Statistics Learning Lectures

Author: Ziyang Gong

Date: July 8, 2021

Version: 0.1.0

Facts are stubborn things, but statistics are pliable. —— Mark Twain

Contents

VI   Statistics Inference   56

IX Machine Learning   93

Part I

Calculus

Chapter 1 Limit Theory

Definition 1.1 (Mapping)

Let $X : \Omega_1 \to \Omega_2$ be a mapping.

1. *For every subset $B \in \Omega_2$, the inverse image of B is*

$$X^{-1}(B) = \{\omega : \omega \in \Omega_1, X(\omega) \in B\} := \{X \in B\}.$$

2. *For every class*

♣

Chapter 2 Differential Calculus

Chapter 3 Integral Calculus

Part II

Real Analysis

Chapter 4 Measure Theory

4.1 Semi-algebras, Algebras and Sigma-algebras

> **Definition 4.1 (Semi-algebra)**
>
> *A nonempty class of $\mathcal{S}$ of subsets of Ω is an **semi-algebra** on Ω that satisfy*
>
> *1. if $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$.*
>
> *2. if $A \in \mathcal{S}$, then A^C is a finite disjoint union of sets in $\mathcal{S}$, i.e.,*
> $$A^C = \sum_{i=1}^{n} A_i, where\, A_i \in \mathcal{S}, A_i \cap A_j = \emptyset, i \neq j.$$
> ♣

> **Definition 4.2 (Algebra)**
>
> *A nonempty class $\mathcal{A}$ of subsets of Ω is an **algebra** on Ω that satisfy*
>
> *1. if $A \in \mathcal{A}$, then $A^C \in \mathcal{A}$.*
>
> *2. if $A_1, A_2 \in \mathcal{A}$, then $A_1 \cup A_2 \in \mathcal{A}$.*
> ♣

> **Definition 4.3 (σ-algebra)**
>
> *A nonempty class $\mathcal{F}$ of subsets of Ω is a σ-**algebra** on Ω that satisfy*
>
> *1. if $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.*
>
> *2. if $A_i \in \mathcal{F}$ is a countable sequence of sets, then $\cup_i A_i \in \mathcal{F}$.*
> ♣

Example 4.1 Special σ-algebra

1. **Trival σ-algebra** := $\{\emptyset, \Omega\}$. This is smallest σ-algebra.
2. **Power Set** := all subsets of σ, denoted by $\mathcal{P}(\Omega)$. This is the largest σ-algebra.
3. **The smallest σ-algebra containing** $A \in \Omega := \{\emptyset, A, A^C, \Omega\}$.

It is easy to define (Lesbegue) measure on the semi-algebra $\mathcal{S}$, and then easily to extend it to the algebra $\overline{\mathcal{S}}$, finally, we can extend it further t o some σ-algebra (mostly consider the smallest one containing $\mathcal{S}$).

> **Lemma 4.1**
>
> *If $\mathcal{S}$ is a semi-algebra, then*
> $$\overline{\mathcal{S}} = \{finite\ disjoint\ unions\ of\ sets\ in\ \mathcal{S}\}$$
> *is an algebra, denoted by $\mathcal{A}(\mathcal{S})$, called **the algebra generated by** $\mathcal{S}$.*
> ♡

Proof Let $A, B \in \overline{\mathcal{S}}$, then $A = \sum_{i=1}^{n} A_i$, $B = \sum_{j=1}^{m} B_j$ with $A_i, B_i \in \mathcal{S}$.

Intersection: For $A_i \cap B_j \in \mathcal{S}$ by the definition of semi-algebra $\mathcal{S}$, thus

$$A \cap B = \sum_{i=1}^{n} \sum_{j=1}^{m} A_i \cap B_j \in \overline{\mathcal{S}}.$$

So $\overline{\mathcal{S}}$ is closed under (finite) intersection.

Complement: For DeMorgan's Law, $A_i^C \in \mathcal{S}$ by the definition of semi-algebra $\mathcal{S}$ and $\overline{\mathcal{S}}$ closed under (finite) intersection that we just shown, thus

$$A^C = (\sum_{i=1}^{n} A_i)^C = \cap_{i=1}^{n} A_i^C \in \overline{\mathcal{S}}.$$

So $\overline{\mathcal{S}}$ is closed under complement.

Union: For DeMorgan's Law and $\overline{\mathcal{S}}$ closed under (finite) intersection and complement that we just shown, thus

$$A \cup B = (A^C \cap B^C)^C \in \overline{\mathcal{S}}.$$

So $\overline{\mathcal{S}}$ is closed under (finite) union.

Hence, $\overline{\mathcal{S}}$ is an algebra.

Theorem 4.1

*For any class $\mathcal{A}$, there exists a unique minimal σ-algebra containing $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$, called **the σ-algebra generated by** $\mathcal{A}$. In other words,*

1. *$\mathcal{A} \subset \sigma(\mathcal{A})$.*
2. *For any σ-algebra $\mathcal{B}$ with $\mathcal{A} \subset \mathcal{B}$, $\sigma(\mathcal{A}) \subset \mathcal{B}$.*

and $\sigma(\mathcal{A})$ is unique.

Proof **Existence**:

Uniqueness:

Example 4.2 Borel σ-algebras generated from semi-algebras

1.

4.2 Measure

Definition 4.4 (Measure)

***Measure** is a nonnegative countably additive set function, that is, a function $\mu : \mathcal{A} \to \mathbf{R}$ with*

1. *$\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{A}$.*

> 2. if $A_i \in \mathcal{A}$ is a countable sequence of disjoint sets, then
>
> $$\mu(\cup_i A_i) = \sum_i \mu(A_i).$$
>
> ♣

Definition 4.5 (Measure Space)

*If μ is a measure on a σ-algebra $\mathcal{A}$ of subsets of Ω, the triplet $(\Omega, \mathcal{A}, \mu)$ is a **measure space**.*

♣

Note *A measure space $(\Omega, \mathcal{A}, \mu)$ is a **probability space**, if $P(\Omega) = 1$.*

Property *Let μ be a measure on a σ-algebra $\mathcal{A}$*

1. ***monotonieity** if $A \subset B$, then $\mu(A) \leq \mu(B)$.*
2. ***subadditivity** if $A \subset \cup_{m=1}^{\infty} A_m$, then $\mu(A) \leq \sum_{m=1}^{\infty} u(A_m)$.*
3. ***continuity from below** if $A_i \uparrow A$ (i.e. $A_1 \subset A_2 \subset \dots$ and $\cup_i A_i = A$), then $\mu(A_i) \uparrow \mu(A)$.*
4. ***continuity from above** if $A_i \downarrow A$ (i.e. $A_1 \supset A_2 \supset \dots$ and $\cap_i A_i = A$), then $\mu(A_i) \downarrow \mu(A)$.*

Proof

Chapter 5 Lebesgue Integration

5.1 Properties of the Integral

> **Theorem 5.1 (Jensen's Inequality)**
>
> *Let (Ω, A, μ) be a probability space. If f is a real-valued function that is μ-integrable, and if φ is a convex function on the real line, then:*
>
> $$\varphi\left(\int_\Omega f \, d\mu\right) \le \int_\Omega \varphi(f) d\mu. \qquad (5.1)$$

Proof Let $x_0 = \int_\Omega f \, d\mu$. Since the existence of subderivatives for convex functions, $\exists a, b \in R$, such that,

$$\forall x \in R, \varphi(x) \ge ax + b \text{ and } ax_0 + b = \varphi(x_0).$$

Then, we got

$$\int_\Omega \varphi(f) d\mu \ge \int_\Omega af + b \, d\mu = a \int_\Omega f \, d\mu + b = ax_0 + b = \varphi\left(\int_\Omega f \, d\mu\right).$$

> **Theorem 5.2 (Hölder's Inequality)**
>
> *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $p, q \in [1, \infty]$ with $1/p + 1/q = 1$. Then, for all measurable functions f and g on Ω,*
>
> $$\int_\Omega |f \cdot g| \, d\mu \le \|f\|_p \|g\|_q. \qquad (5.2)$$

Proof

> **Theorem 5.3 (Minkowski's Inequality)**
>
> *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $p \in [1, \infty]$. Then, for all measurable functions f and g on Ω,*
>
> $$\|f + g\|_p \le \|f\|_p + \|g\|_p. \qquad (5.3)$$

Proof Since $\varphi(x) = x^p$ is a convex function for $p \in [1, \infty)$. By it's definition,

$$|f + g|^p = \left|2 \cdot \frac{f}{2} + 2 \cdot \frac{g}{2}\right|^p \le \frac{1}{2}|2f|^p + \frac{1}{2}|2g|^p = 2^{p-1}\left(|f|^p + |g|^p\right).$$

Therefore,

$$|f + g|^p < 2^{p-1}\left(|f|^p + |g|^p\right) < \infty.$$

By Hölder's Inequality (5.2),

$$\|f+g\|_p^p = \int |f+g|^p \mathrm{d}\mu$$

$$= \int |f+g| \cdot |f+g|^{p-1} \mathrm{d}\mu$$

$$\leq \int (|f|+|g|)|f+g|^{p-1} \mathrm{d}\mu$$

$$= \int |f||f+g|^{p-1} \mathrm{d}\mu + \int |g||f+g|^{p-1} \mathrm{d}\mu$$

$$\leq \left(\left(\int |f|^p \mathrm{d}\mu \right)^{\frac{1}{p}} + \left(\int |g|^p \mathrm{d}\mu \right)^{\frac{1}{p}} \right) \left(\int |f+g|^{(p-1)\left(\frac{p}{p-1}\right)} \mathrm{d}\mu \right)^{1-\frac{1}{p}}$$

$$= (\|f\|_p + \|g\|_p) \frac{\|f+g\|_p^p}{\|f+g\|_p}$$

which means, as $p \in [1, \infty)$,

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p.$$

When $p = \infty$,

$$a$$

Theorem 5.4 (Bounded Convergence Theorem)

♡

Theorem 5.5 (Fatou's Lemma)

♡

Theorem 5.6 (Monotone Convergence Theorem)

♡

5.2 Product Measures

Theorem 5.7 (Fubini's Theorem)

♡

Part III

Functional Analysis

Part IV

Probability Theory

Chapter 6 Random Variables

6.1 Probability Space

> **Definition 6.1 (Probability Space)**
>
> *A probability space is a triple $(\Omega, \mathcal{F}, P)$ consisting of:*
>   1. *the sample space Ω: an arbitrary non-empty set.*
>   2. *the σ-algebra $\mathcal{F} \subseteq 2^\Omega$: a set of subsets of Ω, called events.*
>   3. *the probability measure $P : \mathcal{F} \to [0,1]$: a function on $\mathcal{F}$ which is a measure function.* ♣

6.2 Random Variables

> **Definition 6.2 (Random Variable)**
>
> *A random variable is a measurable function $X : \Omega \to S$ from a set of possible outcomes $(\Omega, \mathcal{F})$ to a measurable space $(S, \mathcal{S})$, that is,*
>
> $$X^{-1}(B) \equiv \{\omega : X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in \mathcal{S}. \tag{6.1}$$
>
> *Typically, $(S, \mathcal{S}) = (R^d, \mathcal{R}^d) \quad (d > 1)$.* ♣

How to prove that functions are measurable?

> **Theorem 6.1**
>
> *If $\{\omega : X(\omega) \in A\} \in \mathcal{F}$ for all $A \in \mathcal{A}$ and $\mathcal{A}$ generates $\mathcal{S}$, then X is measurable.* ♡

1.

6.3 Distributions

6.3.1 Definition of Distributions

> **Definition 6.3 (Distribution)**
>
> A distribution of random variable X is a probability function $P : \mathcal{R} \to \mathbb{R}$ by setting
> $$\mu(A) = P(X \in A) = P\left(X^{-1}(A)\right), \quad for A \in \mathcal{R}. \tag{6.2}$$

> **Definition 6.4 (Distribution Function)**
>
> The distribution of a random variable X is usually described by giving its **distribution function**,
> $$F(x) = P(X \leq x). \tag{6.3}$$

> **Definition 6.5 (Density Function)**
>
> If the distribution function $F(x) = P(X \leq x)$ has the form
> $$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y,$$
> that X has density function f.

6.3.2 Properties of Distributions

> **Theorem 6.2 (Properties of Distribution Function)**
>
> Any distribution function F has the following properties,
> 1. F is nondecreasing.
> 2. $\lim_{x \to \infty} F(x) = 1, \lim_{x \to -\infty} F(x) = 0$.
> 3. F is right continuous, i.e., $\lim_{y \downarrow x} F(y) = F(x)$.
> 4. If $F(x-) = \lim_{y \uparrow x} F(y)$, then $F(x-) = P(X < x)$.
> 5. $P(X = x) = F(x) - F(x-)$.

Proof

> **Theorem 6.3**
>
> If F satisfies (1), (2), and (3) in Theorem 6.2, then it is the distribution function of some random variable.

Proof

> **Theorem 6.4**
>
> *A distribution function has at most countably many discontinuities*

Proof

6.3.3 Families of Distributions

6.4 Expected Value

> **Definition 6.6 (Expectation)**

> **Theorem 6.5 (Bounded Convergence theorem)**

> **Theorem 6.6 (Fatou's Lemma)**
>
> *If $X_n \geq 0$, then*
> $$\liminf_{n \to \infty} E X_n \geq E \left(\liminf_{n \to \infty} X_n \right). \tag{6.4}$$

> **Theorem 6.7 (Monotone Convergence theorem)**
>
> *If $0 \leq X_n \uparrow X$, then*
> $$E X_n \uparrow E X. \tag{6.5}$$

> **Theorem 6.8 (Dominated Convergence theorem)**
>
> *If $X_n \to X$ a.s., $|X_n| \leq Y$ for all n, and $E Y < \infty$, then*
> $$E X_n \to E X. \tag{6.6}$$

6.5 Independence

6.5.1 Definition of Independence

> **Definition 6.7 (Independence)**
>
> 1. *Two events A and B are independent if $P(A \cap B) = P(A) P(B)$.*
> 2. *Two random variables X and Y are independent if for all $C, D \in \mathcal{R}$*
> $$P(X \in C, Y \in D) = P(X \in C) P(Y \in D). \tag{6.7}$$
> 3. *Two σ-fields $\mathcal{F}$ and $\mathcal{G}$ are independent if for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$ the events A*

> *and B are independent.* ♣

The second definition is a special case of the third.

> **Theorem 6.9**
> 1. *If X and Y are independent then $\sigma(X)$ and $\sigma(Y)$ are independent.*
> 2. *Conversely, if $\mathcal{F}$ and $\mathcal{G}$ are independent, $X \in \mathcal{F}$ and $Y \in \mathcal{G}$, then X and Y are independent.*
> ♡

The first definition is, in turn, a special case of the second.

> **Theorem 6.10**
> 1. *If A and B are independent, then so are A^c and B, A and B^c, and A^c and B^c.*
> 2. *Conversely, events A and B are independent if and only if their indicator random variables 1_A and 1_B are independent.*
> ♡

The definition of independence can be extended to the infinite collection.

> **Definition 6.8**
> *An infinite collection of objects (σ-fields, random variables, or sets) is said to be independent if every finite subcollection is,*
> 1. *σ-fields $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are independent if whenever $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$, we have*
> $$P\left(\cap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i). \tag{6.8}$$
> 2. *Random variables $X_1, \ldots, X_n$ are independent if whenever $B_i \in \mathcal{R}$ for $i = 1, \ldots, n$ we have*
> $$P\left(\cap_{i=1}^n \{X_i \in B_i\}\right) = \prod_{i=1}^n P(X_i \in B_i). \tag{6.9}$$
> 3. *Sets $A_1, \ldots, A_n$ are independent if whenever $I \subset \{1, \ldots, n\}$ we have*
> $$P\left(\cap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i). \tag{6.10}$$
> ♣

6.5.2 Sufficient Conditions for Independence

6.5.3 Independence, Distribution, and Expectation

> **Theorem 6.11**
>
> *Suppose $X_1, \ldots, X_n$ are independent random variables and X_i has distribution μ_i, then $(X_1, \ldots, X_n)$ has distribution $\mu_1 \times \cdots \times \mu_n$.*

> **Theorem 6.12**
>
> *If $X_1, \ldots, X_n$ are independent and have*
>   1. *$X_i \geq 0$ for all i, or*
>   2. *$E\,|X_i| < \infty$ for all i.*
>
> *then*
> $$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E X_i \tag{6.11}$$

6.5.4 Sums of Independent Random Variables

> **Theorem 6.13 (Convolution for Random Variables)**
>
>   1. *If X and Y are independent, $F(x) = P(X \leq x)$, and $G(y) = P(Y \leq y)$, then*
> $$P(X + Y \leq z) = \int F(z - y)\mathrm{d}G(y). \tag{6.12}$$
>
>   2. *If X and Y are independent, X with density f and Y with distribution function G, then $X + Y$ has density*
> $$h(x) = \int f(x - y)\mathrm{d}G(y). \tag{6.13}$$
>
>   *Suppose Y has density g, the last formula can be written as*
> $$h(x) = \int f(x - y)g(y)\mathrm{d}y. \tag{6.14}$$
>
>   3. *If X and Y are independent, integral-valued random variables, then*
> $$P(X + Y = n) = \sum_{m} P(X = m)P(Y = n - m). \tag{6.15}$$

6.6 Moments

> **Lemma 6.1**
>
> *If $Y > 0$ and $p > 0$, then*
> $$E(Y^p) = \int_0^\infty py^{p-1} P(Y > y) \, dy. \tag{6.16}$$

6.7 Characteristic Functions

6.7.1 Definition of Characteristic Functions

> **Definition 6.9 (Characteristic Function)**
>
> *If X is a random variable, we define its characteristic function (ch.f) by*
> $$\varphi(t) = E\left(e^{itX}\right) = E\left(\cos tX\right) + iE\left(\sin tX\right). \tag{6.17}$$

Note *Eular Equation.*

6.7.2 Properties of Characteristic Functions

> **Theorem 6.14 (Properties of Characteristic Function)**
>
> *Any characteristic function has the following properties:*
> 1. *$\varphi(0) = 1$,*
> 2. *$\varphi(-t) = \overline{\varphi(t)}$,*
> 3. *$|\varphi(t)| = |Ee^{itX}| \leq E|e^{itX}| = 1$,*
> 4. *$\varphi(t)$ is uniformly continuous on $(-\infty, \infty)$,*
> 5. *$Ee^{it(aX+b)} = e^{itb}\varphi(at)$,*
> 6. *If X_1 and X_2 are independent and have ch.f.'s φ_1 and φ_2, then $X_1 + X_2$ has ch.f. $\varphi_1(t)\varphi_2(t)$.*

Proof

6.7.3 The Inversion Formula

The characteristic function uniquely determines the distribution. This and more is provided by:

Theorem 6.15 (The Inversion Formula)

Let $\varphi(t) = \int e^{itx}\mu(\mathrm{d}x)$ where μ is a probability measure. If $a < b$, then

$$\lim_{T \to \infty} (2\pi)^{-1} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)dt = \mu(a,b) + \frac{1}{2}\mu(\{a,b\}) \qquad (6.18)$$

Proof

Theorem 6.16

If $\int |\varphi(t)|\mathrm{d}t < \infty$, then μ has bounded continuous density

$$f(y) = \frac{1}{2\pi} \int e^{-ity} \varphi(t)\mathrm{d}t. \qquad (6.19)$$

Proof

6.7.4 Convergence in Distribution

Theorem 6.17 (Lèvy's Continuity Theorem)

Let $\mu_n, 1 \le n \le \infty$ be probability measures with ch.f. φ_n.
1. If $\mu_n \xrightarrow{d} \mu_\infty$, then $\varphi_n(t) \to \varphi_\infty(t)$ for all t.
2. If $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at 0, then the associated sequence of distributions μ_n is tight and converges weakly to the measure μ with characteristic function φ.

Proof

6.7.5 Moments and Derivatives

Theorem 6.18

If $\int |x|^n \mu(dx) < \infty$, then its characteristic function φ has a continuous derivative of order n given by

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx} \mu(\mathrm{d}x). \qquad (6.20)$$

Theorem 6.19

If $E|X|^2 < \infty$ then

$$\varphi(t) = 1 + itEX - t^2 E\left(X^2\right)/2 + o\left(t^2\right). \qquad (6.21)$$

> **Theorem 6.20**
>
> *If* $\limsup_{h\downarrow 0}\{\varphi(h) - 2\varphi(0) + \varphi(-h)\}/h^2 > -\infty$, *then*
> $$E|X|^2 < \infty. \tag{6.22}$$

Chapter 7 Convergence of Random Variables

> **Introduction**
>
> ❏ *Convergence in Mean*   ❏ *Convergence in Distribution*
> ❏ *Convergence in Probability*   ❏ *Almost Sure Convergence*
> ❏ *Convergence in Uninform*

7.1 Convergence in Mean

> **Definition 7.1 (Convergence in Mean)**
>
> *A sequence $\{X_n\}$ of real-valued random variables **converges in the r-th mean** ($r \geq 1$) towards the random variable X, if*
>   1. *The r-th absolute moments $E(|X_n|^r)$ and $E(|X|^r)$ of $\{X_n\}$ and X exist,*
>   2. $\lim_{n\to\infty} E\left(|X_n - X|^r\right) = 0.$
>
> *Convergence in the r-th mean is denoted by*
> $$X_n \xrightarrow{L^r} X. \tag{7.1}$$

7.2 Convergence in Probability

> **Definition 7.2 (Convergence in Probability)**
>
> *A sequence $\{X_n\}$ of real-valued random variables **converges in probability** towards the random variable X, if*
> $$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0. \tag{7.2}$$
> *Convergence in probability is denoted by*
> $$X_n \xrightarrow{p} X. \tag{7.3}$$

7.3 Convergence in Uninform

> **Definition 7.3 (Convergence in Uninform)**

7.4 Convergence in Distribution

> **Definition 7.4 (Convergence in Distribution)**
>
> *A sequence $\{X_n\}$ of real-valued random variables is said to **converge in distribution**, or **converge weakly**, or **converge in law** to a random variable X, if*
>
> $$\lim_{n \to \infty} F_n(x) = F(x), \qquad (7.4)$$
>
> *for every number at $x \in \mathbb{R}$ which F is continuous. Here F_n and F are the cumulative distribution functions of random variables X_n and X, respectively.*
>
> *Convergence in distribution is denoted as*
>
> $$X_n \xrightarrow{d} X, \text{ or } X_n \Rightarrow X. \qquad (7.5) \clubsuit$$

Note

- *Convergence in Distribution is the weakest form of convergence typically discussed, since it is implied by all other types of convergence mentioned in this chapter.*
- *Convergence in Distribution does not imply that the sequence of corresponding probability density functions will also converge. However, according to Scheffé's theorem, convergence of the probability density functions implies convergence in distribution.*

> **Lemma 7.1**
>
> *If $F_n \xrightarrow{d} F_\infty$, then there are random variables $Y_n, 1 \leq n \leq \infty$, with distribution F_n so that*
>
> $$Y_n \xrightarrow{a.s.} Y_\infty. \qquad (7.6) \heartsuit$$

> **Theorem 7.1 (Portmanteau Lemma)**
>
> $\{X_n\}$ *converges in distribution to X, if and only if any of the following statements are true,*
>
> - $P(X_n \leq x) \to P(X \leq x)$, *for all continuity points of the distribution of X.*
> - $Ef(X_n) \to Ef(X)$, *for all bounded, continuous (Lipschitz) functions f.*
> - $\liminf_{n \to \infty} P(X_n \in G) \geq P(X_\infty \in G)$, *for all open sets G.*
> - $\limsup_{n \to \infty} P(X_n \in K) \leq P(X_\infty \in K)$, *for all closed sets K.*
> - $\lim_{n \to \infty} P(X_n \in A) = P(X_\infty \in A)$, *for all Borel sets A with $P(X_\infty \in \partial A) = 0$.* $\heartsuit$

Proof

Theorem 7.2 (Continuous Mapping Theorem)

Let g be a measurable function and $D_g = \{x : g$ is discontinuous at $x\}$ with $P(X \in D_g) = 0$, then,

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X),$$
$$X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X), \tag{7.7}$$
$$X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X).$$

If in addition g is bounded, then

$$Eg(X_n) \to Eg(X). \tag{7.8}$$

Proof

Theorem 7.3

If $X_n \xrightarrow{p} X$, then

$$X_n \xrightarrow{d} X, \tag{7.9}$$

and that, conversely, if $X_n \xrightarrow{d} c$, where c is a constant, then

$$X_n \xrightarrow{p} c. \tag{7.10}$$

Proof

1. $\forall \varepsilon > 0$, at fixed point x, since if $X_n \leq x$ and $|X_n - X| \leq \varepsilon$, then $X \leq x + \varepsilon$, then

$$\{X \leq x + \varepsilon\} \subset \{X_n \leq x\} \cup \{|X_n - X| > \varepsilon\},$$

similarily, if $X \leq x - \varepsilon$ and $|X_n - X| \leq \varepsilon$, then $X_n \leq x$, then

$$\{X_n \leq x\} \subset \{X \leq x - \varepsilon\} \cup \{|X_n - X| > \varepsilon\},$$

then, by the union bound,

$$P(X \leq x + \varepsilon) \leq P(X_n \leq x) + P(|X_n - X| > \varepsilon),$$
$$P(X_n \leq x) \leq P(X \leq x - \varepsilon) + P(|X_n - X| > \varepsilon).$$

So, we got

$$P(X \leq x + \varepsilon) - P(|X_n - X| > \varepsilon) \leq P(X_n \leq x)$$
$$\leq P(X \leq x - \varepsilon) + P(|X_n - X| > \varepsilon)$$

As $n \to \infty$, $P(|X_n - X| > \varepsilon) \to 0$, then

$$P(X \leq x - \varepsilon) \leq \lim_{n\to\infty} P(X_n \leq x) \leq P(X \leq x + \varepsilon)$$
$$\Rightarrow F(x - \varepsilon) \leq \lim_{n\to\infty} F_n(x) \leq F(x + \varepsilon)$$

By the property of distribution (Theorem 6.2), as $\varepsilon \to 0$, then

$$\lim_{n \to \infty} F_n(x) = F(x),$$

which means,

$$X_n \xrightarrow{d} X.$$

2. Since $X_n \xrightarrow{d} c$, where c is a constant, then $\forall \varepsilon > 0$,

$$\lim_{n \to \infty} P(X_n \le c + \varepsilon) = 1 \implies \lim_{n \to \infty} P(X_n > c + \varepsilon) = 0$$

$$\lim_{n \to \infty} P(X_n \le c - \varepsilon) = 0.$$

Therefore,

$$P\left(|X_n - c| < \varepsilon\right) = 0,$$

which means

$$X_n \xrightarrow{p} c.$$

Theorem 7.4 (Slutsky's Theorem)

Let X_n, Y_n be sequences of random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then
1. *$X_n + Y_n \xrightarrow{d} X + c$.*
2. *$X_n Y_n \xrightarrow{d} cX$.*
3. *$X_n / Y_n \xrightarrow{d} X/c$, provided that c is invertible.*

Proof

Note *However that convergence in distribution of $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ does in general not imply convergence in distribution of $X_n + Y_n \xrightarrow{d} X + Y$ or of $X_n Y_n \xrightarrow{d} XY$.*

Theorem 7.5 (Cramér-Wold Theorem)

Theorem 7.6 (Helly's Selection Theorem)

For every sequence F_n of distribution functions, there is a subsequence $F_n(k)$ and a right continuous nondecreasing function F so that $\lim_{k \to \infty} F_{n(k)}(y) = F(y)$ at all continuity points y of F.

Theorem 7.7

Every subsequential limit is the distribution function of a probability measure if and only if the sequence F_n is tight, i.e., for all $\epsilon > 0$ there is an M_ϵ so that

$$\limsup_{n \to \infty} 1 - F_n(M_\epsilon) + F_n(-M_\epsilon) \le \epsilon. \tag{7.11}$$

7.5 Almost Sure Convergence

> **Definition 7.5 (Almost Sure Convergence)**
>
> *A sequence $\{X_n\}$ of real-valued random variables converges **almost sure** or **almost everywhere** or **with probability 1** or **strongly** towards the random variable X, if*
>
> $$P\left(\lim_{n\to\infty} X_n = X\right) = 1. \tag{7.12}$$
>
> *Almost sure convergence is denoted by*
>
> $$X_n \xrightarrow{a.s.} X. \tag{7.13}$$ ♣

Note

> **Theorem 7.8**
>
> *If $X_n \xrightarrow{a.s.} X$, then*
>
> $$X_n \xrightarrow{p} X. \tag{7.14}$$ ♡

Proof

> **Theorem 7.9**
>
> $X_n \xrightarrow{p} X$ *if and only if for all subsequence $X_{n(m)}$ exists a further subsequence $X_{n(m_k)}$, such that*
>
> $$X_{n(m_k)} \xrightarrow{a.s.} X. \tag{7.15}$$ ♡

① *if* $r \geq s \geq 1, X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{L^s} X$

② $X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{p} X$

③ $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$

④ *if* $X = c, X_n \xrightarrow{d} X \Rightarrow X_n \xrightarrow{p} X$

⑤ $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$

⑥ $X_n \xrightarrow{p} X \Rightarrow X_{n(m_k)} \xrightarrow{a.s.} X$

⑦ $X_n \xrightarrow{d} X \Rightarrow \exists Y_n \overset{d}{=} X_n, Y \overset{d}{=} X, s.t.\ Y_n \xrightarrow{a.s.} Y$

⑧ $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{d} X$
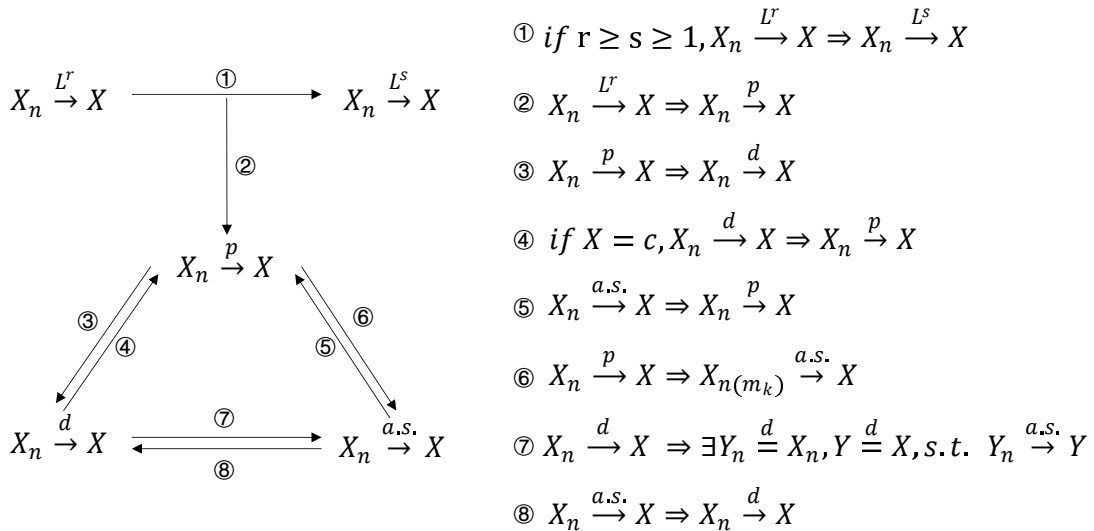
Figure 7.1: Relations of Convergence of Random Variables

7.6 Asymptotic Notation for Random Variables

> **Definition 7.6**
>
> A sequence $\{A_n\}$ of real-valued random variables is of smaller order in probability than a sequence $\{B_n\}$, if
>
> $$\frac{A_n}{B_n} \xrightarrow{p} 0. \tag{7.16}$$
>
> Smaller order in probability is denoted by
>
> $$A_n = o_p(B_n). \tag{7.17}$$
>
> Particularly,
>
> $$A_n = o_p(1) \iff A_n \xrightarrow{p} 0. \tag{7.18}$$ ♣

> **Definition 7.7**
>
> A sequence $\{A_n\}$ of real-valued random variables is of smaller order than or equal to a sequence $\{B_n\}$ in probability, if
>
> $$\forall \varepsilon > 0 \; \exists M_\varepsilon, \quad \lim_{n \to \infty} P\left(|A_n| \leq M_\varepsilon |B_n|\right) \geq 1 - \varepsilon. \tag{7.19}$$
>
> Smaller order than or equal to in probability is denoted by
>
> $$A_n = O_p(B_n). \tag{7.20}$$ ♣

> **Definition 7.8**
>
> A sequence $\{A_n\}$ of real-valued random variables is of the same order as a sequence $\{B_n\}$ in probability, if
>
> $$\forall \varepsilon > 0 \; \exists m_\varepsilon < M_\varepsilon, \quad \lim_{n \to \infty} P\left(m_\varepsilon < \frac{|A_n|}{|B_n|} < M_\varepsilon\right) \geq 1 - \varepsilon. \tag{7.21}$$
>
> Same order in probability is denoted by
>
> $$A_n \asymp_p B_n. \tag{7.22}$$ ♣

Chapter 8 Law of Large Numbers

8.1 Weak Law of Large Numbers

> **Lemma 8.1**
>
> *If $p > 0$ and $E\,|Z_n|^p \to 0$, then*
> $$Z_n \xrightarrow{p} 0. \tag{8.1}$$

Proof

> **Theorem 8.1 (Weak Law of Large Numbers with Finite Variances)**
>
> *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $EX_i = \mu$ and $Var(X_i) \le C < \infty$. Suppose $S_n = X_1 + X_2 + \ldots + X_n$, then*
> $$S_n/n \xrightarrow{L^2} \mu, \quad S_n/n \xrightarrow{p} \mu. \tag{8.2}$$

Proof

> **Theorem 8.2 (Weak Law of Large Numbers without i.i.d.)**
>
> *Let $X_1, X_2, \ldots$ be random variables, Suppose $S_n = X_1 + X_2 + \ldots + X_n$, $\mu_n = E\,S_n$, $\sigma_n^2 = Var(S_n)$, if $\sigma_n^2/b_n^2 \to 0$, then*
> $$\frac{S_n - \mu_n}{b_n} \xrightarrow{p} 0. \tag{8.3}$$

Proof

> **Theorem 8.3 (Weak Law of Large Numbers for Triangular Arrays)**
>
> *For each n, let $X_{n,m}, 1 \le m \le n$, be independent random variables. Suppose $b_n > 0$ with $b_n \to \infty$, $\bar{X}_{n,m} = X_{n,m} I_{(X_{n,m} \le b_n)}$, if*
>
> *1.* $\sum_{m=1}^{n} P\left(|X_{n,m}| > b_n\right) \to 0$, *and*
> *2.* $b_n^{-2} \sum_{m=1}^{n} E\,\bar{X}_{n,m}^2 \to 0$.

Suppose $S_n = X_{n,1} + \cdots + X_{n,n}$ and $a_n = \sum_{m=1}^n E\bar{X}_{n,m}$, then
$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0. \tag{8.4}$$

Proof

Theorem 8.4 (Weak Law of Large Numbers by Feller)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with
$$\lim_{x \to 0} xP(|X_i| > x) = 0. \tag{8.5}$$
Suppose $S_n = X_1 + X_2 + \ldots + X_n$, $\mu_n = E\left(X_1 I_{(|X_1|<n)}\right)$, then
$$S_n/n - \mu_n \xrightarrow{p} 0. \tag{8.6}$$

Proof

Theorem 8.5 (Weak Law of Large Numbers)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E|X_i| < \infty$. Suppose $S_n = X_1 + X_2 + \ldots + X_n$, $\mu = EX_i$, then
$$S_n/n \xrightarrow{p} \mu. \tag{8.7}$$

Proof

Note $E|X_i| = \infty$

8.2 Strong Law of Large Numbers

8.2.1 Borel-Cantelli Lemmas

Lemma 8.2 (Borel-Cantelli Lemma)

If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then
$$P(A_n \text{ i.o. }) = 0. \tag{8.8}$$

Lemma 8.3 (The Second Borel-Cantelli Lemma)

If $\{A_n\}$ are independent with $\sum_{n=1}^{\infty} P(A_n) = \infty$, then,
$$P(A_n \text{ i.o. }) = 1. \tag{8.9}$$

> **Corollary 8.1**
>
> *Suppose $\{A_n\}$ are independent with $P(A_n) < 1, \forall n$. If $P\left(\cup_{n=1}^{\infty} A_n\right) = 1$ then*
>
> $$\sum_{n=1}^{\infty} P(A_n) = \infty, \tag{8.10}$$
>
> *and hence $P(A_n \text{ i.o. }) = 1$*

Proof

8.2.2 Strong Law of Large Numbers

> **Theorem 8.6 (Strong Law of Large Numbers)**
>
> *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E|X_i| < \infty$. Suppose $S_n = X_1 + X_2 + \ldots + X_n$, $\mu = EX_i$, then*
>
> $$S_n/n \stackrel{a.s.}{\to} \mu. \tag{8.11}$$

8.3 Uniform Law of Large Numbers

> **Theorem 8.7 (Uniform Law of Large Numbers)**
>
> *Suppose*
>
> 1. *Θ is compact.*
> 2. *$g(X_i, \theta)$ is continuous at each $\theta \in \Theta$ almost sure.*
> 3. *$g(X_i, \theta)$ is dominated by a function $G(X_i)$, i.e. $|g(X_i, \theta)| \le G(X_i)$.*
> 4. *$EG(X_i) < \infty$.*
>
> *Then*
>
> $$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta) \right| \stackrel{p}{\to} 0. \tag{8.12}$$

Proof Suppose

$$\Delta_\delta(X_i, \theta_0) = \sup_{\theta \in B(\theta_0, \delta)} g(X_i, \theta) - \inf_{\theta \in B(\theta_0, \delta)} g(X_i, \theta).$$

Since (i) $\Delta_\delta(X_i, \theta_0) \stackrel{a.s.}{\to} 0$ by condition (2), (ii) $\Delta_\delta(X_i, \theta_0) \le 2\sup_{\theta \in \Theta} |g(X_i, \theta)| \le 2G(X_i)$ by condition (3) and (4). Then

$$E\Delta_\delta(X_i, \theta_0) \to 0, \text{ as } \delta \to 0.$$

So, for all $\theta \in \Theta$ and $\varepsilon > 0$, there exists $\delta_\varepsilon(\theta)$ such that

$$E\left[\Delta_{\delta_\varepsilon(\theta)}(X_i, \theta)\right] < \varepsilon.$$

Since Θ is compact, we can find a finite subcover, such that Θ is covered by

$$\cup_{k=1}^{K} B\left(\theta_k, \delta_\varepsilon\left(\theta_k\right)\right).$$

$$\sup_{\theta \in \Theta}\left[n^{-1} \sum_{i=1}^{n} g\left(X_i, \theta\right) - Eg\left(X_i, \theta\right)\right]$$

$$= \max_{k} \sup_{\theta \in B\left(\theta_k, \delta_\varepsilon\left(\theta_k\right)\right)}\left[n^{-1} \sum_{i=1}^{n} g\left(X_i, \theta\right) - Eg\left(X_i, \theta\right)\right].$$

$$\leq \max_{k}\left[n^{-1} \sum_{i=1}^{n} \sup_{\theta \in B\left(\theta_k, \delta_\varepsilon\left(\theta_k\right)\right)} g\left(X_i, \theta\right) - E \inf_{\theta \in B\left(\theta_k, \delta_\varepsilon\left(\theta_k\right)\right)} g\left(X_i, \theta\right)\right]$$

Since

$$E\left|\sup_{\theta \in B\left(\theta_k, \delta_c\left(\theta_k\right)\right)} g\left(X_i, \theta\right)\right| \leq EG\left(X_i\right) < \infty,$$

by the Weak Law of Large Numbers (Theorem 8.5),

$$= o_p(1) + \max_{k}\left[E \sup_{\theta \in B\left(\theta_k, \delta_\varepsilon\left(\theta_k\right)\right)} g\left(X_i, \theta\right) - E \inf_{\theta \in B\left(\theta_k, \delta_\varepsilon\left(\theta_k\right)\right)} g\left(X_i, \theta\right)\right]$$

$$= o_p(1) + \max_{k} E \Delta_{\delta_\varepsilon\left(\theta_k\right)}\left(X_i, \theta_k\right)$$

$$\leq o_p(1) + \varepsilon$$

By analogous argument,

$$\inf_{\theta \in \Theta}\left[n^{-1} \sum_{i=1}^{n} g\left(X_i, \theta\right) - Eg\left(X_i, \theta\right)\right] \geq o_p(1) - \varepsilon.$$

The desired result follows from the above equation by the fact that ε is chosen arbitrarily.

Chapter 9 Central Limit Theorems

<div style="border:1px solid #aac; padding:6px;">

Introduction

❏ *Classic Central Limit Theorem*

❏ *Central Limit Theorem for independent non-identical Random Variables*

❏ *Central Limit Theorem for dependent Random Variables*

</div>

9.1 Central Limit Theorem

9.1.1 The De Moivre-Laplace Theorem

> **Lemma 9.1 (Stirling's Formula)**
>
> $$n! \sim \sqrt{2\pi}\, n^{n+\frac{1}{2}} e^{-n} \ \text{ as } n \to \infty. \tag{9.1}$$

Proof

> **Lemma 9.2**
>
> If $c_j \to 0$, $a_j \to \infty$ and $a_j c_j \to \lambda$, then
>
> $$\left(1 + c_j\right)^{a_j} \to e^{\lambda}. \tag{9.2}$$

Proof

> **Theorem 9.1 (The De Moivre-Laplace Theorem)**
>
> *Let $X_1, X_2, \ldots$ be i.i.d. with $P\left(X_1 = 1\right) = P\left(X_1 = -1\right) = 1/2$ and let $S_n = X_1 + \cdots + X_n$. If $a < b$, then as $m \to \infty$*
>
> $$P\left(a \leq S_m/\sqrt{m} \leq b\right) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2}\mathrm{d}x. \tag{9.3}$$

Proof If n and k and integers

$$P\left(S_{2n} = 2k\right) = \binom{2n}{n+k} 2^{-2n}$$

By lemma 9.1, we have

$$\binom{2n}{n+k} = \frac{(2n)!}{(n+k)!(n-k)!}$$

$$\sim \frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \cdot \frac{(2\pi(2n))^{1/2}}{(2\pi(n+k))^{1/2}(2\pi(n-k))^{1/2}}$$

Hence,

$$P\left(S_{2n} = 2k\right) = \binom{2n}{n+k} 2^{-2n}$$

$$\sim \left(1 + \frac{k}{n}\right)^{-n-k} \cdot \left(1 - \frac{k}{n}\right)^{-n+k}$$

$$\cdot (\pi n)^{-1/2} \cdot \left(1 + \frac{k}{n}\right)^{-1/2} \cdot \left(1 - \frac{k}{n}\right)^{-1/2}$$

$$= \left(1 - \frac{k^2}{n^2}\right)^{-n} \cdot \left(1 + \frac{k}{n}\right)^{-k} \cdot \left(1 - \frac{k}{n}\right)^{k}$$

$$\cdot (\pi n)^{-1/2} \cdot \left(1 + \frac{k}{n}\right)^{-1/2} \cdot \left(1 - \frac{k}{n}\right)^{-1/2}$$

Let $2k = x\sqrt{2n}$, i.e., $k = x\sqrt{\frac{n}{2}}$. By lemma 9.2, we have

$$\left(1 - \frac{k^2}{n^2}\right)^{-n} = \left(1 - x^2/2n\right)^{-n} \to e^{x^2/2}$$

$$\left(1 + \frac{k}{n}\right)^{-k} = (1 + x/\sqrt{2n})^{-x\sqrt{n/2}} \to e^{-x^2/2}$$

$$\left(1 - \frac{k}{n}\right)^{k} = (1 - x/\sqrt{2n})^{x\sqrt{n/2}} \to e^{-x^2/2}$$

For this choice of k, $k/n \to 0$, so

$$\left(1 + \frac{k}{n}\right)^{-1/2} \cdot \left(1 - \frac{k}{n}\right)^{-1/2} \to 1.$$

Putting things together, we have

$$P\left(S_{2n} = 2k\right) \sim (\pi n)^{-1/2} e^{-x^2/2}, \text{ as } \frac{2k}{\sqrt{2n}} \to x.$$

Therefore,

$$P\left(a\sqrt{2n} \le S_{2n} \le b\sqrt{2n}\right) = \sum_{m \in [a\sqrt{2n}, b\sqrt{2n}] \cap 2\mathbb{Z}} P\left(S_{2n} = m\right)$$

Let $m = x\sqrt{2n}$, we have that this is

$$\approx \sum_{x \in [a,b] \cap (2\mathbb{Z}/\sqrt{2n})} (2\pi)^{-1/2} e^{-x^2/2} \cdot (2/n)^{1/2}$$

where $2\mathbb{Z}/\sqrt{2n} = \left\{2z/\sqrt{2n} : z \in \mathbb{Z}\right\}$. As $n \to \infty$, the sum just shown is

$$\approx \int_a^b (2\pi)^{-1/2} e^{-x^2/2} dx.$$

To remove the restriction to even integers, observe $S_{2n+1} = S_{2n} \pm 1$.

Let $m = 2n$, as $m \to \infty$,

$$P\left(a \le S_m/\sqrt{m} \le b\right) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2} dx.$$

9.1.2 Classic Central Limit Theorem

Theorem 9.2 (Classic Central Limit Theorem (i.i.d.))

Let $X_1, X_2, \ldots$ be i.i.d. with $E X_i = \mu$, $Var(X_i) = \sigma^2 < \infty$. Let $S_n = X_1 + X_2 + \ldots + X_n$, then

$$\frac{S_n - n\mu}{\sigma n^{\frac{1}{2}}} \xrightarrow{d} \chi, \tag{9.4}$$

where χ has the standard normal distribution.

Proof

Theorem 9.3 (The Linderberg-Feller Central Limit Theorem)

For each n, let $X_{n,m}, 1 \le m \le n$, be independent random variables with $E X_{n,m} = 0$. If
  1. $\sum_{m=1}^n E X_{n,m}^2 \to \sigma^2 > 0$.
  2. $\forall \epsilon > 0, \lim_{n \to \infty} \sum_{m=1}^n E\left(|X_{n,m}|^2 ; |X_{n,m}| > \epsilon\right) = 0$

Then $S_n = X_{n,1} + \cdots + X_{n,n} \xrightarrow{d} \sigma\chi$ as $n \to \infty$.

Theorem 9.4 (Berry-Esseen Theorem)

Proof

9.2 Central Limit Theorem for independent non-identical Random Variables

Theorem 9.5 (The Liapounov Central Limit Theorem)

9.3 Central Limit Theorem for dependent Random Variables

Chapter 10 The Delta Methods

Theorem 10.1 (Delta Method)

Let $\{X_n\}$ be a sequence of random variables with

$$\sqrt{n}\,[X_n - \theta] \xrightarrow{d} \sigma\chi,$$

where θ and σ are finite, then for any function g with the property that $g'(\theta)$ exists and is non-zero valued,

$$\sqrt{n}\,[g\,(X_n) - g(\theta)] \xrightarrow{d} \sigma g'(\theta)\chi.$$

Proof   Under the assumption that $g'(\theta)$ is continuous.

Since, $g'(\theta)$ exists, with the first-order Taylor Approximation:

$$g(X_n) = g(\theta) + g'(\tilde{\theta})(X_n - \theta),$$

where $\tilde{\theta}$ lies between X_n and θ.

Since $X_n \xrightarrow{p} \theta$, and $|\tilde{\theta} - \theta| < |X_n - \theta|$, then

$$\tilde{\theta} \xrightarrow{p} \theta,$$

Since $g'(\theta)$ is continuous, by Continuous Mapping Theorem (7.2),

$$g'(\tilde{\theta}) \xrightarrow{p} g'(\theta).$$

and,

$$\sqrt{n}\,(g(X_n) - g(\theta)) = \sqrt{n}g'(\tilde{\theta})(X_n - \theta),$$

$$\sqrt{n}\,[X_n - \theta] \xrightarrow{d} \sigma\chi,$$

by Slutsky's Theorem (7.4),

$$\sqrt{n}\,[g\,(X_n) - g(\theta)] \xrightarrow{d} \sigma g'(\theta)\chi.$$

Chapter 11 Exercises for Probability Theory and Examples

11.1 Measure Theory

✍ **Exercise 11.1**

1. Show that if $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ are σ-algebras, then $\cup_i \mathcal{F}_i$ is an algebra.
2. Give an example to show that $\cup_i \mathcal{F}_i$ need not be a σ-algebra.

Solution

1. *__Complement__: Suppose $A \in \cup_i \mathcal{F}_i$, since $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$, assume $A \in \mathcal{F}_i$. And each $\mathcal{F}_i$ is σ-algebra,*

$$A^c \in \mathcal{F}_i \subset \cup_i \mathcal{F}_i.$$

   __Finite Union__: Suppose $A_1, A_2 \in \cup_i \mathcal{F}_i$, since $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$, assume $A_1 \in \mathcal{F}_i$, $A_2 \in \mathcal{F}_j$, such that,

$$A_1, A_2 \in \mathcal{F}_{\max(i,j)}.$$

   Since each $\mathcal{F}_i$ is σ-algebra,

$$A_1 \cup A_2 \in \mathcal{F}_i \subset \cup_i \mathcal{F}_i.$$

2. *Let $\mathcal{F}_i$ be a Borel Set of $[1, 2 - \frac{1}{i}]$. Suppose $A_i = [1, 2 - \frac{1}{i}] \in \mathcal{F}_i$,*

$$\cup_i A_i = [1, 2) \notin \cup_i \mathcal{F}_i.$$

11.2 Laws of Large Numbers

11.3 Central Limit Theorems

✍ **Exercise 11.2** Let $g \geq 0$ be continuous. If $X_n \xrightarrow{d} X_\infty$, then

$$\liminf_{n \to \infty} Eg(X_n) \geq Eg(X_\infty).$$

Solution *Let $Y_n \overset{d}{=} X_n, 1 \leq n \leq \infty$ with $Y_n \xrightarrow{a.s.} Y_\infty$ (Lemma 7.1). Since $g \geq 0$ be continuous, $g(Y_n) \xrightarrow{a.s.} g(Y_\infty)$ and $g(Y_n) \geq 0$ (Theorem 7.2), and the Fatou's Lemma (6.6) implies,*

$$\liminf_{n \to \infty} Eg(X_n) = \liminf_{n \to \infty} Eg(Y_n) \geq E\left(\liminf_{n \to \infty} g(Y_n)\right)$$

$$= Eg(Y_\infty) = Eg(X_\infty).$$

✍ **Exercise 11.3** Suppose g, h are continuous with $g(x) > 0$, and $|h(x)|/g(x) \to 0$ as $|x| \to \infty$. If $F_n \xrightarrow{d} F$ and $\int g(x) \mathrm{d}F_n(x) \le C < \infty$, then

$$\int h(x) \mathrm{d}F_n(x) \to \int h(x) \mathrm{d}F(x).$$

Solution

$$\left| \int h(x)\mathrm{d}F_n(x) - \int h(x)\mathrm{d}F(x) \right| = \left| \int_{x \in [-M,M]} h(x)\mathrm{d}F_n(x) + \int_{x \notin [-M,M]} h(x)\mathrm{d}F_n(x) \right.$$
$$\left. - \int_{x \in [-M,M]} h(x)\mathrm{d}F(x) - \int_{x \notin [-M,M]} h(x)\mathrm{d}F(x) \right|$$
$$\le \left| \int_{x \in [-M,M]} h(x)\mathrm{d}F_n(x) - \int_{x \in [-M,M]} h(x)\mathrm{d}F(x) \right|$$
$$+ \left| \int_{x \notin [-M,M]} h(x)\mathrm{d}F_n(x) - \int_{x \notin [-M,M]} h(x)\mathrm{d}F(x) \right|.$$

Let $X_n, 1 \le n < \infty$, with distribution F_n, so that $X_n \xrightarrow{a.s.} X$ (Lemma 7.1).

$$\left| \int_{x \in [-M,M]} h(x)\mathrm{d}F_n(x) - \int_{x \in [-M,M]} h(x)\mathrm{d}F(x) \right| = \left| E\left(h(X_n) - h(X)\right) I_{x \in [-M,M]} \right|.$$

By Continuity Mapping Theorem (7.2), $\lim_{n \to \infty} \left| E\left(h(X_n) - h(X)\right) I_{x \in [-M,M]} \right| = 0$.
Since

$$h(x) I_{x \notin [-M,M]} \le g(x) \sup_{x \notin [-M,M]} \frac{h(x)}{g(x)},$$

and by Exercise 11.2

$$Eg(X) \le \liminf_{n \to \infty} Eg(X_n) = \liminf_{n \to \infty} \int g(x)\mathrm{d}F_n(x) \le C < \infty,$$

$$\left| \int_{x \notin [-M,M]} h(x)\mathrm{d}F_n(x) - \int_{x \notin [-M,M]} h(x)\mathrm{d}F(x) \right| = \left| E\left(h(X_n) - h(X)\right) I_{x \notin [-M,M]} \right|$$
$$\le 2E \max(h(Xn), h(X)) I_{x \notin [-M,M]} \le 2C \sup_{x \notin [-M,M]} \frac{h(x)}{g(x)}.$$

Hence, let $M \to \infty$,

$$\lim_{n \to \infty} \left| \int h(x)\mathrm{d}F_n(x) - \int h(x)\mathrm{d}F(x) \right| \le 2C \sup_{x \notin [-M,M]} \frac{h(x)}{g(x)} \to 0,$$

which means,

$$\int h(x)\mathrm{d}F_n(x) \to \int h(x)\mathrm{d}F(x).$$

✍ **Exercise 11.4** Let $X_1, X_2, \ldots$ be i.i.d. with $EX_i = 0$ and $EX_i^2 = \sigma^2 \in (0, \infty)$. Then

$$\sum_{m=1}^{n} X_m / \left(\sum_{m=1}^{n} X_m^2 \right)^{1/2} \xrightarrow{d} \chi.$$

✎ **Exercise 11.5** Show that if $|X_i| \leq M$ and $\sum_n \text{Var}(X_n) = \infty$, then

$$(S_n - E S_n) / \sqrt{\text{Var}(S_n)} \xrightarrow{d} \chi.$$

✎ **Exercise 11.6** Suppose $EX_i = 0, EX_i^2 = 1$ and $E|X_i|^{2+\delta} \leq C$ for some $0 < \delta, C < \infty$. Show that

$$S_n / \sqrt{n} \xrightarrow{d} \chi.$$

Part V

Stochastic Process

Chapter 12 Martingales

12.1 Conditional Expectation

> **Definition 12.1 (Conditional Expectation)** ♣

Example 12.1

1. If $X \in \mathcal{F}$, then
$$E\left(X \mid \mathcal{F}\right) = X.$$

2. If X is independent of $\mathcal{F}$, then
$$E\left(X \mid \mathcal{F}\right) = E(X).$$

3. If $\Omega_1, \Omega_2, \ldots$ is a finite or infinite partition of Ω into disjoint sets, each of which has positive probability, and let $\mathcal{F} = \sigma\left(\Omega_1, \Omega_2, \ldots\right)$, then
$$E\left(X \mid \mathcal{F}\right) = \frac{E\left(X; \Omega_i\right)}{P\left(\Omega_i\right)} \quad \text{on } \Omega_i.$$

Property

12.2 Martingales

Let $\mathcal{F}_n$ be a filtration, i.e., an increasing sequence of σ-fields.

> **Definition 12.2 (Martingale)**
>
> *A sequence $\{X_n\}$ of real-valued random variables is said to be a martingale with respect to $\mathcal{F}_n$, if*
>
> 1. *X_n is integrable, i.e., $E\left|X_n\right| < \infty$*
> 2. *X_n is adapted to $\mathcal{F}_n$, i.e., $\forall n, X_n \in \mathcal{F}_n$*
> 3. *X_n satisfies the martingale condition, i.e.,*
> $$E\left(X_{n+1} \mid \mathcal{F}_n\right) = X_n, \quad \forall n \tag{12.1}$$ ♣

⚗ **Note** *If in the last definition $=$ is replaced by $\leq$ or $\geq$, then X is said to be a supermartingale or submartingale, respectively.*

Example 12.2 Linear Martingale

Example 12.3 Quadratic Martingale

Example 12.4 Exponential Martingale

Example 12.5 Random Walk Suppose $X_n = X_0 + \xi_1 + \cdots + \xi_n$, where X_0 is constant, ξ_m are independent and have $E\xi_m = 0, \sigma_m^2 = E\xi_m^2 < \infty$. Let $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$ for $n \geq 1$ and take $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Show X_n is a martingale, and X_n^2 is a submartingale.

Proof It is obvious that,

$$E|X_n| < \infty, \quad X_n \in \mathcal{F}_n$$

Since ξ_{n+1} is independent of $\mathcal{F}_n$, so using the linearity of conditional expectation, (4.1.1), and Example 4.1.4,

$$E(X_{n+1} \mid \mathcal{F}_n) = E(X_n \mid \mathcal{F}_n) + E(\xi_{n+1} \mid \mathcal{F}_n) = X_n + E\xi_{n+1} = X_n$$

So X_n is a martingale, and Theorem 4.2.6 implies X_n^2 is a submartingale.

Note *If we let $\lambda = x^2$ and apply Theorem 4.4.2 to X_n^2, we get Kolmogorov's maximal inequality, Theorem 2.5.5:*

$$P\left(\max_{1 \leq m \leq n} |X_m| \geq x\right) \leq x^{-2} \operatorname{var}(X_n) \tag{12.2}$$

Theorem 12.1 (Orthogonality of Martingale Increments)

♡

Theorem 12.2 (Conditional Variance Formula)

♡

Definition 12.3 (Predictable Sequence)

♣

Definition 12.4 (Stopping Time)

♣

Theorem 12.3 (Martingale Convergence Theorem)

♡

12.3 Doob's Inequality

Theorem 12.4 (Doob's Decomposition)

♡

Theorem 12.5 (Doob's Inequality)

♡

Theorem 12.6 (L^p Maximum Inequality)

♡

12.4 Uniform Integrability

12.5 Optional Stopping Theorems

Chapter 13 Markov Chains

13.1 Markov Chain

Definition 13.1 (Markov Chain, Simple)

A sequence $\{X_n\}$ of real-valued random variables is said to be a Markov chain, if for any states $i_0, \ldots i_{n-1}, i,$ and j

$$P\left(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots X_0 = i_0\right) = P\left(X_{n+1} = j \mid X_n = i\right)$$

(13.1)

and the transition probability is

$$p(i, j) = P\left(X_{n+1} = j \mid X_n = i\right)$$

(13.2) ♣

Example 13.1 Random Walk Suppose $X_n = X_0 + \xi_1 + \cdots + \xi_n$, where X_0 is constant, $\xi_m \in \mathbb{Z}^d$ are independent with distribution μ. Show X_n is a Markov chain with transition probability,

$$p(i, j) = \mu(\{j - i\})$$

Proof Since ξ_m are independent with distribution μ,

$$P\left(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots X_0 = i_0\right)$$

$$= P\left(X_n + \xi_{n+1} = j \mid X_n = i\right) = P\left(\xi_{n+1} = j - i\right) = \mu(\{j - i\})$$

Definition 13.2 (Branching Processes)

Let $\xi_i^n, i, n \geq 1,$ be i.i.d. nonnegative integer-valued random variables. Define a sequence $Z_n, n \geq 0$ by $Z_0 = 1$ and

$$Z_{n+1} = \begin{cases} \xi_1^{n+1} + \cdots + \xi_{Z_n}^{n+1} & Z_n > 0 \\ 0 & Z_n = 0 \end{cases}$$

(13.3)

Z_n is called a Branching process. ♣

Note *The idea behind the definitions is that Z_n is the number of individuals in the n-th generation, and each member of the n-th generation gives birth independently to an identically distributed number of children.*

Example 13.2 Branching Processes Show branching process is a Markov chain with transition probability,

$$p(i, j) = P\left(\sum_{k=1}^{i} \xi_k = j\right)$$

Proof Since ξ_k^n are independent with identically distribution,

$$P\left(Z_{n+1} = j \mid Z_n = i, Z_{n-1} = i_{n-1}, \ldots Z_0 = i_0\right)$$

$$= P\left(\sum_{k=1}^{Z_n} \xi_k^{n+1} = j \mid Z_n = i\right) = P\left(\sum_{k=1}^{i} \xi_k = j\right)$$

Suppose $(S, \mathcal{S})$ be a measurable space, which will be the state space for our Markov chain.

Definition 13.3 (Transition Probability)

A function $p : S \times S \to \mathbf{R}$ is said to be a transition probability, if

  1. For each $x \in S$, $A \to p(x, A)$ is a probability measure on $(S, \mathcal{S})$
  2. For each $A \in \mathcal{S}$, $x \to p(x, A)$ is a measurable function

♣

Definition 13.4 (Markov Chain)

A sequence $\{X_n\}$ of real-valued random variables with transition probability p is said to be a Markov chain with respect to $\mathcal{F}_n$, if

$$P\left(X_{n+1} \in B \mid \mathcal{F}_n\right) = p\left(X_n, B\right) \tag{13.4}$$

♣

Remark Given a transition probability p and an initial distribution μ on $(S, \mathcal{S})$, the consistent set of finite dimensional distributions is

$$P\left(X_j \in B_j, 0 \le j \le n\right) = \int_{B_0} \mu\left(\mathrm{d}x_0\right) \int_{B_1} p\left(x_0, \mathrm{d}x_1\right) \cdots \int_{B_n} p\left(x_{n-1}, \mathrm{d}x_n\right) F \tag{13.5}$$

13.2 Markov Properties

Definition 13.5 (Shift Operator)

♣

Theorem 13.1 (Markov Property)

♡

Corollary 13.1 (Chapman-Kolmogorov Equation)

♡

Theorem 13.2 (Strong Markov Property)

♡

13.3 Recurrence and Transience

Let $T_y^0 = 0$, and for $k \geq 1$, and

$$T_y^k = \inf \left\{ n > T_y^{k-1} : X_n = y \right\} \tag{13.6}$$

then T_y^k is the time of the k-th return to y, where $T_y^1 > 0$, so any visit at time 0 does not count.

Let

$$\rho_{xy} = P_x \left(T_y < \infty \right) \tag{13.7}$$

and we have

$$P_x \left(T_y^k < \infty \right) = \rho_{xy} \rho_{yy}^{k-1} \tag{13.8}$$

Proof

Let

$$N(y) = \sum_{n=1}^{\infty} 1_{(X_n = y)} \tag{13.9}$$

be the number of visits to y at positive times.

> **Definition 13.6 (Recurrent)**
>
> *A state y is said to be recurrent if $\rho_{yy} = 1$.* ♣

Property *The recurrent state y has the following properties*

1. y is recurrent if and only if

$$E_y N(y) = \infty.$$

2. If x is recurrent and $\rho_{xy} > 0$, then y is recurrent and $\rho_{yx} = 1$.

> **Definition 13.7**
>
> *A state y is said to be transient if $\rho_{yy} < 1$.* ♣

Property *The transient state y has the following properties*

1. If y is transient, then

$$E_x N(y) < \infty, \quad \forall x.$$

Proof

$$E_x N(y) = \sum_{k=1}^{\infty} P_x(N(y) \geq k) = \sum_{k=1}^{\infty} P_x \left(T_y^k < \infty \right)$$

$$= \sum_{k=1}^{\infty} \rho_{xy} \rho_{yy}^{k-1} = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty$$

Definition 13.8 (Closed State Set)

A set C of states is said to be closed, if

$$x \in C, \rho_{xy} > 0 \Rightarrow y \in C. \tag{13.10}$$

♣

Definition 13.9 (Irreducible State Set)

A set D of states is said to be irreducible, if

$$x, y \in D \Rightarrow \rho_{xy} > 0. \tag{13.11}$$

♣

Theorem 13.3

Let C be a finite closed set, then

1. C contains a recurrent state.
2. If C is irreducible, then all states in C are recurrent.

♡

Theorem 13.4

Suppose $C_x = \{y : \rho_{xy} > 0\}$, then C_x is an irreducible closed set.

♡

Proof If $y, z \in C_x$, then $\rho_{yz} \geq \rho_{yx}\rho_{xz} > 0$. If $\rho_{yw} > 0$, then $\rho_{xw} \geq \rho_{xy}\rho_{yw} > 0$, so $w \in C_x$.

Example 13.3 A Seven-state Chain Consider the transition probability,

	1	2	3	4	5	6	7
1	.3	0	0	0	.7	0	0
2	.1	.2	.3	.4	0	0	0
3	0	0	.5	.5	0	0	0
4	0	0	0	.5	0	.5	0
5	.6	0	0	0	.4	0	0
6	0	0	0	.1	0	.1	.8
7	0	0	0	1	0	0	0

try to identify the states that are recurrent and those that are transient.

Proof $\{2, 3\}$ are transition states, and $\{1, 4, 5, 6, 7\}$ are recurrent states.

Remark Suppose S is finite, for $x \in S$,

1. x is transient, if

$$\exists y, \rho_{xy} > 0, \text{ s.t. } \quad \rho_{yx} = 0$$

2. x is recurrent, if

$$\forall y, \rho_{xy} > 0, \text{ s.t. } \quad \rho_{yx} > 0$$

13.4 Stationary Measures

13.5 Asymptotic Behavior

13.6 Ergodic Theorems

Definition 13.10 (Stationary Sequence)

♣

Theorem 13.5 (Ergodic Theorem)

♡

Example 13.4

Chapter 14 Brownian Motion

> **Definition 14.1 (Brownian Motion (1))**
>
> *A real-valued stochastic process $B(t), t \geq 0$ is said to be Brownian motion, if*
>
> *1. for any $0 = t_0 \leq t_1 \leq \ldots \leq t_n$ the increments*
>
> $$B(t_1) - B(t_0), \ldots, B(t_n) - B(t_{n-1})$$
>
> *are independent*
>
> *2. for any $s, t \geq 0$ and Borel sets $A \in \mathbb{R}$,*
>
> $$P(B(s+t) - B(s) \in A) = \int_A (2\pi t)^{-1/2} \exp\left(-x^2/2t\right) \, \mathrm{d}x \qquad (14.1)$$
>
> *3. the sample paths $t \to B(t)$ are a.s. continuous* ♣

Property *For a one-dimensional Brownian motion, if $B(0) = 0$, then we have the following properties*

1. $EB_t = 0, \operatorname{Var}(B_t) = t, \quad t \geq 0$.
2. $\operatorname{Cov}(B_s, B_t) = s, \operatorname{Corr}(B_s, B_t) = \sqrt{\frac{s}{t}}, \quad \forall 0 \leq s \leq t$.

Proof

1. Since $B_t = B_t - B_0 \sim N(0, t)$, then we have

$$EB_t = 0, \operatorname{Var}(B_t) = t$$

2. Suppose $0 \leq s \leq t$,

$$\operatorname{Cov}(B_s, B_t) = E\left[(B_s - EB_s)(B_t - EB_t)\right] = EB_s B_t$$

Let $B_t = (B_t - B_s) + B_s$, we have

$$EB_s B_t = E\left[B_s \cdot ((B_t - B_s) + B_s)\right]$$
$$= E\left[B_s \cdot (B_t - B_s)\right] + EB_s^2$$

Since $B_s = B_s - B_0$ and $B_t - B_s$ are independent,

$$E\left[B_s \cdot (B_t - B_s)\right] = EB_s \cdot E\left[B_t - B_s\right] = 0$$

Thus

$$\operatorname{Cov}(B_s, B_t) = EB_s^2 = s$$

And

$$\operatorname{Corr}(B_s, B_t) = \frac{\operatorname{Cov}(B_s, B_t)}{\sigma_{B_s} \sigma_{B_t}} = \frac{s}{\sqrt{st}} = \sqrt{\frac{s}{t}}$$

A second equivalent definition of Brownian motion are as followed,

> **Definition 14.2 (Brownian Motion (2))**
>
> *A real-valued stochastic process $B(t), t \geq 0$, **starting from** 0, is said to be Brownian motion, if*
>
> 1. *$B(t)$ is a Gaussian process[a]*
> 2. *$\forall s, t \geq 0, EB_s = 0$ and $EB_s B_t = s \wedge t$*
> 3. *the sample paths $t \to B(t)$ are a.s. continuous*
>
> ---
>
> [a]Gaussian process, i.e., all its finite dimensional distributions are multivariate normal.
> ♣

14.1 Markov Properties

14.2 Martingales

Example 14.1 Quadratic Martingale Suppose B_t is a Brownian motion, then

$$B_t^2 - t$$

is a martingale.

Proof Let $B_t^2 = (B_s + B_t - B_s)^2$, we have

$$
\begin{aligned}
E_x \left(B_t^2 \mid \mathcal{F}_s \right) &= E_x \left(B_s^2 + 2B_s \left(B_t - B_s \right) + \left(B_t - B_s \right)^2 \mid \mathcal{F}_s \right) \\
&= B_s^2 + 2B_s E_x \left(B_t - B_s \mid \mathcal{F}_s \right) + E_x \left(\left(B_t - B_s \right)^2 \mid \mathcal{F}_s \right) \\
&= B_s^2 + 0 + (t - s)
\end{aligned}
$$

since $B_t - B_s$ is independent of $\mathcal{F}_s$ and has mean 0 and variance $t - s$.

Example 14.2 Exponential Martingale Suppose B_t is a Brownian motion, then

$$\exp \left(\theta B_t - \left(\theta^2 t / 2 \right) \right)$$

is a martingale.

Proof Let $B_t = B_t - B_s + B_s$, then

$$
\begin{aligned}
E_x \left(\exp \left(\theta B_t \right) \mid \mathcal{F}_s \right) &= \exp \left(\theta B_s \right) E \left(\exp \left(\theta \left(B_t - B_s \right) \right) \mid \mathcal{F}_s \right) \\
&= \exp \left(\theta B_s \right) \exp \left(\theta^2 (t - s) / 2 \right)
\end{aligned}
$$

since $B_t - B_s$ is independent of $\mathcal{F}_s$ and has mean 0 and variance $t - s$. Thus

$$
\begin{aligned}
E_x \left(\exp \left(\theta B_t - \left(\theta^2 t / 2 \right) \right) \mid \mathcal{F}_s \right) &= E_x \left(\exp \left(\theta B_t \right) \mid \mathcal{F}_s \right) \cdot \exp \left(- \left(\theta^2 t / 2 \right) \right) \\
&= \exp \left(\theta B_s - \left(\theta^2 s / 2 \right) \right)
\end{aligned}
$$

> **Theorem 14.1 (Lévy's Martingale Characterization)**
>
> Let $B(t), t \geq 0$, be a real-valued stochastic process and let $\mathcal{F}_t = \sigma(B_s, s \leq t)$ be the filtration generated by it. Then $B(t)$ is a Brownian motion if and only if
>
> 1. $B(0) = 0$ a.s.
> 2. the sample paths $t \rightarrow B(t)$ are continuous a.s.
> 3. $B(t)$ is a martingale with respect to $\mathcal{F}_t$
> 4. $|B(t)|^2 - t$ is a martingale with respect to $\mathcal{F}_t$

14.3 Sample Paths

Let $0 = t_0^n < t_1^n < \cdots < t_n^n = T$, where $t_i^n = \frac{iT}{n}$ be a partition of the interval $[0, T]$ into n equal parts, and

$$\Delta_i^n B = B\left(t_{i+1}^n\right) - B\left(t_i^n\right) \tag{14.2}$$

be the corresponding increments of the Brownian motion $B(t)$.

> **Theorem 14.2**
>
> $$\lim_{n \to \infty} \sum_{i=0}^{n-1} \left(\Delta_i^n B\right)^2 = T \quad in \quad L^2 \tag{14.3}$$

Proof Since the increments $\Delta_i^n B$ are independent and

$$E\left(\Delta_i^n B\right) = 0, \quad E\left(\left(\Delta_i^n B\right)^2\right) = \frac{T}{n}, \quad E\left(\left(\Delta_i^n B\right)^4\right) = \frac{3T^2}{n^2}$$

it follows that

$$
\begin{aligned}
E\left(\left[\sum_{i=0}^{n-1} \left(\Delta_i^n B\right)^2 - T\right]^2\right) &= E\left(\left[\sum_{i=0}^{n-1} \left(\left(\Delta_i^n B\right)^2 - \frac{T}{n}\right)\right]^2\right) \\
&= \sum_{i=0}^{n-1} E\left[\left(\left(\Delta_i^n B\right)^2 - \frac{T}{n}\right)^2\right] \\
&= \sum_{i=0}^{n-1} \left[E\left(\left(\Delta_i^n B\right)^4\right) - \frac{2T}{n} E\left(\left(\Delta_i^n B\right)^2\right) + \frac{T^2}{n^2}\right] \\
&= \sum_{i=0}^{n-1} \left[\frac{3T^2}{n^2} - \frac{2T^2}{n^2} + \frac{T^2}{n^2}\right] \\
&= \frac{2T^2}{n} \to 0, \quad n \to \infty
\end{aligned}
$$

> **Definition 14.3 (Variation)**
>
> *The variation of a function $f : [0, T] \to \mathbb{R}$ is defined to be*
>
> $$\limsup_{\Delta t \to 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)| \tag{14.4}$$
>
> *where $t = (t_0, t_1, \ldots, t_n)$ is a partition of $[0, T]$, i.e. $0 = t_0 < t_1 < \cdots < t_n = T$, and where*
>
> $$\Delta t = \max_{i=0,\ldots,n-1} |t_{i+1} - t_i| \tag{14.5}$$
>
> ♣

> **Theorem 14.3**
>
> *The variation of the paths of $B(t)$ is infinite a.s..*
>
> ♡

Proof   Consider the sequence of partitions $t^n = (t_0^n, t_1^n, \ldots, t_n^n)$ of $[0, T]$ into n equal parts. Then

$$\sum_{i=0}^{n-1} |\Delta_i^n B|^2 \leq \left(\max_{i=0,\ldots,n-1} |\Delta_i^n B| \right) \sum_{i=0}^{n-1} |\Delta_i^n B|$$

Since the paths of $B(t)$ are a.s. continuous on $[0, T]$,

$$\lim_{n \to \infty} \left(\max_{i=0,\ldots,n-1} |\Delta_i^n B| \right) = 0 \quad \text{a.s.}$$

By Theorem 14.2, we have

$$\lim_{n \to \infty} \sum_{i=0}^{n-1} \left(\Delta_i^n B \right)^2 = T \quad \text{in} \quad L^2$$

Since every sequence of random variables convergent in L^2 has a subsequence convergent a.s. There is a subsequence $t^{n_k} = (t_0^{n_k}, t_1^{n_k}, \ldots, t_{n_k}^{n_k})$ of partitions such that

$$\lim_{k \to \infty} \sum_{i=0}^{n_k-1} \left| \Delta_i^{n_k} B \right|^2 = T \quad \text{a.s.}$$

Since

$$\sum_{i=0}^{n_k-1} \left| \Delta_i^{n_k} B \right| \geq \frac{\sum_{i=0}^{n_k-1} \left| \Delta_i^{n_k} B \right|^2}{\max_{i=0,\ldots,n_k-1} \left| \Delta_i^{n_k} B \right|}$$

hence,

$$\lim_{k \to \infty} \sum_{i=0}^{n_k-1} \left| \Delta_i^{n_k} B \right| = \infty \quad \text{a.s.}$$

while

$$\lim_{k \to \infty} \Delta t^{n_k} = \lim_{k \to \infty} \frac{T}{n_k} = 0$$

14.4 Itô Stochastic Calculus

> **Definition 14.4 (Random Step Process)**
>
> ♣

> **Definition 14.5 (Itô Stochastic Integral)**
>
> *For any $T > 0$ we shall denote by M_T^2 the space of all stochastic processes $f(t), t \geq 0$ such that*
>
> $$1_{[0,T)} f \in M^2$$
>
> *The Itô stochastic integral (from 0 to T) of $f \in M_T^2$ is defined by*
>
> $$I_T(f) = I\left(1_{[0,T)} f\right) \tag{14.6}$$
>
> *which can be denoted by*
>
> $$\int_0^T f(t)\,\mathrm{d}B(t) \tag{14.7}$$
>
> ♣

Property *The Itô Stochastic Integral has the following properties:*

1. Linearity: For $\forall f, g \in M_t^2$, $\forall \alpha, \beta \in \mathbb{R}$,

$$\int_0^t \left(\alpha f(r) + \beta g(r)\right)\mathrm{d}B(r) = \alpha \int_0^t f(r)\,\mathrm{d}B(r) + \beta \int_0^t g(r)\,\mathrm{d}B(r) \tag{14.8}$$

2. Isometry: For $\forall f \in M_t^2$,

$$E\left(\left|\int_0^t f(r)\,\mathrm{d}B(r)\right|^2\right) = E\left(\int_0^t |f(r)|^2\,\mathrm{d}r\right) \tag{14.9}$$

3. Martingale Property: For $\forall f \in M_t^2$ and $\forall 0 \leq s < t$,

$$E\left(\int_0^t f(r)\,\mathrm{d}B(r) \mid \mathcal{F}_s\right) = \int_0^s f(r)\,\mathrm{d}B(r) \tag{14.10}$$

Proof

> **Definition 14.6 (Itô Process)**
>
> *A stochastic process $\xi(t), t \geq 0$ is said to be an Itô process if it has a.s. continuous paths and can be represented as*
>
> $$\xi(T) = \xi(0) + \int_0^T a(t)\,\mathrm{d}t + \int_0^T b(t)\,\mathrm{d}B(t) \quad a.s. \tag{14.11}$$
>
> *where $b(t)$ is a process belonging to M_T^2 for all $T > 0$ and $a(t)$ is a process adapted to the filtration $\mathcal{F}_t$ such that*
>
> $$\int_0^T |a(t)|\,\mathrm{d}t < \infty \quad a.s. \tag{14.12}$$

for all $T \geq 0$. The Itô process is denoted by

$$d\xi(t) = a(t)\,dt + b(t)\,dB(t) \tag{14.13}$$ ♣

Remark The class of all adapted processes $a(t)$ satisfying 14.12 for some $T > 0$ will be denoted by $\mathcal{L}_T^1$.

Theorem 14.4 (Itô Formula)

Suppose that $F(t, x)$ is a real-valued function with continuous partial derivatives $F_t'(t, x)$, $F_x'(t, x)$ and $F_{xx}''(t, x)$ for all $t \geq 0$ and $x \in \mathbb{R}$.

1. If $\xi(t)$ be an Itô process

$$\xi(t) = \xi(0) + \int_0^t a(s)\,ds + \int_0^t b(s)\,dB(s)$$

and the process $b(t)F_x'(t, \xi(t))$ belongs to M_T^2 for all $T \geq 0$. Then $F(t, \xi(t))$ is an Itô process such that

$$dF(t, \xi(t)) = \left(F_t'(t, \xi(t)) + F_x'(t, \xi(t))a(t) + \frac{1}{2}F_{xx}''(t, \xi(t))b(t)^2 \right) dt$$

$$+ F_x'(t, \xi(t))b(t)\,dB(t)$$

$$\tag{14.14}$$

2. If $\xi(t)$ be an Brownian Motion, such that $\xi(t) = B(t)$, and the process $F_x'(t, B(t))$ belongs to M_T^2 for all $T \geq 0$. Then $F(t, B(t))$ is an Itô process such that

$$dF(t, B(t)) = \left(F_t'(t, B(t)) + \frac{1}{2}F_{xx}''(t, B(t)) \right) dt + F_x'(t, B(t))\,dB(t) \tag{14.15}$$ ♡

Example 14.3 Exponential Martingale Show that the exponential martingale

$$X(t) = e^{B(t)}e^{-\frac{t}{2}}$$

is an Itô process, and satisfies the equation

$$dX(t) = X(t)\,dB(t)$$

Proof Let $F(t, x) = e^x e^{-\frac{t}{2}}$, then we have

$$F_t'(t, x) = -\frac{1}{2}F(t, x), \quad F_x'(t, x) = F(t, x), \quad F_{xx}''(t, x) = F(t, x)$$

thus, by Itô Formula, we have

$$\begin{aligned}
dX(t) = dF(t, B(t)) &= \left(F_t'(t, B(t)) + \frac{1}{2}F_{xx}''(t, B(t)) \right) dt + F_x'(t, B(t))\,dB(t) \\
&= \left(-\frac{1}{2}F(t, B(t)) + \frac{1}{2}F(t, B(t)) \right) dt + F(t, B(t))\,dB(t) \\
&= X(t)\,dB(t)
\end{aligned}$$

Example 14.4
Example 14.5

Chapter 15 Exercises for Probability Theory and Examples

Part VI

Statistics Inference

Chapter 16 Introduction

16.1 Populations and Samples

16.2 Statistics

16.2.1 Sufficient Statistics

> **Definition 16.1 (Sufficient Statistics)**
>
> *A statistic T is said to be sufficient for X, or for the family $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ of possible distributions of X, or for θ, if the conditional distribution of X given $T = t$ is independent of θ for all t.* ♣

> **Theorem 16.1 (Fisher–Neyman Factorization Theorem)**
>
> *If the probability density function is $p_\theta(x)$, then T is sufficient for θ if and only if nonnegative functions g and h can be found such that*
>
> $$p_\theta(x) = h(x)g_\theta[T(x)].$$ ♡

Proof

16.2.2 Complete Statistics

> **Definition 16.2 (Complete Statistics)**
>
> *A statistic T is said to be complete, if $Eg(T) = 0$ for all θ and some function g implies that $P(g(T) = 0 \mid \theta) = 1$ for all θ.* ♣

16.3 Estimators

> **Definition 16.3 (Estimator)**
>
> *An estimator is a real-valued function defined over the sample space, that is*
>
> $$\delta : X \to \mathbb{R}. \tag{16.1}$$
>
> *It is used to estimate an estimand, θ, a real-valued function of the parameter.* ♣

Unbiasedness

> **Definition 16.4 (Unbiasedness)**
>
> An estimator $\hat{\theta}$ of θ is unbiased if
> $$E\hat{\theta} = \theta, \quad \forall \theta \in \Theta. \tag{16.2}$$

Note

- *Unbiased estimators of θ may not exist.*
-

Example 16.1 Nonexistence of Unbiased Estimator

Consistency

> **Definition 16.5 (Consistency)**
>
> An estimator $\hat{\theta}_n$ of θ is consistent if
> $$\lim_{n \to \infty} P\left(\left| \hat{\theta}_n - \theta \right| > \varepsilon \right) = 0, \quad \forall \varepsilon > 0, \tag{16.3}$$
> that is,
> $$\hat{\theta}_n \xrightarrow{p} \theta. \tag{16.4}$$

Note

1. *Unbiased But Consistent*
2. *Biased But Not Consistent*

Example 16.2 Common Mean

Asymptotic Normality

> **Definition 16.6 (Asymptotic Normality)**
>
> An estimator $\hat{\theta}_n$ of θ is asymptotic normality if
> $$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \sigma_\theta^2\right). \tag{16.5}$$

Efficiency

Definition 16.7 (Efficiency)

♣

Robustness

Definition 16.8 (Robustness)

♣

Chapter 17 Maximum Likelihood Estimator

Suppose that $\mathbf{X}_n = (X_1, \ldots, X_n)$, where the X_i are i.i.d. with common density $p(x; \theta_0) \in \mathcal{P} = \{p(x; \theta) : \theta \in \Theta\}$.

We assume that

θ_0 is identified in the sense that if $\theta \neq \theta_0$ and $\theta \in \Theta$, then $p(x; \theta) \neq p(x; \theta_0)$ with respect to the dominating measure μ.

For fixed $\theta \in \Theta$, the joint density of $\mathbf{X}_n$ is equal to the product of the individual densities, i.e.,

$$p(\mathbf{X}_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta). \tag{17.1}$$

The maximum likelihood estimate for observed $\mathbf{X}_n$ is the value $\theta \in \Theta$ which maximizes $L(\theta; X_n) := p(\mathbf{X}_n; \theta)$, i.e.,

$$\hat{\theta}(\mathbf{X}_n) = \max_{\theta \in \Theta} L(\theta; X_n). \tag{17.2}$$

Equivalently, the MLE can be taken to be the maximum of the standardized log-likelihood,

$$\frac{l(\theta; \mathbf{X}_n)}{n} = \frac{\log L(\theta; \mathbf{X}_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta; X_i). \tag{17.3}$$

Define

$$Q(\theta; \mathbf{X}_n) := \frac{1}{n} \sum_{i=1}^{n} l(\theta; X_i),$$
$$\hat{\theta}(\mathbf{X}_n) := \max_{\theta \in \Theta} Q(\theta; \mathbf{X}_n). \tag{17.4}$$

17.1 Consistency of MLE

By the Weak Law of Large Numbers (Theorem 8.5), we can get,

$$\frac{1}{n} \sum_{i=1}^{n} l(\theta; X_i) \xrightarrow{p} E[l(\theta; X)]. \tag{17.5}$$

Suppose $Q_0(\theta) = E[l(\theta; X)]$, then we will show that $Q_0(\theta)$ is maximized at θ_0 (i.e., the truth).

> **Lemma 17.1**
>
> *If θ_0 is identified and $E_{\theta_0} [|\log p(X; \theta)|] < \infty, \forall \theta \in \Theta$, then $Q_0(\theta)$ is uniquely maximized at $\theta = \theta_0$.* ♡

Proof

> **Theorem 17.1 (Consistency of MLE)**
>
> *Suppose that $Q(\theta; X_n)$ is continuous in θ and there exists a function $Q_0(\theta)$ such that*
>   1. *$Q_0(\theta)$ is uniquely maximized at θ_0.*
>   2. *Θ is compact.*
>   3. *$Q_0(\theta)$ is continuous in θ.*
>   4. *$Q(\theta; X_n)$ converges uniformly in probability to $Q_0(\theta)$.*
>
> *then*
> $$\hat{\theta}(X_n) \xrightarrow{p} \theta_0. \tag{17.6}$$
> ♡

Proof $\forall \varepsilon > 0$, let
$$\Theta(\epsilon) = \{\theta : \|\theta - \theta_0\| < \epsilon\}.$$

Since $\Theta(\epsilon)$ is an open set, then $\Theta \cap \Theta(\epsilon)^C$ is a compact set (Assumption 2).

Since $Q_0(\theta)$ is a continuous function (Assumption 3), then
$$\theta^* := \sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} \{Q_0(\theta)\}$$
is a achieved for a θ in the compact set.

Since θ_0 is the unique maximized, let
$$Q_0(\theta_0) - Q_0(\theta^*) = \delta > 0.$$

1. For $\theta \in \Theta \cap \Theta(\epsilon)^C$. Let $A_n = \{\sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} |Q(\theta; X_n) - Q_0(\theta)| < \frac{\delta}{2}\}$, then
$$A_n \Rightarrow Q(\theta; X_n) < Q_0(\theta) + \frac{\delta}{2}$$
$$\leq Q_0(\theta^*) + \frac{\delta}{2}$$
$$= Q_0(\theta_0) - \frac{\delta}{2}$$

2. For $\theta \in \Theta(\epsilon)$. Let $B_n = \{\sup_{\theta \in \Theta(\epsilon)} |Q(\theta; X_n) - Q_0(\theta)| < \frac{\delta}{2}\}$, then
$$B_n \Rightarrow Q(\theta; X_n) > Q_0(\theta) - \frac{\delta}{2}, \forall \theta \in \Theta(\epsilon)$$

By Assumption 1,
$$Q(\theta_0; X_n) > Q_0(\theta_0) - \frac{\delta}{2}$$

If both A_n and B_n hold, then
$$\hat{\theta} \in \Theta(\epsilon).$$

By Assumption 4, we can concluded that $P(A_n \cap B_n) \to 1$, so
$$P(\hat{\theta} \in \Theta(\epsilon)) \to 1,$$

which means,
$$\hat{\theta}(\mathbf{X}_n) \xrightarrow{p} \theta_0.$$

17.2 Asymptotic Normality of MLE

17.3 Efficiency of MLE

Chapter 18 Minimum-Variance Unbiased Estimator

> **Definition 18.1 (UMVU Estimators)**
>
> *An unbiased estimator $\delta(X)$ of $g(\theta)$ is the uniform minimum variance unbiased (UMVU) estimator of $g(\theta)$ if*
>
> $$Var_\theta \delta(X) \le Var_\theta \delta'(X), \quad \forall \theta \in \Theta, \tag{18.1}$$
>
> *where $\delta'(X)$ is any other unbiased estimator of $g(\theta)$.* ♣

Note *If there exists an unbiased estimator of g, the estimand g will be called U-estimable.*

1. If $T(\mathbf{X})$ is a complete sufficient statistic, estimator $\delta(\mathbf{X})$ that only depends on $T(\mathbf{X})$, then for any U-estimable function $g(\theta)$ with

$$E_\theta \delta(T(\mathbf{X})) = g(\theta), \quad \forall \theta \in \Theta, \tag{18.2}$$

hence, $\delta(T(\mathbf{X}))$ is the unique UMVU estimator of $g(\theta)$.

2. If $T(\mathbf{X})$ is a complete sufficient statistic and $\delta(\mathbf{X})$ is any unbiased estimator of $g(\theta)$, then the UMVU estimator of $g(\theta)$ can be obtained by

$$E\left[\delta(\mathbf{X}) \mid T(\mathbf{X})\right]. \tag{18.3}$$

Example 18.1 Estimating Polynomials of a Normal Variance Let $X_1, \ldots, X_n$ be distributed with joint density

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \xi)^2\right]. \tag{18.4}$$

Discussing the UMVU estimators of $\xi^r, \sigma^r, \xi/\sigma$.

Solution

1. *σ is known:*

   Since $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the complete sufficient statistic of X_i, and

   $$E(\bar{X}) = \xi,$$

   then the UMVU estimator of ξ is $\bar{X}$.

   Therefore, the UMVU estimator of ξ^r is $\bar{X}^r$ and the UMVU estimator of ξ/σ is $\bar{X}/\sigma$.

2. *ξ is known:*

   Since $s^r = \sum (x_i - \xi)^r$ is the complete sufficient statistic of X_i.

   Assume

   $$E\left[\frac{s^r}{\sigma^r}\right] = \frac{1}{K_{n,r}},$$

where $K_{n,r}$ is a constant depends on n, r.

Since $s^2/\sigma^2 \sim Ga(n/2, 1/2) = \chi^2(n)$, then

$$E\left[\frac{s^r}{\sigma^r}\right] = E\left[\left(\frac{s^2}{\sigma^2}\right)^{\frac{r}{2}}\right] = \int_0^\infty x^{\frac{r}{2}} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \, \mathrm{d}x = \frac{\Gamma\left(\frac{n+r}{2}\right)}{\Gamma(\frac{n}{2})} \cdot 2^{\frac{r}{2}}.$$

therefore,

$$K_{n,r} = \frac{\Gamma(\frac{n}{2})}{2^{\frac{r}{2}} \cdot \Gamma\left(\frac{n+r}{2}\right)}.$$

Hence,

$$E\left[s^r K_{n,r}\right] = \sigma^r \text{ and } E[\xi s^{-1} K_{n,-1}] = \xi/\sigma,$$

which means the UMVU estimator of σ^r is $s^r K_{n,r}$ and the UMVU estimator of ξ/σ is $\xi s^{-1} K_{n,-1}$.

3. **Both ξ and σ is unknown:**

   Since $(\bar{X}, s_x^r)$ are the complete sufficient statistic of X_i, where $s_x^2 = \sum (x_i - \bar{X})^r$.

   Since $s_x^2/\sigma^2 \sim \chi^2(n-1)$, then

$$E\left[\frac{s_x^r}{\sigma^r}\right] = \frac{1}{K_{n-1,r}}.$$

   Hence,

$$E\left[s_x^r K_{n-1,r}\right] = \sigma^r,$$

   which means the UMVU estimator of σ^r is $s_x^r K_{n-1,r}$, and

$$E(\bar{X}^r) = \xi^r,$$

   which means the UMVU estimator of ξ^r is $\bar{X}^r$.

   Since $\bar{X}$ and s_x^r are independent, then

$$E[\bar{X} s_x^{-1} K_{n-1,-1}] = \xi/\sigma$$

   which means the UMVU estimator of ξ/σ is $\bar{X} s_x^{-1} K_{n-1,-1}$.

Example 18.2 Let $X_1, \ldots, X_n$ be i.i.d sample from $U(\theta_1 - \theta_2, \theta_1 + \theta_2)$, where $\theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}^+$. Discussing the UMVU estimators of θ_1, θ_2.

Solution *Let $X_{(i)}$ be the i-th order statistic of X_i, then $(X_{(1)}, X_{(n)})$ is the complete and sufficient statistic for (θ_1, θ_2). Thus it suffices to find a function $(X_{(1)}, X_{(n)})$, which is unbiased of (θ_1, θ_2).*

   Let

$$Y_i = \frac{X_i - (\theta_1 - \theta_2)}{2\theta_2} \sim U(0, 1),$$

and

$$Y_{(i)} = \frac{X_{(i)} - (\theta_1 - \theta_2)}{2\theta_2},$$

be the i-th order statistic of Y_i, then we got

$$E[X_{(1)}] = 2\theta_2 E[Y_{(1)}] + (\theta_1 - \theta_2)$$

$$= 2\theta_2 \int_0^1 ny(1-y)^{n-1}\mathrm{d}y + (\theta_1 - \theta_2)$$

$$= \theta_1 - \frac{3n+1}{n+1}\theta_2$$

$$E[X_{(n)}] = 2\theta_2 E[Y_{(n)}] + (\theta_1 - \theta_2)$$

$$= 2\theta_2 \int_0^1 ny^n\mathrm{d}y + (\theta_1 - \theta_2)$$

$$= \theta_1 + \frac{n-1}{n+1}\theta_2$$

Thus,

$$\theta_1 = E\left[\frac{n-1}{4n}X_{(1)} + \frac{3n+1}{4n}X_{(n)}\right],$$

$$\theta_2 = E\left[-\frac{n+1}{4n}X_{(1)} + \frac{n+1}{4n}X_{(n)}\right],$$

which means the UMVU estimator is

$$\hat{\theta}_1 = \frac{n-1}{4n}X_{(1)} + \frac{3n+1}{4n}X_{(n)}, \quad \hat{\theta}_2 = -\frac{n+1}{4n}X_{(1)} + \frac{n+1}{4n}X_{(n)}.$$

Chapter 19 Bayes Estimator

We shall look for some estimators that make the risk function $R(\theta, \delta)$ small in some overall sense. There are two way to solve it: minimize the average risk, minimize the maximum risk.

This chapter will discuss the first method, also known as, Bayes Estimator.

> **Definition 19.1 (Bayes Estimator)**
>
> *The Bayes Estimator δ with respect to Λ is minimizing the Bayes Risk of δ*
>
> $$r(\Lambda, \delta) = \int R(\theta, \delta) \, d\Lambda(\theta) \tag{19.1}$$
>
> *where Λ is the probability distribution.* ♣

In Bayesian arguments, it is important to keep track of which variables are being conditioned on. Hence, the notations are as followed:

- The density of X will be denoted by $X \sim f(x \mid \theta)$.
- The prior distribution will be denoted by $\Pi \sim \pi(\theta \mid \lambda)$ or $\Lambda \sim \gamma(\lambda)$, where λ is another parameter (sometimes called a hyperparameter).
- The posterior distribution, which calculate the conditional distributions as that of θ given x and λ, or λ given x, which is denoted by $\Pi \sim \pi(\theta \mid x, \lambda)$ or $\Lambda \sim \gamma(\lambda \mid x)$, that is

$$\pi(\theta \mid x, \lambda) = \frac{f(x \mid \theta) \pi(\theta \mid \lambda)}{m(x \mid \lambda)}, \tag{19.2}$$

where marginal distributions $m(x \mid \lambda) = \int f(x \mid \theta) \pi(\theta \mid \lambda) \, d\theta$.

> **Theorem 19.1**
>
> *Let Θ have distribution Λ, and given $\Theta = \theta$, let X have distribution P_θ. Suppose, the following assumptions hold for the problem of estimating $g(\Theta)$ with non-negative loss function $L(\theta, d)$,*
>
> - *There exists an estimator δ_0 with finite risk.*
> - *For almost all x, there exists a value $\delta_\Lambda(x)$ minimizing*
>
> $$E\{L[\Theta, \delta(x)] \mid X = x\}. \tag{19.3}$$
>
> *Then, $\delta_\Lambda(x)$ is a Bayes Estimator.* ♡

⚗ **Note** *Improper prior*

Corollary 19.1

Suppose the assumptions of Theorem 19.1 hold.

1. If $L(\theta, d) = [d - g(\theta)]^2$, then

$$\delta_\Lambda(x) = E[g(\Theta) \mid x]. \tag{19.4}$$

2. If $L(\theta, d) = w(\theta)[d - g(\theta)]^2$, then

$$\delta_\Lambda(x) = \frac{E[w(\theta)g(\Theta) \mid x]}{E[w(\theta) \mid x]}. \tag{19.5}$$

3. If $L(\theta, d) = |d - g(\theta)|$, then $\delta_\Lambda(x)$ is any median of the conditional distribution of Θ given x.

4. If

$$L(\theta, d) = \begin{cases} 0 \text{ when } |d - \theta| \leq c \\ 1 \text{ when } |d - \theta| > c \end{cases},$$

then $\delta_\Lambda(x)$ is the midpoint of the interval I of length $2c$ which maxmizes $P(\Theta \in I \mid x)$.

♡

Proof

Theorem 19.2

Necessiary condition for Bayes Estimator

♡

Methodologies have been developed to deal with the difficulty which sometimes incorporate frequentist measures to assess the choic of Λ.

- Empirical Bayes.
- Hierarchical Bayes.
- Robust Bayes.
- Objective Bayes.

19.1 Single-Prior Bayes

The Single-Prior Bayes model in a general form as

$$\begin{aligned} X \mid \theta &\sim f(x \mid \theta), \\ \Theta \mid \gamma &\sim \pi(\theta \mid \lambda), \end{aligned} \tag{19.6}$$

where we assume that the functional form of the prior and the value of λ is known (we will write it as $\gamma = \gamma_0$).

Given a loss function $L(\theta, d)$, we would then determine the estimator that minimizes

$$\int L(\theta, d(x)) \pi(\theta \mid x) \, d\theta, \tag{19.7}$$

where $\pi\,(\theta\mid x)$ is posterior distribution given by

$$\pi\,(\theta\mid x) = \frac{f\,(x\mid\theta)\,\pi\,(\theta\mid\gamma_0)}{\int f\,(x\mid\theta)\,\pi\,(\theta\mid\gamma_0)\,\mathrm{d}\theta}.$$

In general, this Bayes estimator under squared error loss is given by

$$E\,(\Theta\mid x) = \frac{\int \theta f\,(x\mid\theta)\,\pi\,(\theta\mid\gamma_0)\,\mathrm{d}\theta}{\int f\,(x\mid\theta)\,\pi\,(\theta\mid\gamma_0)\,\mathrm{d}\theta}. \tag{19.8}$$

Example 19.1 Consider

$$X_i \overset{\text{i.i.d}}{\sim} N(\mu, \Gamma^{-1}), \quad i = 1, 2, \ldots, n$$

$$\mu \sim N(0, 1),$$

$$\Gamma \sim \text{Gamma}(2, 1),$$

calculate the Single-Prior Bayes estimator under squared error loss.

Solution

$$p\,(X\mid\mu,\Gamma) = \Gamma^n (2\pi)^{-\frac{n}{2}} \exp\left[-2\Gamma^2 \sum_{i=1}^{n} (x_i - \mu)^2\right],$$

$$p(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2}\right),$$

$$p(\Gamma) = \frac{1}{\Gamma(2)} \Gamma \exp\left(-\Gamma\right).$$

Therefore,

$$h\,(X, \mu, \Gamma) = C\,\Gamma^n \exp\left[-2\Gamma^2 \sum_{i=1}^{n} (x_i - \mu)^2\right] \exp\left(-\frac{\mu^2}{2}\right) \Gamma \exp\left(-\Gamma\right),$$

where $C = \frac{(2\pi)^{-\frac{n+1}{2}}}{\Gamma(2)}$.

For μ, we have

$$\pi\,(\mu\mid X, \Gamma) = \frac{h\,(X, \mu, \Gamma)}{p(\mu\mid X)}$$

For exponential families

<div style="border:1px solid orange; border-radius:8px;">

Theorem 19.3

♡
</div>

19.2 Hierarchical Bayes

In a Hierarchical Bayes model, rather than specifying the prior distribution as a single function, we specify it in a **hierarchy**. Thus, the Hierarchical Bayes model in a general form

as

$$X \mid \theta \sim f\left(x \mid \theta\right),$$
$$\Theta \mid \gamma \sim \pi\left(\theta \mid \lambda\right), \qquad (19.9)$$
$$\Gamma \sim \psi\left(\gamma\right),$$

where we assume that $\psi\left(\cdot\right)$ is known and not dependent on any other unknown hyperparameters.

Note *We can continue this hierarchical modeling and add more stages to the model, but this is not ofthen done in practice.*

Given a loss function $L\left(\theta, d\right)$, we would then determine the estimator that minimizes

$$\int L\left(\theta, d\left(x\right)\right) \pi\left(\theta \mid x\right) \mathrm{d}\theta, \qquad (19.10)$$

where $\pi\left(\theta \mid x\right)$ is posterior distribution given by

$$\pi\left(\theta \mid x\right) = \frac{\int f\left(x \mid \theta\right) \pi\left(\theta \mid \gamma\right) \psi\left(\gamma\right) \mathrm{d}\gamma}{\int \int f\left(x \mid \theta\right) \pi\left(\theta \mid \gamma\right) \psi\left(\gamma\right) \mathrm{d}\theta \mathrm{d}\gamma}.$$

Note *The posterior distribution can also be writed as*

$$\pi\left(\theta \mid x\right) = \int \pi\left(\theta \mid x, \gamma\right) \pi\left(\gamma \mid x\right) \mathrm{d}\gamma,$$

where $\pi\left(\gamma \mid x\right)$ is the posterior distribution of Γ, unconditional on θ. The equation 19.10 can be writed as

$$\int L\left(\theta, d\left(x\right)\right) \pi\left(\theta \mid x\right) \mathrm{d}\theta = \int \left[\int L\left(\theta, d\left(x\right)\right) \pi\left(\theta \mid x, \gamma\right) \mathrm{d}\theta\right] \pi\left(\gamma \mid x\right) \mathrm{d}\gamma.$$

*which shows that **the Hierarchical Bayes estimator can be thought of as a mixture of Single-Prior estimators**.*

Example 19.2 Poisson Hierarchy Consider

$$X_i \mid \lambda \overset{\text{i.i.d}}{\sim} \mathrm{Poisson}\left(\lambda\right), \quad i = 1, 2 \ldots, n$$
$$\lambda \mid b \sim \mathrm{Gamma}\left(a, b\right), \ a \text{ known}, \qquad (19.11)$$
$$\frac{1}{b} \sim \mathrm{Gamma}\left(k, \tau\right),$$

calculate the Hierarchical Bayes estimator under squared error loss.

> **Theorem 19.4**
>
> *For the Hierarchical Bayes model (19.9),*
>
> $$K\left[\pi\left(\lambda \mid x\right), \psi\left(\lambda\right)\right] < K\left[\pi\left(\theta \mid x\right), \pi\left(\theta\right)\right], \qquad (19.12)$$
>
> *where K is the Kullback-Leibler information for discrimination between two densities.*

Proof

Note

19.3 Empirical Bayes

19.4 Bayes Prediction

Chapter 20 Hypothesis Testing

Part VII

Convex Optimization

Chapter 21 Convex Sets

21.1 Affine and Convex Sets

21.1.1 Affine Sets

> **Definition 21.1 (Affine Set)**
>
> *A nonempty set C is said to be **affine set**, if*
> $$\forall x_1, x_2 \in C, \theta \in \mathbf{R}, \theta x_1 + (1 - \theta)x_2 \in C.$$

21.1.2 Convex Sets

> **Definition 21.2 (Convex Set)**
>
> *A nonempty set C is said to be **convex set**, if*
> $$\forall x_1, x_2 \in C, \theta \in [0, 1], \theta x_1 + (1 - \theta)x_2 \in C.$$

> **Definition 21.3 (Convex Hull)**
>
> *The **convex hull** of said to be set C, denoted by conv C is a set of all convex combinations of points in C,*
> $$conv\ C = \{\theta_1 x_1 + \ldots + \theta_k x_k | x_i \in C; \theta_i \geq 0, i = 1, \ldots, k; \theta_1 + \ldots + \theta_k = 1\}.$$

Note *The convex hull conv C is always convex, which is the minimal convex set that contains C.*

21.1.3 Cones

> **Definition 21.4 (Cone)**
>
> *A nonempty set C is said to be **cone**, if*
> $$\forall x \in C, \theta \geq 0, \theta x \in C.$$

> **Definition 21.5 (Convex Cone)**
>
> *A nonempty set C is said to be **convex cone**, if*
> $$\forall x_1, x_2 \in C, \theta_1, \theta_2 \geq 0, \theta_1 x_1 + \theta_2 x_2 \in C.$$

21.2 Some Important Examples

> **Definition 21.6 (Hyperplane)**
>
> *A hyperplane is defined to be*
> $$\{x \,|\, a^T x = b\},$$
> *where $a \in \mathbf{R}^n, a \neq 0, b \in \mathbf{R}$.* ♣

> **Definition 21.7 (Halfspace)**
>
> *A hyperplane is defined to be*
> $$\{x \,|\, a^T x \leq b\},$$
> *where $a \in \mathbf{R}^n, a \neq 0, b \in \mathbf{R}$.* ♣

> **Definition 21.8 ((Euclidean) Ball)**
>
> *A (Euclidean) ball in $\mathbf{R}^n$ with center x_c and radius r is defined to be*
> $$B(x_c, r) = \{x \,|\, \|x - x_c\|_2 \leq r\} = \{x_c + ru \,|\, \|u\|_2 \leq 1\},$$
> *where $r > 0$.* ♣

> **Definition 21.9 (Ellipsoid)**
>
> *A Ellipsoid in $\mathbf{R}^n$ with center x_c is defined to be*
> $$\mathcal{E} = \{x \,|\, (x - x_c)^T P^{-1} (x - x_c) \leq 1\} = \{x_c + Au \,|\, \|u_2\| \leq 1\},$$
> *where $P \in \mathbf{S}^n_{++}$ (symmetric positive definite).* ♣

21.3 Generalized Inequalities

21.3.1 Definition of Generalized Inequalities

> **Definition 21.10 (Proper Cone)**
>
> *A cone $K \subseteq \mathbf{R}^n$ is said to be proper cone, if*
> - *K is convex.*
> - *K is closed.*
> - *K is solid (nonempty interior).*
> - *K is pointed (contains no line).* ♣

Definition 21.11 (Generalized Inequalities)

The partial ordering on $\mathbf{R}^n$ defined by proper cone K, if

$$y - x \in K, \tag{21.1}$$

which can be denoted by

$$x \preceq_K y \ or \ y \succeq_K x. \tag{21.2}$$

The strict partial ordering on $\mathbf{R}^n$ defined by proper cone K, if

$$y - x \in \ int \ K, \tag{21.3}$$

which can be denoted by

$$x \prec_K y \ or \ y \succ_K x. \tag{21.4}$$

Note *When $K = \mathbf{R}_+$, the partial ordering $\preceq_K$ is the usual ordering $\leq$ on $\mathbf{R}$, and the strict partial ordering $\prec_K$ is the usual strict ordering $<$ on $\mathbf{R}$.*

21.3.2 Properties of Generalized Inequalities

Theorem 21.1 (Properties of Generalized Inequalities)

A generalized inequality $\preceq_K$ has the following properties:

- *Preserved under addition:*
- *Transitive:*
- *Preserved under nonnegative scaling:*
- *Reflexive:*
- *Antisymmetric:*
- *Preserved under limits:*

A strict generalized inequality $\prec_K$ has the following properties:

Chapter 22 Convex Optimization Problems

22.1 Generalized Inequality Constraints

> **Definition 22.1 (With Generalized Inequality Constraints)**
>
> *A convex optimization problem with generalized inequality constraints is defined to be*
>
> $$\min_x \quad f_0(x)$$
> $$s.t. \quad f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, m \tag{22.1}$$
> $$Ax = b$$
>
> *where $f_0 : \mathbf{R}^n \to \mathbf{R}$, $K_i \in \mathbf{R}^{k_i}$ are proper conves, and $f_i : \mathbf{R}^n \to \mathbf{R}^{k_i}$ are K_i-convex.* ♣

22.1.1 Conic Form Problems

> **Definition 22.2 (Conic Form Problem)**
>
> *A conic form problem is defined to be*
>
> $$\min \quad c^T x$$
> $$s.t. \quad Fx + g \preceq_K 0 \tag{22.2}$$
> $$Ax = b$$
>
> ♣

22.1.2 Semidefinite Programming

22.2 Vector Optimization

Chapter 23 Unconstrained Minimization

23.1 Definition of Unconstrained Minimization

> **Definition 23.1 (Unconstrained Minimization Problem)**
> *The unconstrained minimization problem is defined to be*
> $$\min_{x} f(x) \qquad\qquad (23.1)$$
> *where $f : \mathbf{R}^n \to \mathbf{R}$ is convex and twice continuously differentiable.* ♣

Note *We assume that the problem is solvable, i.e., there exists an optimal point x^*, such that, $f(x^*) = \inf_x f(x)$.*

Example 23.1 Quadratic Minimization

Example 23.2 Least Square Estimation

Example 23.3 Unconstrained Geometric Programming

Example 23.4 Analytic Center of Linear Inequalities

23.2 General Descent Method

23.3 Gradient Descent Method

23.4 Steepest Descent Method

23.5 Newton's Method

Chapter 24 Exercises for Convex Optimization

24.1 Convex Sets

✎ **Exercise 24.1** Solution set of a quadratic inequality Let $C \subseteq \mathbf{R}^n$ be the solution set of a quadratic inequality,

$$C = \{x \in \mathbf{R}^n | x^T A x + b^T x + c \leq 0\}$$

with $A \in \mathbf{S}^n$, $b \in \mathbf{R}^n$, and $c \in \mathbf{R}$.

1. Show that C is convex if $A \succeq 0$.

Solution

1. *We have to show that $\theta x + (1-\theta)y \in C$ for all $\theta \in [0,1]$ and $x, y \in C$.*

$$(\theta x + (1-\theta)y)^T A(\theta x + (1-\theta)y) + b^T(\theta x + (1-\theta)y) + c$$

$$= \theta^2 x^T A x + \theta(1-\theta)(y^T A x + x^T A y) + (1-\theta)^2 y^T A y + \theta b^T x + (1-\theta)b^T y + c$$

$$= \theta^2(x^T A x + b^T x + c) + (1-\theta)^2(y^T A y + b^T y + c) - \theta^2(b^T x + c)$$

$$\quad - (1-\theta)^2(b^T y + c) + \theta(1-\theta)(y^T A x + x^T A y) + \theta b^T x + (1-\theta)b^T y + c$$

$$\leq -\theta^2(b^T x + c) - (1-\theta)^2(b^T y + c) + \theta(1-\theta)(y^T A x + x^T A y)$$

$$\quad + \theta b^T x + (1-\theta)b^T y + c$$

$$= \theta(1-\theta)[(b^T x + c) + (b^T y + c) + x^T A x + y^T A y]$$

$$\leq \theta(1-\theta)(-x^T A x - y^T A y + x^T A x + y^T A y) \leq 0$$

Therefore, $\theta x + (1-\theta)y \in C$, which shows that C is convex if $A \succeq 0$.

Part VIII

Generalized Linear Model

Chapter 25 Generalized Linear Model

25.1 Exponential Family

> **Definition 25.1 (Exponential Family)**
>
> An exponential family of probability distributions as those distributions whose density is defined to be
> $$f(y \mid \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{25.1}$$
> ♣

Property *The exponential family have the following properties,*
$$E(Y) = b'(\theta) \quad \mathrm{Var}(Y) = b''(\theta)a(\phi).$$

Proof

Table 25.1: Common Distributions of Exponential Family

Distribution	Parameter(s)	θ	ϕ	$b(\theta)$	$a(\phi)$	$c(y, \phi)$	$E(Y)$	$\mathrm{Var}(Y)$
Normal	$N(\mu, \sigma^2)$	μ	σ^2	$\frac{\theta^2}{2}$	ϕ	$-\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]$	θ	ϕ
Bernoulli	$\mathrm{Bern}(p)$	$\log\left(\frac{p}{1-p}\right)$	1	$\log(1 + e^\theta)$	1	0	$\frac{e^\theta}{1+e^\theta}$	$\frac{e^\theta}{(1+e^\theta)^2}$
Poisson	$P(\mu)$	$\log(\mu)$	1	e^θ	1	$-\log(y!)$	e^θ	e^θ

25.2 Model Assumption

Suppose the response Y has a distribution in the exponential family
$$f(y \mid \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$
with link function g, such that,
$$E(Y \mid \mathbf{X}) = \mu = g^{-1}(\eta), \quad \eta = \mathbf{X}'\boldsymbol{\beta} \tag{25.2}$$
where the link function provides the relationship between the linear predictor and the mean of the distribution function. If $\eta = \theta$, the link function is called **canonical link function**.

Note *A generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for the response variable to have an error distribution other than the normal distribution.*

Table 25.2: Commonly Used Link Functions

Distribution	Support of Distribution	Link Function $g(\mu)$	Mean Function $g^{-1}(\eta)$
Normal	real:$(-\infty, +\infty)$	μ	η
Bernoulli	integer: $\{0, 1\}$	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{1}{1+\exp(-\eta)}$
Poisson	integer: $0, 1, 2, \ldots$	$\log(\mu)$	$\exp(\eta)$

25.3 Model Estimation

25.3.1 Maximum Likelihood

Suppose the log-likelihood function be

$$\ell\left(\boldsymbol{\beta} \mid \mathbf{X}, y\right) = \log\left[f\left(y \mid \theta, \phi\right)\right] = \log\left[f\left(y \mid g^{-1}(\eta), \phi\right)\right] \tag{25.3}$$

where g is the canonical link function and $\eta = \mathbf{X}'\boldsymbol{\beta}$.

Let

$$U\left(\boldsymbol{\beta}\right) = \frac{\partial\ell\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}, \quad A\left(\boldsymbol{\beta}\right) = -\frac{\partial^2\ell\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}'\partial\boldsymbol{\beta}}$$

be the score function and observed information matrix.

If $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate, then

$$U\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{0}$$

By mean value theorem,

$$U\left(\hat{\boldsymbol{\beta}}\right) - U\left(\boldsymbol{\beta}_0\right) = \frac{\partial U\left(\boldsymbol{\beta}^*\right)}{\partial\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$$

$$\Rightarrow -U\left(\boldsymbol{\beta}_0\right) = -A\left(\boldsymbol{\beta}^*\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$$

where $\boldsymbol{\beta}^* \in \left[\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}\right]$. Thus,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + A^{-1}\left(\boldsymbol{\beta}^*\right)U\left(\boldsymbol{\beta}_0\right)$$

Suppose $\hat{\boldsymbol{\beta}}_t, \hat{\boldsymbol{\beta}}_{t+1}$ be the maximum likelihood estimate at the t-th and (t+1)-th iterations, respectively. Two algorithms can be used to obtain the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

1. Newton-Raphson Method:

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + A^{-1}\left(\hat{\boldsymbol{\beta}}_t\right)U\left(\hat{\boldsymbol{\beta}}_t\right) \Leftrightarrow A\left(\hat{\boldsymbol{\beta}}_t\right)\hat{\boldsymbol{\beta}}_{t+1} = A\left(\hat{\boldsymbol{\beta}}_t\right)\hat{\boldsymbol{\beta}}_t + U\left(\hat{\boldsymbol{\beta}}_t\right) \tag{25.4}$$

where

$$U\left(\boldsymbol{\beta}\right) = \frac{\partial\ell\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} \tag{25.5}$$

is the score function and

$$A(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}} \tag{25.6}$$

is the observed information matrix.

2. Fisher's Scoring Method:

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + I^{-1}\left(\hat{\boldsymbol{\beta}}_t\right) U\left(\hat{\boldsymbol{\beta}}_t\right) \Leftrightarrow I\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_{t+1} = I\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_t + U\left(\hat{\boldsymbol{\beta}}_t\right) \tag{25.7}$$

where $U(\boldsymbol{\beta})$ is the score function and

$$I(\boldsymbol{\beta}) = E\left[A(\boldsymbol{\beta})\right] = -E\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}}\right] \tag{25.8}$$

is the Fisher information matrix.

25.3.2 Bayesian Methods

Chapter 26 Binary Data

26.1 Model Assumption

Suppose

$$Y \sim b\left(m, \pi\right), \quad i = 1, 2, \ldots, n \tag{26.1}$$

with link function

$$\eta = g\left(\pi\right) = \log\left(\frac{\pi}{1 - \pi}\right) = \mathbf{x}'\boldsymbol{\beta} \tag{26.2}$$

Remark

26.2 Model Estimation

The likelihood function is

$$f\left(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y}\right) = \prod_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} \left(1 - \pi_i\right)^{m_i - y_i} \tag{26.3}$$

and the log-likelihood function is

$$
\begin{aligned}
\ell\left(\boldsymbol{\beta}\right) = \log\left[f\left(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y}\right)\right] &= \sum_{i=1}^{n} \ell_i\left(\boldsymbol{\beta}\right) \\
&= \sum_{i=1}^{n} \left\{ \log\left[\binom{m_i}{y_i}\right] + y_i \log\left(\pi_i\right) + \left(m_i - y_i\right) \log\left(1 - \pi_i\right) \right\} \\
&= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + m_i \log\left(1 - \pi_i\right) \right] + \sum_{i=1}^{n} \log\left[\binom{m_i}{y_i}\right]
\end{aligned} \tag{26.4}
$$

where

$$\pi_i = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)} \tag{26.5}$$

Thus,

$$U_r\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \left(y_i - m_i \pi_i\right) x_{ir}$$

$$I_{sr}\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} m_i \pi_i \left(1 - \pi_i\right) x_{is} x_{ir}$$

Chapter 27 Polytomous Data

> **Definition 27.1 (Polytomous Data)**
>
> *A response is polytomous, if the response of an individual or item in a study is **restricted to one of a fixed set of possible values**.* ♣

Remark There are two types of scales, pure scales and compound scales [1]. For pure scales, there are several types:

1. **Nominal Scale**: a scale used for labeling variables into distinct classifications and does not involve a quantitative value or order.
2. **Ordinal Scale**: a variable measurement scale used to simply depict the order of variables and not the difference between each of the variables.
3. **Interval Scale**: a numerical scale where the order of the variables is known as well as the difference between these variables.

27.1 Model Assumption

Let the category probabilities given $\mathbf{x}_i$ be

$$\pi_j(\mathbf{x}_i) = P\left(Y = y_j \mid \mathbf{X} = \mathbf{x}_i\right) \tag{27.1}$$

and the cumulative probabilities given $\mathbf{x}_i$ be

$$r_j(\mathbf{x}_i) = P\left(Y \leq \sum_{r \leq j} y_r \mid \mathbf{X} = \mathbf{x}_i\right) \tag{27.2}$$

where $i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, k$.

Here,s multinomial distribution is in many ways the most natural distribution to consider in the context of a polytomous response variable. The denisty function of the multinomial distribution is,

$$P\left(Y_1 = y_1, \ldots, Y_k = y_k\right) = \begin{cases} \frac{m!}{y_1! \cdots y_k!} \pi_1^{y_1} \cdot \ldots \cdot \pi_k^{y_k}, & \sum_{i=1}^{k} y_i = m \\ 0 & \text{otherwise} \end{cases}$$

for non-negative integers $y_1, \ldots, y_k$.

As for the link function, we have

---

[1] A bivariate responses with one response ordinal and the other continuous is an example of compound scales.

Nominal Scale

$$\pi_j(\mathbf{x}_i) = \frac{\exp\left[\eta_j(\mathbf{x}_i)\right]}{\sum_{j=1}^{k} \exp\left[\eta_j(\mathbf{x}_i)\right]} \tag{27.3}$$

where $\eta_j(\mathbf{x}_i) = \eta_j(\mathbf{x}_0) + (\mathbf{x}_i - \mathbf{x}_0)' \boldsymbol{\beta}_j + \alpha_i$.

Ordinal Scale

1. Logistic Scale:

$$\log\left[\frac{r_j(\mathbf{x}_i)}{1 - r_j(\mathbf{x}_i)}\right] = \theta_j - \mathbf{x}_i' \boldsymbol{\beta} \tag{27.4}$$

2. Complementary Log-Log Scale:

$$\log\left\{-\log\left[1 - r_j(\mathbf{x}_i)\right]\right\} = \theta_j - \mathbf{x}_i' \boldsymbol{\beta} \tag{27.5}$$

Interval Scale Suppose the j-th category exits a cardinal number or score, s_j, where the difference between scores is a measure of distance between or separation of categories.

1.
$$\log\left[\frac{r_j(\mathbf{x}_i)}{1 - r_j(\mathbf{x}_i)}\right] = \varsigma_0 + \varsigma_1\left(\frac{s_j + s_{j+1}}{2}\right) - \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{x}_i' \boldsymbol{\xi}\left(c_j - \bar{c}\right) \tag{27.6}$$

   where $c_j = \frac{s_j + s_{j+1}}{2}$ or $c_j = \text{logit}\left(\frac{s_j + s_{j+1}}{2}\right)$.

2.
$$\pi_j(\mathbf{x}_i) = \frac{\exp\left[\eta_j(\mathbf{x}_i)\right]}{\sum_{j=1}^{k} \exp\left[\eta_j(\mathbf{x}_i)\right]} \tag{27.7}$$

   where $\eta_j(\mathbf{x}_i) = \eta_j + \left(\mathbf{x}_i' \boldsymbol{\beta}\right) s_j + \alpha_i$.

3.
$$\sum_{j=1}^{k} \pi_j(\mathbf{x}_i) s_j = \mathbf{x}_i \boldsymbol{\beta} \tag{27.8}$$

27.2 Model Estimation

Chapter 28 Count Data

28.1 Model Assumption

Departures from the idealized Poisson model are to be expected. Therefore, we avoid the assumption of Poisson variation and assume only that

$$\text{Var}\,(Y) = \sigma^2 E\,(Y) \tag{28.1}$$

with link function

$$\log\,(\mu) = \eta = \mathbf{x}'\boldsymbol{\beta} \tag{28.2}$$

where $\mu = E\,(Y \mid \mathbf{X})$.

28.2 Model Estimation

For the response in the Poisson distribution, i.e.

$$P(Y = y \mid \mu) = \frac{e^{-\mu}\mu^y}{y!}$$

and the log-likelihood function is

$$\ell\,(\boldsymbol{\beta}) \propto \sum_{i=1}^{n} (y_i \log\,(\mu_i) - \mu_i) \tag{28.3}$$

where $\mu_i = E\,(Y \mid \mathbf{X} = \mathbf{x}_i)$.

Chapter 29 Survival Data

29.1 Survival Data

Definition 29.1 (Survival Function)

The survival function[a] is defined to be

$$S(t) = P(T > t) = \int_t^\infty f(u)\,\mathrm{d}u = 1 - F(t). \tag{29.1}$$

where t is some specified time, T is a random variable denoting the time of death.

---

[a]The survival function is the probability that the time of death is later than some specified time t.

Definition 29.2 (Lifetime Distribution Function)

The lifetime distribution function is defined to be

$$F(t) = P(T \le t) \tag{29.2}$$

If F is differentiable then the derivative, which is the density function of the lifetime distribution[a], is defined to be

$$f(t) = F'(t) = \frac{\mathrm{d}}{\mathrm{d}t} F(t) \tag{29.3}$$

---

[a]The function f is sometimes called the event density; it is the rate of death or failure events per unit time.

Definition 29.3 (Hazard Function)

The Hazard function[a] is defined to be

$$\lambda(t) = \lim_{\varepsilon \to 0^+} \left[\frac{P(t \le T < t + \varepsilon \mid T \ge t)}{\varepsilon} \right] = \frac{f(t)}{S(t)} \tag{29.4}$$

---

[a]The Hazard function is the event rate at time t conditional on survival until time t or later (that is, $T \ge t$).

Property *The relationship among $\lambda(t)$, $f(t)$, $S(t)$,*

  1.

$$\lambda(t) = -\frac{\mathrm{d}\log[S(t)]}{\mathrm{d}t} \tag{29.5}$$

  2.

$$S(t) = \exp\left[-\int_0^t \lambda(x)\,\mathrm{d}x \right] \tag{29.6}$$

3.

$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(x)\,\mathrm{d}x\right] \tag{29.7}$$

Proof

Example 29.1 Constant Hazards Suppose

$$\lambda(t) = \lambda \tag{29.8}$$

then

$$S(t) = \exp\left[-\int_0^t \lambda(x)\,\mathrm{d}x\right] = \exp\left[-\int_0^t \lambda\,\mathrm{d}x\right] = \exp(-\lambda t)$$

$$f(t) = \lambda(t)\exp\left[-\int_0^t \lambda(x)\mathrm{d}x\right] = \lambda\exp\left[-\int_0^t \lambda\mathrm{d}x\right] = \lambda\exp(-\lambda t)$$

which is the exponential distribution.

Example 29.2 Bathtub Hazards

$$\lambda(t) = \alpha t + \frac{\beta}{1 + \gamma t} \tag{29.9}$$

29.2 Estimation of Survival Function

Parametric Approach Suppose $t_1, t_2, \ldots, t_n$ are failure times corresponding to censor indicators $\delta_1, \delta_2, \ldots, \delta_n$. The likelihood function is

$$
\begin{aligned}
f(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}) &= \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\frac{f(t_i)}{S(t_i)}\right)^{\delta_i} S(t_i) \\
&= \prod_{i=1}^{n} [\lambda(t_i)]^{\delta_i} S(t_i)
\end{aligned} \tag{29.10}
$$

where $\lambda(t), S(t)$ depends on some parameter θ.

Example 29.3 Suppose T have exponential density, that,

$$f(t) = \lambda\mathrm{e}^{-\lambda t}, \quad S(t) = \mathrm{e}^{-\lambda t}$$

Thus,

$$
\begin{aligned}
\ell(\lambda) = \log[\ell(\theta)] &= \sum_{i=1}^{n} [\delta_i \log(\lambda) - \lambda t_i] \\
&= \left(\sum_{i=1}^{n} \delta_i\right) \log(\lambda) - \lambda \left(\sum_{i=1}^{n} t_i\right)
\end{aligned}
$$

Hence,

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} t_i = 0 \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}$$

Nonparametric Approach   Then, for $t_{(k)} \le t < t_{(k+1)}$,

$$\begin{aligned}
\hat{S}(t) &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \\
&= \left(1 - \frac{d_1}{n_1} \right) \left(1 - \frac{d_2}{n_2} \right) \cdots \left(1 - \frac{d_k}{n_k} \right) \\
&\approx \left[1 - \hat{\lambda}(t_1) \right] \left[1 - \hat{\lambda}(t_2) \right] \cdots \left[1 - \hat{\lambda}(t_k) \right]
\end{aligned} \tag{29.11}$$

where $\hat{S}(t)$ is referred to as Kaplan-Meier estimate.

29.3 Proportional Hazards Model

Suppose $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ be death times. The number of individuals who alive just before time $t_{(j)}$, including those who are about to die at this time, will be denoted n_j, for $j = 1, 2, \ldots, m$, and d_j will denote the number who die at this time. Thus, we have

29.3.1 Model Assumption

Let $t_1, t_2, \ldots, t_n$ be the failure times associated with censor indicator $\delta_1, \delta_2, \ldots, \delta_n$ and the covariate vectors $\mathbf{x}_i$.

Further, let $t_{(1)} \le t_{(2)} \le \ldots \le t_{(m)}$ be the ordered uncensored failure times corresponding to $\delta_{(j)} = 1$, $j = 1, 2, \ldots, m$, and $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$ are the associated covariate vectors. Note (j) represents the label for the individual who dies at $t_{(j)}$.

The proportional hazards model specifying the hazard at time t for an individual whose covariate vector is $\mathbf{x}$ is given by

$$\lambda(t) = \lambda_0(t) e^{\mathbf{x}' \boldsymbol{\beta}} \tag{29.12}$$

where $\lambda_0(t)$ is referred to as the baseline hazard function.

29.3.2 Model Estimation

The exact likelihood function is

$$\ell[\beta, \lambda_0(t)] = \prod_{i=1}^{n} [\lambda_i(t_i)]^{\delta_i} S(t_i) \tag{29.13}$$

depends on both the nonparametric function $\lambda_0(t)$ and the parameter $\boldsymbol{\beta}$. Thus, it might be difficult to estimate $\lambda_0(t)$ and $\boldsymbol{\beta}$ simultaneously.

The partial likelihood function is

$$\ell_p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{e^{\mathbf{x}'_{(j)}\boldsymbol{\beta}}}{\sum_{l \in R(t_{(j)})} e^{\mathbf{x}'_l \boldsymbol{\beta}}} = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{\sum_{l \in R(t_i)} e^{\mathbf{x}'_l \boldsymbol{\beta}}} \right]^{\delta_i} \tag{29.14}$$

where $R(t)$ is the set of individuals who are alive and uncensored at a time just prior to t_i, which is called the rik set.

Chapter 30 Modified Likelihood

Seek a modified likelihood function that depends on as few of the nuisance parameters as possible while sacrificing as little information as possible.

30.1 Marginal Likelihood

30.2 Conditional Likelihood

Let $\boldsymbol{\theta} = (\boldsymbol{\varphi}, \boldsymbol{\lambda})$, where $\boldsymbol{\varphi}$ is the parameter vector of interest and $\boldsymbol{\lambda}$ is a vector of nuisance parameters. The conditional likelihood can be obtained as follows:

1. Find the complete sufficient statistic $S_{\boldsymbol{\lambda}}$, respectively for $\boldsymbol{\lambda}$.
2. Construct the conditional log-likelihood

$$\ell_c = \log\left(f_{Y|S_{\boldsymbol{\lambda}}}\right) \tag{30.1}$$

where $f_{Y|S_{\boldsymbol{\lambda}}}$ is the conditional distribution of the response Y given $S_{\boldsymbol{\lambda}}$.

Note *Two cases might occur, that, for fixed $\boldsymbol{\varphi}_0$, $S_{\boldsymbol{\lambda}}(\boldsymbol{\varphi}_0)$ depends on $\boldsymbol{\varphi}_0$; or $S_{\boldsymbol{\lambda}}(\boldsymbol{\varphi}_0) = S_{\boldsymbol{\lambda}}$ is independent of $\boldsymbol{\varphi}_0$.*

1. *Independent:*
2. *Dependent:*

Example 30.1

Conditional Likelihood for Exponential Family Suppose that the log-likelihood for $\boldsymbol{\theta} = (\boldsymbol{\varphi}, \boldsymbol{\lambda})$ can be written in the exponential family form

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\theta}'\mathbf{s} - b(\boldsymbol{\theta}) \tag{30.2}$$

Also, suppose $\ell(\boldsymbol{\theta}, \mathbf{y})$ has a decomposition of the form

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\varphi}'\mathbf{s}_1 + \boldsymbol{\lambda}'\mathbf{s}_2 - b(\boldsymbol{\varphi}, \boldsymbol{\lambda}) \tag{30.3}$$

Remark The above decomposition can be achieved only if $\boldsymbol{\varphi}$ is a linear function of θ. The choice of nuisance parameter λ is arbitrary and the inferences regarding $\boldsymbol{\varphi}$ should be unaffected by the parameterization chosen for λ.

The conditional likelihood of the data $\mathbf{Y}$ given $\mathbf{s}_2$ is

$$\ell(\boldsymbol{\varphi} \mid \mathbf{s}_2) = \boldsymbol{\varphi}'\mathbf{s}_1 - b^*(\boldsymbol{\varphi}, \boldsymbol{\lambda}) \tag{30.4}$$

which is independent of the nuisance parameter and may be used for inferences regarding $\boldsymbol{\varphi}$.

Example 30.2 $Y_1 \sim P(\mu_1), Y_2 \sim P(\mu_2)$ are independent. Suppose $\varphi = \log\left(\frac{\mu_2}{\mu_1}\right) = \log(\mu_2) - \log(\mu_1)$ is the parameter of interest and the nuisance parameter is

1. $\lambda_1 = \log(\mu_1)$.

2.

Then, give the conditional log-likelihood for different nuisance parameter.

Solution

1. *The log-likelihood function in the form of (φ, λ) is*

$$\ell(\phi, \lambda_1) \propto \log\left[e^{-(\mu_1+\mu_2)}\mu_1^{y_1}\mu_2^{y_2}\right]$$

$$= -(\mu_1 + \mu_2) + y_1 \log(\mu_1) + y_2 \log(\mu_2)$$

$$= -\mu_1\left(1 + \frac{\mu_2}{\mu_1}\right) + y_1 \log(\mu_1) + y_2 \log(\mu_1)$$

$$\quad - y_2\left[\log(\mu_1) - \log(\mu_2)\right]$$

$$= -e^{\lambda_1}(1 + e^{\varphi}) + (y_1 + y_2)\lambda_1 - y_2\varphi$$

$$= s_1\varphi + s_2\lambda_1 - b(\varphi, \lambda_1)$$

where $s_1 = -y_2, s_2 = y_1 + y_2, b(\varphi, \lambda_1) = e^{\lambda_1}(1 + e^{\varphi})$.

Then, the conditional distribution of Y_1, Y_2 given $S_2 = Y_1 + Y_2$ is $b\left(S_2, \frac{\mu_1}{\mu_1+\mu_2}\right)$, thus,

$$\ell(\varphi \mid S_2 = s_2) \propto y_1 \log\left(\frac{\mu_1}{\mu_1 + \mu_2}\right) + y_2 \log\left(\frac{\mu_2}{\mu_1 + \mu_2}\right)$$

$$= y_1 \log\left(\frac{\mu_1}{\mu_1 + \mu_2}\right) + y_2 \log\left(\frac{\mu_1}{\mu_1 + \mu_2}\right)$$

$$\quad - y_2\left[\log\left(\frac{\mu_1}{\mu_1 + \mu_2}\right) - \log\left(\frac{\mu_2}{\mu_1 + \mu_2}\right)\right]$$

$$= (y_1 + y_2)\log\left(\frac{1}{1 + e^{\varphi}}\right) - y_2\varphi$$

$$= s_1\varphi - b^*(\varphi, s_2)$$

where $b^*(\varphi, s_2) = -s_2 \log\left(\frac{1}{1+\varphi^{-1}}\right)$.

30.3 Profile Likelihood

30.4 Quasi Likelihood

Part IX

Machine Learning

Chapter 31 Kernel Methods

Definition 31.1 (Positive Definite Kernel)

Let $\mathcal{X}$ be a set, a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite kernel on $\mathcal{X}$ iff it is

1. *symmetric, that is,*

$$K\left(\mathbf{x}, \mathbf{x}'\right) = K\left(\mathbf{x}', \mathbf{x}\right), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{31.1}$$

2. *positive definite, that is,*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq 0, \tag{31.2}$$

holds for any $x_1, \dots, x_n \in \mathcal{X}$, given $n \in \mathbb{N}, c_1, \dots, c_n \in \mathbb{R}$. ♣

Theorem 31.1 (Morse-Aronszajn's Theorem)

For any set $\mathcal{X}$, suppose $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite, then there is a unique RKHS $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with reproducing kernel K. ♡

Proof

1. How to build a valid pre-RKHS $\mathcal{H}_0$?

   Consider the vector space $\mathcal{H}_0 \subset \mathbb{R}^{\mathcal{X}}$ spanned by the functions $\{K\left(\cdot, \mathbf{x}\right)\}_{\mathbf{x} \in \mathcal{X}}$. For any $f, g \in \mathcal{H}_0$, suppose

   $$f = \sum_{i=1}^{m} a_i K\left(\cdot, \mathbf{x}_i\right), \quad g = \sum_{j=1}^{n} b_j K\left(\cdot, \mathbf{y}_j\right)$$

   and let the inner product of $\mathcal{H}_0$ be

   $$\langle f, g \rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j K\left(\mathbf{x}_i, \mathbf{y}_j\right) \tag{31.3}$$

   Let $\mathbf{x} \in \mathcal{X}$,

   $$\langle f, K\left(\cdot, \mathbf{x}\right) \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i K\left(\mathbf{x}, \mathbf{x}_i\right) = f(\mathbf{x})$$

   And, we also have

   $$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i g\left(\mathbf{x}_i\right) = \sum_{j=1}^{n} b_j f\left(\mathbf{y}_j\right)$$

Suppose

$$f = \sum_{i=1}^{m} a_i K\left(\cdot, \mathbf{x}_i\right), \quad g = \sum_{j=1}^{n} b_j K\left(\cdot, \mathbf{y}_j\right), \quad h = \sum_{k=1}^{p} c_k K\left(\cdot, \mathbf{z}_k\right)$$

(a). Linearity: For any $\alpha, \beta \in \mathbb{R}$, $\langle \alpha f + \beta g, h \rangle_{\mathcal{H}_0} = \alpha \langle f, h \rangle_{\mathcal{H}_0} + \beta \langle g, h \rangle_{\mathcal{H}_0}$.

$$\langle \alpha f + \beta g, h \rangle_{\mathcal{H}_0} = \left[\alpha \sum_{i=1}^{m} a_i K\left(\cdot, \mathbf{x}_i\right) + \beta \sum_{j=1}^{n} b_j K\left(\cdot, \mathbf{y}_j\right) \right] \cdot \sum_{k=1}^{p} c_k K\left(\cdot, \mathbf{z}_k\right)$$

$$= \alpha \sum_{i=1}^{m} \sum_{k=1}^{p} a_i c_k K\left(\mathbf{x}_i, \mathbf{z}_k\right) + \beta \sum_{j=1}^{n} \sum_{k=1}^{p} b_j c_k K\left(\mathbf{y}_j, \mathbf{z}_k\right)$$

$$= \alpha \langle f, h \rangle_{\mathcal{H}_0} + \beta \langle g, h \rangle_{\mathcal{H}_0}$$

(b). Conjugate Symmetry: $\langle f, g \rangle_{\mathcal{H}_0} = \langle g, f \rangle_{\mathcal{H}_0}$.

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j K\left(\mathbf{x}_i, \mathbf{y}_j\right) = \sum_{j=1}^{n} \sum_{i=1}^{m} b_j a_i K\left(\mathbf{y}_j, \mathbf{x}_i\right)$$

$$= \langle g, f \rangle_{\mathcal{H}_0}$$

(c). Positive Definiteness: $\langle f, f \rangle_{\mathcal{H}_0} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}_0} = 0$ if and only if $f = 0$.

By positive definiteness of K, we have:

$$\langle f, f \rangle_{\mathcal{H}_0} = \| f \|_{\mathcal{H}_0}^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq 0$$

As for, $\langle f, f \rangle_{\mathcal{H}_0} = 0$ if and only if $f = 0$, we have,

"$\Rightarrow$" If $f = 0$, that is $f = \sum_{i=1}^{m} a_i K\left(\cdot, \mathbf{x}_i\right) = 0$, we have

$$\langle f, f \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i f = 0$$

"$\Leftarrow$" For $\forall \mathbf{x} \in \mathcal{X}$, by Cauchy-Schwarwz Inequality, we have,

$$|f(\mathbf{x})| = \left| \langle f, K\left(., \mathbf{x}\right) \rangle_{\mathcal{H}_0} \right| \leq \| f \|_{\mathcal{H}_0} \cdot K\left(\mathbf{x}, \mathbf{x}\right)^{\frac{1}{2}}$$

therefore, if $\| f \|_{\mathcal{H}_0} = 0$, then $f = 0$

Hence, definition in equation 31.3 is a valid inner product, which is a valid pre-RKHS $\mathcal{H}_0$.

Example 31.1 Common Kernels

1. $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right), \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

Proof

1. It is obvious that $K(\mathbf{x}, \mathbf{y})$ is symmetric, we only need to show $K(\mathbf{x}, \mathbf{y})$ is positive definite.

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}\|^2\right) \cdot \exp\left(\frac{1}{\sigma^2}\langle\mathbf{x}, \mathbf{y}\rangle\right) \cdot \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}\|^2\right)$$

By the Taylor expansion of the exponential function, that

$$\exp\left(\frac{x}{\sigma^2}\right) = \sum_{n=0}^{+\infty}\left\{\frac{x^n}{\sigma^{2n} \cdot n!}\right\}$$

Hence,

$$\exp\left(\frac{1}{\sigma^2}\langle\mathbf{x}, \mathbf{y}\rangle\right) = \sum_{n=0}^{+\infty}\left\{\frac{\langle\mathbf{x}, \mathbf{y}\rangle^n}{\sigma^{2n} \cdot n!}\right\}$$

By the Multinomial Theorem, we have

$$\langle\mathbf{x}, \mathbf{y}\rangle^n = \left(\sum_{i=1}^{d} x_i y_i\right)^n = \sum_{k_1+k_2+\ldots+k_d=n}\left[\binom{n}{k_1, k_2, \ldots, k_d}\prod_{i=1}^{d}(x_i y_i)^{k_i}\right]$$

$$= \sum_{k_1+k_2+\ldots+k_d=n}\left[\binom{n}{k_1, k_2, \ldots, k_d}^{\frac{1}{2}}\prod_{i=1}^{d}x_i^{k_i} \cdot \binom{n}{k_1, k_2, \ldots, k_d}^{\frac{1}{2}}\prod_{i=1}^{d}y_i^{k_i}\right]$$

Therefore,

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{\|\mathbf{y}\|^2}{2\sigma^2}\right) \cdot \sum_{n=0}^{+\infty}\left\{\frac{\langle\mathbf{x}, \mathbf{y}\rangle^n}{\sigma^{2n} \cdot n!}\right\}$$

$$= \sum_{n=0}^{+\infty}\frac{\exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right)}{\sigma^n \cdot \sqrt{n!}} \cdot \frac{\exp\left(-\frac{\|\mathbf{y}\|^2}{2\sigma^2}\right)}{\sigma^n \cdot \sqrt{n!}} \cdot \langle\mathbf{x}, \mathbf{y}\rangle^n$$

Let

$$c_{\sigma,n}(\mathbf{x}) = \frac{\exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right)}{\sigma^n \cdot \sqrt{n!}}, \quad f_{n,\mathbf{k}}(\mathbf{x}) = \binom{n}{k_1, k_2, \ldots, k_d}^{\frac{1}{2}}\prod_{i=1}^{d}x_i^{k_i}$$

then,

$$K(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^{+\infty}\sum_{k_1+k_2+\ldots+k_d=n} c_{\sigma,n}(\mathbf{x}) f_{n,\mathbf{k}}(\mathbf{x}) \cdot c_{\sigma,n}(\mathbf{y}) f_{n,\mathbf{k}}(\mathbf{y})$$

$$= \langle\Phi(\mathbf{x}), \Phi(\mathbf{y})\rangle$$

where $\Phi(\mathbf{x})_{\sigma,n,\mathbf{k}} = c_{\sigma,n}(\mathbf{x}) f_{n,\mathbf{k}}(\mathbf{x})$.

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \left\langle \Phi\left(\mathbf{x}_i\right), \Phi\left(\mathbf{x}_j\right) \right\rangle$$

$$= \left\langle \sum_{i=1}^{n} c_i \Phi\left(\mathbf{x}_i\right), \sum_{i=1}^{n} c_i \Phi\left(\mathbf{x}_i\right) \right\rangle \geq 0$$

for any $x_1, \ldots, x_n \in \mathcal{X}$, given $n \in \mathbb{N}, c_1, \ldots, c_n \in \mathbb{R}$, i.e., $K(\mathbf{x}, \mathbf{y})$ is positive definite.

Chapter 32 Support Vector Machine

Theorem 32.1

The minimizer of

$$\arg\min_{g} E\left\{[1 - Yg(X)]_+ \mid X = x\right\}$$

is the sign of $f(x) = \log \frac{p(x)}{1-p(x)}$, i.e.,

$$\text{sgn}\left[p(x) - \frac{1}{2}\right]$$

where $\text{sgn}(\cdot)$ is the sign function.

Proof For the hinge loss function, that,

$$E\left\{[1 - Yg(X)]_+ \mid X = x\right\}$$

$$= [1 - g(x)]_+ P\left(Y = 1 \mid X = x\right) + [1 + g(x)]_+ P\left(Y = -1 \mid X = x\right)$$

$$= [1 - g(x)]_+ p(x) + [1 + g(x)]_+ [1 - p(x)]$$

$$= \begin{cases} [1 - g(x)] \, p(x), & g(x) < -1 \\ 1 + [1 - 2p(x)] \, g(x), & -1 \le g(x) \le 1 \\ [1 + g(x)] [1 - p(x)], & g(x) > 1 \end{cases}$$

When $g(x) < -1$,

$$\arg\min_{g} E\left\{[1 - Yg(X)]_+ \mid X = x\right\} = \arg\min_{g} [1 - g(x)] \, p(x) = -1$$

When $g(x) > 1$,

$$\arg\min_{g} E\left\{[1 - Yg(X)]_+ \mid X = x\right\} = \arg\min_{g} [1 + g(x)] [1 - p(x)] = 1$$

When $-1 \le g(x) \le 1$,

$$\arg\min_{g} E\left\{[1 - Yg(X)]_+ \mid X = x\right\}$$

$$= \arg\min_{g} \{1 + [1 - 2p(x)] \, g(x)\}$$

$$= \begin{cases} -1, & p(x) < \frac{1}{2} \\ 0, & p(x) = \frac{1}{2} \\ 1, & p(x) > \frac{1}{2} \end{cases}$$

Thus, for the $g(x) \in [-1, 1]$ the minimizer of $\arg\min_g E\left\{[1 - Yg(X)]_+ \mid X = x\right\}$ is the sign of $p(x) - \frac{1}{2}$, that is the sign of $f(x) = \log \frac{p(x)}{1-p(x)}$

Chapter 33 Linear Discriminant Analysis

Chapter 34 K-Nearest Neighbor

Chapter 35　Decision Tree

Bibliography

[1]   Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge, UK ; New York: Cambridge University Press, Mar. 8, 2004. 727 pp. ISBN: 978-0-521-83378-3.

[2]   Zdzislaw Brzezniak and Tomasz Zastawniak. *Basic Stochastic Processes*. Oct. 16, 1998.

[3]   Luc Devroye, Laszlo Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Corrected edition. New York: Springer, Apr. 4, 1996. 653 pp. ISBN: 978-0-387-94618-4.

[4]   Rick Durrett. *Probability: Theory and Examples*. 5th Edition. Cambridge ; New York, NY: Cambridge University Press, May 30, 2019. 430 pp. ISBN: 978-1-108-47368-2.

[5]   E. L. Lehmann. *Elements of Large-Sample Theory*. Springer texts in statistics. New York: Springer, 1999. 631 pp. ISBN: 978-0-387-98595-4.

[6]   E. L. Lehmann and George Casella. *Theory of Point Estimation*. 2nd Edition. Springer texts in statistics. New York: Springer, 1998. 589 pp. ISBN: 978-0-387-98502-2.

[7]   P. McCullagh and John A. Nelder. *Generalized Linear Models*. 2nd Edition. Boca Raton: Chapman and Hall/CRC, Aug. 1, 1989. 532 pp. ISBN: 978-0-412-31760-6.