating a parameter θ using an estimator $\hat{\theta}_n$ (a function of your data).

θ_n is consistent when $\hat{\theta}_n \xrightarrow{P} \theta$

1

The mean-squared error criterion is:

$$\|\hat{\theta}_n - \theta\|_2^2 = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \underbrace{\left(\theta - \mathbb{E}[\hat{\theta}_n]\right)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\right)^2\right]}_{\text{variance}}$$

- Remark: we can extend L_p convergence to the case of $p = \infty$ by defining:

a r.v.

$$\|Y\|_\infty = \inf \{ c \in \mathbb{R} : \mathbb{P}(Y > c) = 0 \}$$

↳ essential supremum

- Few results (see any book on probability / analysis for proofs):

i) Minkowski ineq:

$$\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$$

[so if $X, Y \in L_p$ so if $X+Y$]

you can show this using using the "Cr - inequality":

if $x, y \in \mathbb{R}$

$$|x+y|^r \leq \begin{cases} 2^{r-1} (|x|^r + |y|^r) & r \geq 1 \\ |x|^r + |y|^r & 0 < r < 1 \end{cases}$$

ii) Hölder inequality:

$X \in L_p$ $Y \in L_q$ p, q are conjugate $\frac{1}{p} + \frac{1}{q} = 1$

then

$XY \in L_1$ and $\|XY\|_1 \leq \|X\|_p \|Y\|_q$

The notable case of $p = q = 2$ is known as Cauchy - Schwartz inequality:

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]} \sqrt{\mathbb{E}[Y^2]}$$

[Aside if $x, y \in \mathbb{R}^d$ $|x^T y| \leq \|x\| \, \|y\|$ also Cauchy Schwartz]

The proof uses the inequality:

$a, b \in \mathbb{R}$ $$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q}$$ p, q conjugate

and Jensen's inequality: if $f : \mathbb{R} \to \mathbb{R}$ is convex

then $$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$ provided $\mathbb{E}[f(X)]$ exists.

[If g is concave $g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)]$]

because f is convex iff $-f$ is concave

Application of Jensen's inequality: if $1 \leq p \leq q$ then
$$\|X\|_p \leq \|X\|_q$$

Pf/
$$\mathbb{E}[|X|^p] = \mathbb{E}\left[|X|^{p \frac{q}{p} \frac{p}{q}}\right]$$

Jensen $\leftarrow$ $$\leq \left(\mathbb{E}\left[|X|^{p\frac{q}{p}}\right]\right)^{p/q}$$ because $x \to x^{p/q}$, $x \geq 0$ is concave

$$= \left(\mathbb{E}\left[|X|^q\right]\right)^{p/q}$$

$\hookrightarrow$ $$\|X\|_p \leq \|X\|_q \quad \blacksquare$$

- L^p convergence implies convergence in probability.

This follows from Markov's inequality:

[if $X \geq 0$ then $\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon}$ $\forall \varepsilon > 0$]

So if $\{X_n\}$ is a sequence of r.v.'s and X a r.v.

then $\forall \varepsilon > 0$
$$\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \frac{\|X_n - X\|_p^p}{\varepsilon^p} \to 0$$

$$\Big[\quad X_1, X_2 \dots \overset{iid}{\sim} (\mu, \sigma^2)$$

↳ iid from some distribution with mean μ and variance σ^2

Claim $\quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu \qquad$ WLLN

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \le \frac{\mathbb{E}[(\bar{X}_n - \mu)^2]}{\varepsilon^2} = \frac{\mathrm{Var}[\bar{X}_n]}{\varepsilon^2}$$

↓ Chebyshev inequality

$$= \frac{\sigma^2}{n} \frac{1}{\varepsilon^2} \to 0 \text{ as } n \to \infty \Big]$$

- L_p converge does not imply or is implied by convergence wp 1
- Examples:

i) The typewriter sequence: it converges in L_p $\forall p \ge 1$ but not wp 1

ii) Let $U \sim \text{Uniform}(0,1)$ and for each n let

$$X_n = f_n(U) = \begin{cases} 0 & \text{if } 0 < U \le 1/n \\ \frac{1}{U} & \text{otherwise} \end{cases}$$

So $X_n \to \frac{1}{U}$ wp 1 but $\frac{1}{U} \notin L_p$ $\forall p \ge 1$

or

$$X_n = f_n(U) = \begin{cases} n & 0 < U \le 1/n \\ 0 & \text{otherwise} \end{cases}$$

$$\|X_n\|_p^p = n^{p-1} \to \infty \text{ for } p > 1$$

but $X_n \xrightarrow{wp\,1} 0$

Convergence in distribution (aka weak convergence)

This is the weakest form of stochastic convergence.

- Recall that the cdf (cumulative distribution function) of a random variable X over $\mathbb{R}$ is the function

$$x \in \mathbb{R} \longmapsto \mathbb{P}(X \le x) = F(x)$$

It has the following properties:

i) it is non-decreasing

ii) it is right-continuous with left limits

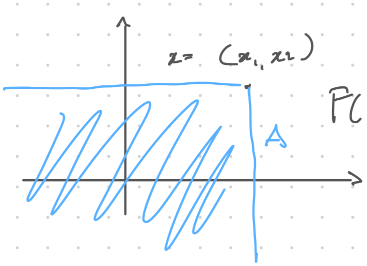
$$\lim_{y \downarrow x} F(y) = F(x) \qquad \lim_{y \uparrow x} F(y) \text{ exists } (:= F(x^-))$$

iii) $\lim_{x \to -\infty} F(x) = 0 \qquad \lim_{x \to +\infty} F(x) = 1$

In fact, any function over $\mathbb{R}$ with these properties defines a probability distribution over $\mathbb{R}$.

- In $\mathbb{R}^d$ the notion of cdf is analogous. The cdf F_X of a random vector X in $\mathbb{R}^d$ is the function

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d \longrightarrow F(x) = \mathbb{P}\left(\bigcap_{j=1}^{d} \{ X_j \le x_j \} \right)$$



$$F(x) = \mathbb{P}(X \in A)$$

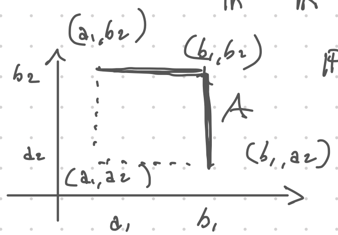
In fact, properties i), ii) and iii) still hold provided that, for $x, y \in \mathbb{R}^d$, $x \le y$ means $x_j \le y_j \ \forall j$

A function F on $\mathbb{R}^d$ satisfying properties i), ii) and iii) does not necessarily define a probability distribution on $\mathbb{R}^d$. We need another property:

Let $A = \prod_{j=1}^{d} (a_j, b_j]$ be a rectangle in $\mathbb{R}^d$

$-\infty < a_j < b_j < \infty \quad \forall j$

In $\mathbb{R}^2$ if F is the cdf of a random vector, say X,

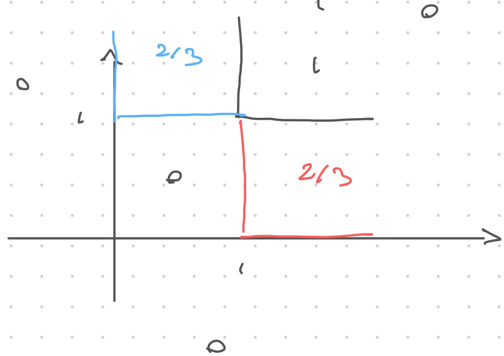


$$\mathbb{P}(X \in A) = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2) \geq 0 \quad (*)$$

If F satisfies properties i), ii) and iii), it is not true that an expression like the one above is non-negative. Example:

Rick Durrett
Probability theory
& examples
5th edition

$$F(x_1, x_2) = \begin{cases} 1 & x_1, x_2 \geq 1 \\ 2/3 & x_1 \geq 1 \quad 0 \leq x_2 \leq 1 \\ 2/3 & x_2 \geq 1 \quad 0 \leq x_1 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



For $\varepsilon \in (0, 1)$ expression $(*)$ evaluates to

$$1 - \frac{2}{3} - \frac{2}{3} + 0 = -1/3 \quad \text{when } \begin{matrix} a_1 = a_2 = 1 - \varepsilon \\ b_1 = b_2 = 1 \end{matrix}$$

To fix this we make the assumption that, for any rectangle $A = \prod_{j=1}^{d} (a_j, b_j]$, $-\infty < a_j < b_j < +\infty$ $\forall j$

= def of A,

iv) $\Delta_A F := \sum_{v \text{ vertex of } A} \text{sgn}(v) F(v) \geq 0$, $\forall A$

2^d vertices

where, for a vertex v, $\text{sgn}(v) = (-1)^{\# a_j\text{'s in } v}$

Then properties i), ii), iii) and iv) guarantee that F defines a prob. distribution on $\mathbb{R}^d$.

Definition: A sequence $\{X_n\}$ of r.v. (or vectors) with c.d.f.'s $\{F_n\}$ converges in distribution to X, with cdf F, when for every $c \in \mathbb{R}^d$ s.t. F is continuous at c,

$$F_n(c) \to F(c) \quad \text{as } n \to \infty$$

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq c)$$

element-wise $\leq$