Contents

1 σ-algebra and independence

A σ-algebra $\mathcal{F}$ of a set Ω is a collection of subsets of Ω such that

1. $\emptyset \in \mathcal{F}$;
2. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$;
3. if $A_j \in \mathcal{F}$ for all $j = 1, 2, \ldots$, then $\bigcup_j^\infty A_j \in \mathcal{F}$.

σ-algebra embodies the set of information. A real-valued random variable is a measurable function $X : (\Omega, \mathcal{F}) \to \mathbb{R}$, meaning that for all $t \in \mathbb{R}$,

$$X^{-1}(-\infty, t] \in \mathcal{F}$$

We can find the smallest σ-algebra that makes X measurable, which is written as $\sigma(X)$. $\sigma(X)$ then can be thought of as quantifying how much we can learn about which event $E \subset \Omega$ has happened, based on an observation of X. For example, we can distinguish two events E_1, E_2, if for all $e_1 \in E_1$ and $e_2 \in E_2$, $X(e_1) > a > X(e_2)$. In this sense, if $\sigma(X) \subset \sigma(Y)$, we might interpret it as Y contains more information than X.

Independent random variables provide different sets of information, so it's good to have independence. Dependent data contain overlapping information, so they are less informative than independent data.

2 Asymptotics and Stochastic Orders

We recall some definitions in case some are not familiar. We say a sequence of real numbers $a_n = o(1)$, if $\lim_{n\to\infty} a_n = 0$. $a_n = o(b_n)$ if $b_n^{-1}a_n = o(1)$. We say $a_n = O(1)$ if there exists C such that for large enough N, $a_n \leq C$.

A sequence of random variables $(X_n) = o_p(1)$, if $X_n \xrightarrow{p} 0$, that is if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n| \geq \varepsilon) = 0$$

$X_n = o_p(Y_n)$ if there exists C_n such that $X_n = Y_n C_n$ and $C_n = o_p(1)$.

A sequence of random variable $X_n = O_p(1)$ if it satifies that for any $\varepsilon > 0$, there exists M such that

$$\sup_n P(|X_n| > M) < \varepsilon$$

We have that a single random variable X is always $O_p(1)$. Also, if X_n converges in distribution, then $X_n = O_p(1)$, so it's a necessary condition for convergence distribution. In fact, it's a fundamental condition called **tightness**.

Theorem 1 (Prokhorov's Theorem). *If $X_n = O_p(1)$, there exists a subsequence X_{n_j} and a random variable X, $X_n \rightsquigarrow X$.*

We write $X_n \rightsquigarrow X$ for convergence in distribution, also known as weak convergence, if for all bounded continuous real-valued functions f, $f(X_n) \to f(X)$.

3 Characteristic Functions

Characteristic functions of a random variable X is defined as $\phi_X(t) = Ee^{it'X}$. It makes a lot of calculations simpler.

Example 3.1. If $X \sim N(\mu, \Sigma)$, then $\phi_X(t) = \exp\left\{it'\mu - \frac{1}{2}t'\Sigma t\right\}$.

First we will establish that $\phi_X(t)$ uniquely determines the distribution of X. Notice that $e^{it'x}$ is a bounded continuous function of x, so if $X_n \rightsquigarrow X$, then $\phi_{X_n}(t) \to \phi_X(t)$.

Theorem 2. *Let X, Y be random vectors, X and Y have the same distribution iff $\phi_X(t) = \phi_Y(t)$.*

The following theorem provides a way to relate characteristic functions to asymptotics,

Theorem 3 (Levy's Continuity Theorem). *If $\phi_{X_n}(t) \to \phi(t)$ pointwise, and $\phi(t)$ is continuous at 0, then $\phi(t)$ is the Characteristic function of some random variable X and $X_n \rightsquigarrow X$.*

Proof. First show $X_n = O_p(1)$ using the continuity. Then the uniqueness of the limiting characteristic function will imply X_n converges in distribution. □

As a result, convergence of random vector can be studied as convergence of random variable, using the Cramer-Wold method:

Proposition 4. *$X_n \in \mathbb{R}^d$ converges in distribution to $X \in \mathbb{R}^d$, iff for all $t \in \mathbb{R}^d$, $t'X_n \rightsquigarrow t'X$.*

For sum of independent random variables, the characteristic functions factorise:

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$

For α-mixing sequence of random variables, we can control the difference between the characteristic functions and the factorisation. Let $(X_n : n \in \mathbb{N})$ be a stationary, α-mixing sequence of random variables, let $\xi_j = \exp\{it'(X_{l_j} + \cdots + X_{l_{j+1}})\}$, where the blocks $[l_j, l_{j+1}]$ are separated by at least s_T distance, then

Theorem 5. *$|E(\xi_1 \dots \xi_k) - E\xi_1 E\xi_2 E\xi_3 \dots E\xi_k| \leq 16(k-1)\alpha(s_T)$.*

this provides a way to prove CLT with α-mixing condition.

4 Hilbert Spaces

Hilbert space has a very nice structure: the inner product. Let S be a vector space over a field $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, an inner product is

$$\langle \cdot | \cdot \rangle : S \times S \to \mathbb{K}$$

satisfying for $a \in \mathbb{K}$ and $x, y \in S$,

$$\begin{aligned} \langle ax + y|z\rangle &= a\langle x|z\rangle + \langle y|z\rangle \\ \langle x|y\rangle &= \overline{\langle y|x\rangle} \\ \langle x|x\rangle \geq 0 \quad \text{with} \quad \langle x|x\rangle = 0 &\iff x = 0 \end{aligned}$$

Such S is called an inner product space. We can define norm on S with $\|x\| = \langle x|x\rangle^{\frac{1}{2}}$.

A Hilbert space is a complete inner product space, meaning that every Cauchy sequence (x_n) in S will converge to a point $x \in S$, where a Cauchy sequence is such that for any $\varepsilon > 0$, there exists N, for all $n, m > N$, $\|x_n - x_m\| \leq \varepsilon$.

One important example of a Hilbert Space is the space of all random variables with finite variance. Let $(\Omega, \mathcal{F}, P)$ be a probability space, we can define the space $L^2(\Omega, \mathcal{F}, P)$ to be the space of all random variables $X : \Omega \to \mathbb{R}$ such that $EX^2 < \infty . L^2(\Omega, \mathcal{F}, P)$ is a Hilbert space with inner product

$$\langle X|Y\rangle := E(XY) \quad \text{and} \quad \|X\| = EX^2$$

With this definition $X_n \to_{L^2} X$ if $E|X_n - X|^2 \to 0$ and we say X_n converges in L^2, or in **mean-square** to X.

It's in this space $L^2(\Omega, \mathcal{F}, P)$ we justify the recursive substitution of AR(1) process, suppose y_t is stationary, so $\operatorname{var}(y_t) < \infty$,

$$\begin{aligned} y_t &= \phi y_{t-1} + \varepsilon_t \\ &= \sum_{j=0}^{k} \phi^j \varepsilon_{t-j} + \phi^k y_{t-k} \end{aligned}$$

each $y_t \in L^2$ and we have for $S_k = \sum_{j=0}^{k} \phi^j \varepsilon_{t-j}$

$$\|y_t - S_k\| = \left\|\phi^k y_{t-k}\right\| \to 0$$

so that $S_k \to_{L^2} y_t$, justifying that $y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$.

Hilbert space has a good *geometric* structure, analougous to the geometry in Euclidean $\mathbb{R}^n$. We say $X \perp Y$, or X is orthogonal to Y if $\langle X|Y\rangle = 0$.

Given any $X, Y \in (S, \langle\cdot|\cdot\rangle)$, define the projection of X on Y as $P_Y X = \frac{\langle X|Y\rangle}{\|Y\|^2} Y$. You can verify that $X - P_Y X \perp Y$. It's more interesting to project $X \in S$ to a subspace $\mathcal{Y}$ of S, which we require to be closed(for all $y_n \in \mathcal{Y}$, if $y_n \to y$, then $y \in \mathcal{Y}$).

Theorem 6 (Projection). *Let $\mathcal{Y} \subset S$ be a closed linear subspace, then for all $x \in S$, there exists a unique projection of x on $\mathcal{Y}$, denoted as $P_{\mathcal{Y}} x \in \mathcal{Y}$ such that*

1. *For all $y \in \mathcal{Y}$, $\langle y|x - P_{\mathcal{Y}} x\rangle = 0$.*

2. *$P_{\mathcal{Y}} x$ solves $\min_{y \in \mathcal{Y}} \|x - y\|$.*

Remark. Many things can be thought of as projection.

1. OLS is the projection of Y vector onto the space spanned by columns of the X matrix.

2. Conditional expectation. Suppose we have two random variables X, Y, instead of projecting X on Y, we consider the projection of X on all functions $f(Y)$ such that $\text{var}(f(Y)) < \infty$. We want to find the function $f^*(Y)$ such that $\|X - f(Y)\|$ is minimised. It turns out that the minimiser is

$$f^*(Y) = E[X \mid Y].$$

   this explains why conditional expectation is the best predictor in terms of mean squared error.

5 Some Matrix Algebra

Let $A = [a_{ij}] = \begin{bmatrix} A_1, \ldots, A_n \end{bmatrix}$ be an $m \times n$ matrix. Some operations:

1. $\text{vec}\, A = \begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}$. $\text{vech}\, A$ is similar, except we discard the elements above the diagonal.

2. $A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & & & \\ \vdots & & & \\ a_{n1}B & \dots & \dots & a_{nn}B \end{bmatrix}$.

And some useful relationships,

1. $(A + B) \otimes (C + D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D$.
2. $(A \otimes B)(C \otimes D) = AC \otimes BD$.
3. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
4. $\mathrm{vec}(ABC) = (C' \otimes B)\,\mathrm{vec}\,A$.
5. $(a \otimes I)A = a \otimes A$ for a vector a.

Practice with the formula on page 39 of lecture 6.